

The Architecture of Open Source Applications

The Architecture of Open Source Applications

Elegance, Evolution, and a Few Fearless Hacks

Edited by Amy Brown & Greg Wilson

オープンソースアプリケーションのアーキテクチャ

編集:Amy Brown and Greg Wilson

翻訳:Arai Kunimitsu and TAKAGI Masahiro

この日本語訳は Creative Commons 表示 3.0 非移植ライセンス (CC BY 3.0) のもとで公開します。あなたは以下の条件に従う限り、自由に

- 本作品を複製、頒布、展示、実演することができます。
- 二次的著作物を作成することができます。
- 本作品を営利目的で利用することができます。

あなたが従うべき条件は以下の通りです。

- 表示—あなたは原著作者のクレジットを表示しなければなりません。

以下のような理解に基づいています。

- 放棄—この作品について著作権者等の権利者から別途許可を得た場合は、上記の許諾条件は適用されません。
- パブリック・ドメイン—作品やその要素が、適用される法律の下でパブリックドメインに属する場合、その状態がこのライセンスによって影響されることはありません。
- そのほかの諸権利—ライセンスによって、以下の諸権利が影響を受けるということは全くありません。
 - あなたのフェア・ディーリングやフェア・ユースの権利、そのほか著作権の例外・制限規定
 - 著作者人格権
 - 他の人がこの作品あるいはその使われ方に関して持つ可能性のある権利、たとえばパブリシティ権やプライバシー権
- Notice—再利用や頒布にあたっては、この作品の使用許諾条件を他の人々に明らかにしなければなりません。一番よい方法は、<http://creativecommons.org/licenses/by/3.0/>へのリンクを示すことです。

このライセンスのコピーを見るには、<http://creativecommons.org/licenses/by/3.0/>を見るか、手紙を Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA. に送ってください。

本書の全文は、<http://www-aosabook.org/> でオンラインで読めます。

(英語版の) 印税はすべて、アムネスティ・インターナショナルに寄付されます。

本書に記載されている製品名や会社名は、各社の登録商標あるいは商標である可能性があります。

本書の製作にあたっては十分に注意を払いましたが、編集者や執筆者そして翻訳者は本書の内容についてなんらかの保証をするものではなく、その内容に基づくいかなる被害に関しても一切の責任を負いません。

表紙の画像は Peter Dutton が撮影した写真です。この写真は Creative Commons 表示 - 非営利 - 繙承 2.0 一般でライセンスされています。このライセンスのコピーを見るには、<http://creativecommons.org/licenses/by-nc-sa/2.0/> を見るか、手紙を Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA. に送ってください。

Revision Date: 平成 24 年 4 月 17 日

我々にすべてを教えてくれた Brian Kernighan に
そして世界中にいる良心の囚人たちに

目次

導入	xi
<i>by Amy Brown & Greg Wilson</i>	
第 1 章 Asterisk	1
<i>by Russell Bryant</i>	
第 2 章 Audacity	17
<i>by James Crook</i>	
第 3 章 The Bourne-Again Shell	35
<i>by Chet Ramey</i>	
第 4 章 Berkeley DB	53
<i>by Margo Seltzer and Keith Bostic</i>	
第 5 章 CMake	79
<i>by Bill Hoffman and Kenneth Martin</i>	
第 6 章 繼続的インテグレーション	91
<i>by C. Titus Brown and Rosangela Canino-Koning</i>	
第 7 章 Eclipse	107
<i>by Kim Moir</i>	
第 8 章 Graphite	133
<i>by Chris Davis</i>	
第 9 章 The Hadoop Distributed File System	147
<i>by Robert Chansler, Hairong Kuang, Sanjay Radia, Konstantin Shvachko, and Suresh Srinivas</i>	
第 10 章 Jitsi	163

by Emil Iovov

第 11 章 LLVM	181
<i>by Chris Lattner</i>	
第 12 章 Mercurial	199
<i>by Dirkjan Ochtman</i>	
第 13 章 NoSQL を取り巻く世界	217
<i>by Adam Marcus</i>	
第 14 章 Python Packaging	245
<i>by Tarek Ziadé</i>	
第 15 章 Riak and Erlang/OTP	271
<i>by Francesco Cesarini, Andy Gross, and Justin Sheehy</i>	
第 16 章 Selenium WebDriver	291
<i>by Simon Stewart</i>	
第 17 章 Sendmail	323
<i>by Eric Allman</i>	
第 18 章 SnowFlock	351
<i>by Roy Bryant and Andrés Lagar-Cavilla</i>	
第 19 章 SocialCalc	365
<i>by Audrey Tang</i>	
第 20 章 Telepathy	389
<i>by Danielle Madeley</i>	
第 21 章 Thousand Parsec	413
<i>by Alan Laudicina and Aaron Mavrinac</i>	
第 22 章 Violet	433
<i>by Cay Horstmann</i>	
第 23 章 VisTrails	449
<i>by Juliana Freire, David Koop, Emanuele Santos, Carlos Scheidegger, Claudio Silva, and Huy T. Vo</i>	

第 24 章 VTK

471

by Berk Geveci and Will Schroeder

第 25 章 Battle for Wesnoth

489

by Richard Shimooka and David White

導入

Amy Brown & Greg Wilson

大工仕事は非常に奥の深いものであり、人はみな、上達するための方法を一生涯かけて学び続けることになる。しかし、大工仕事と建築様式は異なる。ピッチ板や留め継ぎの世界から一歩離れて見渡せば、建造物全体を見た設計が必要になる。そしてそれは、技術的・科学的であるのと同程度に芸術の要素もある。

プログラミングもまた奥の深い作業であり、人はみな、上達するための方法を一生涯かけて学び続けることになる。しかし、プログラミングとソフトウェアアーキテクチャは異なる。多くのプログラマは、何年もかけて大規模な設計の問題に取り組む。「このアプリケーションを拡張可能にすべきだろうか?」「仮にそうだとして、その手法はどうする?スクリプトで拡張できるようにするのかプラグイン的な仕組みを取り入れるのか、あるいはまったく異なる別の方法を考える?」「クライアント側でやるべき処理とサーバー側でやるべき処理の切り分けはどうする?そもそもこのアプリケーションを“クライアント・サーバー”型で考えるのは適切なのか?」といった問題だ。これらの問いは、プログラミングに関するものではない。「階段をどこに配置するか」という問い合わせ大工仕事とは関係ないのと同じことだ。

建築様式とソフトウェアアーキテクチャには共通点も多いが、決定的な違いがひとつある。建築家はその生涯を通じて何千ものビルについて研究を重ねるが、大半のソフトウェア開発者はほんの一握りの大規模ソフトウェアしか知ることがない。しかも、その数少ないソフトウェアは自分たちが書いたものであることが多い。ソフトウェア開発者は歴史上の偉大なプログラムを振り返ることもないし、そういったプログラムの設計に関する熟練者の批評を読むこともない。その結果、先人の成功例を参考にすることもできずに同じ過ちを繰り返す。

そんな状況をどうにかしたいと思って本書を書いた。各章では、オープンソースアプリケーションのアーキテクチャについて解説している。どのような構造になっているのか、各パートがどのように絡み合っているのか、なぜその方式を採用したのか、他の設計上の問題に適用できそうな教訓は何か、といった内容だ。執筆者はそのソフトウェアをもっともよく知る人たちで、何年あるいは何十年もの間、複雑なアプリケーションの設計を経験してきた。本書ではさまざまなアプリケーションを取り上げる。シンプルなドロツールやウェブベースの表計算ソフトもあれば、コンパイラツールキットや数百万行規模の視覚化パッケージもある。数年前に生まれたばかりのアプリケーションもあれば、30周年を迎えるアプリケーショ

ンもある。すべてのアプリケーションに共通しているのは、作者が長い時間をかけて真剣に設計を考えたこと。そしてその考えを皆で分かち合いたいと考えているということだ。きっと読者のみなさんにも楽しんでもらえるだろう。

執筆者

Eric P. Allman (Sendmail): Eric Allman は sendmail や syslog そして trek の原作者であり、Sendmail, Inc. の共同創業者でもある。彼がオープンソースソフトウェアを書き始めたころにはまだ“オープンソース”などという名前はついておらず、ましてや今のような“ブーム”にはなっていなかった。彼は ACM Queue Editorial Review Board および Cal Performances Board of Trustees のメンバーである。個人サイトは <http://www.neophilic.com/~eric> だ。

Keith Bostic (Berkeley DB): Keith はカリフォルニア大学バークレー校の Computer Systems Research Group のメンバーだった。そこで 2.10BSD リリースのアーキテクトや 4.4BSD および関連リリースの開発リーダーをつとめた。彼は USENIX Lifetime Achievement Award (“The Flame”) を受賞した。これは Unix コミュニティへの並はずれた貢献を認められたものだ。また、カリフォルニア大学バークレー校から Distinguished Achievement Award も受賞している。これは 4BSD リリースをオープンソースにしたことに対するものだ。Keith は Berkeley DB のアーキテクトかつ開発者の一員だった。Berkeley DB は、オープンソースの組み込みデータベースシステムである。

Amy Brown (編集担当): Amy はウォータールー大学で数学の学士号を取得し、ソフトウェア業界で 10 年の勤務経験を持つ。現在は、書籍の執筆や編集に携わりつつ時にはソフトウェアも書く。彼女は歌手でもあり、他の人のライブを仕切ったりもする - プロもいれば趣味の人もいる。

C. Titus Brown (Continuous Integration): Titus は、進化的モデリングや物理気象学、発生生物学、ゲノミクス、そしてバイオインフォマティクスを研究している。現在はミシガン州立大学の准教授であり、科学的ソフトウェアの再現性や保守性にまで興味の範囲を広げている。彼は Python Software Foundation のメンバーでもあり、ブログは <http://ivory.idyll.org> にある。

Roy Bryant (Snowflock): ソフトウェアアーキテクトおよび CTO として 20 年の経験を持つ Roy は、Electronics Workbench(現在の National Instruments' Multisim) や Linkwalker Data Pipeline といったシステムを設計した。Linkwalker Data Pipeline は、Microsoft's worldwide Winning Customer Award for High-Performance Computing を 2006 年に受賞した。最後に在籍したスタートアップを売却した彼はトロント大学に戻り、大学院でコンピュータサイエンスを研究している。専門は、ビジュアライゼーションとクラウドコンピューティングだ。最近は、ACM の Eurosys Conference in 2011 で Snowflock 用の Kaleidoscope 拡張について発表した。個人サイトは <http://www.roybryant.net/> である。

Russell Bryant (Asterisk): Russell は Digium, Inc. のオープンソースソフトウェアチームでエンジニアリングマネージャーを務めている。また、2004 年の秋から Asterisk 開発チームのコ

アメンバーとして活動している。これまでに、Asterisk の開発におけるほぼすべての分野に貢献をしてきた。プロジェクトの運営からアーキテクチャ設計、そして開発まで。彼のブログは <http://www.russellbryant.net> である。

Rosangela Canino-Koning (Continuous Integration): ソフトウェア業界の最前線での 13 年間を経て Rosangela は大学に戻り、ミシガン州立大学でコンピュータサイエンスと進化生物学の博士号取得を目指している。空き時間には読書やハイキング、旅行などを楽しむほか、オープンソースのバイオインフォマティクスソフトウェアをハックすることもある。彼女のブログは <http://www voidptr net> である。

Francesco Cesarini (Riak): Francesco Cesarini が Erlang を常用し始めたのは 1995 年のことだった。その後も Ericsson でさまざまなプロジェクトに参加し、OTP R1 リリースにもかかわっている。彼は Erlang Solutions の創設者であり、O'Reilly の *Erlang Programming* の共著者でもある。現在は Erlang Solutions の技術ディレクターとして働いているが、イギリスのオックスフォード大学やスウェーデンのヨーテボリ大学で学生や院生を教えることもある。

Robert Chansler (HDFS): Robert は Yahoo! に在籍するソフトウェア開発のシニアマネージャーである。カーネギーメロン大学の大学院で分散システムを研究した彼はその後、コンパイラ (Tartan Labs)、印刷・画像処理システム (Adobe Systems)、電子商取引 (Adobe Systems, Impresse)、SAN 管理 (SanNavigator, McDATA) などにかかわった。分散システムや HDFS の世界に戻ってきた彼は、解決すべき課題が以前とあまり変わっていないことに気付いた。しかし、登場する数値はどれもみな、ゼロが 2 つか 3 つ多くなっていた。

James Crook (Audacity): James はアイルランドのダブリンに住むソフトウェア開発者。現在は電子工学設計用のツールにかかわっているが、かつてはバイオインフォマティクスソフトウェアを開発していたこともある。彼は Audacity に関する多くの野望を抱えており、その中のいくつかだけでも日の目を見ることを望んでいる。

Chris Davis (Graphite): Chris はソフトウェアコンサルタントであり、Google のエンジニアとしてスケーラブルな監視・自動化ツールの設計と構築に 12 年以上携わっている。Chris が Graphite を書き始めたのは 2006 年で、それ以降ずっとこのプロジェクトを率いている。コードを書いていないときの彼は、料理や作曲そして研究などをしている。彼の研究分野は、知識モデリングや群論、情報理論、カオス理論、そして複雑系などだ。

Juliana Freire (VisTrails): Juliana は、ユタ大学のコンピュータサイエンスの准教授である。それ以前には、ベル研究所 (ルーセント・テクノロジーズ) のデータベースシステム研究部門に在籍したりオレゴン健康科学大学/オレゴン科学技術大学院大学に準教授として在籍したりしていた。彼女の研究分野は、起源や科学データ管理、情報統合、そしてウェブマイニングなどだ。彼女は NSF CAREER および IBM Faculty Award を受賞している。また、彼女の研究に対して国立科学財團やエネルギー省、国立衛生研究所、そして IBM や Microsoft、Yahoo! が資金提供している。

Berk Geveci (VTK): Berk は、Kitware で科学計算のリーダーを務めている。彼は ParaView の開発リーダーでもある。ParaView は、VTK をベースとした視覚化アプリケーションであ

る。彼の研究分野は、大規模なパラレルコンピューティングや計算力学、有限要素、そして視覚化アルゴリズムだ。

Andy Gross (Riak): Andy Gross は Basho Technologies のアーキテクト長であり、Basho のオープンソースおよびエンタープライズデータストレージシステムの設計と開発を仕切っている。Andy が Basho を立ち上げたのは 2007 年 12 月。10 年におよぶソフトウェア開発や分散システムエンジニアリングの経験を経た後のことだった。Basho の前に Andy は、分散システムエンジニアリングの上級技術者として Mochi Media や Apple, Inc.、Akamai Technologies などに勤めていた。

Bill Hoffman (CMake): Bill は Kitware, Inc. の CTO を務める共同創業者である。彼は CMake プロジェクトの主要な開発者であり、大規模な C++ システムに 20 年以上携わってきた経験を持つ。

Cay Horstmann (Violet): Cay はサンノゼ州立大学でコンピュータサイエンスの教授を務めるが、ショットチゅう休暇をとっては業界で働いていたり外国で教えていたりする。プログラミング言語やソフトウェア設計に関する多くの著作があり、オープンソースの Violet や GridWorld の原作者でもある。

Emil Ivov (Jitsi): Emil は Jitsi プロジェクト（かつては SIP Communicator と呼ばれていた）の創設者であり、プロジェクトを率いている。彼は、ice4j.org や JAIN SIP プロジェクトなど他の場所でも活躍している。Emil は 2008 年初めにストラスブル大学で博士号を取得した。それ以降は、Jitsi 関連の活動に重点を置いている。

David Koop (VisTrails): David はユタ大学のコンピュータサイエンスの博士候補（2011 年夏に修了予定）。彼の研究分野は、視覚化や起源そして科学データ管理だ。彼は VisTrails システムのリード開発者であり、VisTrails, Inc. の上級ソフトウェアアーキテクトである。

Hairong Kuang (HDFS) は、貢献者およびコミッターとして長期にわたって Hadoop プロジェクトに長年かかわってきた。かつては Yahoo! で、そして現在は Facebook で働いている。業界で働くようになる前は、彼女はカリフォルニア州立工科大学ポモナ校の准教授だった。カリフォルニア大学アーバイン校でコンピュータサイエンスの博士号を取得している。彼女の研究分野は、クラウドコンピューティングやモバイルエージェント、パラレルコンピューティング、そして分散システムである。

H. Andrés Lagar-Cavilla (Snowflock): Andrés はソフトウェアシステムの研究者で、視覚化やオペレーティングシステム、セキュリティ、クラスタコンピューティング、モバイルコンピューティングなどを対象としている。学士号はアルゼンチンで、そしてコンピュータサイエンスの修士号と博士号はトロント大学で取得した。オンラインでは <http://lagarcavilla.org> で活動している。

Chris Lattner (LLVM): Chris はソフトウェア開発者で、幅広い分野の経験を持つ。コンパイラツール群やオペレーティングシステム、そしてグラフィックや画像レンダリングが得意分野だ。彼は、オープンソースの LLVM プロジェクトの設計者でありリードアーキテクトである。Chris や彼のプロジェクトに関する詳細な情報は <http://nondot.org/~sabre/> で得られる。

Alan Laodicina (Thousand Parsec): Alan はウェイン州立大学の修士課程の学生で、分散コン

ピューティングを学んでいる。空き時間には、コードを書いたりプログラミング言語を学んだり、あるいはポーカーをプレイしたりする。詳細な情報は <http://alanp.ca/> で得られる。

Danielle Madeley (Telepathy): Danielle はオーストラリアのソフトウェアエンジニアで、Colabora Ltd. で Telepathy その他の開発にかかわっている。彼女は電子工学とコンピュータサイエンスの学士号を持っており、Plush Penguin を収集している。ブログは <http://blogs.gnome.org/danni/> である。

Adam Marcus (NoSQL): Adam は博士課程の学生で、データベースシステムとソーシャルコンピューティングの共通部分を MIT コンピュータ科学・人工知能研究所で研究している。最近の研究内容は、伝統的なデータベースシステムと Twitter のようなソーシャルストリーム・Mechanical Turk のようなヒューマンコンピューティング環境との関係だ。研究用のプロトタイプを便利なオープンソースシステムに仕上げるのが好き。オープンソースのストレージシステムを追いかけているほうがビーチを歩くよりも好き。ブログは <http://blog.marcua.net> である。

Kenneth Martin (CMake): Ken は現在 Kitware, Inc. の会長と CFO を務める。Kitware, Inc. は米国に基盤をおく研究開発会社である。彼は Kitware を 1998 年に立ち上げた共同創業者であり、会社を現在のポジションに引き上げるのに貢献した。今や同社は一流の R&D プロバイダであり、政府機関や商業関係などさまざまな分野にまたがるクライアントを抱えている。

Aaron Mavrinac (Thousand Parsec): Aaron は電子工学とコンピュータ工学をウィンザー大学で学ぶ博士候補で、カメラネットワークやコンピュータビジョン、そしてロボット工学を研究している。空き時間には、Thousand Parsec やその他のフリーソフトウェアに関する活動をしたり Python や C のコードを書いたりその他さまざまことに手を出している。彼のウェブサイトは <http://www.mavrinac.com> である。

Kim Moir (Eclipse): Kim はオタワにある IBM Rational Software の研究所で Eclipse や Equinox プロジェクトのリリースエンジニアリングを率いる。また Eclipse Architecture Council のメンバーでもある。彼女が興味を持っている分野は、ビルドの最適化や Equinox そしてコンポーネントベースのソフトウェアを作ることだ。オフのときにはランニング仲間と道路を走り、次のロードレースに備えている。彼女のブログは <http://relengofthererds.blogspot.com/> である。

Dirkjan Ochtman (Mercurial): Dirkjan は 2010 年にコンピュータサイエンスの修士課程を修了した。金融関係のスタートアップ企業での勤務経験は 3 年になる。自由な時間ができると、Mercurial や Python、Gentoo Linux そして Python の CouchDB ライブラリをハックする。彼はアムステルダムの美しい都市に住んでいる。個人サイトは <http://dirkjan.ochtman.nl/> である。

Sanjay Radia (HDFS): Sanjay は Yahoo! で Hadoop プロジェクトのアーキテクトを務める。Hadoop のコミッターであり、Apache Software Foundation の Project Management Committee のメンバーでもある。かつては Cassatt や Sun Microsystems そして INRIA に勤務していた経験もあり、分散システムやグリッドコンピューティング基盤の開発に携わっていた。Sanjay は、カナダのウォータールー大学でコンピュータサイエンスの博士号を取得している。

Chet Ramey (Bash): Chet は 20 年以上 bash にかかわっており、過去 17 年はメイン開発者だった。オハイオ州クリーブランドにあるケース・ウェスタン・リザーブ大学の永年勤続者である彼は、学士号と修士号もそこで取得した。クリーブランド近郊に家族やペットとともに住み、オンラインでは <http://tiswww.cwru.edu/~chet> にいる。

Emanuele Santos (VisTrails): Emanuele はユタ大学で研究をする科学者で、研究分野は科学データ管理や視覚化、起源である。彼女は 2010 年に、ユタ大学でコンピューティングの博士号を取得している。彼女は VisTrails システムのリード開発者でもある。

Carlos Scheidegger (VisTrails): Carlos はユタ大学でコンピューティングの博士号を取得し、今は AT&T Labs の研究部門で研究者として働いている。2007 年の IEEE Visualization と 2008 年の Shape Modeling International では最優秀論文に選ばれた。彼の研究分野は、データの視覚化と解析やジオメトリ処理、そしてコンピュータグラフィックスだ。

Will Schroeder (VTK): Will は Kitware, Inc. の社長で共同創業者でもある。コンピュータサイエンスの教育を受けており、VTK の主要な開発者のひとりだ。彼は美しいコードを書くことを好む。特に計算幾何学やグラフィックに関するコードでは。

Margo Seltzer (Berkeley DB): Margo はハーバード工学・応用科学大学院でコンピュータサイエンスの教授を務め、Oracle Corporation でアーキテクトとしても働いている。彼女は Berkeley DB の設計者のひとりであり、Sleepycat Software の共同創業者でもある。彼女の研究分野は、ファイルシステムやデータベースシステム、トランザクションシステム、そして医療データマイニングである。研究者としての顔は <http://www.eecs.harvard.edu/~margo> で見られ、ブログは <http://mis-misinformation.blogspot.com/> にある。

Justin Sheehy (Riak): Justin は Basho Technologies の CTO。同社は Webmachine や Riak の制作にかかわっている。Basho の前職は、MITRE Corporation の科学者そして Akamai のシステム基盤担当シニアアーキテクトだった。両社で彼が力を注いでいたのは堅牢な分散システムに関するさまざまな内容だった。スケジューリングのアルゴリズムや言語ベースの形式モデル、そして弹性などが含まれる。

Richard Shimooka (Battle for Wesnoth): Richard は、オンタリオ州キングストンにあるクイーンズ大学の Defence Management Studies Program で研究員を務めている。彼はまた、Battle for Wesnoth の管理者代理かつ長官でもある。Richard の著作には、ソーシャルグループ(政府からオープンソースプロジェクトまでの幅広いもの)の組織文化に関して調査したものがいくつかある。

Konstantin V. Shvachko (HDFS) はベテランの HDFS 開発者で、eBay の Hadoop アーキテクトのリーダーである。Konstantin の専門分野は、大規模な分散ストレージシステムのための効率的なデータ構造やアルゴリズムである。彼は平衡木の新しい方式である S-tree を考案した。これは構造化されていないデータの索引付けに最適化されたものだ。また彼は、S-tree ベースの Linux ファイルシステムである treeFS の初期の開発者だった。treeFS は、後の reiserFS の原型となった。Konstantin は、ロシアのモスクワ大学でコンピュータサイエンスの博士号を取得している。また、Apache Hadoop の Project Management Committee のメンバーでもある。

Claudio Silva (VisTrails): Claudio は、ユタ大学のコンピュータサイエンスの正教授である。

彼の研究分野は、視覚化や幾何学的コンピューティング、コンピュータグラフィックス、そして科学データ管理などだ。彼は 1996 年に、ニューヨーク州立大学ストーンブルック校でコンピュータサイエンスの博士号を取得した。2011 年後半には、ニューヨーク大学ポリテクニック研究室にコンピュータサイエンスおよびエンジニアリングの正教授として合流する予定だ。

Suresh Srinivas (HDFS): Suresh は、Yahoo! のソフトウェアアーキテクトとして HDFS にかかわっている。彼は Hadoop のコミッターであり、Apache Software Foundation の PMC のメンバーでもある。Yahoo! の前には Sylantro Systems で働いており、コミュニケーションサービスのホスティングのためのスケーラブルな基盤を開発していた。Suresh は、インドのカルナタカにあるナショナル工科大学でエレクトロニクスと通信の学位を取得した。

Simon Stewart (Selenium): Simon はロンドン在住で、Google のソフトウェアテストエンジニアとして働いている。彼は Selenium プロジェクトの主要な貢献者である。WebDriver の作者でもある彼は、オープンソースにほれ込んでいる。Simon が好きなのはビールを楽しむこととソフトウェアを書くことで、ときにはそれらを同時にすることもある。個人ホームページは <http://www.pubbitch.org/> である。

Audrey Tang (SocialCalc): Audrey は、台湾在住のプログラマーであり翻訳家でもある。現在の勤務先は Socialtext で、彼女のそこでの役職は “Untitled Page” だ。また、Apple のローカライズやリリースエンジニアリングも請け負っている。彼女はかつて Pugs プロジェクトを率いていた。これは実際に動作する Perl 6 の初めての実装だった。また、CPAN や Hackage にも多大な貢献をしている。彼女のブログは <http://pugs.blogs.com/audreyt/> である。

Huy T. Vo (VisTrails): Huy は、2011 年 5 月にユタ大学で博士号を取得した。彼の研究分野は、視覚化やデータフローアーキテクチャ、そして科学データ管理などである。VisTrails, Inc. で上級開発者として働く。彼はまた、ニューヨーク大学ポリテクニック研究室で博士研究員となることも決まっている。

David White (Battle for Wesnoth): David は、Battle for Wesnoth の創設者でありリード開発者である。David はこれまでにもいくつかのオープンソースビデオゲームプロジェクトにかかわってきた。共同で立ち上げた Frogatto もそのひとつだ。彼は Sabre Holdings のパフォーマンスエンジニアであり、旅行技術のリーダーでもある。

Greg Wilson (編集担当): Greg は過去 25 年にわたって高性能科学計算やデータの視覚化、コンピュータセキュリティなどにかかわってきた。数冊のコンピュータ関連書籍(2008 年の Jolt Award を受賞した *Beautiful Code* など)に著者あるいは編集者としてかかわっており、こども向けの本も二冊出版している。Greg は 1993 年にエジンバラ大学でコンピュータサイエンスの博士号を取得した。

Tarek Ziadé (Python Packaging): Tarek はフランスのブルゴーニュに住む。Mozilla の上級ソフトウェアエンジニアであり、サーバーを Python で構築している。空き時間には、Python のパッケージングを率いている。

謝辞

レビューのみなさんに感謝する。

Eric Aderhold	Muhammad Ali	Lillian Angel
Robert Beghian	Taavi Burns	Luis Pedro Coelho
David Cooper	Mauricio de Simone	Jonathan Deber
Patrick Dubroy	Igor Foox	Alecia Fowler
Marcus Hanwell	Johan Harjono	Vivek Lakshmanan
Greg Lapouchnian	Laurie MacDougall Sookraj	Josh McCarthy
Jason Montojo	Colin Morris	Christian Muise
Victor Ng	Nikita Pchelin	Andrew Petersen
Andrey Petrov	Tom Plaskon	Pascal Rapicault
Todd Ritchie	Samar Sabie	Misa Sakamoto
David Scannell	Clara Severino	Tim Smith
Kyle Spaans	Sana Tapal	Tony Targonski
Miles Thibault	David Wright	Tina Yee

また、編集の初期段階での助けとなった Jackie Carter にも感謝する。

貢献

何十人のボランティアのおかげで本書を作ることができたが、まだやり残したことは多い。間違いの指摘、他の言語への翻訳、他のオープンソースプロジェクトのアーキテクチャに関する記述の追加などを歓迎する。協力してくれる場合は aosa@aosabook.org まで連絡してほしい。

Asterisk

Russell Bryant

Asterisk¹は GPLv2 で配布されているオープンソースによる電話通信のプラットフォームである。簡単に言うと、電話を掛けたり、受けたり、カスタム処理を実行したりするためのサーバアプリケーションである。

このプロジェクトは 1999 年に Mark Spencer によって始められた。Mark は自らが興した Linux Support Services 社において、電話システムが必要だったのだが、購入するだけの費用が無かったために自作した。Asterisk の人気が上がると、Asterisk に資源を集中するため社名を Digium, Inc に改めた。

Asterisk の名前の由来は、UNIX におけるワイルドカード文字*にある。Asterisk プロジェクトの目標はすべての電話テクノロジを実装する事である。この目標に向けて、Asterisk は、電話を掛けたり、受けたりするため多くのテクノロジをサポートしている。これらのテクノロジには、従来のアナログおよびデジタルの電話ネットワークである PSTN(Public Switched Telephone Network) はもちろん、多くの VoIP(Voice over IP) プロトコルも含まれる。このような異なる種類の電話テクノロジを実装していること、異なる電話テクノロジ同士を接続できる事が Asterisk の主な強みである。

Asterisk システムに電話が掛かってきたり、Asterisk から電話を掛けたりするとき、通話処理をカスタマイズするのに使える多くの追加機能がある。ボイスメールのような完成されたアプリケーションもある一方、音声ファイルを流したり、数字を読んだり、音声認識といった、組み合せて使うことでカスタムの音声アプリケーションを構築するのに使えるようなり小規模の機能群もある。

1.1 重要なアーキテクチャ・コンセプト

このセクションでは、Asterisk 全体に影響するアーキテクチャのコンセプトについて述べる。ここで述べる考え方は、Asterisk アーキテクチャの根幹を構成する。

¹<http://www.asterisk.org/>

チャネル

Asteriskにおけるチャネルは、Asteriskシステムと電話端末の接続をあらわす(図1.1)。最も単純な例は、電話機がAsteriskシステムを呼び出す時である。この場合の接続は、一つのチャネルであらわされる。Asteriskのコードでは、チャネルは`ast_channel`構造体のインスタンスとして存在する。この呼び出しシナリオは、発信者がボイスメールを使うときの例である。



図1.1: 片方向通話、单一チャネル

チャネル・ブリッジ

おそらく、より馴染のある通話シナリオは電話機同士の接続であろう。このシナリオでは、電話機Aを使っている人が電話機Bを呼び出す。すなわち2台の電話端末がAsteriskシステムに接続しているため、二つのチャネルが存在する(図1.2)。



図1.2: 2つのチャネルによって表される双方向通話

このようにAsteriskチャネルが接続される事をチャネル・ブリッジと呼ぶ。チャネル・ブリッジは、チャネル同士をつなげて、メディアを流す事を目的としている。流れるメディアは音声が主になるが、ビデオやテキストも可能であるし、ひとつ以上のメディア（音声とビデオなど）のこともある。このような場合も、Asteriskでは、一つのチャネルとして扱われ

る。図 1.2 では、2つのチャネルがそれぞれ、電話 A、B に接続されている。このブリッジは、電話機 A から電話機 B へのメディア・ストリームおよび電話機 B から電話機 A へのメディア・ストリームを通すことに責任を持つ。すべてのメディア・ストリームは、Asterisk と交渉されるため、Asterisk が理解できない、またはフルにコントロールできないメディア・ストリームは許可されない。これは、Asterisk において録音や音声処理、異なるテクノロジ間の変換が可能であることを意味する。

2つのチャネルをブリッジにより接続する形式は2通りある。ジェネリック・ブリッジとネイティブ・ブリッジである。ジェネリック・ブリッジは、チャネル・テクノロジに何が使われているか動作する。すべての音声とシグナリングは、Asterisk 抽象チャネル・インターフェースを通してやりとりされる。これは、もっとも柔軟性の高い接続であるが、実現するために高いレベルの抽象化が必要であり、効率性を犠牲にしている。図 1.2 は、ジェネリック・ブリッジを表現している。

ネイティブ・ブリッジは、テクノロジ固有のチャネル同士を接続するときに使われる手法である。2つのチャネルが、同じメディア・トランSPORT・テクノロジを使って Asterisk に接続されるときには、異なるテクノロジ同士を接続する時に使用する Asterisk 抽象レイヤを通しての接続よりも、より効果的な接続方法がある。例えば、電話ネットワークに接続するのに特別のハードウェアが使われるときは、アプリケーションを全く通らずにチャネル同士をハードウェア上でブリッジする事が可能である。いくつかの VoIP プロトコルの場合には、呼制御信号の情報はサーバを通して流れるが、メディア・ストリームはエンドポイント間で相互に直接送受信させることも可能である。

ジェネリック・ブリッジかネイティブ・ブリッジかの選択は、ブリッジが必要になったときに、2つのチャネルの比較で決定される。二つのチャネルが同じネイティブ・ブリッジ技術をサポートする事がわかると、ネイティブ・ブリッジが使われる。その他の場合は、ジェネリック・ブリッジが使われる。2つのチャネルが同じネイティブ・ブリッジをサポートするかどうかの決定は、単純に C 言語の関数ポインタの比較で行われる。これは、もっともエレガントな手法というわけではないが、この方法が我々のニーズを満たさなかつたことは今のところない。ネイティブ・ブリッジ機能については、1.2 節で詳しく説明する。図 1.3 は、ネイティブ・ブリッジの例を表現している。

フレーム

通話中のコミュニケーションは、Asterisk のコードではフレームで表現されている。これは、`ast_frame` 構造体のインスタンスである。フレームはメディア・フレームにもシグナリング・フレームにも使われる。音声メディア・フレームのストリームは基本的にシステムを通過する。シグナリング・フレームは、押された番号や保留中、呼切断などの呼制御イベントを送信するのに使われる。

利用可能なフレーム・タイプのリストは静的に定義される。フレーム・タイプは、タイプおよびサブタイプで記されて、数値の形式で保存される。フレーム・タイプの全リストはソー



図 1.3: ネイティブ・ブリッジの例

スコードの `include/asterisk/frame.h` にある。代表的な例を以下に示す。

- VOICE: オーディオ・ストリーム・フレーム
- VIDEO: ビデオ・ストリーム・フレーム
- MODEM: IP 上で FAX を送信するための T.38 のようなデータに対する符号化を表わす。このフレーム・タイプの主な使用対象は FAX の処理である。この信号が相手側で正しく復号できるように、フレームのデータが絶対に変更されないことが重要である。これは、AUDIO フレームが帯域幅を稼ぐために音声品質を犠牲にしてコーデックを変換することが許されるという点で異なる。
- CONTROL: 呼制御メッセージを表わす。このフレームは、呼制御イベントを示すのに使われる。これらのイベントには、応答や切断、保留などがある。
- DTMF_BEGIN: 番号の始まり。このフレームは、発信者が DTMF キー²を押したとき送信される。
- DTMF_END: 番号の終わり。このフレームは、発信者が DTMF キーを押し終わったときに送信される。

1.2 Asterisk 抽象コンポーネント

Asterisk は高度にモジュール化が進んだアプリケーションである。ソースツリー中の `main/` ディレクトリ以下にコア・アプリケーションがある。しかし、これ自体は、大変有益というわけではない。コア・アプリケーションは、基本的にモジュール・レジストリとして振る舞う。通話をうまく動かすための抽象インターフェースとの接続用のコードも存在する。これらインターフェースの具体的な実装は実行時にローダブルモジュールによって登録される。

²DTMF は Dual-Tone Multi-Frequency を表わす。これは電話機のキーを押したときにオーディオの形式で送信されるトーン信号である。

デフォルトでは、メイン・アプリケーションが起動するときに、ファイルシステム上のあらかじめ定められた場所にある全ての Asterisk モジュールがロードされる。このアプローチは、簡素化のために導入された。しかし、はつきりとロードするモジュールをどの順番でロードするかを指定することができるコンフィグレーション・ファイルもある。これは、コンフィグレーションを少し複雑にするが、不要なモジュールをロードしないように指定できる能力を提供する。このことは、アプリケーションのメモリ・フットプリント削減が主な利点であるが、セキュリティ面での利点もいくつかある。ネットワーク接続に必要なモジュール以外は、ロードしないのがベストである。

モジュールはロードされるときに Asterisk コア・アプリケーションに抽象コンポーネントの実装を登録する。モジュールが Asterisk コアに対して実装および登録できるインターフェースはたくさんある。モジュールは自身が登録したいインターフェースは種別が異なっていれば全て登録することができる。一般的に、関連する機能は一つのモジュールにまとめられる。

チャネル・ドライバ

Asterisk チャネル・ドライバのインターフェースは、最も複雑かつ重要である。Asterisk チャネル API は、電話プロトコルの抽象化を提供する。抽象化により使用する電話プロトコルと独立して Asterisk 機能が動けるようになる。このコンポーネントは、Asterisk 抽象チャネルと実行される電話テクノロジの細部の間を通訳する責務を持つ。

Asterisk チャネル・ドライバ・インターフェースの定義は `ast_channel_tech` インタフェースと呼ばれる。これは、チャネル・ドライバによって実装されなければならない一連のメソッドを定義する。チャネル・ドライバが実装しなければならない最初のメソッドは、`ast_channel` ファクトリ・メソッドであり、具体的には `ast_channel_tech` の中の `requester` メソッドである。Asterisk チャネルが生成されるとき、これが受話であろうが発話であろうが、要求されたチャネルの種類に対応した `ast_channel_tech` の実装は、この電話に対する `ast_channel` をインスタンス化、初期化する責務を持つ。

`ast_channel` が生成される時、生成元の `ast_channel_tech` への参照も作られる。テクノロジ固有の方法で扱われるべき多くのオペレーションがある。これらのオペレーションが `ast_channel` のなかで実行されるときには、オペレーションのハンドリングは、`ast_channel_tech` の適切なメソッドに委ねられる。図 1.2 は、2 つの Asterisk チャネルを示している。図 1.4 は、これを展開して、二つのブリッジされたチャネルとチャネル・テクノロジの実装がどのようにかみ合うのかを図示している。

`ast_channel_tech` の中で最も重要なメソッドを以下に示す。

- `requester`: このコールバックは、チャネル・ドライバにチャネルタイプに対して適切な `ast_channel` オブジェクトのインスタンス化と初期化を要求する。
- `call`: このコールバックは、`ast_channel` に示されているエンドポイントへの発信を開始するのに使われる。



図 1.4: チャネル・テクノロジと抽象チャネル・レイヤ

- **answer:** Asterisk が本 `ast_channel` への着信に応答すると決めたときに呼ばれる。
- **hangup:** システムが呼切断すると決めたときに呼ばれる。チャネル・ドライバは、通話が終わったことをプロトコル特有の方法でエンドポイントに伝える。
- **indicate:** 通話が始まると、エンドポイントに伝えるべき多くの制御イベントが発生する。例えば、デバイスが保留になると、この状態を伝えるために本コールバックが呼ばれる。通話が保留になった事を示す方法は、プロトコル毎に異なる。チャネル・ドライバは、デバイスに保留音を流し始めるだけの事もある。
- **send_digit_begin:** この関数はデバイスへの数字 (DTMF) 送信が始まった事を示すときに呼ばれる。
- **send_digit_end:** この関数はデバイスへの数字 (DTMF) 送信が終わった事を示すときに呼ばれる。
- **read:** この関数は、デバイスが送った `ast_frame` をリードするために Asterisk コアによって呼ばれる。`ast_frame` は、メディア (オーディオ、ビデオなど) や呼制御信号をカプセル化するのに Asterisk で使用される抽象化である。
- **write:** この関数は、デバイスに `ast_frame` を送信するのに使われる。チャネル・ドライバは、データを取得して電話プロトコルに対応した形に、パケット化してエンドポイントに送信する。
- **bridge:** このチャネル・タイプに対するネイティブ・ブリッジのコールバックである。ネイティブ・ブリッジは、前に述べたように、2つのチャネルが同じ種類のときに、呼制御信号やメディアが不要な抽象レイヤを流れるかわりに、チャネル・ドライバがより効率的なブリッジ方法を実行できるときに使用する。これは、性能面で大変重要である。

通話が終了すると、Asterisk コア内の抽象チャネルを扱うコードは、`ast_channel_tech hangup` コールバックを起動し `ast_channel` オブジェクトを破棄する。

ダイヤルプラン・アプリケーション

Asterisk アドミニストレータは、Asterisk ダイヤルプランを使って通話手順を設定する。ダイヤルプランは /etc/asterisk/extensions.conf に存在する。ダイヤルプランはエクステンションと呼ばれる一連の通話ルールを構成する。呼び出しがシステムに到着すると、呼び出し処理のために、ダイヤル番号を使ってダイヤルプラン内の対応するエクステンションを探査する。エクステンションは、チャネル上で実行するダイヤルプラン・アプリケーションのリストを持っている。ダイヤルプランで実行可能なアプリケーションは、アプリケーション・レジストリが保持している。このレジストリはモジュールがロードされるときに登録される。

Asterisk は 200 近いアプリケーションを持つ。アプリケーションの定義は大変ルーズである。アプリケーションはチャネルとインタラクトするために Asterisk の内部 API を自由に使うことができる。発信者にサウンド・ファイルを流す Playback のような単純なタスクを行うアプリケーションもあるし、Voicemail のようなより複雑で大規模なアプリケーションもある。

Asterisk ダイヤルプランを使って複数のアプリケーションを組み合せると、カスタムの通話処理が可能である。提供されたダイヤルプラン言語で実現できるカスタム化よりも複雑なものが必要なときには、都合の良いプログラミング言語を使ってカスタムの通話処理が可能なスクリプト・インターフェースもある。これらのスクリプトのためインターフェースを使って他のプログラム言語が使われる時でも、ダイヤルプラン・アプリケーションはチャネルと相互作用するために起動される。

次の例に入る前に、番号 1234 への通話を扱う Asterisk ダイヤルプランの文法について説明する。1234 は無作為の選択である。これは、3つのダイヤルプラン・アプリケーションを起動する。最初に、呼応答し、次に、サウンド・ファイルを再生し、最後に呼切断を行う。

```
; Define the rules for what happens when someone dials 1234.  
;  
exten => 1234,1,Answer()  
    same => n,Playback(demo-congrats)  
    same => n,Hangup()
```

exten キーワードはエクステンションを定義するのに使われる。exten 行の右側において、1234 は誰かが 1234 を呼び出したときのルールを定義している。次の 1 は、この番号にダイヤルされたときに最初に実行されるステップである事を示している。最後に、Answer はシステムが呼に応答することを指示する。次の same キーワードで始まる 2 行は、直前に指定したエクステンションと同じであることを示す。本例では、1234 に対するエクステンションである事を示している。n は、次のステップであることを示す短縮表現である。各行の最後の項目には実行する動作を指定する。

次は、Asterisk ダイヤルプランを使った別の例である。このケースは次のような流れである。着呼に対して応答する。発信者にビープ音を鳴らし、発信者からの数字を 4 桁読み、DIGITS

変数に格納される。さらに、格納された4桁の数字が、発信者に音声で通知される。最後に終話する。

```
exten => 5678,1,Answer()
    same => n,Read(DIGITS,beep,4)
    same => n,SayDigits(${DIGITS})
    same => n,Hangup()
```

前に触れたように、アプリケーションの定義は、大変ルーズである。登録された関数プロトタイプは非常に単純である：

```
int (*execute)(struct ast_channel *chan, const char *args);
```

しかしながら、アプリケーションの実装は、实际上、include/asterisk/にある全てのAPIsを使用する。

ダイヤルプラン・ファンクション

多くのダイヤルプラン・アプリケーションは、文字列型の引数をとる。決め打ちの場合もあるし、動的な振る舞いの場合には、変数も使われる。次の例は、変数を設定して、Verboseアプリケーションを使って、その値を Asterisk コマンドライン・インターフェースに表示するダイヤルプランの断片である。

```
exten => 1234,1,Set(MY_VARIABLE=foo)
    same => n,Verbose(MY_VARIABLE is ${MY_VARIABLE})
```

ダイヤルプラン・ファンクションは、前の例と同じ構文を使って起動される。Asterisk モジュールは、ダイヤルプラン・ファンクションを登録することができる。ダイヤルプラン・ファンクションは、情報を引き出してダイヤルプランに反映させることができる。ダイヤルプラン・ファンクションは、ダイヤルプランからデータを受けて実行する事もできる。一般的なルールとして、ダイヤルプラン・ファンクションは、チャネル・メタデータをセットしたり引き出したりはできるが、呼制御やメディア処理は行わない。これは、ダイヤルプラン・アプリケーションの仕事として残されている。

次は、ダイヤルプラン・ファンクションの利用例である。最初に、現在のチャネルの発信者IDを Asterisk コマンドライン・インターフェースに表示する。次に Set アプリケーションを使って発信者 ID を変更する。この例では、Verbose と Set はアプリケーションであり、CALLERID がファンクションである。

```
exten => 1234,1,Verbose(The current CallerID is ${CALLERID(num)})
    same => n,Set(CALLERID(num)=<256>555-1212)
```

ここでは、発信者番号情報が、ast_channel のインスタンスのデータ構造体に保存されているため、ダイヤルプラン・ファンクションが必要であった。ダイヤルプラン・ファンクショ

ン・コードは、これらのデータ構造体にデータを設定したり引き出したりする方法を知っている。

もうひとつのダイヤルプラン・ファンクションの例は、通話記録にカスタム情報を加える。これは、CDR(Call Detail Records) と呼ばれる。CDR ファンクションは、通話詳細記録情報の引き出しやカスタム情報の追加を可能にする。

```
exten => 555,1,Verbose(Time this call started: ${CDR(start)})  
    same => n,Set(CDR(mycustomfield)=snickerdoodle)
```

符号化方式変換

VOIPの世界では、異なる符号媒体のネットワークをまたがって送信するために、様々な種類の符号化方式が使われる。符号化方式は、メディアの品質、CPU消費量、帯域幅のトレードオフの中から選択される。Asteriskは多くの符号化方式をサポートし、必要ならば、これらの符号化方式を相互に変換可能である。

通話がセットアップされると、Asteriskは符号化方式変換が不要になるように両端の端末に共通の符号化方式を選択することを試みる。しかし、必ずしも可能というわけではない。共通の符号化方式が使われていても、符号化方式変換が使われる事もある。例えば、Asteriskは、音声がシステムを通過するときに(ボリューム調節などの)信号処理を行うように構成することも可能である。このときに、Asteriskは、信号処理を行う前に、音声を非圧縮形式に変換する必要がある。Asteriskは、通話録音も可能である。録音に指定した符号化方式が通話の符号化方式と異なるときは、符号化方式変換が必要となる。

符号化方式交渉

メディア・ストリームにどの符号化方式を使用するかの交渉は、Asteriskに通話を接続するテクノロジに依存する。従来の電話通信ネットワーク(いわゆる PSTN)上の通話のようなケースでは、交渉の余地は無いと思われる。しかし、特に IP プロトコルを使うようなケースでは、利用可能な符号化方式や優先度を示して、符号化方式を交渉する機構により、符号化方式の合意が取られる。

例えば、SIP(最も一般的な VOIP プロトコル)の場合、通話が Asterisk に届いたときに、符号化方式の交渉が行われる。

1. 端末が Asterisk に通話要求を送信するときに、使いたい符号化方式のリストも含める。
2. Asterisk は、アドミニストレータが用意した、優先度順に並んだ利用可能な符号化方式のリストを参照する。Asterisk はこのリストと端末の要求したりストから最も好ましい符号化方式を選択して応答する。

Asterisk が充分に扱えない分野のひとつに、ビデオのようなより複雑な符号化方式がある。過去 10 年で、符号化方式の交渉に対する要求は、より複雑化してきた。最新の音声符号化方式や、サポートするビデオ符号化方式の改善には、多くの作業が残っている。これは、Asterisk の次のメジャーリリースに向けた開発作業の最優先項目のひとつである。

符号化方式変換モジュールは、一つ以上の `ast_translator` インタフェースの実装を利用する。変換モジュールは、変換元と変換先の属性を持つ。また、変換元フォーマットから変換先フォーマットへのメディア・チャンクの変換に使われるコールバックも実装する。変換

モジュールは、電話の概念については、関知せず、メディアの変換のみに関わる。

符号化方式変換 API のより詳細情報は、`include/asterisk/translate.h` と `main/translate.c` にある。符号化方式変換の抽象化部の実装は、`codecs` ディレクトリにある。

1.3 スレッド

Asterisk は、マルチスレッドを多用するアプリケーションである。Asterisk は、スレッドを管理するのに POSIX スレッド API とロックなどの関連サービスを使っている。スレッドを扱うすべての Asterisk のコードは、デバッグ目的で、一連のラッパーを通してしている。Asterisk 内のほとんどのスレッドは、ネットワーク監視スレッドか、チャネルスレッド (PBX スレッドとも呼ばれる。主目的がチャネルに対する PBX 機能の実行である事に起因する) に分類される。

ネットワーク監視スレッド

ネットワーク監視スレッドは、Asterisk の主要チャネル・ドライバ毎に存在する。ネットワーク監視スレッドは、接続されたネットワーク (IP や PSTN など) を監視し、呼着信や他の着信要求を監視する。本スレッドは、コネクションの初期セットアップを行い、認証やダイヤルされた番号の検証を行う。呼のセットアップが完了すると、監視スレッドは、Asterisk チャネル (`ast_channel`) のインスタンスを生成し、チャネル・スレッドを始動させて呼の切断までの残りの制御を行わせる。

チャネル・スレッド

前に述べたように、チャネルは、Asterisk の基本概念である。チャネルは、着信にも発信にも対応する。着信チャネルは、呼が Asterisk システムに届いたときに生成される。これらのチャネルは、Asterisk ダイヤルプランを実行する。ダイヤルプランを実行する着信チャネル毎に、スレッドが生成される。これらのスレッドは、チャネル・スレッドと呼ばれる。

ダイヤルプラン・アプリケーションは、常にチャネル・スレッドのコンテキストで実行される。ダイヤルプラン・ファンクションもほとんど常に、チャネル・スレッドのコンテキストで実行される。Asterisk CLI のような非同期インターフェースからダイヤルプラン・ファンクションを読んだり書いたりする事が可能である。しかし、`ast_channel` データ構造の所有者は、常にチャネル・スレッドであり、同スレッドが `ast_channel` オブジェクトの生成・消滅もコントロールする。

1.4 通話シナリオ

前の2節では、Asteriskコンポーネントに対する重要なインターフェースとスレッド実行モデルを紹介した。本節では、複数のAsteriskコンポーネントがどのように協調して通話処理するのかを示すために、いくつかの一般的な通話シナリオに分けて示す。

ボイスメールのチェック

一例として、誰かが電話システムを呼び出して、ボイスメールをチェックする通話シナリオを示す。本シナリオでの最初の主要コンポーネントは、チャネル・ドライバである。チャネル・ドライバは、電話からの着信要求の処理に責任を持つ。これは、チャネル・ドライバの監視スレッドで行われる。呼をシステムに運ぶのに使われる電話テクノロジによっては、通話をセットアップするのに必要な交渉が行われる事もある。呼のセットアップのもう一つのステップは、通話先の決定である。これは、通常、発信者がダイヤルした番号により決定される。しかし、あるケースでは、通話の伝送テクノロジがダイヤル番号の明示をサポートしないため、番号特定ができない事もある。アナログ電話の着信が一つの例である。

ダイヤルプラン中のエクステンションにダイヤルされた番号が存在することをチャネル・ドライバが確認すると、チャネル・ドライバはAsteriskチャネル・オブジェクト(`ast_channel`)を割り当て、チャネル・スレッドを起動する。チャネル・スレッドは、残りの通話処理の責任を任せられる(図1.5)。

チャネル・スレッドのメインループは、ダイヤルプランの実行を扱う。ダイヤルされたエクステンションに対して定義されたルールを探索して、定義されたステップに従って順次実行する。次に示すエクステンションの例は、`extensions.conf`ダイヤルプラン内で。このエクステンションは、誰かが、*123にダイヤルしたときに、呼び出しに応答し、`VoicemailMain`アプリケーションを実行する。このアプリケーションは、ユーザが自分宛のメールボックスに残されたメッセージをチェックするものである。

```
exten => *123,1,Answer()
    same => n,VoicemailMain()
```

チャネル・スレッドが`Answer`アプリケーションを実行すると、Asteriskは着信に応答する。呼び出しに応答するには、テクノロジ固有の処理を必要とする。よって、いくつかの一般的な応答処理に加えて、`ast_channel_tech`構造体に関連した`answer`コールバックが応答処理のために呼ばれる。これは、IPネットワーク上に特定のパケットを送信する処理やアナログ電話でのオフフック処理などである。

次のステップでは、チャネル・スレッドは、`VoicemailMain`(図1.6)を実行する。このアプリケーションは、`app_voicemail`モジュールによって供給される。特筆すべき点は、ボイスメールのコードが通話に応対している間、ボイスメール自体は、Asteriskシステムへの呼



図 1.5: 呼設定シーケンス図

び出しを伝送するテクノロジについては何も知らないということである。Asterisk 抽象チャネルは、ボイスメールの実行処理から、これらの詳細を隠蔽している。

発信者が、自身のボイスメールにアクセスする処理には多くの機能が含まれる。しかし、基本的には、発信者からの数字キーの入力に応答して、サウンド・ファイルを読み込んだり、書き込んだりするような処理である。DTMF 信号は、多くの異なる方式で Asterisk に運ばれる。繰り返しになるが、これらの詳細方式の部分は、チャネル・ドライバで行われる。キー入力が Asterisk に届くと、共通のキー入力イベントに変換されてからボイスメールのコードに届く。

これまで述べてきた中で、Asterisk の主要インターフェースの一つは、符号化方式変換である。これらの符号化の実装は、この呼び出しシナリオでは大変重要である。ボイスメールのコードが発信者に対してサウンド・ファイルを再生したいとき、サウンド・ファイルのオーディオデータを直接再生するよりも、それを符号化して転送する方が効率的である。

デイオ形式は、Asterisk と発信者の間で使われているフォーマットと異なるときもある。オーディオ形式を変換する必要がある時、一つ以上の変換器を使って、変換元から変換先の形式への符号化変換のパスを構築する。



図 1.6: VoicemailMain への呼び出し

どこかの時点で発信者は、ボイスメール・システムとやり取りして呼切断を行う。チャネル・ドライバは、これを検出して Asterisk チャネル呼制御の共通イベントへと変換する。この呼制御イベントを受信すると、ボイスメールのコードは残り作業が無いため終了する。制御がチャネル・スレッドのメインループに戻って、ダイヤルプランの実行を続ける。この例では、これ以上のダイヤルプラン処理がないため、チャネル・ドライバはテクノロジ固有の呼切断処理を与えられる。これにより、`ast_channel` オブジェクトは破棄される。

ブリッジ・コール

Asterisk における、もう一つの、よくある通話シナリオは、二つのチャネル間のブリッジ・コールである。これは 2 者間での通話のシナリオである。呼設定処理の最初の部分は、前例と同じである。処理の違いは、通話が設定されて、チャネル・スレッドが、ダイヤルプランを実行し始めるときから生じる。

次のダイヤルプランは、ブリッジ・コールになる典型的な例である。このエクステンションを使うと、電話機が、1234 をダイヤルすると、ダイヤルプランは、Dial アプリケーションを実行する。これは、発信を開始するときのメイン・アプリケーションである。

```
exten => 1234,1,Dial(SIP/bob)
```

Dial アプリケーションへの引数は、システムに、SIP/bob への発信を促す。この引数の SIP の部分は、通話に使われるプロトコルが SIP である事を示している。bob の部分は、SIP プロトコルを実装したチャネル・ドライバ `chan_sip` によって、解釈される。チャネル・ドライバが、bob へのアカウントを正しく設定していると仮定すると、Bob の電話機への呼の伝送方法を知ることになる。

Dial アプリケーションは、SIP/bob 識別子を使って Asterisk コアに対して新しい Asterisk チャネルを割り当てる。コアは、SIP チャネル・ドライバに対してテクノロジ固有の初期化処理を行うように指示する。チャネル・ドライバは、電話の呼び出しを行う処理を起動する。リクエストが進ときに、Asterisk コアに対してイベントの発生を知らせる。これは、さらに、Dial アプリケーションまで届けられる。これらのイベントは、呼応答、ビジー、輻輳、何らかの理由での拒絶など多くの応答種別が含まれる。理想的なケースでは、呼び出しは応答される。インバウンド・チャネルを通じて呼が応答された事は伝わる。システムが、アウトバウンドコールへの応答が完了するまで、Asterisk は、この呼に対しては応答しない。両チャネルが、応答するとチャネル・ブリッジが始まる(図 1.7)。



図 1.7: ジェネリック・ブリッジ・コールにおけるブロック図

チャネル・ブリッジを行っている間、片方のチャネルからのオーディオとシグナリング・イベントは、もう一方のチャネルに通される。これは、片側からの呼切断などのブリッジの終了を示すイベントが起こるまで続けられる。図 1.8 に示したシーケンス図は、ブリッジ・コールの間オーディオ・フレームに対して実行される主な流れを表している。

通話が終わると、切断処理は前の例とほとんど同じように進められる。主な違いは、チャネルが二つあることである。チャネル・テクノロジ固有の切断処理は、チャネル・スレッドが実行を止める前に実行される。

1.5 Final Comments

Asterisk のアーキテクチャは、誕生してから 10 年を越える歳月が経っている。しかし、拡大を続ける産業においても、チャネルの基本概念と Asterisk ダイヤルプランを使った呼制御の柔軟性は、複雑な電話通信システムの開発をサポートし続けている。Asterisk のアーキテクチャで充分でない分野のひとつに、複数のサーバに渡るスケーリングがある。Asterisk 開発コミュニティは、このスケーラビリティを改善するための Asterisk SCF(Scalable Communications Framework)と呼んでいる兄弟プロジェクトを進めている。数年内には、Asterisk と Asterisk SCF が統合されて、より大規模システムへの採用が進み、電話市場における重要性を増すと考えている。



図 1.8: ブリッジ中のオーディオ・フレーム処理のシーケンス図

Audacity

James Crook

Audacity は、よく知られたサウンドレコーダー/オーディオエディタだ。多機能なプログラムでありながら、使いやすさも維持している。大半のユーザーは Windows 上で使っているが、Audacity のソースコードをコンパイルして Linux や Mac でも使える。

Dominic Mazzoni が最初に Audacity を書いたのは 1999 年のこと。当時の彼はカーネギーメロン大学の研究生で、音響処理アルゴリズムの開発やデバッグに使うプラットフォームを作ろうとしていた。その後このソフトは成長し、当初の目的以外にもいろいろな方面で役立つようになった。Audacity がオープンソースソフトウェアとして公開されると、多くの開発者がそれにひきつけられた。熱心なファンたちが参加し、Audacity の改良や保守、テスト、更新、ドキュメント作成、ユーザーのサポートなどを行うようになった。また、長い年月をかけて、ユーザーインターフェイスも他の言語に翻訳された。

Audacity のひとつの目標は、ユーザーインターフェイスを“発見可能”なものにするということだ。特にマニュアルを読まずともすぐに使い始めることができ、使っていくうちにいろいろな機能を発見できるようになることを目指している。この方針もあり、Audacity は他のソフトに比べてユーザーインターフェイスの一貫性をより重視している。多くの人がかかわるプロジェクトでは、このような統一指針が思いのほか重要となる。

Audacity のアーキテクチャにも同様の指針や発見可能性があればどんなによいことか。それに近いこととして我々が言えるのは“試して、そして合わせよ”ということだ。新しいコードを追加するとき、開発者はその周辺のコードを見てその形式や規約に合わせようとする。しかし実際のところ、Audacity のコードベースには、しっかり構成されたコードとそうでないコードが入り混じっている。全体のアーキテクチャを、小さな都市になぞらえて考えるとわかりやすい。印象的な建造物がいくつかある一方で、そのそばには荒廃した貧民街も見受けられるという具合だ。

2.1 Audacity の構造

Audacity は、いくつかのライブラリによる階層構造になっている。Audacity のコードに新しいプログラムを追加するときにはこれらのライブラリに関する詳細な知識は不要だが、ライブラリの API やその役割に親しんでおくことは大切だ。中でも最も重要なライブラリは PortAudio と wxWidgets だ。PortAudio はローレベルのオーディオインターフェイスをクロスプラットフォームで提供し、wxWidgets は GUI コンポーネントを同じくクロスプラットフォームで提供する。

Audacity のコードを読むときには、本質的に不可欠なコードは一部だけであることを知っておけば理解の助けになる。さまざまなオプション機能を提供しているのは各種のライブラリだ—ユーザーはそれをオプション機能だとは考えないかもしれないが。たとえば、組み込みのオーディオエフェクト以外に Audacity は LADSPA (Linux Audio Developer's Simple Plugin API) をサポートしており、オーディオエフェクトのプラグインを動的に読み込むことができる。Audacity の VAMP API は、オーディオ解析用のプラグインのために同じ仕組みを用意している。これらの API がなければ Audacity の機能はあまり豊富とはいえなくなるだろうが、プラグインの機能に依存しないものとなる。

Audacity が使うその他のオプションライブラリには libFLAC や libogg そして libvorbis がある。これらは、さまざまなオーディオ圧縮フォーマットに対応する機能を提供する。MP3 フォーマットへの対応は、LAME あるいは FFmpeg ライブラリを動的に読み込むことで行う。ライセンス上の制約のため、これら主要圧縮フォーマットのライブラリは Audacity 本体に組み込むことができない。

それ以外にも、ライセンスの影響が Audacity のライブラリや構造にあらわれているところがいくつかある。たとえば、VST プラグインに対応していないのはライセンスの制約のためだ。また、非常に効率的な高速フーリエ変換ライブラリである FFTW をいくつかの箇所で使っているが、これは Audacity をソースからコンパイルする人でないと使えない。バイナリ版では、多少速度の落ちる別の実装を使うことになっている。Audacity がプラグインの仕組みを受け付ける限り、Audacity では FFTW を使えない。FFTW の作者は、自分のコードを一般的なサービスとして任意のコードで使わせるということを望んでいないのだ。つまり、プラグインをサポートすると決断することで、FFTW を使えないという代償を払うことになる。プラグインをサポートしたおかげで LADSPA プラグインを使えるようになったが、ビルト済みのバイナリ版では FFTW を使えなくなつた。

アーキテクチャを決める際には、開発者たちの限られた時間をいかにして最大限に生かすかということも検討材料となる。我々開発チームの規模は小さいので、使えるリソースも限られている。たとえば Firefox や Thunderbird の開発チームがセキュリティホールを詳細に調べるのと同じようなことはできない。しかし我々は、Audacity にファイヤウォールを回避するような抜け道を仕込むつもりはない。そこで、Audacity では TCP/IP コネクションを一切扱わないことに決めた。TCP/IP を切り捨てれば、セキュリティに関して考えるべきことを大幅に減らせる。リソースが限られているのを認識したことで、よりよい設計に向かうことが

できた。開発者に必要以上の時間をとらせる機能をカットし、より本質的な機能に集中できるようにしたのだ。

同じく開発者たちの時間を考慮したのが、スクリプト言語の扱いだ。スクリプトで Audacity を操作できるようにしたいが、スクリプト言語を実装するコードは Audacity の中に組み込む必要はない。各言語の処理系を Audacity に組み込んでコンパイルし、ユーザーに好きなものを使わせるというのはあまり意味がない。¹ そのかわりに、スクリプトによる操作はひとつのプラグインモジュールとパイプを使って実装することにした。後ほど説明する。



図 2.1: Audacity の階層

図 2.1 は、Audacity のレイヤーとモジュールを図示したものだ。図を見ると 3 つの重要なクラスが wxWidgets 内で強調されており、それぞれに対応するものが Audacity にも用意されている。高レベルの抽象化を、例レベルのクラスに対して行っているのだ。たとえば BlockFile システムは wxWidgets の wxFiles に対応するもので、このクラス上に構築されている。おそらく、いざれは BlockFiles や ShuttleGUI そしてコマンド処理を仲介ライブラリとして切り出すことになるだろう。そうすれば、より汎用的にできる。

図の下のほうには“Platform Specific Implementation Layers”という細長い部分がある。wxWidgets や PortAudio は、どちらも OS を抽象化するレイヤーだ。どちらにも、対象プラットフォーム

¹唯一の例外は Lisp ベースの言語 Nyquist で、これは Audacity の開発が始まったばかりのころから組み込まれている。できることなら別のモジュールに分割して Audacity にバンドルする形式にしたいところだが、その作業に時間を割く余裕がない。

ムに依存するさまざまな実装にあわせて処理を切り替えるコードが含まれている。

“Other Supporting Libraries” のカテゴリに含まれるのは、さまざまなライブラリの集まりだ。興味深いことに、これらの多くは動的に読み込まれたモジュールを信頼している。これらの動的モジュールは wxWidgets について何も知らない。

Windows プラットフォーム上では、Audacity を单一の一枚岩な実行ファイルとしてコンパイルしていた。つまり wxWidgets と Audacity アプリケーションが同じ実行ファイルに含まれている状態だ。2008 年にこれをモジュラー構造に変更し、wxWidgets は個別の DLL として用意するようにした。これで、さらに別の DLL を実行時に読み込んだときにもそれらの DLL が直接 wxWidgets の機能を使えるようになった。図中の点線より上に組み込むプラグインは、wxWidgets を使うことができる。

wxWidgets を DLL 化したことによるマイナス面もある。まず、配布ファイルのサイズが大きくなった。その理由のひとつは、使ってもいない多くの関数が DLL に組み込まれているからである。DLL 化する前は、これらは最適化されていた。また、Audacity の起動にやや時間がかかるようになった。各 DLL を個別に読み込むためである。しかし、メリットも多い。モジュール化することで、ちょうど Apache のモジュールと同じようなメリットが得られることを期待している。モジュール化したおかげで Apache 本体は非常に安定するし、その一方で実験的な開発や特殊な機能、新たなアイデアなどはモジュールで試すことができる。モジュールのおかげで、プロジェクトをフォークして別の道を行くという衝動に対抗することができる。モジュール化するという決断は、我々にとって非常に重要なアーキテクチャ変更だったと考えている。これらのメリットが得られるだろうと期待してはいるが、今のところはまだ得られていない。wxWidgets の機能を公開することは単なるはじめの一歩に過ぎず、我々はより柔軟なモジュラーシステムに向かってさらに進んでいく。

Audacity のようなプログラムの構造は、前もって明確に設計できるものではない。時間をかけて成長させていくものだ。全体的に見て、現状のアーキテクチャはうまくいっている。ソースファイルの多くの部分に影響する新たな機能を追加するときには、アーキテクチャと対決している気分になる。たとえば、Audacity は現在ステレオとモノラルのトラックをそれぞれ専用の方法で処理している。Audacity を改造してサラウンドサウンドを処理させようにしようすれば、Audacity 内の多くのクラスに手を入れる必要がある。

ステレオのその先に: GetLink の物語

Audacity は、これまでチャンネル数を抽象化したことがなかった。そのかわりに、音声チャンネルに結びつけた抽象化を使っている。GetLink という関数があり、この関数は、2 チャンネルのときはペアのもう一方の音声チャンネルを返し、モノラルのときには NULL を返す。GetLink を使うコードは、まるで最初はモノラル用に書いていたものに後から (GetLink() != NULL) を使ってステレオ処理のコードを継ぎ足したように見える。本当にそうだったのかは不明だが、私はきっとそういうだろうと踏んでいる。GetLink を使って連結リスト内のすべてのチャンネルを反復処理させるといったループはなく、描画やミキシング、そして読み書きのすべてで「ステレオの場合は…」という場合分けが含まれている。もともとのコードが任意の n チャンネルを処理できて、ほとんどの場合は n が 1 か 2 になる、というようになつていいのだ。より汎用的なコードにしようともおもつたら、約 100 か所にある GetLink 関数の呼び出しに手を入れなければならない。ファイル数にして少なくとも 26 以上になる。

GetLink を呼び出しているところを検索して適切に変更するのはそんなに複雑な作業ではない。この“問題”を修正するのは、最初に感じたほど大がかりな作業ではないだろう。GetLink の物語は、修正が困難な構造的欠陥に関するものではない。それよりもむしろ、比較的小規模な欠陥がいかにコードを蝕んでいくかを示すものだ。

今になってみれば、GetLink 関数は private にしてそのかわりにイテレータを提供しておいたほうがよかつたのだろう。そうすれば、あるトラック内のすべてのチャンネルを順に処理することができる。これでステレオの場合の処理を特別に書くことも防げるし、音声チャンネルのリストを使うコードはリストの実装を知らなくても書けるようになる。

設計のモジュール化を進めると、内部構造をうまく隠蔽する方向に我々を導いてくれるだろう。外部向けの API を定義して拡張することで、アプリケーションが提供する機能について、よりしっかり見つめなければならないことになる。それによって我々は、外部向け API にとらわれない抽象化を心がけるようになるだろう。

2.2 wxWidgets GUI ライブラリ

Audacity のユーザーインターフェイスのプログラマーにとって一番重要なライブラリをひとつあげるとすれば、それは wxWidgets GUI ライブラリだ。このライブラリは、ボタンやスライダ、チェックボックス、ウィンドウそしてダイアログなどを提供する。つまり、目に見えるクロスプラットフォームな挙動の大半をこのライブラリが提供していることになる。

wxWidgets ライブラリには自前の文字列クラス `wxString` が用意されており、スレッドやファイルシステムそしてフォントなどをクロスプラットフォームで抽象化する。また他の言語へのローカライズにも対応しており、これらすべての機能を Audacity で使っている。Audacity の開発に新たに参加しようとする人に勧めるのは、まず wxWidgets をダウンロードしてコンパイルし、付属のサンプルをいくつか試してみることだ。wxWidgets は、OS が提供する GUI オブジェクトの上に乗る比較的薄めのレイヤーである。

複雑なダイアログを組み立てるために、wxWidgets では個々のウィジェットの要素だけではなくサイザー (sizer) も用意している。これは要素のサイズや位置を制御するもので、図形要素に対して固定の座標での位置指定をするよりもずっとよい。ユーザーが直接ウィジェットのサイズ変更したり、フォントサイズを変更したりしても、ダイアログ内の要素の位置は自然に更新される。サイザーは、クロスプラットフォームなアプリケーションにとって重要となる。もしこれがなければ、ダイアログのカスタムレイアウトをプラットフォームごとに用意しなければならなくなるだろう。

ダイアログのデザインをリソースファイルに書き出すこともよくある。このリソースファイルをプログラムから読み込んで使うのだ。しかし Audacity では、ダイアログの設計は wxWidgets の関数呼び出しの形でプログラムに埋め込んでいる。これによって最大限の柔軟性を確保している。つまり、ダイアログの正確な内容や振る舞いをアプリケーションレベルのコードで決定できるということだ。

かつては、Audacity の GUI を作成するコードの中に、グラフィカルなダイアログ作成ツールで自動生成したのが明らかなコードを見かけること也有った。この手のツールは基本的なデザインを作るときには便利だった。時を経てそのコードにも手が入り、新たな機能が追加された。その後新たなダイアログを作るときには、既存のコードをコピーして流用することが多くなった。すでに自動生成結果に手が加えられたダイアログのコードをだ。

そんな開発を何年も続けてきた結果、Audacity のソースコードの多く(特に、ユーザーの環境設定用のダイアログまわり)に複雑なコードの重複が目立つようになった。かつてはシンプルだったのかもしれないが、そんなコードを追いかけるのは大変だ。問題のひとつは、ダイアログを組み立てる手順がばらばらだったことだ。小さめの要素を組み合わせて大きな要素を作り、最終的にダイアログが完成するようになっていたが、コードで要素を作成する順番は画面上での要素の配置順と一致していなかった(一致させる必要もなかった)。コードは冗長で、繰り返しも多かった。GUI 関連のコードの中には環境設定のデータをディスクから読み出して中間変数に送るものもあったし、中間変数から表示用 GUI に送るコードもあった。また、GUI から中間変数にデータを送るコードもあれば、中間変数からディスクに保存するコードもあった。そのコードのそばには「//これはひどい」とかいうコメントが入っていたが、何か手を付けようとするまでにはかなり時間がかかってしまった。

2.3 ShuttleGui レイヤー

これらのコードのもつれを解決する手段が新たなクラス ShuttleGui だ。これはダイアログの作成に必要なコードの行数を大幅に減らし、コードをより読みやすくする。ShuttleGui は新たなレイヤーで、wxWidgets ライブラリと Audacity との間に挟まるものだ。その役割は、wxWidgets ライブラリと Audacity との間での情報の受け渡しである。例を示そう。これは図 2.2 のような GUI 要素を作成する。

```
ShuttleGui S;
// GUI Structure
S.StartStatic("Some Title",...);
{
    S.AddButton("Some Button",...);
    S.TieCheckbox("Some Checkbox",...);
}
S.EndStatic();
```

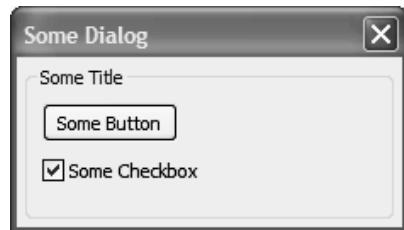


図 2.2: ダイアログの例

このコードはダイアログ内に静的なボックスを定義し、その中にボタンとチェックボックスをひとつずつ置いている。コードとダイアログの対応は明確だ。StartStatic と EndStatic の呼び出しがペアになっている。それ以外にも同様な StartSomething/EndSomething のペアがあり、これらは必ずセットで存在しなければならない。これが、ダイアログ上の配置を制御する。波括弧での囲みやその中の字下げは単にコードの見た目だけにかかるものであり、必須というわけではない。しかし我々は、このように書くよう規約を定めている。コードの構造や StartSomething/EndSomething のペアを明確にするためである。コードが巨大になると、これが可読性の向上にとても役立つ。

先ほど示したソースコードは、単にダイアログを作るだけではない。コメント “//GUI Structure” の後に続くコードを使って、ダイアログのデータを設定保存先に送ったり逆にデータを取得したりすることができる。これまででは、同様のことをするためには、同じようなコードを大量に繰り返さねばならなかった。今では、コードを一度だけ書けばあとは ShuttleGui クラスがうまくやってくれる。

それ以外にも Audacity では、wxWidgets の基本機能を拡張するモジュールを使っている。たとえば、Audacity にはツールバーを管理するための自前のクラスがある。なぜ wxWidget の組み込みのツールバークラスを使わなかつたのかって?それは、歴史的な理由によるものだ。Audacity のツールバーが書かれたのは、wxWidgets にツールバークラスが用意されるよりも前のことだったのだ。

2.4 TrackPanel

Audacity のメインパネルで波形を表示しているのが TrackPanel だ。これは Audacity のカスタムコントロールだ。このコントロールは、いくつかの小さな部品を組み合わせて作られている。トラック情報を表示するパネルや時間軸のルーラー、振幅のルーラー、そしてトラックの波形やテキストラベルなどだ。トラックのサイズ変更や移動は、マウスのドラッグで行える。トラックにはテキストラベルが含まれているが、これは編集可能なテキストボックスを自分で再実装したものであり、組み込みのテキストボックスではない。これらのパネルやトラック、ルーラーは wxWidgets のコンポーネントにすべきだと考える人もいるだろう。しかしそのようにはなっていない。



図 2.3: Audacity のインターフェイスに、Track Panel の要素名を示したもの

スクリーンショット 図 2.3 に Audacity のユーザーインターフェイスを示す。名前がつけられているコンポーネントはすべて、Audacity 用にカスタマイズしたものである。wxWidgets の観点から見れば、ここにある wxWidget のコンポーネントは TrackPanel 用のひとつだけだ。その内部の配置や再描画はすべて、wxWidget ではなく Audacity 側のコードが面倒を見ている。

これらのコンポーネントをうまく取りまとめて TrackPanel を作るのは、本当に恐ろしいことだ（恐ろしいのはあくまでもコードのことであって、実際にユーザーが使う完成品はよくできているけどね）。GUI とアプリケーションのコードが入り混じっていて、きれいに分離でき

ていない。きちんと設計するなら、アプリケーションのコードだけが左右のオーディオチャネルやそのデシベル、ミュート、ソロなどの設定を閲知するべきだ。GUIの要素がそれらを閲知しているようではいけない。GUIの要素はオーディオ関連以外のアプリケーションでも再利用可能でなければならぬ。TrackPanel上の部品は、純粋なGUIでさえもつぎはぎだらけのコードになっている。絶対位置やサイズによる場合分けが含まれており、十分に抽象化できていない。これらのコンポーネントが自身でGUI要素を内包してwxWidgetsのsizerのようなインターフェイスを使っていれば、どんなにきれいで一貫性のあるコードになるだろう。

TrackPanelをそんなふうに改良するために、トラックやその他のウィジエットの移動やサイズ変更を行うwxWidgets用の新たなsizerが必要となった。wxWidgetsのsizerは、それを満たすほど柔軟ではない。新たにsizerを作ったおかげで、それを他の部分でも使えるようになった。我々はツールバー上に保持するボタンにもこれを使い、ツールバー上のボタンの並べ替えをドラッグで簡単にできるようにした。

新たなsizerを作るために事前調査を行ったが、調査が足りなかつたようだ。GUIコンポーネントをwxWidgetsから完全に独立させようとする試みたが、問題が発生した。ウィジエットの再描画がうまく制御できなくなり、コンポーネントのサイズを変更したり移動させたりするとちらつきが発生するようになったのだ。wxWidgetsをベースにして拡張することで再描画時のちらつきを解決し、サイズ変更の処理と再描画の処理をうまく分離させる必要があった。

TrackPanelをこの方式で改良するのを躊躇したもうひとつの理由は、ウィジエット数が増えるとwxWidgetsの起動に時間がかかるなどを既に我々が知っていたからだ。これは、wxWidgetsからはあまり手の施しようのない問題だ。個々のwxWidgetやボタン、テキスト入力ボックスはそれぞれウインドウシステムのリソースを使っている。そして、個々のリソースは、アクセスするためのハンドルを持っている。多数のハンドルを処理するのには時間がかかる。たとえ大半のウィジエットが非表示あるいはスクリーンの外部にある場合であっても処理が遅くなることは変わらない。我々は、小さなウィジエットを大量に使うつもりだったのだ。

最もよい解決策はFlyweightパターンを採用することだ。軽量なウィジエットが自分自身の描画を担当し、対応するオブジェクトを持たないようにすれば、ウインドウシステムのリソースやハンドルを消費せずに済む。我々はwxWidgetsのsizerやコンポーネントウィジエットと同様の構造を使い、同様のAPIを提供するようにした。しかしそれはwxWidgetsのクラス群を派生させたものではない。我々は既存のTrackPanelのコードをリファクタリングによりきれいな構造にした。もしこれが簡単な解決法なら既にそうしていただろうが、自分たちがいittai何を求めているのかについての議論が発散した結果、初期の試みから脱線してしまった。現在のアドホックな方式を一般化するには、大変な設計作業とコーディングが必要となる。複雑ではあるけれども今きちんと動いているコードをそのまま残しておきたいという強い誘惑にもかられた。

2.5 PortAudio ライブラリ: 録音と再生

PortAudio はオーディオライブラリで、Audacity はこれを使って、録音と再生をクロスプラットフォームな方法で提供している。このライブラリがなければ、Audacity は実行環境のサウンドカードを使うことができないだろう。PortAudio が提供する機能にはリングバッファや録音・再生時のサンプルレート変換などがある。重要なのは、その API によって Mac や Linux そして Windows におけるオーディオ処理の差異を隠ぺいできるということだ。PortAudio の内部には、各プラットフォームで API に対応するための実装ファイルが別々に用意されている。

私はこれまで、PortAudio の内部に立ち入って何をしているのか追いかける必要などなかった。しかし、Audacity と PortAudio の間でどのようなやりとりをしているかを知っておくと便利だ。Audacity は、PortAudio からデータのパケットを受信(録音)したり、逆にパケットを PortAudio に送信(再生)したりする。送信や受信が実際のところどのように行われているのか、そしてそれがディスクへの読み書きや描画の更新とどのようにつながっているのか。そのあたりは見る価値があるだろう。

いくつかの異なる処理が、同時に発生する。頻繁に発生し、少量のデータをやりとりし、高速な反応を要するものがあれば、あまり頻繁には発生しないが大量のデータをやりとりしなければならないものもある。こちらについては処理がいつ発生するかはそれほど重要ではない。ここに、処理の内容とその対応に使うバッファとの間のインピーダンスマッチが発生する。もうひとつ見る価値があるのは、オーディオデバイスやハードディスクそして表示画面などを扱う部分だ。末端まで必死になって追いかけるつもりはなく、与えられた API を使って作業をすることになる。各プロセスを同じように見て、たとえばすべて wxThread から立ち上げるようにしたいものだが、そのような贅沢はできない(図 2.4)。

ひとつのオーディオスレッドを PortAudio のコードが立ち上げ、それが直接オーディオデバイスとやりとりする。これは、録音や再生を行うものだ。このスレッドは反応が速いものでなければならず、そうでないとパケットを失ってしまう。PortAudio のコードの配下にあるスレッドが audacityAudioCallback を呼び、録音時には、新たに受け取った小さなパケットを大きめ(5 秒)のキャプチャバッファに追加する。再生時には、5 秒間の再生バッファから小さな塊を取り出す。PortAudio ライブラリは wxWidgets については一切知らない。そのため、PortAudio が作ったこのスレッドは pthread である。

第二のスレッドを立ち上げるのは、Audacity の AudioIO クラスだ。録音の際に、AudioIO がデータをキャプチャバッファから受け取り、それを Audacity のトラックに追加して表示させる。さらに、十分な量のデータが追加された時点で、AudioIO がデータをディスクに書き込む。このスレッドは、再生時のディスクからの読み込みも行う。ここで鍵となるのが関数 `AudioIO::FillBuffers` といくつかの Boolean 変数の設定で、録音と再生をこのひとつの関数で行う。重要なのは、このひとつの関数で双方向の処理をしているということだ。録音部と再生部を同時に使うことがある。“software play through” で、以前に録音された内容に対する多重録音を行うときだ。AudioIO のスレッド内では、完全に OS のディスク IO にしばられた状態となり、ディスクの読み書きでしばらく待たされることもあるかもしれない。これら



図 2.4: 録音・再生時のスレッドとバッファ

の読み書きを `audacityAudioCallback` で行うことはできない。この関数は高速な反応を要求されるからである。

これらふたつのスレッド間の通信は、共有変数を使って行う。どちらの変数がいつ書き込みを行っているのかを制御できているので、ミューテックスを使うようなぜいたくは不要だ。

再生と録音の両方で、さらにもうひとつの要件がある。Audacity は GUI も更新しなければならないということだ。これは、最もタイムクリティカルでない処理である。描画の更新はメインの GUI スレッドで行われ、1 秒間に 20 回発生する定期的なタイマーで実行される。このタイマーが `TrackPanel::OnTimer` を呼び出し、GUI の更新が必要な場所が見つかれば更新する。メインの GUI スレッドは、我々のコードではなく wxWidgets の中から立ち上げる。他のスレッドからは、直接 GUI を更新することはできない。タイマーを使って GUI スレッドを取得し、画面の更新が必要かどうかを調べる。そうすることで、再描画の回数を減らしながらも許容できるレベルの反応を保つ。また、そうすることで、表示用にあまりプロセッサタイムを要求しすぎないようにしている。

オーディオデバイス用のスレッドとバッファ/ディスク用のスレッド、そして定期的なタイマーを持つ GUI スレッドの三つを使ってオーディオデータのやりとりをするというのは、よい設計と言えるだろうか? これら三種類のスレッドを单一の抽象基底クラスから派生させないのは、多少アドホックにも感じる。しかし、このアドホック性の要因の多くは、使っているライブラリによるものである。PortAudio は、自分自身でスレッドを作ることを想定している。wxWidgets フレームワークが GUI スレッドを持っているのは、ごく自然なことだ。バッファを埋めるためのスレッドを要する理由は、オーディオデバイスのスレッドで頻繁に発生する小規模のパケットと、あまり頻繁には発生しないディスクドライブの大規模なパケットとの間のインピーダンスマッチを解決するためだ。これらのライブラリを使うことには明確な利点がある。逆に、これらのライブラリを使うことで必要となるコストは、結局は各ライブラリが提供する抽象化を使うことになるということだ。結果的に、メモリ内でのデータのコピーの回数が、最低限必要な回数だけでなくさらに増えてしまっている。私がこれまで関わってきた高速なデータ交換の処理ではもっと効率的なコードを見たことがある。そのコードでは、割り込みなどで発生するこの手のインピーダンスマッチに対応するのにスレッドなど使っていなかった。データをコピーするのではなく、バッファへのポインタを渡していたのだ。ただ、そんなことができるるのは、使っているライブラリがバッファの抽象化をよりリッチに設計している場合だけである。既存のインターフェイスを使う限りはスレッドを使うことを強いられ、そしてデータのコピーを強いられることになる。

2.6 BlockFile

Audacity が直面する困難のひとつが、録音したオーディオデータへのデータの追加や削除だ。オーディオデータの長さは数時間分になることもある。録音データはどんどん長くなり、利用可能な RAM の容量を簡単に突破してしまうだろう。録音データをディスク上に单一のファイルとして管理していたとすると、ファイルの先頭あたりにデータを挿入しようとしたときには大量のデータ移動が発生することになる。ディスク上でデータのコピーには時間がかかる。つまり、Audacity はちょっとした編集でも長々と待たされるソフトウェアになってしまう。

Audacity がこの問題に対処するために使った方法は、オーディオファイルを多数の BlockFile に分割することだ。個々のファイルの大きさは約 1 MB となる。これが、Audacity が自前のオーディオファイルフォーマットを採用している主な理由だ。マスターファイルの拡張子は .aup となる。これは XML ファイルで、このファイルがさまざまなブロックを取りまとめている。長いオーディオデータの先頭近くを変更した場合でも、影響を受けるのはたったひとつのブロックとマスターファイル.aup だけとなる。

BlockFile は、対立する二つの勢力の調和をうまく保っている。挿入や削除の際に必要以上のコピーが発生することもないし、再生時には、ディスクへのリクエストのたびに適度に大きなデータの塊を取得できることが保証されている。ブロックが小さくなればなるほど、同じ量のオーディオデータを取得するために必要なディスクへのリクエストの回数が増え、逆に大きくなればなるほど挿入や削除の際のコピーの量が増える。

Audacity の BlockFile は決して内部に空き領域を持たないし、最大のブロックサイズを超えることもない。このルールを守るため、挿入や削除をするときには最大 1 ブロックまでのデータのコピーが発生する可能性がある。BlockFile が不要になれば、削除する。BlockFile は参照カウンタで管理されているので、オーディオの一部を削除したとしてもそれに関連する BlockFile はまだ残ったままであり、undo に対応できるようにしている。データを保存するまではその状態になる。Audacity の BlockFile には、空き領域のガベージコレクションはまったく不要だ。我々が対応する必要があるのは、オールインワン型のファイルである。

大きめのデータのマージや分割こそが、データを管理するシステムの本業である。B ツリーから Google の BigTable のタブレットや Unrolled linked list の管理まで、それは変わらない。図 2.5 は、Audacity でオーディオの開始位置付近を削除したときに何が起こるのかを示している。



図 2.5: 削除する前は.aup ファイルと BlockFile が保持するのは ABCDEFGHIJKLMNO。FGHI を削除すると、ふたつの BlockFile がマージされる。

BlockFile が扱うのはオーディオそのものだけではない。それ以外にも、サマリ情報をキャッ

シュする BlockFile もある。Audacity で 4 時間のオーディオデータを表示するときに、画面の再描画のたびにオーディオ全体を読み直すのは考えられない話だ。そういう場合にサマリ情報を代わりに使う。サマリ情報には、その時間の範囲内での最大音量や最小音量が記録されている。表示をズームインしたときは、実際のデータを使って描画を行い、ズームアウトしたときはサマリ情報をもとに描画を行う。

BlockFile システムの特徴のひとつに、各ブロックは Audacity が作ったファイルでなくともかまわないという点がある。たとえば、.wav 形式で保存されたオーディオファイルの特定の時間範囲を指すこともできる。ユーザーが Audacity プロジェクトを立ち上げて.wav ファイルからオーディオをインポートしていくつかのトラックをミックスしたとしよう。このときに作られる BlockFile はサマリ情報を格納するものだけだ。そのおかげでディスク容量も抑えられるし、オーディオデータをコピーする時間も節約できる。しかしながら、これはあまり良い考えではない。これまでに、多くのユーザーがインポート後に元の.wav ファイルを削除してしまっていた。Audacity のプロジェクトフォルダに全部コピーされているものと勘違いしていたのだ。実際はそうではなく、元の.wav ファイルがなければそのプロジェクトは再生できない。そのため、Audacity の現在のデフォルト設定では、インポートしたオーディオを常にコピーして新しい BlockFile を作るようにしている。

BlockFile 方式が、Windows 上で問題となることがあった。Windows 上で大量の BlockFile を扱うと、パフォーマンスが非常に劣化したのだ。原因はおそらく、Windows では同一ディレクトリ上にある大量のファイルの処理速度に難があったからだろう。同様の問題が、ウィジェットをたくさん使ったときの速度低下という形でもあらわれた。その後、サブディレクトリ階層を使うように変更を加え、ひとつのディレクトリには最大 100 個までのファイルしか置かないようにした。

BlockFile の構造を使うときの最大の問題は、その構造がエンドユーザーに公開されてしまうことだ。よく聞く話だが、.aup ファイルを別の場所に移したユーザーが、BlockFile を含むフォルダと一緒に移動させなければならないことに気付かないことがある。Audacity のプロジェクトが単一のファイルにまとまっていて、その内部のスペースやファイル内の利用状況を管理できる状態であればよかつただろう。こうすれば、パフォーマンスはどちらかといえば向上するだろう。追加で必要となるコードは、おそらくガベージコレクションであろう。シンプルなアプローチとしては、ファイルのみ使用領域が設定した割合を上回るときにブロックを新しいファイルへコピーするという方法がある。

2.7 スクリプト

Audacity には、複数のスクリプト言語に対応する実験的なプラグインが付属している。このプラグインは、名前付きパイプを使ったスクリプトインターフェイスを提供する。スクリプト用に公開されたコマンドはテキスト形式で、同じくコマンドの応答もテキストとなる。つまり、名前付きパイプへのテキストの書き出しや名前付きパイプからのテキストの読み込み

に対応しているスクリプト言語ならなんでも、Audacity を動かせるということだ。オーディオそのものなどの大きなデータにパイプ上を行き来させる必要はない(図 2.6)。



図 2.6: スクリプトプラグインが、名前付きパイプを使ったスクリプト処理機能を提供する

プラグイン自身は、自分が運ぶテキストの内容については何も知らない。単にそれを運ぶだけだ。スクリプトプラグインが使うプラグインインターフェイス(あるいは原始的な拡張ポイント)は、Audacity 側でテキスト形式のコマンドとして公開されている。そのため、スクリプトプラグインは小さなプラグインで、コードの大部分はパイプを扱う処理である。

残念ながら、パイプを使うということは TCP/IP 接続を扱うのと同じセキュリティリスクを負うことになる—セキュリティを考慮して Audacity では TCP/IP 接続を扱わないことに決めている。リスクを少しでも下げるために、このプラグインはオプションの DLL としている。これを取得して使う前には熟考が必要だ。また、このプラグインを使うときにはセキュリティに関する警告が出る。

スクリプトの機能を公開した後で、Wiki の機能追加リクエストのページでこんな提案を受けた。KDE の D-Bus を使えば TCP/IP によるプロセス間通信機能を提供できるのではないか、というものだ。既に別のやりかたで始めたところではあるが、最終的に D-Bus をサポートするようになる可能性も残っている。

2.8 リアルタイムエフェクト

Audacity にはリアルタイムエフェクトの機能はない。再生時にその場で計算してエフェクトをかけるような機能のことだ。Audacity で何かのエフェクトを適用すると、処理が完了するまで待たされることになる。リアルタイムエフェクトを可能にすることやエフェクト処理をバックグラウンドで実行させてその間にもユーザーインターフェイスを機能させることは、Audacity への機能追加要望としてもっともよくあげられるものだ。

問題は、あるマシン上でリアルタイムエフェクトが機能したとしても、別の遅いマシンではとてもリアルタイムとは言えないような速度しか出ない可能性があるということだ。Audacity はさまざまなマシン上で動作する。そのため、穏やかな代替機能を用意しておきたい。多少処理速度の遅いマシンでもエフェクトをトラック全体にかけられるようにし、トラックの中

スクリプト機能のはじまり

スクリプト機能ができたきっかけは、ある Audacity ファンから提供された機能だった。それはあるニーズを満たすための機能で、当初は Audacity をフォークする方向に向かっていた。その機能はひとまとめにして CleanSpeech と呼ばれ、教会での説教を mp3 に変換するために作られた。CleanSpeech には無音部分の切り詰め—オーディオ内の長い無音部分を探して切り取る—などの新たなエフェクトやノイズ除去効果のある固定シーケンスを適用する機能があり、オーディオの正規化や録音内容の mp3 への変換機能などもあった。中にはぜひ取り込みたいくなるような素晴らしい機能もあったが、その実装は Audacity 内ではかなり特殊なものだった。それを Audacity の本流に取り込むと、固定シーケンスではなく可変シーケンスのコードを書くことになった。可変シーケンスだと、コマンド名と Shuttle クラスのルックアップテーブル経由で任意のエフェクトを使い、コマンドのパラメータはテキスト形式でユーザー設定項目に格納することができた。この機能は **バッチチェイン** と名付けられた。条件分岐や計算式を追加して楽をする意図的に避け、アドホックなスクリプト言語を作ってしまったないようにした。

今にして思えば、フォークを避けようと努力をする価値はあった。今でも CleanSpeech モードは Audacity に埋め込まれており、環境設定で有効にすることができる。さらにユーザーインターフェイスも減らし、高度な機能は削除した。シンプルにしたバージョンの Audacity は、他の用途で使いたいという要望もくるようになった。中でも特筆すべきなのは学校での採用だった。問題は、どれが高度な機能でどれが必要不可欠な機能なのかについての意見が人によって異なっていたということだ。我々はその後シンプルなハックを行い、翻訳の仕組みを向上させた。メニュー項目の翻訳のうち “#” で始まるものは、メニューに表示しないようにしたのだ。これで、メニューの項目が多すぎると感じる人も再コンパイルなしでメニューを減らせるようになった—これはより汎用的であり、Audacity の mCleanspeech フラグほどには侵略的ではない。このフラグは、いつか取り除いてしまいたいものだ。

CleanSpeech の作業は、我々にバッチチェインと無音切り詰め機能をもたらした。どちらも、コアチーム以外から持ち込まれた魅力的な機能改善だ。バッチチェインはその後のスクリプト機能につながった。それはまた、より汎用的なプラグイン機能を Audacity に持ち込むきっかけにもなった。

央近くまでは処理済みのオーディオを聞けるようにしたい。少しウェイトを入れ、Audacity が最初に処理すべき部分を判断することになる。エフェクトをリアルタイムでレンダリングするには遅すぎるマシンでは、再生がレンダリングに追いつくまではオーディオを聞けるようにしておきたい。そのためには、オーディオエフェクトがユーザーインターフェ

イスを奪ってしまったり、オーディオブロックの処理を左から右へ順にしなければならなかつたりといった制約を取り除くことだ。

比較的最近 Audacity に追加されたオンデマンド読み込み機能には、リアルタイムエフェクトに必要となる要素の多くが含まれている。しかし、オーディオエフェクトにはまったくかかわっていない。オーディオファイルを Audacity にインポートするときに、サマリ情報の BlockFile の作成は今ではバックグラウンドで行われる。Audacity はプレースホルダーとして青とグレーの斜めの縞をオーディオの部分に表示し、まだ処理が完了していないことを示す。そして、オーディオの読み込み中であっても多くのユーザーコマンドに反応することができる。ブロックの処理を左から右へ順に行う必要はない。このコードは、きっとリアルタイムエフェクトにも使われることになるだろうと確信している。

オンデマンド読み込み機能は、リアルタイムエフェクトの実現に向けた一歩となる。エフェクト自身をリアルタイムで行うことに関する複雑性を、いくらか回避してくれるだろう。リアルタイムエフェクトでは、さらにブロック間のオーバーラップが必要となる。そうしないと、エコーのようなエフェクトが正しくつながらない。また、再生中にオーディオのパラメータを変更することにも対応しなければならない。オンデマンド読み込みを最初に実装したおかげで、他に比べて早い段階からコードを使うことができる。実際の使用例からのフィードバックも得られるだろう。

2.9　まとめ

本章の前半で説明したのは、よりよい構造がいかにプログラムの成長につながるか、そして構造に気を使わないことがいかに開発の妨げになるか、ということだった。

- PortAudio や wxWidgets といったサードパーティの API には多大な利点がある。きちんと動作するコードを組み込めるというだけではなく、プラットフォームの差異をうまく抽象化してくれる。サードパーティの API を使う代償は、抽象化の方法を自由に選ぶという柔軟性がなくなることだ。再生や録音のコードはとても美しいとは言えないものだが、これは三種類の異なる方法でスレッド管理する必要があったからだ。また、このコードにはデータのコピーが多いが、もし抽象化を自由にできていればもう少し減らせたはずだ。
- wxWidgets が我々にもたらす API は、ついで冗長で追いづらいコードを書きたくなってしまうものだった。その誘惑から逃れるため、我々は wxWidgets の前に Facade を用意して自分たちの求める抽象化ができるようにした。これによって、アプリケーションのコードがよりきれいになった。
- Audacity の TrackPanel では、既存のウィジェットから容易に得られる機能を超えたものが必要となった。その結果、自前のアドホックなシステムを用意することになった。ウィジェットや sizer と論理的に区別されたアプリケーションレベルのオブジェクトからなるよりきれいなシステムが TrackPanel から出てくるように戦っている。

- 構造に関する決定とは、単に新機能をどのように実装するかを決めるだけにとどまらない。プログラムに何を含めないかを決めるのも重要だ。そうすれば、よりきれいで安全なコードにつながる。Perlのようなスクリプト言語の恩恵を受けるときに自分たちのプログラムをいじらなくて済むのはとてもありがたいことだ。構造に関する決定は、将来の成長戦略に基づくものもある。我々のモジュラーシステムはまだ産まれたばかりのものだが、これを生かしてより多くの実験をより安全に行えることを期待している。また、オンデマンドの読み込み機能は、オンデマンドでのリアルタイムエフェクト処理に進化していくことを期待している。

見れば見るほど明らかなのは、Audacity がコミュニティの尽力の成果だということだ。コミュニティとは、単に Audacity に直接貢献している人たちだけを指すのではない。Audacity はさまざまなライブラリに依存しており、個々のライブラリにはそのコミュニティもあればその分野のドメインエキスパートもいるであろうからだ。Audacity のいろいろ入り混じった構造についての記事を読んだ人にとっては何の驚きもないだろうが、コミュニティは新しい開発者の参入を歓迎しており、スキルレベルに応じていろいろなことをすることができる。

私は、Audacity のコミュニティの質がそのコードの強みや弱みに反映されていると信じて疑わない。より閉じたグループで開発すれば今よりも高品質で一貫性のあるコードが書けるかもしれない。しかし、そんなことをすれば貢献者の数も減り、Audacity が持つ幅広い機能に対応するのは難しくなるだろう。

The Bourne-Again Shell

Chet Ramey

3.1 導入

Unix のシェルは、ユーザーと OS との間のコマンドによるインターフェイスを提供する。しかし、シェルはまた、リッチなプログラミング言語でもある。フロー制御やループそして条件分岐といった制御構造もあるし、基本的な数学演算や関数、文字列変数などもあり、シェルとコマンドの間の双方向の通信もある。

シェルは、ターミナルあるいはターミナルエミュレータ (xterm など) から対話的に使うこともできるし、コマンドをファイルから読み込むこともできる。bash を含むモダンなシェルにはコマンドラインの編集機能があり、コマンドの入力中に emacs 風あるいは vi 風の操作でコマンドラインをいじることができる。また、さまざまな形式でコマンド履歴を記録する。

Bash の処理はシェルのパイプラインとそっくりだ。ターミナルあるいはスクリプトから読み込んだデータはいくつかのステージを通して、各ステージで変換され、シェルが最終的にコマンドを実行してその返り値を受け取る。

本章では、bash の主要なコンポーネントである入力処理やパース、さまざまなワードの展開、その他のコマンド処理、そしてコマンドの実行について、パイプラインの観点から探求する。これらのコンポーネントはキーボードやファイルから読み込んだデータのパイプラインとして働き、それを実行されるコマンドに変える。

Bash

Bash は GNU オペレーティングシステムで使われているシェルであり、一般的には Linux カーネル上で実装されている。また、Mac OS X などその他の主要 OS 上でも動く。過去の歴史上のバージョンである sh に対して、対話的な操作においてもプログラミング機能においても改良が施されている。

名前の由来は Bourne-Again SHell の頭文字をとったもので、Stephen Bourne(現在の Unix シェルの先祖である /bin/sh の作者。このシェルはベル研の Version 7 Unix で登場した)の名



図 3.1: Bash のコンポーネントのアーキテクチャ

前と再実装によって生まれ変わったことをかけている。bash の最初の作者は Brian Fox で、彼は Free Software Foundation のメンバーだった。私は現在の開発者兼メンテナーであり、オハイオ州クリーブランドにあるケースウエスタンリザーブ大学に勤務している。

他の GNU ソフトウェアと同様、bash も移植性がきわめて高い。Unix のほぼすべてのバージョンで動作するし、その他の OS でも動作する—独自にサポートしている移植版には Windows 上の Cygwin や MinGW といった環境もあるし、QNX や Minix といった Unix ライクなシステムへの移植版は配布物に含まれている。ビルドして実行するために必要なのは Posix 環境だけである。つまり、Microsoft の Services for Unix (SFU) などでもよい。

3.2 構文単位およびプリミティブ

プリミティブ

bash には、基本的に三種類のトークンがある。予約語 (reserved word)、単語 (word)、そして演算子 (operator) だ。予約語とはシェルやそのプログラミング言語に対して何らかの意味

を持つ単語のことで、フロー制御構文に使われることが多い。たとえば `if` や `while` がそれにあたる。演算子とはメタ文字を組み合わせたもののこと、メタ文字とはシェル自身に対して特別な意味を持つ文字を指す。`|` や `>` などだ。それ以外のシェルへの入力は普通の単語で、その中にはコマンドライン内での登場位置によって特殊な意味を持つもの—代入文や数値など—もある。

変数およびパラメータ

他のプログラミング言語と同様、シェルにも変数の機能があり、保存したデータを後で参照したり演算に使ったりすることができる。シェルが提供している変数には、ユーザーが設定可能な基本的な変数と、パラメータとして参照できる組み込みの変数がある。シェルのパラメータは一般的にシェルの内部状態を反映するもので、自動的に設定されたり別の操作の副作用として設定されたりする。

変数の値は文字列である。値の中には状況によって特別な意味を持つものもあるが、それについては後で説明する。変数への代入は、`name=value` 形式の文を使う。`value` は必須ではなく、省略した場合は空の文字列を `name` に代入する。`value` を指定すると、シェルはその内容を展開して `name` に代入する。シェルは、変数が設定されているかどうかによって処理を変えることがある。しかし、変数に値を設定するには値を代入する以外の方法はない。値を代入されていない変数は、たとえ事前に宣言されていたとしても参照すると `unset` となる。

ドル記号で始まる単語は、変数あるいはパラメータへの参照を意味する。ドル記号を含めた単語が、その名前の変数の値に置きかえられる。シェルには豊富な展開演算子が用意されており、単純な値の置換だけではなくパターンにマッチする部分を変更したり削除したりすることもできる。

変数には、ローカルとグローバルの二種類がある。デフォルトでは、すべての変数はグローバルとなる。単純なコマンド(最も見なれた形式のコマンド—コマンド名の後にオプションで引数やリダイレクトが続く形式)の前には代入文がくることもあり、そのコマンドのために変数が存在することになる。シェルはストアドプロシージャやシェル関数を実装しており、それぞれ関数ローカルな変数を持つことができる。

変数には最低限の型をつけることができる。単純な文字列値の変数に加えて、静数値と配列が使える。整数型の変数は数値として扱われる。文字列を代入するとそれを計算式とみなして展開し、計算結果を変数の値として代入する。配列は、インデックス型と連想型のどちらかになる。インデックス型の配列は数値を添字として使い、連想配列は任意の文字列を添字として使う。配列の要素は文字列であり、望むなら静数値として扱うこともできる。配列の要素に別の配列が入ることはない。

Bash は、ハッシュテーブルを使ってシェル変数の格納や取得を行う。また、そのハッシュテーブルの連結リストで変数のスコープを実装する。シェル関数の呼び出し用にさまざまなスコープがあり、コマンドの前にある代入文で設定した変数用のテンポラリスコープもある。代入文の後にシェルの組み込みコマンドが続くときは、シェルは変数の参照の解決順序を覚

えておく必要がある。また、連結したスコープが bash にそれを許可しなければならない。実行のネストレベルによっては、走査するスコープの数が驚くほど多くなることもありえる。

シェルプログラミング言語

単純なシェルコマンド、つまり読者の多くが最も見なれているであろうコマンドは、まず echo や cd のようなコマンド名があつてその後にゼロ個以上の引数やリダイレクトが続く。リダイレクトを使うと、起動するコマンドへの入力やコマンドからの出力をシェルのユーザーが制御できるようになる。先ほど説明したように、単純なコマンド内のローカル変数を定義することができる。

予約語を使えば、より複雑なシェルコマンドを実行できる。他の高級言語にもよくある制御構造である if-then-else や while も使えるし、for ループで値のリストを順に処理することもできる。また、C 言語風にカウンタを用いた for ループも使える。これらの複雑なコマンドを使えば、ある条件を調べてその結果によって処理を切り替えるようなコマンドを実行することもできるし、あるコマンドを複数回実行することもできる。

Unix が計算機界にもたらした贈り物のひとつがパイプラインである。これを使えば、一連のコマンド群でひとつのコマンドの出力を次のコマンドへの入力とすることができる。シェルの制御構造はすべてパイプラインの中でも使え、あるコマンドがデータをループに送るようなパイプラインを見るのも珍しくない。

Bash には、あるコマンドの実行時に標準入力や標準出力そして標準エラー出力をリダイレクトして別のファイルやプロセスに送る機能がある。シェルプログラマーは、リダイレクトを使って現在のシェル環境でファイルを開いたり閉じたりすることができる。

Bash では、シェルのプログラムを保存して再利用することができる。シェル関数やシェルスクリプトは、どちらもコマンド群に名前をつけて実行できるようにしたものであり、他のコマンドと同じように実行できる。シェル関数の宣言は特別な構文で行い、同じシェルのコンテキストで使うことができる。シェルスクリプトはコマンドを書いたファイルとして作り、実行するときにはそれを解釈する新たなシェルのインスタンスを立ち上げる。シェル関数は大半の実行時コンテキストを呼び出し元のシェルと共有するが、シェルスクリプトは新たなシェルを立ち上げて動作するので、環境変数で渡された内容しか共有できない。

さらなる注意

さらに読み進めていくうえで覚えておいてほしいのは、シェルがその機能を実装するため使っているデータ構造はほんのわずかであるということだ。配列、ツリー、片方向連結リスト、双方向連結リスト、そしてハッシュテーブル。これだけである。シェルのほぼすべての構造が、これらのプリミティブを用いて実装されている。

あるステージから次のステージに情報を渡したり各処理ステージでデータを操作したりするときに使う基本的なデータ構造が WORD_DESC だ。

```

typedef struct word_desc {
    char *word;           /* Zero terminated string. */
    int flags;            /* Flags associated with this word. */
} WORD_DESC;

```

単語を組み合わせて引数リストなどを作るときには、単純な連結リストを使う。

```

typedef struct word_list {
    struct word_list *next;
    WORD_DESC *word;
} WORD_LIST;

```

WORD_LIST はシェル全体に広がる。単純なコマンドは単語のリストだし、その展開結果も単語のリスト、そして組み込みのコマンドも引数の一覧を単語のリストで受け取る。

3.3 入力の処理

bash のパイプライン処理における最初のステージは、入力の処理である。ターミナルあるいはファイルから文字を受け取り、それを行単位に分け、各行をパーサに渡してコマンドに変換する。想像がつくだろうが、行とは改行文字で終わる文字列のことだ。

Readline およびコマンドラインの編集

Bash は、対話モードのときにはターミナルから入力を読み込み、それ以外の場合は引数で指定したスクリプトファイルから入力を読み込む。対話モードのときは、ユーザーが入力したコマンドラインを編集することができる。編集時には、Unix のエディタ emacs や vi とよく似たキーシーケンスや編集コマンドが使える。

Bash は readline ライブライアリを使ってコマンドラインの編集を実装している。このライブラリが提供する関数を使うと、コマンドラインの編集や入力内容の保存、過去のコマンドの呼び出し、そして csh 風の履歴の展開ができるようになる。Readline はもともと bash 用に開発されたものであり、今でも一緒に開発が進められているが、readline には bash 固有のコードは一切含まれていない。多くのプロジェクトが、readline を使ってターミナルベースの行編集インターフェイスを提供している。

Readline には、任意の長さのキーシーケンスを readline コマンドにバインドする機能もある。Readline には、カーソルの移動やテキストの挿入・削除、前の行の取得、そして途中まで入力した単語の補完などに対応するコマンドがある。これらのコマンドを使い、ユーザーはキーバインドと同じ構文でマクロを定義できる。マクロとは、キーシーケンスに対応して挿入される文字列のことである。マクロのおかげで、readline のユーザーはちょっとした文字列置換や作業の短縮ができるようになる。

Readline の構造

Readline は、読み込み/送出/実行/再表示という基本的なループで構成されている。まず最初に、キーボードからの文字の読み込みを `read` などで行うか、あるいはマクロからの入力を取得する。個々の文字は、キーマップ(ディスパッチテーブル)のインデックスとして使われる。キーマップのインデックスは 8 ビットの 1 文字であるが、その要素はさまざまなものになり得る。たとえば、キーマップの要素の文字列を別のキーマップとして解決することもある。このようにして、複数文字のキーシーケンスを実装している。また、`beginning-of-line` のような `readline` コマンドとして解決させることもあり、これは、そのコマンドを実行する。`self-insert` コマンドにバインドされた文字は、編集バッファに書き込まれる。あるキーシーケンスをひとつのコマンドにバインドすると同時に、そのシーケンスの一部を別のコマンドにバインドすることもできる(これは、比較的最近追加された機能である)。キーマップに特別なインデックスを追加して、これを実現している。キーシーケンスをマクロにバインドすることで、任意の文字列をコマンドラインに追加することから複雑な編集シーケンスのショートカットを作ることまで大きな柔軟性を実現した。Readline は `self-insert` にバインドされた文字を編集バッファに格納する。表示するときには、これが画面の難行かを占めることがある。

Readline が管理する文字バッファや文字列は C の `char` だけによるものであり、必要に応じてそこからマルチバイト文字を組み立てる。内部的には `wchar_t` は使っていない。速度や記憶容量を考慮したことでも理由のひとつだが、編集のコードが書かれた頃にはまだマルチバイト文字のサポートがそれほど広まっていなかったという理由もある。マルチバイト文字をサポートするロケールでは、`readline` が自動的にマルチバイト文字全体を読み込んで編集バッファに追加する。マルチバイト文字を編集コマンドとしてバインドすることも可能だが、バインドするときにはキーシーケンスとして指定しなければならない。可能ではあるが難しいことであり、通常はそんなことをしようとは思わないだろう。たとえば `emacs` や `vi` のコマンドでもマルチバイト文字は使われていない。

キーシーケンスが最終的に編集コマンドに解決されたら、`readline` はターミナルの表示を更新して結果を反映させる。コマンドの結果が文字をバッファに文字を入れるものであったとしても編集位置を移動させるものであったとしても、あるいは行の一部あるいは全体を書き換えるものであったとしても、これは同様に発生する。バインド可能な編集コマンドの中には、履歴ファイルの編集などのように編集バッファには何も変更を加えないものもある。

ターミナルの表示内容の更新は、一見シンプルなようだが実はかなり複雑だ。Readline は三つの内容を気にかけねばならない。画面に表示されている文字バッファの現在の状態、表示バッファの更新後の内容、そして実際に表示されている文字だ。マルチバイト文字があるため、表示されている文字はバッファの内容と正確に一致するとは限らず、再表示エンジンはそのことを考慮しなければならない。再表示するときに `readline` は、現在の表示バッファの内容と更新されたバッファを比較して差分を算出し、更新後のバッファを表示に反映させるのに最適な方法を決めなければならない。この問題には長年悩まされてきた(文字列から

文字列への修正問題)。Readline は次のようにしている。まずバッファの異なる部分の最初と最後の位置を見つけ、その部分だけを更新してカーソルを前後に移動させるコストを算出し(例: ターミナルのコマンドを発行して文字を消してから新しい文字を追加するのと、単に現在の画面表示を上書きしてしまうのとどちらが効率的か?)、もっともコストの低い方法で更新し、必要に応じて最終行の残りの文字を削除してカーソルを正しい位置に移動させる。

再表示エンジンは、readline の中で間違いなく最も頻繁に変更が入っている部分であろう。変更の大半は機能追加である—最も重大なのは、プロンプト内で非表示文字(色の変更など)を扱える機能やマルチバイト文字の対応だ。

Readline は編集バッファの中身を呼び出し元のアプリケーションに返す。そしてアプリケーションが、おそらく変更されているであろう結果を履歴リストに保存する。

アプリケーション側からの **Readline** の拡張

Readline がユーザーに対してその振る舞いをカスタマイズするさまざまな手段を提供しているのと同様に、アプリケーションに対してもその機能群を拡張する仕組みをいくつか用意している。まず、バインド可能な readline の関数は標準の引数を受け取ることができ、指定した結果を返すことができる。これを使えば、アプリケーション側で readline を拡張してそのアプリケーションに合わせた関数を作りやすくなる。たとえば bash では 30 以上のバインドコマンドを追加しており、bash 固有の単語補完からシェルの組み込みコマンドへのインターフェイスまでさまざまなものを使っている。

アプリケーションから readline の振る舞いを変更する二番目の方法は、フック関数へのポインタに既知の名前と呼び出しインターフェイスを使うことだ。アプリケーションが readline の内部動作の一部を置き換え、readline の前に割り込み、アプリケーション固有の変換をさせることができる。

非インタラクティブな入力の処理

シェルが readline を使っていない場合は、stdio あるいは自前のバッファ入力ルーチンを使って入力を取得する。シェルが対話モードでない場合は、stdio よりも bash のバッファ入力パッケージを使うことをお勧めする。なぜなら、Posix は入力の取り込みに奇妙な制約を課すからである。シェルが取り込むのはコマンドのパースに必要な部分だけで、残りはそのまま実行プログラムに渡さなければならない。これは、シェルがスクリプトを標準入力から読み込んでいるときに特に重要となる。シェルは、入力をバッファリングすることを許されている。ただし、ファイルのオフセットをパーサが処理済みの最後の文字の直後にまで戻せる場合に限る。現実的な意味合いで言うと、これはつまり次のような意味である。パイプなどのシーク不能なデバイスからスクリプトを読み込む場合は一文字ずつ読み込まなければならず、ファイルなどから読み込む場合は好きなだけバッファリングできるということだ。

これらの特殊な点を別として、非インタラクティブな入力のシェルでの処理は readline と同様である。つまり、改行文字で区切られる文字列のバッファとして扱う。

マルチバイト文字

マルチバイト文字の処理がシェルに追加されたのは、最初に実装が始まってからかなりの時間がたった後のことだった。この機能は、既存のコードに与える影響を最小限にするよう設計された。マルチバイトをサポートしたロケールにいるときは、シェルへの入力はバイト(Cのchar)のバッファとして格納するが、その中身がマルチバイト文字である可能性も考慮するようになる。Readlineは、マルチバイト文字の表示方法を知っている(鍵となるのは、マルチバイト文字が画面上で一文字あたりどの程度の場所をとるかということと、画面に表示するときにバッファから何バイト取り出すべきかということだ)し、前後に移動するときにもバイト単位ではなく文字単位になることなども知っている。それ以外に、マルチバイト文字がシェルの入力処理に影響を及ぼすことはない。シェルのその他の部分については後ほど説明するが、マルチバイト文字を考慮にいれた処理が必要となる。

3.4 パース

パースエンジンが最初にする仕事は字句解析、つまり文字のストリームを単語に区切ってそれに意味を与えるということだ。単語は、パーサが何らかの操作をするときの基本単位となる。単語とはメタ文字で区切られた文字列のことである。メタ文字には、スペースやタブといったシンプルな区切り文字のほかにシェル言語で特殊な意味を持つ文字(セミコロンやアンパンドなど)がある。

シェルについての歴史的な問題は、Tom Duff が rc(Plan 9 のシェル)に関するペーパーで述べたとおり、Bourne shell の文法を完全に理解している人が誰もいないということである。Posix シェル委員会は Unix シェルの完全な文法を公開するというすばらしい業績を残した。しかしこの文法には、コンテキストに依存する部分が大量にある。この文法に問題がないわけではない—過去の Bourne shell がエラーなしで許容していた構文のいくつかは許可されていない—が、我々が知る限り最善のものだ。

bash のパーサは Posix の文法の初期版に由来するもので、私の知る限りで唯一の、Yacc あるいは Bison で実装された Bourne シェルパーサである。それ故の困難も存在する—シェルの文法は yacc 形式のパースとはあまり相性がよくなくて、複雑な字句解析を必要とするしパーサと字句解析器との連携も多くなる。

いずれにせよ、字句解析器は入力を readline あるいは他のソースから受け取り、メタ文字でトークンに切り分け、コンテキストにあわせてトークンを識別し、それをパーサに渡して文やコマンドとして組み立てることになる。多くの部分はコンテキストに依存する—たとえば for という単語は、予約後かもしれないし識別子かもしれない。あるいは代入文や他の単語の一部かもしれない。次の例はコマンドとしてまったく問題のないものである。

```
for for in for; do for=for; done; echo $for
```

これは `for` と表示する。

ここで、余談としてエイリアスについて説明しよう。Bash では、シンプルなコマンドの先頭の単語をエイリアスで任意のテキストに置き換える。エイリアスは完全に単語なので、エイリアスを使えば(あるいは悪用すれば)シェルの文法を変えてしまうこともできる。たとえば、`bash` が提供していない複合コマンドをエイリアスで実装することもできる。`bash` のパーサはエイリアスを完全に解析フェーズで実装しているので、パーサは解析器に対してエイリアスの展開が許可されたことを通知しなければならない。

他の多くのプログラミング言語と同様に、シェルでも文字をエスケープして特殊な意味を取り除くことができる。エスケープすれば、&のようなメタ文字をコマンド内で使えるようになる。クオートには三種類の方法があり、クオートしたテキストの扱いがそれぞれ少しずつ異なる。バックスラッシュは、それに続く一文字をエスケープする。シングルクオートは、囲まれた文字をすべてそのまま扱う。ダブルクオートもほぼ同様だが、特定の単語の展開は行う(そしてバックスラッシュの扱いが異なる)。字句解析器は、クオートされた文字や文字列をパーサ側で予約語やメタ文字として扱われないようにする。それ以外に特殊な扱いをするのが\$'...' と \$"..."だ。前者はバックスラッシュでエスケープされた文字を ANSI C の文字列と同じように展開し、後者は標準の国際化関数を使って文字を翻訳する。前者は幅広く使われているが、後者はあまり使われていない。実際の使いどころがほとんどないからであろう。

パーサと字句解析器の残りのインターフェイスはそれほど難しいものではない。パーサはある程度の量の状態を符号化して解析器と共有し、文法上必要となるコンテキスト依存の解析を行う。たとえば、字句解析器はトークンの型に応じて単語を分類している。(適切なコンテキストにおける)予約語、通常の単語、代入文などである。これを実現するために、パーサは字句解析器に次のようなことを伝えなければならない。コマンドのペースがどこまで進んだか、複数行の文字列(“ヒアドキュメント”と呼ばれることがある)を処理しているところかどうか、条件分岐の中にいるかどうか、シェルパターンを展開したものを処理しているのか複合代入文を処理しているのかなどである。

ペース段階でのコマンドの置換が終わったことを判断する作業のほとんどは、ひとつの関数(`parse_comsub`)にまとめられている。この関数は恐ろしいほどの量になるシェルの構文を知っており、トークン読み込みのコード以上に重複がある。最適化されているとはとても言えない。この関数はヒアドキュメントやシェルのコメントについて知っていなければならぬし、それだけでなくメタ文字や単語の区切り、クオート処理、予約語が使えるかどうか(つまり、今 `case` 文の中にいるかどうか)なども知っていなければならない。これらを正しく処理できるようになるまでには時間がかかった。

単語の展開の際にコマンド置換を展開するときには、`bash` はパーサを使って言語構造の終了位置を見つける。文字列を `eval` 用のコマンドに変換するのに似ているが、この場合は文字列の最後でコマンドが終わるわけではない。これを正しく動作させるには、パーサが右かっ

こをコマンドの終端を認識しなければならない。これは多くの文法導出に例外条件を追加することにつながり、字句解析器は(適切なコンテキストにおける)右かっこに EOF を表すフラグを立てなければならなくなる。パーサはまた、`yyparse` を再帰的に起動する前にパーサの状態を保存しておかなければならない。コマンドの置換は、コマンドを読み込む際のプロンプト文字列の展開の一部として発生することもあるからである。入力関数は先読みを実装しているので、この関数は最終的に `bash` の入力ポインタを正しい位置まで巻き戻さないといけない。入力を文字列やファイルから読み込んでいるときでもターミナルから `readline` で読み込んでいるときでも同じだ。これが重要なのは、単に入力を読み落とさないようにするためにというだけではない。コマンド置換の展開関数が実行用の正しい文字列を組み立てられるようにするためでもある。

同様の問題が、プログラマブルな単語補完でも発生する。これは、あるコマンドのパース中に別の任意のコマンドを実行できるようにするものだ。この問題を解決するために、起動の館にパーサの状態を保存して後で復元している。

クオート処理もまた、非互換性や論争の元となるものだ。Posix シェルの標準規格が最初に発表されてから 20 年がたつが、標準化ワーキンググループのメンバーはいまだにクオート処理の適切な振る舞いについて議論を続けている。先に述べたように、Bourne シェルがその振る舞いの参考にしているのはリファレンス実装だけである。

パーサが返すのはコマンドを表す C の構造体(ループのような合成コマンドの場合は、その中にさらに別のコマンドが含まれる場合もある)で、それがシェル操作の次のステージ、つまり単語の展開処理に渡される。コマンド構造体は、コマンドオブジェクトおよび単語のリストで構成されている。単語のリストの大半は、コンテキストによってさまざまに変換される。その詳細は次のセクションで説明する。

3.5 単語の展開

パースが終わったら、実行の前に、パース段階で生成された単語の多くを展開することになる。つまり、(たとえば)`$OSTYPE` を文字列 "linux-gnu" に置き換えたりするような処理だ。

パラメータおよび変数の展開

変数の展開は、ユーザーにとってもつともなじみ深いものだ。シェルの変数にはほとんど型付けがなく、わずかな例外を除いて文字列として扱われる。この展開では、パラメータおよび変数の文字列を新たな単語や単語リストに変換する。

展開は、変数の値そのものに対して行われる。プログラマーは、これらを使って変数の値の部分文字列を生成したり値の長さを取得したり、指定したパターンにマッチする部分を先頭あるいは末尾から削除したり、値が指定したパターンにマッチする部分を新しい文字で置き換えたり、アルファベットの大文字小文字を変更したりする。

さらに、変数の状態に依存する展開処理もある。変数に値が設定されているか否かによって、展開や代入の内容が異なってくる。たとえば\${parameter:-word}は、もし設定されていれば parameter と展開されるが、設定されていなかったり空の文字列が設定されている場合は word と展開される。

その他いろいろ

Bash はそれ以外にもさまざまな展開を行い、それぞれについて独自の変な規則に従っている。最初に処理されるのはプレースの展開で、これは

```
pre{one,two,three}post
```

のような文字列を次のように展開する。

```
preonepost pretwopost prethreepost
```

コマンドの置換も行われる。これは、シェルの機能であるコマンドの実行と変数の操作をうまく組み合わせたものだ。シェルがコマンドを実行してその結果を収集し、その出力を使って値の展開をする。

コマンド置換の問題のひとつは、コマンドを直接実行してその処理が完了するまで待ち続けるということだ。シェルからコマンドに対して入力を送る簡単な方法はない。Bash では、プロセス置換という機能を使うことができる。これはコマンド置換とシェルのパイプラインを組み合わせたような機能で、コマンド置換のこれらの欠点を埋め合わせるために使える。コマンド置換と同様に bash がコマンドを実行するが、そのコマンドはバックグラウンドプロセスで動作し、処理が完了するまで待つことはない。この機能の鍵となるのは、bash がコマンドへのパイプを開いて読み書きをしたり、ファイルとして公開して展開の結果を記録したりするという点だ。

次に行われるるのはチルダの展開である。当初の意図は、たとえば~alan を Alan のホームディレクトリに変換するというものであった。しかし年月を経てこの機能は成長し、今ではさまざまなディレクトリを指すようになっている。

最後に行われるのが算術式の展開である。\$((expression)) とすると、expression の部分を C 言語の式と同じルールで評価し、その評価結果を使って展開する。

変数の展開は、シングルクオートとダブルクオートの違いが最も明確にあらわれる処理だ。シングルクオートは一切の展開を禁止する—囲まれた部分は何も変更されずにそのまま展開処理を通過する—が、ダブルクオートの場合はいくつかの展開は許可した上でそれ以外の展開を禁止する。単語の展開やコマンド、算術式、プロセスの置換は行われる—ダブルクオートは、その結果の扱い方にだけ影響を及ぼす—が、プレースやチルダの展開は行われない。

単語の分割

単語を展開した結果は、シェル変数 IFS 内の文字を区切りとして分割される。これを用いて、シェルはひとつの単語を複数の単語に変換する。`$IFS` 内のいずれかの文字¹が結果の中に登場するたびに、bash はそこで単語をふたつに分割する。シングルクオートあるいはダブルクオートで囲まれている場合は、この分割は行われない。

グロブ

結果を分割した後でシェルは、展開された単語をパターンとして解釈し、ファイル名(ディレクトリパスも含む)とのマッチを試みる。

実装

シェルの基本構造がパイプラインにそったものなら、単語の展開は自分自身に向けた小さなパイプラインとなる。単語の展開における各ステージは、単語を受け取って何らかの変換を施し、それを次の展開ステージに渡す。すべての単語展開が終わったら、コマンドを実行する。

bash の単語展開の実装は、既に説明済みの基本的なデータ構造をもとにしている。パーサが出力した単語群は個別に展開され、その結果の単語が入力となる。WORD_DESC 構造体には、单一の単語展開をカプセル化するために必要な情報をすべて保持できる。flags を使って単語展開ステージに必要な情報を符号化し、情報を次のステージに引き継ぐ。たとえば、展開ステージとコマンド実行ステージでは、パーサがフラグを使って特定の単語がシェルの代入文であることを伝える。また、単語展開のコードでは、このフラグを内部的に使って、単語の分割を禁止したりクオートした null 文字列 ("\$x"。ただし \$x は未設定あるいは null 値が設定されている) の存在を示したりする。展開される各単語に対して文字列を用意し、何らかの文字符号化で追加情報を表すようになどしていたら、もっと難しくなっていたことだろう。

パーサと同様、単語展開のコードもマルチバイト文字を正しく扱うことができる。たとえば、変数の長さの展開 (`#$variable`) は、バイト数ではなく文字数を数える。展開のコードは、展開の終わりやマルチバイト文字列で特別な意味を持つ文字を正しく識別する。

3.6 コマンドの実行

bash の内部パイplineにおけるコマンド実行ステージは、実際のアクションが発生する場所である。ほとんどの場合、展開された単語群はコマンド名と引数群に分けられる。OS に

¹たいていの場合は、いずれかの文字の列となる。

渡すときには、コマンド名の部分が読み込んで実行するファイルとなり、残りの単語は `argv` の残りの要素となる。

ここまで説明は意図的に、Posix が単純なコマンド—コマンド名と引数セットからなるコマンド—を呼ぶ場合に重点を置いている。この手のコマンドが最も一般的であるからそうしたのだが、`bash` にはそれ以外のコマンドもある。

コマンド実行ステージへの入力は、パーサが組み立てたコマンド構造と、展開された単語群のセットとなる。ここからが、真の `bash` プログラミング言語の出番だ。このプログラミング言語は先ほど説明したような変数や展開を使うし、高級言語と聞いて一般に思い浮かべるような言語構造を実装している。ループや条件分岐、グルーピング、選択、パターンマッチングによる条件付きの実行、式の評価、そしてシェル特有のいくつかの言語構造などだ。

リダイレクト

OS とのインターフェイスとしてのシェルの役割のひとつを反映しているのが、起動したコマンドの入出力に対するリダイレクト機能である。リダイレクトの構文は、初期のシェル利用者の洗練度を表すものだった。つい最近まで、リダイレクトを使う場合は自分が使っているファイルディスクリプタをきちんと把握しておく必要があったのだ。標準入出力と標準エラー出力だけでなく、それ以外のファイルディスクリプタも番号で明示しなければならなかつた。

最近追加されたリダイレクト構文によって、シェルに適切なファイルディスクリプタを選ばせてそれを指定した変数に代入できるようになり、ユーザーがファイルディスクリプタを指定する必要はなくなった。そのおかげでプログラマがファイルディスクリプタを意識する面倒は減らせたが、新たな処理が増えた。シェルがファイルディスクリプタを正しい場所に複製し、指定した変数にそれを代入しなければならなくなつた。これは、字句解析器からパーサを通してコマンド実行まで情報を渡していく方法を示すもうひとつの例となる。解析器が変数代入を含むリダイレクトとして単語を識別し、パーサは構文の構築時にリダイレクトオブジェクトを作る。このオブジェクトには代入を要するという意味のフラグを立てる。そして、リダイレクトのコードがそのフラグを読み取り、ファイルディスクリプタの番号を正しい変数に代入する。

リダイレクトの実装で最も難しい部分は、リダイレクトを取り消す方法を覚えておくことだ。シェルは、ファイルシステムから実行して新しいプロセスを立ち上げるコマンドとシェル自身が実行する(組み込みの)コマンドの区別を意図的に曖昧にしている。しかし、コマンドの実装がどうであろうと、リダイレクトの効果をそのコマンドが完了した後までひきずつてはいけない²。したがって、シェルはリダイレクトの効力を取り消す方法を覚えておかなければならない。さもないと、シェルの組み込みコマンドの出力をリダイレクトしたときにシェルの標準出力が変わってしまう。`Bash` は、リダイレクトの形式ごとにその取り消し方法を知っている。割り当てられたファイルディスクリプタをクローズするか、あるいは複製さ

²組み込みコマンド `exec` はこのルールの例外だ。

れたファイルディスクリプタを保存してあとで `dup2` を使って復元するかのいずれかだ。これらはいずれもパーサが作った同じリダイレクトオブジェクトを使い、同じ関数で処理される。

複数のリダイレクトを単純にオブジェクトのリストで実装しているので、取り消しに使うリダイレクトは別のリストとして保持する。このリストはコマンドが完了したときに処理されるが、シェルはそれがいつ処理されるのかを気にかけねばならない。シェルの関数や組み込みコマンド“.”に関連づけられたリダイレクトは関数や組み込みコマンドが完了するまで有効にしておかなければならぬからである。コマンドを起動しないときは、組み込みコマンド `exec` は取り消し用のリストを破棄する。`exec` に関連づけられたリダイレクトはシェルの環境に残り続けるからだ。

他にもやつかいなことがあるが、それは `bash` のせいではない。過去のバージョンの Bourne シェルでは、ユーザーが操作できるファイルディスクリプタが 0 から 9 までだけだった。10 以上は、シェルが内部的に使うために予約されていたのだ。Bash ではこの制限を緩め、プロセスのファイルオープンの上限に達するまで任意のディスクリプタを操作できるようにした。つまり、`bash` は自身が保持する内部のファイルディスクリプタ(直接シェルからではなく、外部のライブラリからオープンしたディスクリプタも含む)を覚えておいて必要に応じて移動できるようにしなければならないということになる。管理しなければならないことが増えるし、`close-on-exec` フラグのような仕組みも必要になる。さらに新たにリダイレクトのリストを用意して、コマンドの実行期間や処理済みか破棄されたかなどを管理しなければならない。

組み込みコマンド

Bash には、多数のコマンドがシェル自身の一部として組み込まれている。これらのコマンドはシェルから実行されるもので、新たなプロセスは立ち上げない。

コマンドを組み込みで用意する主な理由は、シェルの内部状態を保ったり変更したりするためだ。`cd` がよい例である。Unix の授業で最初に行う典型的な課題は、なぜ `cd` が外部コマンドとして実装できないのかを説明させることだ。

Bash の組み込みコマンドは、シェルのその他の部分と同じ内部プリミティブを使う。組み込みコマンドは C 言語の関数を使って実装されており、この関数は単語のリストを引数として受け取る。単語リストは単語展開ステージの出力する結果であり、組み込みコマンド側ではそれをコマンド名とその引数として解釈する。組み込みコマンドが使う展開ルールはほとんどが他のコマンドと同じ標準的なものだが、いくつか例外がある。`bash` の組み込みコマンドの中で代入文を引数として受け取れるもの(`declare` や `export` など)は、代入の引数を展開するときにはシェルの変数代入のときと同じルールを使う。ここでも `WORD_DESC` 構造体の `flags` を使い、シェルの内部パイプライン上でステージからステージへと情報を渡す。

単純なコマンドの実行

単純なコマンドは、最もよく見かける形式のコマンドである。ファイルシステムからのコマンドの読み込みやその実行、そして終了ステータスの取得について説明すれば、ここまで説明してこなかったシェルの機能の多くをカバーすることになる。

シェルの変数への代入(つまり、`var=value` 形式の単語)は、それ自身が単純なコマンドの一種である。代入文はコマンド名の前に書くこともできるし、コマンドラインでそれ単体で使うこともできる。コマンドの前に書いた場合は、その変数が後に続くコマンドの実行環境内で渡される(組み込みコマンドやシェル関数の前に書いた場合は、いくつかの例外を除いて、その変数が有効なのは組み込みコマンドあるいは関数の実行中だけとなる)。もし代入文のあとにコマンド名を続けなければ、その代入文はシェルの状態を変更する。

指定したコマンド名がシェルの関数や組み込みコマンドにないものであった場合、`bash` はファイルシステムから、その名前の実行可能なファイルを探す。環境変数 `PATH` の値は、このときの検索先を表すディレクトリをコロン区切りでつなげたリストである。スラッシュ(あるいはそのあのディレクトリ区切り文字)を含むコマンド名は検索対象外となるが、直接実行することはできる。

`PATH` の検索でコマンドが見つかれば、`bash` はそのコマンド名とフルパス名をハッシュテーブルに保存する。ハッシュテーブルの内容は、その後に `PATH` の検索が発生したときに検索の前に参照される。コマンドが見つからない場合、もし特別な名前の関数が定義されていれば `bash` はその関数を実行する。コマンド名と引数を、この関数への引数として渡す。Linux のディストリビューションの中には、この機能を使って存在しないコマンドをインストールさせようとするものもある。

実行するファイルが見つかれば、`bash` はそれをフォークして新しい実行環境を作り、この新しい環境でプログラムを実行する。実行用の環境はシェルの環境を完全に複製したもので、シグナルの処理やリダイレクトでオープンしたりクローズしたりしたファイルなどのちょっとした修正が入っている。

ジョブ制御

シェルによるコマンドの実行には二通りの方法がある。まずはフォアグラウンドでの実行で、これはコマンドが終了するまで待ってその終了ステータスを受け取る。もうひとつはバックグラウンドでの実行で、シェルはすぐに次のコマンドを読み込むことができる。ジョブ制御とは、プロセス(実行されたコマンド)をフォアグラウンドとバックグラウンドの間で移動したり実行の一時停止や再開をしたりする機能のことである。この機能を実装するために `bash` はジョブという概念を導入した。ジョブとは、基本的にはひとつあるいは複数のプロセスから実行されたコマンドのことである。たとえばパイプラインでは、構成する各要素に対してひとつずつプロセスを使う。プロセスグループを使い、個々のプロセスをひとつにまとめて单一のジョブとする。ターミナルは自身に関連付けられたプロセスグループの ID を持つて

いる。つまり、フォアグラウンドプロセスグループとは、ターミナルと同じプロセスグループ IDを持つプロセスグループのことである。

シェルは、いくつかのシンプルなデータ構造を使ってジョブ制御を実装している。子プロセスを表す構造体には、そのプロセス ID や状態、終了時に返すステータスなどが含まれている。パイプラインは、単にこのプロセス構造体をシンプルな連結リストにしただけのものだ。ジョブもまったく同様で、プロセスのリストとジョブの状態(実行中、停止中、終了など)、そしてジョブのプロセスグループ ID で管理されている。プロセスのリストは、通常は単一のプロセスだけで構成されている。パイプラインの場合だけ、複数のプロセスがジョブに関連付けられる。各ジョブはそれぞれ一意なプロセスグループ ID を持つており、ジョブ内のプロセスの中でプロセスグループ ID と同じプロセス ID を持つものがプロセスグループリーダーと呼ばれる。現在のジョブセットは配列に保持されており、この考え方にはユーザーに対する見せ方と非常に似ている。ジョブの状態や終了ステータスは、そのジョブを構成するプロセスの状態や終了ステータスを集約した結果となる。

シェルの他の部分と同様、ジョブ制御を実装するときに複雑になる部分は帳簿管理である。シェルはプロセスを正しいプロセスグループに割り当てねばならないし、子プロセスの作成とプロセスグループへの割り当てを同期させなければならない。またターミナルのプロセスグループも適切に設定しなければならない。ターミナルのプロセスグループがフォアグラウンドジョブを決める(そして、もしシェルのプロセスグループが設定されていなければ、シェル自身がターミナルからの入力を読み込めない)からである。これはとてもプロセス指向な考え方なので、`while` や `for` ループのような複合コマンドを実装してループ全体をひとまとめで開始・停止できるようにするのは難しい。実際、それを実現しているシェルはほとんどない。

複合コマンド

複合コマンドは単純なコマンドのリストで構成されており、`if` や `while` といったキーワードから始まる。複合コマンドのおかげで、シェルのプログラミングの威力を発揮できるようになる。

その実装方法は、特に驚くようなものではない。パーサが複合コマンドに対応するオブジェクトを組み立て、コマンドを解釈するときにはそのオブジェクトを走査していくという方法だ。個々の複合コマンドはそれに対応する C の関数として実装されている。この関数が、適切な展開処理や指定したコマンドの実行、そしてコマンドの返り値に応じた実行フローの切り替えなどを行う。`for` コマンドを実装する関数を例にして説明しよう。この関数は、まず最初に、予約語 `in` に続く単語のリストを展開しなければならない。それから、展開された単語を順にとりあげて適切な変数に代入し、`for` コマンドの本体にあるコマンドのリストを実行することになる。`for` コマンドはコマンドの終了ステータスによって実行を切り替える必要はない。しかし、組み込みコマンド `break` や `continue` の影響には注意する必要がある。

リストにあるすべての単語の処理を終えると、`for` コマンドは処理を返す。これでわかるように、実装の大半はそのコマンドの説明と密接につながっている。

3.7 学んだこと

大切だとわかったこと

私はこれまで 20 年以上 bash にかかわってきて、いくつかのことに気付いた。最も重要なこと—いくら強調してもしそうではないだろう—は、ChangeLog を詳しく書いておくべきだということだ。あとで ChangeLog を読みなおせば、そのときなぜそんな変更をしたのかを思い出せる。さらに、その変更を特定のバグレポートやそれを再現させるテストケースと結び付けられればすばらしい。

もし可能であれば、回帰テストをプロジェクトの立ち上げ時から組み込んでおくことをお勧めする。Bash には数千のテストケースがあり、非インタラクティブな機能のほぼすべてをカバーしている。インタラクティブな機能についてのテストを組み込むことも検討した—Posix には独自の適合性テストスイートが存在する—が、それに必要なフレームワークを配布することになるのは避けたかった。

標準規格は重要だ。Bash は、標準に従った実装をすることで恩恵を受けている。また、自分が実装しているソフトウェアの標準化作業にかかわることも重要だ。さまざまな機能やその振る舞いについて議論できるだけでなく、規格化するときの参考にする標準を持っていればうまくいく。もちろん、それはうまく動かないかもしれない—規格の内容による。

外部の標準も重要だが、内部的な標準を持つこともまた大切だ。私は幸運なことに GNU Project の標準規格に組み込まれ、設計や実装に関する現実的で優れたアドバイスを得ることができた。

よくできたドキュメントも不可欠だ。もし自作のプログラムを他の人に使わせるつもりなら、包括的で明確なドキュメントを作るだけの価値はある。そのソフトウェアが成功して広まると、いろいろなところで大量のドキュメントが作られることになる。そのときには、開発者自身が書いた正式なバージョンがあることが重要となる。

世の中には、よいソフトウェアが豊富にある。使えるものならどんどん使っていこう。たとえば gnulib には、(gnulib フレームワークから取り出せれば) 便利なライブラリ関数が大量に用意されている。BSD や Mac OS X でも同様だ。ピカソも言っていたように「偉大な芸術家は盗む」のだ。

ユーザーコミュニティは大切にしよう。しかし、ときには批判的な声に悩まされることになるかもしれないことに注意。アクティブなユーザーコミュニティからは非常に多くのメリットを得られるが、その結果として人々が熱くなりすぎることもある。個人的に責められていふとは思わないことだ。

私がそれ以外にやったこと

Bash には何百万人ものユーザーがおり、後方互換性の重要性を思い知らされている。ある意味では、後方互換性を保てば決して「ごめんなさい」と言わずに済むともいえる。しかし、世の中はそんなに単純ではない。これまでに、互換性を崩す変更をせざるを得なくなることが何度かあった。そしてそのたびに一部のユーザーから苦情を受けた。しかしそれらの変更は、すべて妥当な理由のあるものだった。間違った決定を正すものだったり、設計時の機能漏れを修正するものだったり、シェルの各パーツ間での非互換性を直すものだったりといった変更だ。初期のうちに、公式な bash 互換性レベル的な何かを出しておけばよかったのだろう。

Bash の開発は、これまでオープンになったことがなかった。私は、マイルストーンリリース (bash-4.2 など) および個別にリリースされるパッチという考え方方に満足している。このようにする理由は次の通りだ。私は、フリーソフトウェアやオープンソースの世界よりは長めのリリース間隔を保つベンダーを相手にしている。そして過去に、ベータ版のソフトウェアが思いのほか広まりすぎてトラブルになることがあった。しかし、もし最初からやり直せるのなら、もう少しこまめにリリースしたり何らかの公開リポジトリを用意したりすることも検討するだろう。

そんなことを挙げていったところで、実現性を考慮しなければ完成しない。今までに何度か検討したけれどもできなかつたことがひとつある。それは、bash のパーサを再帰下降で書きなおし、`bison` を使わないようにすることだ。そうしないとコマンド置換を Posix 準拠にできないのではないかと考えたのだが、最終的にはそんな大規模な変更をしなくても問題を解決できた。もし bash をスクラッチで書きなおすことになつたら、おそらくパーサは自分で書くだろう。そのほうがいろいろ楽になるのは明らかだから。

3.8 結論

Bash は、大規模で複雑なフリーソフトウェアのよい例と言える。20 年を超える開発期間を経て成熟し、かつ強力だ。あらゆる場所で動いており、何百万もの人々が毎日使っている。その多くは bash を使っているということを意識していないだろう。

Bash は多くのソースの影響を受けており、古くは Version 7 Unix のシェル (Stephen Bourne が書いたもの) にまでさかのぼる。最も多くの影響を受けたのは Posix 標準規格で、bash の仕様の大部分は Posix に由来するものだ。過去との互換性を保ちつつ標準に準拠するのは困難なことだった。

GNU Project の一員であることで、Bash は多くの恩恵を受けている。GNU が bash の存在価値を与えてくれたのだ。GNU がなければ bash も存在しなかつただろう。また、bash にはアクティブで活気のあるコミュニティがついている。コミュニティのフィードバックがあったからこそ今日の bash がある—まさにフリーソフトウェアのメリットを体現しているというわけだ。

Berkeley DB

Margo Seltzer and Keith Bostic

コンウェイの法則によると、ソフトウェアの設計はそれを作った組織の構造を反映したものとなるらしい。もう少し話を広げると、こんなことも言えないだろうか。あるソフトウェアがたった二人によって設計されて作り出されたのなら、単に組織の構造を反映しているだけにとどまらず、二人の好みや思想も持ち込まれているのでは? 二人の片割れである Seltzer は、そのキャリアをずっとファイルシステムやデータベース管理システムの世界に捧げている。彼女はきっとこう言うだろう。「その二つは基本的にまったく同じものでしょう? もっと言うなら、OS とデータベース管理システムだってそう。本質的にはどちらも、リソース管理と便利な抽象化をしているに過ぎないのだから」単に、ちょっとした実装の詳細が違う“だけ”だということだ。もう一方の Bostic はツールベースのアプローチによるソフトウェア開発を信じており、シンプルなブロックを組み合わせて部品を作っていくとする。そういうシステムの方が常に、モノリシックなアーキテクチャに比べて各種の「○○性」に優れているからである。そう、理解容易性とか拡張性とか保守性とかテスト可能性とか柔軟性といかいうやつだ。

この考えを組み合わせると、私たち二人がこの二十年間の大半を Berkeley DB に費やしてきたと知っても驚かないだろう。高速で柔軟性・信頼性・拡張性に富むデータ管理用のソフトウェアライブラリ、それが Berkeley DB だ。Berkeley DB は、人々がもっと伝統的なシステム、たとえばリレーションナルデータベースなどに期待する機能の多くを提供する。しかし、そのまとめかたは異なっている。たとえば、Berkeley DB はキーによるアクセスもシーケンシャルアクセスも高速に行え、トランザクションにも対応しているし障害回復機能もある。しかし、そういう機能はライブラリとして提供しており、それを使いたいアプリケーション側で直接リンクして使うようになっている。スタンドアロンのサーバーアプリケーションとして提供しているわけではない。

本章では Berkeley DB の詳細を扱う。さまざまなモジュールの集まりで作られていることや、それぞれが Unix の思想である“ひとつのことをうまくやらせる (do one thing well)”を表現していることもわかるだろう。Berkeley DB を組み込んだアプリケーション側からは、こ

れらのコンポーネントを直接使うこともできるし、あるいはもっとシンプルに、慣れ親しんだ `get`、`put`、`delete` といった操作を通じて暗黙的に使うこともできる。ここでは、そのアーキテクチャに注目する。最初はどう考えたのか、どんな設計をしたのか、そして最終的にどうなったのか、それはなぜか、そういうことだ。設計は、状況に合わせて変化することもあり得る（実際、あったし！）。大切なのは、原則を外さずに一貫したビジョンを持ち続けることだ。また本章では、長期間にわたるソフトウェア開発プロジェクトにおけるコードの成長についても簡単に検討する。Berkeley DB はふたつのディケイドにまたがる現在進行中のプロジェクトであり、必然的に「よい設計」にも打撃を与えている。

4.1 はじまり

Berkeley DB が生まれたのは、まだ Unix オペレーティングシステムが AT&T に独占されていた頃のことだ。何百ものユーティリティやライブラリがあったが、厳しいライセンスの制約に縛られていた。Margo Seltzer はその当時カリフォルニア大学バークレー校の大学院生。一方 Keith Bostic はバークレーの Computer Systems Research Group に属していた。当時の Keith が携わっていたのは、AT&T のプロプライエタリなソフトウェアを Berkeley Software Distribution から取り除く作業だった。

Berkeley DB プロジェクトが立ち上がった当時のささやかな目標は、インメモリのハッシュパッケージである `hsearch` やディスク上でのハッシュパッケージである `dbm/ndbm` を置き換えることだった。新しい、そしてより改善されたハッシュ実装を作り、メモリ上でもディスク上でも扱えるようにする。そしてプロプライエタリなライセンスに縛られずに自由に配布できる。それを目指していた。Margo Seltzer が書いた [SY91]hash ライブラリは、Litwin の Extensible Linear Hashing に関する研究を利用したものだった。巧妙な方法で一定時間でのハッシュ値とページアドレスのマッピングを実現しただけでなく、ハッシュのパケットやファイルシステムのページサイズ（たいていは 4k バイトや 8k バイト）を超える大きなアイテムも扱えるようにした。

ハッシュテーブルがうまくいくのなら、Btree とハッシュテーブルの組み合わせはもつとうまくいくだろう。同じくカリフォルニア大学バークレー校の大学院生だった Mike Olson はこれまでに数多くの Btree を実装しており、新たにまた実装することにも同意してくれた。我々三人は、Margo の hash ソフトウェアと Mike の Btree ソフトウェアをアクセスメソッドに依存しない API に変換した。アプリケーション側からはデータベースハンドルを通じてハッシュテーブルあるいは Btree を参照でき、データベースハンドルがデータの読み込みや変更をするメソッドを保持していた。

これら二つのアクセスメソッドを元にして、Mike Olson と Margo Seltzer は研究論文を共著した ([SO92])。この論文で論じているのは LIBTP で、これはアプリケーションのアドレス空間で動くプログラムのトランザクションライブラリである。

ハッシュと Btree のライブラリは最終的に 4BSD のリリースに組み込まれ、ここで Berkeley DB 1.85 と名付けられた。偽善的には Btree アクセスマソッドが実装しているのは B+link 木

だったが、本章では今後 Btree と呼ぶことにする。アクセスメソッドの名前がそうなっているからだ。Berkeley DB 1.85 の構造や API は、これまでに何らかの Linux や BSD 系システムを使ったことがある人にはなじみやすいものだろう。

Berkeley DB 1.85 ライブライアリはその後数年ほとんど動きがなかったが、1996 年に Netscape が Margo Seltzer や Keith Bostic と契約を結び、LIBTP の論文で示された完全にトランザクショナルな設計による商用に耐えるバージョンを作ることになった。その結果できあがつたのが、初めてトランザクションをサポートしたバージョンである Berkeley DB 2.0 だ。

それ以降の Berkeley DB の歴史はシンプルで、よくありがちな流れになっている。Berkeley DB 2.0 (1997) で、トランザクションが Berkeley DB に導入された。Berkeley DB 3.0 (1999) は新たに設計しなおしたバージョンで、さらに高度な抽象化によってさまざまな機能の増加に対応した。Berkeley DB 4.0 (2001) ではレプリケーションや高可用性に関する機能が導入され、Oracle Berkeley DB 5.0 (2010) では SQL が使えるようになった。

本章の執筆時点では、Berkeley DB は世界で最も広く使われているデータベースツールキットである。何億ものコピーが、ルーター や ブラウザ から メールソフト や OS まであらゆるところで動いている。20 年以上たった現在でも Berkeley DB のツールベースでオブジェクト指向なアプローチは現役であり、自身をインクリメンタルに改良して使う側のソフトウェアの要求に応え続けている。

設計講座 1

複雑なソフトウェアパッケージをテストしたり保守したりしていく上で必須なのが、ソフトウェアをうまくモジュール分割した設計にしておいてモジュール間をよくできた API で連携させることだ。モジュールの境界はニーズにあわせて移動できる(移動できるべき!)が、常に境界が必須だというわけではない。モジュールの境界があれば、ソフトウェアがメンテナンス不能な spaghetti 状態になることを防げる。Butler Lampson はかつてこう言った。「計算機科学の世界のあらゆる問題は、別のレベルの手段で解決できる」。さらに、何かがオブジェクト指向であるとはどういうことなのかと聞かれた Lampson の答えは「API を介して複数の実装を持てるということさ」だった。Berkeley DB の設計や実装はこの考え方を具現化したものであり、複数の実装を共通のインターフェイスで扱えるようになっている。オブジェクト指向の見た目を持っているが、実際のところこのライブラリは C で書かれているのだ。

4.2 アーキテクチャの概要

この節では Berkeley DB ライブライアリのアーキテクチャを取り上げる。まずは LIBTP から始め、そしてその進化の鍵となる局面を強調する。

図 4.1 は Seltzer と Olson の最初の論文から引用したもので、当初の LIBTP のアーキテクチャを示している。一方図 4.2 は、Berkeley DB 2.0 で計画していたアーキテクチャである。



図 4.1: LIBTP プロトタイプシステムのアーキテクチャ



図 4.2: Berkeley DB-2.0 で意図していたアーキテクチャ

LIBTP の実装と Berkeley DB 2.0 の設計との間での唯一の大きな違いは、プロセスマネージャーが削除されたという点である。LIBTP ではコントロールするスレッドが自分自身をライブラリに登録してから個々のスレッド/プロセスを同期させなければならず、サブシステム

レベルでの同期処理は用意していなかった。4.4 節で議論するが、オリジナルの設計のほうがうまく機能したかもしれない。



図 4.3: 実際の Berkeley DB 2.0.6 のアーキテクチャ

当初の設計と、実際にリリースされた db-2.0.6 のアーキテクチャ(図 4.3)の違いを見れば、現実には堅牢なリカバリーマネージャーを実装したことがわかる。図の中でグレーで表記されている部分がリカバリーサブシステムである。リカバリー処理には、ドライバの基盤として図中に「recovery」と記されている部分だけではなく redo や undo のルーチンのリカバリーも含まれる。後者は、アクセスメソッドによる操作をリカバーするものだ。後者については“access method recovery routines”と書かれた円で表記している。このように、Berkeley DB 2.0 では一貫性のある設計でリカバリーを扱っている。LIBTP ではこれと対照的に、ログやリカバリーのルーチンを個々のアクセスメソッドごとに手書きしていた。この汎用的な設計により、さまざまなモジュール間でのよりリッチなインターフェイスができあがった。

図 4.4 に Berkeley DB-5.0.21 のアーキテクチャを示す。図中の番号は、表 4.1 にまとめた API の番号を表す。中には当初から変わらない部分も見受けられるが、現在のアーキテクチャは時を経て変化している。新たなモジュールが追加されたり古いモジュールが分割されたり(たとえば、かつての log は log と dbreg に分かれた)、さらにモジュール間の API が激増したりしている。

この 10 年の進化、十回を超える商用リリース、そして何百もの新機能。これらを経て、以

前のバージョンよりもアーキテクチャがかなり複雑化してきた。中でも特筆すべき点をいくつか取り上げる。まずは、レプリケーション機能がまったく新しいレイヤーとしてシステムに追加されたこと。しかしこれは完全にクリーンな状態で追加されており、おもしろいことにシステムの残りの部分で同じ API を使っているところは以前のコードがそのまま動く。二点目が、log モジュールを log と dbreg (database registration) に分割したことだ。この件については 4.8 節で詳しく説明する。三点目が、すべてのモジュール間呼び出しを先頭にアンダースコアをつけた名前空間にまとめたという点だ。これで、システムを使うアプリケーション側が関数名の重複を気にせずに済むようになった。この件については設計講座 6 で解説する。

四点目が、ログ出力サブシステムの API がカーソルベースとなった (log_get API がなくなって、log_cursor API に置き換えられた) という点だ。Berkeley DB はこれまで、ログの読み書きの制御に複数のスレッドを持つことは一度たりともなかつた。したがって、ライブラリはログ内の現在のシーク位置を一か所気にするだけだった。これは決してうまい抽象化であるとは言えず、レプリケーション環境では動作しなくなつた。アプリケーションの API がカーソルを使った反復処理に対応したことで、ログもカーソルを使った反復処理をサポートするようになった。五点目が、アクセスメソッドの中の fileop モジュールがデータベースの作成・削除・リネームをトランザクション内でできるようにしたことだ。この実装をうまくまとめるために何度か試行錯誤をした (そしてまだ満足のいくレベルにはなっていない) が、何度も作りなおした後でモジュール内に組み込んだ。

設計講座 2

ソフトウェアの設計というものは、単に問題の全体像を把握してから解決を試みようという流れを強制するための手段のひとつにすぎない。熟練したプログラマーは、その目的を達成するためにさまざまなテクニックを使う。まずはとりあえず動くものを書いて、最終的にそれを捨ててしまうという人もいるだろう。詳細なマニュアルや設計文書を書くところから始める人もいるだろう。すべての要件を明確化したコードのテンプレートを作つてから、個別に関数やコメントをはめ込んでいく人もいるだろう。Berkeley DB の場合は、まず最初に完全な Unix マニュアルページを作るところから始めた。各アクセスメソッドやその土台となるコンポーネントのマニュアルを、まだ一切コードを書いていない状態で作ったのだ。どんな手法を使うにせよ、コードを書いてデバッグが始まってからプログラムのアーキテクチャについて考えるのは難しい。大規模なアーキテクチャの変更をしようとすると、それまでのデバッグの労力が無駄になってしまふこともよくあるだろう。ソフトウェアのアーキテクチャを考えるときにはコードのデバッグをするときとは異なる考え方方が要求される。そして、デバッグを始めた時点でのアーキテクチャが、通常はリリースの時まで引き継がれることになる。

なぜアーキテクトはトランザクションライブラリをコンポーネントから外に出したのだろう



図 4.4: Berkeley DB-5.0.21 のアーキテクチャ

う？想定されるたったひとつの使い方にあわせてチューニングすればよかつたのではないだろうか？この問い合わせへの答えは次の三つである。まず、外に出せばよりしっかりととした設計を強要できる。次に、コードの間にきちんとした境界がなければ、複雑なソフトウェアパッケージは必然的に退化して保守不能な状態になってしまふ。最後に、ユーザーがそのソフトウェアをどのように使うかなど、完全に予測するのは不可能だ。ユーザーにソフトウェアコンポーネントへのアクセスを許可すると、きっと思いもよらない方法でそれを使われることになる

アプリケーション API				
1. DBP ハンドル操作		2. DB_ENV リカバリー		3. トランザクション API
open get put del cursor		open(... DB_RECOVER ...)		DB_ENV->txn_begin DB_TXN->abort DB_TXN->commit DB_TXN->prepare
各アクセスメソッドが使う API				
4. Lock へ —lock_downgrade —lock_vec —lock_get —lock_put	5. Mpool へ —memp_nameop —memp_fget —memp_fput —memp_fsync —memp_fopen —memp_fclose —memp_ftruncate —memp_extend_freelist	6. Log へ —log_print_record	7. Dbreg へ —dbreg_setup —dbreg_net_id —dbreg_revoke —dbreg_teardown —dbreg_close_id —dbreg_log_id	
リカバリー API				
8. Lock へ —lock_getlocker —lock_get_list	9. Mpool へ —memp_fget —memp_fput —memp_fset —memp_nameop	10. Log へ —log_compare —log_open —log_earliest —log_backup —log_cursor —log_vtruncate	11. Dbreg へ —dbreg_close_files —dbreg_mark_restored —dbreg_init_recover	12. Txn へ —txn_getckpt —txn_checkpoint —txn_reset —txn_recycle_id —txn_findlastckpt —txn_ckp_read
トランザクションモジュールが使う API				
13. Lock へ —lock_vec —lock_downgrade	14. Mpool へ —memp_sync —memp_nameop	15. Log へ —log_cursor —log_current_lsn	16. Dbreg へ —dbreg_invalidate_files —dbreg_close_files —dbreg_log_files	
レプリケーションシステムへの API				
		17. Log から —rep_send_message —rep_bulk_message		18. Txn から —rep_lease_check —rep_txn_applied —rep_send_message
レプリケーションシステムからの API				
19. Lock へ —lock_vec —lock_get —lock_id	20. Mpool へ —memp_fclose —memp_fget —memp_fput —memp_fsync	21. Log へ —log_get_stable_lsn —log_cursor —log_newfile —log_flush —log_rep_put —log_zero —log_vtruncate	22. Dbreg へ —dbreg_mark_restored —dbreg_invalidate_files —dbreg_close_files	23. Txn へ —txn_recycle_id —txn_begin —txn_recover —txn_getckpt —txn_updateckpt

表 4.1: Berkeley DB 5.0.21 の API

だろう。

これ以降のセクションでは Berkeley DB の各コンポーネントについて検討し、その動きを理解して全体の中でどのような位置づけになっているのかを把握する。

4.3 アクセスマソッド: Btree, Hash, Recno, Queue

Berkeley DB のアクセスマソッドが提供する機能は、キー指定による検索や反復処理を、可変長バイト文字列および固定長バイト文字列に対して行うことだ。Btree と Hash は、可変長のキー/バリューペアに対応している。Recno と Queue は、レコード番号/バリューペアに対応している (Recno は可変長の値に対応しているが、Queue がサポートするのは固定長の値だ

けである)。

Btree と Hash の主な違いは、Btree がキーの参照の局所性を提供するのに対して Hash はそうではないということだ。つまり、Btree はほぼすべてのデータセットに対して適切なアクセスメソッドとなる。一方、Hash アクセスマソッドが適切に使えるのは、データセットが大きすぎて Btree インデックス構造すらメモリ内に収まらないような場面となる。当時は、メモリはインデックス構造ではなくデータのために使うほうがよかったのだ。このトレードオフは、1990 年の時点では十分納得できるものだった。そのころのマシンは、現在のものと比べてメインメモリの量が圧倒的に少なかったのだ。

Recno と Queue の違いは、Queue がレコードレベルのロックに対応しているという点だ。その代償として、Queue は固定長の値しか受け付けない。Recno は可変長のオブジェクトに対応しているが、Btree や Hash と同様にページレベルのロックしかサポートしていない。

もともとの Berkeley DB の設計は、いわゆる CRUD 機能 (Create:作成、Read:読み込み、Update:更新、Delete:削除) をキーベースで行うもので、それがアプリケーションからの主なインターフェイスだった。そこに後付けてカーソルを追加し、反復処理に対応させた。その結果としてコードが混乱してしまい、ライブラリの内部で同じようなコードの流れが重複してしまうことになった。時を経てそのコードは保守不能になってしまい、キーベースの操作をすべてカーソル操作で書き換えるはめになった(キー指定による操作は、現在はキャッシュカーソルを割り当てて操作を実行してからカーソルをカーソルプールに返すという流れになっている)。これは、ソフトウェア開発の世界で何度も繰り返されるルールの適用例のひとつである。「コードの可読性を落としたり複雑化させたりしてしまうような最適化は、本当にそれが必要となるまでは決してしてはいけない」ということだ。

設計講座 3

ソフトウェアのアーキテクチャが優雅に成熟していくなんてことはない。ソフトウェアのアーキテクチャは、ソフトウェアに加えた変更の数に正比例して退化する。バグ修正のおかげでレイヤー化が崩れたり、新機能の追加で設計に無理が出てきたりといった具合だ。徐々に崩れていくソフトウェアアーキテクチャに対して、どの時点でモジュールの設計を見直して書き直すべきか。これはとても難しい決断だ。アーキテクチャが退化するにつれて、そのソフトウェアの保守や開発はより難しくなる。最終的にはリリースのたびに大量の手動テスト部隊を動員しなければいけないようなレガシーコードの塊になってしまう。ソフトウェアの内部構造を理解できる人が誰もいなくなってしまうのだ。一方、根本的に書き直そうとすると、使う側から見れば不安定で互換性のない状態に悩まされることになる。ソフトウェアアーキテクトであるあなたに対して確実に保障できるのは、どちらの道を選んだところで結局誰かに恨まれるということくらいだ。

Berkeley DB の各アクセスメソッドの内部構造に関する議論はここでは省略する。単に、よ

く知られている Btree やハッシュのアルゴリズムを実装しているにすぎない (Recno は Btree のコードの上位に重ねたレイヤーである。また Queue はファイルブロックのロックアップ処理だが、レコードレベルのロックを追加したために多少複雑になっている)。

4.4 ライブラリインターフェイス層

長年にわたって機能追加を続けていくうちに、ようやく気付いたことがある。アプリケーションのコードと内部のコードとで同じ機能を共有する必要があるということだ(たとえばテーブルの JOIN 操作では複数のカーソルを使って行の反復処理を行う。一方アプリケーション側でも、カーソルを使って同じ行を反復処理することになるだろう)。

設計講座 4

変数やメソッドや関数の名前の付けかた、コメントの書きかた、そしてコードの書きかた。どんなスタイルを選んでもかまわない。世の中には「よい書きかた」とされている書式やスタイルがいろいろある。そんなことより大切なのは、どう書くかではなく命名規約やコーディングスタイルの一貫性を保つことだ。熟練したプログラマーは、コードのフォーマットやオブジェクトの命名から大量の情報を引き出す。命名規約やスタイルが一貫していないと、コードの内容を読み取るのに時間と労力がかかってしまい、他のプログラマーに誤読されてしまうことになるだろう。内部的なコーディング規約に反するのは、そのシステムに対して発砲しているのに等しい。

というわけで、アクセスメソッド API を分割してきちんと定義された階層に分けることにした。インターフェイスルーチン層は、必要となる汎用的なエラーチェックや関数固有のエラーチェック、インターフェイスの追跡などをすべて行う。それ以外にも、自動トランザクション管理などもこの層の仕事だ。アプリケーションが Berkeley DB の機能を利用するときには、処理対象のオブジェクトが持つメソッドにもとづいた第一レベルのインターフェイスを呼び出す。(たとえば、`__dbc_put_pp` は Berkeley DB のカーソルの “put” メソッドに対応するインターフェイスの呼び出しで、データアイテムを更新する。最後の “_pp” は、アプリケーションから呼び出せる関数であることを示すサフィックスである)。

Berkeley DB がインターフェイス層で行うタスクの一つが、どのスレッドが Berkeley DB ライブラリ内で動作中なのかを追跡することだ。追跡が必要となる理由は、Berkeley DB の内部的な操作の中にはライブラリ内でスレッドが立ち上がりっていないときにしか実行できないものもあるからである。Berkeley DB がライブラリ内のスレッドの動きを追跡する方法は、ライブラリの API をコールするたびにそのスレッドが実行中であるというフラグを立てておき、API コールの結果が返ってきた時点でフラグをクリアするというものだ。この「入場/退

場のチェック」が、常にインターフェイス層で行われる。これは、そのコールがレプリケーション環境から行われたのかをチェックするのと同じ方法である。

当然、こんな疑問が出てくるはずだ。「スレッド ID を直接ライブラリに渡したほうがずっと簡単なのでは?」答えはイエスだ。そのほうがずっと簡単だろうし、できることならそうしたかった。しかし、そんな変更をしてしまうと、Berkeley DB を使ったアプリケーションはすべて修正が必要となる。アプリケーションからの Berkeley DB の機能の呼び出しの大半を書き換えることになり、たいていの場合はアプリケーションの構造を再考することになるだろう。

設計講座 5

ソフトウェアアーキテクトは、アップグレードに伴ういざこざへの対応に気をつける必要がある。新しいリリースに対応するために要する変更が少しで済むのなら、ユーザーはそれを許容するだろう(コンパイル時にエラーが出るようにしておけば、まだアップグレードが完了していないことが明白になる。アップグレードの際によくわからないエラーで悩まされることもなくなるだろう)。しかし、まったく根本的に変化させるには、新しいコードベースを使わざるを得ない。そしてユーザーにも新たなコードベースへの移行を求める事になる。明らかに、アプリケーションを一から書き直してユーザーの移行を促すのは、時間的にリソース的にもたやすいことではない。しかし、実際にはちょっとしたアップグレードなのに大規模な改修が必要となってユーザーを怒らせるといったことはしなくて済む。

インターフェイス層で行われるもうひとつのタスクが、トランザクションの生成だ。Berkeley DB ライブラリがサポートするモードのひとつに、すべての操作が自動生成されたトランザクション内で行われるというモードがある(これによって、アプリケーション側で明示的にトランザクションを作成したりコミットしたりする手間を省ける)。このモードをサポートするには、アプリケーション側からトランザクションを明示せずに API をコールされるたびに、トランザクションを自動生成することになる。

最後に、すべての Berkeley DB API は引数のチェックを要する。Berkeley DB のエラーチェックには二種類ある。ひとつは汎用的なチェックで、直近の操作でデータベースが破壊されていないかどうかを調べたりレプリケーションの状態が変わっている最中(たとえば、書き込み可能なレプリカの切り替え中など)であるかどうかを調べたりといったものだ。もうひとつは個々の API に固有のチェックで、フラグの用法やパラメータの使い方、オプションの組み合わせ、その他実際にリクエストを処理する前に確認できるあらゆるエラーのチェックを行う。

API 固有のチェックはすべて、最後に`_arg`がつく名前の関数にカプセル化されている。つまり、カーソルの`put`メソッドに関するエラーチェックは`__dbc_put_arg`関数にまとめられており、これが`__dbc_put_pp`関数から呼び出される。

最終的に、引数の検証やトランザクションの生成が完了したら、実際の処理を行うワーカーメソッド(今回の場合は`__dbc_put`)を呼び出す。これは、我々がカーソルの`put`機能を内部的に使うときに利用するのと同じ関数である。

このように分割することで、大がかりな操作があったときにレプリケーション環境で実際にどんなアクションが必要となるのかを判断しやすくなつた。コードベースに手を加える作業を数え切れないほど繰り返した結果、すべての事前チェックを分離することができた。そして、今後もし何か問題が発生したときにもより手を加えやすくなつたのだ。

4.5 基盤となるコンポーネント

アクセスメソッドの基盤となるのが、バッファマネージャー・ロックマネージャー・ログマネージャーそしてトランザクションマネージャーの四つのコンポーネントだ。それについて個別に扱うが、これらすべてに共通するアーキテクチャ上の特性もある。

まず、すべてのサブシステムには自前の API が存在する。当初は各サブシステムが自身のオブジェクトハンドルを保持しており、そのハンドルでサブシステムのすべてのメソッドを扱っていた。たとえば Berkeley DB のロックマネージャーを使うと、自身のロックを処理したりリモートのロックマネージャーに書き出したりできる。あるいは Berkeley DB のバッファマネージャーを使うと、自身のファイルページを共有メモリで扱える。時を経て、これらのサブシステム固有のハンドルは API から削除された。Berkeley DB を使うアプリケーションをシンプルにするためだ。サブシステム群は今でも個別のコンポーネントだし他のサブシステムとは独立して使えるようになっているが、今では共通のオブジェクトハンドルである DB_ENV “environment” ハンドルを共有するようになった。このアーキテクチャが、レイヤー化と汎化を強制する。レイヤーの区切りは時とともに変化するし、あるサブシステムが別のサブシステムの役割に立ち入っている部分もいくつか存在する。しかし、プログラマーにとっては、システムの各部分を個別のソフトウェアプロダクトとして考えるのはよいことだ。

次に、すべてのサブシステム(実際のところ、すべての Berkeley DB の関数)はエラーコードをコールスタックに返す。Berkeley DB はライブラリであり、グローバル変数を宣言したりしてアプリケーションの名前空間に立ち入ることができない。言うまでもないが、すべてのエラーをコールスタック経由の単一のパスで返すことはよい習慣である。

設計講座 6

ライブラリを設計する際には、名前空間を利用することが大切である。そうしなければ、ライブラリを使う側のプログラマーからすれば、関数や定数そして構造体やグローバル変数の名前を何十個も覚えておかないとアプリケーション側で名前が衝突してしまうことになる。

最後に、すべてのサブシステムは共有メモリをサポートする。Berkeley DB は複数のプロ

セスからのデータベースの共有をサポートしているので、すべてのデータ構造は共有メモリ上に存在することになる。この方式を選んだことによる最大の影響は、インメモリのデータ構造がポインタではなくベースアドレスとオフセットのペアを使わなければならないということだ。これは、ポインタベースのデータ構造をマルチプロセスのコンテキストでも動作させるために必要となる。言い換えると、ポインタ経由で間接参照するのではなく、Berkeley DB ライブラリはベースアドレス（メモリ内で共有メモリセグメントがマップされたアドレス）にオフセット（マップされたセグメントの中で、そのデータ構造が存在する位置）を足した場所を参照しなければならないということである。この機能をサポートするために、我々は Berkeley Software Distribution queue パッケージの別バージョンを書いた。これはさまざまな種類のリンクリストを実装したものである。

設計講座 7

実際に共有メモリのリンクリストパッケージを書き始める前に、Berkeley DB エンジニアは共有メモリ内でのさまざまなデータ構造を手書きしていた。しかしこれらの実装はどれも脆く、デバッグしにくいものだった。共有メモリのリストパッケージは BSD のリストパッケージ (queue.h) を参考にして作ったもので、今まで苦労した結果をこれがすべて置き換えてくれた。一度デバッグをしてからは、共有メモリのリンクリストに関する問題は二度と発生しなかった。このことから得られる重要な設計指針は次の三つだ。まず、もし複数回登場する機能があるのなら、共有関数を作つてそれを使うようにする。なぜなら、コードの中に同じような機能が複数回登場するということはそのうちのどれかが間違った実装になっているだろうからだ。次に、汎用目的のルーチンを開発するときにはそのルーチン群用のテストスイートを書く。そうすれば、そのルーチンだけを個別にデバッグできる。最後に、書くのが難しいコードであればあるほど、その部分を個別に書いて保守していくことが重要になる。そうしないと、その周辺のコードの影響を受けて浸食されていることを防げなくなる。

4.6 バッファマネージャー: Mpool

Berkeley DB の Mpool サブシステムは、ファイルページを扱うインメモリのバッファプールである。メインメモリが有限なリソースであるという事実を隠し、メモリの量を超える大きなデータベースを扱うときにはデータベースのページをディスクとの間で移動させる。データベースのページをメモリ内でキャッシュすることで、オリジナルの hash ライブラリの性能は劇的に向上した。かつての hsearch や ndbm の実装の性能を大きく上回る。

Berkeley DB の Btree アクセスマソッドは極めて伝統的な B+木の実装であるが、ツリーノード間のポインタはページ番号で表しており、実際のインメモリのポインタではない。その理

由は、ライブラリの実装がインメモリだけではなくディスク上のフォーマットも扱うからである。このようにする利点は、ページをキャッシュからフラッシュするときにフォーマット変換が不要になることだ。逆に欠点は、インデックスを走査する際にバッファプールのルックアップが必要になるということだ。バッファプールのルックアップは、間接メモリ参照よりもコストの高い操作である。

それ以外にも、パフォーマンスに関する暗黙の影響がある。それは、インメモリでのBerkeley DB のインデックスの表現方法が実際のところはディスク上の永続データのキャッシュであるという前提から得られるものだ。たとえば、Berkeley DB がキャッシュされたページにアクセスするときには、まずメモリ内のページを固定する。固定することで、他のスレッドやプロセスがそのページをバッファプールから削除できないようになる。インデックスが完全にキャッシュ内に取まってディスクへのフラッシュが不要だったとしても、Berkeley DB はアクセスのたびにページを固定する。その理由は、Mpool が提供する基盤モデルがキャッシュであって永続ストレージではないからだ。

Mpool のファイル抽象化

Mpool はファイルシステム上で動かすことを前提としており、ファイルを抽象化した API を公開している。たとえば DB_MPPOOLFILE ハンドルはディスク上のファイルを表し、ファイルからページを取得したりページをファイルに書き出したりといったメソッドを提供する。Berkeley DB ではインメモリのテンポラリデータベースに対応しているが、これらもまた DB_MPPOOLFILE ハンドルから参照されている。これは、ベースになる Mpool の抽象化によるものだ。Mpool の主要な API は get メソッドと put メソッドである。get はページがキャッシュ内にあることを確かめてからそのページを固定し、そしてページへのポインタを返す。ライブラリがそのページでの処理を終えたら、put を呼んでページの固定を解除し、明け渡す。初期の Berkeley DB は、読み込み用にページを固定するときと書き込み用にページを固定するときを区別していなかった。しかし、同時実行性を向上させるために Mpool の API を拡張し、呼び出し側でページの書き込み意図を明示できるようにした。このように読み込みアクセスと書き込みアクセスを区別できることは、多版型同時実行制御を実装するために不可欠だ。読み込み用として固定したページはダーティになったとしてもディスクに書き込めるが、書き込み用に固定した場合はできない。整合性を欠く状態になっている可能性があるからである。

ログ先行書き込み

Berkeley DB は、ログ先行書き込み (WAL) をトランザクションの仕組みとして使っており、処理に失敗した場合の復旧ができるようにしている。ログ先行書き込みという用語は、実際にデータを更新する前に変更内容のログをディスクに書き込ませるという方針を表している。Berkeley DB が WAL をトランザクション機構に採用しているということは Mpool に

とって大切な前提となる。Mpool は、汎用的なキャッシングの機能と WAL プロトコルのサポートとの間でバランスをとった設計でなければならない。

Berkeley DB は、すべてのデータページにログシーケンス番号 (LSN) を書き込む。これを使って、特定のページへの直近の変更に対応するログレコードを記録するのだ。WAL を強制すると、Mpool がページをディスクに書き出す前にそのページの LSN に対応するログレコードがディスク上にあることを確認しなければならない。ここで設計上の難題が発生する。この機能を提供しつつ、Mpool のクライアントで Berkeley DB とは違うページフォーマットを使えるようにするにはどうすればいいだろうか。Mpool がこの難題にこたえるために用意したメソッドが set (および get) のコレクションだ。これを使って振る舞いを指示する。DB_MPOOLFILE のメソッド set_lsn_offset はページ内のバイトオフセットを提供し、どこに LSN があるのかを Mpool に指示する。これは、WAL を確実にするために必要なものだ。このメソッドが一切呼ばれなければ、Mpool は WAL プロトコルを強要しない。同様に set_clearlen メソッドは、ページをキャッシング内で作るときにメタデータを何バイトぶん明示的にクリアする必要があるのかを Mpool に伝える。これらの API のおかげで、Mpool は Berkeley DB のトランザクション要件に対応した機能を提供できるようになる。Mpool を使う全ユーザーに対応を要求せずに済むのだ。

設計講座 8

ログ先行書き込みはカプセル化とレイヤー化の一例でもある。たとえその機能がソフトウェアの他の箇所では一切使わないものだったとしても。結局のところ、キャッシングの中の LSN を気にかけているプログラムがいったいどれほどあるのだというのだろう？にもかかわらず、こういう指針は有用だ。ソフトウェアの保守やテスト、デバッグ、拡張をしやすくしてくれる。

4.7 ロックマネージャー: Lock

Like Mpool, the lock manager was designed as a general-purpose component: a hierarchical lock manager (see [GLPT76]), designed to support a hierarchy of objects that can be locked (such as individual data items), the page on which a data item lives, the file in which a data item lives, or even a collection of files. As we describe the features of the lock manager, we'll also explain how Berkeley DB uses them. However, as with Mpool, it's important to remember that other applications can use the lock manager in completely different ways, and that's OK—it was designed to be flexible and support many different uses.

The lock manager has three key abstractions: a “locker” that identifies on whose behalf a lock is being acquired, a “lock_object” that identifies the item being locked, and a “conflict matrix”.

Lockers are 32-bit unsigned integers. Berkeley DB divides this 32-bit name space into transactional and non-transactional lockers (although that distinction is transparent to the lock manager). When Berkeley DB uses the lock manager, it assigns locker IDs in the range 0 to 0x7fffffff to non-transactional lockers and the range 0x80000000 to 0xffffffff to transactions. For example, when an application opens a database, Berkeley DB acquires a long-term read lock on that database to ensure no other thread of control removes or renames it while it is in-use. As this is a long-term lock, it does not belong to any transaction and the locker holding this lock is non-transactional.

Any application using the lock manager needs to assign locker ids, so the lock manager API provides both `DB_ENV->lock_id` and `DB_ENV->lock_id_free` calls to allocate and deallocate lockers. So applications need not implement their own locker ID allocator, although they certainly can.

ロックオブジェクト

Lock objects are arbitrarily long opaque byte-strings that represent the objects being locked. When two different lockers want to lock a particular object, they use the same opaque byte string to reference that object. That is, it is the application's responsibility to agree on conventions for describing objects in terms of opaque byte strings.

For example, Berkeley DB uses a `DB_LOCK_ILOCK` structure to describe its database locks. This structure contains three fields: a file identifier, a page number, and a type.

In almost all cases, Berkeley DB needs to describe only the particular file and page it wants to lock. Berkeley DB assigns a unique 32-bit number to each database at create time, writes it into the database's metadata page, and then uses it as the database's unique identifier in the Mpool, locking, and logging subsystems. This is the `fileid` to which we refer in the `DB_LOCK_ILOCK` structure. Not surprisingly, the page number indicates which page of the particular database we wish to lock. When we reference page locks, we set the type field of the structure to `DB_PAGE_LOCK`. However, we can also lock other types of objects as necessary. As mentioned earlier, we sometimes lock a database handle, which requires a `DB_HANDLE_LOCK` type. The `DB_RECORD_LOCK` type lets us perform record level locking in the queue access method, and the `DB_DATABASE_LOCK` type lets us lock an entire database.

衝突マトリクス

The last abstraction of the locking subsystem we'll discuss is the conflict matrix. A conflict matrix defines the different types of locks present in the system and how they interact. Let's call the entity holding a lock, the holder and the entity requesting a lock the requester, and let's also assume that the holder and requester have different locker ids. The conflict matrix is an array indexed by `[requester][holder]`, where each entry contains a zero if there is no conflict, indicating that the

設計講座 9

Berkeley DB's choice to use page-level locking was made for good reasons, but we've found that choice to be problematic at times. Page-level locking limits the concurrency of the application as one thread of control modifying a record on a database page will prevent other threads of control from modifying other records on the same page, while record-level locks permit such concurrency as long as the two threads of control are not modifying the same record. Page-level locking enhances stability as it limits the number of recovery paths that are possible (a page is always in one of a couple of states during recovery, as opposed to the infinite number of possible states a page might be in if multiple records are being added and deleted to a page). As Berkeley DB was intended for use as an embedded system where no database administrator would be available to fix things should there be corruption, we chose stability over increased concurrency.

requested lock can be granted, and a one if there is a conflict, indicating that the request cannot be granted.

The lock manager contains a default conflict matrix, which happens to be exactly what Berkeley DB needs, however, an application is free to design its own lock modes and conflict matrix to suit its own purposes. The only requirement on the conflict matrix is that it is square (it has the same number of rows and columns) and that the application use 0-based sequential integers to describe its lock modes (e.g., read, write, etc.). 表 4.2 shows the Berkeley DB conflict matrix.

Requester	Holder No-Lock	Read	Write	Wait	iWrite	iRead	iRW	uRead	wasWrite
No-Lock									
Read			✓		✓		✓		✓
Write		✓	✓	✓	✓	✓	✓	✓	✓
Wait									
iWrite		✓	✓					✓	✓
iRead			✓						✓
iRW		✓	✓					✓	✓
uRead			✓		✓		✓		
wasWrite		✓	✓		✓	✓	✓		✓

表 4.2: Read-Writer Conflict Matrix.

階層化ロックのサポート

Before explaining the different lock modes in the Berkeley DB conflict matrix, let's talk about how the locking subsystem supports hierarchical locking. Hierarchical locking is the ability to lock

different items within a containment hierarchy. For example, files contain pages, while pages contain individual elements. When modifying a single page element in a hierarchical locking system, we want to lock just that element; if we were modifying every element on the page, it would be more efficient to simply lock the page, and if we were modifying every page in a file, it would be best to lock the entire file. Additionally, hierarchical locking must understand the hierarchy of the containers because locking a page also says something about locking the file: you cannot modify the file that contains a page at the same time that pages in the file are being modified.

The question then is how to allow different lockers to lock at different hierarchical levels without chaos resulting. The answer lies in a construct called an intention lock. A locker acquires an intention lock on a container to indicate the intention to lock things within that container. So, obtaining a read-lock on a page implies obtaining an intention-to-read lock on the file. Similarly, to write a single page element, you must acquire an intention-to-write lock on both the page and the file. In the conflict matrix above, the `iRead`, `iWrite`, and `iWR` locks are all intention locks that indicate an intention to read, write or do both, respectively.

Therefore, when performing hierarchical locking, rather than requesting a single lock on something, it is necessary to request potentially many locks: the lock on the actual entity as well as intention locks on any containing entities. This need leads to the Berkeley DB `DB_ENV->lock_vec` interface, which takes an array of lock requests and grants them (or rejects them), atomically.

Although Berkeley DB doesn't use hierarchical locking internally, it takes advantage of the ability to specify different conflict matrices, and the ability to specify multiple lock requests at once. We use the default conflict matrix when providing transactional support, but a different conflict matrix to provide simple concurrent access without transaction and recovery support. We use `DB_ENV->lock_vec` to perform lock coupling, a technique that enhances the concurrency of Btree traversals [Com79]. In lock coupling, you hold one lock only long enough to acquire the next lock. That is, you lock an internal Btree page only long enough to read the information that allows you to select and lock a page at the next level.

4.8 ログマネージャー: Log

The log manager provides the abstraction of a structured, append-only file. As with the other modules, we intended to design a general-purpose logging facility, however the logging subsystem is probably the module where we were least successful.

A log is conceptually quite simple: it takes opaque byte strings and writes them sequentially to a file, assigning each a unique identifier, called a log sequence number (LSN). Additionally, the log must provide efficient forward and backward traversal and retrieval by LSN. There are two tricky parts: first, the log must guarantee it is in a consistent state after any possible failure (where consistent means it contains a contiguous sequence of uncorrupted log records); second, because

設計講座 10

Berkeley DB's general-purpose design was well rewarded when we added concurrent data store functionality. Initially Berkeley DB provided only two modes of operation: either you ran without any write concurrency or with full transaction support. Transaction support carries a certain degree of complexity for the developer and we found some applications wanted improved concurrency without the overhead of full transactional support. To provide this feature, we added support for API-level locking that allows concurrency, while guaranteeing no deadlocks. This required a new and different lock mode to work in the presence of cursors. Rather than adding special purpose code to the lock manager, we were able to create an alternate lock matrix that supported only the lock modes necessary for the API-level locking. Thus, simply by configuring the lock manager differently, we were able to provide the locking support we needed. (Sadly, it was not as easy to change the access methods; there are still significant parts of the access method code to handle this special mode of concurrent access.)

設計講座 11

When you find an architectural problem you don't want to fix "right now" and that you're inclined to just let go, remember that being nibbled to death by ducks will kill you just as surely as being trampled by elephants. Don't be too hesitant to change entire frameworks to improve software structure, and when you make the changes, don't make a partial change with the idea that you'll clean up later—do it all and then move forward. As has been often repeated, "If you don't have the time to do it right now, you won't find the time to do it later." And while you're changing the framework, write the test structure as well.

log records must be written to stable storage for transactions to commit, the performance of the log is usually what bounds the performance of any transactional application.

As the log is an append-only data structure, it can grow without bound. We implement the log as a collection of sequentially numbered files, so log space may be reclaimed by simply removing old log files. Given the multi-file architecture of the log, we form LSNs as pairs specifying a file number and offset within the file. Thus, given an LSN, it is trivial for the log manager to locate the record: it seeks to the given offset of the given log file and returns the record written at that location. But how does the log manager know how many bytes to return from that location?

ログレコードの書式

The log must persist per-record metadata so that, given an LSN, the log manager can determine the size of the record to return. At a minimum, it needs to know the length of the record. We prepend every log record with a log record header containing the record's length, the offset of the previous record (to facilitate backward traversal), and a checksum for the log record (to identify log corruption and the end of the log file). This metadata is sufficient for the log manager to maintain the sequence of log records, but it is not sufficient to actually implement recovery; that functionality is encoded in the contents of log records and in how Berkeley DB uses those log records.

Berkeley DB uses the log manager to write before- and after-images of data before updating items in the database [HR83]. These log records contain enough information to either redo or undo operations on the database. Berkeley DB then uses the log both for transaction abort (that is, undoing any effects of a transaction when the transaction is discarded) and recovery after application or system failure.

In addition to APIs to read and write log records, the log manager provides an API to force log records to disk (`DB_ENV->log_flush`). This allows Berkeley DB to implement write-ahead logging—before evicting a page from Mpool, Berkeley DB examines the LSN on the page and asks the log manager to guarantee that the specified LSN is on stable storage. Only then does Mpool write the page to disk.

設計講座 12

Mpool and Log use internal handle methods to facilitate write-ahead logging, and in some cases, the method declaration is longer than the code it runs, since the code is often comparing two integral values and nothing more. Why bother with such insignificant methods, just to maintain consistent layering? Because if your code is not so object-oriented as to make your teeth hurt, it is not object-oriented enough. Every piece of code should do a small number of things and there should be a high-level design encouraging programmers to build functionality out of smaller chunks of functionality, and so on. If there's anything we have learned about software development in the past few decades, it is that our ability to build and maintain significant pieces of software is fragile. Building and maintaining significant pieces of software is difficult and error-prone, and as the software architect, you must do everything that you can, as early as you can, as often as you can, to maximize the information conveyed in the structure of your software.

Berkeley DB imposes structure on the log records to facilitate recovery. Most Berkeley DB log records describe transactional updates. Thus, most log records correspond to page modifications to a database, performed on behalf of a transaction. This description provides the basis for identifying what metadata Berkeley DB must attach to each log record: a database, a transaction, and a record

type. The transaction identifier and record type fields are present in every record at the same location. This allows the recovery system to extract a record type and dispatch the record to an appropriate handler that can interpret the record and perform appropriate actions. The transaction identifier lets the recovery process identify the transaction to which a log record belongs, so that during the various stages of recovery, it knows whether the record can be ignored or must be processed.

抽象化の打破

There are also a few “special” log records. Checkpoint records are, perhaps, the most familiar of those special records. Checkpointing is the process of making the on-disk state of the database consistent as of some point in time. In other words, Berkeley DB aggressively caches database pages in Mpool for performance. However, those pages must eventually get written to disk and the sooner we do so, the more quickly we will be able to recover in the case of application or system failure. This implies a trade-off between the frequency of checkpointing and the length of recovery: the more frequently a system takes checkpoints, the more quickly it will be able to recover. Checkpointing is a transaction function, so we’ll describe the details of checkpointing in the next section. For the purposes of this section, we’ll talk about checkpoint records and how the log manager struggles between being a stand-alone module and a special-purpose Berkeley DB component.

In general, the log manager, itself, has no notion of record types, so in theory, it should not distinguish between checkpoint records and other records—they are simply opaque byte strings that the log manager writes to disk. In practice, the log maintains metadata revealing that it does understand the contents of some records. For example, during log startup, the log manager examines all the log files it can find to identify the most recently written log file. It assumes that all log files prior to that one are complete and intact, and then sets out to examine the most recent log file and determine how much of it contains valid log records. It reads from the beginning of a log file, stopping if/when it encounters a log record header that does not checksum properly, which indicates either the end of the log or the beginning of log file corruption. In either case, it determines the logical end of log.

During this process of reading the log to find the current end, the log manager extracts the Berkeley DB record type, looking for checkpoint records. It retains the position of the last checkpoint record it finds in log manager metadata as a “favor” to the transaction system. That is, the transaction system needs to find the last checkpoint, but rather than having both the log manager and transaction manager read the entire log file to do so, the transaction manager delegates that task to the log manager. This is a classic example of violating abstraction boundaries in exchange for performance.

What are the implications of this tradeoff? Imagine that a system other than Berkeley DB is using the log manager. If it happens to write the value corresponding to the checkpoint record type in the same position that Berkeley DB places its record type, then the log manager will identify that record as a checkpoint record. However, unless the application asks the log manager for that information

(by directly accessing `cached_ckpt_lsn` field in the log metadata), this information never affects anything. In short, this is either a harmful layering violation or a savvy performance optimization.

File management is another place where the separation between the log manager and Berkeley DB is fuzzy. As mentioned earlier, most Berkeley DB log records have to identify a database. Each log record could contain the full filename of the database, but that would be expensive in terms of log space, and clumsy, because recovery would have to map that name to some sort of handle it could use to access the database (either a file descriptor or a database handle). Instead, Berkeley DB identifies databases in the log by an integer identifier, called a log file id, and implements a set of functions, called `dbreg` (for “database registration”), to maintain mappings between filenames and log file ids. The persistent version of this mapping (with the record type `DBREG_REGISTER`) is written to log records when the database is opened. However, we also need in-memory representations of this mapping to facilitate transaction abort and recovery. What subsystem should be responsible for maintaining this mapping?

In theory, the file to log-file-id mapping is a high-level Berkeley DB function; it does not belong to any of the subsystems, which were intended to be ignorant of the larger picture. In the original design, this information was left in the logging subsystems data structures because the logging system seemed like the best choice. However, after repeatedly finding and fixing bugs in the implementation, the mapping support was pulled out of the logging subsystem code and into its own small subsystem with its own object-oriented interfaces and private data structures. (In retrospect, this information should logically have been placed with the Berkeley DB environment information itself, outside of any subsystem.)

設計講座 13

There is rarely such thing as an unimportant bug. Sure, there’s a typo now and then, but usually a bug implies somebody didn’t fully understand what they were doing and implemented the wrong thing. When you fix a bug, don’t look for the symptom: look for the underlying cause, the misunderstanding, if you will, because that leads to a better understanding of the program’s architecture as well as revealing fundamental underlying flaws in the design itself.

4.9 トランザクションマネージャー: Txn

Our last module is the transaction manager, which ties together the individual components to provide the transactional ACID properties of atomicity, consistency, isolation, and durability. The transaction manager is responsible for beginning and completing (either committing or aborting)

transactions, coordinating the log and buffer managers to take transaction checkpoints, and orchestrating recovery. We'll visit each of these areas in order.

Jim Gray invented the ACID acronym to describe the key properties that transactions provide [Gra81]. Atomicity means that all the operations performed within a transaction appear in the database in a single unit—they either are all present in the database or all absent. Consistency means that a transaction moves the database from one logically consistent state to another. For example, if the application specifies that all employees must be assigned to a department that is described in the database, then the consistency property enforces that (with properly written transactions). Isolation means that from the perspective of a transaction, it appears that the transaction is running sequentially without any concurrent transactions running. Finally, durability means that once a transaction is committed, it stays committed—no failure can cause a committed transaction to disappear.

The transaction subsystem enforces the ACID properties, with the assistance of the other subsystems. It uses traditional transaction begin, commit, and abort operations to delimit the beginning and ending points of a transaction. It also provides a prepare call, which facilitates two phase commit, a technique for providing transactional properties across distributed transactions, which are not discussed in this chapter. Transaction begin allocates a new transaction identifier and returns a transaction handle, DB_TXN, to the application. Transaction commit writes a commit log record and then forces the log to disk (unless the application indicates that it is willing to forego durability in exchange for faster commit processing), ensuring that even in the presence of failure, the transaction will be committed. Transaction abort reads backwards through the log records belonging to the designated transaction, undoing each operation that the transaction had done, returning the database to its pre-transaction state.

チェックポイントの処理

The transaction manager is also responsible for taking checkpoints. There are a number of different techniques in the literature for taking checkpoints [HR83]. Berkeley DB uses a variant of fuzzy checkpointing. Fundamentally, checkpointing involves writing buffers from Mpool to disk. This is a potentially expensive operation, and it's important that the system continues to process new transactions while doing so, to avoid long service disruptions. At the beginning of a checkpoint, Berkeley DB examines the set of currently active transactions to find the lowest LSN written by any of them. This LSN becomes the checkpoint LSN. The transaction manager then asks Mpool to flush its dirty buffers to disk; writing those buffers might trigger log flush operations. After all the buffers are safely on disk, the transaction manager then writes a checkpoint record containing the checkpoint LSN. This record states that all the operations described by log records before the checkpoint LSN are now safely on disk. Therefore, log records prior to the checkpoint LSN are no longer necessary for recovery. This has two implications: First, the system can reclaim any log files

prior to the checkpoint LSN. Second, recovery need only process records after the checkpoint LSN, because the updates described by records prior to the checkpoint LSN are reflected in the on-disk state.

Note that there may be many log records between the checkpoint LSN and the actual checkpoint record. That's fine, since those records describe operations that logically happened after the checkpoint and that may need to be recovered if the system fails.

リカバリー

The last piece of the transactional puzzle is recovery. The goal of recovery is to move the on-disk database from a potentially inconsistent state to a consistent state. Berkeley DB uses a fairly conventional two-pass scheme that corresponds loosely to “relative to the last checkpoint LSN, undo any transactions that never committed and redo any transactions that did commit.” The details are a bit more involved.

Berkeley DB needs to reconstruct its mapping between log file ids and actual databases so that it can redo and undo operations on the databases. The log contains a full history of DBREG_REGISTER log records, but since databases stay open for a long time and we do not want to require that log files persist for the entire duration a database is open, we'd like a more efficient way to access this mapping. Prior to writing a checkpoint record, the transaction manager writes a collection of DBREG_REGISTER records describing the current mapping from log file ids to databases. During recovery, Berkeley DB uses these log records to reconstruct the file mapping.

When recovery begins, the transaction manager probes the log manager's cached_ckpt_lsn value to determine the location of the last checkpoint record in the log. This record contains the checkpoint LSN. Berkeley DB needs to recover from that checkpoint LSN, but in order to do so, it needs to reconstruct the log file id mapping that existed at the checkpoint LSN; this information appears in the checkpoint *prior* to the checkpoint LSN. Therefore, Berkeley DB must look for the last checkpoint record that occurs before the checkpoint LSN. Checkpoint records contain, not only the checkpoint LSN, but the LSN of the previous checkpoint to facilitate this process. Recovery begins at the most recent checkpoint and using the prev_lsn field in each checkpoint record, traverses checkpoint records backwards through the log until it finds a checkpoint record appearing before the checkpoint LSN. Algorithmically:

```
ckp_record = read (cached_ckpt_lsn)
ckp_lsn = ckp_record.checkpoint_lsn
cur_lsn = ckp_record.my_lsn
while (cur_lsn > ckp_lsn) {
    ckp_record = read (ckp_record.prev_ckp)
    cur_lsn = ckp_record.my_lsn
}
```

Starting with the checkpoint selected by the previous algorithm, recovery reads sequentially until the end of the log to reconstruct the log file id mappings. When it reaches the end of the log, its

mappings should correspond exactly to the mappings that existed when the system stopped. Also during this pass, recovery keeps track of any transaction commit records encountered, recording their transaction identifiers. Any transaction for which log records appear, but whose transaction identifier does not appear in a transaction commit record, was either aborted or never completed and should be treated as aborted. When recovery reaches the end of the log, it reverses direction and begins reading backwards through the log. For each transactional log record encountered, it extracts the transaction identifier and consults the list of transactions that have committed, to determine if this record should be undone. If it finds that the transaction identifier does not belong to a committed transaction, it extracts the record type and calls a recovery routine for that log record, directing it to undo the operation described. If the record belongs to a committed transaction, recovery ignores it on the backwards pass. This backward pass continues all the way back to the checkpoint LSN¹. Finally, recovery reads the log one last time in the forward direction, this time redoing any log records belonging to committed transactions. When this final pass completes, recovery takes a checkpoint. At this point, the database is fully consistent and ready to begin running the application.

Thus, recovery can be summarized as:

1. Find the checkpoint prior to the checkpoint LSN in the most recent checkpoint
2. Read forward to restore log file id mappings and construct a list of committed transactions
3. Read backward to the checkpoint LSN, undoing all operations for uncommitted transactions
4. Read forward, redoing all operations for committed transactions
5. Checkpoint

In theory, the final checkpoint is unnecessary. In practice, it bounds the time for future recoveries and leaves the database in a consistent state.

設計講座 14

Database recovery is a complex topic, difficult to write and harder to debug because recovery simply shouldn't happen all that often. In his Turing Award Lecture, Edsger Dijkstra argued that programming was inherently difficult and the beginning of wisdom is to admit we are unequal to the task. Our goal as architects and programmers is to use the tools at our disposal: design, problem decomposition, review, testing, naming and style conventions, and other good habits, to constrain programming problems to problems we *can* solve.

¹Note that we only need to go backwards to the checkpoint LSN, not the checkpoint record preceding it.

4.10 まとめ

Berkeley DB is now over twenty years old. It was arguably the first general-purpose transactional key/value store and is the grandfather of the NoSQL movement. Berkeley DB continues as the underlying storage system for hundreds of commercial products and thousands of Open Source applications (including SQL, XML and NoSQL engines) and has millions of deployments across the globe. The lessons we've learned over the course of its development and maintenance are encapsulated in the code and summarized in the design tips outlined above. We offer them in the hope that other software designers and architects will find them useful.

CMake

Bill Hoffman and Kenneth Martin

In 1999 the National Library of Medicine engaged a small company called Kitware to develop a better way to configure, build, and deploy complex software across many different platforms. This work was part of the Insight Segmentation and Registration Toolkit, or ITK¹. Kitware, the engineering lead on the project, was tasked with developing a build system that the ITK researchers and developers could use. The system had to be easy to use, and allow for the most productive use of the researchers' programming time. Out of this directive emerged CMake as a replacement for the aging autoconf/libtool approach to building software. It was designed to address the weaknesses of existing tools while maintaining their strengths.

In addition to a build system, over the years CMake has evolved into a family of development tools: CMake, CTest, CPack, and CDash. CMake is the build tool responsible for building software. CTest is a test driver tool, used to run regression tests. CPack is a packaging tool used to create platform-specific installers for software built with CMake. CDash is a web application for displaying testing results and performing continuous integration testing.

5.1 CMake History and Requirements

When CMake was being developed, the normal practice for a project was to have a configure script and Makefiles for Unix platforms, and Visual Studio project files for Windows. This duality of build systems made cross-platform development very tedious for many projects: the simple act of adding a new source file to a project was painful. The obvious goal for developers was to have a single unified build system. The developers of CMake had experience with two approaches of solving the unified build system problem.

One approach was the VTK build system of 1999. That system consisted of a configure script for Unix and an executable called pcmaker for Windows. pcmaker was a C program that read in Unix

¹<http://www.itk.org/>

Makefiles and created NMake files for Windows. The binary executable for pcmaker was checked into the VTK CVS system repository. Several common cases, like adding a new library, required changing that source and checking in a new binary. Although this was a unified system in some sense, it had many shortcomings.

The other approach the developers had experience with was a gmake based build system for TargetJr. TargetJr was a C++ computer vision environment originally developed on Sun workstations. Originally TargetJr used the imake system to create Makefiles. However, at some point, when a Windows port was needed, the gmake system was created. Both Unix compilers and Windows compilers could be used with this gmake-based system. The system required several environment variables to be set prior to running gmake. Failure to have the correct environment caused the system to fail in ways that were difficult to debug, especially for end users.

Both of these systems suffered from a serious flaw: they forced Windows developers to use the command line. Experienced Windows developers prefer to use integrated development environments (IDEs). This would encourage Windows developers to create IDE files by hand and contribute them to the project, creating the dual build system again. In addition to the lack of IDE support, both of the systems described above made it extremely difficult to combine software projects. For example, VTK (第 24 章) had very few modules for reading images mostly because the build system made it very difficult to use libraries like libtiff and libjpeg.

It was decided that a new build system would be developed for ITK and C++ in general. The basic constraints of the new build system would be as follows:

- Depend only on a C++ compiler being installed on the system.
- It must be able to generate Visual Studio IDE input files.
- It must be easy to create the basic build system targets, including static libraries, shared libraries, executables, and plugins.
- It must be able to run build time code generators.
- It must support separate build trees from the source tree.
- It must be able to perform system introspection, i.e., be able to determine automatically what the target system could and could not do.
- It must do dependency scanning of C/C++ header files automatically.
- All features would need to work consistently and equally well on all supported platforms.

In order to avoid depending on any additional libraries and parsers, CMake was designed with only one major dependency, the C++ compiler (which we can safely assume we have if we're building C++ code). At the time, building and installing scripting languages like Tcl was difficult on many popular UNIX and Windows systems. It can still be an issue today on modern supercomputers and secured computers with no Internet connection, so it can still be difficult to build third-party libraries. Since the build system is such a basic requirement for a package, it was decided that no additional dependencies would be introduced into CMake. This did limit CMake to creating its own

simple language, which is a choice that still causes some people to dislike CMake. However, at the time the most popular embedded language was Tcl. If CMake had been a Tcl-based build system, it is unlikely that it would have gained the popularity that it enjoys today.

The ability to generate IDE project files is a strong selling point for CMake, but it also limits CMake to providing only the features that the IDE can support natively. However, the benefits of providing native IDE build files outweigh the limitations. Although this decision made the development of CMake more difficult, it made the development of ITK and other projects using CMake much easier. Developers are happier and more productive when using the tools they are most familiar with. By allowing developers to use their preferred tools, projects can take best advantage of their most important resource: the developer.

All C/C++ programs require one or more of the following fundamental building blocks of software: executables, static libraries, shared libraries, and plugins. CMake had to provide the ability to create these products on all supported platforms. Although all platforms support the creation of those products, the compiler flags used to create them vary greatly from compiler to compiler and platform to platform. By hiding the complexity and platform differences behind a simple command in CMake, developers are able to create them on Windows, Unix and Mac. This ability allows developers to focus on the project rather than on the details of how to build a shared library.

Code generators provide added complexity to a build system. From the start, VTK provided a system that automatically wrapped the C++ code into Tcl, Python, and Java by parsing the C++ header files, and automatically generating a wrapping layer. This requires a build system that can build a C/C++ executable (the wrapper generator), then run that executable at build time to create more C/C++ source code (the wrappers for the particular modules). That generated source code must then be compiled into executables or shared libraries. All of this has to happen within the IDE environments and the generated Makefiles.

When developing flexible cross-platform C/C++ software, it is important to program to the features of the system, and not to the specific system. Autotools has a model for doing system introspection which involves compiling small snippets of code, inspecting and storing the results of that compile. Since CMake was meant to be cross-platform it adopted a similar system introspection technique. This allows developers to program to the canonical system instead of to specific systems. This is important to make future portability possible, as compilers and operating systems change over time. For example, code like this:

```
#ifdef linux
// do some linux stuff
#endif
```

Is more brittle than code like this:

```
#ifdef HAS_FEATURE
// do something with a feature
#endif
```

Another early CMake requirement also came from autotools: the ability to create build trees that are separate from the source tree. This allows for multiple build types to be performed on the same source tree. It also prevents the source tree from being cluttered with build files, which often confuses version control systems.

One of the most important features of a build system is the ability to manage dependencies. If a source file is changed, then all products using that source file must be rebuilt. For C/C++ code, the header files included by a .c or .cpp file must also be checked as part of the dependencies. Tracking down issues where only some of the code that should be compiled actually gets compiled as a result of incorrect dependency information can be time consuming.

All of the requirements and features of the new build system had to work equally well on all supported platforms. CMake needed to provide a simple API for developers to create complicated software systems without having to understand platform details. In effect, software using CMake is outsourcing the build complications to the CMake team. Once the vision for the build tool was created with the basic set of requirements, implementation needed to proceed in an agile way. ITK needed a build system almost from day one. The first versions of CMake did not meet all of the requirements set out in the vision, but they were able to build on Windows and Unix.

5.2 How CMake Is Implemented

As mentioned, CMake’s development languages are C and C++. To explain its internals this section will first describe the CMake process from a user’s point of view, then examine its structures.

The CMake Process

CMake has two main phases. The first is the “configure” step, in which CMake processes all the input given to it and creates an internal representation of the build to be performed. Then next phase is the “generate” step. In this phase the actual build files are created.

Environment Variables (or Not)

In many build systems in 1999, and even today, shell level environment variables are used during the build of a project. It is typical that a project has a PROJECT_ROOT environment variable that points to the location of the root of the source tree. Environment variables are also used to point to optional or external packages. The trouble with this approach is that for the build to work, all of these external variables need to be set each time a build is performed. To solve this problem CMake has a cache file that stores all of the variables required for a build in one place. These are not shell or environment variables, but CMake variables. The first time CMake is run for a particular build

tree, it creates a `CMakeCache.txt` file which stores all the persistent variables for that build. Since the file is part of the build tree, the variables will always be available to CMake during each run.

The Configure Step

During the configure step, CMake first reads the `CMakeCache.txt` if it exists from a prior run. It then reads `CMakeLists.txt`, found in the root of the source tree given to CMake. During the configure step, the `CMakeLists.txt` files are parsed by the CMake language parser. Each of the CMake commands found in the file is executed by a command pattern object. Additional `CMakeLists.txt` files can be parsed during this step by the `include` and `add_subdirectory` CMake commands. CMake has a C++ object for each of the commands that can be used in the CMake language. Some examples of commands are `add_library`, `if`, `add_executable`, `add_subdirectory`, and `include`. In effect, the entire language of CMake is implemented as calls to commands. The parser simply converts the CMake input files into command calls and lists of strings that are arguments to commands.

The configure step essentially “runs” the user-provided CMake code. After all of the code is executed, and all cache variable values have been computed, CMake has an in-memory representation of the project to be built. This will include all of the libraries, executables, custom commands, and all other information required to create the final build files for the selected generator. At this point, the `CMakeCache.txt` file is saved to disk for use in future runs of CMake.

The in-memory representation of the project is a collection of targets, which are simply things that may be built, such as libraries and executables. CMake also supports custom targets: users can define their inputs and outputs, and provide custom executables or scripts to be run at build time. CMake stores each target in a `cmTarget` object. These objects are stored in turn in the `cmMakefile` object, which is basically a storage place for all of the targets found in a given directory of the source tree. The end result is a tree of `cmMakefile` objects containing maps of `cmTarget` objects.

The Generate Step

Once the configure step has been completed, the generate step can take place. The generate step is when CMake creates the build files for the target build tool selected by the user. At this point the internal representation of targets (libraries, executables, custom targets) is converted to either an input to an IDE build tool like Visual Studio, or a set of Makefiles to be executed by `make`. CMake’s internal representation after the configure step is as generic as possible so that as much code and data structures as possible can be shared between different built tools.

An overview of the process can be seen in [図 5.1](#).

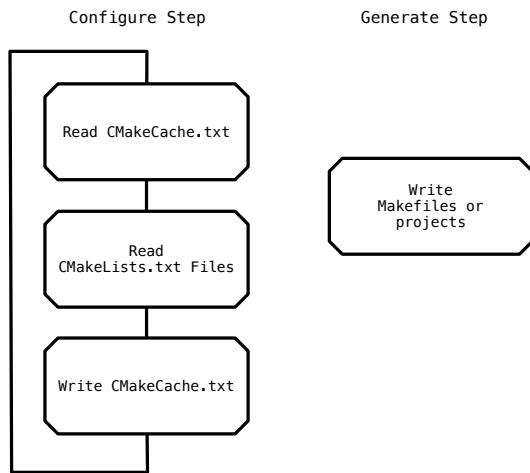


図 5.1: Overview of the CMake Process

CMake: The Code

CMake Objects

CMake is an object-oriented system using inheritance, design patterns and encapsulation. The major C++ objects and their relationships can be seen in 図 5.2.

The results of parsing each `CMakeLists.txt` file are stored in the `cmMakefile` object. In addition to storing the information about a directory, the `cmMakefile` object controls the parsing of the `CMakeLists.txt` file. The parsing function calls an object that uses a lex/yacc-based parser for the CMake language. Since the CMake language syntax changes very infrequently, and lex and yacc are not always available on systems where CMake is being built, the lex and yacc output files are processed and stored in the `Source` directory under version control with all of the other handwritten files.

Another important class in CMake is `cmCommand`. This is the base class for the implementation of all commands in the CMake language. Each subclass not only provides the implementation for the command, but also its documentation. As an example, see the documentation methods on the `cmUnsetCommand` class:

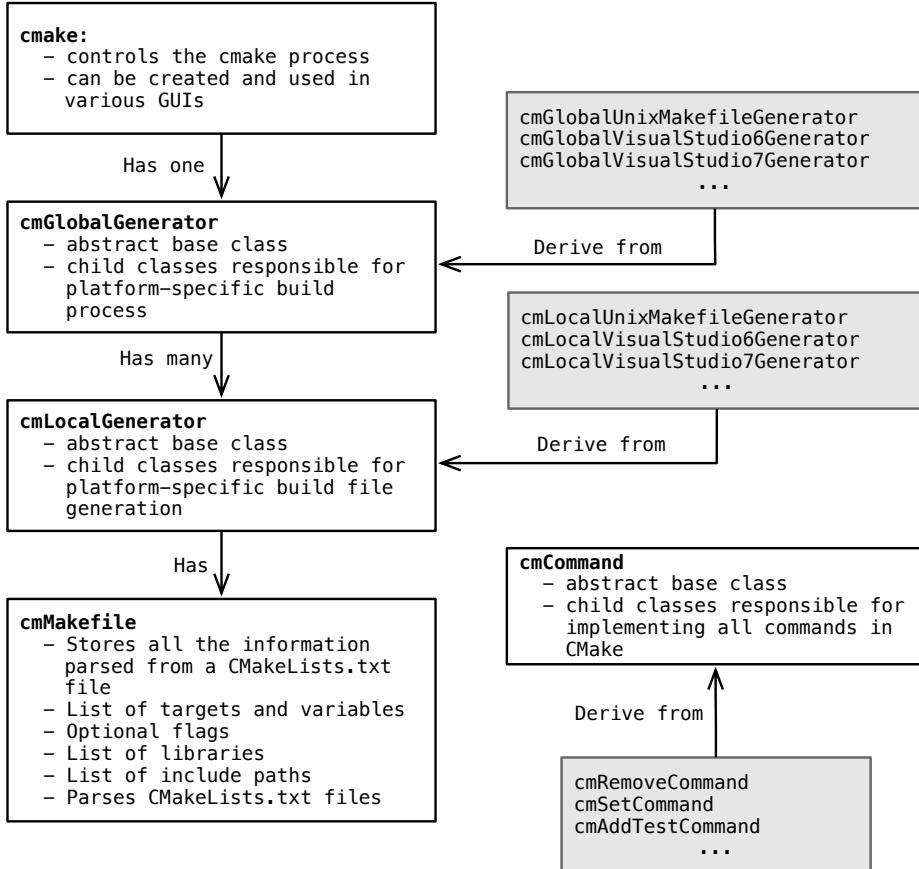


図 5.2: CMake Objects

```

virtual const char* GetTerseDocumentation()
{
    return "Unset a variable, cache variable, or environment variable.";
}

/**
 * More documentation.
 */

virtual const char* GetFullDocumentation()
{
    return
        " unset(<variable> [CACHE])\n"
        "Removes the specified variable causing it to become undefined. "
        "If CACHE is present then the variable is removed from the cache "
        "instead of the current scope.\n"
        "<variable> can be an environment variable such as:\n"

```

```
    "  unset(ENV{LD_LIBRARY_PATH})\n"
    "in which case the variable will be removed from the current "
    "environment.";
}
```

Dependency Analysis

CMake has powerful built-in dependency analysis capabilities for individual Fortran, C and C++ source code files. Since Integrated Development Environments (IDEs) support and maintain file dependency information, CMake skips this step for those build systems. For IDE builds, CMake creates a native IDE input file, and lets the IDE handle the file level dependency information. The target level dependency information is translated to the IDE's format for specifying dependency information.

With Makefile-based builds, native make programs do not know how to automatically compute and keep dependency information up-to-date. For these builds, CMake automatically computes dependency information for C, C++ and Fortran files. Both the generation and maintenance of these dependencies are automatically done by CMake. Once a project is initially configured by CMake, users only need to run `make` and CMake does the rest of the work.

Although users do not need to know how CMake does this work, it may be useful to look at the dependency information files for a project. This information for each target is stored in four files called `depend.make`, `flags.make`, `build.make`, and `DependInfo.cmake`. `depend.make` stores the dependency information for all the object files in the directory. `flags.make` contains the compile flags used for the source files of this target. If they change then the files will be recompiled. `DependInfo.cmake` is used to keep the dependency information up-to-date and contains information about what files are part of the project and what languages they are in. Finally, the rules for building the dependencies are stored in `build.make`. If a dependency for a target is out of date then the depend information for that target will be recomputed, keeping the dependency information current. This is done because a change to a .h file could add a new dependency.

CTest and CPack

Along the way, CMake grew from a build system into a family of tools for building, testing, and packaging software. In addition to command line `cmake`, and the CMake GUI programs, CMake ships with a testing tool CTest, and a packaging tool CPack. CTest and CPack shared the same code base as CMake, but are separate tools not required for a basic build.

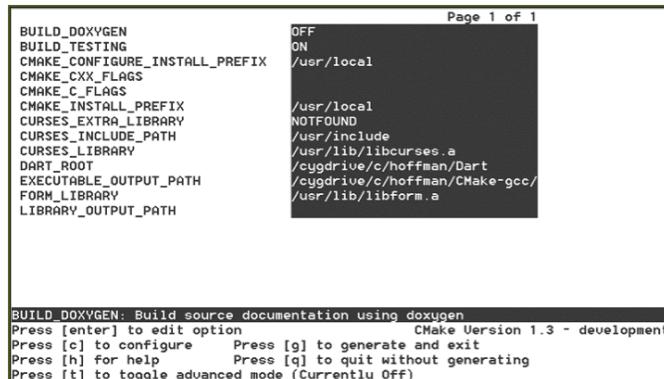
The `ctest` executable is used to run regression tests. A project can easily create tests for CTest to run with the `add_test` command. The tests can be run with CTest, which can also be used to send testing results to the CDash application for viewing on the web. CTest and CDash together are similar to the Hudson testing tool. They do differ in one major area: CTest is designed to allow a

much more distributed testing environment. Clients can be setup to pull source from version control system, run tests, and send the results to CDash. With Hudson, client machines must give Hudson ssh access to the machine so tests can be run.

The cpack executable is used to create installers for projects. CPack works much like the build part of CMake: it interfaces with other packaging tools. For example, on Windows the NSIS packaging tool is used to create executable installers from a project. CPack runs the install rules of a project to create the install tree, which is then given to a an installer program like NSIS. CPack also supports creating RPM, Debian .deb files, .tar, .tar.gz and self-extracting tar files.

Graphical Interfaces

The first place many users first see CMake is one of CMake's user interface programs. CMake has two main user interface programs: a windowed Qt-based application, and a command line curses graphics-based application. These GUIs are graphical editors for the CMakeCache.txt file. They are relatively simple interfaces with two buttons, configure and generate, used to trigger the main phases of the CMake process. The curses-based GUI is available on Unix TTY-type platforms and Cygwin. The Qt GUI is available on all platforms. The GUIs can be seen in [図 5.3](#) and [図 5.4](#).



The screenshot shows the CMake Command Line Interface. It displays a list of cache variables and their current values. The variables listed include:

Variable	Value
BUILD_DOXYGEN	OFF
BUILD_TESTING	ON
CMAKE_CONFIGURE_INSTALL_PREFIX	/usr/local
CMAKE_CXX_FLAGS	
CMAKE_C_FLAGS	
CMAKE_INSTALL_PREFIX	/usr/local
CURSES_EXTRA_LIBRARY	NOTFOUND
CURSES_INCLUDE_PATH	/usr/include
CURSES_LIBRARY	/usr/lib/libcurses.a
DART_ROOT	/cygdrive/c/hoffman/Dart
EXECUTABLE_OUTPUT_PATH	/cygdrive/c/hoffman/CMake-gcc/
FORM_LIBRARY	
LIBRARY_OUTPUT_PATH	/usr/lib/libform.a

At the bottom of the screen, there is a status bar with the following text:

BUILD_DOXYGEN: Build source documentation using doxygen
Press [enter] to edit option CMake Version 1.3 - development
Press [c] to configure Press [g] to generate and exit
Press [h] for help Press [q] to quit without generating
Press [t] to toggle advanced mode (Currently Off)

図 5.3: Command Line Interface

Both GUIs have cache variable names on the left, and values on the right. The values on the right can be changed by the user to values that are appropriate for the build. There are two types of variables, normal and advanced. By default the normal variables are shown to the user. A project can determine which variables are advanced inside the CMakeLists.txt files for the project. This allows users to be presented with as few choices as necessary for a build.

Since cache values can be modified as the commands are executed, the process of converging on a final build can be iterative. For example, turning on an option may reveal additional options. For this reason, the GUI disables the “generate” button until the user has had a chance to see all options

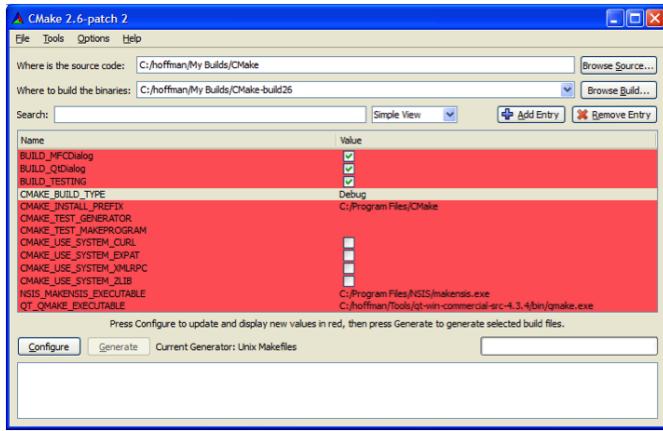


図 5.4: Graphics-based Interface

at least once. Each time the configure button is pressed, new cache variables that have not yet been presented to the user are displayed in red. Once there are no new cache variables created during a configure run, the generate button is enabled.

Testing CMake

Any new CMake developer is first introduced to the testing process used in CMake development. The process makes use of the CMake family of tools (CMake, CTest, CPack, and CDash). As the code is developed and checked into the version control system, continuous integration testing machines automatically build and test the new CMake code using CTest. The results are sent to a CDash server which notifies developers via email if there are any build errors, compiler warnings, or test failures.

The process is a classic continuous integration testing system. As new code is checked into the CMake repository, it is automatically tested on the platforms supported by CMake. Given the large number of compilers and platforms that CMake supports, this type of testing system is essential to the development of a stable build system.

For example, if a new developer wants to add support for a new platform, the first question he or she is asked is whether they can provide a nightly dashboard client for that system. Without constant testing, it is inevitable that new systems will stop working after some period of time.

5.3 Lessons Learned

CMake was successfully building ITK from day one, and that was the most important part of the project. If we could redo the development of CMake, not much would change. However, there are

always things that could have been done better.

Backwards Compatibility

Maintaining backwards compatibility is important to the CMake development team. The main goal of the project is to make building software easier. When a project or developer chooses CMake for a build tool, it is important to honor that choice and try very hard to not break that build with future releases of CMake. CMake 2.6 implemented a policy system where changes to CMake that would break existing behavior will warn but still perform the old behavior. Each `CMakeLists.txt` file is required to specify which version of CMake they are expecting to use. Newer versions of CMake might warn, but will still build the project as older versions did.

Language, Language, Language

The CMake language is meant to be very simple. However, it is one of the major obstacles to adoption when a new project is considering CMake. Given its organic growth, the CMake language does have a few quirks. The first parser for the language was not even lex/yacc based but rather just a simple string parser. Given the chance to do the language over, we would have spent some time looking for a nice embedded language that already existed. Lua is the best fit that might have worked. It is very small and clean. Even if an external language like Lua was not used, I would have given more consideration to the existing language from the start.

Plugins Did Not Work

To provide the ability for extension of the CMake language by projects, CMake has a plugin class. This allows a project to create new CMake commands in C. This sounded like a good idea at the time, and the interface was defined for C so that different compilers could be used. However, with the advent of multiple API systems like 32/64 bit Windows and Linux, the compatibility of plugins became hard to maintain. While extending CMake with the CMake language is not as powerful, it avoids CMake crashing or not being able to build a project because a plugin failed to build or load.

Reduce Exposed APIs

A big lesson learned during the development of the CMake project is that you don't have to maintain backward compatibility with something that users don't have access to. Several times during the development of CMake, users and customers requested that CMake be made into a library so that other languages could be bound to the CMake functionality. Not only would this have fractured

the CMake user community with many different ways to use CMake, but it would have been a huge maintenance cost for the CMake project.

継続的インテグレーション

C. Titus Brown and Rosangela Canino-Koning

継続的インテグレーション (Continuous Integration: CI) システムとは、ソフトウェアのビルドやテストを自動的かつ定期的に行うシステムのことである。CI システムを使う最大のメリットは、ビルドやテストの実行間隔が長くなるのを避けられるということだ。それ以外にも、その他の退屈な作業を単純化して自動化させることもできる。たとえば、クロスプラットフォームでのテスト実行、遅かったりデータを扱ったり設定の難しかったりするテストの定期実行、レガシーな環境での適切なパフォーマンスの確保、ごくまれに失敗するテストの検出、そしてリリースする製品の定期的な作成などといった作業がそれにあたる。また、ビルドやテストの自動化は継続的インテグレーションのために必須となるので、CI は**継続的デプロイ用フレームワーク**に向けた第一歩にもなる。これは、ソフトウェアを更新したら、テストをしてすぐに稼働中のシステムへ展開できるようにする仕組みである。

昨今アジャイルソフトウェア方法論が広まってきたこともあり、継続的インテグレーションはタイムリーな話題である。オープンソースの CI ツールもここ数年急増し、さまざまな言語向けにさまざまな言語で書かれている。そして、さまざまなアーキテクチャモデルに対応した幅広い機能が実装されている。本章では継続的インテグレーションシステムが実装する一般的な機能群について説明し、アーキテクチャに関する選択肢について議論する。そして、選んだアーキテクチャごとにどの機能が実装しやすくてどの機能が実装しづらいのかを検討する。

これ以降では、CI システムを設計する際に選択可能なアーキテクチャのよい例となるシステム群について簡単に説明する。最初に取り上げる Buildbot はマスター/スレーブ型のシステムだ。それに続く CDash はレポートサーバー型、Jenkins はハイブリッド型、そして最後の Pony-Build は Python ベースの分散型レポートサーバーで、これを使ってさらに議論を深めていく。

6.1 概観

継続的インテグレーションシステムのアーキテクチャの世界は二大勢力に支配されているようだ。一方はマスター/スレーブ型のアーキテクチャで、中央のサーバーがリモートのビルドを指揮して制御する。その対極にあるのがレポーティングアーキテクチャで、各クライアントからのレポートを中央のサーバーが集約する。我々の知る範囲では、すべての継続的インテグレーションシステムはこの二種類のアーキテクチャの機能を組み合わせて使っている。

中央集権型のアーキテクチャの例として取り上げる Buildbot は、ふたつのパーツで構成されている。中央サーバーである *buildmaster* がそこに接続しているクライアントのビルドスケジュールを管理し、クライアント側の *buildslaves* が実際のビルドを行う。*buildmaster* はクライアントからの接続先となり、各クライアントがどのコマンドをどの順で実行するのかという設定情報を提供する。*buildslave* は *buildmaster* に接続詞、詳細な指示を受け取る。*buildslave* の設定に含まれるのは、ソフトウェアをインストールすることやマスターサーバーの識別、そしてマスターサーバーに接続するための認証情報などである。ビルド予定をたてるのは *buildmaster* で、その出力は *buildslave* から *buildmaster* に流される。結果はマスターサーバー上に保持され、ウェブ経由で見たり別のレポートシステムや通知システムで見たりすることができる。

アーキテクチャ的にその対極にあるのが CDash で、これは Kitware, Inc. の Visualization Toolkit (VTK)/Insight Toolkit (ITK) プロジェクトで使われている。CDash は本質的にレポートティングサーバーで、CMake および CTest を実行するクライアントコンピューターから受け取った情報を蓄積して表示するように作られている。CDash では、クライアント側がビルドとテストスイートを起動し、ビルドとテストの結果を記録し、それから CDash サーバーに接続して情報をレポートティングサーバーに預ける。

最後に、三番目の例として取り上げる Jenkins (かつては Hudson と呼ばれていたが 2011 年に名前が変わった) は、その両方の操作モードを提供している。Jenkins の場合、ビルドを個別に実行して結果をマスターサーバーに送ることもできるし、ノードをすべて Jenkins マスターサーバーの支配下においてビルドの予定やその実行をマスターサーバーから指示することもできる。

中央集権型モデルと分散型モデルの両方に共通する機能もあり、Jenkins を見てもわかるとおり、両方のモデルをひとつの実装に共存させることもできる。しかし Buildbot と CDash はお互い全く正反対の存在である。ソフトウェアをビルドしてその結果を報告するという点は共通しているが、それ以外の面では全く異なるアーキテクチャを採用している。なぜだろう？

アーキテクチャの選択によって、特定の機能の実装しやすさ（しにくさ）はどの程度の影響を受けるのだろう？中央集権型を採用することで必然的に出てくる機能などがあるのだろうか？既存の実装の拡張しやすさについてはどうだろう—レポートティングの仕組みに手軽に手を入れたり、多数のパッケージを扱うために規模を拡大したり、あるいはビルドやテストをクラウド環境で実行したりといったことはできるのだろうか？

継続的インテグレーションソフトウェアの役割は?

継続的インテグレーションシステムの中核となる機能は単純だ。ソフトウェアをビルドしてテストを実行し、その結果を報告するだけである。ビルドやテストそして結果報告はスクリプトで行える。これは、スケジュールを組み込んだタスクや cron ジョブとして実行する。スクリプトの仕事は、ソースコードの新たなコピーを VCS から取得してビルドし、そしてテストを実行することだ。出力はログファイルに書き込むことになるだろう。ファイルを所定の場所に保存し、ビルドが失敗したときにはメールを送信することになる。この機能を実装するのは簡単だ。UNIXなら、大半の Python パッケージについてたった 7 行のスクリプトでこの機能を実現できる。

```
cd /tmp && \
svn checkout http://some.project.url && \
cd project_directory && \
python setup.py build && \
python setup.py test || \
echo build failed | sendmail notification@project.domain
cd /tmp && rm -fr project_directory
```

図 6.1において影付きでない長方形は、システム内にある個別のサブシステムや機能を表す。矢印は、コンポーネント間の情報の流れを意味する。雲で囲まれている部分は、おそらくリモートで実行されるであろうビルドプロセスを表す。影付きの長方形は、サブシステム間のつながりを表す。たとえば、ビルドの監視にはビルドプロセス自体の監視とシステムの健康状態(CPU の負荷、入出力の負荷、メモリの使用量など)の監視が含まれる。

しかし、単純そうに見えるのは見かけだけである。実際の CI システムは、通常はこれ以上のことを行っている。リモートのビルドプロセスを立ち上げてその結果を受け取ったりするだけでなく、継続的インテグレーションソフトウェアはこのような追加機能に対応していることもある。

チェックアウトと更新: 大規模なプロジェクトでは、ソースコードすべてを新たにチェックアウトするのは帶域的にも時間的にもコストがかかることになる。通常、CI システムは既存の作業コピーをその場で更新することになる。更新の際にやりとりするのは、前回の更新以降の差分だけである。通信量は節約できるが、その代わりにシステム側で作業コピーの状況をわかつていなければならない、更新方法も知る必要がある。つまり、通常は少なくとも VCS とは最小限の統合をすることになる。

ビルドレシピの抽象化: 設定やビルドそしてテストのレシピは、対象となるソフトウェア用に書かなければならない。もとになるコマンドは(Mac OS X と Windows と UNIX など)OS によって異なることが多い。ということは、それぞれの OS に特化したレシピを書く(これはバグのもとになるし、実際のビルド環境とはかけ離れてしまう可能性もある)か、さもなければ何らかの抽象化をしてレシピを CI 構成システムから提供できるようにしなければならない。

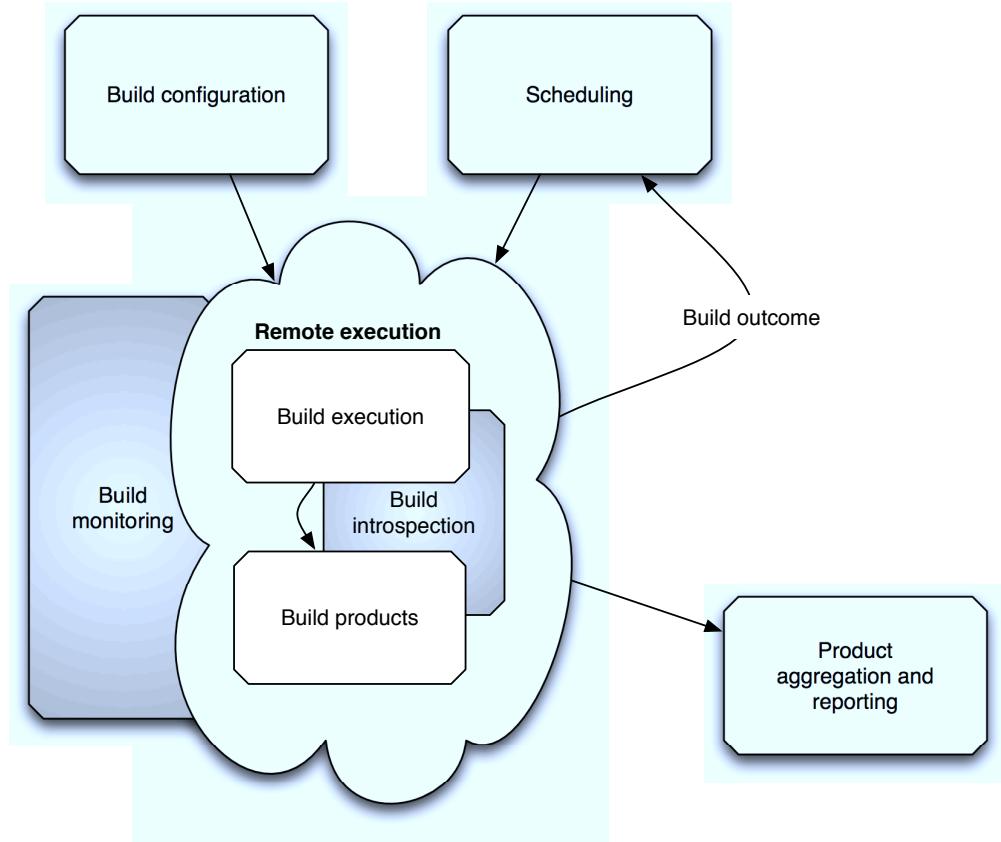


図 6.1: 繙続的インテグレーションシステムの内部構造

チェックアウト/ビルド/テストの状態の保存: チェックアウトの詳細(更新されたファイル、コードのバージョンなど)やビルドの情報(警告やエラー)、そしてテストの結果(コードカバレッジ、パフォーマンス、メモリの使用量)などを保存し、あとで解析に使えるようにしたいという要望もあるだろう。これらの結果を使えば、ビルドアーキテクチャをまたがる質問(最新のチェックインのせいで特定のアーキテクチャのパフォーマンスに問題が出ていないか?)や歴史を超えた質問(コードカバレッジは先月に比べて劇的に上昇したか?)にも答えられるようになる。ビルドレシピと同様、この種の調査の仕組みやデータ形式は、プラットフォームやビルドシステムに依存するものとなる。

パッケージのリリース: ビルドを実行するとバイナリパッケージあるいはその他外部に公開する必要のある何かができるがるかもしれない。たとえば、ビルドマシンに直接アクセスできない開発者が、最新のビルドを特定のアーキテクチャでテストしたくなることもあるだろう。これをサポートするためには、CIシステムがビルドの成果物を中央リポジトリに転送できるようにしておく必要がある。

複数のアーキテクチャでのビルド: 繼続的インテグレーションの目的のひとつは複数のアーキテクチャでビルドしてクロスプラットフォームな機能をテストすることなので、CI ソフトウェアは各ビルトマシンのアーキテクチャを追跡してビルトやビルト結果を各クライアントにリンクしなければならない。

リソース管理: ビルトの手順が特定のマシンのリソースの状況に依存する場合、CI システムはビルトを条件付きで実行させたくなることもある。たとえば、他のビルトやユーザーがいない間はビルトを待ったり、CPU やメモリの使用率が一定に達したらビルトを遅らせたりということがあり得る。

外部リソースとの協調: インテグレーションテストはローカルにないリソースに依存することがある。ステージング環境のデータベースやリモートウェブサービスなどだ。したがって、CI システムは複数のマシン間で協調し、これらのリソースへのアクセスを整理する必要がある。

進捗レポート: 時間がかかるビルト手順については、ビルト状況の定期的な報告も大切である。5 時間におよぶビルトやテストの中で主に知りたいのは最初の 30 分の結果だったとしよう。最後まで実行しないと何も結果を見られないのは時間の無駄になる。

CI システムで必要となりそうな全コンポーネントの概要は図 6.1 に示したとおりだ。CI ソフトウェアは通常、これらのコンポーネントの一部を実装している。

外部とのインタラクション

継続的インテグレーションシステムでは、他のシステムとのやりとりも必要となる。考えるやりとりには、次のような型がある。

ビルト通知: ビルトの結果は、一般的にクライアントとのやりとりを要するだろう。プル形式での取得(ウェブ、RSS、RPC など)あるいはプッシュによる通知(メール、Twitter、PubSubHubbub など)のいずれかとなる。すべてのビルト結果を通知することもあれば失敗したビルトだけを通知することもある。あるいは、所定の時間内に実行できなかつたビルトだけを通知することもある。

ビルト情報: ビルトの詳細やその成果物を取得しなければならないこともあるだろう。通常は、RPC を使うか一括ダウンロードの仕組みを用意する。たとえば、別の解析システムを使ってより詳細な(あるいはより的を絞った) 解析を行ったり、コードカバレッジやパフォーマンスの情報を表示させることができる。さらに、テスト結果のリポジトリを別に用意して、CI システムでの失敗したテストと成功したテストの記録を保存しておくこともあるかもしれない。

ビルト要求: ユーザーあるいはコードリポジトリからのビルト要求を受け、それに対応する必要があるかもしれない。大半の VCS にはコミット後に何らかの処理をフックする仕組みがあり、たとえばビルト処理を起動するような RPC コールを実行することができる。あるいは、ユーザーがウェブインターフェイスあるいは RPC を使って手動でビルト要求を出すかもしれない。

CI システムのリモート制御: より一般化して、実行環境全体の変更をある程度うまく作られた RPC インターフェイスで行いたいものだ。アドホックな拡張あるいは正式に決められたインターフェイスを使って、特定のプラットフォーム上でのビルトの実行や別のブランチにさまざまなパッチを適用したビルトの実行、あるいは条件付きでのビルトの実行などを実行できる必要がある。この機能があれば、より一般的なワークフローにも対応できるので便利だ。たとえば CI テストに完全にパスした変更だけをコミットできるようにしたり、パッチをさまざまなシステムでテストしてから最終的に取り込むようにしたりといったことができる。バグ追跡システムやパッチシステムその他外部のシステムにはさまざまなものがあるので、このロジックを CI システム自体に組み込んでしまうのは意味がない。

6.2 アーキテクチャ

Buildbot と CDash はまったく正反対のアーキテクチャを選択しており、一部重複する部分もあるが別々の機能群を実装している。個々の機能セットについて以下で吟味し、ある機能の実装しやすさ(しにくさ)がアーキテクチャの選択でどのように変わらるのかを確かめる。

実装モデル: Buildbot

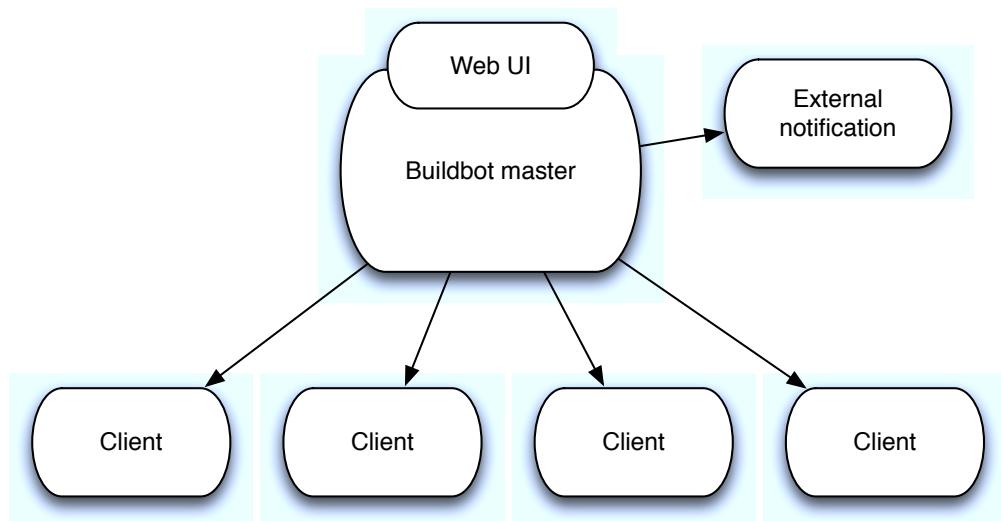


図 6.2: Buildbot のアーキテクチャ

Buildbot はマスター/スレーブ型のアーキテクチャで、一台の中央サーバーと複数台のビルト用スレーブで構成されている。リモートの実行は、完全にマスターサーバー側からの指示

でリアルタイムに行われる。各クライアント上で実行するコマンドをマスター側で設定し、直前のコマンドが終了するとそれを実行する。スケジューリングやビルド要求をサーバー側でとりまとめるだけではなく、実際の指示もマスターが行う。レシピの抽象化機能は組み込まれていない。ただし、基本的なバージョン管理システムとの統合（“我々のコードはこのリポジトリにある”）、そしてビルディレクトリ上で実行されるコマンドとビルディレクトリ内で実行されるコマンドの区別は例外である。OS 固有のコマンドは、通常は設定で直接指定する。

Buildbot は各スレーブとの接続を持続させ、ジョブの管理やスレーブ間での調整を行う。持続的接続を使ったリモートマシンの管理の実装は複雑なものとなり、長年バグの元となり続けている。堅牢なネットワーク接続を長期間保つのは単純なことではなく、ローカルの GUI と対話するアプリケーションのテストをネットワークごしに行うのは大変だ。OS のアラートウィンドウは特に扱いにくいものである。しかし、このように接続を持続させるおかげで、リソースの協調やスケジューリングを直感的に行うことができる。ジョブを実行させるときには、マスターがスレーブを完全に操れるからだ。

Buildbot のモデルの設計に組み込まれたような密な制御は、中央管理型でビルドを管理してリソース間で協調させることができる。Buildbot は buildmaster 上でのマスターのロックとスレーブのロックの両方を実装している。これらを使って、システムグローバルなリソースとマシンローカルなリソースを協調させたビルドが可能となる。この点で、Buildbot が特に適しているのは大規模なシステムのインテグレーションテスト（データベースやその他高価なリソースと組み合わせたテスト）であると言える。

しかし、中央集権型の設定は、分散型の利用モデルでは問題の原因となる。新しい buildslave を追加するには、マスターの設定で明示的に許可しなければならない。つまり、新しい buildslave を動的に中央サーバーにアタッチしてビルドサービスやビルド結果を送るようにするのは不可能だということだ。さらに、個々のスレーブは完全にマスター側からの指示で動いているので、悪意のある設定をされたり設定を間違えてしまったりするといった事故に対して脆弱になる。クライアント OS のセキュリティ制約の範囲内で、マスターはクライアントを文字通り完全に支配する。

Buildbot の機能面での制限のひとつは、ビルドの成果物を中央サーバーに返すシンプルな方法がないことだ。たとえば、コードカバレッジの統計情報やビルド後のバイナリはリモートの buildslave 上に残ったままとなる。中央の buildmaster 側に、それを集約したり配布したりする API は用意されていない。この機能が存在しない理由は定かではない。Buildbot とともに配布されているコマンド群の抽象化の制約のためかもしれない。もとのコマンド群が、スレーブ上でリモートコマンドの実行に焦点を合わせたものだからだ。あるいは、buildmaster と buildslave との間の接続はあくまでも制御システムとして使い、RPC の仕組みとしては使わないよう決めた結果かもしれない。

マスター/スレーブモデルを採用し、かつこのように制限のある通信チャンネルを用意した結果、buildslave 側からはシステムの利用状況を報告できなくなってしまっており、マスター側ではスレーブの負荷対策を組み込むことができない。

ビルド結果の外部 CPU 通知は完全に buildmaster が処理する。新たな通知サービスを追加するには buildmaster 自身の中で実装しなければならない。同様に、新たなビルド要求は直接 buildmaster にしなければならない。

実装モデル: CDash

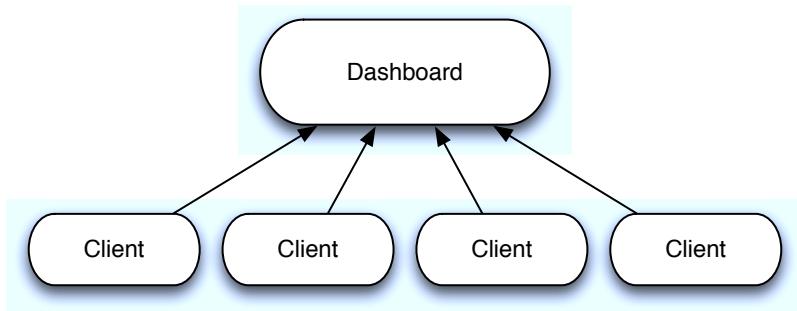


図 6.3: CDash のアーキテクチャ

Buildbot とは対照的に、CDash はレポートイングサーバーモデルを実装している。このモデルにおいて、CDash サーバーは中央リポジトリとしてふるまう。リモートで実行したビルドの情報やビルド・テストの失敗報告、コードカバレッジ解析、そしてメモリ使用状況などがここに集まる。ビルドのスケジュールをたてたり実行したりするのはリモートクライアント側で、ビルドレポートは XML 形式で送信する。ビルド結果の送信は“公式の”ビルドクライアントで行うこともできるし、コアデベロッパー以外の開発者やユーザーが公開したビルドプロセスを自分のマシンで実行することもできる。

このシンプルなモデルが可能が実現できた理由は、CDash とその他の Kitware ビルド基盤の要素(ビルド設定システムである CMake、テストランナーである CTest、そしてパッケージングシステムである CPack)が概念的に密結合していたことである。このソフトウェアが提供する仕組みを使うと、ビルドやテストそしてパッケージングのレシピを高度に抽象化して実装できる。その際に OS を意識する必要はない。

CDash のクライアント主導のプロセスは、クライアント側の CI プロセスを多くの面で単純化する。ビルドを実行するかどうかを決めるのはクライアント側なので、クライアント側の状況(時刻や負荷など)を考慮にいれてビルドを始めることができる。望みに応じてクライアントを増やしたり減らしたりするしてビルドを手伝ったり、ビルドを“クラウドで”行うともできる。ビルドの成果物を中央サーバーに送るのも、単純なアップロードで済む話だ。

しかし、このレポートイングモデルを採用した代償として、CDash には Buildbot が持つ多くの便利な機能が欠けている。中央管理型のリソース制御機能はないし、事前に登録済みでないクライアントを分散環境に投入することもできない。進捗レポート機能も実装されていない。実装するにはビルド状況のインクリメンタルな更新をサーバーが許可しないといけな

い。そしてもちろん、全体にビルド要求を出したり、チェックインに反応して匿名クライアントにビルドさせることもできない—クライアントはすべて信頼できないものとみなさなければならぬ。

最近、CDash に新機能が追加されてクラウドビルドシステム “@Home” が使えるようになつた。これは、クライアントが CDash サーバーに対してビルドサービスを提供する仕組みだ。クライアントがサーバーをポーリングしてビルドリクエストを受けとり、そのリクエストを処理して、結果をサーバーに返す。2010 年 10 月時点の実装では、ビルドのリクエストはサーバー側で手動で行う必要がある。そして、クライアントはサーバーに接続しないとサービスを提供できない。しかし、これを素直に拡張すればより汎用的なビルドモデルが作れる。つまり、サーバー側から自動的にビルドリクエストを送ったら、対応可能なクライアントが処理してくれるというモデルだ。“@Home” システムは、後で説明する Pony-Build システムと非常に似た概念である。

実装モデル: Jenkins

Jenkins は幅広く使われている継続的インテグレーションシステムで、Java で書かれている。2011 年初期までは Hudson という名前で知られていた。スタンドアロンの CI システムとしてローカルシステム上で動かすこともできるし、リモートビルドの調整役として使うこともできる。あるいは、リモートでのビルドの情報を受け取るだけの役割としても使うことができる。JUnit のユニットテストやコードカバレッジレポートで使われている標準の XML をうまく活用し、さまざまなテストツールからのレポートを統合する。Jenkins は元々 Sun が作り始めたものだが、さまざまな場所で使われており、しっかりとしたオープンソースコミュニティがついている。

Jenkins はハイブリッドモードで動作する。デフォルトはマスターサーバーでビルドを実行するが、さまざまなスタイルのリモートビルドも(サーバー側からでもクライアント側からでも)実行できる。Buildbot と同様、本来は中央サーバーが管理するように作られている。しかし、さまざまな分散ジョブ実行機構に対応するようになった。仮想マシンの管理機能も含む。

Jenkins は複数のリモートマシンを管理することができる。接続はマスター側から SSH で確立することもできるし、クライアント側から JNLP (Java Web Start) で確立することもできる。この接続は双方向で、オブジェクトやデータもシリアル化してやりとりすることができる。

Jenkins にはしっかりとプラグイン機構が組み込まれており、この接続の詳細を抽象化している。そのおかげで、多くのサードパーティのプラグインがバイナリビルドや結果のデータを扱えるようになっている。

中央サーバーが管理するジョブ用として、Jenkins には “locks” プラグインが用意されている。このプラグインはジョブを並列実行させないようにするものだが、2011 年 1 月の時点では未完成だ。

実装モデル: Pony-Build

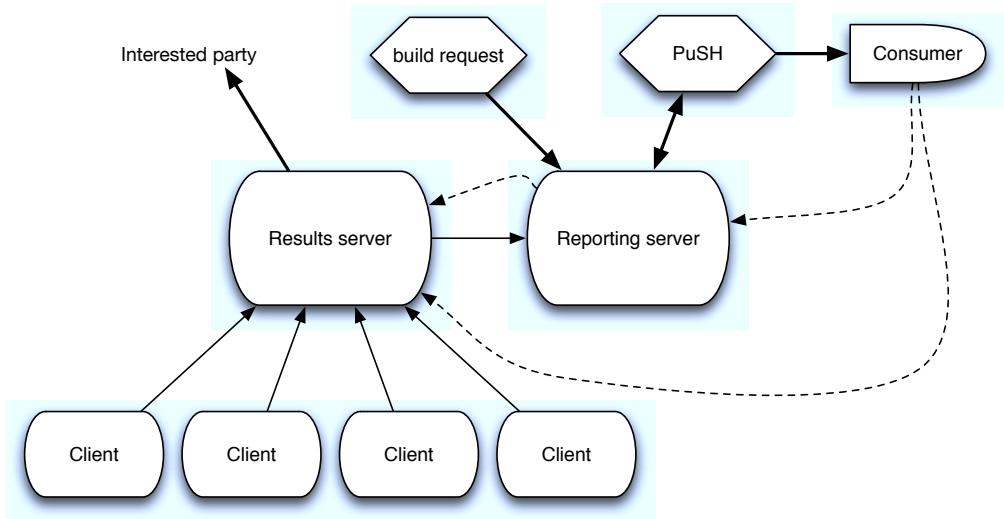


図 6.4: Pony-Build のアーキテクチャ

Pony-Build は、分散型の CI システムの概念実証モデルとして Python で作られた。図 6.4 に示す三つのコアコンポーネントで構成されている。結果サーバーが中央データベースとして働き、個々のクライアントから受け取ったビルド結果を保持する。クライアントはそれぞれ独立してすべての設定情報やビルドコンテキストを保持しており、軽量なクライアントライブラリで VCS のリポジトリにアクセスしたりビルドプロセスを管理したり、結果をサーバーに通信したりといったことができる。レポートサーバーは必須ではない。ここにはシンプルなウェブインターフェイスが組み込まれており、ビルド結果の報告や新しいビルドの要求を行う。我々の実装では、レポートサーバーと結果サーバーは単一のマルチスレッドプロセスで動作する。しかし API レベルでの結合は緩く、個別に動くようにも容易に変更できる。

基本モデルに加えてさまざまな WebHook や RPC 機構が用意されており、ビルドや変更通知そしてビルドに関する調査を支援する。たとえば、VCS のコードリポジトリの変更通知をビルドシステムと直接結び付けるのではなく、リモートからのビルドリクエストを直接レポートシステムに回し、レポートシステムがそれを結果サーバーに伝えるようにする。同様に、メールやインスタントメッセージングなどを使って新しいビルドを直接レポートサーバーにプッシュ通知するのではなく、通知の制御には PubSubHubbub (PuSH) を使っている。これにより、さまざまなアプリケーションが“興味のある”イベント（今のところは新しいビルドと失敗したビルドに限られる）の通知を PuSH WebHook で受け取れるようになっている。

このような疎結合のモデルを採用する利点は多い。

通信の容易性: 基本となるアーキテクチャ上のコンポーネントや WebHook のプロトコルは

極めて容易に実装でき、基本的なウェブプログラミングの知識さえあればよい。

変更の容易性: 新しい通知方式や新しいレポートサーバー用インターフェイスを実装するの
は極めて容易である。

多言語のサポート: さまざまなコンポーネントがお互い WebHook で呼び合っている。大半の
プログラミング言語は WebHook をサポートしているので、コンポーネントごとに別々
の言語で実装することができる。

テストのしやすさ: 各コンポーネントは完全に分離していてモックがあるので、システムの
テストを簡単に実行できる。

設定の容易性: クライアント側の要件は最小限で、Python そのもの以外に必要となるライブ
ラリはひとつだけである。

サーバーの負荷の最小化: 中央サーバーにはクライアントを制御する責任が事実上ないと言
えるので、隔離されたクライアントをサーバーと連携せずに並列に実行させることが
できる。報告時を除いて、サーバーに負荷をかけることはない。

VCS との統合: ビルドの設定は完全にクライアント側で行われるので、VCS に含めること
もできる。

結果へのアクセスの容易性: ビルド結果を取得したいアプリケーションを書くためのプログ
ラミング言語は、XML-RPC リクエストを扱えるものなら何でもよい。ビルドシステム
のユーザーには結果サーバーやレポートサーバーへのネットワークレベルでのアクセ
ス権限を与えることもできるし、レポートサーバーの独自インターフェイスを使って
もよい。ビルドクライアントに必要なアクセス権限は、結果サーバーへの結果の送信
権だけである。

残念ながら、深刻な**弱点**も多い。これは CDash のモデルと同様である。

ビルド要求を送るのが難しい: この弱点の原因は、ビルドクライアントが結果サーバーから
完全に独立しているということだ。クライアント側からサーバー側にビルド要求があ
るかどうかを確認することはできるかもしれない。しかしこれは負荷が高く、待ち時
間も長くなる。それ以外の手段をとるなら、指揮・制御のための接続を確立してサー
バー側からクライアント側にビルドリクエストを直接通知しなければならない。シス
テムはさらに複雑になり、分散型ビルドクライアントの利点を損ねてしまう。

リソースロックのサポートが貧弱: リソースロックを管理するために RPC の仕組みを提供
するのは簡単だが、クライアントポリシーを強制するのはずっと難しい。CDash のよ
うな CI システムはクライアント側を信頼することを前提としているが、クライアント
側はそのつもりがなくとも失敗してしまうかもしれない(ロックの解放を忘れるなど)。
堅牢な分散型ロックシステムを実装するのは困難であり、余計な複雑さを持ち込んで
しまう。たとえば、マスターリソースのロックを信頼できないクライアントに提供する
には、ロックをつかんで離さないクライアントに対するポリシーをマスターロックコ
ントローラー側で決めておく必要がある。これは、クライアントがクラッシュしたり、
デッドロックが発生したりした場合の対策となる。

リアルタイムの監視処理が貧弱: リアルタイムでビルドを監視したりビルドプロセス自体を制御したりする仕組みを実装するのは、常に接続が持続しているシステムでないと困難である。Buildbotがクライアント主導のモデルに比べてはるかに優れている点のひとつは、時間のかかるビルドの途中経過を調べやすいというところだ。これは、ビルドの結果がマスター側のインターフェイスにインクリメンタルに送られてくるからである。さらに、Buildbotは制御用の接続を保持しているので、もし時間のかかるビルドが途中で(設定ミスや間違ったチェックインなどで)失敗した場合は、そこでビルドを中止できる。そのような機能をPony-Build(結果サーバーからクライアント側への連絡ができない)に追加するには、クライアント側から定期的にポーリングさせるかクライアント側への接続を確立するかのどちらかの仕組みが必要となる。

Pony-Buildが提起するCIのその他ふたつの側面は、**レシピ**をいかに実装するかということ、そして**信頼**をどのように管理するのかということだ。これらはともに関連する問題である。というのも、レシピはビルドクライアント上の任意のコードを実行するからである。

ビルドレシピ

ビルドレシピは、使いやすいレベルの抽象化をしてくれるものだ。特に、クロスプラットフォームな言語で作られていたりマルチプラットフォームなビルドシステムを使っていたりする場合に役立つ。たとえばCDashは、非常に厳しいレシピに依存している。CDashを使うソフトウェアのほとんどすべてはCMakeやCTestそしてCPackでビルドされており、これらのツールはマルチプラットフォームな課題に対応するよう作られている。これは、継続的インテグレーションシステムの視点で考えると理想的な状況だ。というのも、CIシステムとしてはすべての課題をビルドツール群に丸投げするだけでよくなるからである。

しかし、これは必ずしもすべての言語やビルド環境で成り立つわけではない。Pythonの世界ではdistutilsやdistutils2を使ってソフトウェアのビルドやパッケージングを行うことが標準になりつつある。しかし、テストを探して実行したりその結果を収集したりする仕組みにはまだ標準が確立されていない。さらに、より複雑なPythonパッケージの多くは自前のビルドロジックをシステムに組み込んでおり、distutilsの拡張機構(任意のコードを実行できる)を通してそれを使っている。たいていのビルドツールで、同じような状況が見られる。標準的なコマンドが用意されていても、常に例外や拡張があるのだ。

ビルドやテストそしてパッケージングのレシピは、このようにいろいろ問題がある。なぜなら次のふたつの問題を解決しなければならないからだ。まず、プラットフォーム非依存な形式で指定しないといけない。単一のレシピを複数のシステム上のソフトウェアのビルドに使えるようにするためだ。そして、ビルドされるソフトウェアにあわせてカスタマイズ可能でなければならない。

信頼

さらに、第三の問題がある。CIシステムのレシピを幅広く使い始めると、信頼しなければならない外部のシステムが増えてしまう。ソフトウェア自身が信頼できる(CIクライアントは任意のコードを実行できるので)だけでは不十分で、さらにレシピも信頼できなければならぬ(レシピもまた任意のコードを実行できるので)。

こういった信頼に関する問題は、しっかりと制御された環境なら扱いやすい。たとえば、ビルドクライアントやCIシステムを内部プロセスの一環として扱うような企業がそれにあたる。しかし、その他の環境では、サードパーティがビルドサービスを提供することもある。たとえばオープンソースプロジェクト向けなどだ。理想的な解決策は、標準のビルドレシピをコミュニティレベルでソフトウェアに含められるようにすることだ。Pythonコミュニティは、これをdistutils2を使って行っている。もうひとつの解決策は、たとえばデジタル署名したレシピを使うことだ。そうすれば、信頼できる個人が書いたレシピを署名付きで配布でき、CIクライアントはそのレシピが信頼に値するかどうかを調べることができる。

どのモデルを選ぶか

経験上、疎結合なRPCやWebhookコールバックベースのモデルによる継続的インテグレーションは非常に実装しやすいものだ。複雑な結合にからむ密接な連携に関する要件を一切無視しさえすれば。基本的なリモートチェックアウトやビルドの実行には、ビルドをローカルで行うかリモートで行うかにかかわらず共通した設計上の制約がある。ビルドに関する情報(成功/失敗など)の収集は、基本的にはクライアント側からの要求に基づいて行う。アーキテクチャによる情報の追跡も結果による情報の追跡も、同じ基本要件を必要とする。したがって、基本的なCIシステムはレポーティングモデルを使えば極めて容易に実装できる。

疎結合のモデルも、非常に柔軟で拡張性のあるものであることがわかった。新たな結果報告の仕組みや通知の仕組み、あるいはビルドレシピを追加したりするのが容易になる。各コンポーネントが明確に分かれており、きちんと独立しているからである。分割されたコンポーネントはそれぞれやるべき作業が明確になっており、テストもしやすければ変更もしやすくなる。

CDashのような疎結合のモデルでリモートビルドをするときの唯一の困難は、ビルドの協調である。ビルドの開始や停止、進行中のビルドの状況報告、そして各クライアント間でのリソースロックの調整などが技術的に要求される。これらは、それ以外のモデルの実装ではあまり問われないものだ。

これらを踏まえて得られる結論は、疎結合のモデルのほうが全般的に“より優れている”ということだ。しかしそう言えるのは、ビルドの協調が不要な場合のみである。ビルドの協調が必要かどうかは、CIシステムを利用するプロジェクトによって決まる。

6.3 将来

Pony-Buildについて検討しているときに、将来の継続的インテグレーションシステムにあつたらしいなという機能をいくつか思いついた。

言語不可知なビルドレシピ群: 現状の継続的インテグレーションシステムはどれも、車輪の再発明をして独自のビルド設定言語を用意している。これは明らかにばかげた話だ。一般的に使われているビルドシステムは十種類にも満たないだろうし、テストランナーだっておそらく数十種類程度だろう。にもかかわらず、どのCIシステムも新しい独自の方法でビルドを指定したり実行するテストコマンドを指定したりする。実際のところ、よく似たCIシステムが氾濫している理由の一つがこれではないだろうか。それぞれのプログラミング言語やコミュニティが構成管理の仕組みを実装し、自分たちの使いなれたビルドシステムやテストシステムに合わせて調整し、同じような機能を持ったシステム上にそのレイヤーをかぶせているのだ。何かドメイン特化言語(DSL)を作り、高々数十種類しかないビルドツールやテストツールで使われるオプションを表せるようにしておけば、長い目で見たときにCIの世界を簡素化できるのではないだろうか。

ビルドやテストの報告用の共通フォーマット: ビルドシステムやテストシステムが、どんな情報をどんなフォーマットで提供すればいいのかに関する決まりが一切ない。もし何らかの共通フォーマットや標準規格ができれば、継続的インテグレーションシステムがビルドの詳細や概要をより簡単に提供できるようになるだろう。今のところそれに近い位置にあるのが、Perlコミュニティで使われているTAP(The Test Anywhere Protocol)やJavaコミュニティで使われているJUnitのXML出力形式である。これらは、実行したテストの数や成功と失敗の数、そしてファイル単位でのコードカバレッジの詳細といった情報を表すことができる。

レポーティングにおける粒度や内部調査機能の向上: あるいは、さまざまなビルドプラットフォームがよくできたフックシステムを提供し、その構成やコンパイルそしてテストのシステムをフックできるようになっていると便利だろう。このフックシステムが(共通のフォーマット以上の)APIを提供してくれれば、CIシステムはそれを使ってより詳細なビルド情報を取り出せる。

まとめ

これまでに解説してきた継続的インテグレーションシステムは、それぞれのアーキテクチャにうまくあてはまる機能を実装してきた。一方、ハイブリッドであるJenkinsは、最初はマスター/スレーブモデルだったが、より疎結合なレポーティングアーキテクチャの機能をそこに追加した。

選んだアーキテクチャによってその機能が決まる、という結論にしたいところだが、もちろんそれはナンセンスだ。それよりは、選んだアーキテクチャがその後の開発の方向性に影

響を与え、特定の機能群を実装する流れになるのではないだろうか。Pony-Build を作っていた我々も驚いた。最初に選んだ CDash 形式のレポートアーキテクチャが、後の設計や実装の決断に大きく影響したのだ。実装上の選択の中には、Pony-Build で中央管理型の構成やスケジューリングシステムを回避した例のように実際の使用例に基づいたものもある。我々が必要としていたのはリモートビルドクライアントを動的に追加できることで、これは Buildbot ではサポートしづらいものだった。それ以外に Pony-Build で実装しなかった機能には、進捗レポートや中央管理型のリソースロックなどがある。これらも実装したかったが、どうしてもという希望もなしに追加するには少し複雑すぎた。

同じような理屈が、おそらく Buildbot や CDash そして Jenkins にもあてはまるだろう。どのツールにも、有用なのに実装されていないという機能がある。おそらくそれはアーキテクチャ上の非互換性が原因だと思われる。しかし、Buildbot や CDash のコミュニティのメンバーとの議論や Jenkins のウェブサイトの記述によると、これらはまず欲しい機能を選ぶところから始めて、それからその機能を実装しやすいアーキテクチャでの開発を進めららしい。たとえば、CDash のコミュニティには比較的小規模なコア開発者チームがあり、中央管理型のモデルでソフトウェアを開発している。最優先で考えるのはソフトウェアをコアマシン群で動作させ続けることで、その次は技術力のあるユーザーからのバグ報告を受け付けることだ。一方 Buildbot が勢力を伸ばしているのは、複雑なビルド環境に多数のクライアントが絡み、共有リソースへのアクセスの調整を要するようなところである。Buildbot には柔軟な設定ファイルフォーマットがあり、スケジューリングや変更通知そしてリソースロックなどのさまざまなオプションを設定できる。そのあたりが好まれているのだろう。Jenkins が目指しているのは、使いやすさとシンプルな継続的インテグレーションだろう。設定用の GUI とローカルサーバーで動かすためのオプションもその一環だ。

オープンソース開発の社会学も、アーキテクチャと機能の相関関係における交絡因子となる。考えてみよう。もし、開発者たちがオープンソースプロジェクトを選ぶときの基準が、そのプロジェクトのアーキテクチャや機能が自分の用途に一致しているかどうかだとしたら？もしそうならば、彼らの貢献が反映されて、そのプロジェクトは今の機能をより伸ばすようになっていくだろう。その結果、プロジェクトは特定の機能セットにロックインされてしまうことになる。もともとその機能を好んで自ら選んだ人たちが開発を手伝うので、彼らが望む機能にうまくマッチしないアーキテクチャを避けるだろう。我々が Buildbot の開発に参加せず新たに Pony-Build を実装する道を選んだのも、まさにこの理由によるものだ。Buildbot のアーキテクチャは、何百何千ものパッケージをビルドするのには適さなかったのだ。

既存の継続的インテグレーションシステムは一般に、本質的に異なる二つのアーキテクチャのいずれかで構築されており、要求される機能のサブセットしか実装していない。CI システムが成熟してユーザー数が増えていくにつれて、さらに機能が追加されていくだろうと期待している。しかし機能を実装するには、選択したアーキテクチャによる制約があるかもしれない。今後どのように発展していくかが楽しみだ。

謝辞

グレッグ・ウィルソンやブレット・キャノン、エリック・ホルシャー、ジェシー・ノラー、ヴィクトリア・レイドラーらとは CI システム全般 (特に Pony-Build) について興味深い議論をさせていただいた。Pony-Build の開発には、ジャック・カールソンやファティマ・シェルカウイ、マックス・ライト、クシュブ・シャキヤらの学生たちが協力してくれた。

Eclipse

Kim Moir

Implementing software modularity is a notoriously difficult task. Interoperability with a large code base written by a diverse community is also difficult to manage. At Eclipse, we have managed to succeed on both counts. In June 2010, the Eclipse Foundation made available its Helios coordinated release, with over 39 projects and 490 committers from over 40 companies working together to build upon the functionality of the base platform. What was the original architectural vision for Eclipse? How did it evolve? How does the architecture of an application serve to encourage community engagement and growth? Let's go back to the beginning.

On November 7, 2001, an open source project called Eclipse 1.0 was released. At the time, Eclipse was described as “an integrated development environment (IDE) for anything and nothing in particular.” This description was purposely generic because the architectural vision was not just another set of tools, but a framework; a framework that was modular and scalable. Eclipse provided a component-based platform that could serve as the foundation for building tools for developers. This extensible architecture encouraged the community to build upon a core platform and extend it beyond the limits of the original vision. Eclipse started as a platform and the Eclipse SDK was the proof-of-concept product. The Eclipse SDK allowed the developers to self-host and use the Eclipse SDK itself to build newer versions of Eclipse.

The stereotypical image of an open source developer is that of an altruistic person toiling late into night fixing bugs and implementing fantastic new features to address their own personal interests. In contrast, if you look back at the early history of the Eclipse project, some of the initial code that was donated was based on VisualAge for Java, developed by IBM. The first committers who worked on this open source project were employees of an IBM subsidiary called Object Technology International (OTI). These committers were paid to work full time on the open source project, to answer questions on newsgroups, address bugs, and implement new features. A consortium of interested software vendors was formed to expand this open tooling effort. The initial members of the Eclipse consortium were Borland, IBM, Merant, QNX Software Systems, Rational Software,

RedHat, SuSE, and TogetherSoft.

By investing in this effort, these companies would have the expertise to ship commercial products based on Eclipse. This is similar to investments that corporations make in contributing to the Linux kernel because it is in their self-interest to have employees improving the open source software that underlies their commercial offerings. In early 2004, the Eclipse Foundation was formed to manage and expand the growing Eclipse community. This not-for-profit foundation was funded by corporate membership dues and is governed by a board of directors. Today, the diversity of the Eclipse community has expanded to include over 170 member companies and almost 1000 committers.

Originally, people knew “Eclipse” as the SDK only but today it is much more. In July 2010, there were 250 diverse projects under development at eclipse.org. There’s tooling to support developing with C/C++, PHP, web services, model driven development, build tooling and many more. Each of these projects is included in a top-level project (TLP) which is managed by a project management committee (PMC) consisting of senior members of the project nominated for the responsibility of setting technical direction and release goals. In the interests of brevity, the scope of this chapter will be limited to the evolution of the architecture of the Eclipse SDK within Eclipse¹ and Runtime Equinox² projects. Since Eclipse has long history, I’ll be focusing on early Eclipse, as well as the 3.0, 3.4 and 4.0 releases.

7.1 Early Eclipse

At the beginning of the 21st century, there were many tools for software developers, but few of them worked together. Eclipse sought to provide an open source platform for the creation of interoperable tools for application developers. This would allow developers to focus on writing new tools, instead of writing to code deal with infrastructure issues like interacting with the filesystem, providing software updates, and connecting to source code repositories. Eclipse is perhaps most famous for the Java Development Tools (JDT). The intent was that these exemplary Java development tools would serve as an example for people interested in providing tooling for other languages.

Before we delve into the architecture of Eclipse, let’s look at what the Eclipse SDK looks like to a developer. Upon starting Eclipse and selecting the workbench, you’ll be presented with the Java perspective. A perspective organizes the views and editors that are specific to the tooling that is currently in use.

Early versions of the Eclipse SDK architecture had three major elements, which corresponded to three major sub-projects: the Platform, the JDT (Java Development Tools) and the PDE (Plug-in Development Environment).

¹<http://www.eclipse.org>

²<http://www.eclipse.org/equinox>

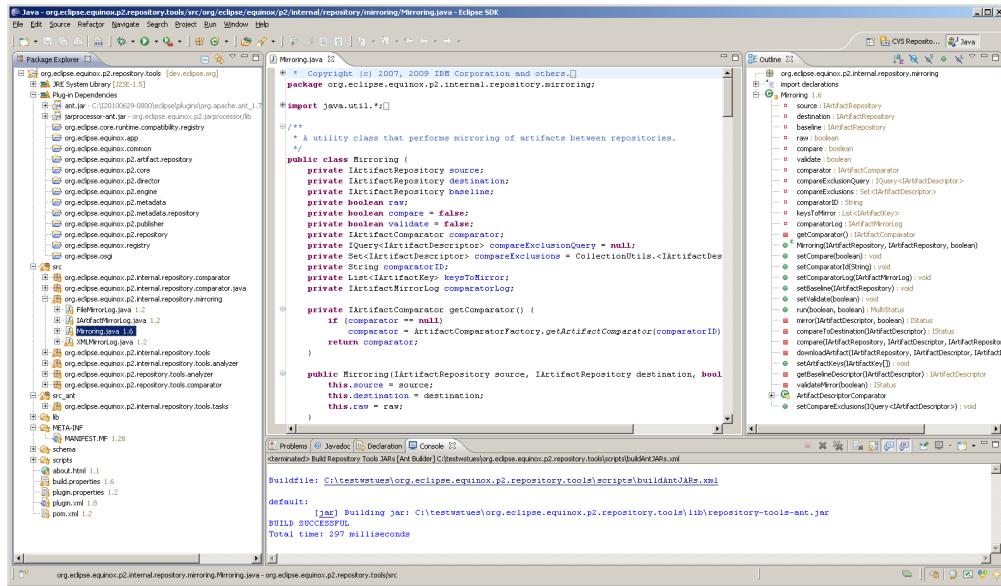


图 7.1: Java Perspective

Platform

The Eclipse platform is written using Java and a Java VM is required to run it. It is built from small units of functionality called plugins. Plugins are the basis of the Eclipse component model. A plugin is essentially a JAR file with a manifest which describes itself, its dependencies, and how it can be utilized, or extended. This manifest information was initially stored in a `plugin.xml` file which resides in the root of the plugin directory. The Java development tools provided plugins for developing in Java. The Plug-in Development Environment (PDE) provides tooling for developing plugins to extend Eclipse. Eclipse plugins are written in Java but could also contain non-code contributions such as HTML files for online documentation. Each plugin has its own class loader. Plugins can express dependencies on other plugins by the use of `requires` statements in the `plugin.xml`. Looking at the `plugin.xml` for the `org.eclipse.ui` plugin you can see its name and version specified, as well as the dependencies it needs to import from other plugins.

```
<?xml version="1.0" encoding="UTF-8"?>
<plugin
    id="org.eclipse.ui"
    name="%Plugin.name"
    version="2.1.1"
    provider-name="%Plugin.providerName"
    class="org.eclipse.ui.internal.UIPlugin">

    <runtime>
```

```

<library name="ui.jar">
    <export name="*"/>
    <packages prefixes="org.eclipse.ui"/>
</library>
</runtime>
<requires>
    <import plugin="org.apache.xerces"/>
    <import plugin="org.eclipse.core.resources"/>
    <import plugin="org.eclipse.update.core"/>
    :
    :
    <import plugin="org.eclipse.text" export="true"/>
    <import plugin="org.eclipse.ui.workbench.texteditor" export="true"/>
    <import plugin="org.eclipse.ui.editors" export="true"/>
</requires>
</plugin>

```

In order to encourage people to build upon the Eclipse platform, there needs to be a mechanism to make a contribution to the platform, and for the platform to accept this contribution. This is achieved through the use of extensions and extension points, another element of the Eclipse component model. The export identifies the interfaces that you expect others to use when writing their extensions, which limits the classes that are available outside your plugin to the ones that are exported. It also provides additional limitations on the resources that are available outside the plugin, as opposed to making all public methods or classes available to consumers. Exported plugins are considered public API. All others are considered private implementation details. To write a plugin that would contribute a menu item to the Eclipse toolbar, you can use the actionSets extension point in the org.eclipse.ui plugin.

```

<extension-point id="actionSets" name="%ExtPoint.actionSets"
                 schema="schema/actionSets.exsd"/>
<extension-point id="commands" name="%ExtPoint.commands"
                 schema="schema/commands.exsd"/>
<extension-point id="contexts" name="%ExtPoint.contexts"
                 schema="schema/contextes.exsd"/>
<extension-point id="decorators" name="%ExtPoint.decorators"
                 schema="schema/decorators.exsd"/>
<extension-point id="dropActions" name="%ExtPoint.dropActions"
                 schema="schema/dropActions.exsd"/> =

```

Your plugin's extension to contribute a menu item to the org.eclipse.ui.actionSet extension point would look like:

```

<?xml version="1.0" encoding="UTF-8"?>
<plugin
    id="com.example.helloworld"
    name="com.example.helloworld"
    version="1.0.0">
    <runtime>
        <library name="helloworld.jar"/>

```

```

</runtime>
<requires>
    <import plugin="org.eclipse.ui"/>
</requires>
<extension
    point="org.eclipse.ui.actionSets">
    <actionSet
        label="Example Action Set"
        visible="true"
        id="org.eclipse.helloworld.actionSet">
        <menu
            label="Example &Menu"
            id="exampleMenu">
            <separator
                name="exampleGroup">
            </separator>
        </menu>
        <action
            label="&Example Action"
            icon="icons/example.gif"
            tooltip="Hello, Eclipse world"
            class="com.example.helloworld.actions.ExampleAction"
            menubarPath="exampleMenu/exampleGroup"
            toolbarPath="exampleGroup"
            id="org.eclipse.helloworld.actions.ExampleAction">
        </action>
    </actionSet>
</extension>
</plugin>

```

When Eclipse is started, the runtime platform scans the manifests of the plugins in your install, and builds a plugin registry that is stored in memory. Extension points and the corresponding extensions are mapped by name. The resulting plugin registry can be referenced from the API provided by the Eclipse platform. The registry is cached to disk so that this information can be reloaded the next time Eclipse is restarted. All plugins are discovered upon startup to populate the registry but they are not activated (classes loaded) until the code is actually used. This approach is called lazy activation. The performance impact of adding additional bundles into your install is reduced by not actually loading the classes associated with the plugins until they are needed. For instance, the plugin that contributes to the org.eclipse.ui.actionSet extension point wouldn't be activated until the user selected the new menu item in the toolbar.



図 7.2: Example Menu

The code that generates this menu item looks like this:

```
package com.example.helloworld.actions;

import org.eclipse.jface.action.IAction;
import org.eclipse.jface.viewers.ISelection;
import org.eclipse.ui.IWorkbenchWindow;
import org.eclipse.ui.IWorkbenchWindowActionDelegate;
import org.eclipse.jface.dialogs.MessageDialog;

public class ExampleAction implements IWorkbenchWindowActionDelegate {
    private IWorkbenchWindow window;

    public ExampleAction() {
    }

    public void run(IAction action) {
        MessageDialog.openInformation(
            window.getShell(),
            "org.eclipse.helloworld",
            "Hello, Eclipse architecture world");
    }

    public void selectionChanged(IAction action, ISelection selection) {
    }

    public void dispose() {
    }

    public void init(IWorkbenchWindow window) {
        this.window = window;
    }
}
```

Once the user selects the new item in the toolbar, the extension registry is queried by the plugin implementing the extension point. The plugin supplying the extension instantiates the contribution, and loads the plugin. Once the plugin is activated, the `ExampleAction` constructor in our example is run, and then initializes a Workbench action delegate. Since the selection in the workbench has changed and the delegate has been created, the action can change. The message dialog opens with the message “Hello, Eclipse architecture world”.

This extensible architecture was one of the keys to the successful growth of the Eclipse ecosystem. Companies or individuals could develop new plugins, and either release them as open source or sell them commercially.

One of the most important concepts about Eclipse is that *everything is a plugin*. Whether the plugin is included in the Eclipse platform, or you write it yourself, plugins are all first class components of the assembled application. 図 7.3 shows clusters of related functionality contributed by plugins in early versions of Eclipse.

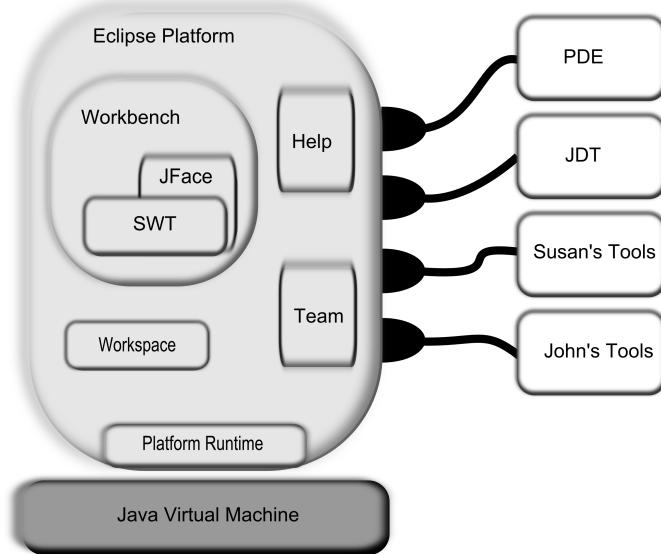


図 7.3: Early Eclipse Architecture

The workbench is the most familiar UI element to users of the Eclipse platform, as it provides the structures that organize how Eclipse appears to the user on the desktop. The workbench consists of perspectives, views, and editors. Editors are associated with file types so the correct editor is launched when a file is opened. An example of a view is the “problems” view that indicates errors or warnings in your Java code. Together, editors and views form a perspective which presents the tooling to the user in an organized fashion.

The Eclipse workbench is built on the Standard Widget Toolkit (SWT) and JFace, and SWT deserves a bit of exploration. Widget toolkits are generally classified as either native or emulated. A native widget toolkit uses operating system calls to build user interface components such as lists and push buttons. Interaction with components is handled by the operating system. An emulated widget toolkit implements components outside of the operating system, handling mouse and keyboard, drawing, focus and other widget functionality itself, rather than deferring to the operating system. Both designs have different strengths and weaknesses.

Native widget toolkits are “pixel perfect.” Their widgets look and feel like their counterparts in other applications on the desktop. Operating system vendors constantly change the look and feel of their widgets and add new features. Native widget toolkits get these updates for free. Unfortunately, native toolkits are difficult to implement because their underlying operating system widget implementations are vastly different, leading to inconsistencies and programs that are not portable.

Emulated widget toolkits either provide their own look and feel, or try to draw and behave like the operating system. Their great strength over native toolkits is flexibility (although modern native

widget toolkits such as Windows Presentation Framework (WPF) are equally as flexible). Because the code to implement a widget is part of the toolkit rather than embedded in the operating system, a widget can be made to draw and behave in any manner. Programs that use emulated widget toolkits are highly portable. Early emulated widget toolkits had a bad reputation. They were often slow and did a poor job of emulating the operating system, making them look out of place on the desktop. In particular, Smalltalk-80 programs at the time were easy to recognize due to their use of emulated widgets. Users were aware that they were running a “Smalltalk program” and this hurt acceptance of applications written in Smalltalk.

Unlike other computer languages such as C and C++, the first versions of Java came with a native widget toolkit library called the Abstract Window Toolkit (AWT). AWT was considered to be limited, buggy and inconsistent and was widely decried. At Sun and elsewhere, in part because of experience with AWT, a native widget toolkit that was portable and performant was considered to be unworkable. The solution was Swing, a full-featured emulated widget toolkit.

Around 1999, OTI was using Java to implement a product called VisualAge Micro Edition. The first version of VisualAge Micro Edition used Swing and OTI’s experience with Swing was not positive. Early versions of Swing were buggy, had timing and memory issues and the hardware at the time was not powerful enough to give acceptable performance. OTI had successfully built a native widget toolkit for Smalltalk-80 and other Smalltalk implementations to gain acceptance of Smalltalk. This experience was used to build the first version of SWT. VisualAge Micro Edition and SWT were a success and SWT was the natural choice when work began on Eclipse. The use of SWT over Swing in Eclipse split the Java community. Some saw conspiracies, but Eclipse was a success and the use of SWT differentiated it from other Java programs. Eclipse was performant, pixel perfect and the general sentiment was, “I can’t believe it’s a Java program.”

Early Eclipse SDKs ran on Linux and Windows. In 2010, there is support for over a dozen platforms. A developer can write an application for one platform, and deploy it to multiple platforms. Developing a new widget toolkit for Java was a contentious issue within the Java community at the time, but the Eclipse committers felt that it was worth the effort to provide the best native experience on the desktop. This assertion applies today, and there are millions of lines of code that depend on SWT.

JFace is a layer on top of SWT that provides tools for common UI programming tasks, such as frameworks for preferences and wizards. Like SWT, it was designed to work with many windowing systems. However, it is pure Java code and doesn’t contain any native platform code.

The platform also provided an integrated help system based upon small units of information called topics. A topic consists of a label and a reference to its location. The location can be an HTML documentation file, or an XML document describing additional links. Topics are grouped together in table of contents (TOCs). Consider the topics as the leaves, and TOCs as the branches of organization. To add help content to your application, you can contribute to the `org.eclipse.help.toc` extension point, as the `org.eclipse.platform.doc.isv plugin.xml` does below.

```

<?xml version="1.0" encoding="UTF-8"?>
<?eclipse version="3.0"?>
<plugin>

<!-- ===== -->
<!-- Define primary TOC -->
<!-- ===== -->
<extension
    point="org.eclipse.help.toc">
    <toc
        file="toc.xml"
        primary="true">
    </toc>
    <index path="index"/>
</extension>
<!-- ===== -->
<!-- Define TOCs -->
<!-- ===== -->
<extension
    point="org.eclipse.help.toc">
    <toc
        file="topics_Guide.xml">
    </toc>
    <toc
        file="topics_Reference.xml">
    </toc>
    <toc
        file="topics_Porting.xml">
    </toc>
    <toc
        file="topics_Questions.xml">
    </toc>
    <toc
        file="topics_Samples.xml">
    </toc>
</extension>

```

Apache Lucene is used to index and search the online help content. In early versions of Eclipse, online help was served as a Tomcat web application. Additionally, by providing help within Eclipse itself, you can also use the subset of help plugins to provide a standalone help server.³

Eclipse also provides team support to interact with a source code repository, create patches and other common tasks. The workspace provided collection of files and metadata that stored your work on the filesystem. There was also a debugger to trace problems in the Java code, as well as a framework for building language specific debuggers.

One of the goals of the Eclipse project was to encourage open source and commercial consumers of this technology to extend the platform to meet their needs, and one way to encourage this adoption is to provide a stable API. An API can be thought of as a technical contract specifying the behavior

³For example: <http://help.eclipse.org>.

of your application. It also can be thought of as a social contract. On the Eclipse project, the mantra is, “API is forever”. Thus careful consideration must be given when writing an API given that it is meant to be used indefinitely. A stable API is a contract between the client or API consumer and the provider. This contract ensures that the client can depend on the Eclipse platform to provide the API for the long term without the need for painful refactoring on the part of the client. A good API is also flexible enough to allow the implementation to evolve.

Java Development Tools (JDT)

The JDT provides Java editors, wizards, refactoring support, debugger, compiler and an incremental builder. The compiler is also used for content assist, navigation and other editing features. A Java SDK isn’t shipped with Eclipse so it’s up to the user to choose which SDK to install on their desktop. Why did the JDT team write a separate compiler to compile your Java code within Eclipse? They had an initial compiler code contribution from VisualAge Micro Edition. They planned to build tooling on top of the compiler, so writing the compiler itself was a logical decision. This approach also allowed the JDT committers to provide extension points for extending the compiler. This would be difficult if the compiler was a command line application provided by a third party.

Writing their own compiler provided a mechanism to provide support for an incremental builder within the IDE. An incremental builder provides better performance because it only recompiles files that have changed or their dependencies. How does the incremental builder work? When you create a Java project within Eclipse, you are creating resources in the workspace to store your files. A builder within Eclipse takes the inputs within your workspace (. java files), and creates an output (. class files). Through the build state, the builder knows about the types (classes or interfaces) in the workspace, and how they reference each other. The build state is provided to the builder by the compiler each time a source file is compiled. When an incremental build is invoked, the builder is supplied with a resource delta, which describes any new, modified or deleted files. Deleted source files have their corresponding class files deleted. New or modified types are added to a queue. The files in the queue are compiled in sequence and compared with the old class file to determine if there are structural changes. Structural changes are modifications to the class that can impact another type that references it. For example, changing a method signature, or adding or removing a method. If there are structural changes, all the types that reference it are also added to the queue. If the type has changed at all, the new class file is written to the build output folder. The build state is updated with reference information for the compiled type. This process is repeated for all the types in the queue until empty. If there are compilation errors, the Java editor will create problem markers. Over the years, the tooling that JDT provides has expanded tremendously in concert with new versions of the Java runtime itself.

Plug-in Development Environment (PDE)

The Plug-in Development Environment (PDE) provided the tooling to develop, build, deploy and test plugins and other artifacts that are used to extend the functionality of Eclipse. Since Eclipse plugins were a new type of artifact in the Java world there wasn't a build system that could transform the source into plugins. Thus the PDE team wrote a component called PDE Build which examined the dependencies of the plugins and generated Ant scripts to construct the build artifacts.

7.2 Eclipse 3.0: Runtime, RCP and Robots

Runtime

Eclipse 3.0 was probably one of the most important Eclipse releases due to the number of significant changes that occurred during this release cycle. In the pre-3.0 Eclipse architecture, the Eclipse component model consisted of plugins that could interact with each other in two ways. First, they could express their dependencies by the use of the `requires` statement in their `plugin.xml`. If plugin A requires plugin B, plugin A can see all the Java classes and resources from B, respecting Java class visibility conventions. Each plugin had a version, and they could also specify the versions of their dependencies. Secondly, the component model provided *extensions* and *extension points*. Historically, Eclipse committers wrote their own runtime for the Eclipse SDK to manage classloading, plugin dependencies and extensions and extension points.

The Equinox project was created as a new incubator project at Eclipse. The goal of the Equinox project was to replace the Eclipse component model with one that already existed, as well as provide support for dynamic plugins. The solutions under consideration included JMX, Jakarta Avalon and OSGi. JMX was not a fully developed component model so it was not deemed appropriate. Jakarta Avalon wasn't chosen because it seemed to be losing momentum as a project. In addition to the technical requirements, it was also important to consider the community that supported these technologies. Would they be willing to incorporate Eclipse-specific changes? Was it actively developed and gaining new adopters? The Equinox team felt that the community around their final choice of technology was just as important as the technical considerations.

After researching and evaluating the available alternatives, the committers selected OSGi. Why OSGi? It had a semantic versioning scheme for managing dependencies. It provided a framework for modularity that the JDK itself lacked. Packages that were available to other bundles must be explicitly exported, and all others were hidden. OSGi provided its own classloader so the Equinox team didn't have to continue to maintain their own. By standardizing on a component model that had wider adoption outside the Eclipse ecosystem, they felt they could appeal to a broader community and further drive the adoption of Eclipse.

The Equinox team felt comfortable that since OSGi already had an existing and vibrant community, they could work with that community to help include the functionality that Eclipse required in a component model. For instance, at the time, OSGi only supported listing requirements at a package level, not a plugin level as Eclipse required. In addition, OSGi did not yet include the concept of fragments, which were Eclipse's preferred mechanism for supplying platform or environment specific code to an existing plugin. For example, fragments provide code for working with Linux and Windows filesystems as well as fragments which contribute language translations. Once the decision was made to proceed with OSGi as the new runtime, the committers needed an open source framework implementation. They evaluated Oscar, the precursor to Apache Felix, and the Service Management Framework (SMF) developed by IBM. At the time, Oscar was a research project with limited deployment. SMF was ultimately chosen since it was already used in shipping products and thus was deemed enterprise-ready. The Equinox implementation serves as the reference implementation of the OSGi specification.

A compatibility layer was also provided so that existing plugins would still work in a 3.0 install. Asking developers to rewrite their plugins to accommodate changes in the underlying infrastructure of Eclipse 3.0 would have stalled the momentum on Eclipse as a tooling platform. The expectation from Eclipse consumers was that the platform should just continue to work.

With the switch to OSGi, Eclipse plugins became known as bundles. A plugin and a bundle are the same thing: They both provide a modular subset of functionality that describes itself with metadata in a manifest. Previously, dependencies, exported packages and the extensions and extension points were described in `plugin.xml`. With the move to OSGi bundles, the extensions and extension points continued to be described in `plugin.xml` since they are Eclipse concepts. The remaining information was described in the `META-INF/MANIFEST.MF`, OSGi's version of the bundle manifest. To support this change, PDE provided a new manifest editor within Eclipse. Each bundle has a name and version. The manifest for the `org.eclipse.ui` bundle looks like this:

```
Manifest-Version: 1.0
Bundle-ManifestVersion: 2
Bundle-Name: %Plugin.name
Bundle-SymbolicName: org.eclipse.ui; singleton:=true
Bundle-Version: 3.3.0.qualifier
Bundle-ClassPath: .
Bundle-Activator: org.eclipse.ui.internal.UIPlugin
Bundle-Vendor: %Plugin.providerName
Bundle-Localization: plugin
Export-Package: org.eclipse.ui.internal;x-internal:=true
Require-Bundle: org.eclipse.core.runtime;bundle-version="[3.2.0,4.0.0)",
  org.eclipse.swt;bundle-version="[3.3.0,4.0.0)";visibility:=reexport,
  org.eclipse.jface;bundle-version="[3.3.0,4.0.0)";visibility:=reexport,
  org.eclipse.ui.workbench;bundle-version="[3.3.0,4.0.0)";visibility:=reexport,
  org.eclipse.core.expressions;bundle-version="[3.3.0,4.0.0)"
Eclipse-LazyStart: true
Bundle-RequiredExecutionEnvironment: CDC-1.0/Foundation-1.0, J2SE-1.3
```

As of Eclipse 3.1, the manifest can also specify a bundle required execution environment (BREE). Execution environments specify the minimum Java environment required for the bundle to run. The Java compiler does not understand bundles and OSGi manifests. PDE provides tooling for developing OSGi bundles. Thus, PDE parses the bundle's manifest, and generates the classpath for that bundle. If you specified an execution environment of J2SE-1.4 in your manifest, and then wrote some code that included generics, you would be advised of compile errors in your code. This ensures that your code adheres to the contract you have specified in the manifest.

OSGi provides a modularity framework for Java. The OSGi framework manages collections of self-describing bundles and manages their classloading. Each bundle has its own classloader. The classpath available to a bundle is constructed by examining the dependencies of the manifest and generating a classpath available to the bundle. OSGi applications are collections of bundles. In order to fully embrace of modularity, you must be able to express your dependencies in a reliable format for consumers. Thus the manifest describes exported packages that are available to clients of this bundle which corresponds to the public API that was available for consumption. The bundle that is consuming that API must have a corresponding import of the package they are consuming. The manifest also allows you to express version ranges for your dependencies. Looking at the `Require-Bundle` heading in the above manifest, you will note that the `org.eclipse.core.runtime` bundle that `org.eclipse.ui` depends on must be at least 3.2.0 and less than 4.0.0.

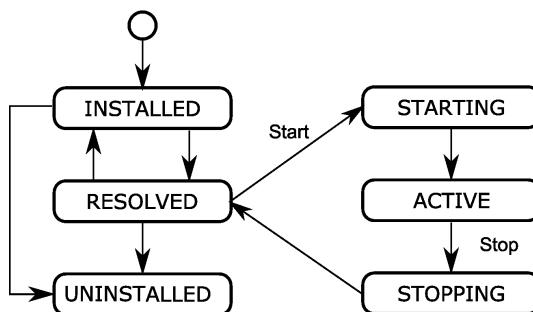


図 7.4: OSGi Bundle Lifecycle

OSGi is a dynamic framework which supports the installation, starting, stopping, or uninstallation of bundles. As mentioned before, lazy activation was a core advantage to Eclipse because plugin classes were not loaded until they were needed. The OSGi bundle lifecycle also enables this approach. When you start an OSGi application, the bundles are in the installed state. If its dependencies are met, the bundle changes to the resolved state. Once resolved, the classes within that bundle can be loaded and run. The starting state means that the bundle is being activated according to its activation policy. Once activated, the bundle is in the active state, it can acquire required resources and interact with other bundles. A bundle is in the stopping state when it is executing

its activator stop method to clean up any resources that were opened when it was active. Finally, a bundle may be uninstalled, which means that it's not available for use.

As the API evolves, there needs to be a way to signal changes to your consumers. One approach is to use semantic versioning of your bundles and version ranges in your manifests to specify the version ranges for your dependencies. OSGi uses a four-part versioning naming scheme as shown in 図 7.5.

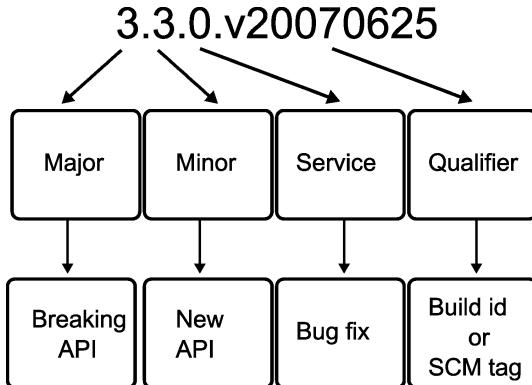


図 7.5: Versioning Naming Scheme

With the OSGi version numbering scheme, each bundle has a unique identifier consisting of a name and a four part version number. An id and version together denote a unique set of bytes to the consumer. By Eclipse convention, if you're making changes to a bundle, each segment of the version signifies to the consumer the type of change being made. Thus, if you want to indicate that you intend to break API, you increment the first (major) segment. If you have just added API, you increment the second (minor) segment. If you fix a small bug that doesn't impact API, the third (service) segment is incremented. Finally, the fourth or qualifier segment is incremented to indicate a build id source control repository tag.

In addition to expressing the fixed dependencies between bundles, there is also a mechanism within OSGi called services which provides further decoupling between bundles. Services are objects with a set of properties that are registered with the OSGi service registry. Unlike extensions, which are registered in the extension registry when Eclipse scans bundles during startup, services are registered dynamically. A bundle that is consuming a service needs to import the package defining the service contract, and the framework determines the service implementation from the service registry.

Like a main method in a Java class file, there is a specific application defined to start Eclipse. Eclipse applications are defined using extensions. For instance, the application to start the Eclipse IDE itself is `org.eclipse.ui.ide.workbench` which is defined in the `org.eclipse.ui.ide.application` bundle.

```

<plugin>
  <extension
    id="org.eclipse.ui.ide.workbench"
    point="org.eclipse.core.runtime.applications">
    <application>
      <run
        class="org.eclipse.ui.internal.ide.application.IDEApplication">
      </run>
    </application>
  </extension>
</plugin>

```

There are many applications provided by Eclipse such as those to run standalone help servers, Ant tasks, and JUnit tests.

Rich Client Platform (RCP)

One of the most interesting things about working in an open source community is that people use the software in totally unexpected ways. The original intent of Eclipse was to provide a platform and tooling to create and extend IDEs. However, in the time leading up to the 3.0 release, bug reports revealed that the community was taking a subset of the platform bundles and using them to build Rich Client Platform (RCP) applications, which many people would recognize as Java applications. Since Eclipse was initially constructed with an IDE-centric focus, there had to be some refactoring of the bundles to allow this use case to be more easily adopted by the user community. RCP applications didn't require all the functionality in the IDE, so several bundles were split into smaller ones that could be consumed by the community for building RCP applications.

Examples of RCP applications in the wild include the use of RCP to monitor the Mars Rover robots developed by NASA at the Jet Propulsion Laboratory, Bioclipse for data visualization of bioinformatics and Dutch Railway for monitoring train performance. The common thread that ran through many of these applications was that these teams decided that they could take the utility provided by the RCP platform and concentrate on building their specialized tools on top of it. They could save development time and money by focusing on building their tools on a platform with a stable API that guaranteed that their technology choice would have long term support.

Looking at the 3.0 architecture in 図 7.6, you will note that the Eclipse Runtime still exists to provide the application model and extension registry. Managing the dependencies between components, the plugin model is now managed by OSGi. In addition to continuing to be able to extend Eclipse for their own IDEs, consumers can also build upon the RCP application framework for more generic applications.

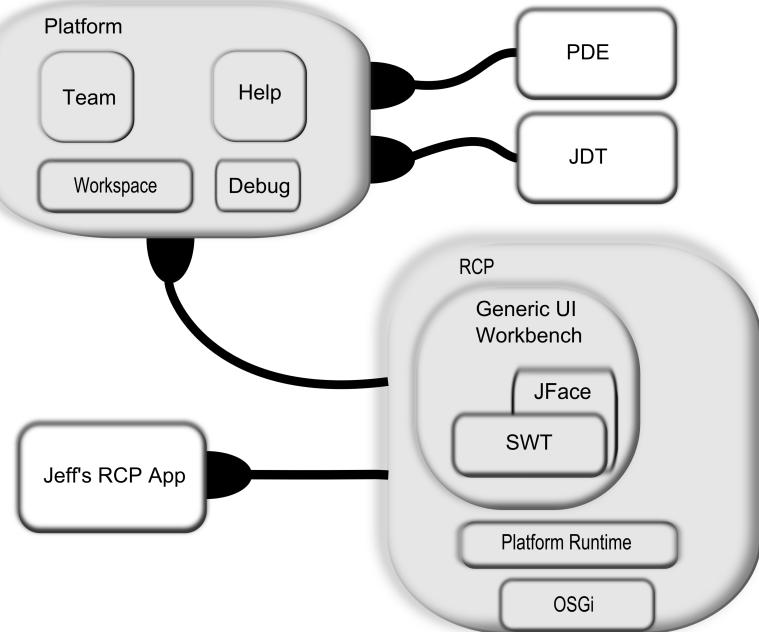


図 7.6: Eclipse 3.0 Architecture

7.3 Eclipse 3.4

The ability to easily update an application to a new version and add new content is taken for granted. In Firefox it happens seamlessly. For Eclipse it hasn't been so easy. Update Manager was the original mechanism that was used to add new content to the Eclipse install or update to a new version.

To understand what changes during an update or install operation, it's necessary to understand what Eclipse means by “features”. A feature is a PDE artifact that defines a set of bundles that are packaged together in a format that can be built or installed. Features can also include other features. (See 図 7.7.)

If you wished to update your Eclipse install to a new build that only incorporated one new bundle, the entire feature had to be updated since this was the coarse grained mechanism that was used by update manager. Updating a feature to fix a single bundle is inefficient.

There are PDE wizards to create features, and build them in your workspace. The `feature.xml` file defines the bundles included in the feature, and some simple properties of the bundles. A feature, like a bundle, has a name and a version. Features can include other features, and specify version ranges for the features they include. The bundles that are included in a feature are listed, along with specific properties. For instance, you can see that the `org.eclipse.launcher.gtk.linux.x86_64`

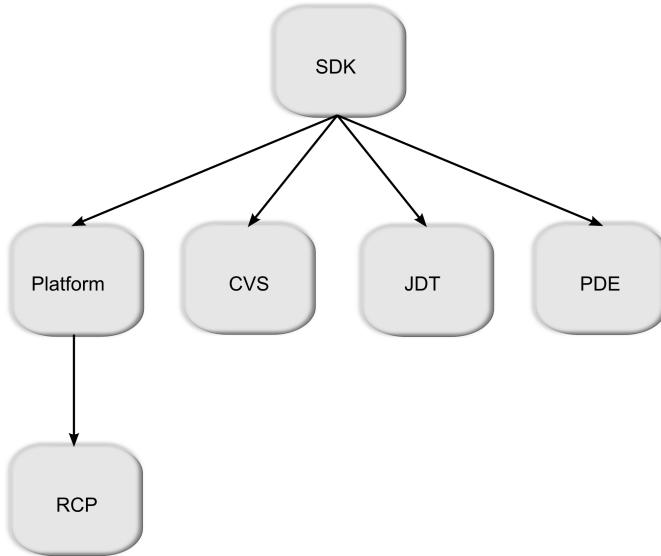


図 7.7: Eclipse 3.3 SDK Feature Hierarchy

fragment specifies the operating system (os), windowing system (ws) and architecture (arch) where it should be used. Thus upgrading to a new release, this fragment would only be installed on this platform. These platform filters are included in the OSGi manifest of this bundle.

```

<?xml version="1.0" encoding="UTF-8"?>
<feature
    id="org.eclipse.rcp"
    label="%featureName"
    version="3.7.0.qualifier"
    provider-name="%providerName"
    plugin="org.eclipse.rcp"
    image="eclipse_update_120.jpg">

    <description>
        %description
    </description>

    <copyright>
        %copyright
    </copyright>

    <license url="%licenseURL">
        %license
    </license>

    <plugin
        id="org.eclipse.equinox.launcher"
        download-size="0"
  
```

```
install-size="0"
version="0.0.0"
unpack="false"/>

<plugin
    id="org.eclipse.equinox.launcher.gtk.linux.x86_64"
    os="linux"
    ws="gtk"
    arch="x86_64"
    download-size="0"
    install-size="0"
    version="0.0.0"
    fragment="true"/>
```

An Eclipse application consists of more than just features and bundles. There are platform specific executables to start Eclipse itself, license files, and platform specific libraries, as shown in this list of files included in the Eclipse application.

```
com.ibm.icu
org.eclipse.core.commands
org.eclipse.core.contenttype
org.eclipse.core.databinding
org.eclipse.core.databinding.beans
org.eclipse.core.expressions
org.eclipse.core.jobs
org.eclipse.core.runtime
org.eclipse.core.runtime.compatibility.auth
org.eclipse.equinox.common
org.eclipse.equinox.launcher
org.eclipse.equinox.launcher.carbon.macosx
org.eclipse.equinox.launcher.gtk.linux.ppc
org.eclipse.equinox.launcher.gtk.linux.s390
org.eclipse.equinox.launcher.gtk.linux.s390x
org.eclipse.equinox.launcher.gtk.linux.x86
org.eclipse.equinox.launcher.gtk.linux.x86_64
```

These files couldn't be updated via update manager, because again, it only dealt with features. Since many of these files were updated every major release, this meant that users had to download a new zip each time there was a new release instead of updating their existing install. This wasn't acceptable to the Eclipse community. PDE provided support for product files, which specified all the files needed to build an Eclipse RCP application. However, update manager didn't have a mechanism to provision these files into your install which was very frustrating for users and product developers alike. In March 2008, p2 was released into the SDK as the new provisioning solution. In the interest of backward compatibility, Update Manager was still available for use, but p2 was enabled by default.

p2 Concepts

Equinox p2 is all about installation units (IU). An IU is a description of the name and id of the artifact you are installing. This metadata also describes the capabilities of the artifact (what is provided) and its requirements (its dependencies). Metadata can also express applicability filters if an artifact is only applicable to a certain environment. For instance, the org.eclipse.swt.gtk.linux.x86 fragment is only applicable if you're installing on a Linux gtk x86 machine. Fundamentally, metadata is an expression of the information in the bundle's manifest. Artifacts are simply the binary bits being installed. A separation of concerns is achieved by separating the metadata and the artifacts that they describe. A p2 repository consists of both metadata and artifact repositories.

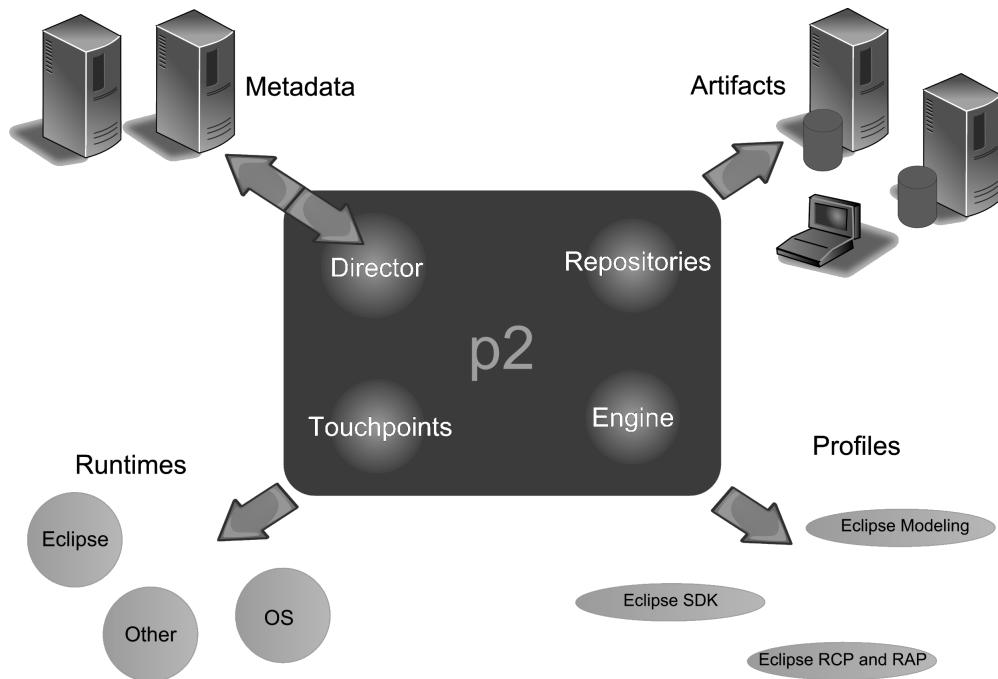


図 7.8: P2 Concepts

A profile is a list of IUs in your install. For instance, your Eclipse SDK has a profile that describes your current install. From within Eclipse, you can request an update to a newer version of the build which will create a new profile with a different set of IUs. A profile also provides a list of properties associated with the installation, such as the operating system, windowing system, and architecture parameters. Profiles also store the installation directory and the location. Profiles are held by a profile registry, which can store multiple profiles. The director is responsible for invoking provisioning operations. It works with the planner and the engine. The planner examines the existing profile, and determines the operations that must occur to transform the install into its new state.

The engine is responsible for carrying out the actual provisioning operations and installing the new artifacts on disk. Touchpoints are part of the engine that work with the runtime implementation of the system being installed. For instance, for the Eclipse SDK, there is an Eclipse touchpoint which knows how to install bundles. For a Linux system where Eclipse is installed from RPM binaries, the engine would deal with an RPM touchpoint. Also, p2 can perform installs in-process or outside in a separate process, such as a build.

There were many benefits to the new p2 provisioning system. Eclipse install artifacts could be updated from release to release. Since previous profiles were stored on disk, there was also a way to revert to a previous Eclipse install. Additionally, given a profile and a repository, you could recreate the Eclipse install of a user that was reporting a bug to try to reproduce the problem on your own desktop. Provisioning with p2 provided a way to update and install more than just the Eclipse SDK, it was a platform that applied to RCP and OSGi use cases as well. The Equinox team also worked with the members of another Eclipse project, the Eclipse Communication Framework (ECF) to provide reliable transport for consuming artifacts and metadata in p2 repositories.

There were many spirited discussions within the Eclipse community when p2 was released into the SDK. Since update manager was a less than optimal solution for provisioning your Eclipse install, Eclipse consumers had the habit of unzipping bundles into their install and restarting Eclipse. This approach resolves your bundles on a best effort basis. It also meant that any conflicts in your install were being resolved at runtime, not install time. Constraints should be resolved at install time, not run time. However, users were often oblivious to these issues and assumed since the bundles existed on disk, they were working. Previously, the update sites that Eclipse provided were a simple directory consisting of JARred bundles and features. A simple `site.xml` file provided the names of the features that were available to be consumed in the site. With the advent of p2, the metadata that was provided in the p2 repositories was much more complex. To create metadata, the build process needed to be tweaked to either generate metadata at build time or run a generator task over the existing bundles. Initially, there was a lack of documentation available describing how to make these changes. As well, as is always the case, exposing new technology to a wider audience exposed unexpected bugs that had to be addressed. However, by writing more documentation and working long hours to address these bugs, the Equinox team was able to address these concerns and now p2 is the underlying provision engine behind many commercial offerings. As well, the Eclipse Foundation ships its coordinated release every year using a p2 aggregate repository of all the contributing projects.

7.4 Eclipse 4.0

Architecture must continually be examined to evaluate if it is still appropriate. Is it able to incorporate new technology? Does it encourage growth of the community? Is it easy to attract new

contributors? In late 2007, the Eclipse project committers decided that the answers to these questions were no and they embarked on designing a new vision for Eclipse. At the same time, they realized that there were thousands of Eclipse applications that depended on the existing API. An incubator technology project was created in late 2008 with three specific goals: simplify the Eclipse programming model, attract new committers and enable the platform to take advantage of new web-based technologies while providing an open architecture.

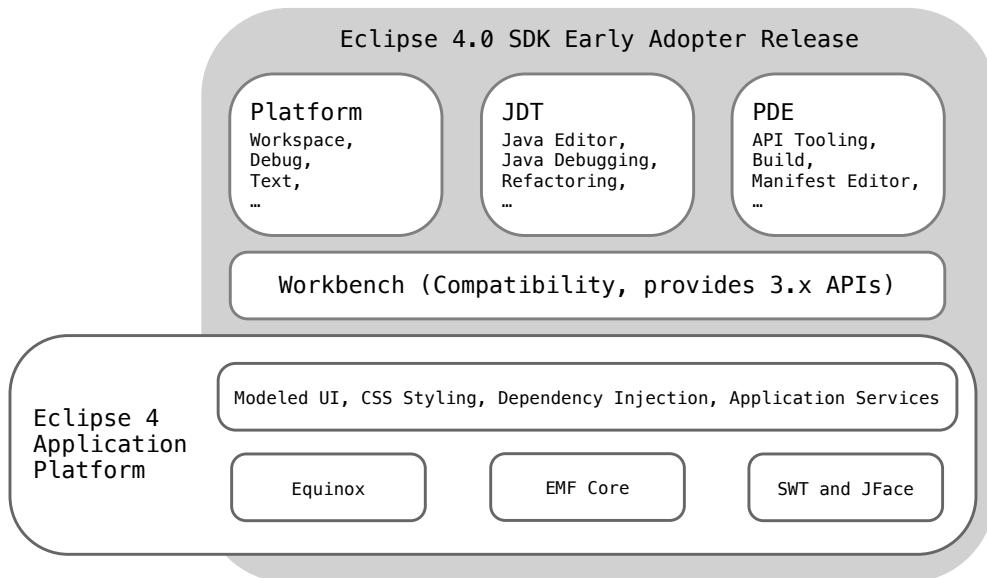


図 7.9: Eclipse 4.0 SDK Early Adopter Release

Eclipse 4.0 was first released in July 2010 for early adopters to provide feedback. It consisted of a combination of SDK bundles that were part of the 3.6 release, and new bundles that graduated from the technology project. Like 3.0, there was a compatibility layer so that existing bundles could work with the new release. As always, there was the caveat that consumers needed to be using the public API in order to be assured of that compatibility. There was no such guarantee if your bundle used internal code. The 4.0 release provided the Eclipse 4 Application Platform which provided the following features.

Model Workbench

In 4.0, a model workbench is generated using the Eclipse Modeling Framework (EMFgc). There is a separation of concerns between the model and the rendering of the view, since the renderer talks to the model and then generates the SWT code. The default is to use the SWT renderers, but other solutions are possible. If you create an example 4.x application, an XMI file will be created for the

defau
to ref
appli

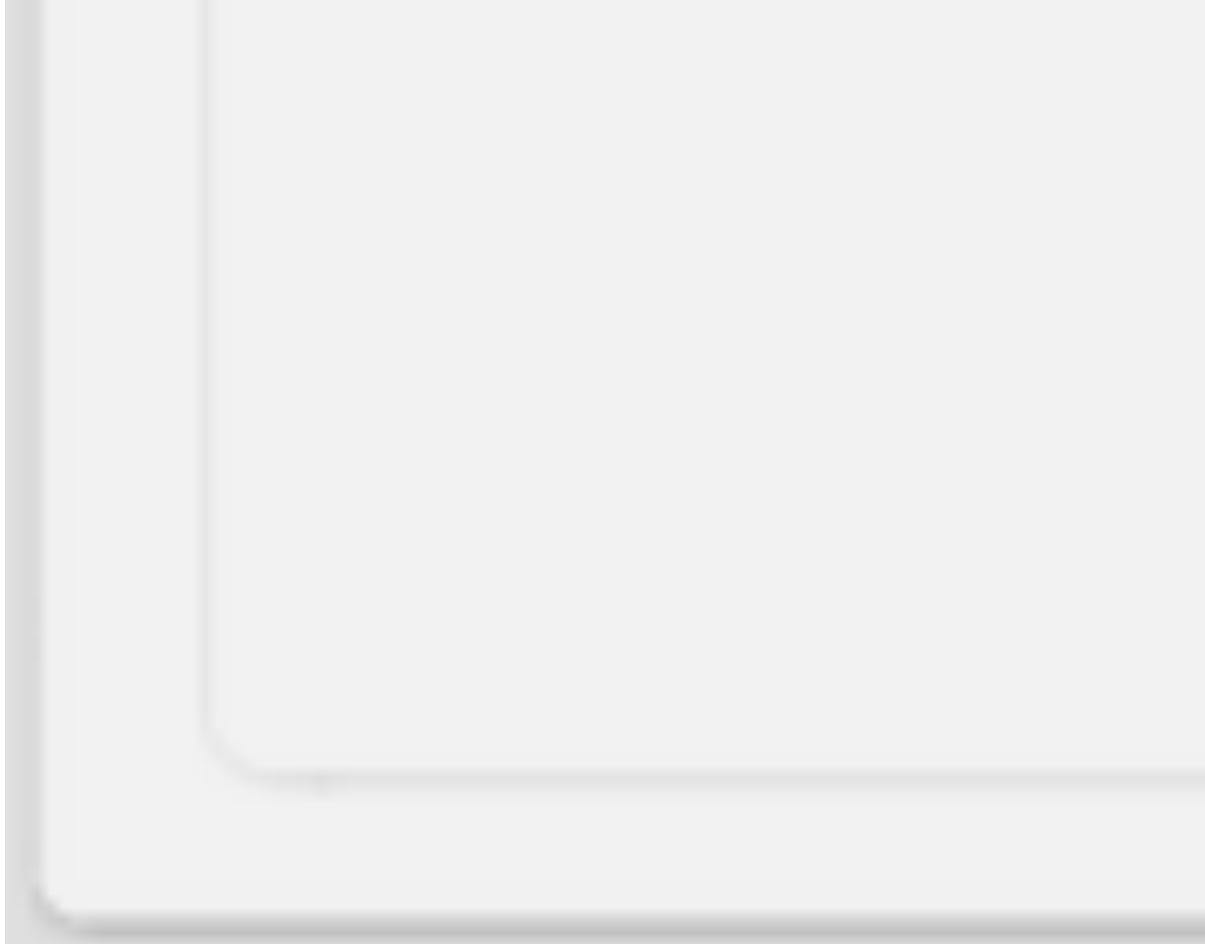


図 7.10: Model Generated for Example 4.x Application

Cascading Style Sheets Styling

Eclipse was released in 2001, before the era of rich Internet applications that could be skinned via CSS to provide a different look and feel. Eclipse 4.0 provides the ability to use stylesheets to easily change the look and feel of the Eclipse application. The default CSS stylesheets can be found in the `css` folder of the `org.eclipse.platform` bundle.

Dependency Injection

Both the Eclipse extensions registry and OSGi services are examples of service programming models. By convention, a service programming model contains service producers and consumers. The broker is responsible for managing the relationship between producers and consumers.

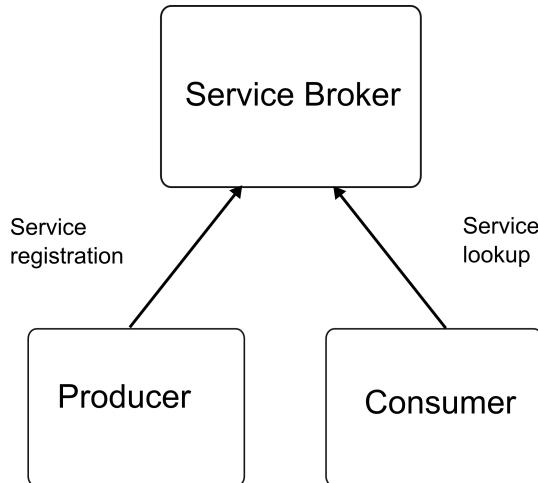


図 7.11: Relationship Between Producers and Consumers

Traditionally, in Eclipse 3.4.x applications, the consumer needed to know the location of the implementation, and to understand inheritance within the framework to consume services. The consumer code was therefore less reusable because people couldn't override which implementation the consumer receives. For example, if you wanted to update the message on the status line in Eclipse 3.x, the code would look like:

```
getViewSite().getActionBars().getStatusLineManager().setMessage(msg);
```

Eclipse 3.6 is built from components, but many of these components are too tightly coupled. To assemble applications of more loosely coupled components, Eclipse 4.0 uses dependency injection to provide services to clients. Dependency injection in Eclipse 4.x is through the use of a custom framework that uses the concept of a context that serves as a generic mechanism to locate services for consumers. The context exists between the application and the framework. Contexts are hierarchical. If a context has a request that cannot be satisfied, it will delegate the request to the parent context. The Eclipse context, called `IEclipseContext`, stores the available services and provides OSGi services lookup. Basically, the context is similar to a Java map in that it provides a mapping of a name or class to an object. The context handles model elements and services. Every element of the model, will have a context. Services are published in 4.x by means of the OSGi service mechanism.

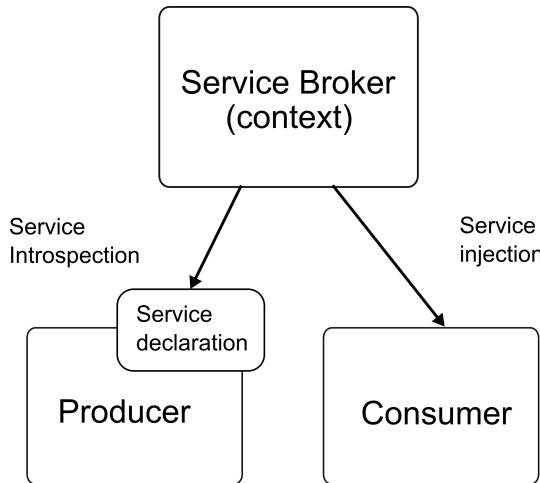


図 7.12: Service Broker Context

Producers add services and objects to the context which stores them. Services are injected into consumer objects by the context. The consumer declares what it wants, and the context determines how to satisfy this request. This approach has made consuming dynamic service easier. In Eclipse 3.x, a consumer had to attach listeners to be notified when services were available or unavailable.

With Eclipse 4.x, once a context has been injected into a consumer object, any change is automatically delivered to that object again. In other words, dependency injection occurs again. The consumer indicates that it will use the context by the use of Java 5 annotations which adhere to the JSR 330 standard, such as @inject, as well as some custom Eclipse annotations. Constructor, method, and field injection are supported. The 4.x runtime scans the objects for these annotations. The action that is performed depends on the annotation that's found.

This separation of concerns between context and application allows for better reuse of components, and absolves the consumer from understanding the implementation. In 4.x, the code to update the status line would look like this:

```
@Inject  
IStatusLineManager statusLine;  
...  
statusLine.setMessage(msg);
```

Application Services

One of the main goals in Eclipse 4.0 was to simplify the API for consumers so that it was easy to implement common services. The list of simple services came to be known as “the twenty things” and are known as the Eclipse Application services. The goal is to offer standalone APIs that clients can use without having to have a deep understanding of all the APIs available. They are structured as individual services so that they can also be used in other languages other than Java, such as Javascript. For example, there is an API to access the application model, to read and modify preferences and report errors and warnings.

7.5 Conclusion

The component-based architecture of Eclipse has evolved to incorporate new technology while maintaining backward compatibility. This has been costly, but the reward is the growth of the Eclipse community because of the trust established that consumers can continue to ship products based on a stable API.

Eclipse has so many consumers with diverse use cases and our expansive API became difficult for new consumers to adopt and understand. In retrospect, we should have kept our API simpler. If 80% of consumers only use 20% of the API, there is a need for simplification which was one of the reasons that the Eclipse 4.x stream was created.

The wisdom of crowds does reveal interesting use cases, such as disaggregating the IDE into bundles that could be used to construct RCP applications. Conversely, crowds often generate a lot of noise with requests for edge case scenarios that take a significant amount of time to implement.

In the early days of the Eclipse project, committers had the luxury of dedicating significant amounts of time to documentation, examples and answering community questions. Over time, this responsibility has shifted to the Eclipse community as a whole. We could have been better at providing documentation and use cases to help out the community, but this has been difficult given the large number of items planned for every release. Contrary to the expectation that software release dates slip, at Eclipse we consistently deliver our releases on time which allows our consumers to trust that they will be able to do the same.

By adopting new technology, and reinventing how Eclipse looks and works, we continue the conversation with our consumers and keep them engaged in the community. If you're interested in becoming involved with Eclipse, please visit <http://www.eclipse.org>.

Graphite

Chris Davis

Graphite¹ performs two pretty simple tasks: storing numbers that change over time and graphing them. There has been a lot of software written over the years to do these same tasks. What makes Graphite unique is that it provides this functionality as a network service that is both easy to use and highly scalable. The protocol for feeding data into Graphite is simple enough that you could learn to do it by hand in a few minutes (not that you'd actually want to, but it's a decent litmus test for simplicity). Rendering graphs and retrieving data points are as easy as fetching a URL. This makes it very natural to integrate Graphite with other software and enables users to build powerful applications on top of Graphite. One of the most common uses of Graphite is building web-based dashboards for monitoring and analysis. Graphite was born in a high-volume e-commerce environment and its design reflects this. Scalability and real-time access to data are key goals.

The components that allow Graphite to achieve these goals include a specialized database library and its storage format, a caching mechanism for optimizing I/O operations, and a simple yet effective method of clustering Graphite servers. Rather than simply describing how Graphite works today, I will explain how Graphite was initially implemented (quite naively), what problems I ran into, and how I devised solutions to them.

8.1 The Database Library: Storing Time-Series Data

Graphite is written entirely in Python and consists of three major components: a database library named `whisper`, a back-end daemon named `carbon`, and a front-end `webapp` that renders graphs and provides a basic UI. While `whisper` was written specifically for Graphite, it can also be used independently. It is very similar in design to the round-robin-database used by `RRDtool`, and only stores time-series numeric data. Usually we think of databases as server processes that client applications talk to over sockets. However, `whisper`, much like `RRDtool`, is a database library used

¹<http://launchpad.net/graphite>

by applications to manipulate and retrieve data stored in specially formatted files. The most basic whisper operations are create to make a new whisper file, update to write new data points into a file, and fetch to retrieve data points.

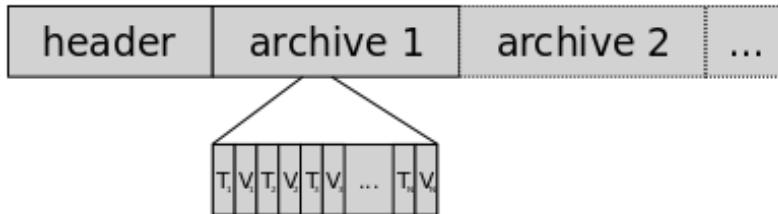


図 8.1: Basic Anatomy of a whisper File

As shown in 図 8.1, whisper files consist of a header section containing various metadata, followed by one or more archive sections . Each archive is a sequence of consecutive data points which are (timestamp, value) pairs. When an update or fetch operation is performed, whisper determines the offset in the file where data should be written to or read from, based on the timestamp and the archive configuration.

8.2 The Back End: A Simple Storage Service

Graphite's back end is a daemon process called carbon-cache, usually simply referred to as carbon. It is built on Twisted, a highly scalable event-driven I/O framework for Python. Twisted enables carbon to efficiently talk to a large number of clients and handle a large amount of traffic with low overhead. 図 8.2 shows the data flow among carbon, whisper and the webapp: Client applications collect data and send it to the Graphite back end, carbon, which stores the data using whisper. This data can then be used by the Graphite webapp to generate graphs.

The primary function of carbon is to store data points for metrics provided by clients. In Graphite terminology, a metric is any measurable quantity that can vary over time (like the CPU utilization of a server or the number of sales of a product). A data point is simply a (timestamp, value) pair corresponding to the measured value of a particular metric at a point in time. Metrics are uniquely identified by their name, and the name of each metric as well as its data points are provided by client applications. A common type of client application is a monitoring agent that collects system or application metrics, and sends its collected values to carbon for easy storage and visualization. Metrics in Graphite have simple hierarchical names, similar to filesystem paths except that a dot is used to delimit the hierarchy rather than a slash or backslash. carbon will respect any legal name and creates a whisper file for each metric to store its data points. The whisper files are stored within carbon's data directory in a filesystem hierarchy that mirrors the

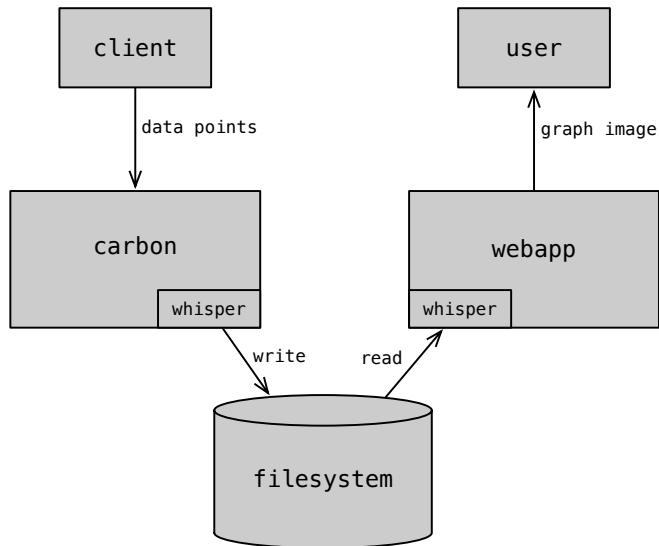


図 8.2: Data Flow

dot-delimited hierarchy in each metric's name, so that (for example) `servers.www01.cpuUsage` maps to `.../servers/www01/cpuUsage.wsp`.

When a client application wishes to send data points to Graphite it must establish a TCP connection to carbon, usually on port 2003². The client does all the talking; carbon does not send anything over the connection. The client sends data points in a simple plain-text format while the connection may be left open and re-used as needed. The format is one line of text per data point where each line contains the dotted metric name, value, and a Unix epoch timestamp separated by spaces. For example, a client might send:

```

servers.www01.cpuUsage 42 1286269200
products.snake-oil.salesPerMinute 123 1286269200
[one minute passes]
servers.www01.cpuUsageUser 44 1286269260
products.snake-oil.salesPerMinute 119 1286269260

```

On a high level, all carbon does is listen for data in this format and try to store it on disk as quickly as possible using whisper. Later on we will discuss the details of some tricks used to ensure scalability and get the best performance we can out of a typical hard drive.

²There is another port over which serialized objects can be sent, which is more efficient than the plain-text format. This is only needed for very high levels of traffic.

8.3 The Front End: Graphs On-Demand

The Graphite webapp allows users to request custom graphs with a simple URL-based API. Graphing parameters are specified in the query-string of an HTTP GET request, and a PNG image is returned in response. For example, the URL:

```
http://graphite.example.com/render?target=servers.www01.cpuUsage&width=500&height=300&from=-24h
```

requests a 500×300 graph for the metric `servers.www01.cpuUsage` and the past 24 hours of data. Actually, only the target parameter is required; all the others are optional and use your default values if omitted.

Graphite supports a wide variety of display options as well as data manipulation functions that follow a simple functional syntax. For example, we could graph a 10-point moving average of the metric in our previous example like this:

```
target=movingAverage(servers.www01.cpuUsage,10)
```

Functions can be nested, allowing for complex expressions and calculations.

Here is another example that gives the running total of sales for the day using per-product metrics of sales-per-minute:

```
target=integral(sumSeries(products.*.salesPerMinute))\&from=midnight
```

The `sumSeries` function computes a time-series that is the sum of each metric matching the pattern `products.*.salesPerMinute`. Then `integral` computes a running total rather than a per-minute count. From here it isn't too hard to imagine how one might build a web UI for viewing and manipulating graphs. Graphite comes with its own Composer UI, shown in 図 8.3, that does this using Javascript to modify the graph's URL parameters as the user clicks through menus of the available features.

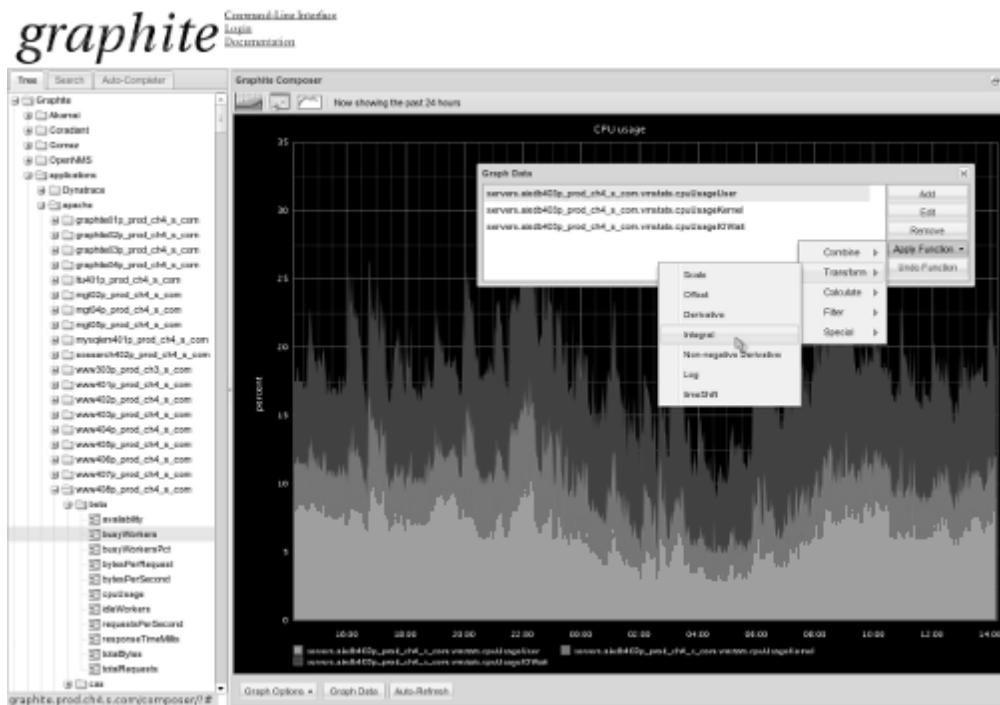


図 8.3: Graphite's Composer Interface

8.4 Dashboards

Since its inception Graphite has been used as a tool for creating web-based dashboards. The URL API makes this a natural use case. Making a dashboard is as simple as making an HTML page full of tags like this:

```

```

However, not everyone likes crafting URLs by hand, so Graphite's Composer UI provides a point-and-click method to create a graph from which you can simply copy and paste the URL. When coupled with another tool that allows rapid creation of web pages (like a wiki) this becomes easy enough that non-technical users can build their own dashboards pretty easily.

8.5 An Obvious Bottleneck

Once my users started building dashboards, Graphite quickly began to have performance issues. I investigated the web server logs to see what requests were bogging it down. It was pretty obvious that the problem was the sheer number of graphing requests. The webapp was CPU-bound, rendering graphs constantly. I noticed that there were a lot of identical requests, and the dashboards were to blame.

Imagine you have a dashboard with 10 graphs in it and the page refreshes once a minute. Each time a user opens the dashboard in their browser, Graphite has to handle 10 more requests per minute. This quickly becomes expensive.

A simple solution is to render each graph only once and then serve a copy of it to each user. The Django web framework (which Graphite is built on) provides an excellent caching mechanism that can use various back ends such as memcached. Memcached³ is essentially a hash table provided as a network service. Client applications can get and set key-value pairs just like an ordinary hash table. The main benefit of using memcached is that the result of an expensive request (like rendering a graph) can be stored very quickly and retrieved later to handle subsequent requests. To avoid returning the same stale graphs forever, memcached can be configured to expire the cached graphs after a short period. Even if this is only a few seconds, the burden it takes off Graphite is tremendous because duplicate requests are so common.

Another common case that creates lots of rendering requests is when a user is tweaking the display options and applying functions in the Composer UI. Each time the user changes something, Graphite must redraw the graph. The same data is involved in each request so it makes sense to put the underlying data in the memcache as well. This keeps the UI responsive to the user because the step of retrieving data is skipped.

8.6 Optimizing I/O

Imagine that you have 60,000 metrics that you send to your Graphite server, and each of these metrics has one data point per minute. Remember that each metric has its own whisper file on the

³<http://memcached.org>

filesystem. This means carbon must do one write operation to 60,000 different files each minute. As long as carbon can write to one file each millisecond, it should be able to keep up. This isn't too far fetched, but let's say you have 600,000 metrics updating each minute, or your metrics are updating every second, or perhaps you simply cannot afford fast enough storage. Whatever the case, assume the rate of incoming data points exceeds the rate of write operations that your storage can keep up with. How should this situation be handled?

Most hard drives these days have slow seek time⁴, that is, the delay between doing I/O operations at two different locations, compared to writing a contiguous sequence of data. This means the more contiguous writing we do, the more throughput we get. But if we have thousands of files that need to be written to frequently, and each write is very small (one whisper data point is only 12 bytes) then our disks are definitely going to spend most of their time seeking.

Working under the assumption that the rate of write operations has a relatively low ceiling, the only way to increase our data point throughput beyond that rate is to write multiple data points in a single write operation. This is feasible because `whisper` arranges consecutive data points contiguously on disk. So I added an `update_many` function to `whisper`, which takes a list of data points for a single metric and compacts contiguous data points into a single write operation. Even though this made each write larger, the difference in time it takes to write ten data points (120 bytes) versus one data point (12 bytes) is negligible. It takes quite a few more data points before the size of each write starts to noticeably affect the latency.

Next I implemented a buffering mechanism in carbon. Each incoming data point gets mapped to a queue based on its metric name and is then appended to that queue. Another thread repeatedly iterates through all of the queues and for each one it pulls all of the data points out and writes them to the appropriate `whisper` file with `update_many`. Going back to our example, if we have 600,000 metrics updating every minute and our storage can only keep up with 1 write per millisecond, then the queues will end up holding about 10 data points each on average. The only resource this costs us is memory, which is relatively plentiful since each data point is only a few bytes.

This strategy dynamically buffers as many datapoints as necessary to sustain a rate of incoming datapoints that may exceed the rate of I/O operations your storage can keep up with. A nice advantage of this approach is that it adds a degree of resiliency to handle temporary I/O slowdowns. If the system needs to do other I/O work outside of Graphite then it is likely that the rate of write operations will decrease, in which case carbon's queues will simply grow. The larger the queues, the larger the writes. Since the overall throughput of data points is equal to the rate of write operations times the average size of each write, carbon is able to keep up as long as there is enough memory for the queues. carbon's queueing mechanism is depicted in 図 8.4.

⁴Solid-state drives generally have extremely fast seek times compared to conventional hard drives.

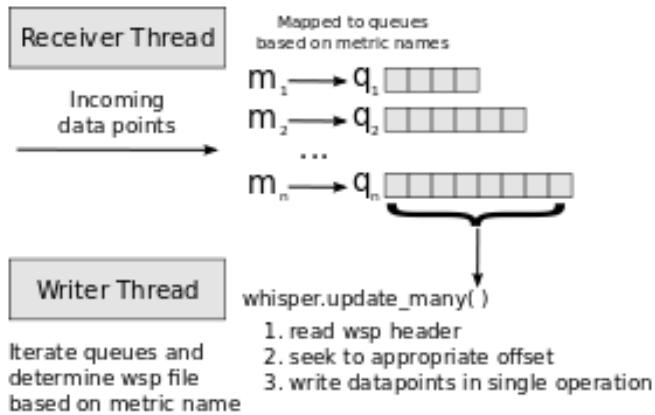


図 8.4: Carbon’s Queueing Mechanism

8.7 Keeping It Real-Time

Buffering data points was a nice way to optimize carbon’s I/O but it didn’t take long for my users to notice a rather troubling side effect. Revisiting our example again, we’ve got 600,000 metrics that update every minute and we’re assuming our storage can only keep up with 60,000 write operations per minute. This means we will have approximately 10 minutes worth of data sitting in carbon’s queues at any given time. To a user this means that the graphs they request from the Graphite webapp will be missing the most recent 10 minutes of data: Not good!

Fortunately the solution is pretty straight-forward. I simply added a socket listener to carbon that provides a query interface for accessing the buffered data points and then modifies the Graphite webapp to use this interface each time it needs to retrieve data. The webapp then combines the data points it retrieves from carbon with the data points it retrieved from disk and voila, the graphs are real-time. Granted, in our example the data points are updated to the minute and thus not exactly “real-time”, but the fact that each data point is instantly accessible in a graph once it is received by carbon is real-time.

8.8 Kernels, Caches, and Catastrophic Failures

As is probably obvious by now, a key characteristic of system performance that Graphite’s own performance depends on is I/O latency. So far we’ve assumed our system has consistently low I/O latency averaging around 1 millisecond per write, but this is a big assumption that requires a little deeper analysis. Most hard drives simply aren’t that fast; even with dozens of disks in a RAID array there is very likely to be more than 1 millisecond latency for random access. Yet if you were to try

and test how quickly even an old laptop could write a whole kilobyte to disk you would find that the write system call returns in far less than 1 millisecond. Why?

Whenever software has inconsistent or unexpected performance characteristics, usually either buffering or caching is to blame. In this case, we're dealing with both. The write system call doesn't technically write your data to disk, it simply puts it in a buffer which the kernel then writes to disk later on. This is why the write call usually returns so quickly. Even after the buffer has been written to disk, it often remains cached for subsequent reads. Both of these behaviors, buffering and caching, require memory of course.

Kernel developers, being the smart folks that they are, decided it would be a good idea to use whatever user-space memory is currently free instead of allocating memory outright. This turns out to be a tremendously useful performance booster and it also explains why no matter how much memory you add to a system it will usually end up having almost zero "free" memory after doing a modest amount of I/O. If your user-space applications aren't using that memory then your kernel probably is. The downside of this approach is that this "free" memory can be taken away from the kernel the moment a user-space application decides it needs to allocate more memory for itself. The kernel has no choice but to relinquish it, losing whatever buffers may have been there.

So what does all of this mean for Graphite? We just highlighted carbon's reliance on consistently low I/O latency and we also know that the write system call only returns quickly because the data is merely being copied into a buffer. What happens when there is not enough memory for the kernel to continue buffering writes? The writes become synchronous and thus terribly slow! This causes a dramatic drop in the rate of carbon's write operations, which causes carbon's queues to grow, which eats up even more memory, starving the kernel even further. In the end, this kind of situation usually results in carbon running out of memory or being killed by an angry sysadmin.

To avoid this kind of catastrophe, I added several features to carbon including configurable limits on how many data points can be queued and rate-limits on how quickly various whisper operations can be performed. These features can protect carbon from spiraling out of control and instead impose less harsh effects like dropping some data points or refusing to accept more data points. However, proper values for those settings are system-specific and require a fair amount of testing to tune. They are useful but they do not fundamentally solve the problem. For that, we'll need more hardware.

8.9 Clustering

Making multiple Graphite servers appear to be a single system from a user perspective isn't terribly difficult, at least for a naïve implementation. The webapp's user interaction primarily consists of two operations: finding metrics and fetching data points (usually in the form of a graph). The find and fetch operations of the webapp are tucked away in a library that abstracts their implementation

from the rest of the codebase, and they are also exposed through HTTP request handlers for easy remote calls.

The `find` operation searches the local filesystem of whisper data for things matching a user-specified pattern, just as a filesystem glob like `*.txt` matches files with that extension. Being a tree structure, the result returned by `find` is a collection of `Node` objects, each deriving from either the `Branch` or `Leaf` sub-classes of `Node`. Directories correspond to branch nodes and whisper files correspond to leaf nodes. This layer of abstraction makes it easy to support different types of underlying storage including RRD files⁵ and gzipped whisper files.

The `Leaf` interface defines a `fetch` method whose implementation depends on the type of leaf node. In the case of whisper files it is simply a thin wrapper around the whisper library's own `fetch` function. When clustering support was added, the `find` function was extended to be able to make remote `find` calls via HTTP to other Graphite servers specified in the webapp's configuration. The node data contained in the results of these HTTP calls gets wrapped as `RemoteNode` objects which conform to the usual `Node`, `Branch`, and `Leaf` interfaces. This makes the clustering transparent to the rest of the webapp's codebase. The `fetch` method for a remote leaf node is implemented as another HTTP call to retrieve the data points from the node's Graphite server.

All of these calls are made between the webapps the same way a client would call them, except with one additional parameter specifying that the operation should only be performed locally and not be redistributed throughout the cluster. When the webapp is asked to render a graph, it performs the `find` operation to locate the requested metrics and calls `fetch` on each to retrieve their data points. This works whether the data is on the local server, remote servers, or both. If a server goes down, the remote calls timeout fairly quickly and the server is marked as being out of service for a short period during which no further calls to it will be made. From a user standpoint, whatever data was on the lost server will be missing from their graphs unless that data is duplicated on another server in the cluster.

A Brief Analysis of Clustering Efficiency

The most expensive part of a graphing request is rendering the graph. Each rendering is performed by a single server so adding more servers does effectively increase capacity for rendering graphs. However, the fact that many requests end up distributing `find` calls to every other server in the cluster means that our clustering scheme is sharing much of the front-end load rather than dispersing it. What we have achieved at this point, however, is an effective way to distribute back-end load, as each carbon instance operates independently. This is a good first step since most of the time the back end is a bottleneck far before the front end is, but clearly the front end will not scale horizontally with this approach.

⁵RRD files are actually branch nodes because they can contain multiple data sources; an RRD data source is a leaf node.

In order to make the front end scale more effectively, the number of remote `find` calls made by the webapp must be reduced. Again, the easiest solution is caching. Just as memcached is already used to cache data points and rendered graphs, it can also be used to cache the results of `find` requests. Since the location of metrics is much less likely to change frequently, this should typically be cached for longer. The trade-off of setting the cache timeout for `find` results too long, though, is that new metrics that have been added to the hierarchy may not appear as quickly to the user.

Distributing Metrics in a Cluster

The Graphite webapp is rather homogeneous throughout a cluster, in that it performs the exact same job on each server. `carbon`'s role, however, can vary from server to server depending on what data you choose to send to each instance. Often there are many different clients sending data to `carbon`, so it would be quite annoying to couple each client's configuration with your Graphite cluster's layout. Application metrics may go to one `carbon` server, while business metrics may get sent to multiple `carbon` servers for redundancy.

To simplify the management of scenarios like this, Graphite comes with an additional tool called `carbon-relay`. Its job is quite simple; it receives metric data from clients exactly like the standard `carbon` daemon (which is actually named `carbon-cache`) but instead of storing the data, it applies a set of rules to the metric names to determine which `carbon-cache` servers to relay the data to. Each rule consists of a regular expression and a list of destination servers. For each data point received, the rules are evaluated in order and the first rule whose regular expression matches the metric name is used. This way all the clients need to do is send their data to the `carbon-relay` and it will end up on the right servers.

In a sense `carbon-relay` provides replication functionality, though it would more accurately be called input duplication since it does not deal with synchronization issues. If a server goes down temporarily, it will be missing the data points for the time period in which it was down but otherwise function normally. There are administrative scripts that leave control of the re-synchronization process in the hands of the system administrator.

8.10 Design Reflections

My experience in working on Graphite has reaffirmed a belief of mine that scalability has very little to do with low-level performance but instead is a product of overall design. I have run into many bottlenecks along the way but each time I look for improvements in design rather than speed-ups in performance. I have been asked many times why I wrote Graphite in Python rather than Java or C++, and my response is always that I have yet to come across a true need for the performance that another language could offer. In [Knu74], Donald Knuth famously said that premature optimization

is the root of all evil. As long as we assume that our code will continue to evolve in non-trivial ways then all optimization⁶ is in some sense premature.

One of Graphite’s greatest strengths and greatest weaknesses is the fact that very little of it was actually “designed” in the traditional sense. By and large Graphite evolved gradually, hurdle by hurdle, as problems arose. Many times the hurdles were foreseeable and various pre-emptive solutions seemed natural. However it can be useful to avoid solving problems you do not actually have yet, even if it seems likely that you soon will. The reason is that you can learn much more from closely studying actual failures than from theorizing about superior strategies. Problem solving is driven by both the empirical data we have at hand and our own knowledge and intuition. I’ve found that doubting your own wisdom sufficiently can force you to look at your empirical data more thoroughly.

For example, when I first wrote `whisper` I was convinced that it would have to be rewritten in C for speed and that my Python implementation would only serve as a prototype. If I weren’t under a time-crunch I very well may have skipped the Python implementation entirely. It turns out however that I/O is a bottleneck so much earlier than CPU that the lesser efficiency of Python hardly matters at all in practice.

As I said, though, the evolutionary approach is also a great weakness of Graphite. Interfaces, it turns out, do not lend themselves well to gradual evolution. A good interface is consistent and employs conventions to maximize predictability. By this measure, Graphite’s URL API is currently a sub-par interface in my opinion. Options and functions have been tacked on over time, sometimes forming small islands of consistency, but overall lacking a global sense of consistency. The only way to solve such a problem is through versioning of interfaces, but this too has drawbacks. Once a new interface is designed, the old one is still hard to get rid of, lingering around as evolutionary baggage like the human appendix. It may seem harmless enough until one day your code gets appendicitis (i.e. a bug tied to the old interface) and you’re forced to operate. If I were to change one thing about Graphite early on, it would have been to take much greater care in designing the external APIs, thinking ahead instead of evolving them bit by bit.

Another aspect of Graphite that causes some frustration is the limited flexibility of the hierarchical metric naming model. While it is quite simple and very convenient for most use cases, it makes some sophisticated queries very difficult, even impossible, to express. When I first thought of creating Graphite I knew from the very beginning that I wanted a human-editable URL API for creating graphs⁷. While I’m still glad that Graphite provides this today, I’m afraid this requirement has burdened the API with excessively simple syntax that makes complex expressions unwieldy. A hierarchy makes the problem of determining the “primary key” for a metric quite simple because a path is essentially a primary key for a node in the tree. The downside is that all of the descriptive data (i.e. column data) must be embedded directly in the path. A potential solution is to maintain

⁶Knuth specifically meant low-level code optimization, not macroscopic optimization such as design improvements.

⁷This forces the graphs themselves to be open source. Anyone can simply look at a graph’s URL to understand it or modify it.

the hierarchical model and add a separate metadata database to enable more advanced selection of metrics with a special syntax.

8.11 Becoming Open Source

Looking back at the evolution of Graphite, I am still surprised both by how far it has come as a project and by how far it has taken me as a programmer. It started as a pet project that was only a few hundred lines of code. The rendering engine started as an experiment, simply to see if I could write one. `whisper` was written over the course of a weekend out of desperation to solve a showstopper problem before a critical launch date. `carbon` has been rewritten more times than I care to remember. Once I was allowed to release Graphite under an open source license in 2008 I never really expected much response. After a few months it was mentioned in a CNET article that got picked up by Slashdot and the project suddenly took off and has been active ever since. Today there are dozens of large and mid-sized companies using Graphite. The community is quite active and continues to grow. Far from being a finished product, there is a lot of cool experimental work being done, which keeps it fun to work on and full of potential.

The Hadoop Distributed File System

Robert Chansler, Hairong Kuang, Sanjay Radia,
Konstantin Shvachko, and Suresh Srinivas

The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size. We describe the architecture of HDFS and report on experience using HDFS to manage 40 petabytes of enterprise data at Yahoo!

9.1 Introduction

Hadoop¹ provides a distributed filesystem and a framework for the analysis and transformation of very large data sets using the MapReduce [DG04] paradigm. While the interface to HDFS is patterned after the Unix filesystem, faithfulness to standards was sacrificed in favor of improved performance for the applications at hand.

An important characteristic of Hadoop is the partitioning of data and computation across many (thousands) of hosts, and the execution of application computations in parallel close to their data. A Hadoop cluster scales computation capacity, storage capacity and I/O bandwidth by simply adding commodity servers. Hadoop clusters at Yahoo! span 40,000 servers, and store 40 petabytes of application data, with the largest cluster being 4000 servers. One hundred other organizations worldwide report using Hadoop.

HDFS stores filesystem metadata and application data separately. As in other distributed filesystems, like PVFS [CIRT00], Lustre², and GFS [GGL03, MQ09], HDFS stores metadata on a dedicated server, called the NameNode. Application data are stored on other servers called DataNodes.

¹<http://hadoop.apache.org>

²<http://www.lustre.org>

All servers are fully connected and communicate with each other using TCP-based protocols. Unlike Lustre and PVFS, the DataNodes in HDFS do not rely on data protection mechanisms such as RAID to make the data durable. Instead, like GFS, the file content is replicated on multiple DataNodes for reliability. While ensuring data durability, this strategy has the added advantage that data transfer bandwidth is multiplied, and there are more opportunities for locating computation near the needed data.

9.2 Architecture

NameNode

The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes. Inodes record attributes like permissions, modification and access times, namespace and disk space quotas. The file content is split into large blocks (typically 128 megabytes, but user selectable file-by-file), and each block of the file is independently replicated at multiple DataNodes (typically three, but user selectable file-by-file). The NameNode maintains the namespace tree and the mapping of blocks to DataNodes. The current design has a single NameNode for each cluster. The cluster can have thousands of DataNodes and tens of thousands of HDFS clients per cluster, as each DataNode may execute multiple application tasks concurrently.

Image and Journal

The inodes and the list of blocks that define the metadata of the name system are called the *image*. NameNode keeps the entire namespace image in RAM. The persistent record of the image stored in the NameNode's local native filesystem is called a checkpoint. The NameNode records changes to HDFS in a write-ahead log called the journal in its local native filesystem. The location of block replicas are not part of the persistent checkpoint.

Each client-initiated transaction is recorded in the journal, and the journal file is flushed and synced before the acknowledgment is sent to the client. The checkpoint file is never changed by the NameNode; a new file is written when a checkpoint is created during restart, when requested by the administrator, or by the CheckpointNode described in the next section. During startup the NameNode initializes the namespace image from the checkpoint, and then replays changes from the journal. A new checkpoint and an empty journal are written back to the storage directories before the NameNode starts serving clients.

For improved durability, redundant copies of the checkpoint and journal are typically stored on multiple independent local volumes and at remote NFS servers. The first choice prevents loss from a single volume failure, and the second choice protects against failure of the entire node. If the NameNode encounters an error writing the journal to one of the storage directories it automatically

excludes that directory from the list of storage directories. The NameNode automatically shuts itself down if no storage directory is available.

The NameNode is a multithreaded system and processes requests simultaneously from multiple clients. Saving a transaction to disk becomes a bottleneck since all other threads need to wait until the synchronous flush-and-sync procedure initiated by one of them is complete. In order to optimize this process, the NameNode batches multiple transactions. When one of the NameNode's threads initiates a flush-and-sync operation, all the transactions batched at that time are committed together. Remaining threads only need to check that their transactions have been saved and do not need to initiate a flush-and-sync operation.

DataNodes

Each block replica on a DataNode is represented by two files in the local native filesystem. The first file contains the data itself and the second file records the block's metadata including checksums for the data and the generation stamp. The size of the data file equals the actual length of the block and does not require extra space to round it up to the nominal block size as in traditional filesystems. Thus, if a block is half full it needs only half of the space of the full block on the local drive.

During startup each DataNode connects to the NameNode and performs a handshake. The purpose of the handshake is to verify the namespace ID and the software version of the DataNode. If either does not match that of the NameNode, the DataNode automatically shuts down.

The namespace ID is assigned to the filesystem instance when it is formatted. The namespace ID is persistently stored on all nodes of the cluster. Nodes with a different namespace ID will not be able to join the cluster, thus protecting the integrity of the filesystem. A DataNode that is newly initialized and without any namespace ID is permitted to join the cluster and receive the cluster's namespace ID.

After the handshake the DataNode registers with the NameNode. DataNodes persistently store their unique storage IDs. The storage ID is an internal identifier of the DataNode, which makes it recognizable even if it is restarted with a different IP address or port. The storage ID is assigned to the DataNode when it registers with the NameNode for the first time and never changes after that.

A DataNode identifies block replicas in its possession to the NameNode by sending a block report. A block report contains the block ID, the generation stamp and the length for each block replica the server hosts. The first block report is sent immediately after the DataNode registration. Subsequent block reports are sent every hour and provide the NameNode with an up-to-date view of where block replicas are located on the cluster.

During normal operation DataNodes send heartbeats to the NameNode to confirm that the DataNode is operating and the block replicas it hosts are available. The default heartbeat interval is three seconds. If the NameNode does not receive a heartbeat from a DataNode in ten minutes the NameNode considers the DataNode to be out of service and the block replicas hosted by that DataNode

to be unavailable. The NameNode then schedules creation of new replicas of those blocks on other DataNodes.

Heartbeats from a DataNode also carry information about total storage capacity, fraction of storage in use, and the number of data transfers currently in progress. These statistics are used for the NameNode's block allocation and load balancing decisions.

The NameNode does not directly send requests to DataNodes. It uses replies to heartbeats to send instructions to the DataNodes. The instructions include commands to replicate blocks to other nodes, remove local block replicas, re-register and send an immediate block report, and shut down the node.

These commands are important for maintaining the overall system integrity and therefore it is critical to keep heartbeats frequent even on big clusters. The NameNode can process thousands of heartbeats per second without affecting other NameNode operations.

HDFS Client

User applications access the filesystem using the HDFS client, a library that exports the HDFS filesystem interface.

Like most conventional filesystems, HDFS supports operations to read, write and delete files, and operations to create and delete directories. The user references files and directories by paths in the namespace. The user application does not need to know that filesystem metadata and storage are on different servers, or that blocks have multiple replicas.

When an application reads a file, the HDFS client first asks the NameNode for the list of DataNodes that host replicas of the blocks of the file. The list is sorted by the network topology distance from the client. The client contacts a DataNode directly and requests the transfer of the desired block. When a client writes, it first asks the NameNode to choose DataNodes to host replicas of the first block of the file. The client organizes a pipeline from node-to-node and sends the data. When the first block is filled, the client requests new DataNodes to be chosen to host replicas of the next block. A new pipeline is organized, and the client sends the further bytes of the file. Choice of DataNodes for each block is likely to be different. The interactions among the client, the NameNode and the DataNodes are illustrated in 図 9.1.

Unlike conventional filesystems, HDFS provides an API that exposes the locations of a file blocks. This allows applications like the MapReduce framework to schedule a task to where the data are located, thus improving the read performance. It also allows an application to set the replication factor of a file. By default a file's replication factor is three. For critical files or files which are accessed very often, having a higher replication factor improves tolerance against faults and increases read bandwidth.

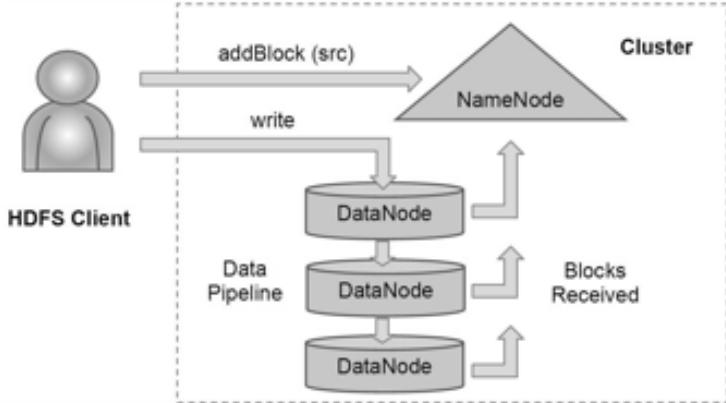


図 9.1: HDFS Client Creates a New File

CheckpointNode

The NameNode in HDFS, in addition to its primary role serving client requests, can alternatively execute either of two other roles, either a CheckpointNode or a BackupNode. The role is specified at the node startup.

The CheckpointNode periodically combines the existing checkpoint and journal to create a new checkpoint and an empty journal. The CheckpointNode usually runs on a different host from the NameNode since it has the same memory requirements as the NameNode. It downloads the current checkpoint and journal files from the NameNode, merges them locally, and returns the new checkpoint back to the NameNode.

Creating periodic checkpoints is one way to protect the filesystem metadata. The system can start from the most recent checkpoint if all other persistent copies of the namespace image or journal are unavailable. Creating a checkpoint also lets the NameNode truncate the journal when the new checkpoint is uploaded to the NameNode. HDFS clusters run for prolonged periods of time without restarts during which the journal constantly grows. If the journal grows very large, the probability of loss or corruption of the journal file increases. Also, a very large journal extends the time required to restart the NameNode. For a large cluster, it takes an hour to process a week-long journal. Good practice is to create a daily checkpoint.

BackupNode

A recently introduced feature of HDFS is the BackupNode. Like a CheckpointNode, the BackupNode is capable of creating periodic checkpoints, but in addition it maintains an in-memory, up-to-date image of the filesystem namespace that is always synchronized with the state of the NameNode.

The BackupNode accepts the journal stream of namespace transactions from the active NameNode, saves them in journal on its own storage directories, and applies these transactions to its own namespace image in memory. The NameNode treats the BackupNode as a journal store the same way as it treats journal files in its storage directories. If the NameNode fails, the BackupNode's image in memory and the checkpoint on disk is a record of the latest namespace state.

The BackupNode can create a checkpoint without downloading checkpoint and journal files from the active NameNode, since it already has an up-to-date namespace image in its memory. This makes the checkpoint process on the BackupNode more efficient as it only needs to save the namespace into its local storage directories.

The BackupNode can be viewed as a read-only NameNode. It contains all filesystem metadata information except for block locations. It can perform all operations of the regular NameNode that do not involve modification of the namespace or knowledge of block locations. Use of a BackupNode provides the option of running the NameNode without persistent storage, delegating responsibility of persisting the namespace state to the BackupNode.

Upgrades and Filesystem Snapshots

During software upgrades the possibility of corrupting the filesystem due to software bugs or human mistakes increases. The purpose of creating snapshots in HDFS is to minimize potential damage to the data stored in the system during upgrades.

The snapshot mechanism lets administrators persistently save the current state of the filesystem, so that if the upgrade results in data loss or corruption it is possible to rollback the upgrade and return HDFS to the namespace and storage state as they were at the time of the snapshot.

The snapshot (only one can exist) is created at the cluster administrator's option whenever the system is started. If a snapshot is requested, the NameNode first reads the checkpoint and journal files and merges them in memory. Then it writes the new checkpoint and the empty journal to a new location, so that the old checkpoint and journal remain unchanged.

During handshake the NameNode instructs DataNodes whether to create a local snapshot. The local snapshot on the DataNode cannot be created by replicating the directories containing the data files as this would require doubling the storage capacity of every DataNode on the cluster. Instead each DataNode creates a copy of the storage directory and hard links existing block files into it. When the DataNode removes a block it removes only the hard link, and block modifications during appends use the copy-on-write technique. Thus old block replicas remain untouched in their old directories.

The cluster administrator can choose to roll back HDFS to the snapshot state when restarting the system. The NameNode recovers the checkpoint saved when the snapshot was created. DataNodes restore the previously renamed directories and initiate a background process to delete block replicas created after the snapshot was made. Having chosen to roll back, there is no provision to roll forward.

The cluster administrator can recover the storage occupied by the snapshot by commanding the system to abandon the snapshot; for snapshots created during upgrade, this finalizes the software upgrade.

System evolution may lead to a change in the format of the NameNode's checkpoint and journal files, or in the data representation of block replica files on DataNodes. The layout version identifies the data representation formats, and is persistently stored in the NameNode's and the DataNodes' storage directories. During startup each node compares the layout version of the current software with the version stored in its storage directories and automatically converts data from older formats to the newer ones. The conversion requires the mandatory creation of a snapshot when the system restarts with the new software layout version.

9.3 File I/O Operations and Replica Management

Of course, the whole point of a filesystem is to store data in files. To understand how HDFS does this, we must look at how reading and writing works, and how blocks are managed.

File Read and Write

An application adds data to HDFS by creating a new file and writing the data to it. After the file is closed, the bytes written cannot be altered or removed except that new data can be added to the file by reopening the file for append. HDFS implements a single-writer, multiple-reader model.

The HDFS client that opens a file for writing is granted a lease for the file; no other client can write to the file. The writing client periodically renews the lease by sending a heartbeat to the NameNode. When the file is closed, the lease is revoked. The lease duration is bound by a soft limit and a hard limit. Until the soft limit expires, the writer is certain of exclusive access to the file. If the soft limit expires and the client fails to close the file or renew the lease, another client can preempt the lease. If after the hard limit expires (one hour) and the client has failed to renew the lease, HDFS assumes that the client has quit and will automatically close the file on behalf of the writer, and recover the lease. The writer's lease does not prevent other clients from reading the file; a file may have many concurrent readers.

An HDFS file consists of blocks. When there is a need for a new block, the NameNode allocates a block with a unique block ID and determines a list of DataNodes to host replicas of the block. The DataNodes form a pipeline, the order of which minimizes the total network distance from the client to the last DataNode. Bytes are pushed to the pipeline as a sequence of packets. The bytes that an application writes first buffer at the client side. After a packet buffer is filled (typically 64 KB), the data are pushed to the pipeline. The next packet can be pushed to the pipeline before receiving the acknowledgment for the previous packets. The number of outstanding packets is limited by the outstanding packets window size of the client.

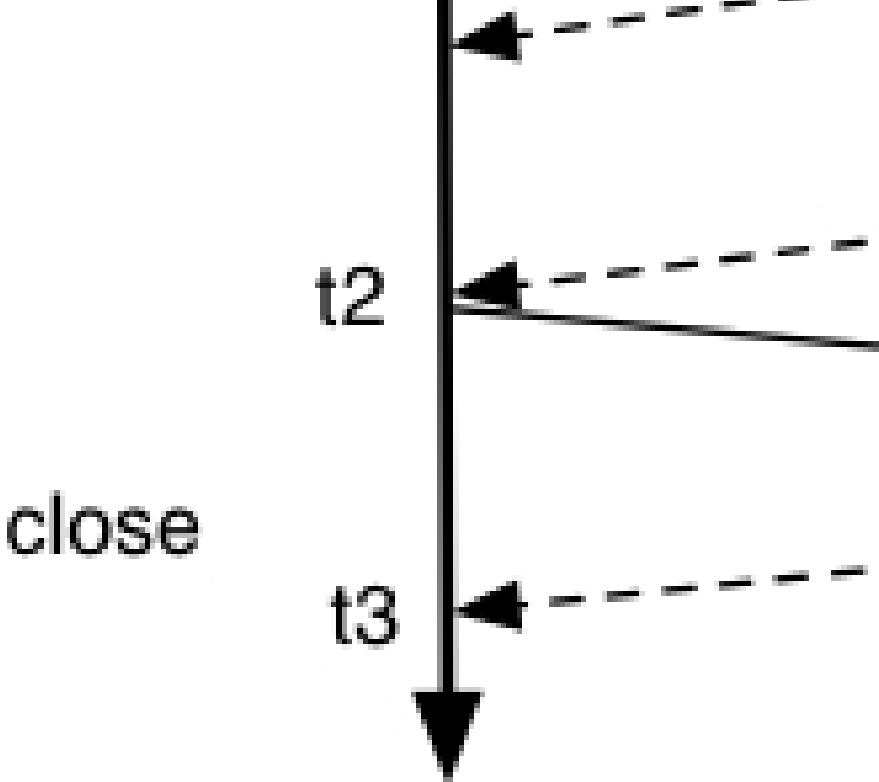


図 9.2: Data Pipeline While Writing a Block

After data are written to an HDFS file, HDFS does not provide any guarantee that data are visible to a new reader until the file is closed. If a user application needs the visibility guarantee, it can explicitly call the `hflush` operation. Then the current packet is immediately pushed to the pipeline, and the `hflush` operation will wait until all DataNodes in the pipeline acknowledge the successful transmission of the packet. All data written before the `hflush` operation are then certain to be visible to readers.

If no error occurs, block construction goes through three stages as shown in 図 9.2 illustrating a pipeline of three DataNodes (DN) and a block of five packets. In the picture, bold lines represent data packets, dashed lines represent acknowledgment messages, and thin lines represent control messages to setup and close the pipeline. Vertical lines represent activity at the client and the three DataNodes where time proceeds from top to bottom. From t_0 to t_1 is the pipeline setup stage. The interval t_1 to t_2 is the data streaming stage, where t_1 is the time when the first data packet gets sent and t_2 is the time that the acknowledgment to the last packet gets received. Here an `hflush` operation transmits packet 2. The `hflush` indication travels with the packet data and is not a separate operation. The final interval t_2 to t_3 is the pipeline close stage for this block.

In a cluster of thousands of nodes, failures of a node (most commonly storage faults) are daily occurrences. A replica stored on a DataNode may become corrupted because of faults in memory, disk, or network. HDFS generates and stores checksums for each data block of an HDFS file. Checksums are verified by the HDFS client while reading to help detect any corruption caused either by client, DataNodes, or network. When a client creates an HDFS file, it computes the checksum sequence for each block and sends it to a DataNode along with the data. A DataNode stores checksums in a metadata file separate from the block's data file. When HDFS reads a file, each block's data and checksums are shipped to the client. The client computes the checksum for the received data and verifies that the newly computed checksums matches the checksums it received. If not, the client notifies the NameNode of the corrupt replica and then fetches a different replica of the block from another DataNode.

When a client opens a file to read, it fetches the list of blocks and the locations of each block replica from the NameNode. The locations of each block are ordered by their distance from the reader. When reading the content of a block, the client tries the closest replica first. If the read attempt fails, the client tries the next replica in sequence. A read may fail if the target DataNode is unavailable, the node no longer hosts a replica of the block, or the replica is found to be corrupt when checksums are tested.

HDFS permits a client to read a file that is open for writing. When reading a file open for writing, the length of the last block still being written is unknown to the NameNode. In this case, the client asks one of the replicas for the latest length before starting to read its content.

The design of HDFS I/O is particularly optimized for batch processing systems, like MapReduce, which require high throughput for sequential reads and writes. Ongoing efforts will improve read/write response time for applications that require real-time data streaming or random access.

Block Placement

For a large cluster, it may not be practical to connect all nodes in a flat topology. A common practice is to spread the nodes across multiple racks. Nodes of a rack share a switch, and rack switches are connected by one or more core switches. Communication between two nodes in different racks has to go through multiple switches. In most cases, network bandwidth between nodes in the same rack is greater than network bandwidth between nodes in different racks.  9.3 describes a cluster with two racks, each of which contains three nodes.

HDFS estimates the network bandwidth between two nodes by their distance. The distance from a node to its parent node is assumed to be one. A distance between two nodes can be calculated by summing the distances to their closest common ancestor. A shorter distance between two nodes means greater bandwidth they can use to transfer data.

HDFS allows an administrator to configure a script that returns a node's rack identification given a node's address. The NameNode is the central place that resolves the rack location of each DataNode.

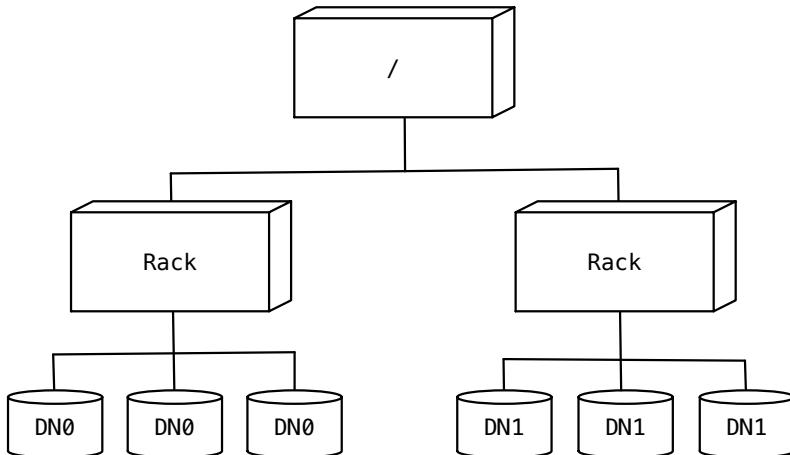


図 9.3: Cluster Topology

When a DataNode registers with the NameNode, the NameNode runs the configured script to decide which rack the node belongs to. If no such a script is configured, the NameNode assumes that all the nodes belong to a default single rack.

The placement of replicas is critical to HDFS data reliability and read/write performance. A good replica placement policy should improve data reliability, availability, and network bandwidth utilization. Currently HDFS provides a configurable block placement policy interface so that the users and researchers can experiment and test alternate policies that are optimal for their applications.

The default HDFS block placement policy provides a tradeoff between minimizing the write cost, and maximizing data reliability, availability and aggregate read bandwidth. When a new block is created, HDFS places the first replica on the node where the writer is located. The second and the third replicas are placed on two different nodes in a different rack. The rest are placed on random nodes with restrictions that no more than one replica is placed at any one node and no more than two replicas are placed in the same rack, if possible. The choice to place the second and third replicas on a different rack better distributes the block replicas for a single file across the cluster. If the first two replicas were placed on the same rack, for any file, two-thirds of its block replicas would be on the same rack.

After all target nodes are selected, nodes are organized as a pipeline in the order of their proximity to the first replica. Data are pushed to nodes in this order. For reading, the NameNode first checks if the client's host is located in the cluster. If yes, block locations are returned to the client in the order of its closeness to the reader. The block is read from DataNodes in this preference order.

This policy reduces the inter-rack and inter-node write traffic and generally improves write performance. Because the chance of a rack failure is far less than that of a node failure, this policy does not impact data reliability and availability guarantees. In the usual case of three replicas, it can

reduce the aggregate network bandwidth used when reading data since a block is placed in only two unique racks rather than three.

Replication Management

The NameNode endeavors to ensure that each block always has the intended number of replicas. The NameNode detects that a block has become under- or over-replicated when a block report from a DataNode arrives. When a block becomes over replicated, the NameNode chooses a replica to remove. The NameNode will prefer not to reduce the number of racks that host replicas, and secondly prefer to remove a replica from the DataNode with the least amount of available disk space. The goal is to balance storage utilization across DataNodes without reducing the block's availability.

When a block becomes under-replicated, it is put in the replication priority queue. A block with only one replica has the highest priority, while a block with a number of replicas that is greater than two thirds of its replication factor has the lowest priority. A background thread periodically scans the head of the replication queue to decide where to place new replicas. Block replication follows a similar policy as that of new block placement. If the number of existing replicas is one, HDFS places the next replica on a different rack. In case that the block has two existing replicas, if the two existing replicas are on the same rack, the third replica is placed on a different rack; otherwise, the third replica is placed on a different node in the same rack as an existing replica. Here the goal is to reduce the cost of creating new replicas.

The NameNode also makes sure that not all replicas of a block are located on one rack. If the NameNode detects that a block's replicas end up at one rack, the NameNode treats the block as mis-replicated and replicates the block to a different rack using the same block placement policy described above. After the NameNode receives the notification that the replica is created, the block becomes over-replicated. The NameNode then will decides to remove an old replica because the over-replication policy prefers not to reduce the number of racks.

Balancer

HDFS block placement strategy does not take into account DataNode disk space utilization. This is to avoid placing new—more likely to be referenced—data at a small subset of the DataNodes with a lot of free storage. Therefore data might not always be placed uniformly across DataNodes. Imbalance also occurs when new nodes are added to the cluster.

The balancer is a tool that balances disk space usage on an HDFS cluster. It takes a threshold value as an input parameter, which is a fraction between 0 and 1. A cluster is balanced if, for each

DataNode, the utilization of the node³ differs from the utilization of the whole cluster⁴ by no more than the threshold value.

The tool is deployed as an application program that can be run by the cluster administrator. It iteratively moves replicas from DataNodes with higher utilization to DataNodes with lower utilization. One key requirement for the balancer is to maintain data availability. When choosing a replica to move and deciding its destination, the balancer guarantees that the decision does not reduce either the number of replicas or the number of racks.

The balancer optimizes the balancing process by minimizing the inter-rack data copying. If the balancer decides that a replica A needs to be moved to a different rack and the destination rack happens to have a replica B of the same block, the data will be copied from replica B instead of replica A.

A configuration parameter limits the bandwidth consumed by rebalancing operations. The higher the allowed bandwidth, the faster a cluster can reach the balanced state, but with greater competition with application processes.

Block Scanner

Each DataNode runs a block scanner that periodically scans its block replicas and verifies that stored checksums match the block data. In each scan period, the block scanner adjusts the read bandwidth in order to complete the verification in a configurable period. If a client reads a complete block and checksum verification succeeds, it informs the DataNode. The DataNode treats it as a verification of the replica.

The verification time of each block is stored in a human-readable log file. At any time there are up to two files in the top-level DataNode directory, the current and previous logs. New verification times are appended to the current file. Correspondingly, each DataNode has an in-memory scanning list ordered by the replica's verification time.

Whenever a read client or a block scanner detects a corrupt block, it notifies the NameNode. The NameNode marks the replica as corrupt, but does not schedule deletion of the replica immediately. Instead, it starts to replicate a good copy of the block. Only when the good replica count reaches the replication factor of the block the corrupt replica is scheduled to be removed. This policy aims to preserve data as long as possible. So even if all replicas of a block are corrupt, the policy allows the user to retrieve its data from the corrupt replicas.

³Defined as the ratio of used space at the node to total capacity of the node.

⁴Defined as the ratio of used space in the cluster to total capacity of the cluster.

Decommissioning

The cluster administrator specifies list of nodes to be decommissioned. Once a DataNode is marked for decommissioning, it will not be selected as the target of replica placement, but it will continue to serve read requests. The NameNode starts to schedule replication of its blocks to other DataNodes. Once the NameNode detects that all blocks on the decommissioning DataNode are replicated, the node enters the decommissioned state. Then it can be safely removed from the cluster without jeopardizing any data availability.

Inter-Cluster Data Copy

When working with large datasets, copying data into and out of a HDFS cluster is daunting. HDFS provides a tool called DistCp for large inter/intra-cluster parallel copying. It is a MapReduce job; each of the map tasks copies a portion of the source data into the destination filesystem. The MapReduce framework automatically handles parallel task scheduling, error detection and recovery.

9.4 Practice at Yahoo!

Large HDFS clusters at Yahoo! include about 4000 nodes. A typical cluster node has two quad core Xeon processors running at 2.5 GHz, 4–12 directly attached SATA drives (holding two terabytes each), 24 Gbyte of RAM, and a 1-gigabit Ethernet connection. Seventy percent of the disk space is allocated to HDFS. The remainder is reserved for the operating system (Red Hat Linux), logs, and space to spill the output of map tasks (MapReduce intermediate data are not stored in HDFS).

Forty nodes in a single rack share an IP switch. The rack switches are connected to each of eight core switches. The core switches provide connectivity between racks and to out-of-cluster resources. For each cluster, the NameNode and the BackupNode hosts are specially provisioned with up to 64 GB RAM; application tasks are never assigned to those hosts. In total, a cluster of 4000 nodes has 11 PB (petabytes; 1000 terabytes) of storage available as blocks that are replicated three times yielding a net 3.7 PB of storage for user applications. Over the years that HDFS has been in use, the hosts selected as cluster nodes have benefited from improved technologies. New cluster nodes always have faster processors, bigger disks and larger RAM. Slower, smaller nodes are retired or relegated to clusters reserved for development and testing of Hadoop.

On an example large cluster (4000 nodes), there are about 65 million files and 80 million blocks. As each block typically is replicated three times, every data node hosts 60 000 block replicas. Each day, user applications will create two million new files on the cluster. The 40 000 nodes in Hadoop clusters at Yahoo! provide 40 PB of on-line data storage.

Becoming a key component of Yahoo!'s technology suite meant tackling technical problems that are the difference between being a research project and being the custodian of many petabytes of

corporate data. Foremost are issues of robustness and durability of data. But also important are economical performance, provisions for resource sharing among members of the user community, and ease of administration by the system operators.

Durability of Data

Replication of data three times is a robust guard against loss of data due to uncorrelated node failures. It is unlikely Yahoo! has ever lost a block in this way; for a large cluster, the probability of losing a block during one year is less than 0.005. The key understanding is that about 0.8 percent of nodes fail each month. (Even if the node is eventually recovered, no effort is taken to recover data it may have hosted.) So for the sample large cluster as described above, a node or two is lost each day. That same cluster will re-create the 60 000 block replicas hosted on a failed node in about two minutes: re-replication is fast because it is a parallel problem that scales with the size of the cluster. The probability of several nodes failing within two minutes such that all replicas of some block are lost is indeed small.

Correlated failure of nodes is a different threat. The most commonly observed fault in this regard is the failure of a rack or core switch. HDFS can tolerate losing a rack switch (each block has a replica on some other rack). Some failures of a core switch can effectively disconnect a slice of the cluster from multiple racks, in which case it is probable that some blocks will become unavailable. In either case, repairing the switch restores unavailable replicas to the cluster. Another kind of correlated failure is the accidental or deliberate loss of electrical power to the cluster. If the loss of power spans racks, it is likely that some blocks will become unavailable. But restoring power may not be a remedy because one-half to one percent of the nodes will not survive a full power-on restart. Statistically, and in practice, a large cluster will lose a handful of blocks during a power-on restart.

In addition to total failures of nodes, stored data can be corrupted or lost. The block scanner scans all blocks in a large cluster each fortnight and finds about 20 bad replicas in the process. Bad replicas are replaced as they are discovered.

Features for Sharing HDFS

As the use of HDFS has grown, the filesystem itself has had to introduce means to share the resource among a large number of diverse users. The first such feature was a permissions framework closely modeled on the Unix permissions scheme for file and directories. In this framework, files and directories have separate access permissions for the owner, for other members of the user group associated with the file or directory, and for all other users. The principle differences between Unix (POSIX) and HDFS are that ordinary files in HDFS have neither execute permissions nor sticky bits.

In the earlier version of HDFS, user identity was weak: you were who your host said you are. When accessing HDFS, the application client simply queries the local operating system for user identity and group membership. In the new framework, the application client must present to the name system credentials obtained from a trusted source. Different credential administrations are possible; the initial implementation uses Kerberos. The user application can use the same framework to confirm that the name system also has a trustworthy identity. And the name system also can demand credentials from each of the data nodes participating in the cluster.

The total space available for data storage is set by the number of data nodes and the storage provisioned for each node. Early experience with HDFS demonstrated a need for some means to enforce the resource allocation policy across user communities. Not only must fairness of sharing be enforced, but when a user application might involve thousands of hosts writing data, protection against applications inadvertently exhausting resources is also important. For HDFS, because the system metadata are always in RAM, the size of the namespace (number of files and directories) is also a finite resource. To manage storage and namespace resources, each directory may be assigned a quota for the total space occupied by files in the sub-tree of the namespace beginning at that directory. A separate quota may also be set for the total number of files and directories in the sub-tree.

While the architecture of HDFS presumes most applications will stream large data sets as input, the MapReduce programming framework can have a tendency to generate many small output files (one from each reduce task) further stressing the namespace resource. As a convenience, a directory sub-tree can be collapsed into a single Hadoop Archive file. A HAR file is similar to a familiar tar, JAR, or Zip file, but filesystem operations can address the individual files within the archive, and a HAR file can be used transparently as the input to a MapReduce job.

Scaling and HDFS Federation

Scalability of the NameNode has been a key struggle [Shv10]. Because the NameNode keeps all the namespace and block locations in memory, the size of the NameNode heap limits the number of files and also the number of blocks addressable. This also limits the total cluster storage that can be supported by the NameNode. Users are encouraged to create larger files, but this has not happened since it would require changes in application behavior. Furthermore, we are seeing new classes of applications for HDFS that need to store a large number of small files. Quotas were added to manage the usage, and an archive tool has been provided, but these do not fundamentally address the scalability problem.

A new feature allows multiple independent namespaces (and NameNodes) to share the physical storage within a cluster. Namespaces use blocks grouped under a Block Pool. Block pools are analogous to logical units (LUNs) in a SAN storage system and a namespace with its pool of blocks is analogous to a filesystem volume.

This approach offers a number of advantages besides scalability: it can isolate namespaces of different applications improving the overall availability of the cluster. Block pool abstraction allows other services to use the block storage with perhaps a different namespace structure. We plan to explore other approaches to scaling such as storing only partial namespace in memory, and truly distributed implementation of the NameNode.

Applications prefer to continue using a single namespace. Namespaces can be mounted to create such a unified view. A client-side mount table provide an efficient way to do that, compared to a server-side mount table: it avoids an RPC to the central mount table and is also tolerant of its failure. The simplest approach is to have shared cluster-wide namespace; this can be achieved by giving the same client-side mount table to each client of the cluster. Client-side mount tables also allow applications to create a private namespace view. This is analogous to the per-process namespaces that are used to deal with remote execution in distributed systems [PPT⁺93, Rad94, RP93].

9.5 Lessons Learned

A very small team was able to build the Hadoop filesystem and make it stable and robust enough to use it in production. A large part of the success was due to the very simple architecture: replicated blocks, periodic block reports and central metadata server. Avoiding the full POSIX semantics also helped. Although keeping the entire metadata in memory limited the scalability of the namespace, it made the NameNode very simple: it avoids the complex locking of typical filesystems. The other reason for Hadoop’s success was to quickly use the system for production at Yahoo!, as it was rapidly and incrementally improved. The filesystem is very robust and the NameNode rarely fails; indeed most of the down time is due to software upgrades. Only recently have failover solutions (albeit manual) emerged

Many have been surprised by the choice of Java in building a scalable filesystem. While Java posed challenges for scaling the NameNode due to its object memory overhead and garbage collection, Java has been responsible to the robustness of the system; it has avoided corruption due to pointer or memory management bugs.

9.6 Acknowledgment

We thank Yahoo! for investing in Hadoop and continuing to make it available as open source; 80% of the HDFS and MapReduce code was developed at Yahoo! We thank all Hadoop committers and collaborators for their valuable contributions.

Jitsi

Emil Ivov

Jitsi はビデオ通話、音声通話、デスクトップ共有、ファイル交換、メッセージングなどを行うアプリケーションである。さらに重要な事は多くのプロトコルを使ってこれらを実装していることである。実装しているプロトコルには、XMPP(Extensible Messaging and Presence Protocol) や SIP(Session Initiation Protocol) のような標準プロトコルから Yahoo! や Windows Live Messenger(MSN) のような専用プロトコルまで含まれる。また、Microsoft Windows や Apple Mac OS X、Linux、FreeBSD 上で動作する。多くの部分を Java で記述しているが、ネイティブで記述している部分もある。本章では、Jitsi の OSGi ベース・アーキテクチャに注目する。プロトコルをどのように実装および管理しているのか、構築する上で学んだことも振り返る¹。

10.1 Jitsi の設計

マルチプロトコル・サポート、クロスプラットフォーム、開発者フレンドリーの 3 点が、Jitsi(かつて SIP Communicator と呼ばれた) の設計に際して制約として重視している項目である。

開発者の視点からみると、マルチプロトコルを採用することは、すべてのプロトコルに対して共通のインターフェースが必要になる事を意味する。具体的には、ユーザがメッセージを送信するときには、グラフィカル・ユーザ・インターフェースは、常に `sendMessage` というメソッドを呼び出す必要がある。実際には、使用しているプロトコルによって、`sendXmppMessage` が呼ばれたり、`sendSipMsg` が呼ばれたりする。

我々のソースコードの多くが Java で記述されているという事実は、かなりの部分、2 点目の制約：クロスプラットフォーム を満たしている。しかし、Java Runtime Environment (JRE) がサポートしていないかったり、我々が望むような方法で実装されていない部分もある。例え

¹ ソースコードの参照は、<http://jitsi.org/source>。Eclipse や NetBeans を使っているのなら、次のサイトにコンフィグレーション方法が載っている：<http://jitsi.org/eclipse>、<http://jitsi.org/netbeans>

ば、ウェブカムからのビデオキャプチャのようなもの SIP Communicator である。そのために、Windows では DirectShow を使ったり Mac OS X では QTKit を使ったり、Linux では Video for Linux 2 を使ったりしている。プロトコルの場合と同じようにビデオ電話を制御する部分のソースコードは、これらの詳細部を隠蔽する(かなり複雑ではある)。

最後の開発者フレンドリーとは、新しい機能の追加が容易であるべきという事を意味する。今日、大勢の人々が VoIP を使っているが、使い方は様々である。多くのサービス・プロバイダやサーバ・ベンダは、異なるユースケースやアイデアを用いて新しい機能を追加している。Jitsi を使う人にとって、これらの望まれる機能の実装が簡単で無ければならない。何か新しい機能を追加する人が、追加・変更に関わる部分のソースコードだけを読んで理解できる必要がある。同様に、ある人の変更が他の人の作業に与える影響が最小限でなければならない。

要約すれば、ソースコードの各部分は、それぞれ独立しているという環境が必要である。オペレーティング・システムに依存している部分が容易に置き換え可能でなければならぬ；プロトコルのように同時に複数が動作しても同じように動作しなければならない。また、各部分が完全にリライト可能でかつ、残りの部分は変更不要でなければならぬ。最後に、インターネット経由でダウンロードしてプラグインとして追加できる機能や各部分を簡単にオン・オフできる機能も望まれる。

我々は、単純に我々自身でフレームワークを記述することを考えた、しかし、直ぐにそのアイデアを捨てた。我々は、できるだけ早く VoIP と IM のソースコードの記述を開始することを望んだ。プラグイン・フレームワークに数ヶ月間費やしたが、エキサイティングには思えなかつた。誰かが OSGi の採用を提案したとき、これが完全にフィットするように思えた。

10.2 Jitsi と OSGi フレームワーク

OSGi の解説本は既にある。よって、このフレームワークの全体を説明するつもりは無い。代わりに、このフレームワークから得られる事および Jitsi での使われ方を説明する。

OSGiにおいて最も大切な事はモジュールである。OSGi アプリケーションの機能はバンドルに分割されている。ひとつの OSGi バンドルは、Java ライブラリや Java アプリケーションを配布するときに使われる標準の JAR ファイルよりも小さい単位である。Jitsi は、これらのバンドルの集合である。Windows Live Messenger と接続する責任を持つバンドルもあれば、XMPP と接続する責任を持つバンドルもある。ほかにも、GUI を扱うバンドルなどもある。これらすべてのバンドルは与えられた環境で動作する。我々の場合は、Apache Felix というオープンソースの OSGi 実装の環境下で動作する。

これらのモジュールは全て一緒に動作する必要がある。GUI バンドルは、プロトコル・バンドルを通じてメッセージを送信する必要がある。さらにメッセージ履歴を取り扱うバンドルを経由してこれらのメッセージを保存する必要がある。これは、OSGi サービスが何であるかを示している：OSGi サービスはバンドルの一部を表していて他バンドルからアクセス可能である。OSGi サービスは、ログ、ネットワーク経由のメッセージ送信、通話履歴の読出

しのような特定機能の利用を許可する Java インタフェースのグループである。実際に機能を実装するクラスは、サービス実装として知られている。サービス実装の多くは実装したサービス・インターフェース名を携える。サービス実装名は、“Impl” の接尾辞を持っている(例えば, ConfigurationServiceImpl)。OSGi フレームワークは開発者にサービスの実装を隠蔽している。また、OSGi フレームワークは、サービス実装がバンドル自身の外側には決して見えないことを確実にしている。この様にして、他のバンドルは、サービス・インターフェースを通してのみ利用できる。

ほとんどのバンドルはアクティベータも持つ。アクティベータは、`start` と `stop` メソッドを定義したシンプルなインターフェースである。Felix が Jitsi のバンドルをロードやリムーブする度に、バンドルが起動やシャットダウンの準備ができるよう、これらのメソッドをコールする。これらのメソッドをコールするとき、Felix は `BundleContext` と言う名前のパラメータを渡す。`BundleContext` は、バンドルに OSGi 環境に接続する方法を与える。この様にして、使用したい OSGi サービスが何であろうと見つける事ができたり、自身を登録したりできる(図 10.1)。

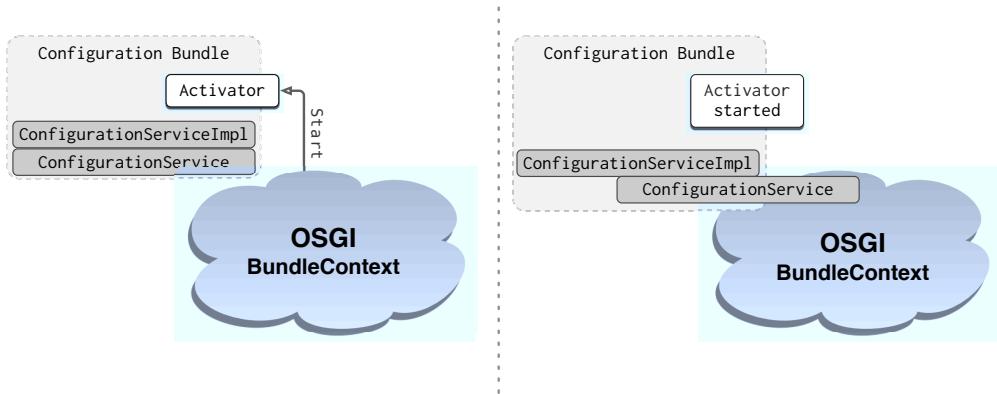


図 10.1: OSGi バンドル・アクティベーション

それでは、実際にこれがどのように動作するのかを見ていこう。プロパティを保存したり引き出したりするだけのサービスを想像してみよう。Jitsi では、これを `ConfigurationService` と呼んでいて次のようなものだ：

```
package net.java.sip.communicator.service.configuration;

public interface ConfigurationService
{
    public void setProperty(String propertyName, Object property);
    public Object getProperty(String propertyName);
}
```

極めてシンプルな ConfigurationService の実装は次のようなものである：

```
package net.java.sip.communicator.impl.configuration;

import java.util.*;
import net.java.sip.communicator.service.configuration.*;

public class ConfigurationServiceImpl implements ConfigurationService
{
    private final Properties properties = new Properties();

    public Object getProperty(String name)
    {
        return properties.get(name);
    }

    public void setProperty(String name, Object value)
    {
        properties.setProperty(name, value.toString());
    }
}
```

net.java.sip.communicator.service パッケージの中でサービスがどのように定義されるか注目すること。また、実装は net.java.sip.communicator.impl にある。Jitsi におけるすべてのサービスと実装は、このように 2つのパッケージに分割される。OSGi は、バンドル自身が含まれる JAR の外部に対して、いくつかのパッケージだけを可視化する事を許可する。こうして、この分割は、バンドルに対してサービス・パッケージだけを *export* して実装を隠蔽する事ができる。

ユーザが、我々の実装を使い始めるのに必要となる最後の作業は、BundleContext に登録する事と ConfigurationService の実装を備えている事を、伝える事である。次は、これがどのように行われるかを示している：

```
package net.java.sip.communicator.impl.configuration;

import org.osgi.framework.*;
import net.java.sip.communicator.service.configuration;

public class ConfigActivator implements BundleActivator
{
    public void start(BundleContext bc) throws Exception
    {
```

```

        bc.registerService(ConfigurationService.class.getName(), // service name
                           new ConfigurationServiceImpl(), // service implementation
                           null);
    }
}

```

ConfigurationServiceImpl クラスが BundleContext に登録されると他のバンドルが使い始める事ができる。いくつかのランダム・バンドルがコンフィグレーション・サービスを使う例を示す：

```

package net.java.sip.communicator.plugin.randombundle;

import org.osgi.framework.*;
import net.java.sip.communicator.service.configuration.*;

public class RandomBundleActivator implements BundleActivator
{
    public void start(BundleContext bc) throws Exception
    {
        ServiceReference cRef = bc.getServiceReference(
            ConfigurationService.class.getName());
        configService = (ConfigurationService) bc.getService(cRef);

        // And that's all! We have a reference to the service implementation
        // and we are ready to start saving properties:
        configService.setProperty("propertyName", "propertyValue");
    }
}

```

再びパッケージに注目する。net.java.sip.communicator.plugin の中で他によって定義されたサービスを使うバンドルをキープするが、自身をエクスポートしたりインプリメントしたりはしない。コンフィグレーション・フォームは、このようなプラグインの良い例である：これらは、ユーザにアプリケーションのある部分をコンフィグする事を許可する Jitsi ユーザインターフェースへの追加である。ユーザが優先権を変更するとコンフィグレーション・フォームは、ConfigurationService や、この機能に対する責任を持つバンドルと直接に作用し合う。しかし、他のバンドルは、作用する必要は無い(図 10.2)。

10.3 バンドルの構築と実行

バンドルのコード記述方法を一通り見てきた所で、次はパッケージングについて説明しよう。実行中、すべてのバンドルは OSGi 環境に次の 3つを示す：他が利用可能な Java パッケージ(すなわちエクスポート・パッケージ)、他が使いたいと思う Java パッケージ(すなわちインポート・パッケージ)、BundleActivator クラスの名前。バンドルは、これを自身が配置される JAR ファイルのマニフェストを通して行う。

上記で定義した ConfigurationService に対するマニフェスト・ファイルは次のようになる：

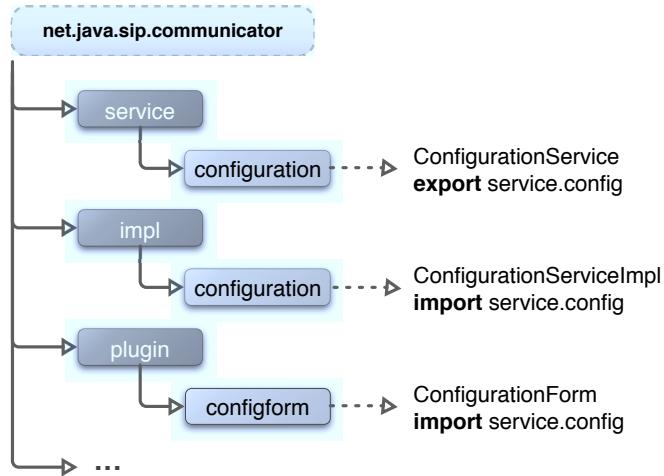


図 10.2: サービス・ストラクチャ

```

Bundle-Activator: net.java.sip.communicator.impl.configuration.ConfigActivator
Bundle-Name: Configuration Service Implementation
Bundle-Description: A bundle that offers configuration utilities
Bundle-Vendor: jitsi.org
Bundle-Version: 0.0.1
System-Bundle: yes
Import-Package: org.osgi.framework,
Export-Package: net.java.sip.communicator.service.configuration
  
```

JAR マニフェストが生成されると、バンドルを生成する準備ができる。Jitsi では、構築関連のタスクには Apache Ant を使う。バンドルを Jitsi ビルド・プロセスに追加するために、プロジェクトのルート・ディレクトリにある build.xml を編集する必要がある。バンドル JAR は、build.xml ファイルの最後の bundle-xxx ターゲットによって生成される。コンフィグレーション・サービスを構築するためには、次のようにする：

```

<target name="bundle-configuration">
  <jar destfile="${bundles.dest}/configuration.jar" manifest=
    "${src}/net/java/sip/communicator/impl/configuration/conf.manifest.mf" >

    <zippedset dir="${dest}/net/java/sip/communicator/service/configuration"
      prefix="net/java/sip/communicator/service/configuration"/>
    <zippedset dir="${dest}/net/java/sip/communicator/impl/configuration"
      prefix="net/java/sip/communicator/impl/configuration" />
  </jar>
</target>
  
```

お分かりのように、Ant ターゲットは、単純にコンフィグレーション・マニフェストを使って JAR ファイルを生成して、service と impl の階層構造から構成されるコンフィグレーション・パッケージに追加する。我々が必要な事は Felix にロードさせる事だけである。

Jitsi は単に OSGi バンドルの集合であることは既に説明した。ユーザがアプリケーションを実行する時、OSGi バンドルは、ロードする必要のあるバンドルのリストとともに、Felix をスタートさせる。このリストは、lib ディレクトリの `felix.client.run.properties` ファイルの中にある。Felix は、スタートレベルによって定義される順番にバンドルをスタートする：すべての、あるレベルのバンドルは、続くレベルのバンドルがロードを開始する前に、ロードが完了していることを保証されている。このことは上記の例では確認できないが、我々のコンフィグレーション・サービスはプロパティをファイルに保存する。よって、`FileAccessService` を使う必要がある。これは、`fileaccess.jar` ファイルに収められている。`ConfigurationService` は、`FileAccessService` のあとに開始することを確実にする：

```
...
felix.auto.start.30= \
    reference:file:sc-bundles/fileaccess.jar

felix.auto.start.40= \
    reference:file:sc-bundles/configuration.jar \
    reference:file:sc-bundles/jmdnslib.jar \
    reference:file:sc-bundles/provdisc.jar \
...
...
```

`felix.client.run.properties` ファイルを見ると、先頭にパッケージのリストを見つける事ができる：

```
org.osgi.framework.system.packages.extra= \
    apple.awt; \
    com.apple.cocoa.application; \
    com.apple.cocoa.foundation; \
    com.apple.eawt; \
...
...
```

このリストは、Felix にシステムのクラス・パスからバンドルを利用するのに必要なパッケージは何であるかを教える。これは、このリスト上のパッケージは、他のバンドルによってエクスポートされる事なしに、バンドルによってインポートされうる(つまり、*Import-Package* マニフェスト・ヘッダに加える)事を意味する。リストは、大抵、OS 依存の JRE 部品に由来するパッケージを含む。そして、Jitsi 開発者は、新しいパッケージを追加する必要はほとんどない；多くの場合、パッケージはバンドルによって利用される事ができる。

10.4 プロトコル・プロバイダ・サービス

Jitsi の `ProtocolProviderService` は、すべてのプロトコルの実装の振る舞いを定義する。これは、Jitsi が接続しているネットワーク上で、(ユーザ・インターフェースのような)他のバンドルがメッセージの送受信、通話、ファイルシェアを行う必要があるときに使うインターフェースである。

プロトコル・サービス・インターフェースは、`net.java.sip.communicator.service.protocol` パッケージ下に見つける事ができる。サービスに対して複数の実装があつたり、サポート・プロトコル毎のサービス実装があつたりする。全ては、`net.java.sip.communicator.impl.protocol.-protocol_name` に保存されている。

`service.protocol` ディレクトリから始めよう。もっとも重要な部分は、`ProtocolProviderService` インタフェースである。プロトコル関連のタスクを行おうとすると、必ず `BundleContext` にあるサービスの実装を調べなくてはならない。サービスとその実装は、Jitsi にサポートするネットワーク、接続状態の確認や詳細確認、最も重要なチャットや通話のような実際の通信タスクを実装したクラスへの参照を取得する事を可能にする。

オペレーション・セット

前に触れたように、`ProtocolProviderService` は、様々な通信プロトコルとその差異を隠して利用するのに必要である。これは、メッセージ送信のようなすべてのプロトコルが持っている機能に対しては極めてシンプルであるが、サポートするプロトコルが少ない機能に対しては手の込んだものになる。これらの違いは、しばしばサービス自身に由来する：例えば、世の中にあるほとんどの SIP サービスは、連絡先リストをサポートしていないが、他のプロトコルでは大抵サポートしている。MSN と AIM は、もうひとつの分かりやすい例である：どちらのプロトコルもオフラインのユーザに対してのメッセージ送信はサポートしていないが、他のプロトコルではサポートしている（現在は変わった）。

`ProtocolProviderService` は、GUI 等の他のバンドルで行っているように、相違を扱う方法が必要である：実際に通話する機能がなければ、AIM コンタクトに通話ボタンを追加する意味がない。

`OperationSets` チェックシート（図 10.3）。当然の事だが、これらは操作のセットであり、Jitsi バンドルが、プロトコル実装を制御するのに使用するインターフェースを提供する。オペレーション・セット・インターフェースにあるメソッドは、すべて特定の機能に関連している。例えば、`OperationSetBasicInstantMessaging` は、インスタント・メッセージの送受信用のメソッドを持っていて、Jitsi が受信したメッセージを引き出すためのリスナーを登録する。もうひとつの例は、`OperationSetPresence` である。`OperationSetPresence` は、連絡先リスト上の状態を問い合わせたり、自分自身の状態を登録するメソッドを持っている。よって、GUI が状態を更新して、連絡先を表示したり、連絡先にメッセージを送信するとき、プレゼンスやメッセージングをサポートしようがしまいが、対応するプロバイダに最初に問い合わせる事ができる。`ProtocolProviderService` がこの目的で定義するメソッドは、次のようになる：

```
public Map<String, OperationSet> getSupportedOperationSets();
public <T extends OperationSet> T getOperationSet(Class<T> opsetClass);
```

`OperationSets` は、新しく追加したプロトコルが `OperationSet` で定義したオペレーションのいくつかしかサポートしないような事がないように設計されなければならない。例えば、相

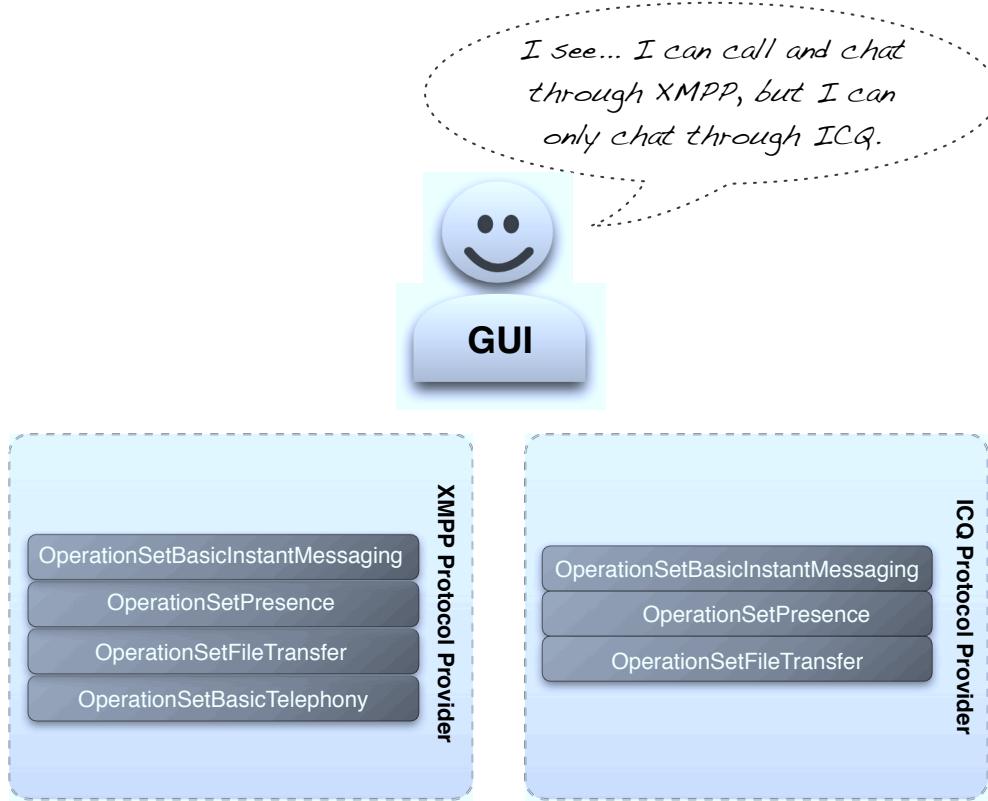


図 10.3: オペレーション・セット

互の状態を問い合わせる機能はあるが、サーバに連絡先を保存する機能をサポートしないプロトコルがある。よって、OperationSetPresence のプレゼンス管理と仲間リストの引き出し機能を結合するよりも、連絡先をオンラインで保存できるプロトコルとだけ使用できる OperationSetPersistentPresence を定義した方が良い。他方、受信することはせずに送信だけできるプロトコルもあり、送信メッセージと受信メッセージが安全に結合できる理由である。

アカウント、ファクトリおよびプロバイダのインスタンス

ProtocolProviderService の重要な特性は、インスタンスとプロトコル・アカウントは 1 対 1 に対応しているという事である。よって、常にユーザが登録したアカウントの数と同じ数の BundleContext のサービス実装がある。

ここで、誰がプロトコル・プロバイダを生成・登録するのだろうかと思うだろう。これには、2 つの異なるエンティティが関わっている。一つ目は、ProtocolProviderFactory である。これ

は、他のバンドルがプロバイダをインスタンス化して、サービスとして登録する事を許可する。プロトコル毎にファクトリがあり、各ファクトリは対応するプロトコルのプロバイダを生成する責任を持つ。ファクトリ実装は、プロトコル内部の残りの部分に保存される。SIPを例にすると、`net.java.sip.communicator.impl.protocol.sip.ProtocolProviderFactorySipImpl`である。

アカウント生成を含むふたつ目のエンティティは、プロトコル・ウィザードである。ファクトリとは違って、ウィザードグラフィカル・ユーザ・インターフェースを含むためプロトコル実装の残りの部分から分離されている。例えば、SIPアカウントを生成を許可するウィザードは、`net.java.sip.communicator.plugin.sipaccregwizz`にある。

10.5 メディア・サービス

IP上でリアルタイム通信を行う場合、理解しておくべき重要な事がある：SIPやXMPPのようなプロトコルは、もっとも一般的なVoIPプロトコルとして認識されている一方、インターネット上で音声やビデオを実際に動かすプロトコルではない。これはリアルタイム・プロトコル(RTP)によって扱われる。SIPとXMPPは、RTPパケットの送信先アドレスの決定、音声、ビデオの符号化方式(つまりコーデック)の交渉などのRTPが必要とするすべての準備に対してだけ責任を持つ。ユーザの位置管理、プレゼンス管理、着信音など他の多くの事に対しても面倒を見る。これは、SIPやXMPPのようなプロトコルはシグナリング・プロトコルと呼ばれる理由である。

これは、Jitsiのコンテキストではどんな意味をもつんだろうか？まず最初に、`sip`や`jabber`のJitsiパッケージには、音声やビデオのフローを操作するソースコードは見つからないと言う事である。この手のソースコードは、`MediaService`にある。`MediaService`とその実装は、`net.java.sip.communicator.service.neomedia`と`net.java.sip.communicator.impl.neomedia`にある。

なぜ“neomedia”？

`neomedia`パッケージ名の“neo”は、初期に使っていた類似のパッケージを置き換える事を意味し、完全に置き換えを行った。これは、我々の経験則「最新を完全に保つ所までアプリケーションを設計するのに多くの時間を使う事には価値がない」を生んだ。単純に、すべてを考慮にいれる方法は無いので、後に変更される運命にある。さらに、綿密な設計フェーズは複雑性を産むが、準備したようなシナリオは決して発生しないために、使われることはない。

MediaService自身に加えて、特に重要なインターフェースが二つある：`MediaDevice`と`MediaStream`である。

キャプチャ、ストリーミング、再生

MediaDevice は、通話時に使うキャプチャ・デバイスと再生デバイスを表わす(図 10.4)。マイクロフォン、スピーカ、ヘッドセット、ウェブカムは全てこのような MediaDevice の例であるが、これだけではない。Jitsi のデスクトップ・ストリーミングとシェアリング・コールは、デスクトップからビデオ・キャプチャを行う。会議通話は、参加者の音声を合成するために AudioMixer デバイスを使う。すべてのケースで、MediaDevice は単一の MediaType を表わす。つまり、音声かビデオのどちらかには成れるが両方にはなれない。これは、例えば、マイクロフォンが統合されたウェブカムを持っている場合、Jitsi はふたつのデバイスと認識する:ひとつはビデオ・キャプチャだけを行い、もうひとつはサウンド・キャプチャだけを行う。

デバイスだけでは、電話やビデオ通話をを行うには、不十分である。メディアを再生したりキャプチャしたりするのに加えて、ネットワーク上に送信できなければならない。これには、MediaStream が使われる。MediaStream インタフェースは、MediaDevice と通話相手をつなげる。通話中に交換する通話相手との受信および送信パケットを表わす。

デバイスと同様に、ひとつのストリームは、一つの MediaType に対してのみ責任を持つ。これは、音声/ビデオ通話の場合、Jitsi はふたつの分離したメディア・ストリームを生成して、それぞれを対応する音声、ビデオの MediaDevice と接続する。

コーデック

メディア・ストリーミングで、もう一つの重要な概念は、コーデックとして知られている MediaFormat である。デフォルトでは、多くのオペレーティング・システムは、オーディオを 48KHz PCM か類似の方式でキャプチャする。これは、我々がしばしば、“raw audio”として参照するもので、WAV ファイルとして取得するオーディオ方式である。WAV ファイルは、高品質であるが莫大なサイズを持つ。この PCM フォーマットでインターネット上にオーディオを転送を試みるのは非現実的である。

これは、コーデックの意味を示す:オーディオやビデオを様々な方法で表現し転送する。iLBC や 8KHz Speex や G.729 のようなオーディオ・コーデックは、狭帯域であるが、こもつたように聞こえる。ワイドバンド Speex や G.722 は、高品質のオーディオを提供するが、より多くの帯域を使用する。高い品質を保ちつつ帯域も合理的であることを狙ったコーデックもある。ポピュラーなビデオコーデックである H.264 が好例である。ここでのトレードオフは、変換時の計算量である。Jitsi で H.264 ビデオ通話を使用すると、高い品質の画像と合理的な帯域幅を痛感するだろう。しかし、CPU 負荷は高い。

単純化すると、コーデック選択は妥協が全てであるという事である。帯域幅、品質、CPU 消費量または、これらの組み合せのどれかを犠牲にする。ほとんどの場合、VoIP 関連の人々はコーデックについて、これ以上知る必要は無い。

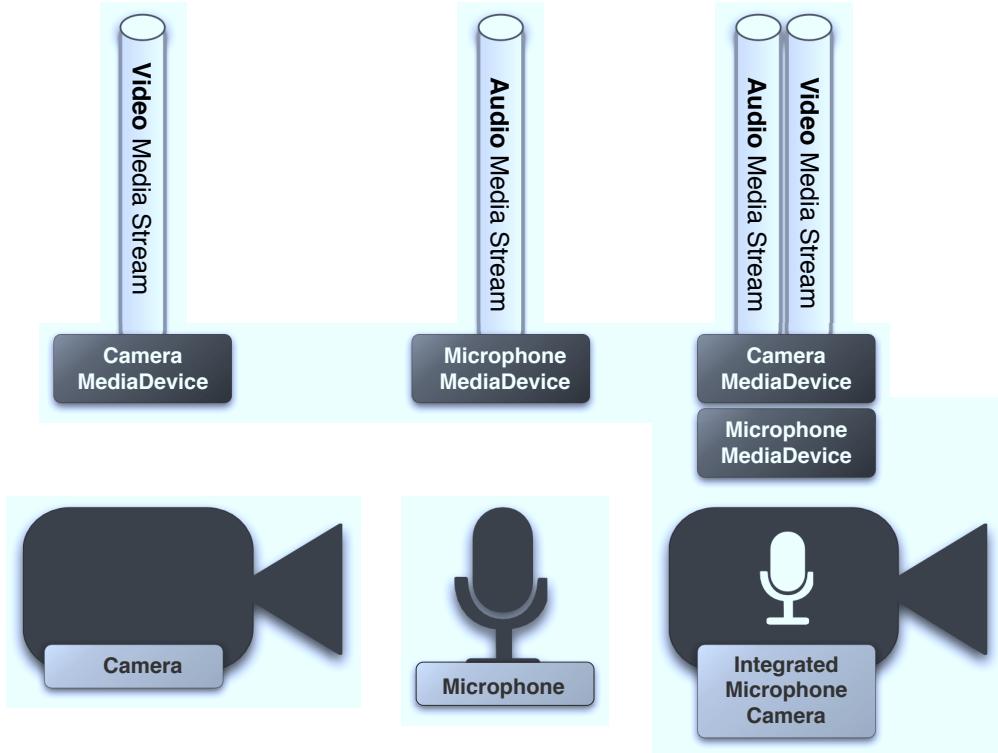


図 10.4: 異なるデバイスに対するメディア・ストリーム

プロトコル・プロバイダとの接続

オーディオ/ビデオをサポートしている Jitsi のプロトコルは全て、MediaService を全く同じ方法で使用する。最初にシステムで使用可能なデバイスを MediaService に問い合わせる：

```
public List<MediaDevice> getDevices(MediaType mediaType, MediaUseCase useCase);
```

MediaType はオーディオ・デバイスかビデオ・デバイスのどちらに興味があるのかを示す。MediaUseCase パラメータは、現在の所、ビデオ・デバイスの場合のみ扱われる。これは、次のようにしてメディア・サービスに利用できるデバイスを知らせる。通常通話 (MediaUseCase.CALL) の場合は、利用可能なウェブカムのリストを返し、デスクトップ・シェアリング・セッション (MediaUseCase.DESKTOP) の場合は、ユーザ・デスクトップへの参照を返す。

次のステップは、デバイスが利用できるフォーマットのリストを取得する事である。これは、`MediaDevice.getSupportedFormats` メソッドを使って次のようにする：

```
public List<MediaFormat> getSupportedFormats();
```

プロトコルの実装は、このリストを取得してリモート・パーティに送る。リモート・パーティは、このリストの中からサポートしているフォーマットのサブセットを作成して応答する。

この交換は、オファー/アンサーモデルとして知られていて、セッション記述プロトコルや同類のプロトコルで利用される。

フォーマット、ポート番号および IP アドレスを交換した後、VoIP プロトコルはメディア・ストリームの生成、コンフィグ、開始を行う。この初期化は、大雑把に次のような流れになる：

```
// 最初にストリーム・コネクタを生成する。ストリームコネクタはメディア・サービス
// に利用するソケットを教える。
// 利用するソケットは、メディア転送時の RTP、フロー制御や統計メッセージの RTCP である。
StreamConnector connector = new DefaultStreamConnector(rtpSocket, rtcpSocket);
MediaStream stream = mediaService.createMediaStream(connector, device, control);

// MediaStreamTarget は、通信相手のメディアが使おうとするアドレスとポート番号を示す。
// この情報を交換する方法は、VoIP プロトコル毎に異なる。
stream.setTarget(target);

// MediaDirection パラメータは、stream に着信なのか発信なのか両方なのかを知らせる。
stream.setDirection(direction);

// そして、ストリーム・フォーマットを設定する。セッション交渉の応答に含まれるリストの
// 最初のフォーマットを使う。
stream.setFormat(format);

// 最後に、メディア・デバイスから media を取得する準備が整い、インターネット上に、
// ストリーミングする。
stream.start();
```

自分のウェブカムで wave でき、マイクを使って、「Hello world!」と言うことができる。

10.6 UI サービス

ここまでで、Jitsi におけるプロトコルの扱い、メッセージ送受信、通話の部分をカバーした。さらに、Jitsi は実際の一般の人々に使われるアプリケーションであり、ユーザインターフェースが最も重要な側面を持つ。多くの時間、ユーザインターフェースは Jitsi の他の全てのバンドルが明示するサービスを使う。しかし、そうとも限らない場面もある。

プラグインは、気に留めておくべき最初の例である。Jitsi におけるプラグインは、ユーザと相互作用できる必要がある。これは、プラグインは、ユーザ・インターフェースのウィンドウやパネルに、コンポーネントをオープンしたり、クローズしたり、ムーブしたり、追加する必要がある事を意味する。これは、UIService が実行される時の話である。Jitsi メイン・ウィンドウの基本制御を可能にすると併に、Mac OS X ドックのアイコンや Windows の通知エリアでアプリケーションをコントロールする方法である。

プラグインは、単純に連絡先リストを使うのに加えて、この機能を拡張できる。Jitsi の暗号チャット (OTR) をサポートするプラグインは、プラグインによる機能拡張の良い例である。OTR バンドルは、ユーザ・インターフェースの様々な部品の中からいくつかの GUI コンポー

メントを登録する必要がある。チャット・ウィンドウに錠前ボタンを追加し、すべての連絡先に対する右クリックメニューにサブセクションを追加する。

良いニュースは、いくつかのメソッド呼び出しだけで、これを行う事ができるという事である。OTR バンドルに対する OSGi アクティベータ OtrActivator は、次のソースコードを含んでいる：

```
Hashtable<String, String> filter = new Hashtable<String, String>();

// Register the right-click menu item.
filter(Container.CONTAINER_ID,
       Container.CONTAINER_CONTACT_RIGHT_BUTTON_MENU.getID());

bundleContext.registerService(PluginComponent.class.getName(),
    new OtrMetaContactMenu(Container.CONTAINER_CONTACT_RIGHT_BUTTON_MENU),
    filter);

// Register the chat window menu bar item.
filter.put(Container.CONTAINER_ID,
           Container.CONTAINER_CHAT_MENU_BAR.getID());

bundleContext.registerService(PluginComponent.class.getName(),
    new OtrMetaContactMenu(Container.CONTAINER_CHAT_MENU_BAR),
    filter);
```

みてわかるように、グラフィカル・ユーザ・インターフェースへのコンポーネント追加は、単純に OSGi サービスの登録に行き着く。反面、UIService の実装は、PluginComponent インターフェースの実装を探索する。新しい実装が登録された事を検出すると直ぐに、実装への参照を取得し、OSGi サービス・フィルタに示されたコンテナに追加する。

右クリック・メニュー項目のときには、これがどのように起こるのかを示す。UI バンドル内で、右クリック・メニューを表わすクラス MetaContactRightButtonMenu は、次のソースコードを含んでいる：

```
// Search for plugin components registered through the OSGI bundle context.
ServiceReference[] serRefs = null;

String osgiFilter = "("
    + Container.CONTAINER_ID
    + "=" + Container.CONTAINER_CONTACT_RIGHT_BUTTON_MENU.getID() + ")";

serRefs = GuiActivator.bundleContext.getServiceReferences(
    PluginComponent.class.getName(),
    osgiFilter);
// Go through all the plugins we found and add them to the menu.
for (int i = 0; i < serRefs.length; i++)
{
    PluginComponent component = (PluginComponent) GuiActivator
        .bundleContext.getService(serRefs[i]);
```

```

        component.setCurrentContact(metaContact);

        if (component.getComponent() == null)
            continue;

        this.add((Component)component.getComponent());
    }
}

```

これで全てである。Jitsi 内にあるウィンドウの多くは、全く同じ事をする：PluginComponet インタフェースを実装するサービスに対するバンドル・コンテキストを探索する。このインターフェースはフィルタを持っていて、このフィルタは対応するコンテナに追加されたい事を示す。プラグインは、行き先を示すボードを持っているヒッチハイカーのようなものである。Jitsi ウィンドウが彼らを拾う運転手である。

10.7 得た教訓

SIP Communicator に取り掛かったときに、最も多かった批判、疑問は、次のようなものである：“なぜ Java を使っている？、動作が遅いでしょう？、音声通話やビデオ通話での音声品質は、たかが知れている！”。 “Java は遅い” という通説は、Jitsi を試さずに Skype を使い続けるための理由として、潜在的なユーザによって、繰り返し述べられている。しかし、このプロジェクトにおける我々が最初に学んだ教訓は、Java における効率性への関心は、既に C++ や他のネイティブ言語における関心と同程度のものである、という事である。

全ての項目を厳密に分析した結果、Java を選択したというつもりはない。我々は、シンプルに Windows や Linux 上で実行する簡単な方法を望んだ。そして、Java と Java Media Framework は、これを行う比較的簡単な方法のように思えた。

この決定を悔やむ理由は、今の所、ほとんどない。それどころか、完全に透過的というわけではないが、Java は、移植性が高く、SIP Communicator のソースコードの 90% は、OS 間で共通である。これは、かなり複雑にも係わらず、すべてのプロトコル・スタック (たとえば、SIP、XMPP、RTP など) の実装を含んでいる。OS に依存するようなソースコード部分に心配を払う必要は無く、大変便利という事が証明された。

さらに、Java の高い評判は、コミュニティを形成する上で大変重要である事が分かった。コントリビュータは希少資源である。人々は、アプリケーションの性質を気に入る必要があり、時間とモチベーションを見つける必要がある—これらのすべてを集めるのは困難である。よって、新しい言語を習う必要がないというのは、利点である。

多くの期待に反して、Java の実効速度が遅いという仮説は、ネイティブ言語に移行する理由にはなっていない。多くの場合、ネイティブ言語にする決断は、OS との融合と Java がどれだけ OS 依存のユーティリティにアクセスできるかに掛かっている。以下で、Java が不十分である 3 大領域について説明する。

Java サウンド対 PortAudio

Java サウンドは、オーディオを取り込んだり、再生したりする Java のデフォルト API である。これは、Java ランタイム環境の一部であるので、Java 仮想マシン上のすべてのプラットフォームで動作する。SIP Communicator としての最初の数年、Jitsi は、もっぱら JavaSound を使っていて、かなり不便であった。

まず第一に、この API は使用するオーディオ・デバイスを選択する方法を提供しなかった。これは大問題である。コンピュータを音声やビデオ通話に使うとき、ユーザは、よく先進の USB ヘッドセットや、他の高品質なオーディオ・デバイスを使う。コンピュータ上で、複数のデバイスが存在するとき、JavaSound は OS がデフォルトと考えるデバイスを辿るが、多くの場合良くない。デフォルトのサウンドカード上のスピーカーを通して音楽を聞きながら、という風に多くのユーザは、他のアプリケーションを走らせながら使う事を好む。さらに重要なことは、多くの場合、SIP Communicator にとって、オーディオ通知を一つのデバイスに送り、実際の通話音声は他のデバイスに送るのが都合が良い。つまり、着信通知をコンピュータの前にはないとしてもユーザに聞こえるようにスピーカーから流し、着信に応答したらヘッドセット側に切り替える。

これは、Java Sound では不可能である。さらに Linux の実装では、今日の Linux ディストリビューションでは廃止予定の OSS を使っている。

我々は、他のオーディオシステムを使うことを決定した。我々は、マルチプラットフォームを諦めたくなかったので、できれば我々自身での実装は避けたかった。これは、PortAudio²が特に得意とする事である。

Java で実現できないことは、クロスプラットフォームのオープンソースプロジェクトを使うのが次善の策である。PortAudio への切り替えは、上記のようなオーディオのレンダリングや取り込みに関する優れたサポートを提供する。また、Windows、Linux、Mac OS X で動作し、FreeBSD など我々がパッケージを提供できないOS もサポートしている。

ビデオ・キャプチャ、レンダリング

ビデオは、オーディオと同様に重要である。しかし、これは、Java クリエータにとっては同様では無いようである。なぜなら、ビデオをキャプチャしたり、レンダリングしたりするデフォルトの JRE API が無いからである。Sun がメンテナンスを止めるまで、しばらく、Java Media Framework がこのような API を目指していたようである。

自然に、我々は PortAudio 方式のビデオの代替手段を探し始めた。しかし、今回はダメだった。最初は、Ken Larson³の LTI-CIVIL フレームワークに決め掛けた。これは、すばらしいプロジェクトで、しばらく⁴使用した。しかし、リアルタイム通信のコンテキストで使うには最適というわけではない事がわかつた。

²<http://portaudio.com/>

³<http://lti-civil.org/>

⁴実際に、まだ、非デフォルトのオプションとして残っている

Jitsi でのビデオ通信を満足に行う唯一の方法は、我々自身でネイティブの取り込みやレンダリングを実装する事だった。これは、簡単な決断ではなかった。複雑性が大幅に増し、潜在的なメンテナンス負荷がプロジェクトに追加される。しかし、他に選択肢はなかった。我々は高品質のビデオ通話を望んだ。そして実現した！

我々のネイティブの取り込みとレンダリングは、Linux、Mac OS X、Windows でそれぞれ、Video4Linux 2、QTKit、DirectShow/Direct3D を直に使った。

ビデオのエンコーディングとデコード

SIP Communicator 従って Jitsi は、最初のリリースからビデオ通話をサポートした。これは、Java Media Framework は、H.263 コーデックと 176x144(CIF) フォーマットをサポートするからである。H.264 CIF の見た目を知っている人は笑うだろう。現在、もしこれしか提供できないのであれば、ビデオ・チャットアプリケーションを使う人はいないだろう。

より高い品質を提供するためには、FFmpeg のような他のライブラリを使う必要があった。ビデオ・エンコーディングは、Java が性能的な限界を示す数少ない分野である。よって、FFmpeg 開発者がビデオを最も効率的な方法で処理するために実際にアセンブラーを多くの場所で使っているという事実が示すように、我々も他の言語を使う。

その他

より良い結果を得るために、ネイティブにする事を決定した箇所が多くある。Mac OS X における Systray 通知の Growl、Linux における libnotify は、こういった例である。他には、Microsoft Outlook や Apple Address Book の連絡先データベースへの問い合わせ、相手先アドレスによる送信元 IP アドレスの決定、既にある Speex や G.722 のコーデック実装の利用、デスクトップ・スクリーンショットの取得、キャラクタのキーコードへの変換がある。

重要な事は、我々がネイティブに解決策を求めるべき可能であり、実際に行ったという事である。これは、次のポイントをもたらした。Jitsi を初めて以来、多くの部分を修正したり、追加したり、完全に書き換えたりしている。これは、見た目、使い心地、性能をより良いものにしたかったからである。しかし、最初によいものを出せなかつた事を後悔はない。疑いを持ったら、選択肢の中から可能なものを選び、実行した。より良い方法を知るまで待つ事もできたが、これを行っていたら、今日までに Jitsi は存在していないだろう。

10.8 謝辞

本章の全ての図を作成してくれた Yana Stamcheva に感謝する。

LLVM

Chris Lattner

This chapter discusses some of the design decisions that shaped LLVM¹, an umbrella project that hosts and develops a set of close-knit low-level toolchain components (e.g., assemblers, compilers, debuggers, etc.), which are designed to be compatible with existing tools typically used on Unix systems. The name “LLVM” was once an acronym, but is now just a brand for the umbrella project. While LLVM provides some unique capabilities, and is known for some of its great tools (e.g., the Clang compiler², a C/C++/Objective-C compiler which provides a number of benefits over the GCC compiler), the main thing that sets LLVM apart from other compilers is its internal architecture.

From its beginning in December 2000, LLVM was designed as a set of reusable libraries with well-defined interfaces [LA04]. At the time, open source programming language implementations were designed as special-purpose tools which usually had monolithic executables. For example, it was very difficult to reuse the parser from a static compiler (e.g., GCC) for doing static analysis or refactoring. While scripting languages often provided a way to embed their runtime and interpreter into larger applications, this runtime was a single monolithic lump of code that was included or excluded. There was no way to reuse pieces, and very little sharing across language implementation projects.

Beyond the composition of the compiler itself, the communities surrounding popular language implementations were usually strongly polarized: an implementation usually provided *either* a traditional static compiler like GCC, Free Pascal, and FreeBASIC, *or* it provided a runtime compiler in the form of an interpreter or Just-In-Time (JIT) compiler. It was very uncommon to see language implementation that supported both, and if they did, there was usually very little sharing of code.

Over the last ten years, LLVM has substantially altered this landscape. LLVM is now used as a common infrastructure to implement a broad variety of statically and runtime compiled languages (e.g., the family of languages supported by GCC, Java, .NET, Python, Ruby, Scheme, Haskell, D, as well as countless lesser known languages). It has also replaced a broad variety of special pur-

¹<http://llvm.org>

²<http://clang.llvm.org>

pose compilers, such as the runtime specialization engine in Apple’s OpenGL stack and the image processing library in Adobe’s After Effects product. Finally LLVM has also been used to create a broad variety of new products, perhaps the best known of which is the OpenCL GPU programming language and runtime.

11.1 A Quick Introduction to Classical Compiler Design

The most popular design for a traditional static compiler (like most C compilers) is the three phase design whose major components are the front end, the optimizer and the back end (図 11.1). The front end parses source code, checking it for errors, and builds a language-specific Abstract Syntax Tree (AST) to represent the input code. The AST is optionally converted to a new representation for optimization, and the optimizer and back end are run on the code.



図 11.1: Three Major Components of a Three-Phase Compiler

The optimizer is responsible for doing a broad variety of transformations to try to improve the code’s running time, such as eliminating redundant computations, and is usually more or less independent of language and target. The back end (also known as the code generator) then maps the code onto the target instruction set. In addition to making *correct* code, it is responsible for generating *good* code that takes advantage of unusual features of the supported architecture. Common parts of a compiler back end include instruction selection, register allocation, and instruction scheduling.

This model applies equally well to interpreters and JIT compilers. The Java Virtual Machine (JVM) is also an implementation of this model, which uses Java bytecode as the interface between the front end and optimizer.

Implications of this Design

The most important win of this classical design comes when a compiler decides to support multiple source languages or target architectures. If the compiler uses a common code representation in its optimizer, then a front end can be written for any language that can compile to it, and a back end can be written for any target that can compile from it, as shown in 図 11.2.

With this design, porting the compiler to support a new source language (e.g., Algol or BASIC) requires implementing a new front end, but the existing optimizer and back end can be reused. If

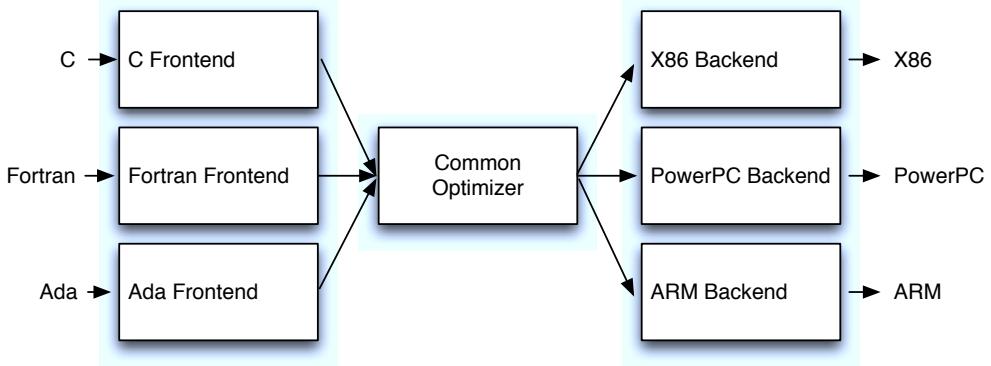


図 11.2: Retargetability

these parts weren't separated, implementing a new source language would require starting over from scratch, so supporting N targets and M source languages would need $N \times M$ compilers.

Another advantage of the three-phase design (which follows directly from retargetability) is that the compiler serves a broader set of programmers than it would if it only supported one source language and one target. For an open source project, this means that there is a larger community of potential contributors to draw from, which naturally leads to more enhancements and improvements to the compiler. This is the reason why open source compilers that serve many communities (like GCC) tend to generate better optimized machine code than narrower compilers like FreePASCAL. This isn't the case for proprietary compilers, whose quality is directly related to the project's budget. For example, the Intel ICC Compiler is widely known for the quality of code it generates, even though it serves a narrow audience.

A final major win of the three-phase design is that the skills required to implement a front end are different than those required for the optimizer and back end. Separating these makes it easier for a "front-end person" to enhance and maintain their part of the compiler. While this is a social issue, not a technical one, it matters a lot in practice, particularly for open source projects that want to reduce the barrier to contributing as much as possible.

11.2 Existing Language Implementations

While the benefits of a three-phase design are compelling and well-documented in compiler textbooks, in practice it is almost never fully realized. Looking across open source language implementations (back when LLVM was started), you'd find that the implementations of Perl, Python, Ruby and Java share no code. Further, projects like the Glasgow Haskell Compiler (GHC) and FreeBASIC are retargetable to multiple different CPUs, but their implementations are very specific to the one source language they support. There is also a broad variety of special purpose compiler tech-

nology deployed to implement JIT compilers for image processing, regular expressions, graphics card drivers, and other subdomains that require CPU intensive work.

That said, there are three major success stories for this model, the first of which are the Java and .NET virtual machines. These systems provide a JIT compiler, runtime support, and a very well defined bytecode format. This means that any language that can compile to the bytecode format (and there are dozens of them³) can take advantage of the effort put into the optimizer and JIT as well as the runtime. The tradeoff is that these implementations provide little flexibility in the choice of runtime: they both effectively force JIT compilation, garbage collection, and the use of a very particular object model. This leads to suboptimal performance when compiling languages that don't match this model closely, such as C (e.g., with the LLJVM project).

A second success story is perhaps the most unfortunate, but also most popular way to reuse compiler technology: translate the input source to C code (or some other language) and send it through existing C compilers. This allows reuse of the optimizer and code generator, gives good flexibility, control over the runtime, and is really easy for front-end implementers to understand, implement, and maintain. Unfortunately, doing this prevents efficient implementation of exception handling, provides a poor debugging experience, slows down compilation, and can be problematic for languages that require guaranteed tail calls (or other features not supported by C).

A final successful implementation of this model is GCC⁴. GCC supports many front ends and back ends, and has an active and broad community of contributors. GCC has a long history of being a C compiler that supports multiple targets with hacky support for a few other languages bolted onto it. As the years go by, the GCC community is slowly evolving a cleaner design. As of GCC 4.4, it has a new representation for the optimizer (known as "GIMPLE Tuples") which is closer to being separate from the front-end representation than before. Also, its Fortran and Ada front ends use a clean AST.

While very successful, these three approaches have strong limitations to what they can be used for, because they are designed as monolithic applications. As one example, it is not realistically possible to embed GCC into other applications, to use GCC as a runtime/JIT compiler, or extract and reuse pieces of GCC without pulling in most of the compiler. People who have wanted to use GCC's C++ front end for documentation generation, code indexing, refactoring, and static analysis tools have had to use GCC as a monolithic application that emits interesting information as XML, or write plugins to inject foreign code into the GCC process.

There are multiple reasons why pieces of GCC cannot be reused as libraries, including rampant use of global variables, weakly enforced invariants, poorly-designed data structures, sprawling code base, and the use of macros that prevent the codebase from being compiled to support more than one front-end/target pair at a time. The hardest problems to fix, though, are the inherent architectural problems that stem from its early design and age. Specifically, GCC suffers from layering problems

³http://en.wikipedia.org/wiki/List_of_JVM_languages

⁴A backronym that now stands for "GNU Compiler Collection".

and leaky abstractions: the back end walks front-end ASTs to generate debug info, the front ends generate back-end data structures, and the entire compiler depends on global data structures set up by the command line interface.

11.3 LLVM's Code Representation: LLVM IR

With the historical background and context out of the way, let's dive into LLVM: The most important aspect of its design is the LLVM Intermediate Representation (IR), which is the form it uses to represent code in the compiler. LLVM IR is designed to host mid-level analyses and transformations that you find in the optimizer section of a compiler. It was designed with many specific goals in mind, including supporting lightweight runtime optimizations, cross-function/interprocedural optimizations, whole program analysis, and aggressive restructuring transformations, etc. The most important aspect of it, though, is that it is itself defined as a first class language with well-defined semantics. To make this concrete, here is a simple example of a .ll file:

```
define i32 @add1(i32 %a, i32 %b) {
entry:
    %tmp1 = add i32 %a, %b
    ret i32 %tmp1
}

define i32 @add2(i32 %a, i32 %b) {
entry:
    %tmp1 = icmp eq i32 %a, 0
    br i1 %tmp1, label %done, label %recurse

recurse:
    %tmp2 = sub i32 %a, 1
    %tmp3 = add i32 %b, 1
    %tmp4 = call i32 @add2(i32 %tmp2, i32 %tmp3)
    ret i32 %tmp4

done:
    ret i32 %b
}
```

This LLVM IR corresponds to this C code, which provides two different ways to add integers:

```
unsigned add1(unsigned a, unsigned b) {
    return a+b;
}

// Perhaps not the most efficient way to add two numbers.
unsigned add2(unsigned a, unsigned b) {
    if (a == 0) return b;
    return add2(a-1, b+1);
}
```

As you can see from this example, LLVM IR is a low-level RISC-like virtual instruction set. Like a real RISC instruction set, it supports linear sequences of simple instructions like add, subtract, compare, and branch. These instructions are in three address form, which means that they take some number of inputs and produce a result in a different register.⁵ LLVM IR supports labels and generally looks like a weird form of assembly language.

Unlike most RISC instruction sets, LLVM is strongly typed with a simple type system (e.g., `i32` is a 32-bit integer, `i32**` is a pointer to pointer to 32-bit integer) and some details of the machine are abstracted away. For example, the calling convention is abstracted through `call` and `ret` instructions and explicit arguments. Another significant difference from machine code is that the LLVM IR doesn't use a fixed set of named registers, it uses an infinite set of temporaries named with a `%` character.

Beyond being implemented as a language, LLVM IR is actually defined in three isomorphic forms: the textual format above, an in-memory data structure inspected and modified by optimizations themselves, and an efficient and dense on-disk binary “bitcode” format. The LLVM Project also provides tools to convert the on-disk format from text to binary: `llvm-as` assembles the textual `.ll` file into a `.bc` file containing the bitcode goop and `llvm-dis` turns a `.bc` file into a `.ll` file.

The intermediate representation of a compiler is interesting because it can be a “perfect world” for the compiler optimizer: unlike the front end and back end of the compiler, the optimizer isn't constrained by either a specific source language or a specific target machine. On the other hand, it has to serve both well: it has to be designed to be easy for a front end to generate and be expressive enough to allow important optimizations to be performed for real targets.

Writing an LLVM IR Optimization

To give some intuition for how optimizations work, it is useful to walk through some examples. There are lots of different kinds of compiler optimizations, so it is hard to provide a recipe for how to solve an arbitrary problem. That said, most optimizations follow a simple three-part structure:

- Look for a pattern to be transformed.
- Verify that the transformation is safe/correct for the matched instance.
- Do the transformation, updating the code.

The most trivial optimization is pattern matching on arithmetic identities, such as: for any integer `X`, `X-X` is 0, `X-0` is `X`, `(X*2)-X` is `X`. The first question is what these look like in LLVM IR. Some examples are:

```
...
%example1 = sub i32 %a, %a
```

⁵This is in contrast to a two-address instruction set, like X86, which destructively updates an input register, or one-address machines which take one explicit operand and operate on an accumulator or the top of the stack on a stack machine.

```

...
%example2 = sub i32 %b, 0
...
%tmp = mul i32 %c, 2
%example3 = sub i32 %tmp, %c
...

```

For these sorts of “peephole” transformations, LLVM provides an instruction simplification interface that is used as utilities by various other higher level transformations. These particular transformations are in the `SimplifySubInst` function and look like this:

```

// X - 0 -> X
if (match(Op1, m_Zero()))
    return Op0;

// X - X -> 0
if (Op0 == Op1)
    return Constant::getNullValue(Op0->getType());

// (X*2) - X -> X
if (match(Op0, m_Mul(m_Specific(Op1), m_ConstantInt<2>())))
    return Op1;

...
return 0; // Nothing matched, return null to indicate no transformation.

```

In this code, `Op0` and `Op1` are bound to the left and right operands of an integer subtract instruction (importantly, these identities don’t necessarily hold for IEEE floating point!). LLVM is implemented in C++, which isn’t well known for its pattern matching capabilities (compared to functional languages like Objective Caml), but it does offer a very general template system that allows us to implement something similar. The `match` function and the `m_` functions allow us to perform declarative pattern matching operations on LLVM IR code. For example, the `m_Specific` predicate only matches if the left hand side of the multiplication is the same as `Op1`.

Together, these three cases are all pattern matched and the function returns the replacement if it can, or a null pointer if no replacement is possible. The caller of this function (`SimplifyInstruction`) is a dispatcher that does a switch on the instruction opcode, dispatching to the per-opcode helper functions. It is called from various optimizations. A simple driver looks like this:

```

for (BasicBlock::iterator I = BB->begin(), E = BB->end(); I != E; ++I)
    if (Value *V = SimplifyInstruction(I))
        I->replaceAllUsesWith(V);

```

This code simply loops over each instruction in a block, checking to see if any of them simplify. If so (because `SimplifyInstruction` returns non-null), it uses the `replaceAllUsesWith` method to update anything in the code using the simplifiable operation with the simpler form.

11.4 LLVM’s Implementation of Three-Phase Design

In an LLVM-based compiler, a front end is responsible for parsing, validating and diagnosing errors in the input code, then translating the parsed code into LLVM IR (usually, but not always, by building an AST and then converting the AST to LLVM IR). This IR is optionally fed through a series of analysis and optimization passes which improve the code, then is sent into a code generator to produce native machine code, as shown in 図 11.3. This is a very straightforward implementation of the three-phase design, but this simple description glosses over some of the power and flexibility that the LLVM architecture derives from LLVM IR.

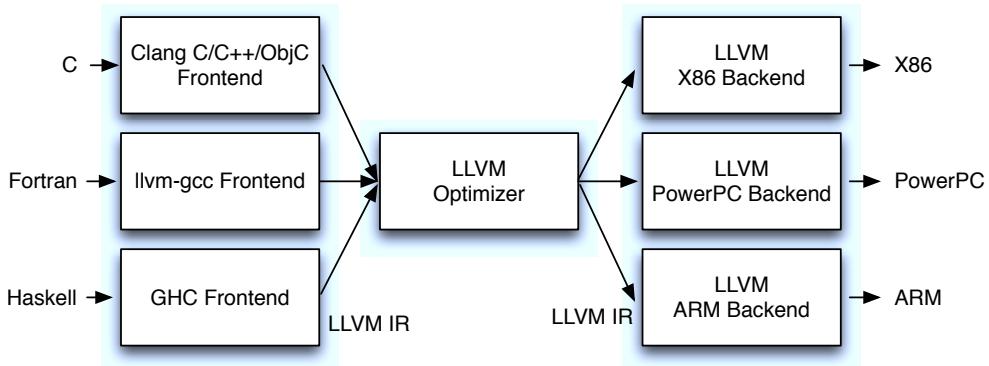


図 11.3: LLVM’s Implementation of the Three-Phase Design

LLVM IR is a Complete Code Representation

In particular, LLVM IR is both well specified and the *only* interface to the optimizer. This property means that all you need to know to write a front end for LLVM is what LLVM IR is, how it works, and the invariants it expects. Since LLVM IR has a first-class textual form, it is both possible and reasonable to build a front end that outputs LLVM IR as text, then uses Unix pipes to send it through the optimizer sequence and code generator of your choice.

It might be surprising, but this is actually a pretty novel property to LLVM and one of the major reasons for its success in a broad range of different applications. Even the widely successful and relatively well-architected GCC compiler does not have this property: its GIMPLE mid-level representation is not a self-contained representation. As a simple example, when the GCC code generator goes to emit DWARF debug information, it reaches back and walks the source level “tree” form. GIMPLE itself uses a “tuple” representation for the operations in the code, but (at least as of GCC 4.5) still represents operands as references back to the source level tree form.

The implications of this are that front-end authors need to know and produce GCC's tree data structures as well as GIMPLE to write a GCC front end. The GCC back end has similar problems, so they also need to know bits and pieces of how the RTL back end works as well. Finally, GCC doesn't have a way to dump out "everything representing my code", or a way to read and write GIMPLE (and the related data structures that form the representation of the code) in text form. The result is that it is relatively hard to experiment with GCC, and therefore it has relatively few front ends.

LLVM is a Collection of Libraries

After the design of LLVM IR, the next most important aspect of LLVM is that it is designed as a set of libraries, rather than as a monolithic command line compiler like GCC or an opaque virtual machine like the JVM or .NET virtual machines. LLVM is an infrastructure, a collection of useful compiler technology that can be brought to bear on specific problems (like building a C compiler, or an optimizer in a special effects pipeline). While one of its most powerful features, it is also one of its least understood design points.

Let's look at the design of the optimizer as an example: it reads LLVM IR in, chews on it a bit, then emits LLVM IR which hopefully will execute faster. In LLVM (as in many other compilers) the optimizer is organized as a pipeline of distinct optimization passes each of which is run on the input and has a chance to do something. Common examples of passes are the inliner (which substitutes the body of a function into call sites), expression reassociation, loop invariant code motion, etc. Depending on the optimization level, different passes are run: for example at -O0 (no optimization) the Clang compiler runs no passes, at -O3 it runs a series of 67 passes in its optimizer (as of LLVM 2.8).

Each LLVM pass is written as a C++ class that derives (indirectly) from the Pass class. Most passes are written in a single .cpp file, and their subclass of the Pass class is defined in an anonymous namespace (which makes it completely private to the defining file). In order for the pass to be useful, code outside the file has to be able to get it, so a single function (to create the pass) is exported from the file. Here is a slightly simplified example of a pass to make things concrete.⁶

```
namespace {
    class Hello : public FunctionPass {
public:
    // Print out the names of functions in the LLVM IR being optimized.
    virtual bool runOnFunction(Function &F) {
        cerr << "Hello: " << F.getName() << "\n";
        return false;
    }
};
```

⁶For all the details, please see *Writing an LLVM Pass manual* at <http://llvm.org/docs/WritingAnLLVMPass.html>.

```
FunctionPass *createHelloPass() { return new Hello(); }
```

As mentioned, the LLVM optimizer provides dozens of different passes, each of which are written in a similar style. These passes are compiled into one or more .o files, which are then built into a series of archive libraries (.a files on Unix systems). These libraries provide all sorts of analysis and transformation capabilities, and the passes are as loosely coupled as possible: they are expected to stand on their own, or explicitly declare their dependencies among other passes if they depend on some other analysis to do their job. When given a series of passes to run, the LLVM PassManager uses the explicit dependency information to satisfy these dependencies and optimize the execution of passes.

Libraries and abstract capabilities are great, but they don't actually solve problems. The interesting bit comes when someone wants to build a new tool that can benefit from compiler technology, perhaps a JIT compiler for an image processing language. The implementer of this JIT compiler has a set of constraints in mind: for example, perhaps the image processing language is highly sensitive to compile-time latency and has some idiomatic language properties that are important to optimize away for performance reasons.

The library-based design of the LLVM optimizer allows our implementer to pick and choose both the order in which passes execute, and which ones make sense for the image processing domain: if everything is defined as a single big function, it doesn't make sense to waste time on inlining. If there are few pointers, alias analysis and memory optimization aren't worth bothering about. However, despite our best efforts, LLVM doesn't magically solve all optimization problems! Since the pass subsystem is modularized and the PassManager itself doesn't know anything about the internals of the passes, the implementer is free to implement their own language-specific passes to cover for deficiencies in the LLVM optimizer or to explicit language-specific optimization opportunities. 囗

11.4 shows a simple example for our hypothetical XYZ image processing system:

Once the set of optimizations is chosen (and similar decisions are made for the code generator) the image processing compiler is built into an executable or dynamic library. Since the only reference to the LLVM optimization passes is the simple create function defined in each .o file, and since the optimizers live in .a archive libraries, only the optimization passes *that are actually used* are linked into the end application, not the entire LLVM optimizer. In our example above, since there is a reference to PassA and PassB, they will get linked in. Since PassB uses PassD to do some analysis, PassD gets linked in. However, since PassC (and dozens of other optimizations) aren't used, its code isn't linked into the image processing application.

This is where the power of the library-based design of LLVM comes into play. This straightforward design approach allows LLVM to provide a vast amount of capability, some of which may only be useful to specific audiences, without punishing clients of the libraries that just want to do simple things. In contrast, traditional compiler optimizers are built as a tightly interconnected mass of code, which is much more difficult to subset, reason about, and come up to speed on. With LLVM

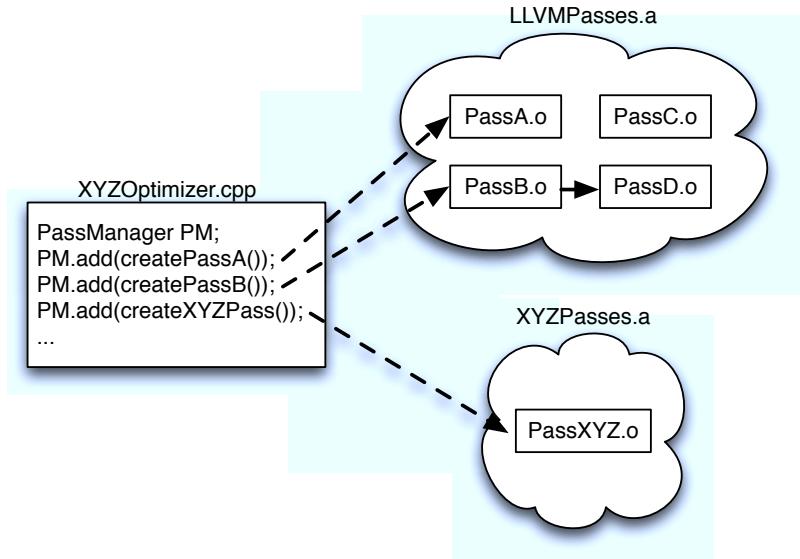


図 11.4: Hypothetical XYZ System using LLVM

you can understand individual optimizers without knowing how the whole system fits together.

This library-based design is also the reason why so many people misunderstand what LLVM is all about: the LLVM libraries have many capabilities, but they don't actually *do* anything by themselves. It is up to the designer of the client of the libraries (e.g., the Clang C compiler) to decide how to put the pieces to best use. This careful layering, factoring, and focus on subset-ability is also why the LLVM optimizer can be used for such a broad range of different applications in different contexts. Also, just because LLVM provides JIT compilation capabilities, it doesn't mean that every client uses it.

11.5 Design of the Retargetable LLVM Code Generator

The LLVM code generator is responsible for transforming LLVM IR into target specific machine code. On the one hand, it is the code generator's job to produce the best possible machine code for any given target. Ideally, each code generator should be completely custom code for the target, but on the other hand, the code generators for each target need to solve very similar problems. For example, each target needs to assign values to registers, and though each target has different register files, the algorithms used should be shared wherever possible.

Similar to the approach in the optimizer, LLVM's code generator splits the code generation problem into individual passes—instruction selection, register allocation, scheduling, code layout optimization, and assembly emission—and provides many builtin passes that are run by default. The

target author is then given the opportunity to choose among the default passes, override the defaults and implement completely custom target-specific passes as required. For example, the x86 back end uses a register-pressure-reducing scheduler since it has very few registers, but the PowerPC back end uses a latency optimizing scheduler since it has many of them. The x86 back end uses a custom pass to handle the x87 floating point stack, and the ARM back end uses a custom pass to place constant pool islands inside functions where needed. This flexibility allows target authors to produce great code without having to write an entire code generator from scratch for their target.

LLVM Target Description Files

The “mix and match” approach allows target authors to choose what makes sense for their architecture and permits a large amount of code reuse across different targets. This brings up another challenge: each shared component needs to be able to reason about target specific properties in a generic way. For example, a shared register allocator needs to know the register file of each target and the constraints that exist between instructions and their register operands. LLVM’s solution to this is for each target to provide a target description in a declarative domain-specific language (a set of .td files) processed by the `tblgen` tool. The (simplified) build process for the x86 target is shown in 図 11.5.

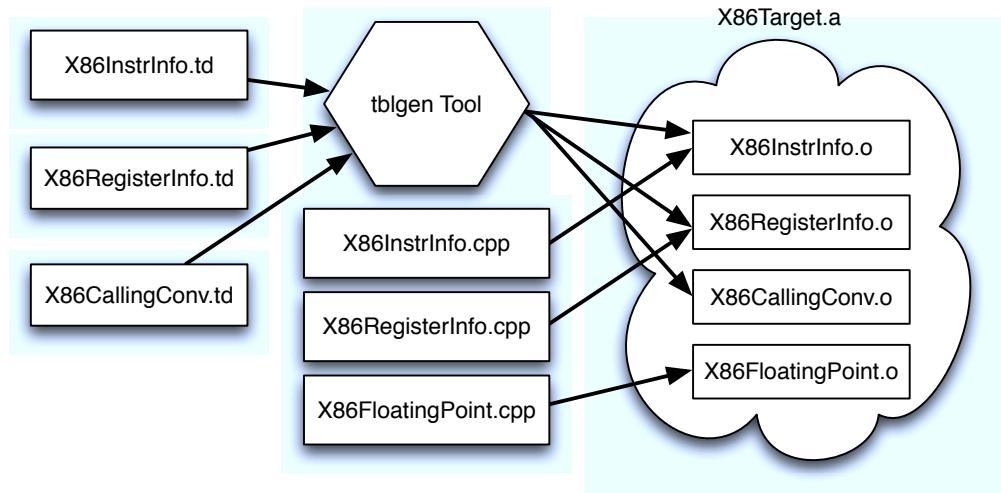


図 11.5: Simplified x86 Target Definition

The different subsystems supported by the .td files allow target authors to build up the different pieces of their target. For example, the x86 back end defines a register class that holds all of its 32-bit registers named “GR32” (in the .td files, target specific definitions are all caps) like this:

```
def GR32 : RegisterClass<[i32], 32,
```

```
[EAX, ECX, EDX, ESI, EDI, EBX, EBP, ESP,
R8D, R9D, R10D, R11D, R14D, R15D, R12D, R13D]> { ... }
```

This definition says that registers in this class can hold 32-bit integer values (“i32”), prefer to be 32-bit aligned, have the specified 16 registers (which are defined elsewhere in the .td files) and have some more information to specify preferred allocation order and other things. Given this definition, specific instructions can refer to this, using it as an operand. For example, the “complement a 32-bit register” instruction is defined as:

```
let Constraints = "$src = $dst" in
def NOT32r : I<0xF7, MRM2r,
  (outs GR32:$dst), (ins GR32:$src),
  "not{l}\t$dst",
  [(set GR32:$dst, (not GR32:$src))];
```

This definition says that NOT32r is an instruction (it uses the I tblgen class), specifies encoding information (0xF7, MRM2r), specifies that it defines an “output” 32-bit register \$dst and has a 32-bit register “input” named \$src (the GR32 register class defined above defines which registers are valid for the operand), specifies the assembly syntax for the instruction (using the {} syntax to handle both AT&T and Intel syntax), specifies the effect of the instruction and provides the pattern that it should match on the last line. The “let” constraint on the first line tells the register allocator that the input and output register must be allocated to the same physical register.

This definition is a very dense description of the instruction, and the common LLVM code can do a lot with information derived from it (by the tblgen tool). This one definition is enough for instruction selection to form this instruction by pattern matching on the input IR code for the compiler. It also tells the register allocator how to process it, is enough to encode and decode the instruction to machine code bytes, and is enough to parse and print the instruction in a textual form. These capabilities allow the x86 target to support generating a stand-alone x86 assembler (which is a drop-in replacement for the “gas” GNU assembler) and disassemblers from the target description as well as handle encoding the instruction for the JIT.

In addition to providing useful functionality, having multiple pieces of information generated from the same “truth” is good for other reasons. This approach makes it almost infeasible for the assembler and disassembler to disagree with each other in either assembly syntax or in the binary encoding. It also makes the target description easily testable: instruction encodings can be unit tested without having to involve the entire code generator.

While we aim to get as much target information as possible into the .td files in a nice declarative form, we still don’t have everything. Instead, we require target authors to write some C++ code for various support routines and to implement any target specific passes they might need (like X86FloatingPoint.cpp, which handles the x87 floating point stack). As LLVM continues to grow new targets, it becomes more and more important to increase the amount of the target that can be

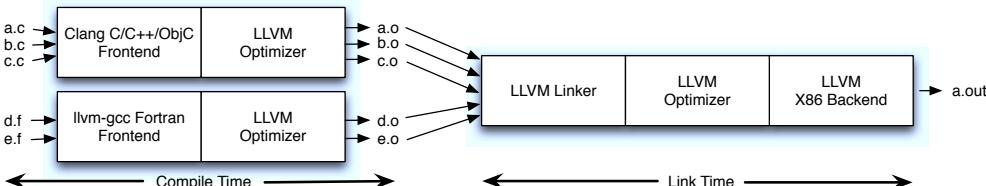
expressed in the .td file, and we continue to increase the expressiveness of the .td files to handle this. A great benefit is that it gets easier and easier write targets in LLVM as time goes on.

11.6 Interesting Capabilities Provided by a Modular Design

Besides being a generally elegant design, modularity provides clients of the LLVM libraries with several interesting capabilities. These capabilities stem from the fact that LLVM provides functionality, but lets the client decide most of the *policies* on how to use it.

Choosing When and Where Each Phase Runs

As mentioned earlier, LLVM IR can be efficiently (de)serialized to/from a binary format known as LLVM bitcode. Since LLVM IR is self-contained, and serialization is a lossless process, we can do part of compilation, save our progress to disk, then continue work at some point in the future. This feature provides a number of interesting capabilities including support for link-time and install-time optimization, both of which delay code generation from “compile time”.

Link-Time Optimization (LTO) addresses the problem where the compiler traditionally only sees one translation unit (e.g., a .c file with all its headers) at a time and therefore cannot do optimizations (like inlining) across file boundaries. LLVM compilers like Clang support this with the -fipa or -O4 command line option. This option instructs the compiler to emit LLVM bitcode to the .o file instead of writing out a native object file, and delays code generation to link time, shown in  11.6.

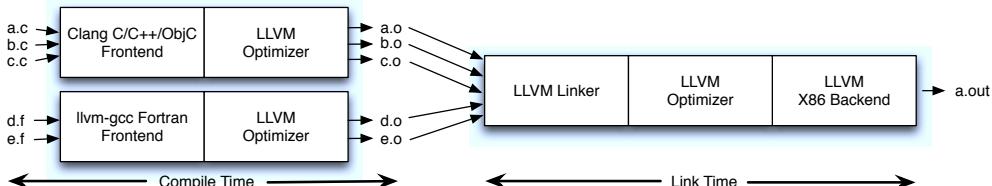


图 11.6: Link-Time Optimization

Details differ depending on which operating system you’re on, but the important bit is that the linker detects that it has LLVM bitcode in the .o files instead of native object files. When it sees this, it reads all the bitcode files into memory, links them together, then runs the LLVM optimizer over the aggregate. Since the optimizer can now see across a much larger portion of the code, it can inline, propagate constants, do more aggressive dead code elimination, and more across file boundaries. While many modern compilers support LTO, most of them (e.g., GCC, Open64, the Intel compiler, etc.) do so by having an expensive and slow serialization process. In LLVM, LTO

falls out naturally from the design of the system, and works across different source languages (unlike many other compilers) because the IR is truly source language neutral.

Install-time optimization is the idea of delaying code generation even later than link time, all the way to install time, as shown in 図 11.7. Install time is a very interesting time (in cases when software is shipped in a box, downloaded, uploaded to a mobile device, etc.), because this is when you find out the specifics of the device you're targeting. In the x86 family for example, there are broad variety of chips and characteristics. By delaying instruction choice, scheduling, and other aspects of code generation, you can pick the best answers for the specific hardware an application ends up running on.

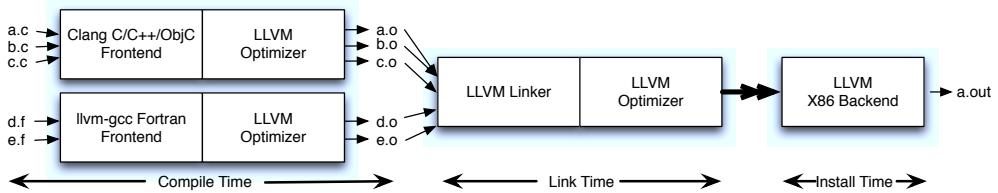


図 11.7: Install-Time Optimization

Unit Testing the Optimizer

Compilers are very complicated, and quality is important, therefore testing is critical. For example, after fixing a bug that caused a crash in an optimizer, a regression test should be added to make sure it doesn't happen again. The traditional approach to testing this is to write a .c file (for example) that is run through the compiler, and to have a test harness that verifies that the compiler doesn't crash. This is the approach used by the GCC test suite, for example.

The problem with this approach is that the compiler consists of many different subsystems and even many different passes in the optimizer, all of which have the opportunity to change what the input code looks like by the time it gets to the previously buggy code in question. If something changes in the front end or an earlier optimizer, a test case can easily fail to test what it is supposed to be testing.

By using the textual form of LLVM IR with the modular optimizer, the LLVM test suite has highly focused regression tests that can load LLVM IR from disk, run it through exactly one optimization pass, and verify the expected behavior. Beyond crashing, a more complicated behavioral test wants to verify that an optimization is actually performed. Here is a simple test case that checks to see that the constant propagation pass is working with add instructions:

```
; RUN: opt < %s -constprop -S | FileCheck %s
define i32 @test() {
  %A = add i32 4, 5
  ret i32 %A
```

```
; CHECK: @test()  
; CHECK: ret i32 9  
}
```

The RUN line specifies the command to execute: in this case, the opt and FileCheck command line tools. The opt program is a simple wrapper around the LLVM pass manager, which links in all the standard passes (and can dynamically load plugins containing other passes) and exposes them through to the command line. The FileCheck tool verifies that its standard input matches a series of CHECK directives. In this case, this simple test is verifying that the constprop pass is folding the add of 4 and 5 into 9.

While this might seem like a really trivial example, this is very difficult to test by writing .c files: front ends often do constant folding as they parse, so it is very difficult and fragile to write code that makes its way downstream to a constant folding optimization pass. Because we can load LLVM IR as text and send it through the specific optimization pass we're interested in, then dump out the result as another text file, it is really straightforward to test exactly what we want, both for regression and feature tests.

Automatic Test Case Reduction with BugPoint

When a bug is found in a compiler or other client of the LLVM libraries, the first step to fixing it is to get a test case that reproduces the problem. Once you have a test case, it is best to minimize it to the smallest example that reproduces the problem, and also narrow it down to the part of LLVM where the problem happens, such as the optimization pass at fault. While you eventually learn how to do this, the process is tedious, manual, and particularly painful for cases where the compiler generates incorrect code but does not crash.

The LLVM BugPoint tool⁷ uses the IR serialization and modular design of LLVM to automate this process. For example, given an input .ll or .bc file along with a list of optimization passes that causes an optimizer crash, BugPoint reduces the input to a small test case and determines which optimizer is at fault. It then outputs the reduced test case and the opt command used to reproduce the failure. It finds this by using techniques similar to “delta debugging” to reduce the input and the optimizer pass list. Because it knows the structure of LLVM IR, BugPoint does not waste time generating invalid IR to input to the optimizer, unlike the standard “delta” command line tool.

In the more complex case of a miscompilation, you can specify the input, code generator information, the command line to pass to the executable, and a reference output. BugPoint will first determine if the problem is due to an optimizer or a code generator, and will then repeatedly partition the test case into two pieces: one that is sent into the “known good” component and one that is sent into the “known buggy” component. By iteratively moving more and more code out of the partition that is sent into the known buggy code generator, it reduces the test case.

⁷<http://llvm.org/docs/Bugpoint.html>

BugPoint is a very simple tool and has saved countless hours of test case reduction throughout the life of LLVM. No other open source compiler has a similarly powerful tool, because it relies on a well-defined intermediate representation. That said, BugPoint isn't perfect, and would benefit from a rewrite. It dates back to 2002, and is typically only improved when someone has a really tricky bug to track down that the existing tool doesn't handle well. It has grown over time, accreting new features (such as JIT debugging) without a consistent design or owner.

11.7 Retrospective and Future Directions

LLVM's modularity wasn't originally designed to directly achieve any of the goals described here. It was a self-defense mechanism: it was obvious that we wouldn't get everything right on the first try. The modular pass pipeline, for example, exists to make it easier to isolate passes so that they can be discarded after being replaced by better implementations⁸.

Another major aspect of LLVM remaining nimble (and a controversial topic with clients of the libraries) is our willingness to reconsider previous decisions and make widespread changes to APIs without worrying about backwards compatibility. Invasive changes to LLVM IR itself, for example, require updating all of the optimization passes and cause substantial churn to the C++ APIs. We've done this on several occasions, and though it causes pain for clients, it is the right thing to do to maintain rapid forward progress. To make life easier for external clients (and to support bindings for other languages), we provide C wrappers for many popular APIs (which are intended to be extremely stable) and new versions of LLVM aim to continue reading old .ll and .bc files.

Looking forward, we would like to continue making LLVM more modular and easier to subset. For example, the code generator is still too monolithic: it isn't currently possible to subset LLVM based on features. For example, if you'd like to use the JIT, but have no need for inline assembly, exception handling, or debug information generation, it should be possible to build the code generator without linking in support for these features. We are also continuously improving the quality of code generated by the optimizer and code generator, adding IR features to better support new language and target constructs, and adding better support for performing high-level language-specific optimizations in LLVM.

The LLVM project continues to grow and improve in numerous ways. It is really exciting to see the number of different ways that LLVM is being used in other projects and how it keeps turning up in surprising new contexts that its designers never even thought about. The new LLDB debugger is a great example of this: it uses the C/C++/Objective-C parsers from Clang to parse expressions, uses the LLVM JIT to translate these into target code, uses the LLVM disassemblers, and uses LLVM targets to handle calling conventions among other things. Being able to reuse this existing code allows

⁸I often say that none of the subsystems in LLVM are really good until they have been rewritten at least once.

people developing debuggers to focus on writing the debugger logic, instead of reimplementing yet another (marginally correct) C++ parser.

Despite its success so far, there is still a lot left to be done, as well as the ever-present risk that LLVM will become less nimble and more calcified as it ages. While there is no magic answer to this problem, I hope that the continued exposure to new problem domains, a willingness to reevaluate previous decisions, and to redesign and throw away code will help. After all, the goal isn't to be perfect, it is to keep getting better over time.

Mercurial

Dirkjan Ochtman

Mercurial はモダンな分散型バージョン管理システム (VCS) で、大半は Python で書かれて いる。ただ、ごく一部、パフォーマンスのために C で書いているところもある。本章では、 Mercurial のアルゴリズムやデータ構造を設計したときのいくつかの判断について説明する。 まず最初に、バージョン管理システムの歴史を簡単に振り返っておこう。これが、本章を読 み進めるための前提知識となる。

12.1 バージョン管理システムの簡単な歴史

本章で主に扱うのは Mercurial のアーキテクチャについてである。しかし、その概念の多くは他のバージョン管理システムと共通している。Mercurial についての議論を実のあるものにするために、まずはさまざまなバージョン管理システムの概念や動作に名前をつけるところからはじめよう。さらに、それらすべてを見渡すために、この世界の歴史についても簡単 に説明する。

バージョン管理システムが作られたのは、複数人によるソフトウェアシステムの開発作業 を支援するためだった。ソースの完全なコピーをやりとりして変更履歴を各自で管理させるなどということをしなくて済むように、と作られたのだ。ここで、単なるソフトウェアのソースコードだけではなく、任意のファイルツリーに一般化して考えよう。バージョン管理シス テムの主要機能のひとつは、変更をツリーに渡すことだ。基本的な流れは、このようになる。

1. 最新のファイルツリーを、どこか別の場所から取得する
2. このバージョンのツリー上で、何らかの変更を加える
3. 変更内容をどこかに公開し、他の人がそれを取得できるようにする

最初の操作、つまりファイルツリーをローカルに取得する作業のことを **チェックアウト** と呼ぶ。データの取得元であり、かつ変更点の公開先でもある格納場所のことは **リポジトリ** と呼び、チェックアウトしたものることは **作業ディレクトリ** あるいは **作業ツリー**、**作業コピー** な

どと呼ぶ。作業コピーの内容をリポジトリ上の最新の状態に更新することは、単に**更新**と呼ぶ。このとき、場合によっては**マージ**が必要になることもある。マージとは、同一ファイル上での別のユーザーによる変更をとりまとめることだ。`diff` コマンドを使うと、ツリーあるいはファイルについて複数のリビジョン間での変更点を確認できる。最もよくある使いかたは、作業コピー上にあるローカルの(まだ公開していない)変更を確認することだ。変更を公開するには `commit` コマンドを実行する。これは、作業ディレクトリ上での変更をリポジトリに記録する。

中央集中型バージョン管理

バージョン管理システムの元祖は Source Code Control System だ。略して SCCS と呼ばれるこのシステムが最初に登場したのは 1975 年のことだった。差分をひとつのファイルに記録するという方式でバージョンを管理しており、各バージョンのコピーを取っておくよりも効率的だった。その変更を他人に公開する仕組みは用意されていなかった。その後 1982 年に登場した Revision Control System(略して RCS) は、さらに進化しており、SCCS の代替として使えるフリーソフトウェアだった(RCS は、今でも GNU プロジェクトが保守を続けている)。

RCS に続いて登場したのが CVS。これは Concurrent Versioning System の略で、1986 年にリリースされた。当初は、RCS のリビジョンファイルを操作するためのスクリプト群という形式だった。CVS における大きなイノベーションは、複数のユーザーによる同時編集と、その後での編集のマージという概念だった。ここで、編集時の衝突という概念も登場する。開発者が何らかのファイルの新しいバージョンをコミットするためには、手元のファイルをリポジトリ内の最新版に基づいたものにしておく必要があった。リポジトリ内のファイルと作業ディレクトリ上のファイルの両方で(同じ行の)変更があった場合は、両者の変更の衝突を解消しなければならなかった。

CVS は、**ブランチ**や**タグ**という概念を初めて取り入れたシステムでもある。ブランチのおかげで別々の作業を並行して行えるようになり、タグのおかげでスナップショットに名前をつけて容易に参照できるようになった。CVS の差分のやりとりは、当初は共有ファイルシステム上のリポジトリを使って行っていた。その後、CVS もクライアントサーバー型のアーキテクチャを採用し、(インターネットのような)大規模ネットワーク上でも使えるようになった。

2000 年に、3 人の開発者が新しい VCS の開発をはじめた。それは Subversion と名付けられ、CVS の大きな問題点を修正することを目標として開発を進めた。最も重要なのは、Subversion がツリー全体を一括して扱うことだった。つまり、リビジョン間での変更はアトミックであり、一貫性があつてそれぞれ隔離されており、永続するということだ。Subversion の作業コピーには、作業ディレクトリのリビジョンをチェックアウトした当初の状態が残っている。つまり、よくある `diff` 操作(ローカルのツリーをチェックアウトしたときの状態と比べる作業)がローカルで高速に実行できるということだ。

Subversion の興味深い概念のひとつが、タグやブランチもプロジェクトのツリーの一部とみなしたことである。Subversion のプロジェクトは、tags、branches そして trunk の三つ

のエリアに分かれていることが多い。この設計は、バージョン管理システムに不慣れなユーザーにとっては非常にわかりやすかった。しかし、この設計による柔軟性は、変換ツールを作る側にとってはさまざまな問題のもととなつた。他のシステムでは、タグやブランチはそれ専用の構造で表すことが多かつたからである。

ここまで取り上げたすべてのシステムは、いわゆる**中央集中型**である。CVS 以降のツールでは変更点をやりとりする方法を知っているが、それは他のコンピュータが変更の履歴を保持しているということに依存している。**分散型**のバージョン管理システムは、リポジトリの履歴のすべて(あるいは大半)を各コンピュータに保持しており、そのリポジトリの作業ディレクトリも保持している。

分散型バージョン管理

Subversion は CVS よりもずっときれいな実装だったが、それでも弱点はいくつか残っていた。たとえば、これは中央集中型のバージョン管理システムに共通する問題だが、変更をコミットすると同時に事実上その変更が公開されてしまうという点もそのひとつだ。リポジトリの履歴が中央で一元管理されている以上、どうしてもそうなってしまう。これはつまり、ネットワークに接続できない環境ではコミットが不可能であるということを意味する。第二に、中央集中型のシステムでリポジトリにアクセスする際には、最低一度はネットワーク上のやりとりが発生する。そのため、分散型システムの場合のローカルアクセスに比べて、比較的動作が遅くなってしまう。第三に、これまで取り上げたシステムは、どれもマージが下手である(中にはいくらかましになったものもあるが)。大規模なグループで並行作業をしていると、新しいバージョンがバージョン管理システムのどの変更を含むものかを把握することが重要となる。とりこぼしがないかを確認し、その後のマージをうまく行うためにこの情報を使うことが必要となる。第四に、昔ながらの VCS が求める中央管理は、時に人工的に見えることもある。統合のための場所を一ヵ所用意することになる。分散型の VCS を支持する人々は、分散型のシステムのほうがより有機的な組織に対応できると言う。各開発者による変更のプッシュや統合を、必要に応じてその場で行えるというわけだ。

これらの問題を解決しようと、新たなツールがいろいろ登場した。私の周囲の世界(オープンソースの世界)を見る限り、2011 年の時点での三大巨頭は Git と Mercurial そして Bazaar である。Git と Mercurial は、どちらも 2005 年に開発が始まった。Linux カーネルの開発者が、プロプライエタリな BitKeeper を使わないようにすると決断したところである。どちらのツールも Linux カーネルの開発者が最初に作り始めた(Git は Linus Torvalds、そして Mercurial は Matt Mackall だ)。何万ものファイル上で繰り広げられる何十万ものチェンジセット(そう、たとえば Linux カーネルみたいなもの)もきちんと扱えるように作られている。Matt も Linus も、Monotone VCS の影響を強く受けていた。Bazaar の開発はそれとは別に進んでいたが、Git や Mercurial と同時期に広く使われるようになった。Canonical がすべてのプロジェクトを Bazaar に移行したのがちょうどこの時期である。

分散型バージョン管理システムの構築にはいろいろ困難もあった。その多くは、分散型システムという性質によるものだった。一例を挙げよう。中央集中型のシステムにおけるソース管理サーバーは常に本流の履歴を提供できるが、分散型の VCS にはそもそも「本流」の履歴など存在しない。チェンジセットは各地で並行にコミットできるので、特定のリポジトリ上で一時的にリビジョンを並べることは不可能だ。

この問題に対して一般的に広く受け入れられている解決策は、チェンジセットを一列に並べるのではなくチェンジセットの無閉路有向グラフ (directed acyclic graph: DAG) を使うという方法だ(図 12.1)。つまり、新たにコミットしたチェンジセットは別のリビジョンの子リビジョンとなり、自分自身あるいは自分の子孫の子リビジョンになることはできないという構造である。この考え方を採用するために、三種類の特殊なリビジョンを用意した。親を持たないリビジョンである **root リビジョン**(ひとつのリポジトリに複数の root があつてもかまわない)、複数の親を持つ**マージリビジョン**、そして子を持たない **head リビジョン**である。各リポジトリは、まず空の root リビジョンがひとつだけある状態から始まる。それを起点としてチェンジセットが連なり、最終的にひとつあるいは複数の head を持つことになる。二人のユーザーがそれぞれ別のコミットをしたときに、一方がもう一方のコミットを取り込みたくなったとしよう。そんな場合は、別のユーザーの変更を新しいリビジョンに明示的にマージしなければならない。この新しいリビジョンを、マージリビジョンとしてコミットする。

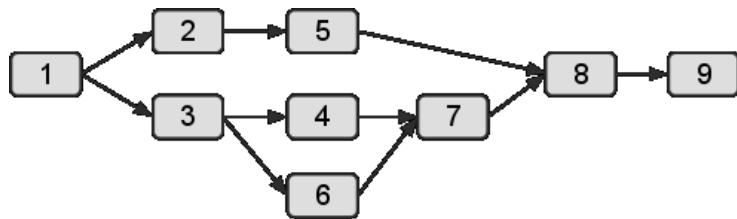


図 12.1: 各リビジョンの無閉路有向グラフ

DAG モデルを使うと、中央集中型のバージョン管理システムでは解決が困難な問題を解決する手助けになる。つまり、マージリビジョンを使えば、新たに DAG にマージされたブランチの情報を記録できるということだ。その結果としてできあがるグラフは並行するブランチ全体の大きなまとまりもうまく表せており、それをより小さなグループにマージしたり、最終的に「本流」ブランチにまとめたりもできる。

この手法を使うには、バージョン管理システムがチェンジセット間の親子関係を保持している必要がある。これを実現するため、チェンジセットのデータをやりとりする際には自分の親のチェンジセットを知っている状態にしておくことになる。当然、そのためにはチェンジセットに何らかの識別子が必要となる。システムによっては UUID やそれに類する仕組みを使った識別子を用意しているものもあるが、Git や Mercurial の場合はチェンジセットの内容の SHA1 ハッシュを使っている。この方式を使うと、チェンジセットの ID を使ってチェンジセットの中身が改ざんされていないことを検証できるという利点もある。実際、親の情報

もハッシュされたデータに含んでいるので、どのリビジョンからであってもそのハッシュだけで履歴の正当性を検証できる。コミットの作者名やコミットメッセージ、タイムスタンプ、他のメタデータは、新しいリビジョンの実際の内容をハッシュするのと同様にハッシュされる。つまり、メタデータも同様に検証できるということだ。また、タイムスタンプはコミット時に記録されるので、任意のリポジトリ内でリニアに進む必要もない。

これまでに中央集中型のVCSしか使ってこなかった人たちにとっては、こういった考え方はなかなかじめないものだ。わかりやすい番号でリビジョンを表すのではなく、単なる40文字の16進文字列を使うだって？さらに、リビジョンの並び順など存在しない。単に個々のローカルリビジョン内での並び順があるだけで、全体的な“並び順”は一直線ではなくDAGで表している。既に別の子チェンジセットを持つリビジョンに対して、新しいheadをコミットして開発を続けることもできる。慣れ親しんだ中央集中型のVCSなら、そんなことをすれば警告が出ていたことだろう。

幸いにも、ツリーの並び順を可視化してくれるツールも存在する。またMercurialでは、チェンジセットのハッシュの(曖昧さのない)短縮版も使えるし、さらに、ローカル限定ではあるが一連の番号でチェンジセットを識別することもできる。この一連の番号は単調増加する整数値で、クローンに投入されたチェンジセットの順番を表す。この順番はクローンごとに異なるので、リモートリポジトリに対する操作ではこの番号は使えない。

12.2 データ構造

DAGの概念がわかつてきたところで、それが実際どのようにMercurialに格納されているのかを見ていこう。DAGモデルはMercurialの内部動作の中核であり、いくつかの異なるDAGを使ってリポジトリをディスク上に格納している(コードのメモリ内の構造も同じだ)。このセクションではそれぞれのDAGについて説明し、さらにそれらがどのように連携するかも説明する。

課題

実際のデータ構造に関する話題の前に、Mercurialの成長の歴史について簡単にまとめておく。Mercurialのそもそもの始まりは、Matt Mackallが2005年4月20日にLinux Kernel メーリングリストに投稿した一通のメールだった。その何日か前に、カーネルの開発でBitKeeperを使わないようにするという決断があったころの話である。Mattはそのメールで、いくつかの目標を示した。シンプルであること、スケーラブルであること、そして効率的であること。

[Mac06]においてMattが指摘したのは次のようなことである。今時のVCSは何百万ものファイルを含むツリーを扱う必要がある。チェンジセットの数もそれと同じくらいだろうし、何千ものユーザーが並行して新しいリビジョンを作ったりという作業を何十年ものスパンで行うことになるだろう。Mercurialが目指す目標を設定するにあたって、彼は現在の技術的な制約についてもまとめた。

- 速度: CPU
- 容量: ディスクやメモリ
- 帯域: メモリ、LAN、ディスクそして WAN
- ディスクのシーク率

ディスクのシーク率や WAN の帯域は今でも制約となっており、最適化を要する。この論文ではさらに、この手のシステムのパフォーマンスをファイルレベルで評価する際の一般的なシナリオや制約についての考察が続く。

- ストレージの圧縮: ファイルの履歴をディスク上に保存するのに最適の圧縮方法は何か? どんなアルゴリズムを使えば、CPU 時間をボトルネックにしない範囲で I/O のパフォーマンスを最大化できるか?
- 任意のリビジョンのファイルの取得: 多くのバージョン管理システムが採用している格納方法だと、過去のリビジョンのファイルを大量に読み込まないと新しいリビジョンを再現できない(差分を格納している)。我々としては、それを改善しつつ過去のリビジョンも高速に取得できるようにしたい。
- ファイルのリビジョンの追加: 新たなリビジョンの追加は頻発する。新しいリビジョンを追加するたびに古いリビジョンを上書きするようなことは避けたい。そんなことをすれば、リビジョンが増えるにつれて速度が低下するからである。
- ファイルの履歴の表示: ある特定のファイルを変更したすべてのチェンジセットの歴史を見直せるようにしたい。そうすれば、アノテーション(あるファイルの個々の行が、どのチェンジセットで変更されたものかを確認すること)を行えるようになる(これは CVS の時代には `blame` と呼ばれていたが、その言葉の否定的な意味を排除するために、後のシステムでは `annotate` という名前に変わったものもある)。

この論文ではさらに、同様のシナリオについてプロジェクトレベルでの考察を続ける。プロジェクトレベルでの基本的な操作といえば、あるリビジョンのチェックアウトや新たなるリビジョンのコミット、そして作業ディレクトリとの差分の検出などである。差分の検出などは特に、ツリーが大きくなると速度が低下しがちだ(Mozilla や NetBeans プロジェクトを見るとよい。どちらも必要に迫られたバージョン管理に Mercurial を採用している)。

高速リビジョンストレージ: Revlog

Matt が Mercurial で用いた解決策が `revlog` (revision log を縮めたもの) だ。`revlog` 方式を使えば、各リビジョンのファイルの中身(そして、前のバージョンからの差分)を効率的に管理できる。アクセス時間を効率化(ディスクのシークを最適化)し、ストレージのスペースも効率的に使い、先に述べた一般的なシナリオをうまく扱えなければならない。そのために、`revlog` はディスク上ではふたつのファイルに分かれている。インデックスファイルとデータファイルだ。

6 バイト	ハングオフセット
2 バイト	フラグ
4 バイト	ハング長
4 バイト	非圧縮時の長さ
4 バイト	ベースリビジョン
4 バイト	リンクリビジョン
4 バイト	親リビジョン 1
4 バイト	親リビジョン 2
32 バイト	ハッシュ値

表 12.1: Mercurial のレコードフォーマット

インデックスは固定長のレコードで構成されており、その詳細は表 12.1 のとおりである。固定長のレコードを使う利点は、ローカルのリビジョン番号がわかればそのリビジョンにダイレクトアクセスできる(常に一定の時間でアクセスできる)ということだ。つまり、インデックスファイルの所定の位置(インデックス長 × リビジョン)を読めば、データの場所がわかる。インデックスをデータから分離したことで、インデックスデータの読み込みが高速化した。インデックスを読む際にファイルのデータをすべてシークする必要がなくなる。

ハングオフセットとハング長で指定するのは、読む込むデータファイルの量である。これを読めば、そのリビジョンの圧縮データを取得できる。元データを取得するには、まず最初にベースリビジョンを読み込んでから当該リビジョンまでの差分を適用していけばよい。ベースリビジョンをどのタイミングで更新するかについてはちょっと考慮した。このタイミングを決める基準は、累積した差分のサイズとそのリビジョンの非圧縮時のサイズの比較である(データの圧縮には zlib を使い、ディスクスペースを節約している)。差分群のサイズをこの方式で一定に制限することで、特定のリビジョンを再構築するときの手間をできるだけ軽減している。そのおかげで、大量の差分を適用するはめになることを回避しているのだ。

リンクリビジョンは、そのリビジョンが依存する revlog を最高レベルまでたどるために利用する(あとでもう少し詳しく説明する)。そして親リビジョンには、ローカルのリビジョン番号(整数値)が格納されている。これを使えば、関連する revlog のデータを見つけるのも容易になる。ハッシュに保存するのは、このエンジセットを指す一意な識別子だ。SHA1 ハッシュに必要なのは 20 バイトだが、ここでは 32 バイトを確保している。これは、将来の拡張性を考慮したものである。

三種類の revlog

revlog が履歴データに関する汎用的な構造を提供しているので、その上位レイヤーとしてファイルツリーを表すデータモデルを作成できる。このデータモデルは次の三種類の revlog で構成されている。*changelog*、*manifests* そして *filelogs* だ。*changelog* には各リビジョンのメ

タデータが含まれ、さらに manifest revlog へのポインタ (つまり、manifest revlog 内のあるリビジョンのノード id) も含まれている。また、manifest ファイルにはファイル名の一覧が記録されており、さらに各ファイルのノード id も含まれている。このノード id は、各ファイルの filelog 内のリビジョンを指すものである。Mercurial のコード内では、changelog や manifest そして filelog を表すクラスがそれぞれ用意されている。これらはどれも汎用の revlog クラスを継承したものであり、個々の概念をきれいに層化している。

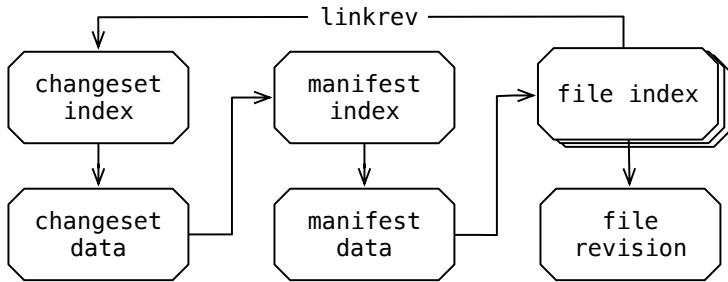


図 12.2: ログの構造

changelog のリビジョンは、このようになる。

```

0a773e3480fe58d62dcc67bd9f7380d6403e26fa
Dirkjan Ochtman <dirkjan@ochtman.nl>
1276097267 -7200
mercurial/discovery.py
discovery: fix description line

```

これこそが、revlog を層化したおかげで得られる価値である。changelog 層がこのようにシンプルな値のリストで表せるようになった。最初の行は manifest のハッシュである。それに続くのは作者名、そして日付と時刻 (Unix タイムスタンプとタイムゾーンオフセットをあわせた形式)、変更されたファイルの一覧、そして最後が内容を表すメッセージとなる。さらにもうひとつ隠し機能がある。実は任意のメタデータを changelog に含めることができるのだ。過去との互換性を維持するために、メタデータはタイムスタンプの後に追加することになる。

次は manifest だ。

```

.hgignore\x006d2dc16e96ab48b2fcc44f7e9f4b8c3289cb701
.hgsigs\x00de81f258b33189c609d299fd605e6c72182d7359
.hgtags\x00b174a4a4813ddd89c1d2f88878e05acc58263efa
CONTRIBUTORS\x007c8afb9501740a450c549b4b1f002c803c45193a
COPYING\x005ac863e17c7035f1d11828d848fb2ca450d89794
...

```

この manifest リビジョンは、changeset 0a773e が指している先である (Mercurial の UI は、曖昧にならない範囲で後半を省略できるようになっている)。これはツリーないのすべてのファイルを単純にならべた一覧であり、一行がひとつのファイルを表す。各行は、先頭にファイ

ル名があつてその後に NULL バイトを置き、さらに 16 進エンコードしたノード id が続く。このノード id が各ファイルの filelog を指している。ツリーに含まれるディレクトリを特別に区別することはない。しかし、ファイルのパスにスラッシュが含まれていれば、それがディレクトリであろうと判断できる。manifest はその他の revlog と同様に差分を格納していたことを思い出そう。つまり、manifest は revlog 層から容易にアクセスできる構造でなければならぬ。任意のリビジョンに対して、そのリビジョンで変更があつたファイルとその新しいハッシュだけを格納することになる。manifest は、通常は Mercurial の Python コードの中ではハッシュテーブル風のデータ構造で表される。ファイル名がキー、そしてノードが値である。

三種類めの revlog が、filelog だ。filelog は、Mercurial の内部的な store ディレクトリに格納されている。その名前は、追跡対象のファイルとほぼ同じである。ただし完全に同じではなく、ちょっとしたエンコードを行つてゐる。主要な OS すべてできちんと動作させるようするためである。たとえば、Windows や Mac OS X のように大文字小文字を区別しないファイルシステムも扱う必要があるし、Windows ではファイル名に使えない文字もある。また、ファイルシステムによってさまざまな文字エンコーディングが使われている。ご想像の通り、すべての OS で確実に動作させようとするとかなり大変なことになる。一方、filelog リビジョンの中身自体は、何の変哲もないものである。単にファイルの中身をそのまま保持しており、その先頭にちょっとしたオプションのメタデータが付属しているだけである（このメタデータを使って、ファイルのコピーやリネームなどの操作を追跡する）。

このデータモデルを使えば Mercurial リポジトリ内に格納されているどのデータにもアクセスできるようになる。しかし、このモデルがいつでも便利だとは限らない。実際の内部モデルは垂直指向（すなわち、ファイルごとにひとつの filelog が存在する）である。だが、Mercurial の開発者がよくやりたくなる作業は、特定のリビジョンに関してすべてのファイルの詳細を扱うというものだった。そんな場合は、まず changelog から特定の changeset を探し、そのリビジョンに関連する manifest や filelog に簡単にアクセスできるようにしたい。そんな要望に応えるため、別のクラス群が用意された。これは revlog の上位層に位置するもので、まさに期待通りに役割を果たす。その名は contexts だ。

revlog の設定が分割されている利点のひとつが、順序づけをするときにあらわれる。順序づけをするには、まず最初に filelog を追加し、その後に manifest を続け、最後に changelog を追加する。その間リポジトリは常に一貫した状態となる。changelog を読み込むあらゆるプロセスは、その他の revlog へのすべてのポインタが有効であることを保証されていることになり、そのおかげでさまざまな問題に対応できるようになる。それでもなお、Mercurial には明示的なロックも存在する。このロックで、複数のプロセスが同時に並行して revlog に追記できないようにしている。

作業ディレクトリ

最後にもうひとつ、重要なデータ構造が残っている。その名は *dirstate* だ。dirstate とは、ある特定の時点での作業ディレクトリ内に何があるのかを表したものである。最も重要なこと

として、`dirstate` はどのリビジョンがチェックアウトされているのかを保持している。これが `status` や `diff` といったコマンドでのすべての比較の基準となり、さらに次にチェックインするコミットの親を決めるうことになる。`dirstate` は、`merge` コマンドを発行したときには二つの親を持つことになる。そして、一方の変更群を他方にマージしようと試みる。

`status` や `diff` はとても頻繁に使う操作なので(直近のチェンジセットから現時点でどれくらい進んでいるのかを確かめられる)、`dirstate` には Mercurial が最後に作業ディレクトリを走査したときの状態のキャッシュも残している。最終更新時刻とファイルサイズを追跡することで、ツリーの走査を高速化しているのだ。また、ファイルの状態も同時に追跡しなければならない。つまり、そのファイルが作業ディレクトリに追加されたのあるいはそこから削除されたのか、変更が加わったのかといったことだ。これも作業ディレクトリの走査を高速化させ、コミット時にこれらの情報を取得しやすくしている。

12.3 バージョン管理の仕組み

ここまでで、Mercurial の内部的なデータモデルや低レベルでのコードの構造についてなじみが出てきたことだろう。さらにもう一歩進んでみよう。ここでは、これまでに説明した基盤の上で Mercurial がどのようにバージョン管理の概念を実装しているのかを考えていく。

ブランチ

ブランチの一般的な利用法は、開発のラインを別々に分割して進めて後で統合するというものだ。その理由として考えられるのは、たとえば誰かが新しい手法を試すときに開発のメインラインは常に出荷可能な状態にしておくこと(フィーチャブランチ)、あるいは旧バージョン用のバグフィックスを素早くリリースできるようにしておくこと(保守ブランチ)などである。どちらの手法もよく使われるものであり、今時のバージョン管理システムならすべてこれらの手法に対応している。暗黙のブランチは DAG ベースのバージョン管理では一般的なものだが、名前付きブランチ(ブランチ名をチェンジセットのメタデータに記録する方式)はそれほど一般的ではない。

当初は、Mercurial ではブランチの名前を明示することができなかった。ブランチの処理は、単に別々のクローンを作つてそれを個別に公開するというだけのことだった。これは効率的だし理解しやすく、特にフィーチャブランチに関しては非常に便利だった。というのも、ブランチを作るオーバーヘッドがほとんどなかったからである。しかし大規模なプロジェクトでは、クローンであつてもかなり重たい作業となつた。大半のファイルシステム上にリポジトリのハードリンクが格納されていてもなお、個別の作業ツリーを作るのには時間がかかるし大量のディスク容量を必要とする。

これらの弱点に対応するため、Mercurial には第二のブランチ方式が追加された。チェンジセットのメタデータにブランチ名を含める方式である。`branch` コマンドが追加され、現在の

作業ディレクトリにブランチ名を設定できるようになった。このブランチ名は、次のコミットのときに用いられる。通常の `update` コマンドを使ってある特定の名前のブランチにアップデートすることもできるし、ブランチ上でコミットしたチェンジセットは常にそのブランチに関連するコミットとなる。この手法は名前付きブランチ (*named branches*) と呼ばれる。しかし当初はブランチを作るだけでそれを閉じる（ブランチ一覧からそのブランチを見えなくする）ことはできなかった。ブランチを閉じられるようになったのは、それから数リリース後のことだった。ブランチを閉じる機能の実装は、チェンジセットのメタデータにフィールドを追加することで行った。追加フィールドに、このチェンジセットがブランチを閉じるという情報を記録したのだ。あるブランチが複数の `head` を持つ場合は、それらをすべて閉じてからでないとブランチ一覧からそのブランチを消すことはできない。

もちろん、何かを成し遂げる方法はたったひとつではない (there's more than one way to do it)。Git における名前付きブランチの実装は Mercurial とは異なり、参照を使っている。参照とは Git の歴史上で別のオブジェクト（通常はチェンジセット）を指す名前のことである。これは、Git のブランチがその場限りの短期的なものであることを意味する。つまり、いったん参照を削除してしまえば、そこにブランチが存在した形跡は一切なくなってしまう。これはちょうど、別々の Mercurial のクローンを作つて、一方のクローンをもう一方にマージしたのと同じ状態だ。この方式だとローカルでのブランチの操作が非常に簡単かつ軽量になり、ブランチ一覧がごちゃごちゃすることも避けられる。

この方式でのブランチ管理はどんどん広まり、今や名前付きブランチ方式や Mercurial 風のクローンによるブランチよりもずっと幅広く使われている。その流れを受けて、Mercurial にも `bookmarks` 拡張が用意された。将来のバージョンでおそらく Mercurial に組み込まれることだろう。この拡張は、バージョン管理対象外のシンプルなファイルを使って参照を追跡する。Mercurial のデータ交換に使っている `wire` プロトコルを拡張してブックマークも通信できるようにし、ブックマークもプッシュできるようにした。

タグ

ちょっと見た限りだと、Mercurial におけるタグの実装は奇妙に感じことだろう。`tag` コマンドを使って最初にタグを追加するときに、`.hgtags` というファイルがリポジトリに追加され、それをコミットする。このファイルの各行には、チェンジセットのノード `id` とそのチェンジセットノードに対応するタグ名が記録される。このようにして、タグ情報ファイルはリポジトリ内の他のファイルとまったく同様の扱いとなるのである。

このようにした重要な理由が次の三つだ。まず最初に、タグは変更できなければならない。間違いは必ず発生するものだし、間違ったタグの修正や削除ができなければならない。次に、タグ自体もチェンジセットの履歴の一部でなければならない。そのタグがいつ誰によってどういう理由で作られたのか、あるいは変更されたのかといった情報を見られれば有用だ。最後に、過去のチェンジセットにさかのぼってタグを設定できなければならない。たとえば、

バージョン管理システムからエクスポートしたリリース用の成果物に対して大規模なテストをしてから初めてリリースするというプロジェクトも存在する。

これらの特性はすべて、.hgtags の設計から容易に得られる。人によっては作業ディレクトリ内に.hgtags ファイルが存在するのを見て困惑するかもしれない。しかし、そのおかげでタグ付けの仕組みと Mercurial のその他の部分との統合(同じリポジトリの別のクローンとの同期など)が非常にシンプルに行えるのだ。もしタグがソースツリーとは別になっているのなら(Git などがそうだ)、そのタグの出自を調べたり重複したタグの衝突を解決したりするための仕組みが別途必要になる。後者の状況はめったにないかもしれないが、そもそもそんな状況が問題になりすらしない設計があるのならそっちのほうがよいだろう。

これらすべてをうまく機能させるために、Mercurial は.hgtags ファイルに新しい行を追加することしかしない。これは、別々のクローンでタグが並行して作られたときのマージにも役立つ。任意のタグに対して最新のノード id が優先され、空のノード id を追加すると(空のノード id は空の root リビジョンを表し、すべてのリポジトリに共通して存在する)、それはタグを削除するのと同じ意味になる。Mercurial はリポジトリ内のすべてのブランチのタグの関係についても考慮し、新しいタグのほうを優先する。

12.4 全体的な構造

Mercurial の大半は Python で書かれている。ごく一部で C が使われているところもあるが、これはアプリケーション全体としてのパフォーマンスを考慮したものである。そのごく一部を除いた部分では、Python で書くほうが適していると考えた。なぜなら、上位レベルの概念を表現するには Python のような動的言語のほうがずっと容易だったからである。コードの多くは特にパフォーマンスを最優先に考えたものではないが、そんなことはあまり気にしていない。それと引き替えに、大半の部分のコーディングが簡単になっているのだから。

Python のモジュールは、各モジュールがひとつのファイルに対応している。モジュールの中には必要に応じていくらでもコードを書くことができ、このモジュールがコードの構造をまとめる鍵となる。モジュールの中で型を使ったり他のモジュールの関数を呼び出したりするときには、他のモジュールを明示的にインポートすることがある。`__init__.py` モジュールを含むディレクトリはパッケージと呼ばれ、そこに含まれるすべてのモジュールやパッケージが Python のインポーターに公開される。

Mercurial がデフォルトでインストールする Python のパッケージは `mercurial` と `hgext` の二つである。`mercurial` パッケージには Mercurial を実行するために必要なコアコードが含まれており、もう一方の `hgext` にはさまざまな拡張機能が含まれている。コアと一緒に配布すると有用だろと思われるものである。しかし、これらの拡張を使いたい場合は、設定ファイルを自分で編集して拡張を有効化させなければならない(詳細は後述する)。

念のために言うが、Mercurial はコマンドラインアプリケーションである。つまり、インターフェイスはシンプルなものだということである。ユーザーは、`hg` スクリプトにコマンド

を指定して呼び出すことになる。各種コマンド (log や diff、あるいは commit など) にはいくつかのオプションや引数を受け取るものもある。また、すべてのコマンドに共通するオプションもある。次に、このインターフェイスを通して起こることは次の三つに分類できる。

- hg は、ユーザーに何かを問い合わせたり状態を示したりするメッセージを出力する。
- hg は、コマンドラインプロンプトでさらに入力を待ち受けることもできる。
- hg は、別のプログラム (コミットメッセージ用のエディタやコードの衝突を解消するためのマージを支援するプログラムなど) を起動することもある。

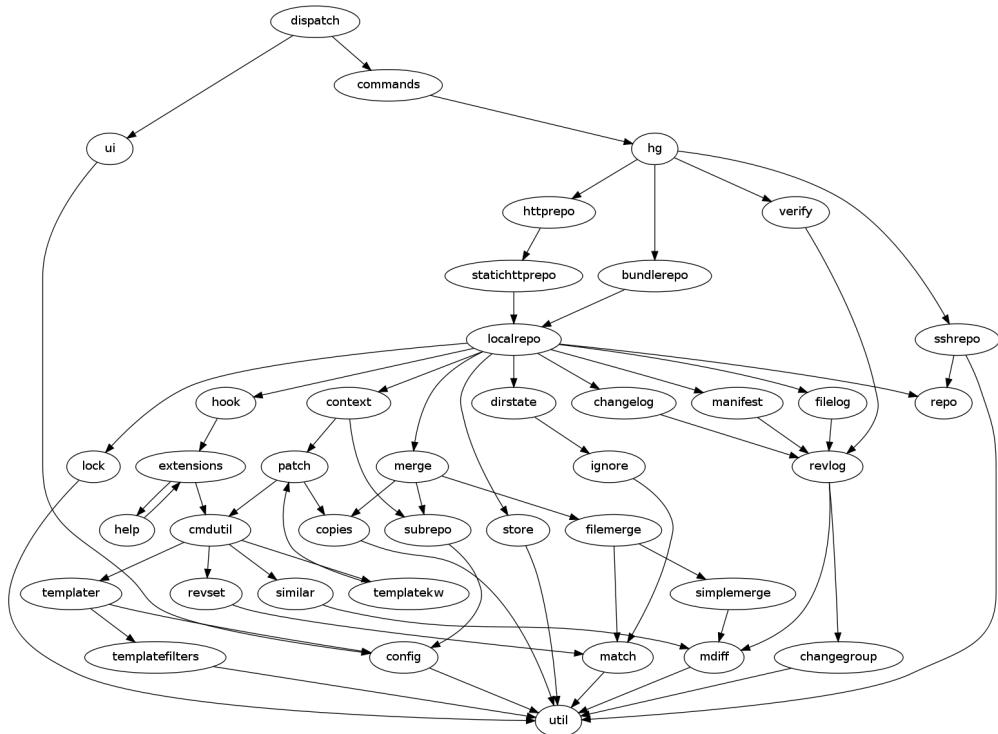


図 12.3: インポートグラフ

このプロセスのはじまりは、図 12.3 のインポートグラフから読み取れる。すべてのコマンドライン引数は、まず `dispatch` モジュールの関数に渡される。最初に行うのは `ui` オブジェクトのインスタンスを作ることである。`ui` クラスは、まず最初にいろいろな既知の場所 (ホームディレクトリなど) から設定ファイルを探し、設定オプションを `ui` オブジェクト内に保存する。設定ファイルには拡張機能へのパスが含まれていることもある。その場合は、この時点で拡張を読み込まねばならない。コマンドラインオプションで渡されたグローバルオプションもすべて、ここで `ui` オブジェクトに保存される。

これが終わると、次にリポジトリオブジェクトを作るかどうかを判断しなければならない。ほとんどのコマンドはローカルリポジトリ (`localrepo` モジュールの `localrepo` クラスで表

されるもの)が必要だが、なかにはリモートリポジトリ (HTTP や SSH、あるいはその他登録済みの形式でアクセスするもの) でも動作するコマンドもある。また、リポジトリを参照しなくても動作するコマンドもある。リポジトリを参照しないコマンドの例としては `init` があり、これは新しいリポジトリを初期化するときに使う。

すべてのコアコマンドは、`commands` モジュール内のひとつの関数で表される。そのため、何か特定のコマンドのソースコードを探すのも容易である。`commands` モジュールにはハッシュテーブルも含まれており、このハッシュテーブルではコマンド名とそれに対応する関数そしてそのオプションに関する説明を対応させている。こうすることで、一般的なオプション群(たとえば、たいていのコマンドは `log` コマンドのオプションと同じようなオプションを使えるようになっている)を共有できるようになる。オプションに関する説明は、`dispatch` モジュールがそのコマンドのオプションをチェックするときに使う。そして、渡された値をそのコマンドの関数が期待する型に変換する。ほとんどすべての関数は、`ui` オブジェクトと操作対象の `repository` オブジェクトを受け取る。

12.5 拡張性

Mercurial をさらに強化できるように用意されたのが、拡張を書く機能である。Python は比較的身につけやすい言語であり、Mercurial の API はとてもうまく作られている(きちんとドキュメントも用意されている)。そんなこと也有って「Mercurial の拡張を書きたいから Python を勉強はじめました」という人も多い。

拡張機能の作成

拡張を有効にするには、Mercurial が起動時に読み込む設定ファイルの中に一行追加する必要がある。拡張を表すキーと、Python モジュールへのパスをそこに記述する。機能を追加するには次のようにいくつかの方法がある。

- 新しいコマンドの追加
- 既存のコマンドのラッピング
- 使用するリポジトリのラッピング
- Mercurial の関数のラッピング
- 新しいリポジトリ型の追加

新しいコマンドを追加するには、単に `cmdtable` という名前のハッシュテーブルを拡張モジュールに追加するだけである。このハッシュテーブルを拡張ローダーが読み込んで、コマンドテーブルに追加する。コマンドのディスパッチ時に、このコマンドテーブルを用いる。同じく、拡張内で `uisetup` および `reposetup` 関数を定義することもできる。これらはそれぞれ、UI やリポジトリのインスタンスを生成した後にディスパッチのコードから呼び出される。共通のふるまいが一つある。それは、`reposetup` 関数を使い、拡張が提供する `repository`

のサブクラスにリポジトリをラップすることだ。こうすることで、拡張側ですべての基本的なふるまいを変更できるようになる。たとえば、私がかつて書いたある拡張では、`uisetup` をフックして `ui.username` の内容を書き換えていた。環境から取得できる SSH 認証情報に基づいたユーザー名を設定していたのだ。

新しいリポジトリ型を追加するといった、さらに思い切った拡張を書くこともできる。たとえば `hgsubversion` プロジェクト (Mercurial 本体には組み込まれていない) は、Subversion のリポジトリを扱うためのリポジトリ型を登録する。これを使えば Subversion のリポジトリもクローンでき、まるで Mercurial のリポジトリであるかのように扱えるようになる。変更を Subversion のリポジトリにプッシュすることも可能だ。しかし両者のシステムの間にはインピーダンスミスマッチがあるため、かなりのエッジケースが存在する。一方、ユーザーインターフェイスに関しては完全に透過的になっている。

Mercurial を根本的に書き換える人には “モンキーパッチ” という手段がある。動的言語の世界ではおなじみの方法だ。拡張のコードは Mercurial と同じアドレス空間で実行され、また Python はリフレクション機能を持つ極めて柔軟な言語なので、Mercurial 内部で定義されている関数やクラスの書き換えだって容易に行える。見苦しいハックになってしまう可能性もあるが、非常に強力な仕組みでもある。たとえば、`hgext` にある `highlight` 拡張は、組み込みのウェブサーバーを書き換えてリポジトリブラウザにシンタックスハイライト機能を追加し、ファイルの中身を読みやすくしている。

Mercurial を拡張する方法は、もう一つ存在する。よりシンプルな方法である **エイリアス** だ。すべての設定ファイルでエイリアスを定義できる。エイリアスを使えば、既存のコマンドと設定済みのオプション群をあわせたものに新しい名前をつけることができる。これを使えば、既存のコマンドの短縮形を定義することもできる。最近のバージョンの Mercurial では、シェルのコマンドもエイリアスとして呼べるようになった。これを使えば、シェルスクリプトを使ってより複雑なコマンドを作ることもできる。

フック

各種バージョン管理システムは、これまでずっとフック機構を提供してきた。この仕組みを使って VCS 上でのイベントとその他の世界とのやりとりができるようにしていたのだ。よくある使い道としては、継続的インテグレーションシステム¹に通知を送ったり、ウェブサーバー上の作業ディレクトリを更新して最新版を一般公開したりといったものがある。もちろん Mercurial にも同様に、フックを起動するサブシステムが組み込まれている。

実際のところ、フックにもまた二通りの仕組みが用意されている。一つは昔ながらの仕組みで他のバージョン管理システムにもよくあるもの、つまり、シェル内でスクリプトを実行するという仕組みだ。もう一方はもう少し興味深い仕組みである。というのも、ユーザー側で Python のフックを起動させる仕組みとして、Python のモジュール名とそのモジュールから呼び出す関数名を指定させているのだ。Mercurial と同じプロセス内で動くためにより高速

¹ 第 6 章を参照。

になるというだけでなく、この方式では repo オブジェクトや ui オブジェクトも渡せるので VCS ないでより複雑な操作を容易に行えるということになる。

Mercurial のフックは pre-command、post-command、controlling そして miscellaneous の四種類に分類できる。最初の二つは単純なもので、設定ファイルの hooks セクションに *pre-command* あるいは *post-command* というキーを設定してそこに任意のコマンドを定義するだけである。残りの二つについては、定義済みのイベントが用意されているのでそれを利用する。*controlling* フックがその他と異なる点は、何かが起こる直前にそのフックが実行され、場合によってはそのイベントの発生を中断させることもできるというところだ。よくある利用法は、中央サーバー上で何らかの方法でチェンジセットを検証するというものだ。Mercurial は分散型のシステムなので、コミット時にはこの手のチェックをすることができない。たとえば Python プロジェクトでは、フックを使ってコーディング規約のチェックを行っている。あるチェンジセットで追加しようとしているコードが所定のスタイルを満たしていない場合には、中央リポジトリがそれを却下するという仕組みだ。

それ以外のフックの使い道としては、プッシュログがある。これは Mozilla やその他多数の企業組織で使われている。プッシュログは個々のプッシュを記録し(ひとつのプッシュには複数のチェンジセットが含まれることもある)、さらにそのプッシュを誰がいつ行ったのかも記録する。これを、そのリポジトリの監査証跡とするのだ。

12.6 学んだこと

Matt が Mercurial の開発を始めるにあたって最初に決めたことのひとつが、Python で開発することだった。Python は拡張性に優れており、非常にお手軽にコードを書ける。また、さまざまなプラットフォームでの互換性もきちんと考慮されているので、Mercurial を主要三大 OS で動くようにするのも比較的容易だった。その一方で、Python は他の(コンパイル型の)言語に比べて実行速度が遅い。特にインタプリタの起動には時間がかかる。長期間実行しつづけるプロセスならこれは気にならないが、バージョン管理システムのように短期間に何度も起動するツールにとっては問題だ。

開発初期の方針として、いったんコミットした後のチェンジセットの変更をしにくくようとした。あるリビジョンを変更するにはその id ハッシュの変更が必須になるので、いったんインターネット上に公開したチェンジセットの“取り消し”は面倒な作業となる。そこで Mercurial では、その作業が難しくなるようにしたのだ。しかし、まだ公開していないリビジョンを変更するぶんには特に問題はない。そこで、Mercurial がリリースされて間もなく、未公開のリビジョンの変更をしやすくするようコミュニティが動き出した。この問題を解決しようとする拡張も存在するが、利用方法を身につけるのが難しく、それまでふつうに Mercurial をしてきたユーザーにとってはあまり直感的ではないものだった。

revlog はディスクのシーク処理を減らすのに役立ったし、changelog や manifest そして filelogs のレイヤー化アーキテクチャはうまく機能した。コミットは高速に行え、各リビジョンが利

用するディスク領域は比較的少なめになる。しかし、ファイルのリネームのような一部の作業は、各ファイルのリビジョンを別々に格納しているせいで効率的に行えない。最終的にはこの問題も修正されるだろうが、レイヤー構造に反するちょっとしたハックが必要になるだろう。同じく、ファイル単位の DAG を使って filelog ストレージを管理していることから、あまり大量のファイルを扱うのは現実的ではない。大量のファイルを管理するコードの一部がオーバーヘッドになってしまっている。

Mercurial が重視しているもう一つのことは、使い方を簡単に身につけられるようにすることだ。必須機能の大半を少数のコアコマンド群にまとめ、各コマンドで一貫性のあるオプションを用意している。その狙いは、特に他の VCS を使ったことがあるユーザーが Mercurial を段階的に学べるようにすることである。この思想の一環として、拡張機能で Mercurial をカスタマイズできるようにした。単に特定の使い道にあわせて拡張するだけではなく、それ以外にも使えるようにしたのだ。この狙いがあるため、Mercurial の開発者はその UI をできるだけ他の VCS(特に Subversion) と合わせようとしている。同様に、開発チームではドキュメントも重視している。Mercurial 自身に付属するドキュメントでは、他のトピックやコマンドへのクロスリファレンスも提供されている。また、エラーメッセージも有用なものになるよう心がけており、単に「操作が失敗しました」ではなく、何をすればいいのかというヒントを提供できるようにしている。

もう少し小さめの選択の中には、新しいユーザーにとっては少しひっくりするかもしれないものもある。たとえば、タグの処理を(先のセクションで述べたように)作業ディレクトリ内の個別のファイルで扱うという方式を好みない人も多いだろう。しかしこ的方式には好みの特性もあるのだ(もちろん欠点もあるが)。また、他の VCS ではチェックアウトしたチェンジセットとその先祖だけをリモートホストに送るのがデフォルトだが、Mercurial ではリモートに存在しないすべてのチェンジセットを送信することを選んだ。どちらの方式にもそれなりの理由があり、あなたにとってどちらが最適かは、その開発スタイルに依存する。

他のソフトウェア開発プロジェクトと同様、Mercurial の開発でもさまざまなトレードオフが発生した。Mercurial は今までもよい選択をしてきたと思っているが、今になって考えると「もっと適切な選択肢があった」と思えるものもある。歴史的には、Mercurial は一般向けに使えるようになった第一世代の分散型バージョン管理システムである。個人的には、次の世代のシステムがどのようなものになるのかを楽しみにしている。

NoSQLを取り巻く世界

Adam Marcus

他の大半の章とは異なり、NoSQL は単体のツールを表す言葉ではない。時には補完しあったり時には競合したりするさまざまなツール群からなる生態系を指す用語である。NoSQL と名付けられたツール群は、SQL ベースのリレーションナルデータベースとは異なる方式でデータを格納する仕組みを提供する。NoSQL を理解するには、まずどんなツールが存在するのかを理解し、そしてそれぞれのツールがどのようにデータを格納するのかという設計を知る必要がある。

NoSQL をストレージシステムとして使おうと検討するときにまず理解すべきことは、NoSQL の中にもさまざまなシステムが存在することである。NoSQL システムは、伝統的なりレーションナルデータベースシステムが持つ快適な機能の多くを廃止した。そして、これまでデータベース側に隠蔽されていた操作をアプリケーションの設計側に押し出したのだ。つまり、システムアーキテクトの立場で考えると、これらのシステムの仕組みをより深く知っておく必要があるということだ。

13.1 その名の由来は？

NoSQL の世界を語る前に、まずはその名前をきちんと定義してみよう。NoSQL システムとは、文字通りにとらえると SQL ではない問い合わせインターフェイスを持つシステムのことである。ただ、NoSQL コミュニティではもう少し包括的にとらえている。NoSQL システムとは今までのリレーションナルデータベースの代替となるものであり、開発者がシステムを設計するときに、SQL にとらわれずに (*Not Only SQL*) いろいろなインターフェイスを使えるようにするものだという考え方である。リレーションナルデータベースをその代替となる NoSQL で完全に置き換えることがあるかもしれないし、両者を組み合わせてアプリケーション開発時のさまざまな問題に対応していくこともあるかもしれない。

NoSQL の世界に飛び込む前に少し考えてみよう。SQL や関係モデルが適するのはどのような場面だろうか。そして NoSQL システムのほうがより適しているのはどんな場合だろうか。

SQL と関係モデル

SQL は、データを問い合わせるための宣言型の言語である。宣言型の言語とは、そのシステムに何をさせたいのかをプログラマーが指定する言語のことである。そのシステムが**どのように動くべき**なのかを手続き的に定義するのではない。いくつか例を示そう。39 番の社員を探したり、レコード全体から社員名と電話番号だけを取り出したり、経理部門に属する社員のレコードだけに絞り込んだり、部署ごとの社員数を調べたり、社員テーブルのデータを管理職テーブルと連結させたりといった操作だ。

おおざっぱに言うと、SQL を使えば、データのディスク上での配置や使うインデックスそしてデータを処理するアルゴリズムを知らなくてもこれらの問い合わせに答えられるということである。多くのリレーションナルデータベースのアーキテクチャ上で重要となるパツツが**クエリオプティマイザ**だ。これは、論理的に等価であるいろいろな問い合わせプランの中から最も効率的な問い合わせができるものを見つける。このオプティマイザは、たいていの場合ふつうのデータベースユーザーよりも賢い。しかし時には、必要な情報が不足していたりデータモデルがあまりにもシンプルすぎたりといった理由で最適な実行計画を生成できないこともある。

現在もっとも一般的に使われているデータベースがリレーションナルデータベースで、これは**関係データモデル**に従っている。このモデルは、現実世界のさまざまなエンティティをそれぞれ別のテーブルに格納する。たとえば、社員の情報は Employees テーブルに格納し、部署の情報は Departments テーブルに格納するといったものだ。テーブルの各行は、さまざまな項目を保持している。たとえば社員の持つ情報としては社員 ID や給与、生年月日、そして姓名などである。これらの項目が、Employees テーブルの各カラムに格納される。

関係モデルは SQL と密接に関連している。フィルタのような単純な SQL 問い合わせは、あるフィールドが何らかの条件(例: `employeeid = 3`、`salary > $20000`)にマッチするレコードをすべて取得する。層少し複雑な構造を使って、データベースに他の作業をさせることもできる。複数のテーブルからのデータの結合(例: 社員番号 3 番の社員が所属する部署の部署名は?)などである。それ以外にも、集約(例: 従業員の給与の平均額は?)などの操作もでき、この場合はテーブル全体のスキヤンが発生する。

関係モデルでは高度に構造化されたエンティティを定義し、さらにそれらの間に厳格なリレーションシップも定義する。このようなモデルに対して SQL で問い合わせをすれば、カスタム開発なしで複雑なデータの取得もできるようになる。しかし、このように複雑なモデルや問い合わせにも制限はある。

- 複雑さは予測不可能性につながる。SQL は表現力がありすぎるので、個々のクエリのコスト、つまりその作業量に関するコストをきちんと考へないといけなくなる。問い合わせ言語をシンプルにするとアプリケーションのロジックが複雑になってしまい、データストレージシステムを用意するのは簡単になる。単にシンプルなリクエストに応答できればそれで済むのだから。

- 問題をモデル化する方法は一つではない。関係データモデルは厳格なモデルであり、各テーブルに設定されたスキーマが個々の行のデータを規定する。あまり構造化されていないデータを格納する場合や行によって格納するカラムにはばらつきがある場合などは、関係モデルだと制約が大きすぎることになる。アプリケーションの開発者の視点で考えても、関係モデルがあらゆる種類のデータを完璧にモデリングできるとは言えない。たとえば、アプリケーションのロジックの多くはオブジェクト指向の言語で書かれており、より高レベルの概念であるリストやキュー、セットなどを使っている。プログラマーの中には、永続層でこれらをモデリングしたいという人もいるだろう。
- データの量が増加して一つのサーバーでは保持しきれなくなると、データベース内のテーブルをパーティションに区切って複数のコンピューターで管理する必要がある。別のテーブルにあるデータを取得するためにネットワークをまたがる JOIN をするなどということは回避するには、テーブルを非正規化しなければならない。非正規化とは、さまざまなテーブルにあるデータの中で一度に使いたいものをすべて一か所にまとめて格納することである。これにより、データベースはまるでキーで検索する方式のストレージシステムのようになるが、他にもっと適したデータモデルがないのかという思いは残る。

長年にわたって検討してきた設計を独断で切り捨ててしまうのは、あまりよい考えではない。データをデータベースに格納するときには SQL と関係モデルを検討しよう。これは何十年にもわたる研究と開発のたまものであり、高度なモデリングができる。また、複雑な操作も容易に理解できることは請け合う。NoSQL が選択肢にあがるのは、何か特有の問題がある場合だ。たとえば大量のデータを扱う必要があったり作業量が膨大になったり、SQL とりレーションナルデータベースではうまく最適化できないようなデータモデリングを採用した場合などである。

NoSQL のはじまり

NoSQL ムーブメントの起源をたどれば、その大半は研究コミュニティの論文に行き着く。NoSQL システムの設計に関する決断には多くの論文が絡んでいるが、中でも特筆すべきなのが次の二つである。

Google の BigTable [CDG⁺06] は興味深いデータモデルを提示した。このモデルでは、複数列からなる履歴データを分類して格納する。データを複数のサーバーに分散させるために階層型のレンジベースパーティショニング方式を用い、データは厳密な整合性（この概念については 13.5 節で定義する）のもとで更新される。

Amazon の Dynamo [DHJ⁺07] は、キー指向の分散型データストアを用いる。Dynamo のデータモデルはシンプルで、アプリケーション固有のデータの blob にキーをマッピングする。パーティショニング方式は障害からの回復機能を持つが、それを実現するためにデータの整合性はより緩やかな手法（結果整合性）で管理している。

これら二つの概念についてこれから詳細を説明するが、その概念の多くは互いに組み合わせて使えるものだと理解しておくことが大切だ。たとえば、NoSQL システムのひとつである HBase¹は、BigTable の設計に忠実な作りになっている。別の NoSQL システムである Voldemort²は、Dynamo の機能の多くを再現している。さらに別の NoSQL システムである Cassandra³の場合は、BigTable から引き継いだ機能（データモデル）もあれば Dynamo から引き継いだ機能（パーティショニングや整合性管理の方式）もある。

特徴と検討事項

NoSQL システムは、大掛かりな SQL 標準規格と決別して、ストレージの設計に関してシンプルながらも段階的なソリューションを提供する。その思想は、「データベースがデータを操作する方法を単純化すればするほど、アーキテクトは問い合わせのパフォーマンスを予測しやすくなる」というものだ。NoSQL システムの多くは、複雑な問い合わせロジックをアプリケーション側に任せている。その結果、データストア側では問い合わせのパフォーマンスを予測しやすくなる。問い合わせの種類が限られてくるからである。

NoSQL システムは、リレーションナルデータに単に宣言型の問い合わせ機能を追加するものとは一線を画する。トランザクション特性や整合性、永続性といった機能は、銀行などの組織のデータベースに求められる要件を満たすことを保証するものだ。トランザクションは、複数の込み入った操作をひとまとめにして「すべて成功」か「すべて失敗」かのいずれかになることを保証する。たとえば、ある口座から引き落とした資金を別の口座に入金するなどといった操作がトランザクションの対象となる。整合性は、ある値が更新されたときにそれ以降の問い合わせが更新後の値を見られることを保証する。永続性は、ある値が更新されたときにそれが安定したストレージ（ハードディスクドライブなど）に書き込まれ、データベースがクラッシュしても復旧可能であることを保証する。

NoSQL システムでは、これらの保証のうちのいくつかをより緩やかにした。金融関連のアプリケーションを除く多くのアプリケーションではそれでも受け入れ可能なレベルであり、保証を緩める引き替えにパフォーマンスを向上させている。このように保証を緩め、そしてデータモデルと問い合わせ言語を変更することで、データベースをパーティショニングして複数サーバーに分散させることも簡単になった。データの量が増えて一つのマシンではさばききれなくなつてもだいじょうぶである。

NoSQL システムは、まだ生まれたばかりの幼年期にある。本章で取り上げるシステムにおけるアーキテクチャ上の判断は、さまざまなユーザーの要求をとりまとめたものである。さまざまなオープンソースプロジェクトのアーキテクチャをまとめようとしたときに最も大変だったのは、どのシステムも変わりつつあるというところだった。個々のシステムの詳細は変わるものだということを頭に入れておこう。何かの NoSQL システムを使うことになった

¹<http://hbase.apache.org/>

²<http://project-voldemort.com/>

³<http://cassandra.apache.org/>

ときに、本章はどう考えればいいかの指針にはなるだろう。しかし、その製品にどんな機能があるかを知るといった目的には使えない。

NoSQL システムについて考えるにあたって、検討すべき点を次にまとめる。

データモデルとクエリモデル: データの表現方法は? 行? それともオブジェクト? あるいはデータ構造とかドキュメントとか? データベースに対する問い合わせで複数のレコードを集約できる?

永続性: 値を変更したときに、それはすぐにストレージ上に反映される? ひとつのマシンがクラッシュした場合に備えて複数のマシンに格納する?

スケーラビリティ: データは単一のサーバー上に置く? データの読み書きをさばくために複数のディスクを扱う必要がある?

パーティショニング: スケーラビリティや可用性、永続性の観点から、データを複数のサーバーに置く必要がある? どのレコードがどのサーバーにあるのかということをどうやって知る?

整合性: 複数のサーバーにまたがってデータのパーティショニングや複製を行ったときに、レコードの変更に対して各サーバーはどのように協調する?

トランザクションの特性: 一連の操作を実行するときに、それをトランザクションとしてまとめるこことできるデータベースもある。トランザクションは、実行中の他の操作との間の ACID (Atomicity: 原子性、Consistency: 整合性、Isolation: 独立性、Durability: 永続性) の一部あるいはすべてを保証する。これらの保証は一般的にパフォーマンスとのトレードオフとなるが、あなたが扱う業務ロジックはこれらの保証を要するものか?

単一サーバーでのパフォーマンス: データを安全にディスク上に格納したい場合に、ディスク上のデータ構造として最適なものは? 読み込みが多い場合と書き込みが多い場合とではどうなる? ディスクへの書き込みがボトルネックになっている?

作業量の分析: ユーザーにとって使いやすいウェブアプリケーションを作るときには、作業量のチェックに細心の注意を払うことになる。多くの場合、欲しいのはデータセットサイズのレポートで、たとえば複数のユーザーの統計情報を集約したものだろう。あなたの利用法や使うツールは、その手の機能を必要としている?

これらすべての検討事項について取り上げる予定だが、後半の三点については(どれも同様に重要ではあるけれども)本章ではありません詳しく扱わない。

13.2 NoSQL のデータモデルおよびクエリモデル

データベースの**データモデル**とは、データをどのような論理構造で管理するかを示すものである。一方**クエリモデル**は、データの取得や更新をどのように行うのかを決定づける。一般的なデータモデルとしては、関係モデルやキー指向ストレージモデル、そして各種のグラフモデルなどがある。クエリ言語としては、SQL やキーワードアッピングそして MapReduce などがおなじみだろう。NoSQL システムはさまざまなデータモデルとクエリモデルの組み合わせでできており、アーキテクチャ上の検討事項もそれぞれ異なる。

キー・ベースの NoSQL データモデル

NoSQL システムの多くは、関係モデルや SQL の機能性とは一線を画して、データセットの検索を单一フィールドによるものだけに制限している。たとえば、社員情報にはさまざまな項目があるにもかかわらず、ID による検索しかできないといった具合だ。その結果として、NoSQL システムにおける問い合わせの大半は、キーによる検索をベースにしたものとなる。プログラマーは、各データを識別するためのキーを選択する。そしてたいていの場合、できることといえばデータベース内でそのキーによる検索をしてアイテムを取得することくらいなのだ。

キーリックアップ方式のシステムでは、複雑な結合操作や複数キーによる同一データの取得などを実現しようとすると、キーの名前にちょっとした工夫を要する。社員を検索する際に「社員 ID での検索」「ある部署に所属する全社員の検索」の二通りを行いたい場合は、二種類のキーを作成することになる。たとえば、キー `employee:30` は社員 ID が 30 の社員レコードを指し、キー `employee_departments:20` は部署番号 20 に属する全社員のリストを含むといった具合だ。結合操作はアプリケーションのロジック側に追い出される。部署番号 20 に属する全社員を取得するには、アプリケーション側でまず `employee_departments:20` を使って社員 ID のリストを取得し、そのリストをループさせて各 ID に対して `employee:ID` による検索を行う。

キーリックアップモデルの利点は、データベースへの問い合わせのパターンが一貫したものになるというところである。データベースで行う作業はキーリックアップだけであり、これは比較的一様で予測しやすいものである。アプリケーションのボトルネックを探すプロファイリングもシンプルになる。というのも、複雑な操作はアプリケーションのコード側に押し出されているからである。その反面、データモデルのロジックと業務ロジックが密結合してしまい、抽象化は崩れてしまう。

各キーに関連づけられたデータについて見ていく。各種 NoSQL システムは、さまざまなソリューションを使っている。

キー・バリューストア

NoSQL のデータ格納方式の中で最もシンプルなのがキー・バリューストアだ。個々のキーを、任意のデータを含む値に対応させる。NoSQL システム自体はその値の中身については何も知らず、単にデータをアプリケーションに渡すだけとなる。先ほどの社員データベースの例で考えると、キー `employee:30` を blob に対応させることになる。その中身は JSON かもしれないし、Protocol Buffers⁴ や Thrift⁵ あるいは Avro⁶ のようなバイナリフォーマットかもしれない。何らかの形式で、社員 ID が 30 の社員の情報をカプセル化したものとなる。

⁴<http://code.google.com/p/protobuf/>

⁵<http://thrift.apache.org/>

⁶<http://avro.apache.org/>

構造化されたフォーマットで複雑なデータを表してそれをキーに関連づけた場合は、取り出したデータの処理はアプリケーション側で行う必要がある。キー・バリュー形式のデータストアでは、キーを指定して取り出した値の中の特定の項目を取り出すような仕組みは用意されていない。キー・バリューストアの強みは、そのシンプルなクエリモデルにある。通常は set、get そして delete といったプリミティブで構成されている。一方、データベース内でのシンプルな絞り込み機能を追加する仕組みは持っていない。これは、扱うデータの中身が把握できないからである。Voldemort は Amazon の Dynamo をベースにしたシステムであり、分散型キー・バリューストアを提供する。BDB⁷は、キー・バリュー型のインターフェイスを持つ永続化ライブラリを提供する。

キー・データ構造ストア

キー・データ構造ストアは Redis⁸によって有名になった方式で、値として型を割り当てる。Redis では、値として使える型は integer、string、list、set そして sorted set である。set/get/delete に加えて型ごとに固有のコマンドも用意されている。たとえば整数型ならインクリメントやデクリメント、リストならプッシュやポップなどだ。さらに、クエリモデルにも機能が追加されている。リクエストの種類によってパフォーマンスが劇的に落ちるなどということはない。シンプルな、型ごとの機能を提供する一方で、集約や結合といった複数キーの操作はできない。このようにして、Redis は機能とパフォーマンスのバランスをとっている。

キー・ドキュメントストア

キー・ドキュメントストアには CouchDB⁹、MongoDB¹⁰そして Riak¹¹などがある。これらは、構造化された情報を含むドキュメントをキーにマップする。これらのシステムは、ドキュメントを JSON 形式 (あるいはその類似形式) で格納する。リストや辞書も格納し、あるドキュメントの中に別のドキュメントを再帰的に埋め込むこともできる。

MongoDB はキー空間をコレクション内で分離しているので、たとえば Employees のキーと Department のキーが衝突することはない。CouchDB や Riak は、型の追跡を開発者側に任せている。ドキュメントの格納方法の自由さと複雑さは諸刃の剣である。アプリケーションの開発者はドキュメントのモデリングに関してかなりの自由を与えられているが、アプリケーション側での問い合わせのロジックは著しく複雑化する。

⁷<http://www.oracle.com/technetwork/database/berkeleydb/overview/index.html>

⁸<http://redis.io/>

⁹<http://couchdb.apache.org/>

¹⁰<http://www.mongodb.org/>

¹¹http://www.basho.com/products_riak_overview.php

BigTable カラムファミリーストア

HBase や Cassandra は、Google の BigTable が使っているモデルをベースにしたデータモデルを採用している。このモデルでは、キーがひとつの行を指し示す。その行に含まれるデータは、ひとつあるいは複数のカラムファミリー (CF) に格納されている。CF の中で、各行は複数の列を保持できる。列内の値にはタイムスタンプがついており、複数のバージョンの行・列マッピングを一つの CF に保持することができる。

概念的には、カラムファミリーを次のようにとらえることができる。つまり、ある形式(行 ID、CF、列、タイムスタンプ)の複合キーを格納し、それをキーで並べ替えた複数の値にマップするというものだ。この設計が、多くの機能をキー空間に持たせるというデータモデリングにつながる。この方式は、タイムスタンプを持つヒストリカルなデータのモデリングに適している。もちろん、このモデルはスペース列に対応している。というのも、何も列を持たない行 ID が、必ずしも各列に NULL 値を持つ必要がないからである。その反面、NULL 値をほとんど(あるいはまったく)持たない列でも各行に列 ID が必要となる。そのため、要する領域は多くなる。

どのプロジェクトのデータモデルもオリジナルの BigTable のモデルとはいろいろな面で異なるが、中でも特筆すべきなのが Cassandra における変更だろう。Cassandra は、各 CF の中にスーパーカラムという概念を導入した。スーパーカラムを使って、違ったレベルでのマッピングやモデリングそしてインデキシングを行えるようにしたのである。また、複数のカラムファミリーをひとまとめに格納してパフォーマンスを稼ぐ仕組みであるローカルグループという概念を廃止した。

グラフストレージ

NoSQL のデータ格納方式のひとつに、グラフストレージがある。データを作る方法はただ一つとは限らないし、リレーションナルモデルやキー指向のモデルがデータの格納や問い合わせに対して常に最適であるとは限らない。グラフは計算機科学では基本的なデータ構造であり、HyperGraphDB¹²や Neo4J¹³はグラフ構造のデータを格納する NoSQL ストレージシステムとしてよく知られている。グラフストレージは、これまで取り上げた他のストレージとはあらゆる点で異なる。データモデル、データの走査や問い合わせのパターン、ディスク上の物理的なデータ配置、複数マシンへの分散、クエリのトランザクション特性などがすべて異なるのだ。このように全く異なるものを公正に評価するにはページが足りない。しかし、これだけは意識しておこう。ある種のデータに関しては、グラフとして格納したほうがずっとうまく扱えるということを。

¹²<http://www.hypergraphdb.org/index>

¹³<http://neo4j.org/>

複雑な問い合わせ

NoSQL システムにおけるキーだけを使ったロックアップには、特筆すべき例外がある。MongoDB では任意の数のプロパティを使った索引付けができるようになっており、比較的高レベルの問い合わせ言語を使って取得したいデータを指定することもできる。BigTable ベースのシステムでは、スキーマーを使ったカラムファミリーの反復処理に対応しており、特定のアイテムを洗濯する際に、カラム上での絞り込みができる。CouchDB ではデータに対してさまざまなビューを作ることができ、テーブルに対して MapReduce タスクを実行させ、より複雑なロックアップや更新もできるようにしている。ほとんどのシステムには Hadoop あるいはその他の MapReduce フレームワークへのバインディングがあり、データセットに対して解析的な問い合わせを実行できる。

トランザクション

NoSQL システムは全般的に、**トランザクションの特性**よりもパフォーマンスを重視している。SQL ベースのシステムでは、複数の文の組み合わせ—主キーを指定して行を取得するといった単純な処理から、複数のテーブルを連結していくつかのフィールドの平均を算出するなどの複雑な処理まで—をひとつのトランザクションにまとめられるようになっている。

SQL データベースは、トランザクション間での ACID を保証している。複数の操作をひとまとめにして実行するトランザクションは原子性 (Atomic: ACID の A) がある。つまり、すべての操作が行われるか、なにも行われないかのいずれかになる。整合性 (Consistency: ACID の C) が保証するのは、そのトランザクションがデータベースの一貫性を保ち、状態を破壊しないということだ。独立性 (Isolation: ACID の I) とは、二つのトランザクションが同時に同じレコードを操作したとしてもお互い相手側に影響を及ぼさないということである。永続性 (Durability: ACID の D、次の節で詳述する) が保証するのは、トランザクションが確定したらそれが安全な場所に格納されるということだ。

ACID 準拠のトランザクションは、開発者に安心を与えてくれる。データの状態の確認が容易になるからだ。こんな場面を想像してみよう。複数のトランザクションが並行稼働しており、それぞれが複数のステップからなる処理をしている（たとえば、まず銀行口座の残高を確認してその次に \$60 を引き落とし、最後に値を更新するなど）。ACID 準拠のデータベースはこれらの処理順序に関して何らかの制限がかかることが多いが、すべてのトランザクションで正しい結果が得られる。正確さを重視した結果、パフォーマンス特性に予期せぬ影響が出ることが多い。処理が遅いトランザクションが一つあるせいで、他のトランザクションもそれにあわせて処理を待つ羽目になるといったものだ。

大半の NoSQL システムは、ACID に完全に準拠することよりもパフォーマンスを向上させることを優先している。しかし、キーのレベルでは ACID を保証しており、同じキーに対する二つの操作があればそれは直列化され、キーと値のペアに深刻な被害が及ばないようにしている。多くのアプリケーションではこのレベルで十分で、データの正確性に目立った問題

が発生することもない。また、より規則性のある操作を素早く実行できるようになる。しかしこの方針のおかげで、アプリケーションの設計やデータの正確性に関して開発者側で考慮しなければならない点はより多くなる。

トランザクションを軽視する傾向に反する例外として特筆すべきなのがRedisだ。単一サーバー上で、RedisはMULTIコマンドを提供する。これは、原子性と整合性を保証した状態で複数の操作を組み合わせて実行するものだ。またWATCHコマンドは独立性を保証する。それ以外のシステムでは、より低レベルな*test-and-set*¹⁴機能を提供しており、これで同様の独立性を保証している。

スキーマフリー ストレージ

多くのNoSQLシステムに共通する特徴が、データベース内でスキーマを強要しないということである。ドキュメントやカラムファミリーを格納する方式であっても、各エンティティのプロパティが同じである必要はない。その利点は、構造化されたデータに関してサポートすべき要件が減るということ。そして、スキーマをオンザフライで修正するときにもパフォーマンスの劣化はあまり起こらないということである。そのぶん、アプリケーションの開発者側にはより多くの責務が課せられる。より身構えたプログラムが必要になるのだ。たとえば、社員レコードにlastnameプロパティがなかったとして、それは修正すべきエラーなのだろうか。それとも、スキーマの更新がシステムを通じて伝搬している最中なのだろうか。*sloppy-schema*なNoSQLシステムを使うプロジェクトの場合、数回のイテレーションを終えた後のデータやスキーマのバージョン管理は、アプリケーションレベルのコードで行うことになる。

13.3 データの永続性

理想を言えば、データに何か変更があったときにはそれをすぐに安全な場所に永続化させたいし、複数の場所にレプリカを作るなどしてデータのロスを防ぎたい。しかし、データの安全性を保証しようとするとパフォーマンスに影響が及ぶ。そこで、各種NoSQLシステムはそれとは異なる手法で**データの永続性**を保証しつつパフォーマンスを向上させている。データを失う場面にはさまざまなものがあるし、すべてのNoSQLシステムがこういった問題からあなたを守ってくれるというわけでもない。

いちばんシンプルかつありがちなシナリオは、サーバーの再起動や停電などだ。このような場合を想定してデータの永続性を確保するには、データをメモリからハードディスクに移すことになる。ハードディスクは、電源を落としてもデータを失わない。ハードディスク障害に対応するには、データを別のデバイスにコピーする。コピー先は、同一マシン上の別のハードディスク(RAIDミラー)だったりネットワーク上の別のマシンだったりする。しかし、データセンターも、ハリケーンなどの自然災害にあれば使えなくなる可能性がある。そこで、

¹⁴[訳注]「更新前に確かめる」

ひとつのハリケーンで同時に被害にあうことのない程度に離れた場所にあるデータセンターにバックアップを取るという組織も存在する。データをハードディスクに書き込んで複数のサーバーやデータセンターにコピーするという作業は高くつく。そこで、さまざまな NoSQL システムはデータの永続性の保証とパフォーマンスを天秤にかけてバランスを取っている。

单一サーバーの永続化

永続化の型として最もシンプルなのが**单一サーバーの永続化**で、サーバーを再起動したり電源を落としたりしても変更したデータが生き残ることを保証する。通常これは、変更したデータをディスクに書き込むことを意味する。そしてこの処理がボトルネックになることが多い。ディスク上のファイルにデータを書き込むよう OS に指示を出したとしても、OS は書き込みをバッファリングしてすぐには書き込まないことがある。複数の書き込み操作を一括処理するためである。`fsync` システムコールを実行すれば、バッファにたまつた更新を OS ディスクに永続化させようと試みる。

一般的なハードディスクドライブの性能は、一秒あたり 100 から 200 のランダムアクセス(シーク)といったものであり、シーケンシャルライトもたかだか 30-100 MB/sec 程度である。どちらについても、メモリのほうが桁違いに高速である。单一サーバーでの永続化を保証する効率を上げるには、あなたのシステムによるランダムライトの回数を減らし、ハードディスクごとのシーケンシャルライトの回数を増やすようにすればよい。理想的には、一回の `fsync` コールあたりの書き込み回数を最小化してシーケンシャルライトの回数を最大化したいものだ。そして、書き込みを `fsync` させるまではユーザーに対して書き込み成功を伝えないようにしておきたい。单一サーバーでの永続化を保証するときにパフォーマンスを改善するためのテクニックを、いくつか紹介する。

`fsync` の頻度の制御

Memcached¹⁵は、ディスク上での永続化の保障を放棄する引き換えとして、極めて高速なインメモリでの操作を提供するシステムのひとつである。サーバーを再起動すると、memcached 上のデータは消えてしまう。つまり、キャッシングとしてはよくできているが永続化には難があるデータストアとなる。

Redis の場合は、どのタイミングで `fsync` をコールするかについていくつかのオプションを選べるようになっている。更新のたびに `fsync` をコールするような設定もできる。これは、低速だが安全な選択肢となる。よりパフォーマンスを稼ぐために、N 秒おきに書き込みを `fsync` させることも可能だ。この場合、最悪で N 秒ぶんの操作をロストしてしまうことになるが、その程度なら受け入れられるという使い方もあるだろう。最後に、永続化を重視しないような場面(おおざっぱな統計情報の保守や、Redis をキャッシングとして使うといった場面)では、

¹⁵<http://memcached.org/>

`fsync` をまったくコールしないようにもできる。適当なタイミングで OS がデータをディスクに書き込むだろうが、それがいつ発生するかはまったく保証しないという選択肢だ。

ログ出力によるシーケンシャルライトの増加

NoSQL システムがディスクから高速にデータを取得するために、B+木などのデータ構造が使われている。こういったデータ構造のデータを更新するときには、ファイル内のランダムな場所を更新する。更新のたびに `fsync` しようとすると、一回の更新に対して複数のランダムライトが発生することになる。ランダムライトを減らすために、Cassandra や HBase、Redis、そして Riak といったシステムは、更新操作を *log* というファイルにシーケンシャルに書き込んでいる。システムで使っている他のデータ構造は言っての期間ごとに `fsync` するのに対して、ログだけは頻繁に `fsync` を行う。データベースがクラッシュしたときにはログを正式な状態として扱うことで、ランダムな更新をシーケンシャルライトにまとめている。

NoSQL システムの中には、MongoDB のようにその場でデータ構造に書き込みを行うものもあれば、さらにロギングを行うものもある。Cassandra や HBase が使っているテクニックは BigTable を参考にしたもので、ログとルックアップデータ構造をひとつの *log-structured merge tree* にまとめている。Riak は、同様の機能を *log-structured hash table* で提供する。CouchDB は伝統的な B+木に手を加えたものを使っており、すべての変更を物理ストレージ上の構造に追記する。これらのテクニックによって書き込みのスループットは向上するが、定期的にログの最適化をしないとログのサイズがどんどん膨れ上がってしまう。

書き込みのグルーピングによるスループットの向上

Cassandra は、複数の更新を並行してまとめ、一回の `fsync` コールの間に実行する。このような設計は**グループコミット**と呼ばれており、更新あたりの待ち時間が長くなってしまう。ユーザーによる更新が受け付けられたかどうかを知るには、並行するいくつかの更新が完了するまで待たなければならない。待ち時間が増える一方で、これはスループットの向上につながる。複数のログ追記処理が一回の `fsync` で処理されるからだ。本章の執筆時点では、HBase の更新は Hadoop Distributed File System (HDFS)¹⁶が提供するストレージに永続化される。最近適用されたパッチで、`fsync` やグループコミットを尊重する追記もサポートするようになった。

複数サーバーの永続化

ハードディスクドライブだってマシンだって、壊れてしまって復旧不能になることがある。重要なデータは別のマシンにもコピーしておくことが必須である。多くの NoSQL システムには、複数のサーバーを使ってデータを永続化する仕組みが存在する。

¹⁶<http://hadoop.apache.org/hdfs/>

Redis は、伝統的なマスター/スレーブ型の手法でデータを複製する。マスターに対してなされたすべての操作がログ風の仕組みでスレーブ機に送られ、スレーブ機の上で同じ操作を再現させる。マスター上で障害が発生したら、マスターから受け取ったオペレーションログの状態に基づいてスレーブがデータを提供することになる。この構成では、何らかのデータロスが発生する可能性がある。マスターへの更新の結果をユーザーに返す前に、スレーブへのログの永続化が完了したかどうかを確認しないからである。CouchDB は、同様の形式で双方向のレプリケーションを行う。ドキュメントに対する変更を他のマシンに複製するよう、サーバーを設定するのだ。

MongoDB にはレプリカセットという仕組みがあり、何台かのサーバーで各ドキュメントの格納にかかる。MongoDB のオプションで、すべてのレプリカが更新を受け取ったことを保証させることもできる。一方、最新のデータがすべてのレプリカに行きわたるのを確認せずに処理を進めることもできる。その他多くの分散型 NoSQL ストレージシステムは、データのマルチサーバーレプリケーションに対応している。HDFS 上に構築される HBase は、複数サーバーの永続化を HDFS 経由で実現する。すべての書き込みは、ふたつ以上の HDFS ノードに複製されるまでユーザーに制御を返さない。これによって、複数サーバーでの永続化を保証している。

Riak や Cassandra そして Voldemort では、より細やかにレプリケーションを設定できる。それぞれ微妙な違いはあるが、これらのシステムでは N と W のふたつの値を設定できる。 N は最終的にデータのコピーを保持することになるマシンの台数、そして W は $W < N$ を満たす数で、少なくともこれだけの台数のマシンにデータが書き込まれた時点でユーザーに制御を戻す。

データセンター全体のサービスが停止してしまう事態に対応するには、複数のデータセンターにまたがるマルチサーバーのレプリケーションが必要になる。Cassandra や HBase そして Voldemort には *rack-aware* な設定があり、さまざまなマシンがどのラック (あるいはどのデータセンター) に配置されているのかを指定することができる。一般に、リモートサーバーでの処理が完了するまでユーザーのリクエストをブロックすると、待ち時間が長くなってしまう。そこで、WAN による別のデータセンターへのバックアップのときは、処理の完了を確認せずに更新処理を終える。

13.4 パフォーマンス向上のためのスケーリング

エラー処理について語る前に、もっと楽観的な状況を考えてみよう。やった!大成功!という場面だ。あなたが構築したシステムがうまく動き出すと、データストアはそのコンポーネントのひとつとなり、それなりの負荷にさらされることになる。お手軽だがあまり美しくない解決策は、既存の機器を **スケールアップ** することだ。RAM とディスクをさらに調達して、ひとつのマシンでさばける量を増やせばよい。あなたのシステムがさらに成功を収めると、ハードウェアをよいものにして高価なメモリをどんどん投入することにも限界が出てくる。ここまでくると、データをレプリケートして複数マシンで負荷分散をさせるしか方法がなく

なる。これは**スケールアウト**とよばれる手法で、システムの**水平スケーラビリティ**の指標となる。

水平スケーラビリティの理想的な目標は**リニアなスケーラビリティ**、つまり、ストレージシステムのマシン数を二倍にすればそのシステムのクエリ処理性能も二倍になるというものだ。これを実現するためのポイントは、データを複数マシンにどのように振り分けるかということになる。シャーディングという手法は、読み込みと書き込みを複数のマシンに分散させてストレージシステムをスケールアウトさせるものである。シャーディングは多くのシステムで設計の基本となっている。具体的には Cassandra や HBase、Voldemort、Riak、そして最近では MongoDB や Redis もそうだ。中には、CouchDB のように単一サーバーでのパフォーマンスを重視してシステムではシャーディング機能を提供しないというプロジェクトもある。しかし、セカンダリプロジェクトがコーディネーターとなって、複数のマシンにそれぞれ独立にインストールした環境に処理を振り分けることもできる。

ここで、いくつかの用語についてまとめておこう。ここでは、**シャーディング**と**パーティショニング**同じ意味で使う。また、**マシン**や**サーバー**そして**ノード**は、分割されたデータを格納する物理的な計算機を指すものとする。最後に、**クラスタ**あるいは**リング**という用語は、ストレージシステムを構成するマシン群を指すものとする。

シャーディングするということは、どのマシンもそれ単体ではデータセット上のすべての書き込みを処理する必要がなくなるということだ。しかしそれと同時に、どのマシンもそれ単体ではデータセットへのすべての問い合わせには対応しきれないということでもある。多くの NoSQL システムではキー指向のデータモデルやクエリモデルを採用しているので、いずれにせよデータセット全体にまたがるような問い合わせはほとんどない。これらのシステムではデータにアクセスする主要な方法がキーに基づいているので、シャーディングもまたキーに基づいて行うのがよい。キーに対する何らかの関数が、そのキー・バリューペアをどのマシンに格納するのかを決定する。キーとマシンのマッピング方法を二通り紹介しよう。ハッシュパーティショニングとレンジパーティショニングだ。

必要になるまでシャーディングを避ける

シャーディングはシステムを複雑化させるものであり、可能な限り避けるべきである。ここでは、シャーディングを使わずにスケールさせる方法を二通り取り上げる。リードレプリカとキャッシュだ。

リードレプリカ

ストレージシステムの多くは、読み込みリクエストのほうが書き込みリクエストより多くなる。そんな場合のシンプルな解決策は、データのコピーを複数のマシン上に置くことだ。書き込みリクエストはすべてマスターノードに任せる。読み込みリクエストはデータのレプ

リカを持つマシンにまわす。ただしこのレプリカは、書き込みサーバー上のデータよりも若干古いものになることが多い。

もし既にマスター・スレーブ公正で複数サーバーでのデータの永続化を実現している (Redis や CouchDB そして MongoDB ではそれが一般的) のなら、読み込み用のスレーブが書き込み用のマスターの負荷を多少軽減させることができる。ある種のクエリ、たとえばデータセットのサマリーの集計などは、コストのかかる処理である一方で必ずしも最新のデータを必要とはしないことがある。そんなクエリは、スレーブのレプリカに対して実行すればよい。一般に、データが最新である必要性が少なければ少ないほど、その処理を読み込み用スレーブに任せてクエリのパフォーマンスを稼ぎやすくなる。

キャッシュ

システム上でよく使われるコンテンツをキャッシュすると、たいていの場合は驚くほどうまく機能する。Memcached は、複数サーバー上にメモリロックを確保してデータストアのデータをキャッシュする。Memcached クライアントは、水平スケーラビリティの技を使って別のサーバー上にある Memcached に負荷を分散する。キャッシュプールにメモリを追加するには、単に Memcached が動くホストを追加するだけでよい。

Memcached はキャッシュ用に設計されているので、処理をスケールさせるための永続化ソリューションのアーキテクチャはそれほど複雑ではない。複雑なソリューションの前に、キャッシュでスケーラビリティの問題を解決できないかどうかを検討してみよう。キャッシュ処理は単なるその場しのぎのバンドエイドではない。Facebook では Memcached で何と数十テラバイトものメモリを各 h おしているのだってさ!

リードレプリカやキャッシュを使えば、読み込み処理をスケールアップさせることができます。しかし、書き込みやデータ更新の頻度が上がり始めたら、最新状態を保持するマスターサーバーへの負荷が増加することになる。このセクションの後半では、書き込み処理を複数サーバーにシャーディングする方法を扱う。

コーディネーターによるシャーディング

CouchDB プロジェクトは、单一サーバー上での挙動を重視している。Lounge と BigCouch のふたつのプロジェクトは外部のプロキシを通じて CouchDB への負荷をシャーディングし、単独の CouchDB インスタンスのフロントエンドとして機能する。この構成では、個々の CouchDB がお互いを意識することはない。コーディネーターが、リクエストされたドキュメントのキーに応じて個々の CoucdDB インスタンスにリクエストを分散させる。

Twitter は、シャーディングやレプリケーションの概念をまとめた Gizzard¹⁷ というフレームワークを構築した。Gizzard は、任意の型のスタンドアロンデータストア (SQL システムあるいは NoSQL システムのラッパーを作れる) を受け取り、キーのレンジでパーティショニング

¹⁷<http://github.com/twitter/gizzard>

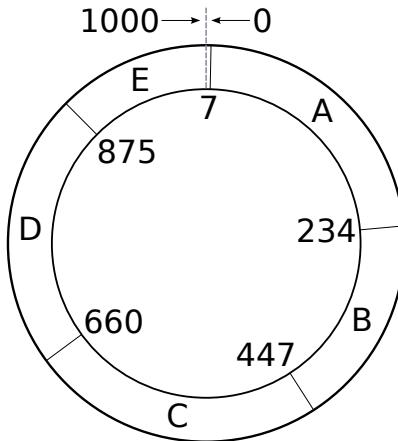


図 13.1: 分散ハッシュテーブルリング

して任意の深さのツリーとしてまとめる。耐障害性を高めるため、Gizzard は同じキーレンジのデータを複数の物理マシンにレプリケートするようにも設定できる。

コンシスティントハッシュリング

よくできたハッシュ関数は、キーのセットを統一された形式で分散させる。これは、キー・バリューのペアを複数のサーバーに分散させるための便利なツールとして使える。学術論文上で**コンシスティントハッシュ**と呼ばれる技術は広範囲にわたる。このテクニックをデータストアに応用したはじめての例が**分散ハッシュテーブル(DHT: Distributed Hash Tables)**というシステムだ。Amazon の Dynamo の動作原理を参考にした NoSQL システムではこの分散テクニックを採用しており、Cassandra や Voldemort そして Riak などにそれが見られる。

ハッシュリングの実例

コンシスティントハッシュリングは、次のように動作する。まず、ハッシュ関数 H があるとしよう。この関数は、キーを均等に分布する大きな整数値にマップする。比較的大きな整数値 L をとって範囲 $[1, L]$ の数のリングを作れば、 $H(\text{key}) \bmod L$ をそのリングに含めることができる。これで、各キーが $[1, L]$ の範囲に収まることになる。サーバーのコンシスティントハッシュリングは、各サーバーの固有な識別子(その IP アドレスなど)に H を適用したものを使って構成される。この動作原理を理解しやすくするために、5 台のサーバー (A-E) からなるハッシュリングの例を図 13.1 に示す。

ここでは、 $L = 1000$ とした。 $H(A) \bmod L = 7$ 、 $H(B) \bmod L = 234$ 、 $H(C) \bmod L = 447$ 、 $H(D) \bmod L = 660$ 、そして $H(E) \bmod L = 875$ であるとしよう。これで、キーがどのサーバーに配置されるかがわかるようになった。リング内で、あるサーバーとその次のサーバーの間

にキーが取まる場合に、そのサーバーにキーを格納することになる。たとえば、A が受け持つキーはそのハッシュが [7, 233] の範囲になるものであり、E が受け持つキーはそのハッシュが [875, 6] の範囲(これは、値が 1000 の部分をまたがっている)になるものである。つまり、もし $H('employee30') \bmod L = 899$ になるのならこのデータはサーバー E に格納されるし、 $H('employee31') \bmod L = 234$ になるとしたらこのデータはサーバー B に格納される。

データのレプリケーション

複数サーバーでの永続化のためにレプリケーションをするには、あるサーバーの担当範囲に割り当てられたキーと値のペアをリング内のその次のサーバーに渡せばよい。たとえば、三重のレプリケーションを行うには、範囲 [7,233] にマップされたキーをサーバー A、B、そして C に格納する。仮に A がダウンしたら、隣にある B と C がその範囲を受け持つ。設計によっては、E にレプリケートして A の担当分を一時的に受け持つこともある。この場合は担当範囲を拡張して A の範囲も含めるようとする。

よりよい振り分け

ハッシュはキー空間を均等分布させるという点では統計的に有効であるが、均等に分布させるには通常はある程度多くのサーバーを必要とする。残念ながら、たいていは少数のサーバーからスタートすることが多い。そしてその段階では、ハッシュ関数がうまくキーを分散させてはくれない。先ほどの例で見ると、A の担当範囲の長さは 227 であるのに対して E の担当範囲は 132 しかない。こんな状態だと、サーバーによって負荷が違うということになるだろう。また、どれかひとつのサーバーがダウンしたときに代わりを受け持つのも難しくなる。隣のサーバーが、ダウンしたサーバーの担当範囲全体を制御しなければならなくなるからである。

担当するキーの範囲にばらつきが出てしまう問題を解決するために、Riak を含む多くの DHT は、物理マシン単位でいくつか ‘仮想’ ノードを作成する。たとえば、4 つの仮想ノードを作ったサーバー A が、サーバー A_1、A_2、A_3、そして A_4 として動作するといった具合だ。各仮想ノードには異なるハッシュ値が割り当てられ、キー空間の各部分によりうまく分散させられるようにする。Voldemort も同様の手法を採用しており、パーティションの数を手動で設定できるようになっている。通常はサーバーの数よりパーティションの数を多くするので、結果的に各サーバーが複数のパーティションを受け持つことになる。

Cassandra は、各サーバーで複数の小さなパーティションを担当するということはしない。そのため、時にはキーの範囲の分散が均一にならないこともある。ロードバランス用として、Cassandra には非同期プロセスが用意されている。このプロセスは、これまでの負荷の履歴に基づいてリング上のサーバーの配置を調整する。

レンジパーティショニング

レンジパーティショニング方式によるシャーディングでは、システム内の何台かのマシンが「どのサーバーがどのキー範囲を受け持つか」というメタ情報を保持する。このメタ情報への問い合わせによって、キー範囲の検索と適切なサーバーへの振り分けを行う。コンシスティントハッシュリングの手法と同様、レンジパーティショニングもキー空間をいくつかの範囲に分割する。そして各範囲をひとつのマシンが管理し、場合によっては他のマシンにレプリケートしたりする。コンシスティントハッシュ方式と違うところは、キーのソート順で隣同士になるキーがほぼ同じパーティションに収まるという点だ。これにより、ルーティング用のメタデータのサイズを軽減できる。範囲を表すには単に [開始位置, 終了位置] の印があればよいだけだからである。

キー範囲とサーバーのマッピング情報を更新し続ける際に、レンジパーティショニング方式では高負荷なサーバーの負荷分散をよりきめ細やかに制御できるようになる。特定のキー範囲が他の範囲に比べてトランザクションが多くなるようなら、ロードマネージャーはそのサーバーが担当する範囲を狭めることもできるし、そのサーバーが担当するシャードの数を減らすこともできる。動的に負荷を調整できるという新たな自由を得るために使ったのは、シャードを監視したりルーティングしたりするための追加コンポーネントだ。

BigTable の手法

Google による BigTable に関する論文には、階層化レンジパーティショニングでデータをタブレットにシャーディングする手法が解説されている。一つのタブレットが、一定範囲の行のキーとカラムファミリー内の値を保持する。タブレットは必要なログをすべて保持し、自分が担当する範囲のキーに関する問い合わせに答えるためのデータ構造もすべて保持する。タブレットサーバーは、各タブレットにかかる負荷に応じて複数のタブレットを担当する。

各タブレットは、100 から 200MB のサイズを保つ。タブレットのサイズが変われば、隣り合うキー範囲の二つの小さなタブレットを一つにまとめたり、あるいは大きなタブレットを二つに分割したりする。マスターサーバーが、タブレットのサイズや負荷そしてタブレットサーバーの稼働状態を解析する。そして、マスターサーバーは、どのタブレットサーバーがどのタブレットを担当するのかを常時調節する。

マスターサーバーは、タブレットの割り当てをメタデータテーブルで管理する。このメタデータはかなり大きくなる可能性があるので、メタデータテーブル自体も複数のタブレットにシャーディングしてキー範囲とタブレットを関連づける。タブレットサーバーが、この範囲の管理を行う。その結果、クライアントは三階層の走査を経てキーの保存先のタブレットサーバーを知ることになる。その様子を図 13.2 に示す。

実際の例を見てみよう。クライアントがキー 900 を検索しようとすると、サーバー A に問い合わせを行う。このサーバーには、レベル 0 のメタデータ用のタブレットが格納されている。このタブレットを見れば、レベル 1 のメタデータがサーバー 6 上のタブレットにあること

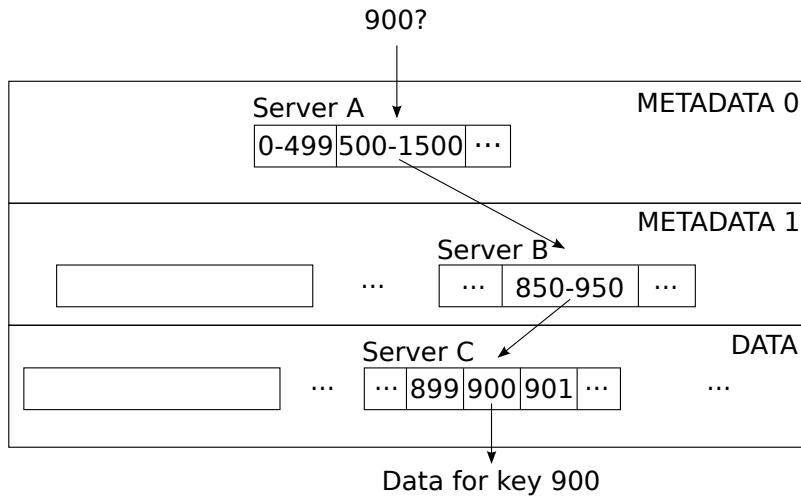


図 13.2: BigTable ベースのレンジパーティショニング

がわかる。このタブレットには 500-1500 の範囲のキーが含まれる。クライアントはサーバー B に対してこのキーでリクエストを送る。そして、キー 850-950 を含むタブレットがサーバー C 上にあるという応答を得る。最終的に、クライアントはそのキーについてのリクエストをサーバー C に送り、問い合わせ結果の行を取得する。レベル 0 とレベル 1 のメタデータタブレットはクライアント側でキャッシュしてもかまわない。そうすれば、タブレットサーバーに対する同じような問い合わせの繰り返しを回避できる。BigTable の論文では、この三段階の階層によって 2^{61} バイトのストレージを 128MB のタブレットで対応できるとしている。

障害の処理

BigTable の設計では、マスターが单一障害点になる。しかし、タブレットサーバーへのリクエストに影響が及ばないように一時的にダウンさせることも可能だ。タブレットへのリクエストを処理するタブレットサーバーがダウンすれば、マスターがそれを認識してそのタブレットの担当を切り替えるまではリクエストが一時的に失敗するようになる。

マシンの障害を認識して対応するための方法として、BigTable の論文では Chubby を使っている。これは分散ロックシステムで、サーバーの死活監視を行うものである。ZooKeeper¹⁸は Chubby のオープンソース実装で、Hadoop ベースのプロジェクトの中には、これを使ってセカンドリマスターサーバーやタブレットサーバーの再配置を行うものもある。

¹⁸<http://hadoop.apache.org/zookeeper/>

レンジパーティショニングベースの NoSQL プロジェクト

HBase は、BigTable の階層方式を使ってレンジパーティショニングを行う。ベースとなるタブレットのデータは Hadoop の分散ファイルシステム (HDFS) に格納する。HDFS がデータのレプリケーションやレプリカ間の整合性を管理する。そして、タブレットサーバー側ではリクエストの処理やストレージの更新、タブレットの分割や統合を管理する。

MongoDB も、BigTable と同様の方法でレンジパーティショニングを処理する。いくつかの設定ノードがルーティングテーブルを管理して、このルーティングテーブルを使ってどのストレージノードがどの範囲のキーを担当するかを決定する。設定ノード間の同期は**二相コミット**というプロトコルで行い、BigTable のマスターが受け持つ範囲指定機能と Chubby が受け持つ高可用性のための構成管理機能の両方を提供する。ステートレスで動くルーティングプロセスとは別に、直近のルーティング設定を追跡し続けることでキーへのリクエストを適切なストレージノードに振り分ける。ストレージノードはレプリカセット内に置かれ、レプリカセットでレプリケーションを行う。

Cassandra では順序を保持したパーティショニング機能を提供しており、データに対するレンジスキヤンを高速に行いたい場合に利用できる。Cassandra のノードもコンシスティントハッシュを使ってリング内に配置される。しかし、キー・バリューのペアをハッシュして割り当て先のサーバーを決めるのではなく、キーそのものを単純にサーバーにマップする。そうすることで、キーが必然的にフィットする範囲を制御できるようになる。たとえばキー 20 と 21 は、図 13.1 のコンシスティントハッシュリング上でどちらもサーバー A にマップされる。ハッシュしてリング内でランダムに分散させるわけではない。

Twitter の Gizzard フレームワークは、分割され、レプリケートされたデータをさまざまなバックエンドにまたがって管理するもので、レンジパーティショニングを使ってデータをシャーディングする。ルーティングサーバーは任意のレベルの階層構造を構成し、キー範囲をその階層下のサーバーに割り当てる。各サーバーは自分に割り当てられたキー範囲のデータを格納したり、別の層のルーティングサーバーに処理を振ったりする。このモデルにおけるレプリケーションは、あるキー範囲の更新を複数のマシンに送信することで実現する。Gizzard のルーティングノードは、書き込みに失敗したときの対応方法が他の NoSQL システムとは異なる。Gizzard を使ったシステムでは、すべての更新が幂等 (何度も実行しても結果が変わらない) でなければならない。ストレージノードがダウンしているときは、ルーティングノードがその処理をキャッシュし、更新が確認できるまで何度もその処理を送信し続ける。

どのパーティショニング方式を採用するか

ハッシュ方式とレンジ方式、シャーディングの手法としてどちらが適切なのかって? それは状況による。明らかにレンジパーティショニングを選ぶべきなのは、たとえばキーによる検索よりも範囲指定による検索が多発するような場面だ。キーの順に値を読み込むことになるので、そうしておけばネットワーク上のあちこちのノードを飛び回る必要がなくなる。ノード

ドを移るときのネットワークのオーバーヘッドは馬鹿にできない。しかし、範囲指定の検索が不要な場合には、どちらの方式を選べばよいのだろう？

ハッシュパーティショニングを使えば、データを複数のノードに適切に分散させることができる。そして、データの非対称性も仮想ノードで軽減できる。ハッシュパーティショニング方式では、ルーティングもシンプルになる。ほとんどの場合は、クライアント側でハッシュ関数を実行すれば問い合わせ先のサーバーを確定できる。バランスを調整するための仕組みを導入している場合は、あるキーに対して適切なノードを見つけるのは少し難しくなる。

レンジパーティショニングは、ノードのルーティングや構成を管理するための事前コストが必要となる。この処理の負荷は高くなりがちで、耐障害性をきちんと考慮していないと障害の主因になるだろう。しかし、うまくやれば、レンジパーティショニングで分割したデータは小さいチャンクで負荷分散できるようになり、負荷が高くなったときの調整も可能となる。かりにひとつのサーバーがダウンしたとしても、そこに割り当てられていた範囲を多数のサーバーに分散させることができ、ダウンしたサーバーだけに負荷をかけることはない。

13.5 整合性

これまで、「データを複数マシンにレプリケートすれば永続化できるし負荷分散もできる」と、その利点だけを説明してきた。このあたりでその実態も書いておこう。複数マシンにデータのレプリカを置いてお互いの整合性を保つというのは、大変な作業だ。実際のところ、レプリカが壊れて同期できなくなることもあるだろうし壊れたデータを復旧できなくなることもあるだろう。ネットワーク上で複数のレプリカセットができてしまったり、マシン間のメッセージが遅延したり途中で消えてしまったりといったこともあり得る。NoSQLの世界でデータの整合性を保つための方法としてよく使われるのは、次の二通りの手法だ。ひとつは強整合性 (strong consistency) で、これはすべてのレプリカを同期させる。もうひとつが結果整合性 (eventual consistency) で、こちらの方式ではレプリカが同期されていなくてもかまわないが、最終的にはお互い相手側の状態に追いつけるようにしておかなければならない。まず最初に、どんな場合に結果整合性が選択肢に入るのかを分散コンピューティングの本質から考える。その後で、それぞれの手法の詳細を見ていく。

CAPについて

データに対する強整合性を保証できなくなることについて、なぜそんなに考慮するのだろう？ すべては、今どきのネットワーク機器上で構築された分散システムの特性によるものだ。Eric Brewer が初めて提唱した CAP 定理は、後に Gilbert と Lynch [GL02] によって証明された。この定理では、まず最初に分散システムの三つの特性を提示する。この頭文字をまとめた頭字語が CAP である。

整合性 (Consistency): あるデータのすべてのレプリカについて、どれを読んでも同じバージョンのデータを得られるか?(ここでいう整合性は、ACID の C とは異なる)

可用性 (Availability): アクセス不能なレプリカがいくつあっても、読み書きのリクエストに対応できるか?

耐分断性 (Partition tolerance): レプリカの一部が一時的にネットワーク上で他と分断されたときに、それでもシステムを稼働させ続けられるか?

そして、定理はさらにこう続く。複数台のコンピュータで構成されるストレージシステムは、この三つのうちの二つまでしか達成できず、その二つを達成するためには残りの一つが犠牲になってしまうのだ、と。また我々は、耐分断性を保証するシステムを実装せざるを得ない。現状のネットワーク機器やメッセージングプロトコルでは、パケットをロストすることもあればスイッチが故障することもある。そして、ネットワークがダウンしていたりメッセージの送信先のサーバーが存在しなかつたりといったことを知るすべもない。すべての NoSQL システムで、耐分断性が必須となる。残された選択肢は、整合性と可用性のどちらを妥協するかである。両方を保証できるような NoSQL システムは存在しない。

整合性を保証するというのは、レプリカ間でのデータの同期をきちんと保つということだ。これを実現する簡単な方法は、すべてのレプリカに対する更新を確認することである。もしどれか一つのレプリカがダウンして更新確認を受け取れなかった場合は、そのキーの可用性を下げる。つまり、ダウンしたレプリカが復旧して確認の応答を返すまでは、その更新処理が成功したとは見なさないということだ。つまり、整合性を保証しようとすると、各データアイテムに対する 24 時間体制の可用性は保証できなくなる。

可用性を保証するというのは、ユーザーが何らかの操作を実行したときには、他のレプリカの状態がどうであるかにかかわらず自身が持つデータ上で操作を受け付けなければならぬということだ。これは、レプリカ間でのデータの整合性を失ってしまうことにつながる。というのも、各レプリカにすべての更新が行き渡っているかどうかの確認がないので、中には一部の更新を受け取っていないレプリカも発生しうるからである。

CAP 定理からの帰結として、強整合性あるいは結果整合性のいずれかの手法で NoSQL データストアが作られることになった。それ以外の手法も存在して、たとえば Yahoo! の PNUTS [CRS⁺⁰⁸] システムでは、緩い整合性と緩い可用性 (relaxed consistency and relaxed availability) という手法をとっている。しかし、本章で扱っているオープンソースの NoSQL システムの中にはまだこの手法を採用しているものが存在しないので、本章ではこの手法は扱わない。

強整合性

強整合性を謳うシステムが保証するのは、データ項目のレプリカが常にそのキーの値を保持しているということだ。レプリカの中には、他のレプリカとの同期がとれなくなっているものもあるかもしれない。しかしそんな場合であっても、たとえばユーザーが employee30:salary

の値を問い合わせれば、ユーザーに返す値は常に一貫性のあるものとなる。その原理を、数字で説明しよう。

ひとつのキーを N 台のマシンにレプリケートしたものとする。あるマシン、おそらく N 台のなかのどれかが、コーディネーターとしてユーザーからのリクエストを処理する。このコーディネーターは、N 台のマシンのうちの一定台数以上が各リクエストを受け付けたことを保証する。あるキーに対する書き込みや更新があれば、少なくとも W 台のマシンがその更新を処理し終えたことを確認するまでコーディネーターはユーザーに応答を返さない。ユーザーが何らかのキーの値を読もうとしたときには、少なくとも R 台から同じ値を受け取るまでコーディネーターは応答を返さない。このとき、 $R+W>N$ であればシステムの強整合性を実証できるものとする。

この考え方を、実際に数字を入れて確認しよう。各キーを $N=3$ でレプリケートする(それぞれ A、B、C とする)。キー `employee30:salary` の初期値は \$20,000 だったが、ここで `employee30` を \$30,000 に昇給させることになった。要件として、 $W=2$ つまり A、B、C のうちで少なくとも 2 台が書き込みリクエストを受け付けることとする。このとき A と B が書き込みリクエスト (`employee30:salary, $30,000`) を受け付けると、コーディネーターはユーザーに対して `employee30:salary` が更新できたと伝える。マシン C が仮に `employee30:salary` へのリクエストを受け取れなかったとしよう。つまり、マシン C の値は \$20,000 のままである。ここでコーディネーターがキー `employee30:salary` に対する読み込みリクエストを受け取ると、そのリクエストを 3 台のマシンすべてに送信する。

- もし仮に $R=1$ で最初に応答したのが C だったとしたら、結果は \$20,000 となってこの社員はあまりうれしくないだろう。
- しかし、もし $R=2$ にしておけばコーディネーターは最初に C の値を受け取っても A あるいは B からの二番目の応答を待ち続ける。どちらが先に来ても先ほどの C の値と食い違っているのでさらに待ち続け、最終的に三番目のマシンからの応答を受け取った時点で \$30,000 が多数派であることを確認できる。

したがって、この場合に強整合性を満たすには、 $R \geq 2$ にして $R+W \geq 3$ を満たす必要がある。書き込みリクエストで W 台のレプリカから応答が返ってこなかったり、読み込みリクエストで一貫性のある結果が R 台以上のレプリカから得られなかったりした場合には何が起こるのだろう? コーディネーター側としては、最終的にタイムアウトを発生させてユーザーにエラーを返すこともできるし、要件を満たすまでずっと待ち続けることもできる。どちらにしても、そのリクエストについては少なくとも一定期間はアクセス不能とみなされる。

R と W をどのように設定するかによって、何台のマシンが不調になってしまってキーに対するさまざまな操作が可能になるかが決まる。たとえば書き込み操作はすべてのレプリカにきちんと反映させたいのなら、 $W=N$ とすることになる。この場合、どれか一台でもレプリカが応答しなければ、書き込みはハングしたり失敗したりする。よくある選択肢は $R + W = N + 1$ とするもので、こうすれば強整合性を維持するために最低限必要な稼働台数を最小に抑えられ

る。多くの強整合性システムは $W=N$ そして $R=1$ という設定を選んでいる。そうすれば、同期に失敗したノードをどうするかを考えずに済むからである。

HBase は HDFS 上でストレージをレプリケートする。これは分散型のストレージ層だ。HDFS は強整合性を保証する。HDFS では、全 N 台(通常は 2 あるいは 3)のレプリカに書き込み終えるまで書き込みは成功しない。つまり $W = N$ である。読み込みはひとつのレプリカだけで応答できるので、 $R = 1$ となる。大量の書き込みによるダウンを避けるため、ユーザーから各レプリカへのデータの転送は、非同期で並列処理される。すべてのレプリカがデータのコピーを受け取ったら、最後にシステム上のデータを新しいものに置き換える処理が行われる。これはアトミックな処理で、全レプリカの整合性を保つたものだ。

結果整合性

Dynamo ベースのシステムである Voldemort や Cassandra そして Riak などでは、ユーザーが必要に応じて N や R 、 W を指定できるようにしている。 $R + W \leq N$ であってもかまわない。つまり、強整合性と結果整合性のどちらを達成するのかをユーザーが選択できるということだ。ユーザーが結果整合性を選択した場合、仮にプログラマーが強整合性を望んだとしても、 $W < N$ ならレプリカを安心して扱えない。レプリカ間での結果整合性を提供するには、システム側でさまざまなツールを使ってレプリカを最新状態に保つことになる。まずは、さまざまなシステムがどのようにしてデータが古くなったことを検出するのかを見ていこう。それから、レプリカを同期する方法を考える。そして最後に、同期プロセスを高速化するための Dynamo の影響を受けた方法をいくつか紹介する。

バージョニングと衝突

あるキーについて、二つのレプリカが違うバージョンの値を返す可能性がある以上、データのバージョン管理や衝突の検出が重要になる。Dynamo ベースのシステムで使っているバージョニング方式が**ベクタークロック**だ。ベクタークロックとは、各キーにベクターを割り当て、そこにレプリカのカウンタを含める方式である。たとえば、何らかのキーのレプリカを A と B そして C で扱うとする。このときベクタークロックには三つのエントリ (N_A , N_B , N_C) があり、その初期値は $(0, 0, 0)$ となる。

レプリカ上でキーの値が変更されるたびに、ベクター内でそれに対応するカウンタが加算される。直前のバージョンが $(39, 1, 5)$ だったときに B がキーの値を変更すると、ベクタークロックは $(39, 2, 5)$ に書き換わる。別のレプリカ、たとえば C が B からそのキーのデータの更新を受け取るときには、B からのベクタークロックを自分のものと比較する。自分のベクタークロックカウンタのほうが B から受け取ったものよりも小さい場合、自分のデータは古いバージョンなのであるから B の内容で上書きできる。B より C のほうが大きいカウンタとその逆のカウンタが両方ある場合、たとえば $(39, 2, 5)$ と $(39, 1, 6)$ などのようになつた場合は、矛盾する更新があったとサーバーが判断して衝突したとみなす。

衝突の解決

衝突の解決方法は、システムによってさまざまである。Dynamo の論文では、衝突の解決はストレージシステムを使うアプリケーション側に任せている。二つのバージョンのショッピングカートを一つにまとめるのはそれほど面倒ではないだろうが、共同作業で編集していた文書の複数バージョンをまとめる際の衝突は人間のレビューがないと解決できないだろう。Voldemort はこのモデルを採用しており、あるキーに対して複数のコピーを返し、クライアントアプリケーション側での対応を求める。

Cassandra は各キーのタイムスタンプを格納しており、二つのバージョンが衝突する場合は一番タイムスタンプの新しいバージョンを採用する。これによってクライアントとのやりとりの必要をなくし、API を単純化している。この設計では、衝突したデータを先ほどのショッピングカートの例のように自動マージすることは難しいし、分散カウンタを実装するのも困難だ。Riak は、Voldemort と Cassandra のどちらの手法でも使える。CouchDB はハイブリッド方式だ。衝突を検出したら、ユーザー側でそのキーを手動で修復させるよう問い合わせ、衝突が解決するまでは特定のバージョンを確定的に採用してユーザーに返すようになっている。

リードリペア

R 台のレプリカが衝突していないデータをコーディネーターに返せたら、コーディネーターはその衝突していない値を安全にアプリケーションに返せる。それでもコーディネーターは、同期ができていないレプリカがあることを検出するかもしれない。そんな場合に使えるテクニックとして Dynamo の論文で提案されており、Cassandra や Riak そして Voldemort が実装しているのがリードリペアだ。コーディネーターが読み込み時に衝突を検出すると、たとえ整合性のある結果をユーザーに返せたとしても、コーディネーターは衝突したレプリカの衝突解決プロトコルを開始する。これで、追加作業を最小限にしながら能動的に衝突を解消できる。各レプリカは既に自分の持つデータをコーディネーターに送っているので、早期に衝突を解決すればシステム内でのデータの相違を抑えられる。

Hinted Handoff

Cassandra や Riak そして Voldemort はすべて、*Hinted Handoff* というテクニックを使っている。これは、どれか一つのノードが一時的に使えなくなっている状態での書き込みのパフォーマンスを向上させるテクニックだ。あるキーに対応するレプリカの一つが書き込みリクエストに反応しなかったときに、別のノードを選択して一時的にその書き込み処理を引き継がせる。反応しなかったノードへの書き込みは個別に続けられ、反応しなかったノードが復旧したことをバックアップノードが知った時点で、そのレプリカに新しい書き込みをすべて転送する。Dynamo の論文では ‘sloppy quorum’ という手法を利用しておらず、Hinted Handoff で書き込まれたノードも書き込みの成功判断基準である W にカウントできるようにしている。

Cassandra や Voldemort は Hinted Handoff のぶんを W にカウントせず、本来割り当てられているレプリカの中で W に満たない場合は書き込みに失敗する。それでもなお Hinted Handoff は有用だ。というのも、反応しなくなったノードが復旧したときのリカバリーを高速に行えるからである。

Anti-Entropy

あるレプリカのダウン長期間になつたり、ダウンしたレプリカを Hinted Handoff で引き継いだマシン自体もまたダウンしてしまった場合、レプリカはお互いに同期しなければならない。この場合に Cassandra や Riak が使う手順は、Dynamo に影響を受けた *Anti-Entropy* というものだ。Anti-Entropyにおいて、各レプリカはマークル木 (*Merkle Trees*) を交換する。これは、担当するキー範囲のうち最新状態と同期できていない部分を識別するものだ。マークル木とは、ハッシュの検証を階層的に行うものだ。もしキー空間全体のハッシュが二つのレプリカで一致しなければ、レプリケートしているキー空間のより小さい部分のハッシュを順に交換していく、同期できていないキーが特定できるまでそれを続ける。この手法により、ほとんど同じ状態で一部だけ違うというレプリカの間でのデータ交換の量を減らせる。

Gossip

分散システムが成長するにつれ、システム内の個々のノードが何をしているのかを追跡するのが難しくなってくる。Dynamo ベースの三つのシステムが他のノードを追跡するために使っているのは、*Gossip* という昔ながらのテクニックだ。定期的(毎秒など)に、あるノードがランダムに別のノードを選んでお互いに通信し、自分が知っている他のノードの健康状態を交換する。このようにすることで各ノードは他のノードがダウンしているかどうかを知ることができ、クライアントからのキーの検索要求をどこに振ればいいかもわかるようになる。

13.6 最後に

NoSQLを取り巻く世界はまだ成熟しておらず、今回議論したシステムの多くもそのアーキテクチャや設計そしてインターフェイスを変えていくかもしれない。本章を読んで得られるものは、個々の NoSQL システムが現時点で何をしているかということではない。これらのシステムが、どのような決断を経て現在の機能セットに至ったのかということだ。NoSQL は、設計作業の多くをアプリケーション側の設計に委ねた。これらのシステムのアーキテクチャについて理解すれば、単に次世代のすばらしい NoSQL システムを作るにとどまらず、現時点のバージョンを責任を持って使えるようにする手助けになることだろう。

13.7 謝辞

Jackie Carter や Mihir Kedia、そして匿名のレビューのみなさんに感謝する。みなさんのコメントや提案が本章の改善に大いに役立った。また本章は、NoSQL コミュニティの長年の作業がなければ存在し得なかった。これからもぜひこの調子で!

Python Packaging

Tarek Ziadé

14.1 Introduction

There are two schools of thought when it comes to installing applications. The first, common to Windows and Mac OS X, is that applications should be self-contained, and their installation should not depend on anything else. This philosophy simplifies the management of applications: each application is its own standalone “appliance”, and installing and removing them should not disturb the rest of the OS. If the application needs an uncommon library, that library is included in the application’s distribution.

The second school, which is the norm for Linux-based systems, treats software as a collection of small self-contained units called *packages*. Libraries are bundled into packages, any given library package might depend on other packages. Installing an application might involve finding and installing particular versions of dozens of other libraries. These dependencies are usually fetched from a central repository that contains thousands of packages. This philosophy is why Linux distributions use complex package management systems like dpkg and RPM to track dependencies and prevent installation of two applications that use incompatible versions of the same library.

There are pros and cons to each approach. Having a highly modular system where every piece can be updated or replaced makes management easier, because each library is present in a single place, and all applications that use it benefit when it is updated. For instance, a security fix in a particular library will reach all applications that use it at once, whereas if an application ships with its own library, that security fix will be more complex to deploy, especially if different applications use different versions of the library.

But that modularity is seen as a drawback by some developers, because they’re not in control of their applications and dependencies. It is easier for them to provide a standalone software appliance to be sure that the application environment is stable and not subject to “dependency hell” during system upgrades.

Self-contained applications also make the developer’s life easier when she needs to support several operating systems. Some projects go so far as to release portable applications that remove *any* interaction with the hosting system by working in a self-contained directory, even for log files.

Python’s packaging system was intended to make the second philosophy—multiple dependencies for each install—as developer-, admin-, packager-, and user-friendly as possible. Unfortunately it had (and has) a variety of flaws which caused or allowed all kinds of problems: unintuitive version schemes, mishandled data files, difficulty re-packaging, and more. Three years ago I and a group of other Pythoners decided to reinvent it to address these problems. We call ourselves the Fellowship of the Packaging, and this chapter describes the problems we have been trying to fix, and what our solution looks like.

Terminology

In Python a *package* is a directory containing Python files. Python files are called *modules*. That definition makes the usage of the word “package” a bit vague since it is also used by many systems to refer to a *release* of a project.

Python developers themselves are sometimes vague about this. One way to remove this ambiguity is to use the term “Python packages” when we talk about a directory containing Python modules. The term “release” is used to define one version of a project, and the term “distribution” defines a source or a binary distribution of a release as something like a tarball or zip file.

14.2 The Burden of the Python Developer

Most Python programmers want their programs to be usable in any environment. They also usually want to use a mix of standard Python libraries and system-dependent libraries. But unless you package your application separately for every existing packaging system, you are doomed to provide Python-specific releases—a Python-specific release is a release aimed to be installed within a Python installation no matter what the underlying Operating System is—and hope that:

- packagers for every target system will be able to repackage your work,
- the dependencies you have will themselves be repackaged in every target system, and
- system dependencies will be clearly described.

Sometimes, this is simply impossible. For example, Plone (a full-fledged Python-powered CMS) uses hundreds of small pure Python libraries that are not always available as packages in every packaging system out there. This means that Plone *must* ship everything that it needs in a portable application. To do this, it uses `zc.buildout`, which collects all its dependencies and creates a

portable application that will run on any system within a single directory. It is effectively a binary release, since any piece of C code will be compiled in place.

This is a big win for developers: they just have to describe their dependencies using the Python standards described below and use `zc.buildout` to release their application. But as discussed earlier, this type of release sets up a fortress within the system, which most Linux sysadmins will hate. Windows admins won't mind, but those managing CentOS or Debian will, because those systems base their management on the assumption that every file in the system is registered, classified, and known to admin tools.

Those admins will want to repackage your application according to their own standards. The question we need to answer is, "Can Python have a packaging system that can be automatically translated into other packaging systems?" If so, one application or library can be installed on any system without requiring extra packaging work. Here, "automatically" doesn't necessarily mean that the work should be fully done by a script: RPM or `dpkg` packagers will tell you that's impossible—they always need to add some specifics in the projects they repackage. They'll also tell you that they often have a hard time re-packaging a piece of code because its developers were not aware of a few basic packaging rules.

Here's one example of what you can do to annoy packagers using the existing Python packaging system: release a library called "MathUtils" with the version name "Fumanchu". The brilliant mathematician who wrote the library have found it amusing to use his cats' names for his project versions. But how can a packager know that "Fumanchu" is his second cat's name, and that the first one was called "Phil", so that the "Fumanchu" version comes after the "Phil" one?

This may sound extreme, but it can happen with today's tools and standards. The worst thing is that tools like `easy_install` or `pip` use their own non-standard registry to keep track of installed files, and will sort the "Fumanchu" and "Phil" versions alphanumerically.

Another problem is how to handle data files. For example, what if your application uses an SQLite database? If you put it inside your package directory, your application might fail because the system forbids you to write in that part of the tree. Doing this will also compromise the assumptions Linux systems make about where application data is for backups (`/var`).

In the real world, system administrators need to be able to place your files where they want without breaking your application, and you need to tell them what those files are. So let's rephrase the question: is it possible to have a packaging system in Python that can provide all the information needed to repackage an application with any third-party packaging system out there without having to read the code, and make everyone happy?

14.3 The Current Architecture of Packaging

The `Distutils` package that comes with the Python standard library is riddled with the problems described above. Since it's the standard, people either live with it and its flaws, or use more advanced

tools like `Setuptools`, which add features on the top of it, or `Distribute`, a fork of `Setuptools`. There's also `Pip`, a more advanced installer, that relies on `Setuptools`.

However, these newer tools are all based on `Distutils` and inherit its problems. Attempts were made to fix `Distutils` in place, but the code is so deeply used by other tools that any change to it, even its internals, is a potential regression in the whole Python packaging ecosystem.

We therefore decided to freeze `Distutils` and start the development of `Distutils2` from the same code base, without worrying too much about backward compatibility. To understand what changed and why, let's have a closer look at `Distutils`.

Distutils Basics and Design Flaws

`Distutils` contains commands, each of which is a class with a `run` method that can be called with some options. `Distutils` also provides a `Distribution` class that contains global values every command can look at.

To use `Distutils`, a developer adds a single Python module to a project, conventionally called `setup.py`. This module contains a call to `Distutils`' main entry point: the `setup` function. This function can take many options, which are held by a `Distribution` instance and used by commands. Here's an example that defines a few standard options like the name and version of the project, and a list of modules it contains:

```
from distutils.core import setup

setup(name='MyProject', version='1.0', py_modules=['mycode.py'])
```

This module can then be used to run Distutils commands like `sdist`, which creates a source distribution in an archive and places it in a `dist` directory:

```
$ python setup.py sdist
```

Using the same script, you can install the project using the `install` command:

```
$ python setup.py install
```

Distutils provides other commands such as:

- `upload` to upload a distribution into an online repository.
- `register` to register the metadata of a project in an online repository without necessary uploading a distribution,
- `bdist` to creates a binary distribution, and
- `bdist_msi` to create a `.msi` file for Windows.

It will also let you get information about the project via other command line options.

So installing a project or getting information about it is always done by invoking Distutils through this file. For example, to find out the name of the project:

```
$ python setup.py --name
MyProject
```

`setup.py` is therefore how everyone interacts with the project, whether to build, package, publish, or install it. The developer describes the content of his project through options passed to a function, and uses that file for all his packaging tasks. The file is also used by installers to install the project on a target system.

Having a single Python module used for packaging, releasing, *and* installing a project is one of Distutils' main flaws. For example, if you want to get the name from the `lxml` project, `setup.py` will do a lot of things besides returning a simple string as expected:

```
$ python setup.py --name
Building lxml version 2.2.
NOTE: Trying to build without Cython, pre-generated 'src/lxml/lxml.etree.c'
needs to be available.
Using build configuration of libxslt 1.1.26
Building against libxml2/libxslt in the following directory: /usr/lib/lxml
```

It might even fail to work on some projects, since developers make the assumption that `setup.py` is used only to install, and that other Distutils features are only used by them during development. The multiple roles of the `setup.py` script can easily cause confusion.

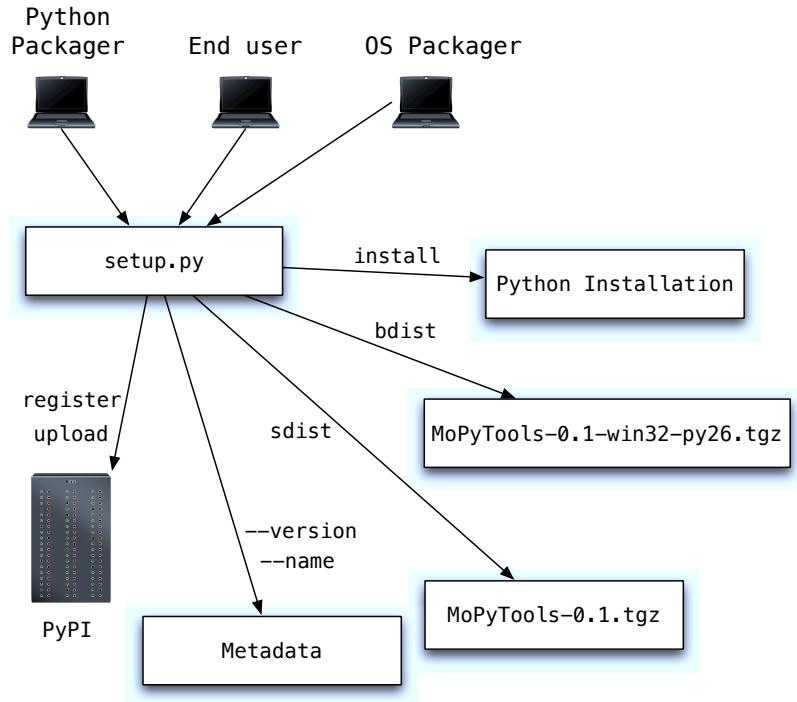


図 14.1: Setup

Metadata and PyPI

When Distutils builds a distribution, it creates a Metadata file that follows the standard described in PEP 314¹. It contains a static version of all the usual metadata, like the name of the project or the version of the release. The main metadata fields are:

- Name: The name of the project.
- Version: The version of the release.
- Summary: A one-line description.
- Description: A detailed description.
- Home-Page: The URL of the project.
- Author: The author name.
- Classifiers: Classifiers for the project. Python provides a list of classifiers for the license, the maturity of the release (beta, alpha, final), etc.
- Requires, Provides, and Obsoletes: Used to define dependencies with modules.

¹The Python Enhancement Proposals, or PEPs, that we refer to are summarized at the end of this chapter

These fields are for the most part easy to map to equivalents in other packaging systems.

The Python Package Index (PyPI)², a central repository of packages like CPAN, is able to register projects and publish releases via Distutils' `register` and `upload` commands. `register` builds the `Metadata` file and sends it to PyPI, allowing people and tools—like installers—to browse them via web pages or via web services.

The screenshot shows the PyPI repository interface. At the top, there's a navigation bar with 'Package Index > MoPyTools 0.1'. Below it, the project name 'MoPyTools 0.1' is displayed, along with a subtitle 'Set of tools to build Mozilla Services apps'. To the right, a 'Not Logged In' sidebar offers links for 'Login', 'Register', 'Lost Login?', 'Use OpenID', and 'Ip'. A table lists the package details: File (MoPyTools-0.1.tar.gz (md5)), Type (Source), Py Version (empty), Uploaded on (2011-02-04), Size (3KB), and # downloads (28). Below the table, project metadata is shown: Author (Tarek Ziade <tarek at mozilla.com>), Home Page (<http://bitbucket.org/tarek/mopytools>), Package Index Owner (tarek), and DOAP record (MoPyTools-0.1.xml). A note at the bottom encourages users to 'Log in to rate this package.'

図 14.2: The PyPI Repository

You can browse projects by `Classifiers`, and get the author name and project URL. Meanwhile, `Requires` can be used to define dependencies on Python modules. The `requires` option can be used to add a `Requires` metadata element to the project:

```
from distutils.core import setup

setup(name='foo', version='1.0', requires=['ldap'])
```

Defining a dependency on the `ldap` module is purely declarative: no tools or installers ensure that such a module exists. This would be satisfactory if Python defined requirements at the module level through a `require` keyword like Perl does. Then it would just be a matter of the installers browsing the dependencies at PyPI and installing them; that's basically what CPAN does. But that's not possible in Python since a module named `ldap` can exist in any Python project. Since Distutils allows people to release projects that can contain several packages and modules, this metadata field is not useful at all.

Another flaw of `Metadata` files is that they are created by a Python script, so they are specific to the platform they are executed in. For example, a project that provides features specific to Windows could define its `setup.py` as:

²Formerly known as the CheeseShop.

```
from distutils.core import setup

setup(name='foo', version='1.0', requires=['win32com'])
```

But this assumes that the project only works under Windows, even if it provides portable features. One way to solve this is to make the `requires` option specific to Windows:

```
from distutils.core import setup
import sys

if sys.platform == 'win32':
    setup(name='foo', version='1.0', requires=['win32com'])
else:
    setup(name='foo', version='1.0')
```

This actually makes the issue worse. Remember, the script is used to build source archives that are then released to the world via PyPI. This means that the static Metadata file sent to PyPI is dependent on the platform that was used to compile it. In other words, there is no way to indicate statically in the metadata field that it is platform-specific.

Architecture of PyPI

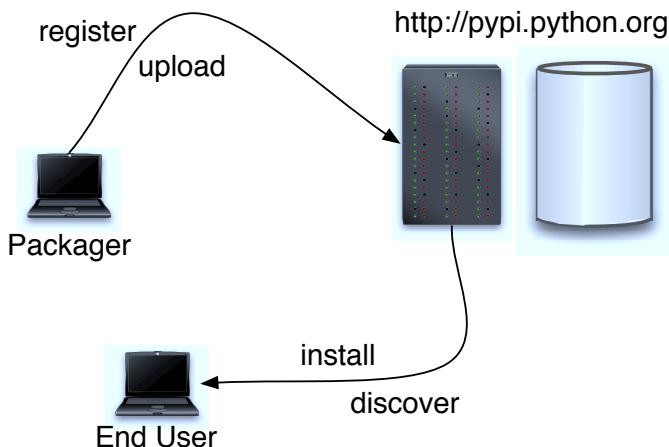


図 14.3: PyPI Workflow

As indicated earlier, PyPI is a central index of Python projects where people can browse existing projects by category or register their own work. Source or binary distributions can be uploaded and added to an existing project, and then downloaded for installation or study. PyPI also offers web services that can be used by tools like installers.

Registering Projects and Uploading Distributions

Registering a project to PyPI is done with the `Distutils register` command. It builds a POST request containing the metadata of the project, whatever its version is. The request requires an Authorization header, as PyPI uses Basic Authentication to make sure every registered project is associated with a user that has first registered with PyPI. Credentials are kept in the local `Distutils` configuration or typed in the prompt every time a `register` command is invoked. An example of its use is:

```
$ python setup.py register
running register
Registering MPTools to http://pypi.python.org/pypi
Server response (200): OK
```

Each registered project gets a web page with an HTML version of the metadata, and packagers can upload distributions to PyPI using `upload`:

```
$ python setup.py sdist upload
running sdist
...
running upload
Submitting dist/mopytools-0.1.tar.gz to http://pypi.python.org/pypi
Server response (200): OK
```

It's also possible to point users to another location via the `Download-URL` metadata field rather than uploading files directly to PyPI.

Querying PyPI

Besides the HTML pages PyPI publishes for web users, it provides two services that tools can use to browse the content: the Simple Index protocol and the XML-RPC APIs.

The Simple Index protocol starts at `http://pypi.python.org/simple/`, a plain HTML page that contains relative links to every registered project:

```
<html><head><title>Simple Index</title></head><body>
...
<a href='MontyLingua/'>MontyLingua</a><br/>
<a href='mootiro_web/'>mootiro_web</a><br/>
<a href='Mopidy/'>Mopidy</a><br/>
<a href='mopowg/'>mopowg</a><br/>
<a href='MOPPY/'>MOPPY</a><br/>
<a href='MPTools/'>MPTools</a><br/>
<a href='morbid/'>morbid</a><br/>
<a href='Morelia/'>Morelia</a><br/>
<a href='morse/'>morse</a><br/>
...
</body></html>
```

For example, the MPTools project has a `MPTools/` link, which means that the project exists in the index. The site it points at contains a list of all the links related to the project:

- links for every distribution stored at PyPI
- links for every Home URL defined in the Metadata, for each version of the project registered
- links for every Download-URL defined in the Metadata, for each version as well.

The page for MPTools contains:

```
<html><head><title>Links for MPTools</title></head>
<body><h1>Links for MPTools</h1>
<a href=".../../packages/source/M/MPTools/MPTools-0.1.tar.gz">MPTools-0.1.tar.gz</a><br/>
<a href="http://bitbucket.org/tarek/mopytools" rel="homepage">0.1 home_page</a><br/>
</body></html>
```

Tools like installers that want to find distributions of a project can look for it in the index page, or simply check if `http://pypi.python.org/simple/PROJECT_NAME/` exists.

This protocol has two main limitations. First, PyPI is a single server right now, and while people usually have local copies of its content, we have experienced several downtimes in the past two years that have paralyzed developers that are constantly working with installers that browse PyPI to get all the dependencies a project requires when it is built. For instance, building a Plone application will generate several hundreds queries at PyPI to get all the required bits, so PyPI may act as a single point of failure.

Second, when the distributions are not stored at PyPI and a Download-URL link is provided in the Simple Index page, installers have to follow that link and hope that the location will be up and will really contain the release. These indirections weakens any Simple Index-based process.

The Simple Index protocol's goal is to give to installers a list of links they can use to install a project. The project metadata is not published there; instead, there are XML-RPC methods to get extra information about registered projects:

```
>>> import xmlrpclib
>>> import pprint
>>> client = xmlrpclib.ServerProxy('http://pypi.python.org/pypi')
>>> client.package_releases('MPTools')
['0.1']
>>> pprint.pprint(client.release_urls('MPTools', '0.1'))
[{'comment_text': '',
 'downloads': 28,
 'filename': 'MPTools-0.1.tar.gz',
 'has_sig': False,
 'md5_digest': '6b06752d62c4bffe1fb65cd5c9b7111a',
 'packagetype': 'sdist',
 'python_version': 'source',
 'size': 3684,
 'upload_time': <DateTime '20110204T09:37:12' at f4da28>,
```

```

'url': 'http://pypi.python.org/packages/source/M/MPTools/MPTools-0.1.tar.gz']}
>>> pprint.pprint(client.release_data('MPTools', '0.1'))
{'author': 'Tarek Ziade',
'author_email': 'tarek@mozilla.com',
'classifiers': [],
'description': 'UNKNOWN',
'download_url': 'UNKNOWN',
'home_page': 'http://bitbucket.org/tarek/mopytools',
'keywords': None,
'license': 'UNKNOWN',
'maintainer': None,
'maintainer_email': None,
'name': 'MPTools',
'package_url': 'http://pypi.python.org/pypi/MPTools',
'platform': 'UNKNOWN',
'release_url': 'http://pypi.python.org/pypi/MPTools/0.1',
'requires_python': None,
'stable_version': None,
'summary': 'Set of tools to build Mozilla Services apps',
'version': '0.1'}

```

The issue with this approach is that some of the data that the XML-RPC APIs are publishing could have been stored as static files and published in the Simple Index page to simplify the work of client tools. That would also avoid the extra work PyPI has to do to handle those queries. It's fine to have non-static data like the number of downloads per distribution published in a specialized web service, but it does not make sense to have to use two different services to get all static data about a project.

Architecture of a Python Installation

If you install a Python project using `python setup.py install`, `Distutils`—which is included in the standard library—will copy the files onto your system.

- *Python packages* and modules will land in the Python directory that is loaded when the interpreter starts: under the latest Ubuntu they will wind up in `/usr/local/lib/python2.6/dist-packages/` and under Fedora in `/usr/local/lib/python2.6/sites-packages/`.
- *Data files* defined in a project can land anywhere on the system.
- The *executable script* will land in a `bin` directory on the system. Depending on the platform, this could be `/usr/local/bin` or in a `bin` directory specific to the Python installation.

Ever since Python 2.5, the metadata file is copied alongside the modules and packages as `project-version.egg-info`. For example, the `virtualenv` project could have a `virtualenv-1.4.9.egg-info` file. These metadata files can be considered a database of installed projects, since it's possible to iterate over them and build a list of projects with their versions. However, the `Distutils` installer does not record the list of files it installs on the system. In other words, there is no way to remove all

files that were copied in the system. This is a shame since the `install` command has a `--record` option that can be used to record all installed files in a text file. However, this option is not used by default and `Distutils`' documentation barely mentions it.

SetupTools, Pip and the Like

As mentioned in the introduction, some projects tried to fix some of the problems with `Distutils`, with varying degrees of success.

The Dependencies Issue

`PyPI` allowed developers to publish Python projects that could include several modules organized into Python packages. But at the same time, projects could define module-level dependencies via `Require`. Both ideas are reasonable, but their combination is not.

The right thing to do was to have project-level dependencies, which is exactly what `SetupTools` added as a feature on the top of `Distutils`. It also provided a script called `easy_install` to automatically fetch and install dependencies by looking for them on `PyPI`. In practice, module-level dependency was never really used, and people jumped on `SetupTools`' extensions. But since these features were added in options specific to `SetupTools`, and ignored by `Distutils` or `PyPI`, `SetupTools` effectively created its own standard and became a hack on a top of a bad design.

`easy_install` therefore needs to download the archive of the project and run its `setup.py` script again to get the metadata it needs, and it has to do this again for every dependency. The dependency graph is built bit by bit after each download.

Even if the new metadata was accepted by `PyPI` and browsable online, `easy_install` would still need to download all archives because, as said earlier, metadata published at `PyPI` is specific to the platform that was used to upload it, which can differ from the target platform. But this ability to install a project and its dependencies was good enough in 90% of the cases and was a great feature to have. So `SetupTools` became widely used, although it still suffers from other problems:

- If a dependency install fails, there is no rollback and the system can end up in a broken state.
- The dependency graph is built on the fly during installation, so if a dependency conflict is encountered the system can end up in a broken state as well.

The Uninstall Issue

`SetupTools` did not provide an uninstaller, even though its custom metadata could have contained a file listing the installed files. `Pip`, on the other hand, extended `SetupTools`' metadata to record installed files, and is therefore able to uninstall. But that's yet another custom set of metadata, which

means that a single Python installation may contain up to four different flavours of metadata for each installed project:

- Distutils' egg-info, which is a single metadata file.
- Setuptools' egg-info, which is a directory containing the metadata and extra Setuptools specific options.
- Pip's egg-info, which is an extended version of the previous.
- Whatever the hosting packaging system creates.

What About Data Files?

In Distutils, data files can be installed anywhere on the system. If you define some package data files in `setup.py` script like this:

```
setup(...,  
    packages=['mypkg'],  
    package_dir={'mypkg': 'src/mypkg'},  
    package_data={'mypkg': ['data/*.dat']},  
)
```

then all files with the `.dat` extension in the `mypkg` project will be included in the distribution and eventually installed along with the Python modules in the Python installation.

For data files that need to be installed outside the Python distribution, there's another option that stores files in the archive but puts them in defined locations:

```
setup(...,  
    data_files=[('bitmaps', ['bm/b1.gif', 'bm/b2.gif']),  
               ('config', ['cfg/data.cfg']),  
               ('/etc/init.d', ['init-script'])]  
)
```

This is terrible news for OS packagers for several reasons:

- Data files are not part of the metadata, so packagers need to read `setup.py` and sometimes dive into the project's code.
- The developer should not be the one deciding where data files should land on a target system.
- There are no categories for these data files: images, man pages, and everything else are all treated the same way.

A packager who needs to repackage a project with such a file has no choice but to patch the `setup.py` file so that it works as expected for her platform. To do that, she must review the code and change every line that uses those files, since the developer made an assumption about their location. Setuptools and Pip did not improve this.

14.4 Improved Standards

So we ended up with with a mixed up and confused packaging environment, where everything is driven by a single Python module, with incomplete metadata and no way to describe everything a project contains. Here's what we're doing to make things better.

Metadata

The first step is to fix our Metadata standard. PEP 345 defines a new version that includes:

- a saner way to define versions
- project-level dependencies
- a static way to define platform-specific values

Version

One goal of the metadata standard is to make sure that all tools that operate on Python projects are able to classify them the same way. For versions, it means that every tool should be able to know that “1.1” comes after “1.0”. But if project have custom versioning schemes, this becomes much harder.

The only way to ensure consistent versioning is to publish a standard that projects will have to follow. The scheme we chose is a classical sequence-based scheme. As defined in PEP 386, its format is:

`N.N[.N]+[{a|b|c|rc}N[.N]+][.postN][.devN]`

where:

- N is an integer. You can use as many N s as you want and separate them by dots, as long as there are at least two (MAJOR.MINOR).
- a, b, c and rc are *alpha*, *beta* and *release candidate* markers. They are followed by an integer. Release candidates have two markers because we wanted the scheme to be compatible with Python, which uses rc . But we find c simpler.
- *dev* followed by a number is a dev marker.
- *post* followed by a number is a post-release marker.

Depending on the project release process, dev or post markers can be used for all intermediate versions between two final releases. Most process use dev markers.

Following this scheme, PEP 386 defines a strict ordering:

- alpha < beta < rc < final
- dev < non-dev < post, where non-dev can be alpha, beta, rc or final

Here's a full ordering example:

```
1.0a1 < 1.0a2.dev456 < 1.0a2 < 1.0a2.1.dev456  
< 1.0a2.1 < 1.0b1.dev456 < 1.0b2 < 1.0b2.post345  
< 1.0c1.dev456 < 1.0c1 < 1.0.dev456 < 1.0  
< 1.0.post456.dev34 < 1.0.post456
```

The goal of this scheme is to make it easy for other packaging systems to translate Python projects' versions into their own schemes. PyPI now rejects any projects that upload PEP 345 metadata with version numbers that don't follow PEP 386.

Dependencies

PEP 345 defines three new fields that replace PEP 314 Requires, Provides, and Obsoletes. Those fields are Requires-Dist, Provides-Dist, and Obsoletes-Dist, and can be used multiple times in the metadata.

For Requires-Dist, each entry contains a string naming some other Distutils project required by this distribution. The format of a requirement string is identical to that of a Distutils project name (e.g., as found in the Name field) optionally followed by a version declaration within parentheses. These Distutils project names should correspond to names as found at PyPI, and version declarations must follow the rules described in PEP 386. Some example are:

```
Requires-Dist: pkginfo  
Requires-Dist: PasteDeploy  
Requires-Dist: zope.interface (>3.5.0)
```

Provides-Dist is used to define extra names contained in the project. It's useful when a project wants to merge with another project. For example the ZODB project can include the transaction project and state:

```
Provides-Dist: transaction
```

Obsoletes-Dist is useful to mark another project as an obsolete version:

```
Obsoletes-Dist: OldName
```

Environment Markers

An environment marker is a marker that can be added at the end of a field after a semicolon to add a condition about the execution environment. Some examples are:

```
Requires-Dist: pywin32 (>1.0); sys.platform == 'win32'  
Obsoletes-Dist: pywin31; sys.platform == 'win32'  
Requires-Dist: foo (1,!<1.3); platform.machine == 'i386'  
Requires-Dist: bar; python_version == '2.4' or python_version == '2.5'  
Requires-External: libxslt; 'linux' in sys.platform
```

The micro-language for environment markers is deliberately kept simple enough for non-Python programmers to understand: it compares strings with the == and in operators (and their opposites), and allows the usual Boolean combinations. The fields in PEP 345 that can use this marker are:

- Requires-Python
- Requires-External
- Requires-Dist
- Provides-Dist
- Obsoletes-Dist
- Classifier

What's Installed?

Having a single installation format shared among all Python tools is mandatory for interoperability. If we want Installer A to detect that Installer B has previously installed project Foo, they both need to share and update the same database of installed projects.

Of course, users should ideally use a single installer in their system, but they may want to switch to a newer installer that has specific features. For instance, Mac OS X ships `Setuptools`, so users automatically have the `easy_install` script. If they want to switch to a newer tool, they will need it to be backward compatible with the previous one.

Another problem when using a Python installer on a platform that has a packaging system like RPM is that there is no way to inform the system that a project is being installed. What's worse, even if the Python installer could somehow ping the central packaging system, we would need to have a mapping between the Python metadata and the system metadata. The name of the project, for instance, may be different for each. That can occur for several reasons. The most common one is a conflict name: another project outside the Python land already uses the same name for the RPM. Another cause is that the name used include a python prefix that breaks the convention of the platform. For example, if you name your project `foo-python`, there are high chances that the Fedora RPM will be called `python-foo`.

One way to avoid this problem is to leave the global Python installation alone, managed by the central packaging system, and work in an isolated environment. Tools like `Virtualenv` allows this.

In any case, we do need to have a single installation format in Python because interoperability is also a concern for other packaging systems when they install themselves Python projects. Once a third-party packaging system has registered a newly installed project in its own database on the system, it needs to generate the right metadata for the Python installaton itself, so projects appear to be installed to Python installers or any APIs that query the Python installation.

The metadata mapping issue can be addressed in that case: since an RPM knows which Python projects it wraps, it can generate the proper Python-level metadata. For instance, it knows that `python26-webob` is called `WebOb` in the PyPI ecosystem.

Back to our standard: PEP 376 defines a standard for installed packages whose format is quite similar to those used by `Setuptools` and `Pip`. This structure is a directory with a `dist-info` extension that contains:

- **METADATA**: the metadata, as described in PEP 345, PEP 314 and PEP 241.
- **RECORD**: the list of installed files in a csv-like format.
- **INSTALLER**: the name of the tool used to install the project.
- **REQUESTED**: the presence of this file indicates that the project installation was explicitly requested (i.e., not installed as a dependency).

Once all tools out there understand this format, we'll be able to manage projects in Python without depending on a particular installer and its features. Also, since PEP 376 defines the metadata as a directory, it will be easy to add new files to extend it. As a matter of fact, a new metadata file called `RESOURCES`, described in the next section, might be added in a near future without modifying PEP 376. Eventually, if this new file turns out to be useful for all tools, it will be added to the PEP.

Architecture of Data Files

As described earlier, we need to let the packager decide where to put data files during installation without breaking the developer's code. At the same time, the developer must be able to work with data files without having to worry about their location. Our solution is the usual one: indirection.

Using Data Files

Suppose your `MPTools` application needs to work with a configuration file. The developer will put that file in a Python package and use `__file__` to reach it:

```
import os

here = os.path.dirname(__file__)
cfg = open(os.path.join(here, 'config', 'mopy.cfg'))
```

This implies that configuration files are installed like code, and that the developer *must* place it alongside her code: in this example, in a subdirectory called config.

The new architecture of data files we have designed uses the project tree as the root of all files, and allows access to any file in the tree, whether it is located in a Python package or a simple directory. This allowed developers to create a dedicated directory for data files and access them using `pkgutil.open`:

```
import os
import pkgutil

# Open the file located in config/mopy.cfg in the MPTools project
cfg = pkgutil.open('MPTools', 'config/mopy.cfg')
```

`pkgutil.open` looks for the project metadata and see if it contains a RESOURCES file. This is a simple map of files to locations that the system may contain:

```
config/mopy.cfg {confdir}/{distribution.name}
```

Here the `confdir` variable points to the system's configuration directory, and `distribution.name` contains the name of the Python project as found in the metadata.

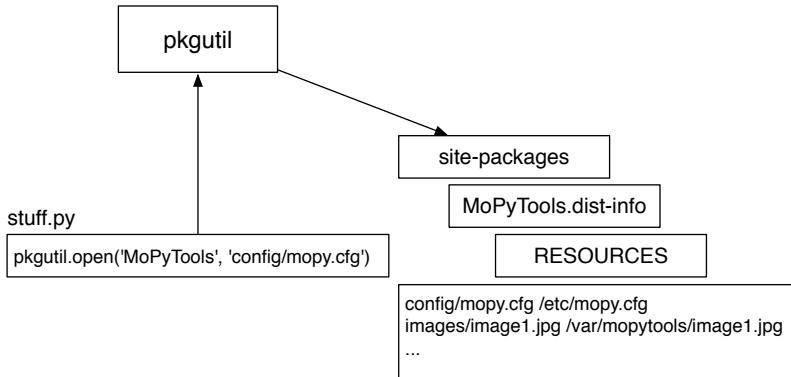


図 14.4: Finding a File

As long as this RESOURCES metadata file is created at installation time, the API will find the location of `mopy.cfg` for the developer. And since `config/mopy.cfg` is the path relative to the project tree, it means that we can also offer a development mode where the metadata for the project are generated in-place and added in the lookup paths for `pkgutil`.

Declaring Data Files

In practice, a project can define where data files should land by defining a mapper in their `setup.cfg` file. A mapper is a list of (`glob-style pattern`, `target`) tuples. Each pattern points to one of

several files in the project tree, while the target is an installation path that may contain variables in brackets. For example, MPTools's `setup.cfg` could look like this:

```
[files]
resources =
    config/mopy.cfg {confdir}/{application.name}/
    images/*.jpg     {datadir}/{application.name}/
```

The `sysconfig` module will provide and document a specific list of variables that can be used, and default values for each platform. For example `confdir` is `/etc` on Linux. Installers can therefore use this mapper in conjunction with `sysconfig` at installation time to know where the files should be placed. Eventually, they will generate the `RESOURCES` file mentioned earlier in the installed metadata so `pkgutil` can find back the files.

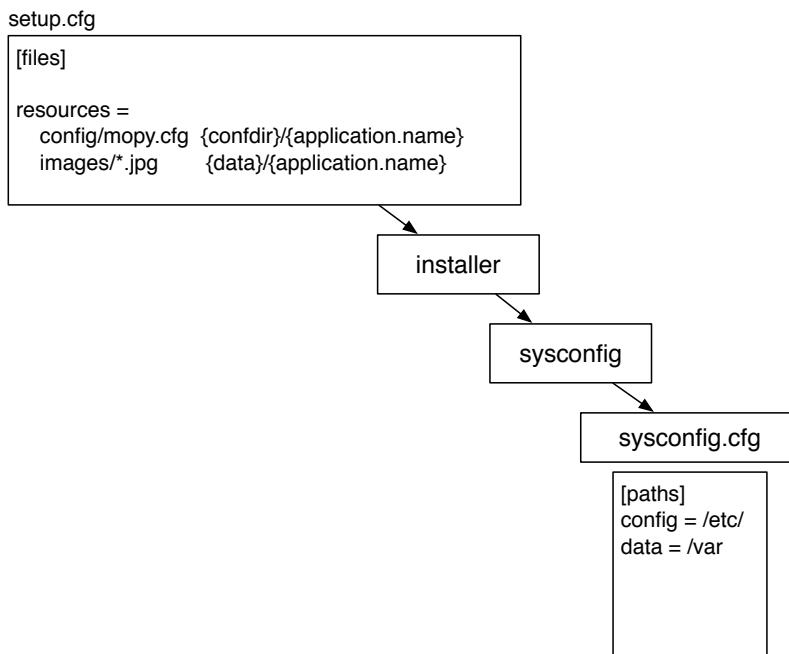


図 14.5: Installer

PyPI Improvements

I said earlier that PyPI was effectively a single point of failure. PEP 380 addresses this problem by defining a mirroring protocol so that users can fall back to alternative servers when PyPI is down. The goal is to allow members of the community to run mirrors around the world.

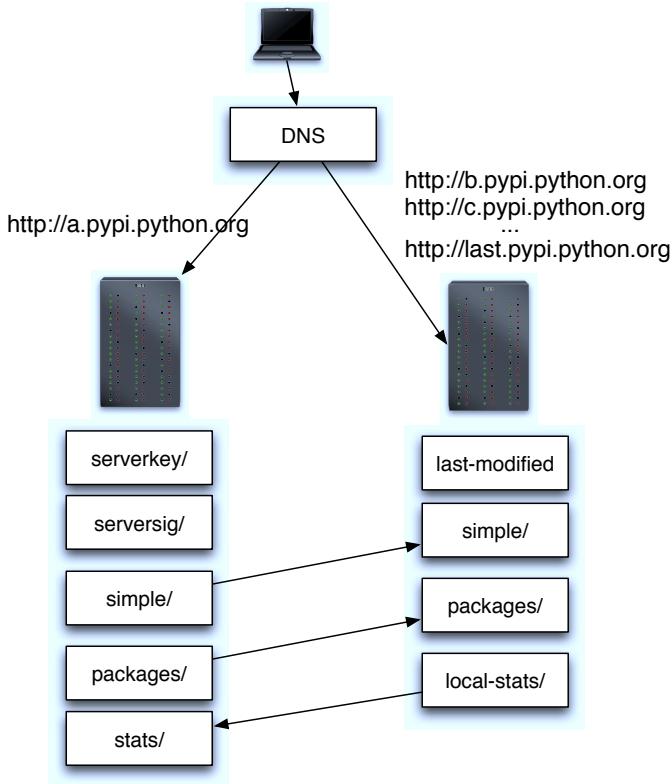


図 14.6: Mirroring

The mirror list is provided as a list of host names of the form `X.pypi.python.org`, where `X` is in the sequence `a, b, c, ..., aa, ab, ...`. `a.pypi.python.org` is the master server and mirrors start with `b`. A CNAME record `last.pypi.python.org` points to the last host name so clients that are using PyPI can get the list of the mirrors by looking at the CNAME.

For example, this call tells use that the last mirror is `h.pypi.python.org`, meaning that PyPI currently has 6 mirrors (`b` through `h`):

```
>>> import socket
>>> socket.gethostname_ex('last.pypi.python.org')[0]
'h.pypi.python.org'
```

Potentially, this protocol allows clients to redirect requests to the nearest mirror by localizing the mirrors by their IPs, and also fall back to the next mirror if a mirror or the master server is down. The mirroring protocol itself is more complex than a simple rsync because we wanted to keep downloads statistics accurate and provide minimal security.

Synchronization

Mirrors must reduce the amount of data transferred between the central server and the mirror. To achieve that, they *must* use the changelog PyPI XML-RPC call, and only refetch the packages that have been changed since the last time. For each package P, they *must* copy documents /simple/P/ and /serversig/P.

If a package is deleted on the central server, they *must* delete the package and all associated files. To detect modification of package files, they may cache the file's ETag, and may request skipping it using the If-None-Match header. Once the synchronization is over, the mirror changes its /last-modified to the current date.

Statistics Propagation

When you download a release from any of the mirrors, the protocol ensures that the download hit is transmitted to the master PyPI server, then to other mirrors. Doing this ensures that people or tools browsing PyPI to find out how many times a release was downloaded will get a value summed across all mirrors.

Statistics are grouped into daily and weekly CSV files in the stats directory at the central PyPI itself. Each mirror needs to provide a local-stats directory that contains its own statistics. Each file provides the number of downloads for each archive, grouped by use agents. The central server visits mirrors daily to collect those statistics, and merge them back into the global stats directory, so each mirror must keep /local-stats up-to-date at least once a day.

Mirror Authenticity

With any distributed mirroring system, clients may want to verify that the mirrored copies are authentic. Some of the possible threats include:

- the central index may be compromised
- the mirrors might be tampered with
- a man-in-the-middle attack between the central index and the end user, or between a mirror and the end user

To detect the first attack, package authors need to sign their packages using PGP keys, so that users can verify that the package comes from the author they trust. The mirroring protocol itself only addresses the second threat, though some attempt is made to detect man-in-the-middle attacks.

The central index provides a DSA key at the URL /serverkey, in the PEM format as generated by openssl dsa -pubout³. This URL must not be mirrored, and clients must fetch the official

³I.e., RFC 3280 SubjectPublicKeyInfo, with the algorithm 1.3.14.3.2.12.

serverkey from PyPI directly, or use the copy that came with the PyPI client software. Mirrors should still download the key so that they can detect a key rollover.

For each package, a mirrored signature is provided at `/serversig/package`. This is the DSA signature of the parallel URL `/simple/package`, in DER form, using SHA-1 with DSA⁴.

Clients using a mirror need to perform the following steps to verify a package:

1. Download the `/simple` page, and compute its SHA-1 hash.
2. Compute the DSA signature of that hash.
3. Download the corresponding `/serversig`, and compare it byte for byte with the value computed in step 2.
4. Compute and verify (against the `/simple` page) the MD5 hashes of all files they download from the mirror.

Verification is not needed when downloading from central index, and clients should not do it to reduce the computation overhead.

About once a year, the key will be replaced with a new one. Mirrors will have to re-fetch all `/serversig` pages. Clients using mirrors need to find a trusted copy of the new server key. One way to obtain one is to download it from <https://pypi.python.org/serverkey>. To detect man-in-the-middle attacks, clients need to verify the SSL server certificate, which will be signed by the CACert authority.

14.5 Implementation Details

The implementation of most of the improvements described in the previous section are taking place in `Distutils2`. The `setup.py` file is not used anymore, and a project is completely described in `setup.cfg`, a static .ini-like file. By doing this, we make it easier for packagers to change the behavior of a project installation without having to deal with Python code. Here's an example of such a file:

```
[metadata]
name = MPTools
version = 0.1
author = Tarek Ziade
author-email = tarek@mozilla.com
summary = Set of tools to build Mozilla Services apps
description-file = README
home-page = http://bitbucket.org/tarek/pypi2rpm
project-url: Repository, http://hg.mozilla.org/services/server-devtools
classifier = Development Status :: 3 - Alpha
License :: OSI Approved :: Mozilla Public License 1.1 (MPL 1.1)
```

⁴I.e., as a RFC 3279 Dsa-Sig-Value, created by algorithm 1.2.840.10040.4.3.

```

[files]
packages =
    mopytools
    mopytools.tests

extra_files =
    setup.py
    README
    build.py
    _build.py

resources =
    etc/mopytools.cfg {confdir}/mopytools

```

`Distutils2` use this configuration file to:

- generate META-1.2 metadata files that can be used for various actions, like registering at PyPI.
- run any package management command, like `sdist`.
- install a `Distutils2`-based project.

`Distutils2` also implements `VERSION` via its `version` module.

The `INSTALL-DB` implementation will find its way to the standard library in Python 3.3 and will be in the `pkgutil` module. In the interim, a version of this module exists in `Distutils2` for immediate use. The provided APIs will let us browse an installation and know exactly what's installed.

These APIs are the basis for some neat `Distutils2` features:

- installer/uninstaller
- dependency graph view of installed projects

14.6 Lessons learned

It's All About PEPs

Changing an architecture as wide and complex as Python packaging needs to be carefully done by changing standards through a PEP process. And changing or adding a new PEP takes in my experience around a year.

One mistake the community made along the way was to deliver tools that solved some issues by extending the Metadata and the way Python applications were installed without trying to change the impacted PEPs.

In other words, depending on the tool you used, the standard library `Distutils` or `Setuptools`, applications were installed differently. The problems were solved for one part of the community that used these new tools, but added more problems for the rest of the world. OS Packagers for instance,

had to face several Python standards: the official documented standard and the de-facto standard imposed by `Setuptools`.

But in the meantime, `Setuptools` had the opportunity to experiment in a realistic scale (the whole community) some innovations in a very fast pace, and the feedback was invaluable. We were able to write down new PEPs with more confidence in what worked and what did not, and maybe it would have been impossible to do so differently. So it's all about detecting when some third-party tools are contributing innovations that are solving problems and that should ignite a PEP change.

A Package that Enters the Standard Library Has One Foot in the Grave

I am paraphrasing Guido van Rossum in the section title, but that's one aspect of the batteries-included philosophy of Python that impacts a lot our efforts.

`Distutils` is part of the standard library and `Distutils2` will soon be. A package that's in the standard library is very hard to make evolve. There are of course deprecation processes, where you can kill or change an API after 2 minor versions of Python. But once an API is published, it's going to stay there for years.

So any change you make in a package in the standard library that is not a bug fix, is a potential disturbance for the eco-system. So when you're doing important changes, you have to create a new package.

I've learned it the hard way with `Distutils` since I had to eventually revert all the changes I had done in it for more than a year and create `Distutils2`. In the future, if our standards change again in a drastic way, there are high chances that we will start a standalone `Distutils3` project first, unless the standard library is released on its own at some point.

Backward Compatibility

Changing the way packaging works in Python is a very long process: the Python ecosystem contains so many projects based on older packaging tools that there is and will be a lot of resistance to change. (Reaching consensus on some of the topics discussed in this chapter took several years, rather than the few months I originally expected.) As with Python 3, it will take years before all projects switch to the new standard.

That's why everything we are doing has to be backward-compatible with all previous tools, installations and standards, which makes the implementation of `Distutils2` a wicked problem.

For example, if a project that uses the new standards depends on another project that don't use them yet, we can't stop the installation process by telling the end-user that the dependency is in an unknown format!

For example, the `INSTALL-DB` implementation contains compatibility code to browse projects installed by the original `Distutils`, `Pip`, `Distribute`, or `Setuptools`. `Distutils2` is also able to install projects created by the original `Distutils` by converting their metadata on the fly.

14.7 References and Contributions

Some sections in this paper were directly taken from the various PEP documents we wrote for packaging. You can find the original documents at <http://python.org>:

- PEP 241: Metadata for Python Software Packages 1.0: <http://python.org/peps/pep-0214.html>
- PEP 314: Metadata for Python Software Packages 1.1: <http://python.org/peps/pep-0314.html>
- PEP 345: Metadata for Python Software Packages 1.2: <http://python.org/peps/pep-0345.html>
- PEP 376: Database of Installed Python Distributions: <http://python.org/peps/pep-0376.html>
- PEP 381: Mirroring infrastructure for PyPI: <http://python.org/peps/pep-0381.html>
- PEP 386: Changing the version comparison module in Distutils: <http://python.org/peps/pep-0386.html>

I would like to thank all the people that are working on packaging; you will find their name in every PEP I've mentioned. I would also like to give a special thank to all members of The Fellowship of the Packaging. Also, thanks to Alexis Mitaireau, Toshio Kuratomi, Holger Krekel and Stefane Fermigier for their feedback on this chapter.

The projects that were discussed in this chapter are:

- `Distutils`: <http://docs.python.org/distutils>
- `Distutils2`: <http://packages.python.org/Distutils2>
- `Distribute`: <http://packages.python.org/distribute>
- `Setuptools`: <http://pypi.python.org/pypi/setuptools>
- `Pip`: <http://pypi.python.org/pypi/pip>
- `Virtualenv`: <http://pypi.python.org/pypi/virtualenv>

Riak and Erlang/OTP

Francesco Cesarini, Andy Gross, and Justin Sheehy

Riak is a distributed, fault tolerant, open source database that illustrates how to build large scale systems using Erlang/OTP. Thanks in large part to Erlang’s support for massively scalable distributed systems, Riak offers features that are uncommon in databases, such as high-availability and linear scalability of both capacity and throughput.

Erlang/OTP provides an ideal platform for developing systems like Riak because it provides inter-node communication, message queues, failure detectors, and client-server abstractions out of the box. What’s more, most frequently-used patterns in Erlang have been implemented in library modules, commonly referred to as OTP behaviors. They contain the generic code framework for concurrency and error handling, simplifying concurrent programming and protecting the developer from many common pitfalls. Behaviors are monitored by supervisors, themselves a behavior, and grouped together in supervision trees. A supervision tree is packaged in an application, creating a building block of an Erlang program.

A complete Erlang system such as Riak is a set of loosely coupled applications that interact with each other. Some of these applications have been written by the developer, some are part of the standard Erlang/OTP distribution, and some may be other open source components. They are sequentially loaded and started by a boot script generated from a list of applications and versions.

What *differs* among systems are the applications that are part of the release which is started. In the standard Erlang distribution, the boot files will start the *Kernel* and *StdLib* (Standard Library) applications. In some installations, the *SASL* (Systems Architecture Support Library) application is also started. SASL contains release and software upgrade tools together with logging capabilities. Riak is no different, other than starting the Riak specific applications as well as their runtime dependencies, which include *Kernel*, *StdLib* and *SASL*. A complete and ready-to-run build of Riak actually embeds these standard elements of the Erlang/OTP distribution and starts them all in unison when `riak start` is invoked on the command line. Riak consists of many complex applications, so this chapter should not be interpreted as a complete guide. It should be seen as an introduction

to OTP where examples from the Riak source code are used. The figures and examples have been abbreviated and shortened for demonstration purposes.

15.1 An Abridged Introduction to Erlang

Erlang is a concurrent functional programming language that compiles to byte code and runs in a virtual machine. Programs consist of functions that call each other, often resulting in side effects such as inter-process message passing, I/O and database operations. Erlang variables are single assignment, i.e., once they have been given values, they cannot be updated. The language makes extensive use of pattern matching, as shown in the factorial example below:

```
-module(factorial).
-export([fac/1]).
fac(0) -> 1;
fac(N) when N>0 ->
    Prev = fac(N-1),
    N*Prev.
```

Here, the first clause gives the factorial of zero, the second factorials of positive numbers. The body of each clause is a sequence of expressions, and the final expression in the body is the result of that clause. Calling the function with a negative number will result in a run time error, as none of the clauses match. Not handling this case is an example of non-defensive programming, a practice encouraged in Erlang.

Within the module, functions are called in the usual way; outside, the name of the module is prepended, as in `factorial:fac(3)`. It is possible to define functions with the same name but different numbers of arguments—this is called their *arity*. In the export directive in the `factorial` module the `fac` function of arity one is denoted by `fac/1`.

Erlang supports tuples (also called product types) and lists. Tuples are enclosed in curly brackets, as in `{ok, 37}`. In tuples, we access elements by position. Records are another data type; they allow us to store a fixed number of elements which are then accessed and manipulated by name. We define a record using the `-record(state, {id, msg_list=[]}).` To create an instance, we use the expression `Var = #state{id=1}`, and we examine its contents using `Var#state.id`. For a variable number of elements, we use lists defined in square brackets such as in `[23, 34]`. The notation `[X|Xs]` matches a non-empty list with head `X` and tail `Xs`. Identifiers beginning with a lower case letter denote atoms, which simply stand for themselves; the `ok` in the tuple `{ok, 37}` is an example of an atom. Atoms used in this way are often used to distinguish between different kinds of function result: as well as `ok` results, there might be results of the form `{error, "Error String"}`.

Processes in Erlang systems run concurrently in separate memory, and communicate with each other by message passing. Processes can be used for a wealth of applications, including gateways to

databases, as handlers for protocol stacks, and to manage the logging of trace messages from other processes. Although these processes handle different requests, there will be similarities in how these requests are handled.

As processes exist only within the virtual machine, a single VM can simultaneously run millions of processes, a feature Riak exploits extensively. For example, each request to the database—reads, writes, and deletes—is modeled as a separate process, an approach that would not be possible with most OS-level threading implementations.

Processes are identified by process identifiers, called PIDs, but they can also be registered under an alias; this should only be used for long-lived “static” processes. Registering a process with its alias allows other processes to send it messages without knowing its PID. Processes are created using the `spawn(Module, Function, Arguments)` built-in function (BIF). BIFs are functions integrated in the VM and used to do what is impossible or slow to execute in pure Erlang. The `spawn/3` BIF takes a Module, a Function and a list of Arguments as parameters. The call returns the PID of the newly spawned process and as a side effect, creates a new process that starts executing the function in the module with the arguments mentioned earlier.

A message `Msg` is sent to a process with process id `Pid` using `Pid ! Msg`. A process can find out its PID by calling the BIF `self`, and this can then be sent to other processes for them to use to communicate with the original process. Suppose that a process expects to receive messages of the form `{ok, N}` and `{error, Reason}`. To process these it uses a receive statement:

```
receive
    {ok, N} ->
        N+1;
    {error, _} ->
        0
end
```

The result of this is a number determined by the pattern-matched clause. When the value of a variable is not needed in the pattern match, the underscore wild-card can be used as shown above.

Message passing between processes is asynchronous, and the messages received by a process are placed in the process’s mailbox in the order in which they arrive. Suppose that now the receive expression above is to be executed: if the first element in the mailbox is either `{ok, N}` or `{error, Reason}` the corresponding result will be returned. If the first message in the mailbox is not of this form, it is retained in the mailbox and the second is processed in a similar way. If no message matches, the receive will wait for a matching message to be received.

Processes terminate for two reasons. If there is no more code to execute, they are said to terminate with reason *normal*. If a process encounters a run-time error, it is said to terminate with a *non-normal* reason. A process terminating will not affect other processes unless they are linked to it. Processes can link to each other through the `link(Pid)` BIF or when calling the `spawn_link(Module, Function, Arguments)`. If a process terminates, it sends an EXIT signal to processes in its link

set. If the termination reason is non-normal, the process terminates itself, propagating the EXIT signal further. By calling the `process_flag(trap_exit, true)` BIF, processes can receive the EXIT signals as Erlang messages in their mailbox instead of terminating.

Riak uses EXIT signals to monitor the well-being of helper processes performing non-critical work initiated by the request-driving finite state machines. When these helper processes terminate abnormally, the EXIT signal allows the parent to either ignore the error or restart the process.

15.2 Process Skeletons

We previously introduced the notion that processes follow a common pattern regardless of the particular purpose for which the process was created. To start off, a process has to be spawned and then, optionally, have its alias registered. The first action of the newly spawned process is to initialize the process loop data. The loop data is often the result of arguments passed to the spawn built-in function at the initialization of the process. Its loop data is stored in a variable we refer to as the process state. The state, often stored in a record, is passed to a receive-evaluate function, running a loop which receives a message, handles it, updates the state, and passes it back as an argument to a tail-recursive call. If one of the messages it handles is a ‘stop’ message, the receiving process will clean up after itself and then terminate.

This is a recurring theme among processes that will occur regardless of the task the process has been assigned to perform. With this in mind, let’s look at the differences between the processes that conform to this pattern:

- The arguments passed to the spawn BIF calls will differ from one process to another.
- You have to decide whether you should register a process under an alias, and if you do, what alias should be used.
- In the function that initializes the process state, the actions taken will differ based on the tasks the process will perform.
- The state of the system is represented by the loop data in every case, but the contents of the loop data will vary among processes.
- When in the body of the receive-evaluate loop, processes will receive different messages and handle them in different ways.
- Finally, on termination, the cleanup will vary from process to process.

So, even if a skeleton of generic actions exists, these actions are complemented by specific ones that are directly related to the tasks assigned to the process. Using this skeleton as a template, programmers can create Erlang processes that act as servers, finite state machines, event handlers and supervisors. But instead of re-implementing these patterns every time, they have been placed in library modules referred to as behaviors. They come as part as the OTP middleware.

15.3 OTP Behaviors

The core team of developers committing to Riak is spread across nearly a dozen geographical locations. Without very tight coordination and templates to work from, the result would consist of different client/server implementations not handling special borderline cases and concurrency-related errors. There would probably be no uniform way to handle client and server crashes or guaranteeing that a response from a request is indeed the response, and not just any message that conforms to the internal message protocol.

OTP is a set of Erlang libraries and design principles providing ready-made tools with which to develop robust systems. Many of these patterns and libraries are provided in the form of “behaviors.”

OTP behaviors address these issues by providing library modules that implement the most common concurrent design patterns. Behind the scenes, without the programmer having to be aware of it, the library modules ensure that errors and special cases are handled in a consistent way. As a result, OTP behaviors provide a set of standardized building blocks used in designing and building industrial-grade systems.

Introduction

OTP behaviors are provided as library modules in the `stdlib` application which comes as part of the Erlang/OTP distribution. The specific code, written by the programmer, is placed in a separate module and called through a set of predefined callback functions standardized for each behavior. This callback module will contain all of the specific code required to deliver the desired functionality.

OTP behaviors include worker processes, which do the actual processing, and supervisors, whose task is to monitor workers and other supervisors. Worker behaviors, often denoted in diagrams as circles, include servers, event handlers, and finite state machines. Supervisors, denoted in illustrations as squares, monitor their children, both workers and other supervisors, creating what is called a supervision tree.

Supervision trees are packaged into a behavior called an application. OTP applications are not only the building blocks of Erlang systems, but are also a way to package reusable components. Industrial-grade systems like Riak consist of a set of loosely coupled, possibly distributed applications. Some of these applications are part of the standard Erlang distribution and some are the pieces that make up the specific functionality of Riak.

Examples of OTP applications include the Corba ORB or the Simple Network Management Protocol (SNMP) agent. An OTP application is a reusable component that packages library modules together with supervisor and worker processes. From now on, when we refer to an application, we will mean an OTP application.

The behavior modules contain all of the generic code for each given behavior type. Although it is possible to implement your own behavior module, doing so is rare because the ones that come with

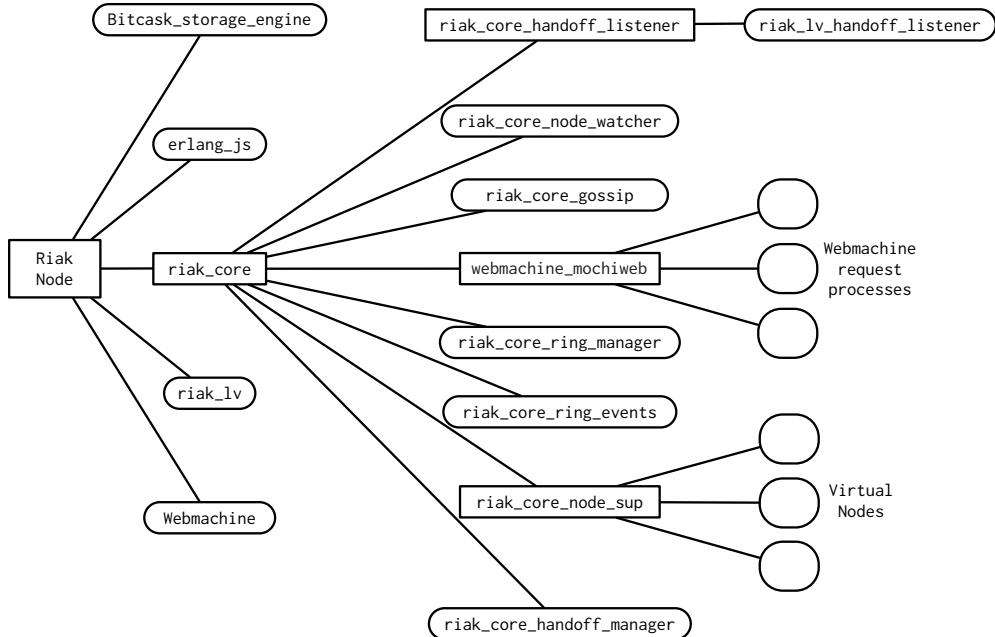


図 15.1: OTP Riak Supervision Tree

the Erlang/OTP distribution will cater to most of the design patterns you would use in your code. The generic functionality provided in a behavior module includes operations such as:

- spawning and possibly registering the process;
- sending and receiving client messages as synchronous or asynchronous calls, including defining the internal message protocol;
- storing the loop data and managing the process loop; and
- stopping the process.

The loop data is a variable that will contain the data the behavior needs to store in between calls. After the call, an updated variant of the loop data is returned. This updated loop data, often referred to as the new loop data, is passed as an argument in the next call. Loop data is also often referred to as the behavior state.

The functionality to be included in the callback module for the generic server application to deliver the specific required behavior includes the following:

- Initializing the process loop data, and, if the process is registered, the process name.
- Handling the specific client requests, and, if synchronous, the replies sent back to the client.
- Handling and updating the process loop data in between the process requests.
- Cleaning up the process loop data upon termination.

Generic Servers

Generic servers that implement client/server behaviors are defined in the `gen_server` behavior that comes as part of the standard library application. In explaining generic servers, we will use the `riak_core_node_watcher.erl` module from the `riak_core` application. It is a server that tracks and reports on which sub-services and nodes in a Riak cluster are available. The module headers and directives are as follows:

```
-module(riak_core_node_watcher).
-behavior(gen_server).
%% API
-export([start_link/0,service_up/2,service_down/1,node_up/0,node_down/0,services/0,
        services/1,nodes/1,avsn/0]).
%% gen_server callbacks
-export([init/1,handle_call/3,handle_cast/2,handle_info/2,terminate/2, code_change/3]).


-record(state, {status=up, services=[], peers=[], avsn=0, bcast_tref,
                bcast_mod={gen_server, abcast}}).
```

We can easily recognize generic servers through the `-behavior(gen_server)`. directive. This directive is used by the compiler to ensure all callback functions are properly exported. The record state is used in the server loop data.

Starting Your Server

With the `gen_server` behavior, instead of using the `spawn` and `spawn_link` BIFs, you will use the `gen_server:start` and `gen_server:start_link` functions. The main difference between `spawn` and `start` is the synchronous nature of the call. Using `start` instead of `spawn` makes starting the worker process more deterministic and prevents unforeseen race conditions, as the call will not return the PID of the worker until it has been initialized. You call the functions with either of:

```
gen_server:start_link(ServerName, CallbackModule, Arguments, Options)
gen_server:start_link(CallbackModule, Arguments, Options)
```

`ServerName` is a tuple of the format `{local, Name}` or `{global, Name}`, denoting a local or global `Name` for the process alias if it is to be registered. Global names allow servers to be transparently accessed across a cluster of distributed Erlang nodes. If you do not want to register the process and instead reference it using its PID, you omit the argument and use a `start_link/3` or `start/3` function call instead. `CallbackModule` is the name of the module in which the specific callback functions are placed, `Arguments` is a valid Erlang term that is passed to the `init/1` callback function, while `Options` is a list that allows you to set the memory management flags `fullsweep_after` and `heapsize`, as well as other tracing and debugging flags.

In our example, we call `start_link/4`, registering the process with the same name as the callback module, using the `?MODULE` macro call. This macro is expanded to the name of the module it is defined in by the preprocessor when compiling the code. It is always good practice to name your behavior with an alias that is the same as the callback module it is implemented in. We don't pass any arguments, and as a result, just send the empty list. The options list is kept empty:

```
start_link() ->
    gen_server:start_link({local, ?MODULE}, ?MODULE, [], []).
```

The obvious difference between the `start_link` and `start` functions is that `start_link` links to its parent, most often a supervisor, while `start` doesn't. This needs a special mention as it is an OTP behavior's responsibility to link itself to the supervisor. The `start` functions are often used when testing behaviors from the shell, as a typing error causing the shell process to crash would not affect the behavior. All variants of the `start` and `start_link` functions return `{ok, Pid}`.

The `start` and `start_link` functions will spawn a new process that calls the `init(Arguments)` callback function in the `CallbackModule`, with the `Arguments` supplied. The `init` function must initialize the `LoopData` of the server and has to return a tuple of the format `{ok, LoopData}`. `LoopData` contains the first instance of the loop data that will be passed between the callback functions. If you want to store some of the arguments you passed to the `init` function, you would do so in the `LoopData` variable. The `LoopData` in the Riak node watcher server is the result of the `schedule_broadcast/1` called with a record of type `state` where the fields are set to the default values:

```
init([]) ->

    %% Watch for node up/down events
    net_kernel:monitor_nodes(true),

    %% Setup ETS table to track node status
    ets:new(?MODULE, [protected, named_table]),

    {ok, schedule_broadcast(#state{})}.
```

Although the supervisor process might call the `start_link/4` function, a different process calls the `init/1` callback: the one that was just spawned. As the purpose of this server is to notice, record, and broadcast the availability of sub-services within Riak, the initialization asks the Erlang runtime to notify it of such events, and sets up a table to store this information in. This needs to be done during initialization, as any calls to the server would fail if that structure did not yet exist. Do only what is necessary and minimize the operations in your `init` function, as the call to `init` is a synchronous call that prevents all of the other serialized processes from starting until it returns.

Passing Messages

If you want to send a synchronous message to your server, you use the `gen_server:call/2` function. Asynchronous calls are made using the `gen_server:cast/2` function. Let's start by taking two functions from Riak's service API; we will provide the rest of the code later. They are called by the client process and result in a synchronous message being sent to the server process registered with the same name as the callback module. Note that validating the data sent to the server should occur on the client side. If the client sends incorrect information, the server should terminate.

```
service_up(Id, Pid) ->
    gen_server:call(?MODULE, {service_up, Id, Pid}).

service_down(Id) ->
    gen_server:call(?MODULE, {service_down, Id}).
```

Upon receiving the messages, the `gen_server` process calls the `handle_call/3` callback function dealing with the messages in the same order in which they were sent:

```
handle_call({service_up, Id, Pid}, _From, State) ->
    %% Update the set of active services locally
    Services = ordsets:add_element(Id, State#state.services),
    S2 = State#state { services = Services },

    %% Remove any existing mrefs for this service
    delete_service_mref(Id),

    %% Setup a monitor for the Pid representing this service
    Mref = erlang:monitor(process, Pid),
    erlang:put(Mref, Id),
    erlang:put(Id, Mref),

    %% Update our local ETS table and broadcast
    S3 = local_update(S2),
    {reply, ok, update_avsn(S3)}.

handle_call({service_down, Id}, _From, State) ->
    %% Update the set of active services locally
    Services = ordsets:del_element(Id, State#state.services),
    S2 = State#state { services = Services },

    %% Remove any existing mrefs for this service
    delete_service_mref(Id),

    %% Update local ETS table and broadcast
    S3 = local_update(S2),
    {reply, ok, update_avsn(S3)};
```

Note the return value of the callback function. The tuple contains the control atom `reply`, telling the `gen_server` generic code that the second element of the tuple (which in both of these cases is the atom `ok`) is the reply sent back to the client. The third element of the tuple is the new State, which, in a new iteration of the server, is passed as the third argument to the `handle_call/3` function; in both cases here it is updated to reflect the new set of available services. The argument `_From` is a tuple containing a unique message reference and the client process identifier. The tuple as a whole is used in library functions that we will not be discussing in this chapter. In the majority of cases, you will not need it.

The `gen_server` library module has a number of mechanisms and safeguards built in that operate behind the scenes. If your client sends a synchronous message to your server and you do not get a response within five seconds, the process executing the `call/2` function is terminated. You can override this by using `gen_server:call(Name, Message, Timeout)` where `Timeout` is a value in milliseconds or the atom `infinity`.

The timeout mechanism was originally put in place for deadlock prevention purposes, ensuring that servers that accidentally call each other are terminated after the default timeout. The crash report would be logged, and hopefully would result in the error being debugged and fixed. Most applications will function appropriately with a timeout of five seconds, but under very heavy loads, you might have to fine-tune the value and possibly even use `infinity`; this choice is application-dependent. All of the critical code in Erlang/OTP uses `infinity`. Various places in Riak use different values for the timeout: `infinity` is common between coupled pieces of the internals, while `Timeout` is set based on a user-passed parameter in cases where the client code talking to Riak has specified that an operation should be allowed to time out.

Other safeguards when using the `gen_server:call/2` function include the case of sending a message to a nonexistent server and the case of a server crashing before sending its reply. In both cases, the calling process will terminate. In raw Erlang, sending a message that is never pattern-matched in a receive clause is a bug that can cause a memory leak. Two different strategies are used in Riak to mitigate this, both of which involve “catchall” matching clauses. In places where the message might be user-initiated, an unmatched message might be silently discarded. In places where such a message could only come from Riak’s internals, it represents a bug and so will be used to trigger an error-alerting internal crash report, restarting the worker process that received it.

Sending asynchronous messages works in a similar way. Messages are sent asynchronously to the generic server and handled in the `handle_cast/2` callback function. The function has to return a tuple of the format `{reply, NewState}`. Asynchronous calls are used when we are not interested in the request of the server and are not worried about producing more messages than the server can consume. In cases where we are not interested in a response but want to wait until the message has been handled before sending the next request, we would use a `gen_server:call/2`, returning the atom `ok` in the reply. Picture a process generating database entries at a faster rate than Riak can consume. By using asynchronous calls, we risk filling up the process mailbox and make the node

run out of memory. Riak uses the message-serializing properties of synchronous gen_server calls to regulate load, processing the next request only when the previous one has been handled. This approach eliminates the need for more complex throttling code: in addition to enabling concurrency, gen_server processes can also be used to introduce serialization points.

Stopping the Server

How do you stop the server? In your handle_call/3 and handle_cast/2 callback functions, instead of returning {reply, Reply, NewState} or {noreply, NewState}, you can return {stop, Reason, Reply, NewState} or {stop, Reason, NewState}, respectively. Something has to trigger this return value, often a stop message sent to the server. Upon receiving the stop tuple containing the Reason and State, the generic code executes the terminate(Reason, State) callback.

The terminate function is the natural place to insert the code needed to clean up the State of the server and any other persistent data used by the system. In our example, we send out one last message to our peers so that they know that this node watcher is no longer up and watching. In this example, the variable State contains a record with the fields status and peers:

```
terminate(_Reason, State) ->
    %% Let our peers know that we are shutting down
    broadcast(State#state.peers, State#state { status = down }).
```

Use of the behavior callbacks as library functions and invoking them from other parts of your program is an extremely bad practice. For example, you should never call riak_core_node_watcher:init(Args) from another module to retrieve the initial loop data. Such retrievals should be done through a synchronous call to the server. Calls to behavior callback functions should originate only from the behavior library modules as a result of an event occurring in the system, and never directly by the user.

15.4 Other Worker Behaviors

A large number of other worker behaviors can and have been implemented using these same ideas.

Finite State Machines

Finite state machines (FSMs), implemented in the gen_fsm behavior module, are a crucial component when implementing protocol stacks in telecom systems (the problem domain Erlang was originally invented for). States are defined as callback functions named after the state that return a tuple containing the next State and the updated loop data. You can send events to these states

synchronously and asynchronously. The finite state machine callback module should also export the standard callback functions such as `init`, `terminate`, and `handle_info`.

Of course, finite state machines are not telecom specific. In Riak, they are used in the request handlers. When a client issues a request such as `get`, `put`, or `delete`, the process listening to that request will spawn a process implementing the corresponding `gen_fsm` behavior. For instance, the `riak_kv_get_fsm` is responsible for handling a `get` request, retrieving data and sending it out to the client process. The FSM process will pass through various states as it determines which nodes to ask for the data, as it sends out messages to those nodes, and as it receives data, errors, or timeouts in response.

Event Handlers

Event handlers and managers are another behavior implemented in the `gen_event` library module. The idea is to create a centralized point that receives events of a specific kind. Events can be sent synchronously and asynchronously with a predefined set of actions being applied when they are received. Possible responses to events include logging them to file, sending off an alarm in the form of an SMS, or collecting statistics. Each of these actions is defined in a separate callback module with its own loop data, preserved between calls. Handlers can be added, removed, or updated for every specific event manager. So, in practice, for every event manager there could be many callback modules, and different instances of these callback modules could exist in different managers. Event handlers include processes receiving alarms, live trace data, equipment related events or simple logs.

One of the uses for the `gen_event` behavior in Riak is for managing subscriptions to “ring events”, i.e., changes to the membership or partition assignment of a Riak cluster. Processes on a Riak node can register a function in an instance of `riak_core_ring_events`, which implements the `gen_event` behavior. Whenever the central process managing the ring for that node changes the membership record for the overall cluster, it fires off an event that causes each of those callback modules to call the registered function. In this fashion, it is easy for various parts of Riak to respond to changes in one of Riak’s most central data structures without having to add complexity to the central management of that structure.

Most common concurrency and communication patterns are handled with the three primary behaviors we’ve just discussed: `gen_server`, `gen_fsm`, and `gen_event`. However, in large systems, some application-specific patterns emerge over time that warrant the creation of new behaviors. Riak includes one such behavior, `riak_core_vnode`, which formalizes how virtual nodes are implemented. Virtual nodes are the primary storage abstraction in Riak, exposing a uniform interface for key-value storage to the request-driving FSMs. The interface for callback modules is specified using the `behavior_info/1` function, as follows:

```
behavior_info(callbacks) ->
```

```
[{init,1},
 {handle_command,3},
 {handoff_starting,2},
 {handoff_cancelled,1},
 {handoff_finished,2},
 {handle_handoff_command,3},
 {handle_handoff_data,2},
 {encode_handoff_item,2},
 {is_empty,1},
 {terminate,2},
 {delete,1}];
```

The above example shows the `behavior_info/1` function from `riak_core_vnode`. The list of `{CallbackFunction, Arity}` tuples defines the contract that callback modules must follow. Concrete virtual node implementations must export these functions, or the compiler will emit a warning. Implementing your own OTP behaviors is relatively straightforward. Alongside defining your callback functions, using the `proc_lib` and `sys` modules, you need to start them with particular functions, handle system messages and monitor the parent in case it terminates.

15.5 Supervisors

The supervisor behavior’s task is to monitor its children and, based on some preconfigured rules, take action when they terminate. Children consist of both supervisors and worker processes. This allows the Riak codebase to focus on the correct case, which enables the supervisor to handle software bugs, corrupt data or system errors in a consistent way across the whole system. In the Erlang world, this non-defensive programming approach is often referred to the “let it crash” strategy. The children that make up the supervision tree can include both supervisors and worker processes. Worker processes are OTP behaviors including the `gen_fsm`, `gen_server`, and `gen_event`. The Riak team, not having to handle borderline error cases, get to work with a smaller code base. This code base, because of its use of behaviors, is smaller to start off with, as it only deals with specific code. Riak has a top-level supervisor like most Erlang applications, and also has sub-supervisors for groups of processes with related responsibilities. Examples include Riak’s virtual nodes, TCP socket listeners, and query-response managers.

Supervisor Callback Functions

To demonstrate how the supervisor behavior is implemented, we will use the `riak_core_sup.erl` module. The Riak core supervisor is the top level supervisor of the Riak core application. It starts a set of static workers and supervisors, together with a dynamic number of workers handling the HTTP and HTTPS bindings of the node’s RESTful API defined in application specific configuration files. In a similar way to `gen_servers`, all supervisor callback modules must include the

`-behavior(supervisor)`. directive. They are started using the `start` or `start_link` functions which take the optional `ServerName`, the `CallBackModule`, and an `Argument` which is passed to the `init/1` callback function.

Looking at the first few lines of code in the `riak_core_sup.erl` module, alongside the behavior directive and a macro we will describe later, we notice the `start_link/3` function:

```
-module(riak_core_sup).
-behavior(supervisor).
%% API
-export([start_link/0]).
%% Supervisor callbacks
-export([init/1]).
-define(CHILD(I, Type), {I, {I, start_link, []}, permanent, 5000, Type, [I]}).
start_link() ->
    supervisor:start_link({local, ?MODULE}, ?MODULE, []).
```

Starting a supervisor will result in a new process being spawned, and the `init/1` callback function being called in the callback module `riak_core_sup.erl`. The `ServerName` is a tuple of the format `{local, Name}` or `{global, Name}`, where `Name` is the supervisor's registered name. In our example, both the registered name and the callback module are the atom `riak_core_sup`, originating from the `?MODULE` macro. We pass the empty list as an argument to `init/1`, treating it as a null value. The `init` function is the only supervisor callback function. It has to return a tuple with format:

```
{ok, {SupervisorSpecification, ChildSpecificationList}}
```

where `SupervisorSpecification` is a 3-tuple `{RestartStrategy, AllowedRestarts, MaxSeconds}` containing information on how to handle process crashes and restarts. `RestartStrategy` is one of three configuration parameters determining how the behavior's siblings are affected upon abnormal termination:

- `one_for_one`: other processes in the supervision tree are not affected.
- `rest_for_one`: processes started after the terminating process are terminated and restarted.
- `one_for_all`: all processes are terminated and restarted.

`AllowedRestarts` states how many times any of the supervisor children may terminate in `MaxSeconds` before the supervisor terminates itself (and its children). When ones terminates, it sends an `EXIT` signal to its supervisor which, based on its restart strategy, handles the termination accordingly. The supervisor terminating after reaching the maximum allowed restarts ensures that cyclic restarts and other issues that cannot be resolved at this level are escalated. Chances are that the issue is in a process located in a different sub-tree, allowing the supervisor receiving the escalation to terminate the affected sub-tree and restart it.

Examining the last line of the `init/1` callback function in the `riak_core_sup.erl` module, we notice that this particular supervisor has a one-for-one strategy, meaning that the processes are

independent of each other. The supervisor will allow a maximum of ten restarts before restarting itself.

`ChildSpecificationList` specifies which children the supervisor has to start and monitor, together with information on how to terminate and restart them. It consists of a list of tuples of the following format:

```
{Id, {Module, Function, Arguments}, Restart, Shutdown, Type, ModuleList}
```

`Id` is a unique identifier for that particular supervisor. `Module`, `Function`, and `Arguments` is an exported function which results in the behavior `start_link` function being called, returning the tuple of the format `{ok, Pid}`. The `Restart` strategy dictates what happens depending on the termination type of the process, which can be:

- transient processes, which are never restarted;
- temporary processes, are restarted only if they terminate abnormally; and
- permanent processes, which are always restarted, regardless of the termination being normal or abnormal.

`Shutdown` is a value in milliseconds referring to the time the behavior is allowed to execute in the `terminate` function when terminating as the result of a restart or shutdown. The atom `infinity` can also be used, but for behaviors other than supervisors, it is highly discouraged. `Type` is either the atom `worker`, referring to the generic servers, event handlers and finite state machines, or the atom `supervisor`. Together with `ModuleList`, a list of modules implementing the behavior, they are used to control and suspend processes during the runtime software upgrade procedures. Only existing or user implemented behaviors may be part of the child specification list and hence included in a supervision tree.

With this knowledge at hand, we should now be able to formulate a restart strategy defining inter-process dependencies, fault tolerance thresholds and escalation procedures based on a common architecture. We should also be able to understand what is going on in the `init/1` example of the `riak_core_sup.erl` module. First of all, study the `CHILD` macro. It creates the child specification for one child, using the callback module name as `Id`, making it permanent and giving it a shut down time of 5 seconds. Different child types can be workers or supervisors. Have a look at the example, and see what you can make out of it:

```
-define(CHILD(I, Type), {I, {I, start_link, []}, permanent, 5000, Type, [I]}).

init([]) ->
    RiakWebs = case lists:flatten(riak_core_web:bindings(http),
                                    riak_core_web:bindings(https)) of
        [] ->
            %% check for old settings, in case app.config
            %% was not updated
            riak_core_web:old_binding();
```

```

Binding ->
    Binding
end,
Children =
    [?CHILD(riak_core_vnode_sup, supervisor),
     ?CHILD(riak_core_handoff_manager, worker),
     ?CHILD(riak_core_handoff_listener, worker),
     ?CHILD(riak_core_ring_events, worker),
     ?CHILD(riak_core_ring_manager, worker),
     ?CHILD(riak_core_node_watcher_events, worker),
     ?CHILD(riak_core_node_watcher, worker),
     ?CHILD(riak_core_gossip, worker) |
      RiakWebs
    ],
{ok, {{one_for_one, 10, 10}, Children}}.

```

Most of the `Children` started by this supervisor are statically defined workers (or in the case of the `vnode_sup`, a supervisor). The exception is the `RiakWebs` portion, which is dynamically defined depending on the HTTP portion of Riak’s configuration file.

With the exception of library applications, every OTP application, including those in Riak, will have their own supervision tree. In Riak, various top-level applications are running in the Erlang node, such as `riak_core` for distributed systems algorithms, `riak_kv` for key/value storage semantics, `webmachine` for HTTP, and more. We have shown the expanded tree under `riak_core` to demonstrate the multi-level supervision going on. One of the many benefits of this structure is that a given subsystem can be crashed (due to bug, environmental problem, or intentional action) and only that subtree will in a first instance be terminated.

The supervisor will restart the needed processes and the overall system will not be affected. In practice we have seen this work well for Riak. A user might figure out how to crash a virtual node, but it will just be restarted by `riak_core vnode_sup`. If they manage to crash that, the `riak_core` supervisor will restart it, propagating the termination to the top-level supervisor. This failure isolation and recovery mechanism allows Riak (and Erlang) developers to straightforwardly build resilient systems.

The value of the supervisory model was shown when one large industrial user created a very abusive environment in order to find out where each of several database systems would fall apart. This environment created random huge bursts of both traffic and failure conditions. They were confused when Riak simply wouldn’t stop running, even under the worst such arrangement. Under the covers, of course, they were able to make individual processes or subsystems crash in multiple ways—but the supervisors would clean up and restart things to put the whole system back into working order every time.

Applications

The application behavior we previously introduced is used to package Erlang modules and resources into reusable components. In OTP, there are two kinds of applications. The most common form, called normal applications, will start a supervision tree and all of the relevant static workers. Library applications such as the Standard Library, which come as part of the Erlang distribution, contain library modules but do not start a supervision tree. This is not to say that the code may not contain processes or supervision trees. It just means they are started as part of a supervision tree belonging to another application.

An Erlang system will consist of a set of loosely coupled applications. Some are written by the developers, some are available as open source, and others are part of the Erlang/OTP distribution. The Erlang runtime system and its tools treat all applications equally, regardless of whether they are part of the Erlang distribution or not.

15.6 Replication and Communication in Riak

Riak was designed for extreme reliability and availability at a massive scale, and was inspired by Amazon’s Dynamo storage system [DHJ⁺07]. Dynamo and Riak’s architectures combine aspects of both Distributed Hash Tables (DHTs) and traditional databases. Two key techniques that both Riak and Dynamo use are *consistent hashing* for replica placement and a *gossip protocol* for sharing common state.

Consistent hashing requires that all nodes in the system know about each other, and know what partitions each node owns. This assignment data could be maintained in a centrally managed configuration file, but in large configurations, this becomes extremely difficult. Another alternative is to use a central configuration server, but this introduces a single point of failure in the system. Instead, Riak uses a gossip protocol to propagate cluster membership and partition ownership data throughout the system.

Gossip protocols, also called epidemic protocols, work exactly as they sound. When a node in the system wishes to change a piece of shared data, it makes the change to its local copy of the data and gossips the updated data to a random peer. Upon receiving an update, a node merges the received changes with its local state and gossips again to another random peer.

When a Riak cluster is started, all nodes must be configured with the same partition count. The consistent hashing ring is then divided by the partition count and each interval is stored locally as a {HashRange, Owner} pair. The first node in a cluster simply claims all the partitions. When a new node joins the cluster, it contacts an existing node for its list of {HashRange, Owner} pairs. It then claims (partition count)/(number of nodes) pairs, updating its local state to reflect its new ownership. The updated ownership information is then gossiped to a peer. This updated state then spread throughout the entire cluster using the above algorithm.

By using a gossip protocol, Riak avoids introducing a single point of failure in the form of a centralized configuration server, relieving system operators from having to maintain critical cluster configuration data. Any node can then use the gossiped partition assignment data in the system to route requests. When used together, the gossip protocol and consistent hashing enable Riak to function as a truly decentralized system, which has important consequences for deploying and operating large-scale systems.

15.7 Conclusions and Lessons Learned

Most programmers believe that smaller and simpler codebases are not only easier to maintain, they often have fewer bugs. By using Erlang’s basic distribution primitives for communication in a cluster, Riak can start out with a fundamentally sound asynchronous messaging layer and build its own protocols without having to worry about that underlying implementation. As Riak grew into a mature system, some aspects of its networked communication moved away from use of Erlang’s built-in distribution (and toward direct manipulation of TCP sockets) while others remained a good fit for the included primitives. By starting out with Erlang’s native message passing for everything, the Riak team was able to build out the whole system very quickly. These primitives are clean and clear enough that it was still easy later to replace the few places where they turned out to not be the best fit in production.

Also, due to the nature of Erlang messaging and the lightweight core of the Erlang VM, a user can just as easily run 12 nodes on 1 machine or 12 nodes on 12 machines. This makes development and testing much easier when compared to more heavyweight messaging and clustering mechanisms. This has been especially valuable due to Riak’s fundamentally distributed nature. Historically, most distributed systems are very difficult to operate in a “development mode” on a single developer’s laptop. As a result, developers often end up testing their code in an environment that is a subset of their full system, with very different behavior. Since a many-node Riak cluster can be trivially run on a single laptop without excessive resource consumption or tricky configuration, the development process can more easily produce code that is ready for production deployment.

The use of Erlang/OTP supervisors makes Riak much more resilient in the face of subcomponent crashes. Riak takes this further; inspired by such behaviors, a Riak cluster is also able to easily keep functioning even when whole nodes crash and disappear from the system. This can lead to a sometimes-surprising level of resilience. One example of this was when a large enterprise was stress-testing various databases and intentionally crashing them to observe their edge conditions. When they got to Riak, they became confused. Each time they would find a way (through OS-level manipulation, bad IPC, etc) to crash a subsystem of Riak, they would see a very brief dip in performance and then the system returned to normal behavior. This is a direct result of a thoughtful “let it crash” approach. Riak was cleanly restarting each of these subsystems on demand, and the

overall system simply continued to function. That experience shows exactly the sort of resilience enabled by Erlang/OTP's approach to building programs.

Acknowledgments

This chapter is based on Francesco Cesarini and Simon Thompson's 2009 lecture notes from the central European Functional Programming School held in Budapest and Komárno. Major contributions were made by Simon Thompson of the University of Kent in Canterbury, UK. A special thank you goes to all of the reviewers, who at different stages in the writing of this chapter provided valuable feedback.

Selenium WebDriver

Simon Stewart

Selenium はブラウザの自動化ツールで、ウェブアプリケーションのエンドツーエンドテストを書くときによく使われる。ブラウザの自動化ツールが行うのは、その名前から想像できるとおりのことだ。ブラウザを制御して、繰り返し行われるタスクを自動化できる。解決しようとしている問題は単純なものだが、本章で説明するとおり、その裏側ではさまざまなことが起こっている。

Selenium のアーキテクチャについて語る前に、プロジェクト内のさまざまなパーツがどのように関連するのかを説明しておこう。上位レベルから見ると、Selenium は三つのツールを組み合わせたものである。そのうちの一つである Selenium IDE は Firefox の拡張で、これを使うとテストの記録や再生ができる。この「記録/再生」のパラダイムは限定的なものであり、多くのユーザーにとっては物足りないものだ。そこで登場するのが第二のツールである Selenium WebDriver だ。さまざまな言語向けの API を提供しており、より細やかな制御を行うことができるし、標準のソフトウェア開発の流れに組み込むこともできる。最後のツールは Selenium Grid だ。これを使うと、Selenium API で分散環境にあるブラウザのインスタンスを制御でき、テストを並列に実行できるようになる。Selenium プロジェクト内では、これらのツールはそれぞれ “IDE”、“WebDriver” そして “Grid” と呼ばれている。本章で扱うのは Selenium WebDriver のアーキテクチャである。

本章の内容は Selenium 2.0 のベータ版に基づいており、2010 年の後半に執筆された。もしそれより後にこれを読んでいるのなら、世間はさらに進んでいることだろう。ここで取り上げたアーキテクチャに関する選択が実際にどのような形で実装されたのかを見られるかもしれない。もし 2010 年後半よりも前にこれを読んでいるのなら…おめでとう! あなたはついにタイムマシンを手に入れたんだね! 頼むから、宝くじの当選番号を教えてくれないかな?

16.1 歴史

ジェイソン・ハギンズが Selenium プロジェクトを立ち上げたのは 2004 年のこと。そのとき彼は ThoughtWorks で社内用の勤怠管理 (T&E: Time and Expenses) システムを開発していた。それは、Javascript を使いまくるシステムだった。Internet Explorer がシェアを支配していた時代だったが、ThoughtWorks ではそれ以外のブラウザも使われており (特に Mozilla 系が多かった)、T&E アプリがそういったブラウザで動かないというバグ報告も受けていた。当時のオープンソースのテストツールといえば、特定のブラウザ (たいてい IE) に絞ったものかブラウザをシミュレートするもの (HttpUnit など) しかなかった。商用のツールのライセンスを購入するというのは、限られた予算の社内システムでは無理な話だった。そのため、そもそも選択肢にすら入っていなかった。

自動化が困難なときには手動でのテストに頼るのが一般的だ。しかし、チームの規模が小さいときやリリース頻度が極端に高いときなどにはこの方式はスケールしない。また、自動化できるはずの手順をいちいちやってもらうように頼むのも、彼らの人間性を浪費しているように思える。はつきり言って、人はくだらない繰り返し作業をするのが苦手だ。機械に比べて能率も悪いし間違いも多い。手動でのテストも選択肢から消えた。

幸いなことに、テスト対象のブラウザはすべて Javascript をサポートしていた。ジェイソンやチームのメンバーが Javascript を使ってテストツールを書こうとしたのも自然なことだった。そのツールを使って、アプリケーションのふるまいを検証しようとしたのだ。既にあった FIT¹ の影響を受け、生の Javascript ではなく表形式の構文を採用した。そのおかげで、プログラミングの経験が少ない人でも HTML ファイルにキーワードを書き込んでテストを書けるようになった。このツールは当初 “Selenium” と呼ばれていたが、後に “Selenium Core” という名前に変わり、Apache 2 ライセンスで 2004 年に公開された。

Selenium の表の書式は、FIT の ActionFixture と似ており、テーブルの各行が三つのカラムに分かれている。最初のカラムには実行するコマンド名を指定し、次のカラムには要素の ID、そして最後のカラムにはオプションの値を指定する。たとえばこれは、“Selenium WebDriver” という文字列を name が “q” である要素に入力する例だ。

```
type      name=q      Selenium WebDriver
```

Selenium は Javascript だけで書かれているので、初期の設計では Core やテスト群をテスト対象のアプリケーション (AUT) と同じサーバーに置く必要があった。ブラウザのセキュリティポリシーや Javascript のサンドボックスの制限を回避するためである。しかし、現実的にそれが不可能な場合だってある。さらに悪いことに、開発者用の IDE には巨大なコードベースを縦横無尽に渡り歩くための機能が用意されているのに、HTML 用のツールにはそういうものが存在しない。程なくわかったことだが、そんなに大きくなないテストスイートの保守ですら面倒つきつい作業になった。²

¹<http://fit.c2.com>

²これは FIT でも同じだった。プロジェクトのメンバーの一人であるジェイムズ・ショアが、その弱点について <http://jamesshore.com/Blog/The-Problems-With-Acceptance-Testing.html> で説明している。

この問題やその他の問題を解決するために、HTTP プロキシが書かれた。これを使い、すべての HTTP リクエストを Selenium で捕捉できるようにしたのだ。このプロキシを使えば、“同一生成元ポリシー”の制約の多くを回避できるようになった。このポリシーは、Javascript からそのページを提供するサーバー以外への呼び出しを許可しないというものである。これを回避できたことで、最初の弱点は何とかしのげるようになった。この設計のおかげで、Selenium のバインディングを複数の言語で書けるようになつた。必要なのは、単に特定の URL に HTTP リクエストを送信する機能だけである。連結用の書式は Selenium Core の表形式の構文をもとにして作られ、その表形式の構文と併せて後に “Selenese” として知られるようになった。他言語のバインディングはブラウザを遠隔操作するものだったので、このツールは “Selenium Remote Control” あるいは “Selenium RC” と呼ばれることになった。

Selenium の開発が進む一方で、別のブラウザ自動化フレームワークも ThoughtWorks で生まれていた。それが WebDriver で、最初のコードは 2007 年始めに公開された。WebDriver は、あるプロジェクトから派生したツールである。そのプロジェクトでは、エンドツーエンドテストをテストツールから分離しようとしていたのだ。一般的な例にならい、分離する方法として Adapter パターンを使った。WebDriver は、これまでに多数のプロジェクトに適用されてきた結果から生み出されたもので、当初は HtmlUnit に対するラッパーだった。その後、Internet Explorer や Firefox のサポートもすぐに追加された。

WebDriver が公開されたとき、WebDriver と Selenium RC の間には大きな違いがあった。しかしどちらも、ブラウザの自動化のための API を提供するというニッチを狙っているという点では共通していた。ユーザーから見たときの最大の相違点は、Selenium RC は辞書ベースの API ですべてのメソッドを单一のクラスで公開しているのに対して WebDriver はよりオブジェクト指向な API を提供していることだった。さらに、WebDriver がサポートするのが Java だけであるのに対して Selenium RC はさまざまな言語に対応していた。それ以外にも技術的な違いがあった。Selenium Core (RC の元になっているもの) は基本的に Javascript アプリケーションで、ブラウザのセキュリティサンドボックス内で動作する。WebDriver はネイティブにブラウザにバインドすることを試みており、フレームワーク自身の開発に要する労力を犠牲にしてでもブラウザのセキュリティモデルを回避しようとしている。

2009 年 8 月、二つのプロジェクトが合流することが発表された。その結果として登場したのが Selenium WebDriver だ。執筆時点では、WebDriver がサポートしている言語は Java と C#、Python、そして Ruby である。また、Chrome や Firefox、Internet Explorer、Opera、そして Android や iPhone のブラウザに対応している。Selenium WebDriver には姉妹プロジェクトもある。ソースコードリポジトリは別だが本体のプロジェクトと密接に連携して開発が進められており、Perl のバインディングや BlackBerry のブラウザ用の実装を用意している。また、“headless(画面なし)” の WebKit にも対応する。これはテストを継続的インテグレーションサーバーで画面なしで実行しなければならないときに便利だ。元々の Selenium RC の仕組みは今でも保守されており、WebDriver が未対応なブラウザもこれでサポートすることができる。

16.2 ジャーゴンについての余談

残念ながら、Selenium プロジェクトにはジャーゴンが氾濫している。これまでに登場したものをまとめてみよう。

- *Selenium Core* は当初の Selenium の実装の中心となる部分のこと、ブラウザを制御するための Javascript 群である。時には “Selenium” と呼ばれることもあるし、“Core” と呼ばれることがある。
- *Selenium RC* は、Selenium Core 用の言語バインディングを表す。紛らわしいことに、これもまた “Selenium” と呼ばれることもあるし “RC” と呼ばれることもある。現在は Selenium WebDriver に置き換わっており、RC の API は “Selenium 1.x API” と呼ばれる。
- *Selenium WebDriver* は、かつて RC が扱っていたのと同じニッチを埋めるものであり、1.x 系のバインディングを包含している。この言葉は、言語バインディングだけでなく個別のブラウザ制御用コードの実装のことも指す。一般的には単に “WebDriver” と呼ばれており、時には Selenium 2 と呼ばれることがある。そのうち、さらに短縮して “Selenium” と呼ばれるようになるであろうことは疑う余地もない。

賢明な読者のみなさんはお気づきだろうが、“Selenium” という用語があまりにもいろんな場面で使われすぎている。しかし幸いなことに、その「Selenium」が何を表すのかは話の流れで明確になることが多い。

最後に、これから使うであろうフレーズをもうひとつ紹介しておく。“ドライバ”だ。この用語は、WebDriver API の特定の実装を指す名前として使う。たとえば、Firefox ドライバや Internet Explorer ドライバといった使い方をする。

16.3 アーキテクチャについて

個々のピースを見てそのつながり具合を理解する前に、まずはプロジェクト全体のアーキテクチャや開発について知っておくと有用だ。簡潔にまとめると、このようになる。

- コストを抑える。
- ユーザーをエミュレートする。
- ドライバが動作することを実証する…
- …が、それがどのように動作するのかを知る必要はないようとする。
- バス係数を上げる。
- Javascript の実装を尊重する。
- すべてのメソッドコールは RPC である。
- 我々は、オープンソースプロジェクトである。

コストを抑える

○○プラットフォーム上の■■ブラウザをサポートするというのは本質的にコストのかかる提案である。最初の開発の面でも、その後の保守の面でも。その品質を高く保ちつつその他の原則を破らずに済む手段がもしあれば、それこそが我々の目指す道となる。我々が可能な限り Javascript を採用しているのも、それが理由だ。後ほど詳しく説明する。

ユーザーをエミュレートする

WebDriver は、ユーザーがウェブアプリケーションとやりとりする内容をそのまま正確にシミュレートするように作られている。ユーザーの入力をシミュレートする一般的な手段は、Javascript を使って一連のイベントを合成することだ。アプリケーション側からは、ユーザーがその操作をしたのと同じように見える。このような“イベント合成”方式は、面倒なことだけである。ブラウザによって、また場合によっては同じブラウザでもバージョンによって、発火するイベントが微妙に異なったり微妙に違う値をとったりする。さらに問題を複雑にしているのが、多くのブラウザがこのような手段でのフォーム要素の操作を許可していないという事実だ。たとえばファイルアップロード要素などは、セキュリティの観点から“イベント合成”ができないようになっている。

WebDriver では、可能な場合は Javascript を使わず OS レベルでイベントを発火させるという手法をとる。このような“ネイティブイベント”はブラウザが生成するものではないので、イベント合成方式に立ちはだかったセキュリティの制約を回避できる。そして、OS の機能を使っているので、特定のプラットフォーム上のあるブラウザで動作するようになれば、そのコードを再利用して同じプラットフォーム上の別のブラウザに対応させるのも容易である。不幸にも、この手法が使える環境は限られている。WebDriver とブラウザを密接にバインドでき、かつウインドウにフォーカスを合わせずにブラウザへネイティブイベントを送る方法を開発チームが見つけた場合 (Selenium のテストは実行に時間がかかるので、テストの実行中にはマシン上で他のタスクを実行できるようにしておきたい) にしか使えない。執筆時点では、この条件を満たしてネイティブイベントが使える環境は Linux と Windows である。Mac OS X では使えない。

WebDriver がユーザーの入力をどうやってエミュレートするかにかかわらず、我々は可能な限りユーザーのふるまいを再現できるよう努力している。これは RC とは対照的で、RC が提供する API は実際のユーザーの作業よりずっと低レベルな操作をするものである。

ドライバが動作することを実証する

ある意味理想主義的かもしれないが、動作しないコードなどまったく無意味である。Selenium プロジェクトでドライバが正しく動作することを実証するために、自動テストのテストケースを豊富にそろえている。その多くは“インテグレーションテスト”であり、コードがコンパ

イルされていることとブラウザを使ってウェブサーバーとやりとりすることを前提としている。しかし、書ける場面では“ユニットテスト”も書いている。これはインテグレーションテストとは異なり、完全に再コンパイルしなくとも実行できる。執筆時点では、約 500 のインテグレーションと約 250 のユニットテストがあつてそれぞれすべてのブラウザで実行できる。バグ修正や機能追加のときには新たなテストを追加するし、我々としてはもっとユニットテストを書くように力を注いでいる。

すべてのテストがすべてのブラウザで動作するというわけではない。いくつかのブラウザでは対応していない特定の機能を確認するテストもあるし、ブラウザによって処理方法が変わる機能のテストもある。たとえば HTML5 の機能のテストがこれにあたる。HTML5 の中には、すべてのブラウザが対応しているわけではない機能もある。それにもかかわらず、主要なデスクトップブラウザには、それぞれ大量のテストのサブセットがある。ご存じの通り、500 を超えるテストを複数のプラットフォームでブラウザごとに実行するのは大変な作業だ。我々が常に鬱陶続いている課題でもある。

すべての動作原理を理解する必要はない

我々が使っているすべての言語や技術について熟練しているという人はほとんどいない。したがって、我々のアーキテクチャは各開発者が自分の得意分野だけに注力できるようにすべきだ。満足に扱えない不得手な部分をいじらなくても、やりたいことができるようとする。

バス係数を上げる

ソフトウェア開発の世界には“バス係数 (bus factor)”と呼ばれる概念がある（まじめなものではない）。この係数は、プロジェクトのコア開発者の数を表す。仮にプロジェクトに何らかの災難（バスに追突されるなど）があったときに、その人が欠けてしまえばプロジェクトが続行不能になるような人のことである。ブラウザの自動化のように複雑な作業は、特にこのようになりやすい。そこで、我々のアーキテクチャに関する決断は、この係数を可能な限り上げる方向に進めている。

Javascript の実装を尊重する

WebDriver は、他に何も手段がない場合は最終的にピュア Javascript を使ってブラウザを操作する。つまり、我々が用意する API は Javascript の実装にあわせたものでなければいけないということだ。ここで具体例を示す。HTML5 で導入された LocalStorage は、構造化されたデータをクライアント側に保存するための API である。この API のブラウザでの実装には、通常は SQLite が使われている。我々としての自然な実装は、ベースとなっているデータストアへのデータベース接続を (JDBC などで) 提供することだろう。しかし、最終的に我々が用

意した API は背後にある Javascript の実装を模したものだった。なぜなら、典型的なデータベースアクセス用 API を参考にすると Javascript の実装とは食い違ってしまうからだ。

すべてのメソッドコールは **RPC** である

WebDriver が制御するブラウザは、別のプロセスで動いている。見落としがちなことだが、これはつまり、API からのすべての呼び出しは RPC コールになるということだ。したがって、フレームワークのパフォーマンスはネットワークのレイテンシに大きな影響を受ける。通常の操作では、特に気になるほどの問題は発生しない (たいていの OS は、localhost へのルーティングを最適化している)。しかしブラウザとテストコードの間のネットワークレイテンシが増加すると、API の設計側も API の利用者側も使いづらくなる。

このせいで、API の設計に関してちょっとした対立が発生する。大きな API で荒い関数群を用意すると、複数のコールをひとまとめにしてレイテンシを下げることができる。しかしこれは、API を表現力があって使いやすいものにすることとのトレードオフとなる。たとえば、ある要素がエンドユーザーに見えるかどうかを調べるには、いくつかのチェックが必要となる。CSS のさまざまなプロパティを (親要素のプロパティも含めて) 考慮するだけでなく、おそらく要素の大きさも調べなければならないだろう。ミニマルな API なら、これらのチェックをそれぞれ個別に行うことになるだろう。WebDriver では、これらのチェックを单一のメソッド `isDisplayed` にまとめた。

結論: オープンソースです

アーキテクチャの観点からは少しずれるが、Selenium はオープンソースプロジェクトである。ここまであげてきたポイントをまとめると、我々が望むのは新しい開発者がプロジェクトに参加しやすいようにするということだ。参加するために必要な知識ができるだけ抑え、知っていてほしい言語の数も最小にして、さらに自動テストを活用して何も壊さないことを確かめる。これらはすべて、新規参入を促すのが狙いである。

最初は、このプロジェクトは複数のモジュール群に分かれていた。特定のブラウザ用のモジュールがいくつかあって、さらに全体に共通するコードやユーティリティコードを含むモジュールがあったのだ。各バインディングのソースツリーもこれらのモジュールの配下にあった。この方式は Java や C# 使いの人たちにとっては便利だったが、Rubyist や Pythonista にとってはやりにくいものだった。これはそのまま、各言語の貢献者数の違いとなって現れた。Python や Ruby のバインディングを開発できる人 (あるいは開発に興味を持つ人) がほんの少ししかいなかつたのだ。これを解決するために、2010 年の 10 月から 11 月にかけてソースコードの構成を見直し、Ruby や Python のコードはトップレベルに言語ごとのディレクトリを置いてそこで管理するようにした。Ruby や Python の世界にいるオープンソース開発者には、このほうが見慣れた構造だった。そして、それからすぐに、これらの言語のコミュニティからの貢献が目に見えて增加了。

16.4 複雑性への対応

ソフトウェアは、こぶだらけでこぼこな構造である。その原因は複雑性であり、APIを設計する側の立場で考えると、その複雑性をどこに押し込めるかを判断しなければならない。究極の選択として、複雑性を可能な限りまんべんなく広げるという方法がある。これは、APIの利用者に複雑性をすべて押しつけてしまうというものである。その対極にあるのが、API側で可能な限り複雑性を囲い込み、一か所に隔離してしまうという方法である。囲い込んだ場所は暗黒地帯となり、そこに手を入れるのはとても恐ろしい作業になるだろう。しかしその見返りとして、APIのユーザーは実装に深入りせずに済むようになる。つまり、ユーザーが複雑性に立ち向かうためのコストを設計者側で前払いしたことになる。

WebDriver の開発者は、複雑性を見つけたらそれをできるだけ特定のいくつかの場所に隔離する心がけている。そのまま広めることはしない。その理由のひとつはユーザー層である。彼らはバグや問題点を見つけることには並はずれた力を持っている。それは我々のバグリストを見れば明らかだ。しかし、彼らの多くは開発者ではないので、あまり複雑な API はうまく使いこなせない。我々が求める API は、利用者を正しい方向に導けるようなものだ。たとえば、オリジナルの Selenium API のこれらのメソッドについて考えてみよう。これらはどれも、input 要素に値を設定するために使うものだ。

- type
- typeKeys
- typeKeysNative
- keydown
- keypress
- keyup
- keydownNative
- keypressNative
- keyupNative
- attachFile

同じことを実現するための WebDriver の API は、これだ。

- sendKeys

前述のとおり、これは RC と WebDriver の思想の大きな違いのひとつを表している。WebDriver が目指すのがユーザーをエミュレートすることであるのに対して、RC が提供する API はより低レベルのものであり、ユーザーの視点では見つけづらかったり到達不能だったりするものである。typeKeys と typeKeysNative の違いは、前者が常に合成イベントを使うのにに対して後者は AWT の Robot を使ったキータイプを試みるという点である。残念なことに、AWT の Robot がキープレスイベントを送信する先はフォーカスがあたっているウィンドウになる。これは必ずしもブラウザであるとは限らない。WebDriver は、それとは対照的に、ウィンドウハンドルに対して直接イベントを送信する。そのため、ブラウザのウィンドウにフォーカスがあたっている必要はない。

WebDriver の設計

開発チームでは、WebDriver の API が“オブジェクトベース”であるよう心がけている。インターフェイスを明確に定義して单一のロールあるいは責務だけを受け持たせようとする。しかし、考えうるすべての HTML タグを個別にモデル化するわけではなく、我々が用意したのは一つの WebElement インターフェイスだけである。この方針で進めると、自動補完機能を持つ IDE を使う開発者に対して次に進むべき道を示せる。コーディングは、(たとえば Java の場合) このような具合になる。

```
WebDriver driver = new FirefoxDriver();
driver.<ここでスペースキーを打つ>
```

このときに現れるのは、比較的少なめな 13 種類のメソッドである。その中から適切なものを選ぶことになる。

```
driver.findElement(<ここでスペースキーを打つ>)
```

たいていの IDE は、ここで引数の型に関するヒントを表示する。この場合は “By” である。“By” オブジェクトには定義済みのファクトリーメソッドが数多く用意されており、By 自身のスタティックメソッドとして宣言されている。ユーザーは、あっという間にこのような状態までたどりつけるだろう。

```
driver.findElement(By.id("some_id"));
```

UnsupportedOperationExceptions やその仲間たちはとても不愉快なものだったが、その機能を必要とする一部のユーザーに対して何らかの手段で機能を公開する必要があった。大多数の他のユーザー向けの他の API には影響を及ぼさない形式で。そのための手段として、WebDriver ではロールベースのインターフェイスを活用した。たとえば JavascriptExecutor インターフェイスは、任意の Javascript コード片を現在のページのコンテキストで実行する機能を提供する。WebDriver のインスタンスをこのインターフェイスにあわせてキャストすれば、このインターフェイスの持つメソッドを使える。

組み合わせの激増への対処

ちょっと考えればすぐにわかることがあるが、WebDriver がさまざまなブラウザや言語に対応しているということは、よっぽど注意しないと保守のコストが激増してしまうということである。X 種類のブラウザと Y 種類の言語に対応しようとすると、最悪 $X \times Y$ 種類の実装を保守するはめになる。

WebDriver でサポートする言語を減らすのもひとつの方法はあるが、その道をゆくつもりはなかった。理由はふたつある。まず、ある言語から別の言語への切り替えには認識を切り替えるコストがかかる。つまり、フレームワークのユーザーにとっては、開発時に使って

ロールベースのインターフェイス

単純な Shop クラスを考えてみよう。この店では、毎日のように在庫の補充が必要となる。そこで、Stockist と協調して新たな在庫を配達する。また、従業員への給与の支払いや税金の支払いが毎月発生する。議論の都合上、これらの処理には Accountant を使うものとしよう。これをモデリングした一例を次に示す。

```
public interface Shop {  
    void addStock(StockItem item, int quantity);  
    Money getSalesTotal(Date startDate, Date endDate);  
}
```

Shop と Accountant そして Stockist の間のインターフェイスを定義するときにどのあたりに境界線を引くか。この問い合わせにはふたつの選択肢がある。理論上の線を図 16.1 のように引ける。

これはつまり、Accountant や Stockist がそれぞれのメソッドの引数として Shop を受け付けることを意味する。しかし、この方式には問題もある。Accountant がほんとうに在庫棚を使いたがっているとは思えないし、Stockist にとっては Shop が追加したさまざまな商品の値上げ情報など知らせる必要もない。そこで、もう少しましな線の引き方を図 16.2 に示す。

この場合、Shop はふたつのインターフェイスを実装しなければならない。しかし、これらのインターフェイスは、Shop が Accountant と Stockist に対して満たすべきロールを明確に定義している。これが、ロールベースのインターフェイスである。

```
public interface HasBalance {  
    Money getSalesTotal(Date startDate, Date endDate);  
}  
  
public interface Stockable {  
    void addStock(StockItem item, int quantity);  
}  
  
public interface Shop extends HasBalance, Stockable {  
}
```

いるのと同じ言語でテストを書ければ好都合になる。次に、ひとつのプロジェクトに複数の言語を混在させるとチーム内で不満が出る可能性がある。場合によっては、会社のコーディング規約などで特定の技術だけを使うよう強制されているかもしれない（しかし、ありがたいことに、最近は二番目の理由はあまり聞かなくなってきた）。したがって、サポートする言語を減らすという選択肢は取れなかった。

サポートするブラウザを減らすというのも、あり得ない話だ。かつて WebDriver で Firefox

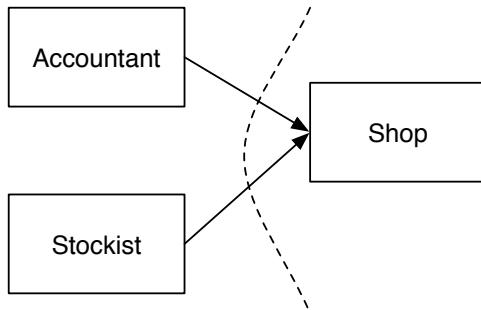


図 16.1: Accountant と Stockist が Shop に依存する

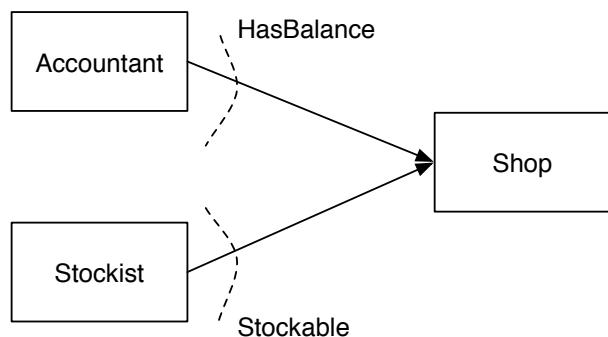


図 16.2: Shop は HasBalance と Stockable を実装する

2 のサポートを打ち切ったときには大騒ぎになった。当時のブラウザ市場での Firefox 2 のシェアは既に 1%を割っていたというのに。

残された道はひとつ。各言語のバインディングから、すべてのブラウザを同じように扱えるようにするという道だ。統一形式のインターフェイスを用意し、さまざまな言語から容易にアクセスできるようにする。さらに考えたのは、言語バインディングそのものもできる限り書きやすくするということだった。そのためには、言語バインディングができるだけスリムにしておくことを考えた。これを実現するため、できる限りのロジックをベースとなるドライバ側に押し込めた。ドライバに組み込めなかった機能はすべての言語のバインディング側で実装する必要があり、大変な作業となってしまう。

たとえば IE ドライバでは、IE を探して起動する責務をドライバ本体のロジックに組み込めた。その結果としてドライバ側のコードの行数はおそらくほどに膨れ上がったが、新しいインスタンスを作るための言語バインディングはドライバのメソッドをひとつ呼び出すだけで済むことになった。ちなみに、Firefox ドライバの場合はこの変更に失敗した。Java の世界だけでも、三つの巨大なクラスで Firefox の設定や起動を処理していた。行数にして 1300 行ほどである。Java サーバーを立てずに FirefoxDriver をサポートしようとすると、すべての

言語のバインディングでこれらのクラスが重複することになる。つまり、同じようなコードを大量に保守しなければならないということだ。

WebDriver の設計の問題点

この方式で機能を公開したことによるマイナスは、あるインターフェイスの存在に気づくまでは WebDriver がそんな機能を持っていることに気づけないということである。これは、API の探索性を損ねてしまう。実際、WebDriver が登場したばかりのころは、利用者にそのインターフェイスの存在に気づかせるために多くの時間を費やしていた。最近ではドキュメントにより力を注ぐようになり、API も幅広く使われるようになったので、ユーザーにとっては必要な情報を見つけやすくなってきた。

自分たちの API って本当に貧弱だなあと感じる箇所をひとつ紹介しよう。RenderedWebElement というインターフェイスがあるのだが、これはいろんなメソッドがごちゃ混ぜになったものである。たとえばレンダリング後の要素の状態を取得したり (isDisplayed や getSize そして getLocation)、その要素に対して何らかの操作をしたり (hover や ドラッグ&ドロップなど)、CSS のプロパティの値を取得したりといったことができる。そもそもこのインターフェイスが作られた理由は HtmlUnit ドライバが必要な情報を公開していなかったことだが、Firefox や IE のドライバは情報を公開していた。最初は要素の状態を取得するメソッド群しか提供していなかつたが、どんどん他のメソッドが増えていった。メソッドを増やす前に、API をどのように育てていくかを熟考すべきだったと後悔している。このインターフェイスは今や広く使われているため、どのように対処するかが難しい。見苦しい API ではあるけれども広く使われているのだからそのまま維持し続けるのか、いつのこと削除してしまうのか。私は“割れ窓”を放置しておきたくなかった。そこで、Selenium 2.0 のリリースまでにこれを何とかすることが重要となった。その結果どうなったか。みなさんがこれを読むころには RenderedWebElement は削除されているはずだ。

実装側の観点で考えると、ブラウザと密に結合してしまう設計にも問題がある。たとえそれが不可避なものであったとしてもである。新しいブラウザに対応するのは大変な作業になるし、うまく動かすにはたいてい試行錯誤が必要となるだろう。具体例を挙げると、Chrome ドライバは 4 回ほどゼロから書き直したし、IE ドライバも 3 回は大幅に書き直している。ブラウザと密結合する利点は、ブラウザをより細かく制御できるという点である。

16.5 レイヤーと Javascript

ブラウザの自動化ツールは、基本的にこれらの三つの部分から構成されている。

- DOM への問い合わせ手段。
- Javascript を実行する仕組み。
- ユーザーの入力をエミュレートする何らかの手段。

このセクションで扱うのは最初の項目、つまり DOM への問い合わせの仕組みである。ブラウザの世界の共通語は Javascript であり、これを使って DOM の問い合わせができれば理想的であろう。この選択は一見明らかなようだが、Javascript を採用するにはいくつかの問題や競合する要件があり、バランスをとる必要がある。

多くの大規模プロジェクトと同様、Selenium のライブラリ群は階層構造になっている。最下層にあるのが Google の Closure Library で、これはプリミティブやモジュール化機構を提供する。これを使うと、ソースファイルを集中的に保持しつつサイズを小さくできる。その上に乗るのがユーティリティライブラリで、このライブラリではさまざまな関数を提供する。属性の値を取得したりある要素がユーザーに見えるかどうかを調べたりといった単純なタスクをこなす関数もあれば、同期イベントを使ってクリック操作をシミュレートするなどの複雑な関数もある。プロジェクト全体から見たときに、このライブラリはブラウザ自動化の最小単位の機能を提供するものに見える。そのため、このライブラリは Browser Automation Atoms あるいは atoms と呼ばれる。最後に、その上に乗るのがアダプタ層である。これが、WebDriver およびコアとの API の役割をする。

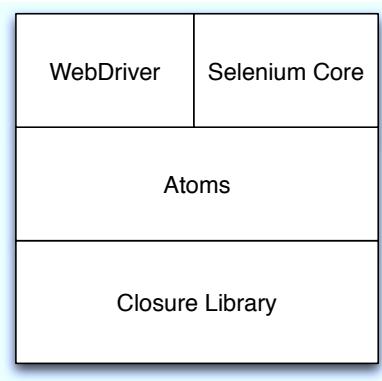


図 16.3: Selenium Javascript ライブラリの階層

Closure Library が選ばれたのにはいくつか理由がある。最大の理由は、Closure Compiler がそのライブラリの使っているモジュール化技術を理解するということだ。Closure Compiler は、出力言語として Javascript を対象としたコンパイラである。“コンパイル”とひとことで言ってもその幅は広い。單に入力ファイルを依存関係の順で並べ替えて連結・整形することもあれば、高度な最適化や未使用コードの削除まで含めることもある。Closure Library には、それ以外にもまぎれもない利点があった。Javascript のコードを開発するチームのメンバーの中に、Closure Library にとても詳しい人が何人かいたのだ。

この“核となる”ライブラリは、プロジェクト全般で DOM を扱う必要のある場面に幅広く用いられている。RC や主に Javascript で書かれたドライバの場合は、このライブラリを直接使う。ひとつのスクリプトにまとめてしまうことが多い。Java で書かれたドライバの場合は、WebDriver アダプタ層の個々の関数を完全に最適化した状態でコンパイルし、生成され

たJavascriptをJARにリソースとして組み込む。C系の言語で書かれたドライバ、たとえばiPhone用やIE用のドライバの場合は、個々の関数を最適化してコンパイルするだけではなく、生成された出力をヘッダで定義する定数に変換する。そしてそれを、ドライバの標準的なJavascript実行機能を使って必要に応じて実行する。奇妙なことに思えるかもしれないが、Javascriptを基盤のドライバに押し込むこともできる。そうすれば、複数の場所で生のソースを公開する必要がなくなる。

`atoms` は広範囲で使われているので、さまざまなブラウザで一貫したふるまいを保証できる。また、ライブラリが Javascript で書かれていて実行するのに権限の昇格が不要なので、開発サイクルを手軽にすばやく回せるようになる。Closure Library は動的に依存を読み込めるので、Selenium の開発者はただ単にテストを書いてそれをブラウザに読ませるだけでよい。コードを修正したら、必要に応じて再読み込みさせる。あるブラウザでテストが通れば、あとはそれを他のブラウザでも読ませてテストが通ることを確認するだけである。Closure Library がうまい具合にブラウザ間の差異を取り去ってくれているので通常はこれで十分だが、サポート対象の全ブラウザ用のテストスイートを実行する継続的ビルド環境もあることを知ればさらに安心できるだろう。

もともと、Core と WebDriver で同じようなことをしているコードは多かった。同じ機能を微妙に異なる方法で実現していたのだ。atoms を使った開発を始める際に、我々はコードを念入りに調べて“最適なもの”を残そうとした。結局のところ、どちらのプロジェクトについても、幅広く使われていたこともあるってコードは非常に堅牢だった。すべての投げ捨ててゼロから書き直すのは無駄だし馬鹿げていた。個々の atom をピックアップして、それを使うであろう場面を調べて atom を使うように切り替えた。たとえば Firefox ドライバの `getAttribute` メソッドは、もともと 50 行ほどあったのが空行込みで 6 行にまで縮まった。

```
FirefoxDriver.prototype.getElementAttribute =
  function(respond, parameters) {
    var element = Utils.getElementAt(parameters.id,
                                    respond.session.getDocument());
    var attributeName = parameters.name;

    respond.value = webdriver.element.getAttribute(element, attributeName);
    respond.send();
};
```

最後から 2 行目の `respond.value` に代入しているところで、WebDriver 用の `atom` を使って いる。

`atoms` は、このプロジェクトのアーキテクチャに関する主題のいくつかを現実的に実証したものである。当然、API の実装は Javascript の実装に従うという要件を強要する。さらによい点は、同じライブラリをコードベース全体で共有しているというところだ。複数のプラットフォームにまたがるバグが検出されて修正することになっても、一か所でバグを修正するだけである。変更に対するコストは下がり、安定性や効率性も向上した。`atoms` を使えば、プロジェクトのバス係数をより好ましい方向に進められる。通常の Javascript のユニット

テストツールをつかえば修正がうまく機能するかどうかを検証できるので、プロジェクトへの新規参入障壁はかなり下がるだろう。参加する前に各ドライバの実装の詳細を知っておく必要がなくなる。

`atoms` を使うメリットはそれ以外にもある。既存の RC の実装をエミュレートした（しかし裏側には WebDriver がいる）レイヤーをうまく使えば、きちんと管理された方法で新しい WebDriver の API に移行させることができる。Selenium Core は atom 化されているので、その各関数を個別にコンパイルできるようになった。そのおかげで、このエミュレートレイヤーを書く作業は容易になり、より正確に書けるようになった。

もちろん、このアプローチには弱点もあることは言うまでもない。中でも最も重大なのが、Javascript をコンパイルして C の定数にしてしまうという奇妙な作業だ。これは、C のコードでプロジェクトに貢献したいと考えている人の気持ちを挫折させてしまうに十分だ。また、すべてのブラウザの全バージョンを持っていてそのすべてで全テストを実行できるような開発者などめったにいない。誰かが不注意で予期せぬ場所にバグを作りこんでしまうこともあり得るし、もし継続的ビルドがうまく機能していないければ、そんなバグが見つかるまでにはある程度の時間を要してしまうだろう。

`atoms` はブラウザごとの返り値を正規化するので、予期せぬ値が返ってくることもあり得る。たとえば、このような HTML を考えよう。

```
<input name="example" checked>
```

`checked` 属性の `value` が何になるかはブラウザに依存する。`atoms` はこれを正規化し、その他 HTML5 仕様で定義されているすべての Boolean 属性の値を “true” か “false” に揃える。この atom をはじめてコードベースに投入したときに気付いたのは、返り値に関してブラウザ依存の値を前提としているコードがいかに多いかということだった。この値は今は一貫しているが、それまでにはコミュニティに対して「何が起きたのか」「なぜそうなったのか」を説明するための期間を要した。

16.6 リモートドライバ、中でも特に Firefox ドライバについて

リモート WebDriver は、当初から RPC の仕組みを尊重していた。それは徐々に成長して WebDriver の主要な仕組みのひとつとなり、保守コストの軽減に役立った。統一インターフェイスを提供することで、各言語のバインディングがそれを使えるようにしたのだ。ロジックのほとんどを言語バインディングからドライバ側に移してはいるが、もしドライバ側から独自のプロトコルでの通信が必要になったときは、それに対応するコードがすべての言語バインディングにまたがって用意されている。

リモート WebDriver プロトコルは、別プロセスで動いているブラウザのインスタンスとの通信が必要になったときに使うものである。このプロトコルの設計にあたっては、さまざまな内容を検討した。ほとんどは技術的なものだが、オープンソースプロジェクトであることを鑑みて、ソーシャルな面も検討材料のひとつになった。

あらゆる RPC 機構は、二つの部分に分かれる。トランスポートとエンコーディングだ。リモート WebDriver プロトコルをどのように実装したところで、クライアントとして想定するすべての言語に対してこれら両方のサポートが必要になることはわかっていた。設計の最初のイテレーションでは、Firefox ドライバについて検討した。

Mozilla は(つまり Firefox も)常に、マルチプラットフォームアプリケーションとして開発されてきた。その開発を促進するために、Mozilla は Microsoft の COM に似たフレームワークを作った。このフレームワークを使うと、コンポーネントの組み立てや連携が XPCOM(クロスプラットフォーム COM) という仕組みで可能となる。XPCOM のインターフェイスは IDL で宣言されており、C や Javascript だけでなくその他の言語のバインディングも用意されている。XPCOM は Firefox 自体を作るためにも使われており、さらに Javascript バインディングを持っているので、XPCOM オブジェクトを Firefox 拡張で使うことができる。

Win32 COM には、リモートからアクセスするためのインターフェイスが用意されている。XPCOM にも同様の機能を追加する計画があった。実際、ダーリン・フィッシャーが XPCOM ServerSocket の実装を追加してこれを実現しようとしていた。結局この D-XPCOM 計画が日の目を見ることはなかったが、その基盤の痕跡は今でも残っている。我々はこの機能を使い、基本的なサーバー機能を Firefox の拡張機能として実装した。そしてそこに、Firefox の制御のためのすべてのロジックを閉じ込めた。利用したプロトコルは、テキストベースで行指向なもので、すべての文字列を UTF-2 でエンコードしていた。個々のリクエストやレスポンスは数値から始まる。この数値は、リクエストやレスポンスに達するまでに改行文字がいくつ現れるかを指している。最も重要なのは、このスキームは Javascript で SeaMonkey(当時の Firefox が採用していた Javascript エンジン) として実装しやすかったということだ。これは、Javascript の文字列を内部的に符号なし 16 ビット整数値で格納している。

独自の符号化プロトコルを生のソケットにのせてやりとりするのは、暇つぶしとしてはいいだろう。しかし、いくつかの問題点もある。まず、自前のプロトコルに対応したライブラリは広く出回っていないので、サポートしたいすべての言語についてライブラリをゼロから自前で実装する必要がある。実装すべきコードの量もそのぶん多くなるので、新しい言語のバインディングを作つて貢献しようと考へる人たちにとっての敷居が高くなってしまう。また、また、行指向のプロトコルはテキストデータを扱う限りは便利なのだが、スクリーンショットなど画像を扱おうとし始めると問題になる。

最初に採用した RPC の仕組みは実践的ではないということが、間もなくはつきりとしてきた。幸いにも、それ以外にもよく知られたトランスポートの仕組みがあった。幅広く採用されており、ほぼすべての言語でサポートされているという、まさに我々の望み通りのもの。それが HTTP だ。

我々は HTTP をトランスポートとして採用することに決めた。次に決める必要があったのが、エンドポイントを单一にする (SOAP 風) かあるいは複数にする (REST 風) かということだった。当初の Selenium プロトコルは単一エンドポイント方式で、コマンドと引数を符号化したクエリ文字列を使っていました。この手法はうまく機能していたが、どうも“気分的に”間違っている感じがした。我々が思い描いていたのは、リモートの WebDriver インスタンスに

ブラウザ内から接続して、サーバーの状態を見られるようにすることだった。そして最終的に採用したのが、“REST的な”手法だった。複数のエンドポイント URL を用意し、HTTP メソッドを使って意味づけをする。しかし、真に RESTful なシステムに求められる制約のいくつかは満たしていない。状態を保持する場所やキャッシュなどである。その大きな理由は、アプリケーションの状態が存在する箇所はひとつだけだからである。

HTTP を採用したおかげで、符号化されたデータも（コンテンツネゴシエーションに基づく）さまざまな方法でサポートできるようになった。しかし、正式な形式をひとつ用意して、すべてのリモート WebDriver プロトコルがそれに対応できるように実装すべきだと判断した。選択肢としてすぐ候補にあがるのが、HTML や XML あるいは JSON である。XML は真っ先に候補から外した。データフォーマットとしてはよくできているライブラリのサポートもほぼすべての言語でそろっているとはいえ、経験上オープンソースのコミュニティではあまり XML が好まれないこともわかつていた。さらに、返されるデータは共通の“形式”ではあるものの、フィールドが追加されることも十分あり得た³。これらの拡張機能で XML 名前空間を使ったモデリングもできたが、そんなことをしたらクライアント側のコードが無駄に複雑化してしまう。それだけは避けたかった。そこで、XML は「オプションで使うこともできる」という扱いになった。HTML はまったくもってうまい選択だと言えないだろう。我々は自前のデータフォーマットを定義しなければならない。マイクロフォーマットで無理やりフォーマットを埋め込むこともできるが、それはまるで、卵を割るときにハンマーを使うようなものだ。

最終的に残った候補が Javascript Object Notation (JSON) だった。ブラウザ側での文字列からオブジェクトへの変換は直接 eval を呼ぶだけでいいし、最近のブラウザなら Javascript オブジェクトと文字列の相互変換を安全に副作用なしで行うプリミティブが用意されている。現実的な観点からも、JSON はよく使われているデータフォーマットであり、ほとんどすべての言語で JSON 処理用のライブラリが用意されている。また、開発者にも人気が高い。無難な選択肢と言える。

リモート WebDriver プロトコルの第 2 イテレーションでは、HTTP をトранスポートとして採用し、UTF-8 でエンコードした JSON をデフォルトの符号化スキームとした。UTF-8 はデフォルトの符号化方式として選ばれたものであり、Unicode のサポートがあまり充実していない言語でクライアントを書くのも容易である。というのも、UTF-8 は ASCII と後方互換性があるからだ。サーバーに送信するコマンドは、URL を使ってどのコマンドが送信されたのかを判断し、コマンドへのパラメータは配列形式でエンコードする。

たとえば WebDriver.get("http://www.example.com") の呼び出しは、セッション ID をエンコードして最後に “/url” をつけた URL への POST リクエストにマップされる。このとき、パラメータの配列は [’http://www.example.com’] のようになる。返される結果はもう少し構造化されており、返り値やエラーコード用のプレースホルダーが用意されている。この形式は、リモートプロトコルの第 3 イテレーションまでしか続かなかった。リクエストにおける

³たとえば、リモートサーバーは base64 でエンコードしたスクリーンショットに加えて発生した例外もすべて返し、デバッグの助けにしている。しかし Firefox ドライバはそれを返さない。

るパラメータの配列は、そのイテレーションで名前つきパラメータの辞書に変わった。この変更によって、デバッグ用のリクエストがとても簡単に実行できるようになった。また、クライアントがパラメータの順番を間違えてしまう可能性をなくし、システム全体としてより堅牢になった。必然的に、通常の HTTP ステータスコードを使って特定の返り値や応答を表すようになった。それが最も適切な方法だったからである。たとえば、どこにもマップされていない URL を呼ぼうとしたときや “空のレスポンス” を表したいときなどに使える。

リモート WebDriver プロトコルには二段階のエラー処理がある。無効なリクエストを扱うものと、コマンドが失敗した場合を扱うものである。無効なリクエストの例としては、サーバー上に存在しないリソースへのリクエストあるいはそのリソースが処理できないメソッドでのリクエスト(たとえば、現在のページの URL を指すリソースに対する DELETE コマンド)などがある。このような場合は、通常の HTTP 4xx レスポンスが送出される。コマンドが失敗した場合には、レスポンスのエラーコードは 500 (“Internal Server Error”) となり、返すデータの中により詳しい情報を含めて何が悪かったのかをわかりやすくする。

データを含むレスポンスがサーバーから返されるときには、JSON オブジェクト形式となる。

キー	説明
sessionId	不透過なハンドル。サーバーがセッション固有のコマンドの送り先を決めるために使う。
status	コマンドの結果を表す数値のステータスコード。ゼロ以外の値は、コマンドが失敗したことを表す。
value	レスポンスの JSON データ。

レスポンスは、たとえばこのようになる。

```
{  
  sessionId: 'BD204170-1A52-49C2-A6F8-872D127E7AE8',  
  status: 7,  
  value: 'Unable to locate element with id: foo'  
}
```

見てわかるとおり、ステータスコードをレスポンス内で符号化しており、ゼロではない値が入っていることから何かがうまくいかなかったことがわかる。IE ドライバはまず最初にステータスコードを使い、プロトコル内で使う値はこの値をミラーしている。すべてのエラーコードは各ドライバで共通なので、エラー処理のコードは特定の言語で書いてすべてのドライバで共有できる。これにより、クライアント側の実装がより容易になる。

Remote WebDriver Server は単なる Java サーブレットである。これはマルチプレクサとして動作し、受け付けたコマンドを適切な WebDriver インスタンスに振り向ける。まあ大学院の二年目くらいでも書けるレベルのものだ。Firefox ドライバでもリモート WebDriver プロトコルを実装しており、そのアーキテクチャのほうがずっと興味深い。そこで、言語バインディングから受け取ったリクエストがバックエンドに到達してからユーザーに応答を返すまでの流れを追いかけてみよう。

ここでは Java を使っているものとする。また “element” が WebElement のインスタンスである。すべてはこの行からはじまる。

```
element.getAttribute("row");
```

内部的に、element は不透過な “id” を保持している。サーバーサイドではこれを使い、対話相手の要素を識別する。ここでは仮に、id の値が “some_opaque_id” であるものとして話を進める。これは Java の Command オブジェクトに Map として符号化されており、(名前付きの) パラメータ id に要素の ID、そして name に問い合わせ対象の属性の名前を保持する。

正しい URL を指すテーブル内での検索は、このようになる。

```
/session/:sessionId/element/:id/attribute/:name
```

URL のセクションの中でコロンから始まるものはすべて、変数であって後で何かの値で置き換えるものである。パラメータ id と name は既に指定している。sessionId はもうひとつの不透過なハンドルで、ルーティングのために使う。これを使えば、サーバーが複数のセッションを同時に扱えるようになる (Firefox ドライバではこれができない)。この URL を展開したものは、たとえばこのようになる。

```
http://localhost:7055/hub/session/XXX/element/some_opaque_id/attribute/row
```

余談だが、WebDriver のリモートワイヤプロトコルの開発が始まったのは、URL Templates が RFC 草案として提案されたのとほぼ同時期だった。我々が考えた URL の指定方法も URL Templates も、どちらも URL 内での変数の展開 (そして派生) を許していた。残念なことに、URL Templates が提案されていることを我々が知ったのはかなり後になってからのことだった。そのため、ワイヤプロトコルの記述に URL Templates を使うことができなかつた。

我々の実装するメソッドは幕等⁴なので、ここで使うべき正しい HTTP メソッドは GET である。このあたりの処理は、HTTP を話せる Java ライブラリ (Apache HTTP Client) に委譲して、そのライブラリにサーバーを呼ばせる。

Firefox ドライバは、Firefox の拡張機能として実装されている。その基本的な設計は図 16.4 のとおりだ。多少風変わりなところがあるとすれば、それは HTTP サーバーを組み込んでいくところだ。元々は自前でこれを実装していたのだが、HTTP サーバーを XPCOM で書くというのは我々の得意分野ではない。そこで、Mozilla 自身が用意している基本的な HTTPD でそれを置き換えた。リクエストはこの HTTPD が受け取り、ほぼそのままの形で dispatcher オブジェクトに渡される。

ディスパッチャは、リクエストを受け取ってからサポートする既知の URL リストを順にたどり、リクエストにマッチする URL を探す。このマッチングは、クライアント側で行われた変数の置換に基づいて行われる。リクエストメソッドも含めて完全に一致するものが見つかれば、実行するコマンドを表す JSON オブジェクトを組み立てる。今回の場合は、このようなオブジェクトになる。

⁴すなわち、何度実行しても同じ値を返す。

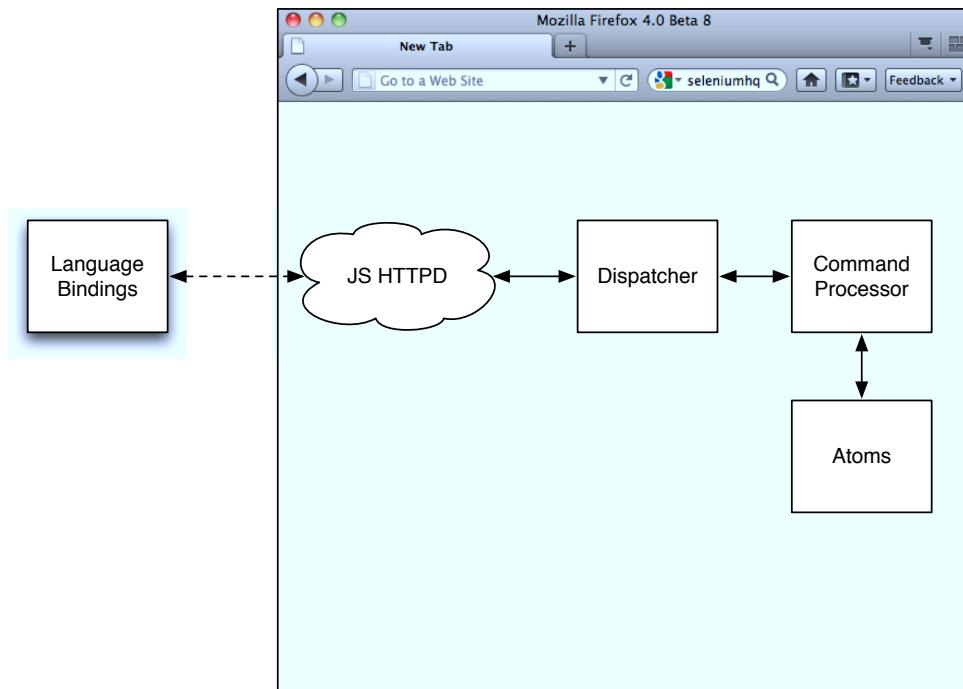


図 16.4: Firefox ドライバのアーキテクチャ概要

```
{
  'name': 'getElementAttribute',
  'sessionId': { 'value': 'XXX' },
  'parameters': {
    'id': 'some_opaque_key',
    'name': 'rows'
  }
}
```

次にこれが、JSON 文字列として我々の書いた XPCOM コンポーネントに渡される。このコンポーネントは CommandProcessor と呼ばれている。このようなコードだ。

```
var jsonResponseString = JSON.stringify(json);
var callback = function(jsonResponseString) {
  var jsonResponse = JSON.parse(jsonResponseString);

  if (jsonResponse.status != ErrorCode.SUCCESS) {
    response.setStatus(Response.INTERNAL_ERROR);
  }

  response.setContentType('application/json');
  response.setBody(jsonResponseString);
```

```

        response.commit();
    };

    // コマンドをディスパッチする
    Components.classes['@googlecode.com/webdriver/command-processor;1'].
        getService(Components.interfaces.nsICommandProcessor).
        execute(jsonString, callback);

```

やたら大量のコードがあるが、ポイントとなるのは次のふたつだ。まず、上のオブジェクトを JSON 文字列に変換する。次に、コールバックを渡してメソッドを実行させる。これが、送出する HTTP レスポンスを作る。

コマンドプロセッサで実行するメソッドは、“name”を見てどの関数を呼ぶかを判断し、そしてそれを実行する。この実装関数に渡す最初のパラメータは respond オブジェクトで(このように呼ばれているのは、もともとこの関数は単にユーザーにレスポンスを返すためだけのものだったからである)、これは返す値をカプセル化するだけではなく、レスポンスをユーザーに送り返せるようにするメソッドや DOM の情報を見つけるための仕組みも用意されている。二番目のパラメータは、先述の parameters オブジェクトの値(この場合は id と name)である。この方式のメリットは、各関数が統一インターフェイスを持っていて、それがクライアント側で使うデータ構造を反映しているということだ。つまり、コードを書くときに頭の中で考えるモデルがクライアント側でもサーバー側でも同じようになるのだ。ここにgetAttribute の実装を示す。これは先ほど 16.5 節で見たものである。

```

FirefoxDriver.prototype.getAttribute = function(respond, parameters) {
    var element = Utils.getElementAt(parameters.id,
                                    respond.session.getDocument());
    var attributeName = parameters.name;

    respond.value = webdriver.element.getAttribute(element, attributeName);
    respond.send();
};

```

要素の参照に矛盾を生じさせないようにするために、最初の行は単にキャッシュ内の不透過な ID が参照する要素を探すだけになっている。Firefox ドライバの場合、この不透過な ID は UUID であり、“キャッシュ”は単なるマップである。getElementAt メソッドは、要素への参照が既知のものであるかどうかとそれが DOM にアタッチされているかどうかをチェックする。どちらかのチェックに失敗すると、その ID は(必要に応じて)キャッシュから削除され、例外を投げてそれをユーザーに返す。

最後から二行目では、先述のブラウザ自動化用 atom を使っている。ここでは一つのスクリプトとしてコンパイルされ、拡張機能の一部として読み込まれている。

最後の行で呼ばれているのが send メソッドだ。このメソッドはシンプルなチェックを行い、execute メソッドで指定したコールバックを呼んでからレスポンスを送出する。レスポンスは JSON 文字列形式でユーザーに戻され、それがこのような形式のオブジェクトに移される(getAttribute の返り値が “7”、つまりその要素が見つからなかったものと仮定する)。

```
{  
  'value': '7',  
  'status': 0,  
  'sessionId': 'XXX'  
}
```

その後、Java クライアントが status フィールドの値をチェックする。もし値がゼロでなければ、数値のステータスコードを適切な型の例外オブジェクトに変換して投げる。その際に、“value” フィールドの値を使ってユーザー向けのメッセージを設定する。status がゼロの場合は、“value” フィールドの値をユーザーに返す。

ほとんどは、ごく自然な処理だろう。しかし一ヵ所だけ、賢明な読者なら疑問に思うところがあるはずだ。なぜディスピッチャは、execute メソッドを呼ぶ前にわざわざオブジェクトを文字列に変換するのだろう？

なぜそうしているかというと、Firefox ドライバは Javascript だけで書かれたテストの実行もサポートしているからである。普通は、これをサポートするのはかなり難しい。テストが実行されるのはブラウザの Javascript セキュリティサンドボックスの中であり、テストで有用な多くの作業（別ドメインへの移動やファイルのアップロードなど）に制限が出てしまうからである。しかし Firefox の WebDriver 拡張機能では、サンドボックスから脱出するためのハッチを提供している。document 要素に webdriver というプロパティを追加しているのだ。WebDriver の Javascript API はこれを見て、次のような操作ができると判断する。JSON でシリализしたコマンドオブジェクトを document 要素の command プロパティに追加して webdriverCommand イベントを発火させ、何らかの要素での webdriverResponse イベントの発生を監視する。このイベントが、response プロパティの値が設定されたことを示す。

これはつまり、WebDriver 拡張機能をインストールした Firefox でウェブをブラウズするの非常に危険だということだ。悪意のある人なら、リモートから容易にブラウザを乗っ取れてしまうからである。

その裏側では DOM メッセンジャーが動いていて webdriverCommand を待機している。シリализした JSON オブジェクトをこれが読み込み、コマンドプロセッサの execute メソッドを呼び出す。このとき、コールバックは単に document 要素の response 属性に設定されるだけである。それが、期待する webdriverResponse イベントを発火させる。

16.7 IE ドライバ

Internet Explorer は興味深いブラウザである。さまざまな COM インターフェイスが一齊に動く構造になっている。Javascript エンジンもまったく同じで、なじみのある Javascript の変数も実際はその背後にある COM インターフェイスへの参照となっている。たとえば Javascript の window の実体は IHTMLWindow である。つまり、document は COM インターフェイス IHTMLDocument のインスタンスとなる。Microsoft は、既存のふるまいを維持しつつブラ

ウザに機能を追加するというすばらしい仕事をやってのけた。つまり、IE6 が公開する COM クラスを使ったアプリケーションは、そのまま IE9 でも動くということである。

Internet Explorer ドライバのアーキテクチャは、時を経て成長してきた。その設計の根底にある大きな目標は、インストーラーを使わないということだ。あまり聞き慣れない要件だと思うので、ここで少し補足しておこう。インストーラーを不要にしたい第一の理由は、WebDriver が“5 分でテストできる”という目標を満たせなくなるからだ。開発者がパッケージをダウンロードしてインストーラーを実行するまでに、ある程度の時間を要してしまう。さらに重要なのは、WebDriver のユーザーの中には自分のマシンにソフトウェアをインストールする権限を持っていない人も比較的多いということである。インストーラーをなくせば、あるプロジェクトで IE を使ったテストを始めるときに、継続的インテグレーションサーバーにログオンしてインストーラーを実行するという必要もなくなる。最後に、インストーラーを実行するという習慣を持たない言語もあるということだ。たとえば Java なら、通常は CLASSPATH の通った場所に JAR を置くだけだ。経験上、インストーラーがついているライブラリはあまり好まれず、使われることもない。

だから、インストーラーはやめた。その結果、次のようにになった。

Windows 上でのプログラミングに使う言語として自然な選択は、.Net 上で動くもの(おそらく C#)になるだろう。IE ドライバは IE の COM オートメーションインターフェイスを使っており、IE 自体と密に結合している。IE の COM オートメーションインターフェイスは、すべてのバージョンの Windows に組み込まれている。特に使っているのはネイティブの MSHTML および ShDocVw DLL であり、これらは IE の一部である。C# 4 より前のバージョンでは、CLR/COM の相互運用性は Primary Interop Assembly (PIA) を分離することで実現していた。PIA は本質的に、CLR が管理する世界と COM が管理する世界の橋渡しをするために作られたものである。

残念なことに、C# 4 を使おうとすると .Net ランタイムの新しいバージョンを使うことになってしまう。多くの企業では最先端を使うのを避け、問題は残っているが安定している旧リリースを使うことを好む。C# 4 を使ってしまうと、無視できない割合のユーザー層を排除してしまうことになる。PIA を使うことにはそれ以外のデメリットもある。ライセンスの制約を考えてみよう。Microsoft に問い合わせた結果明らかになつたことは、Selenium プロジェクトには MSHTML や ShDocVw の PIA を配布する権限がないということだった。仮に配布する権利が認められたとしても、世間にインストールされている Windows と IE のライブラリには無数の組み合わせがある。それらすべてに対応する PIA を配布しなければならないということだ。クライアントマシン上で、その場で PIA を作るという方法もあるが、それは使えない。開発者向けのツールが必要になるし、通常のユーザーのマシン上にはそんなものはインストールされていないからだ。

したがって、大規模なコードを書くには C# はとても魅力的な言語だが、今回の選択肢からは消えた。我々が必要としたのは何かネイティブな言語で、少なくとも IE との通信ができるものだった。この目的にかなう次の選択肢は C++ であり、最終的に我々が選んだのもこの言語だった。C++ を使えば PIA が不要になるというメリットがあるが、それと同時に、Visual

Studio C++ランタイム DLL を再配布するかあるいは静的にリンクするかの選択を迫られることになってしまう。DLL を配布するにはインストーラーが必要になってしまうので、IE と通信するライブラリを静的にリンクすることにした。

インストーラーを使わないという要件を満たすために、かなりのコストがかかっている。しかし、当初のテーマであった「複雑性をどこに押し込めるか」を考えると、ユーザーにとって使いやすくするための投資としては十分に意味のあるものだ。C++を採用するという決断については、現時点で改めて評価しなおしているところである。というのも、ユーザーにとっての使いやすさと引き替えに、協力してくれる開発者の母数が減ってしまうという事実があるからだ。高度な C++で書かれたオープンソースプロジェクトに貢献してくれる可能性のある開発者の数は、C#のプロジェクトに比べて大幅に少ないように感じる。

IE ドライバの当初の設計を図 16.5 に示す。

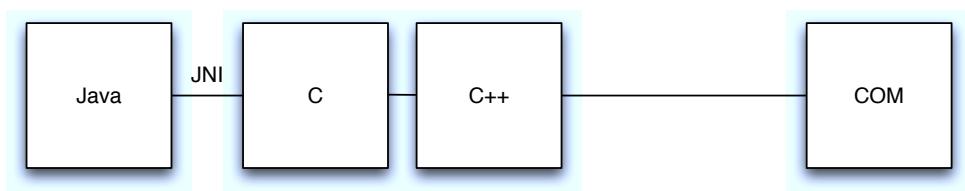


図 16.5: 当初の IE ドライバ

スタックの最下層から話を始めよう。ここでは、IE の COM オートメーションインターフェイスを使っていることがわかる。概念的な意味での使いやすさを向上させるために、この生のインターフェイスを C++のクラス群でラッパーした。このラッパーは、WebDriver API の構造に似せてある。Java のクラスに C++との通信をさせるために JNI を使い、JNI のメソッドで COM インターフェイスの C++ラッパーと通信している。

このアプローチは、クライアント側の言語が Java だけに限られる場合はうまくいく。しかし、サポートすることになったすべての言語に対してバックにあるライブラリを変更するのは大変だし、複雑になってしまう。したがって、JNI だけでもうまくいくとは言え、これだけではまだ適切な抽象化が実現できていない。

適切な抽象化とは何だったのか? 我々がサポートしようとしていたすべての言語には、C のコードを直接呼び出す仕組みがあった。C#の場合は PInvoke 形式がそれにあたる。Ruby には FFI があるし、Python には ctypes がある。Java の世界には、すばらしいライブラリである JNA (Java Native Architecture) が存在する。我々が必要としていたのは、これらの共通項を使う API を公開することだった。これを実現するためにオブジェクトモデルを平坦化し、シンプルな 2 文字か 3 文字のプレフィックスを使って各メソッドの“ホームインターフェイス”を示した。たとえば “wd” は “WebDriver” を表し、“wde” は WebDriver Element を表す。つまり WebDriver.get は wdGet となり、WebElement.getText は wdeGetText となった。各メソッドはステータスコードを表す整数値を返し、同時に “out” パラメータを使ってそれ以外のデー

タも返せるようにした。最終的に、メソッドのシグネチャはこのようになる。

```
int wdeGetAttribute(WebDriver*, WebElement*, const wchar_t*, StringWrapper**)
```

呼び出す側のコードでは、WebDriver や WebElement そして StringWrapper は不透過型となる。API ではこれらの違いを表し、どの値をどのパラメータで使うのかを明確にしている。しかし、単に “void *” とすることもできる。また、テキストにはワイド文字を使っていることもわかるだろう。これは、国際化したテキストを適切に扱えるようにするためにある。

Java 側では、この関数のライブラリをインターフェイス経由で公開している。それを使って WebDriver が用意する通常のオブジェクト指向のインターフェイスと同じようにすることができる。たとえば、Java でのgetAttribute メソッドの定義はこのようになる。

```
public String getAttribute(String name) {  
    PointerByReference wrapper = new PointerByReference();  
    int result = lib.wdeGetAttribute(  
        parent.getDriverPointer(), element, new WString(name), wrapper);  
  
    errors.verifyErrorCode(result, "get attribute of");  
  
    return wrapper.getValue() == null ? null : new StringWrapper(lib, wrapper).toString();  
}
```

これが、図 16.6 で示す設計につながる。

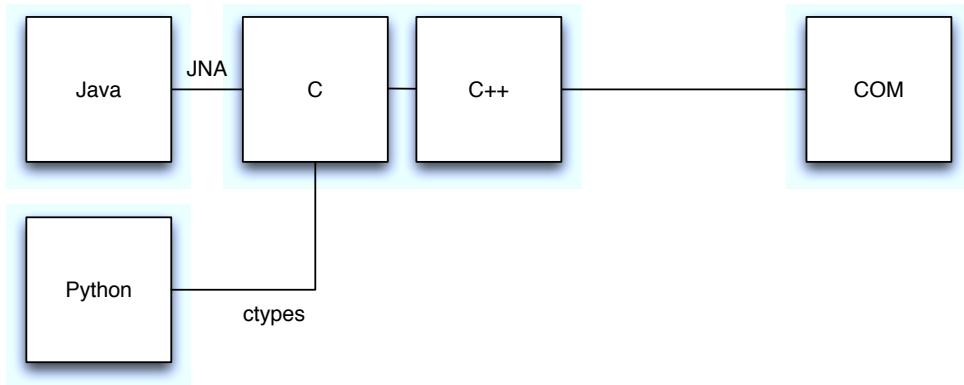


図 16.6: 手を加えられた IE ドライバ

すべてのテストをローカルマシンで動かしている間はこれでうまくいく。しかし、IE ドライバをリモートの WebDriver で使い始めると、ランダムなロックが発生してしまうようになった。原因を追跡していくと、最後は IE の COM オートメーションインターフェイスの制約に行き着いた。このインターフェイスは、“Single Thread Apartment” モデルで使うように作られていたのだ。これは本質的に、インターフェイスを毎回同じスレッドから呼び出す必

要があるということである。ローカルで動かしている場合は、デフォルトでこの挙動となる。しかし Java アプリケーションサーバーは、負荷に対応するために複数スレッドをまとめて使う。で、どうなったって？ どんな場合にも同じスレッドを使って IE ドライバにアクセスさせる方法なんか見つからなかった。

この問題に対するひとつの解決策が、IE ドライバをシングルスレッドのエグゼキュータ内で実行してアプリケーションサーバー内の Future 経由のアクセスをすべてシリアル化することで、しばらくの間は我々もその方法をとっていた。しかしこれは、複雑性を呼び出し側のコードに残してしまうという点でアンフェアであり、不注意で IE ドライバを複数のスレッドから呼んでしまうという自体が容易に想像できた。そこで、この複雑性はドライバ自身に押し込めるに決めた。IE のインスタンスを個別のスレッドに持たせ、Win32 API の PostThreadMessage を使ってスレッド越しの通信を行った。というわけで、執筆時点での IE ドライバの設計は図 16.7 のようになっている。

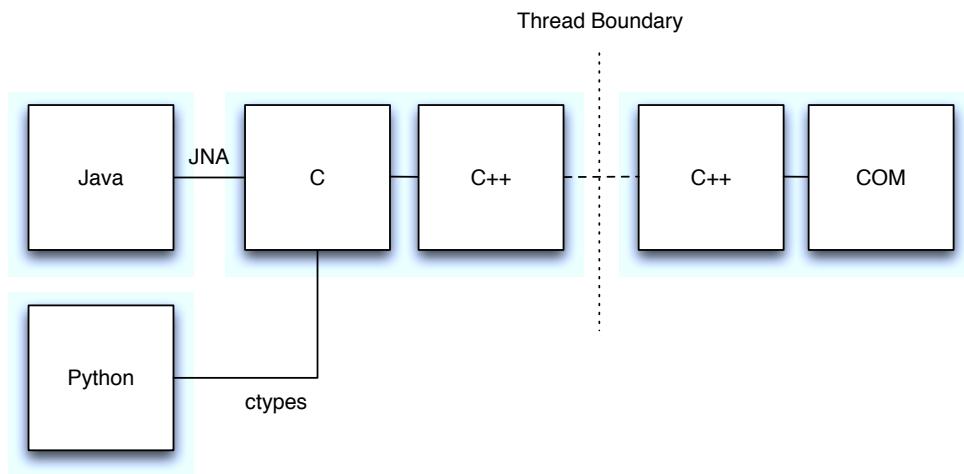


図 16.7: Selenium 2.0 alpha 7 の時点の IE ドライバ

これは自ら進んで選んだ設計ではない。しかしうまく動いており、ユーザーに不便な思いをさせずに済んでいる。

この方式の問題のひとつが、IE のインスタンスがロックされているかどうかを自分で判断しづらいことだ。これが問題となるのは、DOM を操作している際にモーダルダイアログが開いたりスレッド境界のはるか向こう側で致命的な障害が発生したりした場合である。その対策として、送信するすべてのスレッドメッセージにはタイムアウトを設定しており、その値は太っ腹にも 2 分となっている。メーリングリストでのユーザーの声を見る限り、この値はほぼ適切なようだ。しかし、常にこれが正解というわけではないので、IE ドライバの今後のバージョンではタイムアウトの値を変更可能にする予定だ。

もうひとつの問題は、内部的なデバッグが非常に難しくなるということだ。まずスピード

を問われる(要するに、2分以内にコードを追いかけきらないといけない)し、適切なブレークポイントを設定して、スレッドをまたがるコードの流れをきちんと追いかけなければならない。言うまでもないが、オープンソースの世界には興味をそそられる課題が満載だ。こんな汚れ仕事を進んでやろうとする人はまずいないだろう。そのおかげでシステムのバス係数が劇的に低下してしまう。プロジェクトのメンテナとして、これは気がかりなことだ。

その対策として、IE ドライバをどんどん Automation Atoms 化している。Firefox ドライバや Selenium Core と同様に、だ。使おうとしている atom をコンパイルして C++ のヘッダファイルを作り、個々の関数を定数として公開する。実行時には、Javascript を使ってこれらの定数を実行する。このようにすると、IE ドライバのコードの大部分は C コンパイラがなくても開発・テストできるようになる。そのぶん、バグを見つけたり修正してくれたりする人たちにとっても敷居が低くなるだろう。最終的な目標は、インターフェイス API だけをネイティブコードのまま残し、それ以外はできる限り atom に任せることだ。

もうひとつのアプローチとして検討中なのが、IE ドライバを書き直して軽量 HTTP サーバーを使うようにすることだ。そうすれば、IE ドライバをリモート WebDriver として扱えるようになる。もしこれが実現すれば、スレッドの境界にまつわる複雑性の多くを取り除け、必要なコードの量も減らせるうえに、処理の流れも非常に追いややすくなる。

16.8 Selenium RC

特定のブラウザとの密な結合が常に可能だとは限らない。そんな場合、WebDriver は次善の策としてもともと Selenium が使っていた仕組みを利用する。つまり Selenium Core を使うということだ。これはピュア Javascript フレームワークであり、Javascript サンドボックスのコンテキストで動くこともあっていろいろ制約も多い。WebDriver の API を使う側から見ると、サポート対象とされているブラウザの中にもいくつかのレベルがあるということだ。いくつかのブラウザは密に統合されているので期待通りの制御ができるが、そうでないブラウザは Javascript によるサポートしかなく、もともとの Selenium RC が持っていた機能と同程度の制御しかできないということになる。

概念的には、ここで採用する設計は図 16.8 のように極めてシンプルなものである。

ご覧の通り、全体が大きく三つに分かれている。クライアントのコード、中間サーバー、そしてブラウザ内で動作する Selenium Core の Javascript コードだ。クライアント側は単なる HTTP クライアントであり、シリアル化したコマンドをサーバー側に送る。リモート WebDriver とは違って単にエンドポイントがひとつあるだけであり、どの HTTP メソッドを使うかはあまり関係ない。その理由のひとつは、Selenium RC のプロトコルが Selenium Core のテーブルベースな API を元にしていることだ。そのため、URL のクエリパラメータを三つ使うだけで API 全体をカバーできてしまう。

クライアントが新しいセッションを開始すると、Selenium Server はリクエストされた“ブラウザ文字列”に対応するブラウザランチャーを探す。そしてランチャーが、指定されたブ

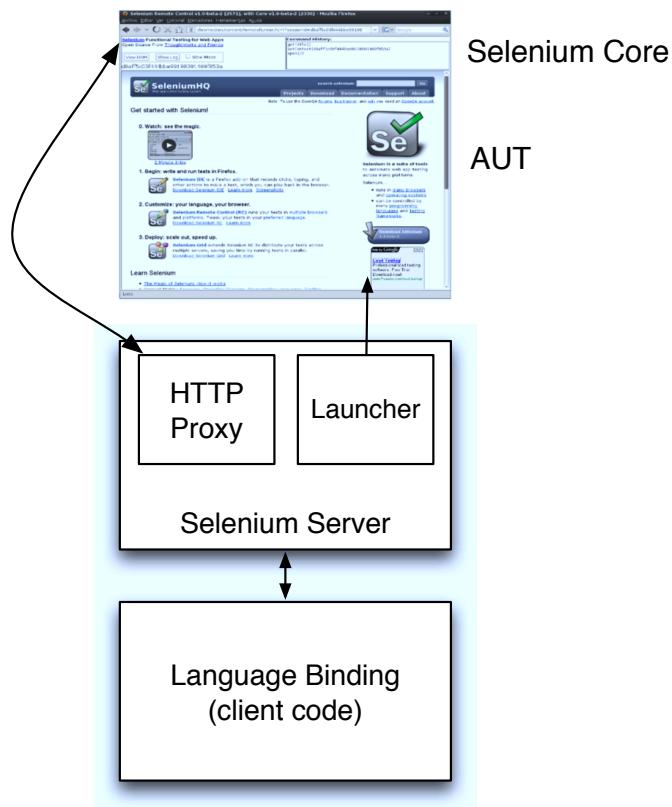


図 16.8: Selenium RC のアーキテクチャの概要

ラウザの構成を設定してブラウザのインスタンスを立ち上げる。Firefox の場合は、事前に用意したプロファイルやプレインストールの拡張機能 (“quit” コマンドを処理するための拡張や、“document.readyState” に対応する拡張など。“document.readyState” は、Selenium がサポートする Firefox のバージョンの中でも古いものには存在しない) を展開するだけのことだ。ここで行われる設定の中でもポイントとなるのが、Selenium Server が自分自身をブラウザのプロキシに設定することだ。これにより、少なくとも一部のリクエスト (“/selenium-server” に向けたもの) を Selenium Server でルーティングできるようになる。Selenium RC は、次の三つのモードのうちのいずれかで動作する。一つのウィンドウ内のフレームを制御するモード (“singlewindow” モード)、別のウィンドウを開いて AUT をそのウィンドウで制御するモード (“multiwindow” モード)、そして自分自身をプロキシ経由でページに差し込むモード (“proxyinjection” モード) だ。処理モードによっては、すべてのリクエストがプロキシ経由となることがある。

ブラウザの設定を終えたらブラウザを立ち上げる。そのときの初期 URL は、Selenium Server が稼働するページ—RemoteRunner.html である。このページがすべての初期化処理を受け持

ち、Selenium Core に必要な Javascript ファイル群をすべてここで読み込む。読み込みが完了すると、“runSeleniumTest” 関数が呼ばれる。この関数は、Selenium オブジェクトのリフレクションを使って利用可能なコマンド群を初期化をしてから、メインのコマンド処理ループを起動する。

ブラウザ内で動く Javascript が XMLHttpRequest を待ち受けサーバーの URL(/selenium-server/driver) に送る。このサーバーがすべてのリクエストのプロキシとなり、リクエストが正しい場所に届くことを保証する。リクエストを送るだけではなく、その前に前回実行したコマンドのレスポンスを送信したり、ブラウザが立ち上がったときに “OK” を送ったりもする。サーバーはその後リクエストをオープンし続け、クライアントからユーザーのテストのコマンドを受け取ったら、オープンしていたリクエストのレスポンスを Javascript に返す。この方式は俗に “Response/Request” と呼ばれていたが、最近では “Comet with AJAX long polling” と呼ばれることが多い。

なぜ RC の挙動がそうなっているのかって? Server をプロキシとして設定しなければならない理由は、Javascript の “同一生成元” ポリシーに違反せずにすべてのリクエストを横取りするためだ。“同一生成元” ポリシーとは、Javascript ではスクリプトがある場所と同じサーバー上のリソースしかリクエストできないという制限のことである。そもそもこれはセキュリティを考慮して用意されたポリシーだが、ブラウザ自動化フレームワークの開発者から見ると極めて邪魔なものであり、こうでもしないとどうにもならないのだ。

XMLHttpRequest コールをサーバーに向けて行う理由は次のふたつだ。まずは最も重要な理由から。HTML5 の一部である WebSockets が大半のブラウザに実装されるようになるまでは、サーバープロセスをブラウザ側から立ち上げるための信頼できる方法が存在しないということである。つまり、サーバーはどこか別の場所で立ち上げないといけないということだ。もうひとつの理由は、XMLHttpRequest がレスポンスコールバックを非同期で呼び出すことだ。これはつまり、次のコマンドを待ち受けている間にも、操作中のブラウザの動きには何の影響も及ぼさないと言うことである。次のコマンドを待ち受ける方法は、それ以外にも二通り存在する。ひとつは、定期的にサーバーをポーリングして、何か別のコマンドが実行されていないかを調べる方法だ。しかしこの方法だと、ユーザーのテストにある程度の遅延が生じてしまう。もうひとつの方法は、Javascript をビギループの中に置いてしまうことだ。しかしこの方法は CPU を食いつぶしてしまい、他の Javascript をブラウザ内から実行できなくなってしまう（ひとつのウィンドウ内のコンテキストには、Javascript のスレッドをひとつしか実行できないからである）。

Selenium Core の中は、大きく二つの部分に分かれている。本体の selenium オブジェクトは利用可能なすべてのコマンドのホストとして機能し、その API をユーザー向けに提供する。もうひとつの部分が browserbot だ。これは、Selenium オブジェクトが各ブラウザの差異を取り扱うための抽象化として使うもので、一般的なブラウザの機能を理想的な状態で表している。これによって selenium の関数はよりきれいで保守しやすくなり、一方で browserbot にすべてを集中させることになる。

Selenium Core も、徐々に Automation Atoms を使うように書き換えているところだ。selenium

と browserbot はおそらく残さざるを得ないだろう。大量のコードがこれらの公開する API に依存しているからである。しかし、最終的にはこれらも単なるガワだけのクラスにしてしまって、内部の実装はすぐにでも atoms に委譲させてしまいたいところだ。

16.9 過去を振り返って

ブラウザの自動化フレームワークを作るという作業は、部屋の塗装に似ている。始める前にはとても簡単な作業に思えるのだ。「要するに、ざっと塗ってしまえばそれでおしまいでしょ?」実際はそうではなく、作業をすればするほどやるべきことが増えてきて、ひとつひとつのタスクが面倒なものになっていく。部屋の塗装の例だと、照明周りや幅木などを手掛け始めるととたんに時間がかかるようになる。ブラウザの自動化フレームワークでこれにあたるのが、ブラウザごとの機能の違いや微妙な動きだ。そのおかげで作業はさらに複雑になる。極めつけが、私の隣の席で Chrome ドライバを開発するダニエル・ワグナー=ホールだ。あまりにもイライラした彼は、デスクに両手をたたきつけてこう叫んだ。「なんでこんなあり得ないことばかり起こるんだよ!」できることなら過去にさかのぼって当時の自分にこう伝えてやりたい。「そのプロジェクト、たぶん思ってるよりもずっと時間がかかるよ」と。

今さらどうにもならないことではあるが、もし Automation Atom のようなレイヤーの必要性をもっと早くに認識して対応していたらどうなっただろうかと考えることもある。きっと、これまでに我々が直面してきた問題(内部的な問題も外部的な問題も、そして技術的な問題も社会的な問題も)のいくつかは、もっと容易に対処できたことだろう。Core や RC は、特定の言語—基本的に Javascript と Java—に注力しすぎた実装になっている。ジェイソン・ハギンズはかつて、これを指摘して Selenium の“ハッカビリティ”を改善し、そのおかげでプロジェクトに参加するための敷居が下がった。atom があったからこそ、WebDriver が幅広いハッカビリティを確保できたのだ。我々が atom を幅広く適用できたのは、Closure コンパイラのおかげである。Closure コンパイラがオープンソースでリリースされるとすぐに、それを採用した。

「こうすればよかった」だけではなく、「これはうまくできた」ということを振り返るのもいいだろう。フレームワークを書くときにユーザーの視点を重視したという判断は、今でも間違いではなかったと思っている。初期の段階では、改善すべき点をアーリーアダプターが指摘してくれたおかげで、ツールの使いやすさを急速に改善することができた。後に、WebDriver がより高度な作業を行うようになるにつれて、利用する開発者の数も増えてきた。新たな API を追加するときには今まで以上に注意を払うようになり、それがプロジェクトを引き締めることにもつながった。我々がやろうとしていることを考えると、これは非常に大切なことだ。

ブラウザと密に結合させたことには、功罪の両面がある。よい面は、忠実にユーザーをエミュレートしてブラウザを完璧に制御できるようになったということだ。逆に悪い面は、この手法をとると技術的な要求が厳しくなるということだ。特に、ブラウザのフックポイントを見つけ出すのが大変になる。その大変さは、IE ドライバの開発の進み具合を見ればよくわ

かる。ここでは紹介できなかったが Chrome ドライバも同様で、これもまた話せば長くなる経緯がある。いつの日か、この複雑性とうまく向き合う方法を見つけ出したいものだ。

16.10 今後に向けて

WebDriver との密な結合ができないブラウザってのが、常に存在する。なので、Selenium Core は今後もいつだって必要になるだろう。当初から続く現在の設計を変更して、atoms も使っている Closure Library と同じものを使うように変更しようという動きが進行中である。また、既存の WebDriver の実装にも、atoms をより深く埋め込んでいこうという動きもある。

WebDriver の当初の目標のひとつが、他の API やツールを組み立てるためのブロックとして使えるようにすることだった。もちろん、Selenium が唯一のブラウザ自動化ツールというわけではない。それ以外にも、オープンソースのブラウザ自動化ツールはいくらでも存在する。そのひとつが Watir (Web Application Testing In Ruby) であり、Selenium と Watir の開発者が協力して、Watir API を WebDriver のコアにのせようという動きも始まっている。我々は、他のプロジェクトとの協力を望んでいる。すべてのブラウザに対応し続けるという作業はたいへんなものだからである。しっかりとした中核があつて、他の開発者たちがその上に何かを構築しやすいようにできれば、すばらしいことだろう。我々としては、その中核が WebDriver となって欲しい。

そんな将来をうかがわせるような申し出が Opera Software からあった。彼らは独自に WebDriver API を実装し、WebDriver のテストスイートを使ってそのコードのふるまいを検証し、そして OperaDriver としてリリースする。Selenium チームのメンバーは Chromium チームとも共同作業をしており、WebDriver のサポートやよりよいフックを Chromium に追加しようとしている。Chrome に対しても同様だ。Mozilla とも良好な関係を築いている。彼らは Firefox-Driver のコードに貢献してくれたし、あの有名な Java ブラウザエミュレータである HtmlUnit の開発者も提供してくれた。

ひとつの見方として、この傾向は今後も続くだろう。さまざまなブラウザで、統一された方法で自動化のフックができるようになるという流れだ。ウェブアプリケーションのテストを書く開発者にとってのメリットは明白だし、ブラウザを作る側にとってもその利点は明らかだ。たとえば、手動テストのコストと比較して、多くの大規模プロジェクトでは自動テストにより依存している。もし特定のブラウザでのテストが不可能 (あるいはばかみたいに高いコストがかかる) なら、そのブラウザに対するテストは行われないだろう。アプリケーションが複雑になればなるほど、テストしていないブラウザでそれがうまく動くかどうかは怪しくなる。最終的に統一された自動化フックが WebDriver ベースのものになるかどうかはわからない。でも、そうあって欲しいものだね!

今後数年の動きが楽しみだ。我々はオープンソースプロジェクトなので、いつでもみなさんの参加を歓迎する。さあ、一緒に <http://selenium.googlecode.com/> への旅に出ないかい?

Sendmail

Eric Allman

電子メールのプログラムと聞いて、たいていの人が思い浮かべるのはメールクライアントだろう。厳密には、Mail User Agent (MUA) と呼ばれているものだ。しかし、電子メールを扱うソフトウェアには、もうひとつ重要なものがある。それが、実際に送信者から受信者にメールを配達するソフトウェア—Mail Transfer Agent (MTA) である。インターネット上で最初に登場した MTA であり、現在でも幅広く使われているものが sendmail だ。

Sendmail が最初に作られたのは、まだインターネットが公式には存在しなかったころである。そして、それから大成功を収めた。1981 年に最初に作られたころ、インターネットはまだ学術的な実験段階で、接続されているホストはたかだか数百台に過ぎなかつた。それが今や、2011 年 1 月の時点でインターネットに接続されているホストは 8 億を超える¹。当時、今のこの状況を想像できた人はあまりいなかつただろう。Sendmail は、インターネット上の SMTP の実装として、今でも最も使われているものである。

17.1 むかしむかし...

後に sendmail として知られることになるプログラムの最初のバージョンが書かれたのは 1980 年のことだった。もともとは、メッセージを別のネットワークに転送するためのちょっとしたハックだった。インターネットは構築されていたが、当時はまだそれほど機能的ではなかつた。実際のところ、当時はさまざまなネットワークがコンセンサスを得ないままに提案されていた。アメリカでは Arpanet が使われており、インターネットはその上位版として設計された。しかしヨーロッパでは OSI (Open Systems Interconnect) の取り組みを支持しており、おそらく OSI のほうが勝ちを取めるだろうとみられていたこともあった。どちらも電話会社から借りた専用線を使っており、米国での速度は 56 Kbps だった。

おそらく、(接続するマシンやユーザーの数で考えると) 当時最も成功していたネットワークは UUCP ネットワークだろう。このネットワークの特徴は、中央管理型の権限がどこにも

¹<http://ftp.isc.org/www/survey/reports/2011/01/>

なかったことだ。ある意味では、ピアツーピアのネットワークの元祖と言えるだろう。このネットワークは電話回線でのダイアルアップで動いており、最大で 9600 bps の速度しか出なかつた。その当時の最速のネットワーク (3 Mbps) は Xerox の Ethernet をベースとしたもので、これは XNS (Xerox Network Systems) というプロトコルで動いていた—しかし、ローカルにインストールした環境以外では動作しなかつた。

当時の環境は、今とは異なつていた。コンピュータはそれぞれ異質なものであり、そもそも 1 バイトを 8 ビットにするかどうかさえ完全には合意されていなかつた。たとえば PDP-10(1 ワードが 36 ビット、1 バイトが 9 ビット)、PDP-11(1 ワードが 16 ビット、1 バイトが 8 ビット)、CDC 6000 シリーズ(1 ワードが 60 ビット、1 文字が 6 ビット)、IBM 360(1 ワードが 32 ビット、1 バイトが 8 ビット)、そして XDS 940 や ICL 470、Sigma 7 などがあつた。そのころ赤丸急上昇中だったプラットフォームのひとつが Unix で、これはベル研究所が発表したものだつた。Unix ベースのマシンの大半は 16 ビットアドレス空間を保持していた。当時の Unix マシンの主流は PDP-11 で、Data General 8/32 や VAX-11/780 が登場してきたころだつた。スレッドはまだ存在しなかつた—実際のところ、ダイナミックプロセスという概念自体がまだまだ出たてのものだつたのだ (Unix にはスレッドがあつたが、IBM の OS/360 みたいな“本物の”システムにはまだ実装されていなかつた)。ファイルのロックはまだ Unix カーネルではサポートされていなかつた(ファイルシステムのリンクを使うという裏技はあつた)。

とりあえずあるにはあつたものの、ネットワークは一般的に低速だつた(その多くは 9600 ポーの TTY 回線を使つていて。お金持ちの中には Ethernet を使つてゐるところもあつたが、それでもローカルネットワークだけの話だつた)。あの由緒あるソケットインターフェイスが発明されるのは、それから何年か後のことである。公開鍵による暗号化の仕組みもまだ発明されていなかつたので、ネットワークセキュリティに関して我々の知る仕組みのほとんどは実現不可能だつた。

ネットワーク上を流れる電子メールは Unix で既に実現されていたが、そのためにはちょっとしたハックを要した。当時のユーザーエージェントとして一番よく使われていたのが /bin/mail コマンド(現在では binmail や v7mail と呼ばれることもある)だが、それ以外のユーザーエージェントを使つてゐるところもあつた。たとえばバークレーの Mail は、メッセージを個別のアイテムとして扱う方法を知つていて。単に cat プログラムに任せるだけではなかつた。どのユーザーエージェントも、/usr/spool/mail ディレクトリを直接読んだり(そして直接書き込んだり!)してゐた。実際に格納されるメッセージを抽象化することなどなかつたのだ。

メッセージをネットワークに送るかローカルの電子メールに送るかの振り分けロジックは、単にアドレスを見て中に感嘆符 (UUTP の場合) あるいはコロン (BerkNET の場合) が含まれているかどうかを見るだけのことでしかなかつた。Arpanet にアクセスする人たちは、完全に個別のメールプログラムを使う必要があつた。これは他のネットワークとの相互運用ができず、ローカルメールも別の場所に異なる書式で保存してゐた。

さらにおもしろいのが、この時点ではまだ、メッセージ自体の書式についても事実上標準化されていなかつたということだ。大まかに決まつてゐるのは、メッセージの先頭にヘッダフィールドを置き、各ヘッダフィールドの後には改行を置いて、フィールド名とその値はコ

ロンで区切るということくらいだった。それ以外のこと、たとえばどんなヘッダフィールド名を使うかや個々のフィールドの構文などについてはほとんど標準化されていなかった。たとえば `Subject:` の代わりに `Subj:` を使うシステムもあったし `Date:` フィールドの構文もそれぞればらばらだったし、システムによっては `From:` フィールドのフルネームを解釈できないものもあった。さらに、ドキュメントもあいまいなものばかりだったし、ドキュメントが実際の内容と食い違っていることもあった。特に RFC 733(Arpanet メッセージについての説明)は実際に使われている内容と微妙に異なっていた(しかし重要な違いであることもあった)。そして、実際のメッセージ送信の仕組みは実際のところ公式には文書化されていなかった(その仕組みを扱う RFC はいくつかあったが、きちんとした定義はどこにもなかった)。メッセージングシステムまわりはこんなひどい状況だった。

1979年、INGRES Relational Database Management Project(私の昼間の業務だった)は DARPA の資金援助を受け、9600bps の Arpanet 接続を我々の PDP-11 につなげることになった。その当時、コンピュータサービス部門で Arpanet に接続しているマシンはそれだけだったので、誰もが我々のマシンを使って Arpanet を使ったがった。しかし、そのマシンはすでに最大限に使い切っており、使えるログインポートはあとふたつしか残っていなかった。そこで、そのふたつを部門内で共有した。そのため、頻繁に争いが発生した。しかし、私は気づいた。我々の大半はリモートログインやファイル転送を求めているのではなく、単に電子メールを使いたいだけだったのだ。

そんなところに、`sendmail`(当初は `delivermail` という名前だった)が登場した。このカオスを何とか統一しようという試みだ。すべての MUA(Mail User Agent、あるいはメールクライアント)は、単に `delivermail` を呼び出すだけでメールを配達することができた。いちいち調査をしてアドホックな(そして、たいてい互換性のない)対応をする必要もない。`Delivermail/sendmail` は、ローカルメールの保存方法や配達方法に関しては一切関知しなかった。つまり、単に他のプログラムとの間でメールを入れ替える以外のことは何もしなかったのだ(これは、後に SMTP が追加されたことで変わった。詳しくは後ほど)。ある意味では、自分自身がメールシステムであるというよりは、さまざまなメールシステムの間をつなぐ糊に過ぎなかったのだ。

Sendmail の開発を進めるうちに、Arpanet はインターネットに姿を変えた。その変更点は幅広く、ローレベルのパケットからその上にのるアプリケーションプロトコルまですべて変化した。そして、その変化は一斉に起こったわけではなかった。Sendmail の開発は標準規格の策定とまさに文字通り並行して進んでおり、Sendmail が標準規格に影響を及ぼすこともあった。もうひとつ特筆すべき点は、いま我々が思い浮かべる“ネットワーク”がまだ数百台規模だったころから何億台規模に成長するまでの間、sendmail がずっと生き残って成長を続けたという事実である。

もうひとつのネットワーク

ちなみに、当時まったく別のメール標準規格も提案されていた。それは X.400 と呼ばれ、ISO/OSI (International Standards Organization/Open Systems Interconnect) の一部だった。X.400 はバイナリプロトコルで、メッセージは ASN.1(Abstract Syntax Notation 1) で符号化する。この仕組みは、今でも LDAPなどの一部のインターネットプロトコルで使われている。LDAP は X.500 を単純化したものであり、X.500 は X.400 が使っているディレクトリサービスだ。Sendmail は X.400 との直接の互換性に関しては一切考慮していない。しかし、両者の橋渡しをするゲートウェイが存在する。X.400 は当初から多くの商用ベンダーに採用されていたが、最終的に市場を制したのはインターネットメールと SMTP だった。

17.2 設計の原則

Sendmail の開発中にこだわっている設計指針をいくつか紹介する。これらは結局ひとことでまとめることができて、要するに「無駄なことはしない」ということだ。これは、その当時の流れに反するものだった。当時は、より幅広い目標を達成するために実装をどんどん膨らませる方向に進んでいる人が多かった。

プログラマには限界があることを受け入れる

もともと sendmail は、業務としてではなく空き時間を使って書いたものだ。Arpanet メールをカリフォルニア大学バークレー校の人たちがより使いやすくするための手っ取り早い手段として作られた。その肝となるのが既存のネットワーク間でのメールの転送だった。すべての処理はスタンダードアロンのプログラムとして実装されており、複数のネットワークが存在することを想定していなかった。既存のソフトウェアにそれなりに手を入れるというのは、一人のプログラマーが空き時間だけでこなすには不可能な作業である。既存のコードをできるだけ変更せずに済み、かつ新しいコードもできるだけ書かずに済むような設計を目指す必

要があった。後述する指針の大半も、この考えに基づいている。仮に大規模なチームがいたとしても、この指針は間違いではなかっただろう。

ユーザーエージェントの再設計はしない

Mail User Agent (MUA) とは、エンドユーザーの多くが“メールソフト”と聞いて真っ先に思い浮かべるもの—メールの読み書きや返信のために使うプログラムである。Mail Transfer Agent (MTA) はそれとはまったく別で、電子メールを送信者から受信者に向けて配達する役割を果たす。Sendmail が書かれた当時、多くの実装はこれらふたつの機能を組み合わせたものであり、協力して開発を進めていた。両方同時に作業を進めるのはキツかったので、Sendmail ではユーザーインターフェイスの問題を完全に投げ捨てた。MUA に対して行った唯一の変更は、自前のルーティング処理の代わりに sendmail を起動させるようにする変更だった。実際、ユーザーエージェントは既にいろいろなものがあつたし、実際にメールを操作する部分なのでユーザー好みも人それぞれだ。MUA を MTA から切り離すという判断は今でこそ受け入れられるだろうが、当時の常識からは完全にかけ離れていた。

ローカルメールストアの再設計はしない

ローカルメールストア(受信者がメールを読むまでの受信メールの保存場所)は、正式には標準化されていなかった。/usr/mail や/var/mail あるいは/var/spool/mail のような場所で中央管理する方式をとっているところもあれば、受信者のホームディレクトリに(.mailなどといった名前のファイルで)保存しているところもあった。大半のサイトでは“From”で始まってその後スペースが続く行があればそこからメッセージが始まるものとしていた(あまりにも筋が悪すぎるが、当時はそういう決まりだった)が、Arpanet に注力していたサイトでは、control-A が4文字続く行でメッセージを区切って保管していた。一部のサイトではメールボックスをロックして衝突を回避していたが、サイトによってロックの規約が異なっていた(ファイルロックのプリミティブは当時まだなかった)。要するに、やれることがあるとすれば、ローカルのメールストレージは完全にブラックボックスとして扱うことくらいだった。

ほぼすべてのサイトで、実際にローカルメールボックスのストレージを扱う処理は/bin/mail に埋め込まれていた。このプログラムは、(極めて原始的な)ユーザーインターフェイスとルーティング、そしてストレージ操作の処理をひとつにまとめたものである。sendmail と組み合わせるために、ルーティング部分は外に切り出して sendmail を呼び出す処理に切り替えた。最終的な配達を強制するための-d フラグが追加された。つまり、/bin/mail が sendmail を呼び出してルーティングするのをやめさせることだ。後になって、物理的なメールボックスにメッセージを配達するコードが切り出され、mail.local という別のプログラムになつた。現在の/bin/mail は、メールを送信するための最小限の共通処理だけを残したものになつている。

Sendmail が世界に合わせるのであって、世界を Sendmail に合わせるのではない

UUCP や BerkNET といったプロトコルが既に個別のプログラムとして実装されており、それぞれ自前の(時にちょっと奇妙な)コマンドライン構造をもっていた。時には、それらのプログラムも sendmail と一緒に活発な開発が進むこともあった。sendmail で再実装してしまう(たとえば、標準の呼び出し規約に変換する)のは、どう見てもつらいだろう。ここから、次の原則が得られる。つまり、sendmail のほうが他の世界に合わせるのであって、他の世界のほうを sendmail の流儀に合わせようとしてはいけないということだ。

変更はできる限り少なく

最大限の注意を払い、sendmail の開発中には、触る必要のないものには一切触らないようにした。単に時間がないというだけの理由ではない。当時のバークレーには、コードの所有者を厳密に定めるのではなく“最後にコードを触った人が、そのプログラムの担当になる”(簡単に言うと“触ったら、責任を持て”)という文化があったのだ。今どきの常識で考えるとあり得ない話だが、これでうまく回っていた。当時のバーカークレーにはフルタイムで Unix に関わっている人はいなかったのだ。各人が、それぞれ自分が興味を持った部分の作業をしてコミットし、それ以外の部分についてはよっぽどの緊急事態でない限り手を触れなかつた。

信頼性をまず考える

sendmail 以前のメールシステム(大半のメール配達システムも含む)は、信頼性についてあまりにも無神経すぎた。たとえば、4.2BSD よりも前のバージョンの Unix にはネイティブなファイルロックの仕組みがなかった。その代わりの手段として、テンポラリファイルを使ってファイルロックをシミュレートしていた。つまり、テンポラリファイルを作成してそれをロックファイルにリンクさせたのだ(もしロックファイルが既に存在すれば、リンクのコールが失敗する)。しかし、時には別のプログラムが、ロックファイルとしての扱い方を知らないまま同じファイルに書き込みをしてしまうこともあり(たとえばロックファイル名が違ったり、そもそもロックを考慮していないなど)、メールを失う可能性もあり得るようになつてしまふ。sendmail では、メールを失うことなどあり得ないような手法を採用した(これは、もともと私がデータベース屋だったことも関係するかもしれない。データを失うなんてとんでもないという考えだった)。

その他

初期のバージョンでは、まだできていないことがたくさんあった。私はメールシステムを一から作り直そうとはしなかったし、完璧なソリューションを構築しようとも思わなかつた。

単に、必要になったときにその機能を追加するようにしたのだ。最初期のバージョンでは、設定を変更したければソースコードをいじってコンパイルし直せといった具合だった(さすがにそれはすぐになくなつたが)。sendmail のやり方は、だいたいこんな感じだ。まず、何か動くものを手早く作る。そして、必要に応じて機能拡張し、問題があれば対応する。

17.3 開発フェーズ

古くからあるソフトウェアはだいたいそうだが、sendmail の開発もいくつかのフェーズに分けることができ、個々のフェーズについて基本テーマや考えがある。

第一波: delivermail

sendmail が最初に登場したときには、delivermail と呼ばれていた。単純化しすぎとは言わないまでも、シンプル極まりないものだった。唯一の機能は、あるプログラムから別のプログラムにメールを転送することだった。実際のところ SMTP すらサポートしておらず、直接のネットワーク接続も一切行わなかった。キューなんか不要だった。だって、それぞれのネットワークが自分でキューを持っているんだから。プログラムは、単にクロスバースイッチになればよいだけのことだ。delivermail はネットワークプロトコルを直接にはサポートしていなかったので、デーモンとして動かす理由などなかった。メッセージが投稿されるたびに起動してメッセージを仕分け、次のホップを実装する適切なプログラムにそれを渡し、あとは終了するだけでよい。また、ヘッダを書き換えてメッセージの配達先のネットワークにマッチさせることもしなかった。そのせいで、転送したメッセージに返信することができないということになりがちだった。さすがにそれはまずかったので、メールの処理だけを扱う本が書かれたりした(*!%:: A Directory of Electronic Mail Addressing & Networks [AF94]* という、ぴったりな名前がついている)。

delivermail はすべてコンパイルされる。また、すべてアドレス内に埋め込まれた特殊文字を利用していた。特殊文字には優先順位があった。たとえば、ホストの設定に関しては“@”記号を探し、もし見つかれば、そのアドレス全体を割り当てられた Arpanet リレーホストに送信する。見つからなければ次にコロンを探し、割り当てられたホストとユーザー(見つかれば)に BerkNET でメッセージを送る。そして次に感嘆符(“!”)を探す。これは、そのメッセージを UUCP リレーに転送することを表す。見つからなければ、ローカルへの配達を試みる。この設定をまとめると、次のようになる。

アドレスの区切り文字はその組み合わせによって異なり、結果的に曖昧な状態になってしまって経験則でしか解決できなくなることもある。たとえば最後の例は、別のサイトなら{Uucp, foo, bar@baz}と解釈されることもあり得るだろう。

設定をコンパイルする理由はいくつもある。まず、アドレス空間が 16 ビットでメモリも限られているので、実行時に設定をパースするのはコストがかかりすぎだった。次に、当時の

入力	送信先 ネットワーク、ホスト、ユーザー
foo@bar	{ Arpanet, bar, foo }
foo:bar	{ Berknet, foo, bar }
foo!bar!baz	{ Uucp, foo, bar!baz }
foo!bar@baz	{ Arpanet, baz, foo!bar }

システムは個々のサイトで非常にカスタマイズされていたので、再コンパイルをするのも理にかなっていた。ローカルに、必要なバージョンのライブラリがあることを確実にするためである（共有ライブラリは、Unix 6th Edition の時点ではまだなかった）。

*D*elivermail は 4.0 BSD および 4.1 BSD に同梱され、予想以上の好評を得た。ハイブリッドなネットワークアーキテクチャを採用しているのはバークレーだけではなかったというわけだ。その証拠に、いろいろな要求が出てきた。

第二波: **sendmail 3、4、そして 5**

バージョン 1 とバージョン 2 は、*delivermail* という名前で公開された。1981 年 3 月にバージョン 3 の開発が始まる。このバージョンからは、*sendmail* という名前で公開されるようになった。この時点ではまだ 16 ビットの PDP-11 が一般的に使われていたが、32 ビットの VAX-11 もポピュラーになりつつあった。当初の制約の中でもアドレス空間による制約については、これで和らぎはじめた。

sendmail の初期の目標は、実行時の設定に変換して別のネットワークへのメッセージの転送をできるよう容易することだった。また、ルーティングを設定するためのリッチな言語を提供することだった。利用していたテクニックは基本的にテキストレベルでのアクセスの書き換え（文字ではなくトークンにもとづいたもの）で、当時の大規模システムでも一般的に使われていたものだ。アドホックなコードを使い、コメント文字列（括弧で囲まれたもの）を切り出して保存した後でプログラムによる書き換えを終えてからもう一度挿入する。ヘッダフィールドを追加したり拡張したり（たとえば Date ヘッダフィールドを追加したり、送信者のフルネームがわかっている場合は From ヘッダにそれを含めるなど）できるようにしておくことも大切だ。

SMTP の開発が始まったのは 1981 年 11 月のことだった。カリフォルニア大学バークレー校の The Computer Science Research Group (CSRG) が DARPA と契約を結び、Unix ベースのプラットフォーム作って DARPA が出資する研究をサポートした。その意図は、プロジェクト間の共有をしやすくすることだった。TCP/IP スタックを初めて手がけたのがちょうどこの頃だが、ソケットのインターフェイスの詳細はまだ確定していなかった。Telnet や FTP といった基本的なアプリケーションプロトコルはできていたが、SMTP はまだ実装されていなかったのだ。実際、その時点ではまだ SMTP プロトコルの策定に決着がついていなかった。メールをどのように送信すべきかの議論は収束しておらず、プロトコルの名前は Mail Transfer

Protocol (MTP) と呼ばれていた。議論は白熱し、MTP はどんどん複雑になっていった。結局、業を煮やして SMTP (Simple Mail Transfer Protocol) の提案が出された。多かれ少なかれ、恣意が入っていた(公式に公開されるのは 1982 年 8 月のことだ)。公式には、私がかかわっているのは INGRES Relational Database Management System だった。しかし、当時のバークレーの中で私以上にメールシステムを知っている人はいなかったので、SMTP の実装にもかかわることになった。

私が最初に考えたのは、SMTP を話すメーラーを別に作って自前のキューを持たせ、データベースとして動かすことだった。このサブシステムを sendmail にアタッチして、ルーティングを任せる。しかし、SMTP の機能のいくつかのせいで、この案は困難になった。たとえば、EXPN コマンドや VRFY コマンドは、パース処理やエイリアス処理そしてローカルアドレスの検証モジュールにアクセスできなければならぬ。また、当時の私が重視していたのが、RCPT コマンドは未知のアドレスを受け取ったときにすぐに結果を返すということだった。いったんメッセージを受け付けてから後で配達失敗のメッセージを返すのではいけないと考えたのだ。これは後に、先見の明があったと判明する。皮肉なことに、後の MTA の多くはここを誤り、spam の氾濫を招いてしまっている。これらの問題があったので、SMTP は sendmail 本体に含めることに決めた。

Sendmail 3 は 4.1a BSD および 4.1c BSD (ベータ版) に同梱され、sendmail 4 は 4.2 BSD に含まれた。そして sendmail 5 は 4.3 BSD に含まれることになった。

第三波：カオスな日々

バークレーを去ってスタートアップ企業に転職してから、私が sendmail に割ける時間は大幅に減少した。しかし、インターネットの世界はどんどん拡大し続け、sendmail もさまざまな新しい(そして巨大な)環境で使われるようになった。主要 Unix ベンダーのほとんど(Sun、DEC、そして IBM など)は自前の sendmail を用意しており、お互いに互換性がなくなっていた。オープンソースの sendmail を作ろうという試みもあった。特筆すべきは IDA sendmail と KJS だ。

IDA sendmail は、リンシェーピン大学によるものだ。IDA が含めた拡張は、大規模な環境やまったく新しい構成のシステムへのインストールや管理を容易にするものだった。主な新機能のひとつは dbm(3) データベースを含めたことで、これで活発なサイトもサポートできるようになった。設定ファイルに新しい構文を導入しており、外部のシステムでのアドレスの構文とのマッピング(たとえば johnd@example.com のかわりに john_doe@example.com 宛てにメールを送るなど)やルーティングなどができるようになった。

King James Sendmail(略して KJS。ポール・ヴィクシーが作ったものは、そこら中にわき出したさまざまなバージョンの sendmail を統一しようとした試みである。残念ながら、期待していたほどの影響を及ぼすにはとうてい及ばなかった。新しい技術をメールシステムに取り込もうとしそうで失敗したのだ。たとえば、Sun が作ったディスクレスクラスタに、YP(後の NIS) ディレクトリサービスや NFS(Network File System) を追加したりした。特に、YP は

sendmail に見えなければならない。エイリアスをローカルファイルではなく YP に保存していたからである。

第四波: sendmail 8

数年を経て、私はふたたびスタッフとしてバークレーに戻ってきた。当時の業務は、計算機科学部の研究用共有基盤のインストールやサポートを管理することだった。そのためには、個々の研究グループでアドホックに構築した大規模な環境を何らかのきちんとした方法で統合する必要があった。インターネットの黎明期と同様、研究グループごとにまったく違うプラットフォームを使っており、中には古すぎるものもあった。一般に、すべての研究グループは独自のシステムを持っていた。しかしどんどんうまく管理されておらず、大半が“繰越維持費”に悩まされていた。

たいていの場合、電子メールはみな整っていない。各個人の電子メールアドレスは“`person@host.berkeley.edu`”のようになっており、`host` はオフィスのワークステーション名あるいは共有サーバー名(キャンパス内では内部サブドメインは使っていなかった)だった。例外として、一部の人は`@berkeley.edu` アドレスを持っていた。目標は、内部サブドメインを使うようにする(つまり、すべてのホストを `cs.berkeley.edu` サブドメインに置く)ことと、統一されたメールシステムを使う(つまり、すべての人が`@cs.berkeley.edu` なアドレスを持つ)ことだった。この目標を実現するための最も簡単な方法は、新しいバージョンの sendmail を学部全体で使わせることだった。

まずは、多数出回っている sendmail の類似品の中でもよく使われているものについて調べることにした。違いを探すというよりはむしろ、他の類似品にはない便利な機能について理解することを心がけた。その過程で見つかったアイデアを元にして sendmail 8 を作っていった。関連するアイデアをひとまとめにしたり、より汎用的にしたりなどの変更を加えることもよくあった。たとえば、sendmail の類似品の中には dbm(3) や NIS のような外部のデータベースにアクセスできる機能を持つものもあった。sendmail 8 では、これらをひとまとめにした“マップ”という仕組みを導入し、複数の形式のデータベース(データベース以外にもどんな変換方式も使える)を処理できるようにしている。同様に、“汎用”データベース(外部の名前マッピングを利用する)機能は、IDA の sendmail から導入した。

Sendmail 8 には新たな設定パッケージも導入された。これは m4(1) マクロプロセッサを利用したものだ。sendmail 5 の設定パッケージよりもさらに宣言的に記述できるよう心がけ、大部分は手続き型で書けるようになっている。sendmail 5 の場合は、管理者が設定ファイルをすべて手で書き換える必要があった。単に m4 からのショートカットとして “include” を使うだけであってもだ。sendmail 8 の設定ファイルの場合、管理者が宣言できるのは、必要な機能やメーラーなどだけで、最終的な設定ファイルは m4 が作成する。

17.7 節で、sendmail 8 での機能追加について解説する。

第五波：ビジネスの日々

インターネットが成長するにつれて sendmail を使うサイトも増加し、大量のユーザーをサポートするのが徐々につらくなってきた。しばらくの間は、ボランティアの集まり（非公式にだが、“Sendmail Consortium”あるいは sendmail.org と呼ばれていた）のおかげで無料サポートを続けることができていた。サポートには電子メールやニュースグループを使っていた。しかし 1990 年代後半には、インストール数が増えすぎて、もはやボランティアベースでのサポートを続けるのはほぼ無理な状態になっていた。そこで、よりビジネスよりな友人とともに Sendmail, Inc.²を設立し、コードの面倒を見る新たな人材を確保しようとした。

そこで扱っていた商品は、設立当初は設定管理ツールが主だった。しかし、オープンソースの MTA にも、ビジネスの世界の要望に対応するためにさまざまな新機能が追加された。特筆すべき点としては、TLS (connection encryption) や SMTP Authentication のサポート、Denial of Service 対策などサイトのセキュリティ対策の向上、そして最も重要なものとしてはメールのフィルタリング用プラグイン（後述する Milter インターフェイス）などがある。

本章の執筆時点では、扱う商品はさらに幅広くなり、電子メールベースの大規模なアプリケーションスイートも含まれるようになっている。そのほぼすべては sendmail の拡張機能として作られており、創業当初の数年間で作られたものだ。

sendmail 6 や 7 はどこに行ったの？

Sendmail 6 は、本質的には sendmail 8 のベータ版だった。公式にはリリースされなかつたが、かなり広範囲に広まった。Sendmail 7 が存在することはなかった。バージョン 6 からバージョン 8 に一気に上げたのだ。というのも、1993 年 6 月に 4.4 BSD がリリースされたことにあわせて、BSD ディストリビューションのファイルをすべてそれにあわせたからである。

17.4 設計に関する決定

正しい判断ができたこともあった。当時は正しい判断だったが、その後状況が変わってまづい判断になったものもあった。そして、どちらとも言えない曖昧なものもあった。

設定ファイルの構文

設定ファイルの構文は、次のふたつの要因によって決まった。ひとつは、アプリケーション全体を 16 ビットアドレス空間におさめるために、パーサを小さくする必要があるということ。もうひとつは、初期の設定はごく少なかった（1 ページにおさまる程度）ので、多少構文をあいまいにしても、ファイルがそんなに読みにくくはならなかつたということ。しかし、

²<http://www.sendmail.com>

時を経て、より多くの判断が C のコードから設定ファイルに追い出されるようになり、設定ファイルが肥大化し始めた。そして、設定ファイルは「難解なもの」という悪評が広まった。多くの人にとってフラストレーションの元となったのが、タブ文字を構文的に意味のある要素とした判断だ。これは、当時他のツール (make など) で使われていた構文をそのまま使っただけだが、間違いだった。この問題が深刻化したのは、ウィンドウシステムが登場したころ(つまり、カットアンドペーストが多用されるようになったころ。カットアンドペーストではタブ文字の情報が残らない)からである。

今思えば、設定ファイルが巨大化して世の中が 32 ビットマシンに取って代わられた頃に、設定ファイルの構文を検討しなおしてもよかつた。そのように考えた時期もあったのだが、結局そうしなかった。当時すでに“大量に”インストールされていた sendmail 環境を壊してしまいたくなかったからである(実際のところ、その時点では実際にインストールされていたマシンは、おそらく数百台程度だっただろう)。この判断は間違っていた。当時の私は、その後インストール数がどれほど伸びるかを想像できなかつたし、早めに構文を変更しておくことで今後どれだけの時間を節約できるのかにも思いが及ばなかつたのだ。また、標準規格がある程度安定してきた時点で、一般的な項目はもう一度 C のコード側に戻せば、設定ファイルはもう少しシンプルにできただろう。

当時特に気についていたのは、どれだけの機能を設定ファイルに追い出せるかということだった。私が sendmail の開発を進めていたのは、ちょうど SMTP の標準規格を定めようとしつつある頃だった。SMTP 側で設計の変更があれば、すぐに—通常は 24 時間以内に—それを設定ファイルに追い出すようにしていた。これは、SMTP の策定にも貢献したと思っている。何か設計の変更が提案されればそれをすぐに実際に試すことができたし、試すためには(難解な)設定ファイルを書くだけで済んだからだ。

ルールの書き換え

sendmail を書く際に決めづらかったのが、ネットワーク越しの転送を許可するために必要な書き換えを、受信側のネットワークの規約に違反しない方法で行うためにどうすればよいかということだった。ネットワーク越しの通信では、メタ文字の変更(たとえば BerkNET ではコロンを区切り文字に使っていたが、コロンは SMTP のアドレスには使えない)やアドレスコンポーネントの並べ替え、そしてコンポーネントの追加や削除などが必要となる。たとえば、状況に応じてこのような書き換えが必要となった。

From	To
a:foo	a.foo@berkeley.edu
a!b!c	b!c@a.uucp
<@a.net,@b.org:user@c.com>	<@b.org:user@c.com>

正規表現はあまり良い選択ではなかった。というのも、正規表現ではワードの区切りや

クオートなどにうまく対応できなかつたからである。すぐに明らかになつたことだが、これに対応する正規表現を書くのは事実上ほぼ不可能で、とてもわかりにくいものになつてしまつた。正規表現では、いくつかのメタ文字を予約語として使つてはいる。たとえば“.”や“*”、“+”、“[”そして“]”がそれにあたるのだが、これらはみな電子メールアドレスの中に登場しうる文字である。設定ファイルでこれらの文字をエスケープしてしまつてもよかつたのだが、それは複雑で混乱の元になるし、ちょっと見苦しいと思つた(ベル研のUPASがこの方式を採用していた。これはUnix Eighth Editionのメーラーとして採用されたが、まったくヒットしなかつた³)。そのかわりにスキヤンフェイズが必要となり、そこでトークンを切り出して正規表現のように文字を操作することにした。“オペレータ文字”を表すひとつのパラメータをトークンやトークンの区切りとみなせば十分だつた。空白文字はトークンを区切るが、それ自身はトークンにはならない。書き換えルールは単なるパターンマッチと置換の組み合わせであり、原則的にサブルーチンとして組み込まれた。

大量のメタ文字をエスケープしてそれらの“特殊”機能(正規表現で使われているもの)を消し去るかわりに、私は单一の“エスケープ”文字を使うことにした。これを通常の文字と組み合わせて、ワイルドカードパターン(任意の単語にマッチする、など)を表すのだ。伝統的なUnixのアプローチなら、ここでバックスラッシュを使うところだつた。しかし、バックスラッシュは既に一部のアドレス構文でクオート文字として使われていた。いろいろ調べた結果見つかった数少ない候補のひとつが“\$”で、これなら電子メールの構文で特殊文字としては使われていなかつた。

当初の方針の中で間違つていたと思えるのは、皮肉にも、空白の使い方だつた。空白は区切り文字で、これはスキヤン対象の入力の大半と同様だつた。そこで、パターン内のトークンの間で自由に空白を使うことができた。しかし、元々配布されていた設定ファイルには空白は含まれておらず、結果としてパターンが必要以上に読みづらいものとなつてしまつた。次のふたつのパターンを比較してみよう(文法的にはどちらも同じ意味だ)。

```
$+ + $* @ $+ . ${mydomain}  
$+$*@$+. ${mydomain}
```

書き換えを使ったパース

sendmailでは文法学にもとづいてアドレスをパースすべきであり、書き換えルールを使うべきではないという提案もあつた。書き換えルールを使うのはあくまでもアドレスの書き換えだけにすべきだ、と。聞く限りでは理にかなつてゐるように見える。ただし標準規格でのアドレスの定義が文法学に乗つ取つていればの話だが。書き換えルールを使ひ回している主な理由は、場合によつてはヘッダフィールドのアドレスをパースする必要があるということだ(正式なエンベロープを持たないネットワークから受信したメールで、ヘッダから送信者エンベロープを取り出す場合など)。この手のアドレスのパースは、YACCのようなLALR(1)パーサや伝統的なスキヤナでは困難である。というのも、相当な先読みを要求されるからだ。た

³http://doc.cat-v.org/bell_labs/upas_mail_system/upas.pdf

とえば、`allman@foo.bar.baz.com <eric@example.com>`のようなアドレスをパースすることを考えると、スキヤナあるいはパーサで先読みが必要となる。つまり、最初の “`allman@...`” がアドレスではないということは、少なくとも “`<`”まで読まなければわからない。LALR(1) パーサには先読みトークンがひとつしかないので、これはスキヤナで行う必要がある。相当複雑な作業だ。書き換えルールには既にいくらでも後戻りできる(つまり、いくらでも先読みできる)ので、こちらのほうが適している。

第二の理由は、パターンの認識がしやすいうえに壊れた入力も修正しやすかったということだ。最後の理由は、やりたいことをこなすには書き換えでも十分すぎるくらい高機能だったからである。それに、コードを再利用できるならそうするほうが賢いというものだ。

書き換えルールに関して特筆すべき点がひとつある。パターンマッチを行うときには入力とパターンとともにトークンに分割したほうが有用だ。そうすれば、入力アドレスとパターンそのものに対して同じスキヤナを使える。そのためには、スキヤナを呼び出す際に、入力ごとに文字タイプテーブルを切り替えられるようにしなければならない。

SMTP やキューの `sendmail` への埋め込み

SMTP の送出側(クライアント側)の実装として“自明な”方法は、UUTP と同様に外部のメーラーとして実装することだ。しかしこの場合、いくつか疑問が出てくる。たとえば、キューイングは `sendmail` でするのだろうか? それとも SMTP クライアントモジュールで行うのだろうか? `sendmail` で行うのなら、メッセージのコピーを各受信者に送信する(つまり、“piggybacking” はせず、そこで一つの接続を開き、複数の RCPT コマンドを送る)か、あるいはずっとリッチな逆方向のコミュニケーション手段が必要になる。受信者ごとの状況を知るには、単純に Unix の終了コードを使うだけでは不十分だからである。クライアントモジュール側で行うのなら、大量の複製が発生する可能性が出てくる。特に、当時は XNS など他のネットワークもまだ候補にあった。さらに、キューを `sendmail` 側に含めれば、よりエレガントな方法で障害に対応できるようになる。特に、リソースの枯渇などの一時的な問題に対応しやすい。

受け入れ側(サーバー側)の SMTP については、難しい決断があった。当時の私は、VRFY や EXPN といった SMTP コマンドも忠実に実装することを重視していた。これらのコマンドは、エイリアスの仕組みにアクセスできなければならない。これを実現するには、SMTP サーバーモジュールと `sendmail` との間でよりリッチなプロトコルの交換が必要となる。これは、単にコマンドラインと終了コードだけで実現できるものではない—実際のところ、SMTP 自体はそのようなプロトコルだったのだが。

今なら、キューイングは `sendmail` のコアに残すだろうが SMTP のクライアント側、サーバー側の実装はそれぞれ別のプロセスに切り出してしまいたいところだ。その理由のひとつはセキュリティの向上である。サーバー側でいったん 25 番ポートを開いたインスタンスを持てば、もはや root 権限は不要になる。TLS や DKIM 署名のようなモダンな拡張のせいでクライアント側は複雑になるが(権限を持たないユーザーに秘密鍵へのアクセスを許してはいけないから)、厳密に言えばこちらも root 権限は不要だ。しかし、セキュリティの問題は依然

問題として残る。クライアント側の SMTP が非 root ユーザーで稼働していたとしても、秘密鍵を読めるということは特別な権限を持っていることになる。つまり、他のサイトと直接通信すべきではないということだ。これらの問題はすべて、多少の手間でなんとかできる。

キューの実装

sendmail は、当時の規約に従ってキューファイルを格納していた。実際のところ、採用したフォーマットは当時の lpr サブシステムと極めて似たものだった。各ジョブに対してふたつのファイルがあり、ひとつが制御情報でもうひとつが実際のデータとなっていた。制御ファイルはフラットなテキストファイルで、各行の最初の文字がその行の意味を表していた。

sendmail がキューを処理するときは、制御ファイルをすべて読んだ上で関連する情報をメモリ内に格納し、そしてリストをソートしなければならない。キュー内のメッセージ数が比較的少ない場合は、これはうまく機能した。しかしキュー内のメッセージが 10,000 前後になつた時点から不調になり始めた。特に、ディレクトリが肥大化してファイルシステム内の間接ブロックを要するようになると、そこがパフォーマンスに深刻な影響を及ぼす。大幅にパフォーマンスが落ちてしまうこともあり得る。この問題を改善する手段として sendmail で複数のキューディレクトリを扱えるようにすることもできた。しかしそうしたところで、せいぜいちょっとしたハック程度の効果しか得られないだろう。

別の実装としては、すべての制御ファイルをひとつのデータベースファイルにまとめるともできただろう。そうしなかった理由は、sendmail を書き始めた頃にはまだ汎用的に使えるデータベースパッケージが存在しなかつたことだ。後に登場した dbm(3) にはいくつかの問題があった。たとえば、領域の再配置(すべてのキーを 512 バイトの単一ページにまとめるために必要)ができないことやロックの仕組みがないことだ。堅牢なデータベースパッケージは、なかなか登場しなかつた。

それ以外にも、別のデーモンを用意してキューの状態をメモリ上に保持させるという方法もあつただろう。そのデーモンがログを書いておけば、リカバリも可能だ。当時はまだ電子メールのトラフィックがそれほど多くなかつたこと、そしてメモリを潤沢に搭載したマシンが少なかつたこと、バックグラウンドプロセスのコストが比較的高いこと、プロセスの実装が複雑になることなどを考慮すると、当時としてはこの選択肢は割に合わなかつた。

もうひとつの設計上の判断として、メッセージヘッダをキューのデータファイルではなく制御ファイル側に格納するようにした。その根拠は、ほとんどのヘッダはそれなりの書き換えを要し、書き換え方法も配達先によってさまざまに変わること(そしてメッセージの配達先が複数になることもあり、複数回のカスタマイズが必要になる)。そしてヘッダのペースのコストが高そうだったことだ。そこで、ペース済みの形式でヘッダを格納しておけばコストを抑えられるだろうと考えた。今思えば、これはあまりよい判断ではなかつた。ちょうどメッセージ本文の格納に Unix 標準フォーマット(行末が newline)を使い、受信したメッセージのフォーマット(行末は newline かもしれないし carriage-return/line-feed かもしれない。単に carriage-return だけかもしれないし line-feed/carriage-return かもしれない)を使わなかつた

ように。電子メールの世界が成長して標準が定まっていくにつれて、書き換えの必要性は少なくなった。さらに、どうってことないよう見える書き換えであってもエラーのリスクがある。

誤入力の受け入れと修正

sendmail が作られた当時の世界にはさまざまなプロトコルが乱立しており、標準規格もほとんど定まっていなかった。そのため、不正な形式のメッセージはできる限りきれいにしようと決めた。これは、RFC 793 に明記された“ロバストネス原則”(またの名をポステルの法則)⁴にもマッチしている。これらの変更の中には、明白であり必須なものもある。UUCP メッセージを Arpanet に送信するときには、UUCP アドレスを Arpanet アドレスに変換しないと単なる“reply”コマンドすら正常に動作しない。また、行末文字をさまざまなプラットフォームの決まりにあわせて変換したりなどといった作業も必要だ。中には自明だとまでは言えないものもある。受信したメッセージに From: ヘッダーフィールドがない場合はどうすればいいだろう? このフィールドは Internet の仕様では必須のものだが、ここで From: ヘッダーフィールドを追加してしまうべきだろうか? それとも From: ヘッダーフィールドを追加せずにそのまま通してしまうべきだろうか? あるいはそのメッセージを拒否すべきだろうか? 当時の私が最重要視していたのは相互運用性だった。そこで sendmail では、メッセージにパッチをあてて From: ヘッダーフィールドを追加するなどしていた。しかしそのせいで、壊れている他のメールシステムがそのままの状態で長く生き続けることになってしまった。本来なら、ずっと前に修正されるなりこの世から消えてしまうなりすべきだったはずなのに。

私の判断は、その当時は正しかったと確信している。しかし今となっては問題もある。高度な相互運用性を維持することは、メールの流れを妨げないためにも重要だった。もし不正なメッセージを拒否していたら、当時のメッセージの大半は拒否されてしまっていただろう。もし何も手を入れずに素直にしていたら、受け取ったメッセージに返信できなくなってしまう。というか、そもそも誰がそのメッセージを送ったのかさえわからなくなる—あるいはそのメッセージが別のメールに拒否されたのかどうかもわからない。

現在は標準規格がきちんと定まっており、そのほとんどの部分は正確で完全になっている。もはや、大半のメッセージが拒否されてしまうという状況ではなくなつた。しかし今もなお、不正な形式のメッセージを送るメールソフトが残っている。そのせいで、インターネット上の他のソフトウェアの間でも無駄に多くの問題が発生している。

設定ファイルでの M4 の使用

ある時期、sendmail の設定ファイルに頻繁に変更を加えて個人的にさまざまなマシンに対応させようとしていたことがあった。設定ファイルの大半はマシンが異なつても同じだったた

⁴“自分には保守的であれ。他者はリベラルに受け入れよ。”

め、できれば何かツールを使って設定ファイルを作れたらいいなと考えていた。m4 マクロプロセッサは Unix に含まれるツールである。もともとは、プログラミング言語(特に RATFOR)のフロントエンドとして設計された。最も重要なのは、m4 が“include”機能に対応していたことだ。これは C 言語における “#include” と同様の機能である。最初の設定ファイルはこの機能に毛が生えた程度で、ちょっとしたマクロによる拡張をしただけのものであった。

IDA sendmail も m4 を使っていたが、その使い方はまったく異なっていた。今思えば、当時の私はこれらのプロトタイプをもっと調査するべきだったのだろう。彼らはいろいろうまい使い方をしており、特にクオートの処理方法がすばらしかった。

sendmail 6 以降、m4 設定ファイルは新たに書き直され、より宣言的なスタイルで分量も少なくなった。この設定ファイルは m4 プロセッサのパワーをさらに利用するものとなつたが、そのために、GNU m4 がその構文を微妙に変更しただけでも問題になることがあつた。

最初に考えていたのは、m4 の設定は 80/20 ルールに従つて使おうということだった。つまり、m4 を使うのは作業全体の 20%だけにしておけばファイルがシンプルになるし、その 20%が問題の 80%をカバーしてくれるということだ。この計画はすぐに頓挫した。その理由はふたつ。まずは些細な理由のほうから話す。少なくとも最初のうちはこの計画はうまくいき、問題の大半は比較的楽に処理できていた。しかし sendmail やそれを取り巻く世界が拡大するにつれて、徐々に難しくなってきた。TLS 暗号化や SMTP 認証などの機能が組み込まれても、それに対応するのに時間がかかるようになったのだ。

もうひとつの重要な理由は、生の設定ファイルを直接さわるのがもはやほとんどの人にとって難しくなりつつあったということだ。要するに、生の.cf フォーマットはアセンブラーのコードに等しい状態になつた。細心の注意を払えば編集できるが、現実的な話ではない。そこで、m4 スクリプトで書かれたその“ソースコード”が.mc ファイルとして格納されることになった。

もうひとつの重要な特徴は、生の設定ファイルは実際のところプログラミング言語だったということだ。手続き型のコード(ルールセット)やサブルーチンの呼び出し、パラメータの展開、そしてループなどの機能があった(しかし goto は使えない)。その文法は曖昧なものものだったが、おおまかには sed や awk と似ていた。少なくとも概念的には。m4 フォーマットは宣言型だった。低レベルの言語機能を使うこともできたが、現実的にはこれらの詳細はユーザーからは隠されていた。

m4 を使うという判断が正しかったのかどうかはよくわからない。当時の私が考えていた(そして今もそう思っている)のは、複雑なシステムを便利に使うためには、何らかの DSL(ドメイン特化言語)を実装してシステムを構築できるようにするとよいということだ。しかし、この DSL を設定項目としてエンドユーザーに公開すると、システムの設定がプログラミングの問題になってしまう。DSL は強力なものだが、それを使うためのコストも無視できない。

17.5 その他の検討事項

それ以外にも、アーキテクチャや開発に関して話しておくべきポイントがある。

インターネット全体に広がるシステムの最適化

ネットワークベースのシステムのほとんどには、クライアント側とサーバー側の対立がある。クライアント側にとって都合のいい戦略はサーバー側にとってあまりよろしくないものであることもあるし、その逆もまたあり得る。たとえば、サーバー側での処理コストを最小限に抑えようとすれば、できる限りの情報をクライアント側に PUSH することになる。一方クライアント側でも同じように考えたとすると、やることは同じだがその方向が正反対になる。たとえば、サーバー側では spam 処理の間も接続を維持したままにしておきたいだろう。そうすればメッセージの配達を拒否する際のコストが下がるからである(最近は、配達を拒否することが一般的になった)。しかし、クライアント側としてはできるだけ早く次に進みたいはずだ。システム全体を見て、インターネット全体のことを考えれば、クライアント側とサーバー側のニーズに対してうまくバランスをとるのが最適なソリューションとなる。

クライアント側あるいはサーバー側のいずれか一方を明示的に重視した戦略をとる MTA もいくつかあった。そんなことができたのは、それらの MTA があまり広まっていなかつたからというだけの理由にすぎない。自分のシステムがインターネット上の大部分で使われるようになると、両者にかかる負荷のバランスを考慮した設計が必要となる。インターネット全体を最適化するためだ。これは複雑で込み入った作業となる。なぜなら、MTA というものは常にどちらか一方向に完全に傾いてしまっているものだからである。たとえば大量メール配信システムはメールの送出側の最適化しか気にしない。

接続の両サイドにかかるシステムを設計する際には、どちらか一方に注意を向けすぎないようにすることが大切だ。これは、よくあるクライアントとサーバーの不調和とはまったく対照的な話であることに注意しよう—たとえばウェブサーバーとそのクライアントは、通常は別々のグループが開発しているものだ。

Milter

sendmail への機能追加のうち最も重要なもののひとつが milter (*mail filter*) インターフェイスだ。milter はオフボードのプラグインとして(つまり、別プロセスで実行させて)メールの処理に使える。もともとはスパム対策のために作られたものだった。milter のプロトコルは、サーバーの SMTP プロトコルと同期して動作する。新たな SMTP コマンドをクライアントから受信するたびに、sendmail は milter を呼び出してそのコマンドからの情報を渡す。milter は、その内容にあわせてコマンドを受け入れたり拒否したりする。拒否した場合は SMTP プロトコルでのそのコマンドの実行が却下される。milter はコールバックとして設計されており、SMTP コマンドを受け取ると適切なサブルーチンが呼び出される。milter はスレッド化されている。接続単位のコンテキストポインタを保持しており、各ルーチンに状態を渡すことができる。

理屈上は、milter は sendmail のアドレス空間内にロードできるモジュールとしても動かせる。しかしそうしなかった。理由は次の三点である。まず第一に、セキュリティの問題が重

大だった。仮に sendmail 専用に root 権限のないアカウントを作つて運用していたとしても、そのユーザーはすべてのメッセージの状態にアクセスできるようになる。同様に、milter の作者が sendmail の内部状態にアクセスしようとすることも避けられない。

第二に、我々は sendmail と milter の間にファイアウォールを作つたかった。仮に milter がクラッシュしたとしてもそこで被害を食い止め、メールの流れは妨げられないようにしたかった。第三に、milter の作者にとって、スタンドアロンのプロセスをデバッグするほうが sendmail 全体を相手にするよりもずっと楽だった。

milter がスパム対策以外にも有用であることに、間もなく気づきだした。実際、milter.org⁵のサイトにはさまざまな milter が掲載されている。スパム対策以外にもウィルス対策やアーカイブ、コンテンツ監視、ログ出力、トライフィックの削減などさまざまなカテゴリがあり、商用製品もあればオープンソースのプロジェクトも存在する。postfix⁶も、同じインターフェイスで milter をサポートするようになった。milter は、sendmail の大成功を示す一例と言える。

リリーススケジュール

よく議論になるのが、“早めに、そして頻繁にリリース”派と“安定したシステムをリリース”派の対立である。sendmail は、このどちらの手法も繰り返し使つてきた。かなり大量の変更をするときには、一日に何度もリリースすることもあった。基本的な考えは、何か変更するたびにリリースするというものだった。これは、ソース管理システムのツリーを一般向けに公開するのと同じようなことだ。私は個人的に、頻繁にリリースするほうがソースツリーを公開するよりも好きだ。少なくともその理由の一部になっているのは、私のソース管理システムの使い方があまり一般的ではないということだ。大規模な変更をするときなど、コードを書いている途中できちんと動作しない状態のスナップショットをチェックインすることもある。もしツリーを公開することになったら私はブランチを切つてスナップショットを扱うことになるだろう。しかし、いずれにしてもそれが全世界に晒されてしまうわけで、かなり戸惑わせてしまうことになる。また、リリースをするということはそれに番号を振るということであり、番号をつけておけばバグレポートに対して変更を追いかけやすくなる。もちろんこのやりかたで進めるにはリリース作業が簡単でなければならないが、常にそうだとは限らない。

sendmail がクリティカルな本番環境で使われるようになり始めると、この方式では問題が出てきた。私が行つた変更が「みんなにちょっと試してもらいたいもの」なのか「本番環境に適用して欲しいもの」なのか、それが通じないことが出てきたのだ。リリースするときに“alpha”とか“beta”とか名付けておけばいくらかましになるが、それでも問題は解決しない。その結果どうなつたかというと、sendmail が成熟するにつれてリリースの頻度が下がつてリリースあたりの変更が大きくなつた。これが特に問題となつたのは、sendmail が営利企業に

⁵<http://milter.org>

⁶<http://postfix.org>

組み込まれたときだった。顧客は最新のイケてるバージョンを望むが、同時に安定版も望む。そして、その二つが両立しないという事実を認めようとしないのだ。

この手のオープンソース開発者のニーズと商用製品のニーズとの対立は、決してなくならない。早めに頻繁にリリースすることには多くのメリットがある。特に、勇気のある(時に無謀な)テスターを多く獲得できるのが大きい。彼らによるテストは、標準的な開発環境では決して再現できないだろう。しかし、プロジェクトがうまく進むとそれは徐々に製品に姿を変える傾向がある(オープンソースであろうとフリーであろうと関係ない)。そして製品になってしまふとそのニーズはプロジェクトのニーズとは変わってくる。

17.6 セキュリティ

sendmail のこれまでの生き様は、セキュリティの面では波乱に満ちたものだった。さまざまな波乱の中には、起きてしかるべきものもあればそうでないものもあった。そして我々が考える“セキュリティ”的概念も変わってきた。インターネットが始まった頃のユーザー数はたかだか数千人程度で、その大半は学術研究に関わる人たちだった。古き良き時代。いろんな意味で、今のインターネットよりも親切で紳士的だった。ネットワークの設計は情報の共有を推奨する作りになっていたし、ファイヤウォールを構築するなどという考えはなかった(そんな概念は、そもそも初期のインターネットには存在しなかった)。今やネット上には危険がいっぱい。悪意に満ちた場所となり、スパマーやクラッカーがそこらじゅうにあふれている。インターネットが紛争地帯にたとえられることも増えてきた。紛争地帯には、民間人の犠牲者もつきものだ。

ネットワークサーバーをセキュアに書くのはとても難しい。プロトコルがほんの少しでも複雑になると、なおさらだ。どんなプログラムだって、少なくとも些細な問題は抱えている。一般的な TCP/IP の実装でさえ、外部からのアタックを受けてしまった。より上位レベルの実装言語になると、万能薬は存在しない。そしてその言語自身が脆弱性を作ることもある。必ずといっていいほど見るフレーズが、どこから来たものかにかかわらず“すべての入力は疑ってかかれ”というものだ。疑うべき入力には、二次的な入力も含まれる。たとえば DNS サーバーや milter からの入力もそうだ。初期のネットワークソフトウェアの大半がそうだったのだが、sendmail の初期のバージョンでは入力を信頼しすぎていた。

しかし、sendmail で最大の問題だったのは、初期のバージョンが root 権限で動作していたという点だ。root 権限を要した理由は、SMTP をリスンするソケットを開いたり各ユーザーの転送する情報を読んだり、各ユーザーのメールボックスやホームディレクトリにメッセージを配達したりするためにだった。しかし、今どきのシステムの大半では、メールボックスの名前とシステム上のユーザーの概念とが切り離されている。そのため、SMTP をリスンするソケットを開く以外の操作では root 権限は事实上不要となった。現在の sendmail では、接続を処理する前に root 権限を放棄できるようになっている。これをサポートしている環境なら、root 権限に関する問題は気にしなくてもよくなつた。さらに、ユーザーのメールボック

スへの直接の配送をしないシステム上では、sendmail を chroot 環境で動かすこともできる。そうすれば、さらに権限を隔離できる。

不幸にも、sendmail のセキュリティが貧弱だという評判が広まるにつれて、まったく関係のない問題までも sendmail のせいにされることが出てきた。たとえば、あるシステム管理者は、自分が/etc ディレクトリに書き込み権限を与えておきながら、誰かが/etc/passwd ファイルを書き換えたときにそれを sendmail のせいにしたりした。そうした事件を経て、我々もセキュリティについてより真剣に考えるようになった。そして、sendmail がアクセスするファイルやディレクトリの所有者やモードを、明示的にチェックするようにした。このチェックは非常に厳しいものだったので、DontBlameSendmail オプションを用意してこのチェックを無効化できるようにもした。

それ以外の観点からのセキュリティ問題もあった。プログラム自身のアドレス空間を守るのとは直接関係しないものだ。たとえば、spam が増加するにつれて出てきた、メールアドレス収集に関する問題がこれにあたる。SMTP の VRFY コマンドや EXPN コマンドはそれぞれ、個別のアドレスを検証したりメーリングリストのメンバーを展開したりするために用意されたコマンドである。スパマーたちに悪用されることがあまりにも多発したので、ほとんどのサイトでは今やこのコマンドは無効化されている。少なくとも VRFY に関しては、これは残念なことだ。このコマンドは、アンチスパムエージェントが送信者のアドレスを検証するときに使うこともあるからである。

同様に、ウィルス対策の保護もかつてはデスクトップ側の問題とみられていた。しかし、その重要性が増すにつれて、商用レベルの MTA ではウィルス対策のチェックができることが当然になってきた。その他、最近の設定でセキュリティ関連の必須要件となったものとしては、重要なデータを強制的に暗号化させることやデータの喪失に対する保護、HIPPA などの法的規制への対応などがある。

初期の sendmail が最重要視していたのは信頼性だった。つまり、あらゆるメッセージをきちんと配達する（あるいは送信者に差し戻す）ということだ。しかしじョージョブ問題（発信元アドレスを偽装したメールによる攻撃。多くの人はセキュリティの問題ととらえている）のせいで、多くのサイトがバウンスマッセージの作成を無効にしてしまった。もし SMTP 接続が開いている間に失敗を検出できれば、サーバー側で SMTP コマンドを失敗させることで何か問題が起こったことを伝えられる。しかし、SMTP 接続が閉じてしまった後だと、アドレスが間違っていたメッセージはただ黙って消えてしまうだけになってしまう。最近のメールの大半は 1 ホップで届くので、何か問題があればわかるだろう。しかし、少なくとも理屈上は、この世界では信頼性よりセキュリティを重視するようになったということだ。

17.7 Sendmail の進化

激動の環境でソフトウェアが生き残り続けるには、環境の変化にあわせて自らを進化させなければならない。新たなハードウェア技術が登場すればそれにあわせて OS も変化するし、

OS が変わればライブラリやフレームワークも変わる。つまりそれはアプリケーションにも変化を促すことになる。アプリケーションが存続すればするほど、問題のある環境で使われることも多くなる。変更は避けられない。生き延びるために変更を受け入れてそれを取り込まねばならない。このセクションでは、これまで sendmail に起こった変化の中で重要なものをいくつか取り上げる。

設定はより冗長に

当初の sendmail の設定は、極めて簡潔だった。たとえば、オプションやマクロの名前はすべて一文字だった。その理由は次の三つである。まず第一に、そうすればパース処理をシンプルにできるということ (16 ビット環境ではこれが重要だった)。第二に、オプションの数がそれほど多くなかったこと。一文字のオプションを考えるのにそれほど苦労しなかった。第三に、単一文字の規約が既にコマンドラインのフラグで確立されていたこと。

同様に、書き換えルールセットも、当初は名前ではなく番号で指定していた。ルールセットの数が少ないいうちは、これでもまだ耐えられた。しかし、その数が増えてくると、よりわかりやすい名前をつけることが大切になった。

sendmail の稼働環境が複雑になるにつれ、また同時に 16 ビット環境が去りゆくにつれ、よりリッチな設定方法の必要性が明らかになってきた。幸いにも、後方互換性を維持したままでの変更が可能だった。この変更によって、設定ファイルのわかりやすさが劇的に向上した。

他のサブシステムとの接続: さらなる統合

sendmail が書かれたころは、メールシステムといえば大抵は OS の他の部分から隔離されているものだった。統合を要するサービスはごく一部で、たとえば /etc/passwd や /etc/hosts といったファイルくらいだった。サービスを斬り合える機能はまだ発明されていなかつたし、ディレクトリサービスも存在しなかつた。そして設定ファイルはまだ小さく、手で書けるレベルだった。

状況はすぐに変わった。最初に追加した中のひとつが DNS だ。システムが持つホストルックアップの抽象化機能 (gethostbyname) は、IP アドレスを探すには充分だったが、メールでは、それ以外にも MX などを問い合わせる必要があったのだ。後に、IDA sendmail では外部データベースを使ったルックアップ機能を組み込んだ。このデータベースには dbm(3) ファイルを使った。sendmail 8 ではそれをさらに拡張し、汎用的なマッピングサービスを用意して他の形式のデータベースも使えるようにした。外部のデータベースも使えるし、内部での変換も使える。内部での変換は、書き換え機能 (アドレスのクオート解除など) なしには実現できなかっただろう。

今やメールシステムは多数の外部サービスに依存しており、一般的に、電子メール専用として設計されることなくなつた。そのため、sendmail のコードもより抽象化する方向に進

んだ。そのおかげで、メールシステムの開発や保守はより難しいものとなった。“可動部品”がいろいろ増えたからである。

ギスギスした世界

sendmail の開発が始まったころの世界は、現在の常識からするとまったく異質の世界に見えることだろう。初期のネットワークにかかわっていた人の大半は研究者で、比較的温和な人が多かった。学術思想についての争いでたちの悪い振る舞いが起こることもあったが、すぐに収まった。sendmail は当時の世情を反映して作られており、メール配送の信頼性を上げることを最重要視していた。たとえユーザー側で間違いがあったとしても、可能な限りメールを配送できるように心がけた。

今の世の中は、当時に比べてはるかにギスギスしている。飛び交う電子メールの大半は、悪意のあるものだ。MTA が目指すゴールは、メールをきちんと配送することから悪意のあるメールを排除することに変わった。いまどきの MTA で最重要視されるのは、きっとフィルタリング機能だろう。それに対応するため、sendmail にも数多くの変更が必要となった。

一例として、数多くのルールセットが追加された。これらを使って SMTP コマンドのパラメータをチェックし、問題を早期に発見できるようになった。エンベロープを読んだ段階でメッセージを拒否するほうが、メッセージ全体を読んでから拒否するよりもはるかにコストが低くなる。ましてや、メッセージの配送を受け入れた後で拒否するようにするとずっと高くついてしまう。初期のフィルタリングは、メッセージをいったん受け入れた後でフィルタプログラムに渡し、フィルタを通過したものだけを別の sendmail インスタンスに送るという仕組みだった(いわゆる“サンドウィッチ”構成である)。いま同じようなことをすれば、非常にコストがかかることだろう。

また、sendmail での TCP/IP 接続の使い方も変わった。もともとはごく標準的に使っているだけだったが、最近はより洗練されている。ネットワーク入力を“のぞき見”して、前のコマンドが受理されていないのに送信者が次のコマンドを送信していないかどうかを調べたりもする。このため、sendmail が複数のネットワークを扱えるように作られた抽象化の中にはうまく動かないものも出てきた。今でも、sendmail がたとえば XNS や DECnet のネットワークにも接続できるようにかなりの作業をしている。しかし、TCP/IP に固有の知見が多くのコードに取り込まれてしまっている。

ギスギスした世界に立ち向かうために、さまざまな設定項目が追加された。アクセステーブルへの対応やリアルタイムブラックリストへの対応、アドレス収集対策、DoS 対策、そして spam のフィルタリングなどだ。そのおかげでメールシステムの設定はかなり複雑な作業になってしまったが、今の世の中で生きていくためにはそうするしかなかった。

新技術の採用

新たな標準規格が続々と誕生し、それらもまた sendmail に大きく手を加える要因となった。たとえば TLS(暗号化) を追加するときには、大半のコードを変更しなければならなかつた。SMTP のパイプラインを実現するには、ローレベルの TCP/IP ストリームを注視してデッドロックを回避する必要があつた。サブミッションポート(587)をサポートするには、複数の入力ポートを待ち受ける機能が必要となつた。そして、ポートによって挙動を変えなければならなかつた。

それ以外にも、標準規格ではなくその場の状況に押されて追加した機能もある。たとえば、milter インターフェイスを追加したのは spam に耐えられなくなつたからだ。milter は標準規格として確立されたものではなかつたが、大きな新技術だつた。

どの場合についても、これらの変更によって何らかの面でメールシステムは強化された。セキュリティの向上やパフォーマンスの向上、あるいは新機能の追加などによってである。しかし、どの場合についてもそれなりのコストがかかつており、ほぼすべての変更がコードや設定ファイルを複雑化させている。

17.8 もし今やりなおせるとしたら?

後からなら何とでも言える。というわけで、今だったら違うやりかたにしただらうということもたくさんある。当時としては予測不能だったこと(spam のせいでメールに対する視点がどれほど変わるかや、最新のツール軍がどんなものになるかなど)もあれば、どう見ても予測できたはずだらうということもある。また、sendmail を書いている間に私自身さまざまことを学んだ。電子メールのことや TCP/IP のこと、そしてプログラミング自体のことなど。誰もがコードとともに成長するということだ。

しかし、今でも同じようにするだらうということもたくさんあって、その中には一般的な常識に反するものもある。

違うやりかたにしたいところ

おそらく sendmail 史上最大の過ちは、後に自分がどれだけ重要な存在になるのかを早期に気付けなかつたことだらう。世界が正しい方向に向かうようほんの少し突っつける機会が何度かあつたはずなのに、そうしなかつた。実際、たとえば sendmail の入力チェックを厳しくして不正な入力を拒否するようにもできたはずだが、そうすべきときに実際は何もしなかつた。同じく、設定ファイルの構文に改善の余地があることもわかつっていた。まだ世界中で数百インスタンス程度しか sendmail が動いていなかつたころの話だ。わかつてはいたのだが、結局そのときは変更しなかつた。その時点での既存ユーザーに負担を強いることになるからである。今思えば、何かを改善するなら早めにしておくべきだつた。一時的につらいこともあるが、長い目で見ればよりよい結果になつただらう。

バージョン 7 の Mailbox の構文

その一例が、バージョン 7 のメールボックスでのメッセージの分割方法である。当時は “From_U” (“_U” は ASCII の空白文字、つまり 0x20) で始まる行でメッセージを分割していた。もしメッセージ本文の中に “From_U” で始まる行があれば、ローカルのメールボックスソフトウェアはそれを “>From_U” に変換していた。いくつかのシステムでは「その前に空行を含むこと」を要件とできたが、すべてではなかったのでそれに依存はできなかつた。今日に至るまで、“>From” はありとあらゆる箇所で予期せず登場している。それはメールとは直接関係のない(しかし、いつだったかそれがメールで処理されたことがあった)場面も含む。今思えば、BSD メールシステムから別の形式の構文に変換したほうがよかつた。変更した直後には多くの人から恨まれるだろうが、そうしておけばさまざまな問題からこの世界を救えたはずだ。

設定ファイルの構文とその中身

おそらく、設定ファイルの構文における最大のミスは、書き換えルールの記述でパターンと置換内容の区切りにタブ (HT, 0x09) を採用したことだろう。その当時は、make の流儀をまねたのだった。しかしその数年後に make の作者である Stuart Feldman に聞いた話によると、彼もまたタブを採用したことをいちばん後悔しているとのことだった。設定ファイルを画面上で見たときにタブがあるかどうかがわかりにくくいうだけでなく、大半のウインドウシステムではカットアンドペーストでタブが消えてしまうという問題もあった。

書き換えルールという考え方自体は間違っていなかつたと思っている(以下を参照)。しかし、設定ファイルの全体構造はもう少しなんとかできただろう。たとえば、設定ファイルで階層構造が必要になることを想定できていなかつた(SMTP リスナーのポートごとに異なるオプションを設定したりなど)。当時、設定ファイルを設計するにあたっては、“標準” フォーマットなど存在しなかつた。今なら Apache 形式の設定構文を採用するだろう。すっきりしているし、充分な表現力もある。あるいは、Lua などの組み込み言語を使ったかもしれない。

sendmail の開発が進んでいたころは、アドレス空間も小さいしプロトコルもまだ流動的だつた。可能な限り設定ファイル側に押し出しておくのが無難だったので。今の状況では、これは間違いだろう。今や MTA は広大なアドレス空間を使えるし、標準規格もきちんと固まっている。さらに、“設定ファイル”的一部は実際のところプログラムのコードのようになってしまっており、新しいリリースが出るたびに更新が必要になるほどだ。.mc 設定ファイルを使えばこの問題は解決できるが、ソフトウェアを更新するたびに設定ファイルをビルドしなおす必要があるというのもつらいことだ。この問題に対するシンプルな対策は、sendmail が読む設定ファイルを二つに分割することだろう。一方は利用者から見えないようにしてソフトウェアの更新時に自動的にインストールされるようにし、もう一方を公開してローカル設定用に使わせるという方法だ。

ツールの活用

今ではさまざまな新しいツールが登場している。たとえばソフトウェアの構成やビルドひとつとってもそうである。必要に応じてツールを活用すればその力を生かせるだろうが、時にやり過ぎてしまうこともある。そうなれば、システムを理解するが必要以上に困難になってしまう。たとえば、単に `strtok(3)` があれば充分なときにわざわざ `yacc(1)` の構文を使うのはばかげている。しかし、車輪の再発明をするのもあまりよい考えではない。例を挙げると、私はよっぽどの場合を除いて `autoconf` を使うようにしている。

後方互換性

もし将来の姿が見えていて、`sendmail` がいかに普及するかがわかっていたなら、開発初期の段階で既存の環境との互換性を崩してしまうことを躊躇しなかつただろう。もし既存の習慣がうまくいかなくなってしまうのならそれを修正すべきであって、なんとか対応させることのではいけない。とはいえる、私はまだメッセージフォーマットの厳格なチェックをしていない。単に無視できたり簡単に手直しできたりするような問題もあるからである。たとえば、`Message-Id:` ヘッダーがないメッセージには今でもきっとヘッダーを追加するだろう。しかし、`From:` ヘッダーがないメッセージについては、わざわざエンベロープの情報からヘッダーを作るよりもそのメッセージを拒否してしまいたい。

内部的な抽象化

内部的な抽象化の中には、今ならそんなことはしないだろうというものもある。また、当時はそうしなかったが、今なら抽象化するだろうというものもある。たとえば、`null` 終端文字列を長さ/値のペアのかわりに使うことはしないだろう。こうしたほうが標準 C ライブライアリで使いやすくなることはわかっていても、である。セキュリティに関する問題ひとつとっても、`null` 終端文字列を使わない意味がある。逆に言うと、例外処理を C で書こうとは思わない。しかし、一貫性のあるステータスコード体系は作って、それを使うようにするだろう。ルーチンの返り値が `null` や `false` あるいは負の数だったらエラーを意味するなどということはない。

メールボックスの名前を Unix のユーザー id から切り離すという抽象化は、今ならきっと行うだろう。`sendmail` を書いていた当時は、Unix のユーザーにメールを送ることしか想定していなかった。今やそんな前提は成立しない。仮に Unix のユーザー管理と同じモデルのシステムであっても、決してメールを受け取ることのないシステムアカウントが存在する。

同じようにしたいところ

もちろん、**うまくいってたこと** だってあったんだよ…

Syslog

sendmail の派生プロジェクトの中でも成功したうちのひとつが、syslog である。sendmail が書かれた当時は、プログラムからログを出力しようとすれば、何かファイルを作つてそこに書き込むしかなかった。その手のログファイルがファイルシステム上に散乱していたのだ。当時は syslog に書き込むのはなかなか難しかった（まだ UDP は存在しなかったので、mpx ファイルとかいうものを使っていた）が、よくやってくれた。しかし、一ヵ所だけ変更したいところがある。ログに記録されるメッセージの構文にもっと注意を払い、機械可読性を向上させたい。当時の私には、ログ監視ツールの登場を見できなかつたのだ。

書き換えルール

書き換えルールにはいろいろ問題もあるが、今でもきっと採用するだろう（今使われているほど多くはならないだろうが）。タブ文字を使ったことは大きな間違いだった。しかし、ASCII の制約やメールアドレスの構文を考慮すると、何らかのエスケープ文字は必要となる⁷。一般に、パターン置換のパラダイムはうまく機能するし、非常に柔軟である。

ツールに頼りすぎない

先ほどは「もっと既存のツールを活用する」と書いたが、いまどきのランタイムライブラリの多くは、あまり使いたいとは思わない。私見だが、多くのライブラリは肥大化しすぎて危険になっているように感じる。ライブラリの選択は慎重に行うべきだ。再利用することにメリットと、必要以上に高機能なツールを使うことによる問題とのトレードオフになる。これだけは避けようと思っているツールのひとつが XML で、少なくとも設定ファイルを XML にしようとは思わない。XML ファイルの構文は、設定の記述に使うにはごてごてしそうでいる。もちろん XML が役立つ場面もあるのだろうが、現状は必要以上に使われすぎている。

コードは C で書く

もっと自然に書ける言語、たとえば Java とか C++ を使えばいいのではないかと助言してくれる人もいる。C 言語にはいろいろ問題もあるが、それでも私は実装言語に C を使うだろう。まあ個人的な理由もある。C のほうが、Java や C++ よりもよく知っているからだ。しかしそれだけではない。いまどきのオブジェクト指向言語にはがつかりさせられているのだ。その多くはメモリ管理に無頓着すぎて、無尽蔵にメモリを使いつぎでいる。メモリを確保するときにはパフォーマンス上の問題もいろいろ考慮しなければならないのだが、それはここでは書ききれない。sendmail は、内部的にはオブジェクト指向の概念を採用しているところもある（たとえば、マップクラスなど）が、個人的な意見としては、すべてオブジェクト指向化してしまうのは無駄が多く、制約をかけすぎだと思っている。

⁷ 設定ファイルで Unicode を使うことはあまり広まらないだろうと思っている。

17.9 まとめ

sendmail MTA が誕生したころの世界は、限りなく混乱していた。まるで“西部開拓時代”のようなものだ。電子メールはアドホックな仕組みとしてしか存在しなかつたし、今のような標準規格もまだ正式には定まっていなかった。この 31 年の間に“電子メールの問題”の種類も変わった。昔は単に、巨大なメッセージを高負荷な環境でいかにきちんと配達するかというのが問題だったのだが、徐々に spam 対策やウィルス対策のほうが大切になってきた。今や、電子メールの活躍の場はさらに広がっている。電子メールベースのアプリケーションのプラットフォームとして使われることもある。sendmail はこの世界の主力製品に成長した。リスク管理にうるさい企業でさえも、今や電子メールを受け入れている。電子メールは、單なるテキストベースでの一対一のコミュニケーションツールではなく、ミッションクリティカルな部分を担うマルチメディアベースのツールに成長したのだ。

なぜここまで成功できたのか、よくわからない面もある。激動の世界で生き残り続け、かつ成長していくプログラムを、ごく少数のパートタイムの開発者で作り上げるということ。これは、きちんとまとまったソフトウェア開発の方法論があつてこそ実現できたことだ。sendmail の成功の要因について、少しでもみなさんにお伝えいただろうか。

SnowFlock

Roy Bryant and Andrés Lagar-Cavilla

Cloud computing provides an attractively affordable computing platform. Instead of buying and configuring a physical server, with all the associated time, effort and up front costs, users can rent “servers” in the cloud with a few mouse clicks for less than 10 cents per hour. Cloud providers keep their costs low by providing virtual machines (VMs) instead of physical computers. The key enabler is the virtualization software, called a virtual machine monitor (VMM), that emulates a physical machine. Users are safely isolated in their “guest” VMs, and are blissfully unaware that they typically share the physical machine (“host”) with many others.

18.1 Introducing SnowFlock

Clouds are a boon to agile organizations. With physical servers, users are relegated to waiting impatiently while others (slowly) approve the server purchase, place the order, ship the server, and install and configure the Operating System (OS) and application stacks. Instead of waiting weeks for others to deliver, the cloud user retains control of the process and can create a new, standalone server in minutes.

Unfortunately, few cloud servers stand alone. Driven by the quick instantiation and pay-per-use model, cloud servers are typically members of a variable pool of similarly configured servers performing dynamic and scalable tasks related to parallel computing, data mining, or serving web pages. Because they repeatedly boot new instances from the same, static template, commercial clouds fail to fully deliver on the promise of true on-demand computation. After instantiating the server, the cloud user must still manage cluster membership and broker the addition of new servers.

SnowFlock addresses these issues with VM Cloning, our proposed cloud API call. In the same way that application code routinely invokes OS services through a syscall interface, it could now also invoke cloud services through a similar interface. With SnowFlock’s VM Cloning, resource

allocation, cluster management, and application logic can be interwoven programmatically and dealt with as a single logical operation.

The VM Cloning call instantiates multiple cloud servers that are identical copies of the originating parent VM up to the point of cloning. Logically, clones inherit all the state of their parent, including OS- and application-level caches. Further, clones are automatically added to an internal private network, thus effectively joining a dynamically scalable cluster. New computation resources, encapsulated as identical VMs, can be created on-the-fly and can be dynamically leveraged as needed.

To be of practical use, VM cloning has to be applicable, efficient, and fast. In this chapter we will describe how SnowFlock’s implementation of VM Cloning can be effectively interwoven in several different programming models and frameworks, how it can be implemented to keep application runtime and provider overhead to a minimum, and how it can be used to create dozens of new VMs in five seconds or less.

With an API for the programmatic control of VM Cloning with bindings in C, C++, Python and Java, SnowFlock is extremely flexible and versatile. We’ve successfully used SnowFlock in prototype implementations of several, quite different, systems. In parallel computation scenarios, we’ve achieved excellent results by explicitly cloning worker VMs that cooperatively distribute the load across many physical hosts. For parallel applications that use the Message Passing Interface (MPI) and typically run on a cluster of dedicated servers, we modified the MPI startup manager to provide unmodified applications with good performance and much less overhead by provisioning a fresh cluster of clones on demand for each run. Finally, in a quite different use case, we used SnowFlock to improve the efficiency and performance of elastic servers. Today’s cloud-based elastic servers boot new, cold workers as needed to service spikes in demand. By cloning a running VM instead, SnowFlock brings new workers online 20 times faster, and because clones inherit the warm buffers of their parent, they reach their peak performance sooner.

18.2 VM Cloning

As the name suggests, VM clones are (nearly) identical to their parent VM. There are actually some minor but necessary differences to avoid issues such as MAC address collisions, but we’ll come back to that later. To create a clone, the entire local disk and memory state must be made available, which brings us to the first major design tradeoff: should we copy that state up-front or on demand?

The simplest way to achieve VM cloning is to adapt the standard VM “migration” capability. Typically, migration is used when a running VM needs to be moved to a different host, such as when the host becomes overloaded or must be brought down for maintenance. Because the VM is purely software, it can be encapsulated in a data file that can then be copied to a new, more appropriate host, where it picks up execution after a brief interruption. To accomplish this, off-the-shelf VMMs

create a file containing a “checkpoint” of the VM, including its local filesystem, memory image, virtual CPU (VCPU) registers, etc. In migration, the newly booted copy replaces the original, but the process can be altered to produce a clone while leaving the original running. In this “eager” process, the entire VM state is transferred up front, which provides the best initial performance, because the entire state of the VM is in place when execution begins. The disadvantage of eager replication is that the laborious process of copying the entire VM must happen before execution can begin, which significantly slows instantiation.

The other extreme, adopted by SnowFlock, is “lazy” state replication. Instead of copying everything the VM might ever need, SnowFlock transfers only the vital bits needed to begin execution, and transfers state later, only when the clone needs it. This has two advantages. First, it minimizes the instantiation latency by doing as little work as possible up front. Second, it increases the efficiency by copying only the state that is actually used by the clone. The yield of this benefit, of course, depends on the clone’s behavior, but few applications access every page of memory and every file in the local filesystem.

However, the benefits of lazy replication aren’t free. Because the state transfer is postponed until the last moment, the clone is left waiting for state to arrive before it can continue execution. This situation parallels swapping of memory to disk in time-shared workstation: applications are blocked waiting for state to be fetched from a high latency source. In the case of SnowFlock, the blocking somewhat degrades the clone’s performance; the severity of the slowdown depends on the application. For high performance computing applications we’ve found this degradation has little impact, but a cloned database server may perform poorly at first. It should be noted that this is a transient effect: within a few minutes, most of the necessary state has been transferred and the clone’s performance matches that of the parent.

As an aside, if you’re well versed in VMs, you’re likely wondering if the optimizations used by “live” migration are useful here. Live migration is optimized to shorten the interval between the original VM’s suspension and the resumption of execution by the new copy. To accomplish this, the Virtual Machine Monitor (VMM) pre-copies the VM’s state while the original is still running, so that after suspending it, only the recently changed pages need to be transferred. This technique does not affect the interval between the migration request and the time the copy begins execution, and so would not reduce the instantiation latency of eager VM cloning.

18.3 SnowFlock’s Approach

SnowFlock implements VM cloning with a primitive called “VM Fork”, which is like a standard Unix fork, but with a few important differences. First, rather than duplicating a single process, VM Fork duplicates an entire VM, including all of memory, all processes and virtual devices, and the local filesystem. Second, instead of producing a single copy running on the same physical host,

VM Fork can simultaneously spawn many copies in parallel. Finally, VMs can be forked to distinct physical servers, letting you quickly increase your cloud footprint as needed.

The following concepts are key to SnowFlock:

- Virtualization: The VM encapsulates the computation environment, making clouds and machine cloning possible.
- Lazy Propagation: The VM state isn't copied until it's needed, so clones come alive in a few seconds.
- Multicast: Clone siblings have similar needs in terms of VM state. With multicast, dozens of clones start running as quickly as one.
- Page Faults: When a clone tries to use missing memory, it faults and triggers a request to the parent. The clone's execution is blocked until the needed page arrives.
- Copy on Write (CoW): By taking a copy of its memory and disk pages before overwriting them, the parent VM can continue to run while preserving a frozen copy of its state for use by the clones.

We've implemented SnowFlock using the Xen virtualization system, so it's useful to introduce some Xen-specific terminology for clarity. In a Xen environment, the VMM is called the hypervisor, and VMs are called domains. On each physical machine (host), there is a privileged domain, called "domain 0" (dom0), that has full access to the host and its physical devices, and can be used to control additional guest, or "user", VMs that are called "domain U" (domU).

In broad strokes, SnowFlock consists of a set of modifications to the Xen hypervisor that enable it to smoothly recover when missing resources are accessed, and a set of supporting processes and systems that run in dom0 and cooperatively transfer the missing VM state, and some optional modifications to the OS executing inside clone VMs. There are six main components.

- VM Descriptor: This small object is used to seed the clone, and holds the bare-bones skeleton of the VM as needed to begin execution. It lacks the guts and muscle needed to perform any useful work.
- Multicast Distribution System (`mcdist`): This parent-side system efficiently distributes the VM state information simultaneously to all clones.
- Memory Server Process: This parent-side process maintains a frozen copy of the parent's state, and makes it available to all clones on demand through `mcdist`.
- Memtap Process: This clone-side process acts on the clone's behalf, and communicates with the memory server to request pages that are needed but missing.
- Clone Enlightenment: The guest kernel running inside the clones can alleviate the on-demand transfer of VM state by providing hints to the VMM. This is optional but highly desirable for efficiency.
- Control Stack: Daemons run on each physical host to orchestrate the other components and manage the SnowFlock parent and clone VMs.



図 18.1: SnowFlock VM Replication Architecture

Pictorially speaking, 図 18.1 depicts the process of cloning a VM, showing the four main steps: (1) suspending the parent VM to produce an architectural descriptor; (2) distributing this descriptor to all target hosts; (3) initiating clones that are mostly empty state-wise; and (4) propagating state on-demand. The figure also depicts the use of multicast distribution with `mcdist`, and fetch avoidance via guest enlightenment.

If you're interested in trying SnowFlock, it's available in two flavors. The documentation and open source code for the original University of Toronto SnowFlock research project are available¹. If you'd prefer to take the industrial-strength version for a spin, a free, non-commercial license is available

¹<http://sysweb.cs.toronto.edu/projects/1>

from GridCentric Inc.² Because SnowFlock includes changes to the hypervisor and requires access to dom0, installing SnowFlock requires privileged access on the host machines. For that reason, you'll need to use your own hardware, and won't be able to try it out as a user in a commercial cloud environment such as Amazon's EC2.

Throughout the next few sections we'll describe the different pieces that cooperate to achieve instantaneous and efficient cloning. All the pieces we will describe fit together as shown in 図 18.2.

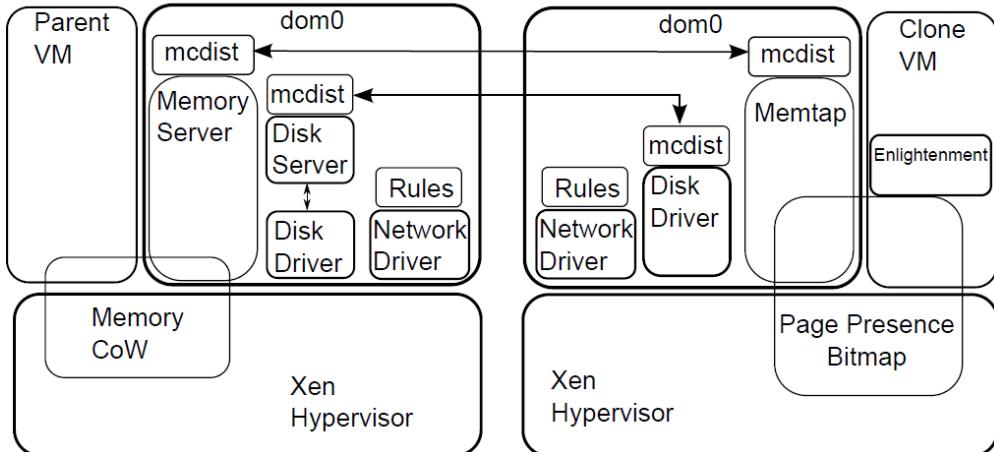


図 18.2: Software Components of SnowFlock

18.4 Architectural VM Descriptor

The key design decision for SnowFlock is to postpone the replication of VM state to a lazy runtime operation. In other words, copying the memory of a VM is a late binding operation, allowing for many opportunities for optimization.

The first step to carry out this design decision is the generation of an architectural descriptor of the VM state. This is the seed that will be used to create clone VMs. It contains the bare minimum necessary to create a VM and make it schedulable. As the name implies, this bare minimum consists of data structures needed by the underlying architectural specification. In the case of SnowFlock, the architecture is a combination of Intel x86 processor requirements and Xen requirements. The architectural descriptor thus contains data structures such as page tables, virtual registers, device metadata, wallclock timestamps, etc. We refer the interested reader to [LCWB⁺11] for an in-depth description of the contents of the architectural descriptor.

²<http://www.gridcentriclabs.com/architecture-of-open-source-applications>

An architectural descriptor has three important properties: First, it can be created in little time; 200 milliseconds is not uncommon. Second, it is small, typically three orders of magnitude smaller than the memory allocation of the originating VM (1 MB for a 1 GB VM). And third, a clone VM can be created from a descriptor in less than a second (typically 800 milliseconds).

The catch, of course, is that the cloned VMs are missing most of their memory state by the time they are created from the descriptor. The following sections explain how we solve this problem—and how we take advantage of the opportunities for optimization it presents.

18.5 Parent-Side Components

Once a VM is cloned it becomes a parent for its children or clones. Like all responsible parents, it needs to look out for the well-being of its descendants. It does so by setting up a set of services that provision memory and disk state to cloned VMs on demand.

Memserver Process

When the architectural descriptor is created, the VM is stopped in its tracks throughout the process. This is so the VM memory state settles; before actually pausing a VM and descheduling from execution, internal OS drivers quiesce into a state from which clones can reconnect to the external world in their new enclosing VMs. We take advantage of this quiescent state to create a “memory server”, or `memserver`.

The memory server will provide all clones with the bits of memory they need from the parent. Memory is propagated at the granularity of an x86 memory page (4 kbytes). In its simplest form, the memory server sits waiting for page requests from clones, and serves one page at a time, one clone at a time.

However, this is the very same memory the parent VM needs to use to keep executing. If we would allow the parent to just go ahead and modify this memory, we would serve corrupted memory contents to clone VMs: the memory served would be different from that at the point of cloning, and clones would be mightily confused. In kernel hacking terms, this is a sure recipe for stack traces.

To circumvent this problem, a classical OS notion comes to the rescue: Copy-on-Write, or CoW memory. By enlisting the aid of the Xen hypervisor, we can remove writing privileges from all pages of memory in the parent VM. When the parent actually tries to modify one page, a hardware page fault is triggered. Xen knows why this happened, and makes a copy of the page. The parent VM is allowed to write to the original page and continue execution, while the memory server is told to use the copy, which is kept read-only. In this manner, the memory state at the point of cloning remains frozen so that clones are not confused, while the parent is able to proceed with execution. The overhead of CoW is minimal: similar mechanisms are used by Linux, for example, when creating new processes.

Multicasting with Mcdist

Clones are typically afflicted with an existential syndrome known as “fate determinism.” We expect clones to be created for a single purpose: for example, to align X chains of DNA against a segment Y of a database. Further, we expect a set of clones to be created so that all siblings do the same, perhaps aligning the same X chains against different segments of the database, or aligning different chains against the same segment Y. Clones will thus clearly exhibit a great amount of temporal locality in their memory accesses: they will use the same code and large portions of common data.

We exploit the opportunities for temporal locality through `mcdist`, our own multicast distribution system tailored to SnowFlock. Mcdist uses IP multicast to simultaneously distribute the same packet to a set of receivers. It takes advantage of network hardware parallelism to decrease the load on the memory server. By sending a reply to all clones on the first request for a page, each clone’s requests act as a prefetch for its siblings, because of their similar memory access patterns.

Unlike other multicast systems, `mcdist` does not have to be reliable, does not have to deliver packets in an ordered manner, and does not have to atomically deliver a reply to all intended receivers. Multicast is strictly an optimization, and delivery need only be ensured to the clone explicitly requesting a page. The design is thus elegantly simple: the server simply multicasts responses, while clients time-out if they have not received a reply for a request, and retry the request.

Three optimizations specific to SnowFlock are included in `mcdist`:

- Lockstep Detection: When temporal locality does happen, multiple clones request the same page in very close succession. The `mcdist` server ignores all but the first of such requests.
- Flow Control: Receivers piggyback their receive rate on requests. The server throttles its sending rate to a weighted average of the clients’ receive rate. Otherwise, receivers will be drowned by too many pages sent by an eager server.
- End Game: When the server has sent most pages, it falls back to unicast responses. Most requests at this point are retries, and thus blasting a page through the wire to all clones is unnecessary.

Virtual Disk

SnowFlock clones, due to their short life span and fate determinism, rarely use their disk. The virtual disk for a SnowFlock VM houses the root partition with binaries, libraries and configuration files. Heavy data processing is done through suitable filesystems such as HDFS (第 9 章) or PVFS. Thus, when SnowFlock clones decide to read from their root disk, they typically have their requests satisfied by the kernel filesystem page cache.

Having said that, we still need to provide access to the virtual disk for clones, in the rare instance that such access is needed. We adopted the path of least resistance here, and implemented the disk

by closely following the memory replication design. First, the state of the disk is frozen at the time of cloning. The parent VM keeps using its disk in a CoW manner: writes are sent to a separate location in backing storage, while the view of the disk clones expect remains immutable. Second, disk state is multicast to all clones, using `mcdist`, with the same 4 KB page granularity, and under the same expectations of temporal locality. Third, replicated disk state for a clone VM is strictly transient: it is stored in a sparse flat file which is deleted once the clone is destroyed.

18.6 Clone-Side Components

Clones are hollow shells when they are created from an architectural descriptor, so like everybody else, they need a lot of help from their parents to grow up: the children VMs move out and immediately call home whenever they notice something they need is missing, asking their parent to send it over right away.

Memtap Process

Attached to each clone after creation, the `memtap` process is the lifeline of a clone. It maps all of the memory of the clone and fills it on demand as needed. It enlists some crucial bits of help from the Xen hypervisor: access permission to the memory pages of the clones is turned off, and hardware faults caused by first access to a page are routed by the hypervisor into the `memtap` process.

In its simplest incarnation, the `memtap` process simply asks the memory server for the faulting page, but there are more complicated scenarios as well. First, `memtap` helpers use `mcdist`. This means that at any point in time, any page could arrive by virtue of having been requested by another clone—the beauty of asynchronous prefetching. Second, we allow SnowFlock VMs to be multi-processor VMs. There wouldn't be much fun otherwise. This means that multiple faults need to be handled in parallel, perhaps even for the same page. Third, in later versions `memtap` helpers can explicitly prefetch a batch of pages, which can arrive in any order given the lack of guarantees from the `mcdist` server. Any of these factors could have led to a concurrency nightmare, and we have all of them.

The entire `memtap` design centers on a page presence bitmap. The bitmap is created and initialized when the architectural descriptor is processed to create the clone VM. The bitmap is a flat bit array sized by the number of pages the VM's memory can hold. Intel processors have handy atomic bit mutation instructions: setting a bit, or doing a test and set, can happen with the guarantee of atomicity with respect to other processors in the same box. This allows us to avoid locks in most cases, and thus to provide access to the bitmap by different entities in different protection domains: the Xen hypervisor, the `memtap` process, and the cloned guest kernel itself.

When Xen handles a hardware page fault due to a first access to a page, it uses the bitmap to decide whether it needs to alert `memtap`. It also uses the bitmap to enqueue multiple faulting virtual

processors as dependent on the same absent page. Memtap buffers pages as they arrive. When its buffer is full or an explicitly requested page arrives, the VM is paused, and the bitmap is used to discard any duplicate pages that have arrived but are already present. Any remaining pages that are needed are then copied into the VM memory, and the appropriate bitmap bits are set.

Clever Clones Avoid Unnecessary Fetches

We just mentioned that the page presence bitmap is visible to the kernel running inside the clone, and that no locks are needed to modify it. This gives clones a powerful “enlightenment” tool: they can prevent the fetching of pages by modifying the bitmap and pretending they are present. This is extremely useful performance-wise, and safe to do when pages will be completely overwritten before they are used.

There happens to be a very common situation when this happens and fetches can be avoided. All memory allocations in the kernel (using `vmalloc`, `kzalloc`, `get_free_page`, user-space `brk`, and the like) are ultimately handled by the kernel page allocator. Typically pages are requested by intermediate allocators which manage finer-grained chunks: the slab allocator, the glibc malloc allocator for a user-space process, etc. However, whether allocation is explicit or implicit, one key semantic implication always holds true: no one cares about what the page contained, because its contents will be arbitrarily overwritten. Why fetch such a page, then? There is no reason to do so, and empirical experience shows that avoiding such fetches is tremendously advantageous.

18.7 VM Cloning Application Interface

So far we have focused on the internals of cloning a VM efficiently. As much fun as solipsistic systems can be, we need to turn our attention to those who will use the system: applications.

API Implementation

VM Cloning is offered to the application via the simple SnowFlock API, depicted in [図 18.1](#). Cloning is basically a two-stage process. You first request an allocation for the clone instances, although due to the system policies that are in effect, that allocation may be smaller than requested. Second, you use the allocation to clone your VM. A key assumption is that your VM focuses on a single operation. VM Cloning is appropriate for single-application VMs such as a web server or a render farm component. If you have a hundred-process desktop environment in which multiple applications concurrently call VM cloning, you’re headed for chaos.

The API simply marshals messages and posts them to the XenStore, a shared-memory low-throughput interface used by Xen for control plane transactions. A SnowFlock Local Daemon

<code>sf_request_ticket(n)</code>	Requests an allocation for n clones. Returns a ticket describing an allocation for $m \leq n$ clones.
<code>sf_clone(ticket)</code>	Clones, using the ticket allocation. Returns the clone ID, $0 \leq ID < m$.
<code>sf_checkpoint_parent()</code>	Prepares an immutable checkpoint C of the parent VM to be used for creating clones at an arbitrarily later time.
<code>sf_create_clones(C, ticket)</code>	Same as <code>sf_clone</code> , uses the checkpoint C . Clones will begin execution at the point at which the corresponding <code>sf_checkpoint_parent()</code> was invoked.
<code>sf_exit()</code>	For children ($1 \leq ID < m$), terminates the child.
<code>sf_join(ticket)</code>	For the parent ($ID = 0$), blocks until all children in the ticket reach their <code>sf_exit</code> call. At that point all children are terminated and the ticket is discarded.
<code>sf_kill(ticket)</code>	Parent only, discards ticket and immediately kills all associated children.

表 18.1: The SnowFlock VM Cloning API

(SFLD) executes on the hypervisor and listens for such requests. Messages are unmarshalled, executed, and requests posted back.

Programs can control VM Cloning directly through the API, which is available for C, C++, Python and Java. Shell scripts that harness the execution of a program can use the provided command-line scripts instead. Parallel frameworks such as MPI can embed the API: MPI programs can then use SnowFlock without even knowing, and with no modification to their source. Load balancers sitting in front of web or application servers can use the API to clone the servers they manage.

SFLDs orchestrate the execution of VM Cloning requests. They create and transmit architectural descriptors, create cloned VMs, launch disk and memory servers, and launch memtap helper processes. They are a miniature distributed system in charge of managing the VMs in a physical cluster.

SFLDs defer allocation decisions to a central SnowFlock Master Daemon (SFMD). SFMD simply interfaces with appropriate cluster management software. We did not see any need to reinvent the wheel here, and deferred decisions on resource allocation, quotas, policies, etc. to suitable software such as Sun Grid Engine or Platform EGO.

Necessary Mutations

After cloning, most of the cloned VM's processes have no idea that they are no longer the parent, and that they are now running in a copy. In most aspects, this just works fine and causes no issues. After all, the primary task of the OS is to isolate applications from low-level details, such as the network identity. Nonetheless, making the transition smooth requires a set of mechanisms to be put in place. The meat of the problem is in managing the clone's network identity; to avoid conflicts and confusion, we must introduce slight mutations during the cloning process. Also, because these tweaks may necessitate higher-level accommodations, a hook is inserted to allow the user to configure any necessary tasks, such as (re)mounting network filesystems that rely on the clone's identity.

Clones are born to a world that is mostly not expecting them. The parent VM is part of a network managed most likely by a DHCP server, or by any other of the myriad ways sysadmins find to do their job. Rather than assume a necessarily inflexible scenario, we place the parent and all clones in their own private virtual network. Clones from the same parent are all assigned a unique ID, and their IP address in this private network is automatically set up upon cloning as a function of the ID. This guarantees that no intervention from a sysadmin is necessary, and that no IP address collisions will ever happen.

IP reconfiguration is performed directly by a hook we place on the virtual network driver. However, we also rig the driver to automatically generate synthetic DHCP responses. Thus, regardless of your choice of distribution, your virtual network interface will ensure that the proper IP coordinates are propagated to the guest OS, even when you are restarting from scratch.

To prevent clones from different parents colliding with each others' virtual private networks—and to prevent mutual DDoS attacks—clone virtual network are isolated at the Ethernet (or layer 2) level. We hijack a range of Ethernet MAC OUIs³ and dedicate them to clones. The OUI will be a function of the parent VM. Much like the ID of a VM determines its IP address, it also determines its non-OUI Ethernet MAC address. The virtual network driver translates the MAC address the VM believes it has to the one assigned as a function of its ID, and filters out all traffic to and from virtual private network with different OUIs. This isolation is equivalent to that achievable via ebtables, although much simpler to implement.

Having clones talk only to each other may be fun, but not fun enough. Sometimes we will want our clones to reply to HTTP requests from the Internet, or mount public data repositories. We equip any set of parent and clones with a dedicated router VM. This tiny VM performs firewalling, throttling and NATing of traffic from the clones to the Internet. It also limits inbound connections to the parent VM and well-known ports. The router VM is lightweight but represents a single point of centralization for network traffic, which can seriously limit scalability. The same network rules could be applied in a distributed fashion to each host upon which a clone VM runs. We have not released that experimental patch.

SFLDs assign IDs, and teach the virtual network drivers how they should configure themselves: internal MAC and IP addresses, DHCP directives, router VM coordinates, filtering rules, etc.

18.8 Conclusion

By tweaking the Xen hypervisor and lazily transferring the VM's state, SnowFlock can produce dozens of clones of a running VM in a few seconds. Cloning VMs with SnowFlock is thus instantaneous and live—it improves cloud usability by automating cluster management and giving applications greater programmatic control over the cloud resources. SnowFlock also improves cloud agility by speeding up VM instantiation by a factor of 20, and by improving the performance of most newly created VMs by leveraging their parent's warm, in-memory OS and application caches. The keys to SnowFlock's efficient performance are heuristics that avoid unnecessary page fetches, and the multicast system that lets clone siblings cooperatively prefetch their state. All it took was the clever application of a few tried-and-true techniques, some sleight of hand, and a generous helping of industrial-strength debugging.

We learned two important lessons throughout the SnowFlock experience. The first is the often-underestimated value of the KISS theorem. We were expecting to implement complicated prefetching techniques to alleviate the spate of requests for memory pages a clone would issue upon startup. This was, perhaps surprisingly, not necessary. The system performs very well for many workloads based on one single principle: bring the memory over as needed. Another example of the value of

³OUI, or Organizational Unique ID, is a range of MAC addresses assigned to a vendor.

simplicity is the page presence bitmap. A simple data structure with clear atomic access semantics greatly simplifies what could have been a gruesome concurrency problem, with multiple virtual CPUs competing for page updates with the asynchronous arrival of pages via multicast.

The second lesson is that scale does not lie. In other words, be prepared to have your system shattered and new bottlenecks uncovered every time you bump your scale by a power of two. This is intimately tied with the previous lesson: simple and elegant solutions scale well and do not hide unwelcome surprises as load increases. A prime example of this principle is our `mcdist` system. In large-scale tests, a TCP/IP-based page distribution mechanism fails miserably for hundreds of clones. Mcdist succeeds by virtue of its extremely constrained and well-defined roles: clients only care about their own pages; the server only cares about maintaining a global flow control. By keeping `mcdist` humble and simple, SnowFlock is able to scale extremely well.

If you are interested in knowing more, you can visit the University of Toronto site⁴ for the academic papers and open-source code licensed under GPLv2, and GridCentric⁵ for an industrial strength implementation.

⁴<http://sysweb.cs.toronto.edu/projects/1>

⁵<http://www.gridcentriclabs.com/>

SocialCalc

Audrey Tang

The history of spreadsheets spans more than 30 years. The first spreadsheet program, VisiCalc, was conceived by Dan Bricklin in 1978 and shipped in 1979. The original concept was quite straightforward: a table that spans infinitely in two dimensions, its cells populated with text, numbers, and formulas. Formulas are composed of normal arithmetic operators and various built-in functions, and each formula can use the current contents of other cells as values.

Although the metaphor was simple, it had many applications: accounting, inventory, and list management are just a few. The possibilities were practically limitless. All these uses made VisiCalc into the first “killer app” of the personal computer era.

In the decades that followed successors like Lotus 1-2-3 and Excel made incremental improvements, but the core metaphor stayed the same. Most spreadsheets were stored as on-disk files, and loaded into memory when opened for editing. Collaboration was particularly hard under the file-based model:

- Each user needed to install a version of the spreadsheet editor.
- E-mail ping-pong, shared folders, or setting up a dedicated version-control system all added bookkeeping overhead.
- Change tracking was limited; for example, Excel does not preserve history for formatting changes and cell comments.
- Updating formatting or formulas in templates required painstaking changes to existing spreadsheet files that used that template.

Fortunately, a new collaboration model emerged to address these issues with elegant simplicity. It is the wiki model, invented by Ward Cunningham in 1994, and popularized by Wikipedia in the early 2000s.

Instead of files, the wiki model features server-hosted pages, editable in the browser without requiring special software. Those hypertext pages can easily link to each other, and even include

portions of other pages to form a larger page. All participants view and edit the latest version by default, with revision history automatically managed by the server.

Inspired by the wiki model, Dan Bricklin started working on WikiCalc in 2005. It aims to combine the authoring ease and multi-person editing of wikis with the familiar visual formatting and calculating metaphor of spreadsheets.

19.1 WikiCalc

The first version of WikiCalc (図 19.1) had several features that set it apart from other spreadsheets at the time:

- Plain text, HTML, and wiki-style markup rendering for text data.
- Wiki-style text that includes commands to insert links, images, and values from cell references.
- Formula cells may reference values of other WikiCalc pages hosted on other websites.
- Ability to create output to be embedded in other web pages, both static and live data.
- Cell formatting with access to CSS style attributes and CSS classes.
- Logging of all edit operations as an audit trail.
- Wiki-like retention of each new version of a page with roll-back capability.

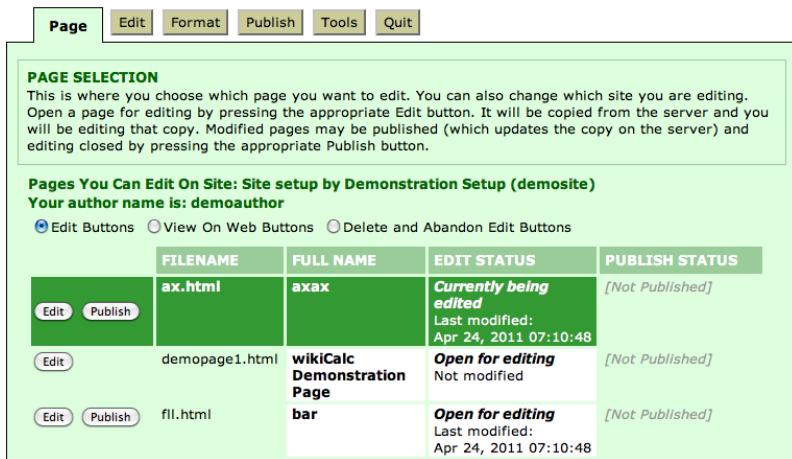


図 19.1: WikiCalc 1.0 Interface

WikiCalc 1.0's internal architecture (図 19.2) and information flow (図 19.3) were deliberately simple, but nevertheless powerful. The ability to compose a master spreadsheet from several smaller spreadsheets proved particularly handy. For example, imagine a scenario where each salesperson keeps numbers in a spreadsheet page. Each sales manager then rolls up their reps' numbers into a

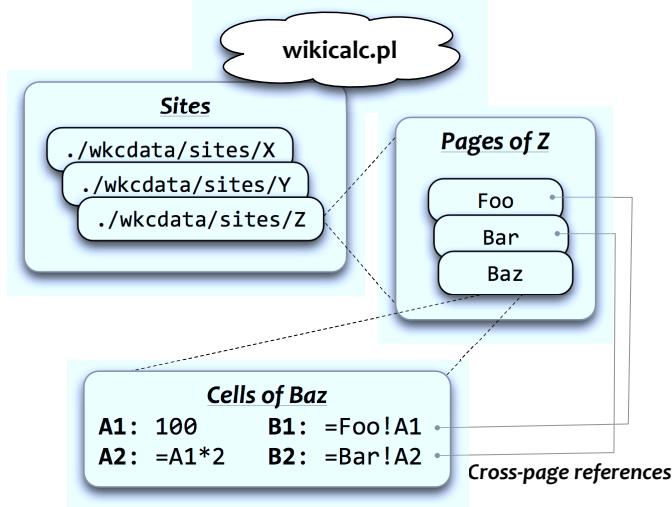


図 19.2: WikiCalc Components

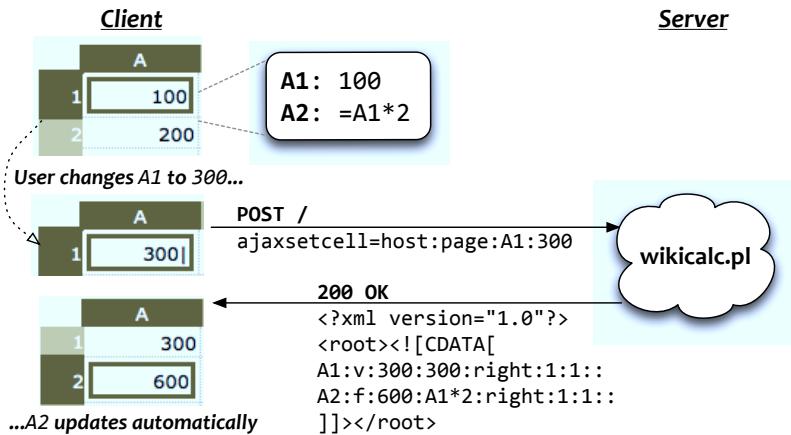


図 19.3: WikiCalc Flow

regional spreadsheet, and the VP of sales then rolls up the regional numbers into a top-level spreadsheet.

Each time one of the individual spreadsheets is updated, all the roll-up spreadsheets can reflect the update. If someone wants further detail, they simply click through to view the spreadsheet behind the spreadsheet. This roll-up capability eliminates the redundant and error-prone effort of updating numbers in multiple places, and ensures all views of the information stay fresh.

To ensure the recalculations are up-to-date, WikiCalc adopted a thin-client design, keeping all the state information on the server side. Each spreadsheet is represented on the browser as a `<table>`

element; editing a cell will send an ajaxsetcell call to the server, and the server then tells the browser which cells need updating.

Unsurprisingly, this design depends on a fast connection between the browser and the server. When the latency is high, users will start to notice the frequent appearance of “Loading...” messages between updating a cell and seeing its new contents as shown in 図 19.4. This is especially a problem for users interactively editing formulas by tweaking the input and expecting to see results in real time.

A	B	C	D
1	Loading...		
2			
3	Sample financial calculation in a table with borders	Year	2006 2007
4		Sales	Loading... 170.5
5		Cost	124.0 136.4
6		Profit	31.0 34.1

図 19.4: Loading Message

Moreover, because the <table> element had the same dimensions as the spreadsheet, a 100×100 grid would create 10,000 <td> DOM objects, which strains the memory resource of browsers, further limiting the size of pages.

Due to these shortcomings, while WikiCalc was useful as a stand-alone server running on localhost, it was not very practical to embed as part of web-based content management systems.

In 2006, Dan Bricklin teamed up with Socialtext to start developing SocialCalc, a ground-up rewrite of WikiCalc in Javascript based on some of the original Perl code.

This rewrite was aimed at large, distributed collaborations, and sought to deliver a look and feel more like that of a desktop app. Other design goals included:

- Capable of handling hundreds of thousands of cells.
- Fast turnaround time for edit operations.
- Client-side audit trail and undo/redo stack.
- Better use of Javascript and CSS to provide full-fledged layout functionality.
- Cross-browser support, despite the more extensive use of responsive Javascript.

After three years of development and various beta releases, Socialtext released SocialCalc 1.0 in 2009, successfully meeting the design goals. Let’s now take a look at the architecture of the SocialCalc system.

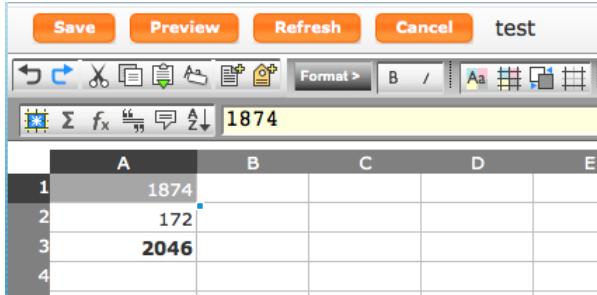


图 19.5: SocialCalc Interface

19.2 SocialCalc

图 19.5 and 图 19.6 show SocialCalc's interface and classes respectively. Compared to WikiCalc, the server's role has been greatly reduced. Its only responsibility is responding to HTTP GETs by serving entire spreadsheets serialized in the save format; once the browser receives the data, all calculations, change tracking and user interaction are now implemented in Javascript.

The Javascript components were designed with a layered MVC (Model/View/Controller) style, with each class focusing on a single aspect:

Sheet is the data model, representing an in-memory structure of a spreadsheet. It contains a dictionary from coordinates to *Cell* objects, each representing a single cell. Empty cells need no entries, and hence consume no memory at all.

Cell represents a cell's content and formats. Some common properties are shown in 表 19.1.

RenderContext implements the view; it is responsible for rendering a sheet into DOM objects.

TableControl is the main controller, accepting mouse and keyboard events. As it receives view events such as scrolling and resizing, it updates its associated *RenderContext* object. As it receives update events that affects the sheet's content, it schedules new commands to the sheet's command queue.

SpreadSheetControl is the top-level UI with toolbars, status bars, dialog boxes and color pickers.

SpreadSheetViewer is an alternate top-level UI that provides a read-only interactive view.

We adopted a minimal class-based object system with simple composition/delegation, and make no use of inheritance or object prototypes. All symbols are placed under the `SocialCalc.*` namespace to avoid naming conflicts.

Each update on the sheet goes through the `ScheduleSheetCommands` method, which takes a command string representing the edit. (Some common commands are shown in 表 19.2.) The application embedding SocialCalc may define extra commands on their own, by adding named callbacks into the `SocialCalc.SheetCommandInfo.CmdExtensionCallbacks` object, and use the `startcmdextension` command to invoke them.

SocialCalc Class Diagram

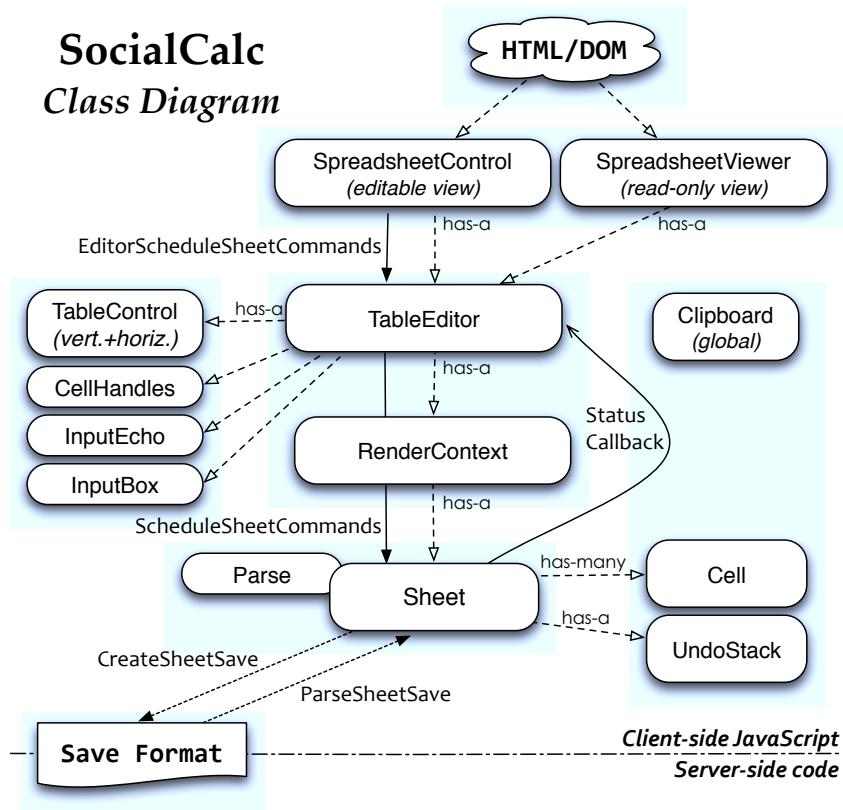


図 19.6: SocialCalc Class Diagram

19.3 Command Run-loop

To improve responsiveness, SocialCalc performs all recalculation and DOM updates in the background, so the user can keep making changes to several cells while the engine catches up on earlier changes in the command queue.

When a command is running, the **TableEditor** object sets its busy flag to true; subsequent commands are then pushed into the `deferredCommands` queue, ensuring a sequential order of execution. As the event loop diagram in 図 19.7 shows, the **Sheet** object keeps sending `StatusCallback` events to notify the user of the current state of command execution, through each of the four steps:

ExecuteCommand: Sends `cmdstart` upon start, and `cmdend` when the command finishes execution. If the command changed a cell's value indirectly, enter the *Recalc* step. Otherwise, if the command changed the visual appearance of one or more on-screen cells, enter the *Render* step. If neither of the above applies (for example with the copy command), skip to the

datatype	t
datavalue	1Q84
color	black
bgcolor	white
font	italic bold 12pt Ubuntu
comment	Ichi-Kyu-Hachi-Yon

表 19.1: Cell Contents and Formats

set	sheet defaultcolor blue	erase	A2
set	A width 100	cut	A3
set	A1 value n 42	paste	A4
set	A2 text t Hello	copy	A5
set	A3 formula A1*2	sort	A1:B9 A up B down
set	A4 empty	name	define Foo A1:A5
set	A5 bgcolor green	name	desc Foo Used in formulas like SUM(Foo)
merge	A1:B2	name	delete Foo
unmerge	A1	startcmdextension	UserDefined args

表 19.2: SocialCalc Commands

PositionCalculations step.

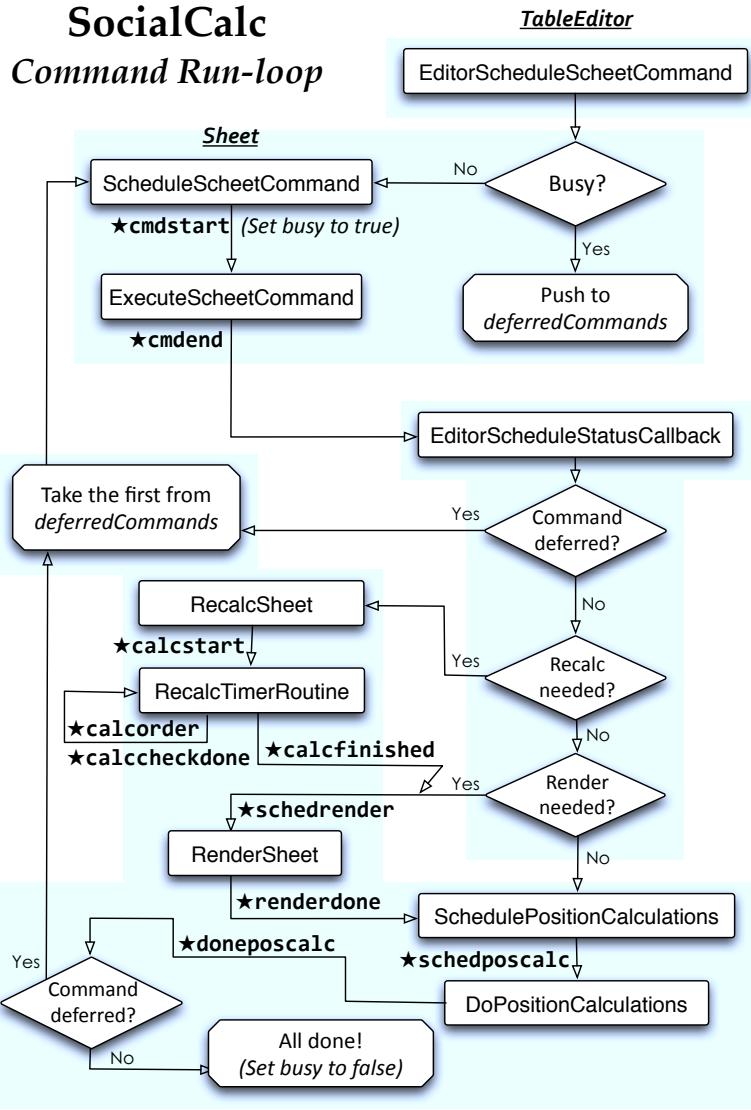


図 19.7: SocialCalc Command Run-loop

Recalc (as needed): Sends calcstart upon start, calccorder every 100ms when checking the dependency chain of cells, calccheckdone when the check finishes, and calcfinished when all affected cells received their re-calculated values. This step is always followed by the *Render* step.

Render (as needed): Sends schedrender upon start, and renderdone when the <table> element is updated with formatted cells. This step is always followed by *PositionCalculations*.

PositionCalculations: Sends schedposcalc upon start, and doneposcalc after updating the scroll-

bars, the current editable cell cursor, and other visual components of the TableEditor.

Because all commands are saved as they are executed, we naturally get an audit log of all operations. The Sheet.CreateAuditString method provides a newline-delimited string as the audit trail, with each command in a single line.

ExecuteSheetCommand also creates an undo command for each command it executes. For example, if the cell A1 contains “Foo” and the user executes set A1 text Bar, then an undo-command set A1 text Foo is pushed to the undo stack. If the user clicks Undo, then the undo-command is executed to restore A1 to its original value.

19.4 Table Editor

Now let’s look at the TableEditor layer. It calculates the on-screen coordinates of its RenderContext, and manages horizontal/vertical scroll bars through two TableControl instances.

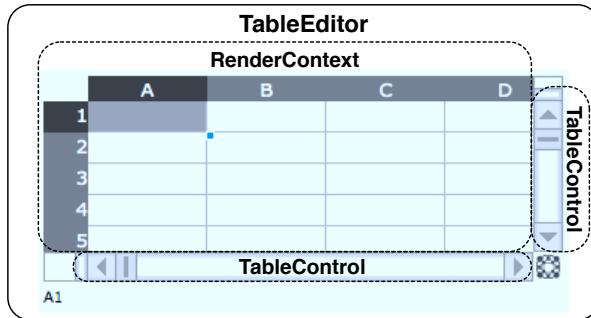


図 19.8: TableControl Instances Manage Scroll Bars

The view layer, handled by the RenderContext class, also differs from WikiCalc’s design. Instead of mapping each cell to a `<td>` element, we now simply create a fixed-size `<table>` that fits the browser’s visible area, and pre-populate it with `<td>` elements.

As the user scrolls the spreadsheet through our custom-drawn scroll bars, we dynamically update the `innerHTML` of the pre-drawn `<td>` elements. This means we don’t need to create or destroy any `<tr>` or `<td>` elements in many common cases, which greatly speeds up response time.

Because RenderContext only renders the visible region, the size of Sheet object can be arbitrarily large without affecting its performance.

TableEditor also contains a CellHandles object, which implements the radial fill/move/slide menu attached to the bottom-right corner to the current editable cell, known as the ECell, shown in 図 19.9.

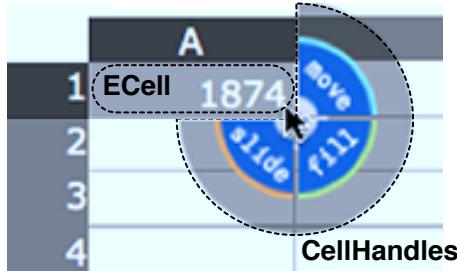


図 19.9: Current Editable Cell, Known as the ECell

The input box is managed by two classes: `InputBox` and `InputEcho`. The former manages the above-the-grid edit row, while the latter shows an updated-as-you-type preview layer, overlaying the ECell's content (図 19.10).

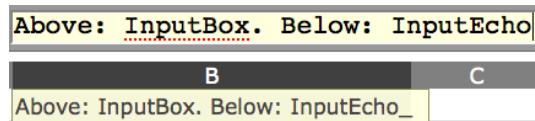


図 19.10: The Input Box is Managed by Two Classes

Usually, the SocialCalc engine only needs to communicate to the server when opening a spreadsheet for edit, and when saving it back to server. For this purpose, the `Sheet.ParseSheetSave` method parses a save format string into a `Sheet` object, and the `Sheet.CreateSheetSave` method serializes a `Sheet` object back into the save format.

Formulas may refer to values from any remote spreadsheet with a URL. The `recalc` command re-fetches the externally referenced spreadsheets, parses them again with `Sheet.ParseSheetSave`, and stores them in a cache so the user can refer to other cells in the same remote spreadsheets without re-fetching its content.

19.5 Save Format

The save format is in standard MIME `multipart/mixed` format, consisting of four `text/plain; charset=UTF-8` parts, each part containing newline-delimited text with colon-delimited data fields. The parts are:

- The `meta` part lists the types of the other parts.
- The `sheet` part lists each cell's format and content, each column's width (if not default), the sheet's default format, followed by a list of fonts, colors and borders used in the sheet.

- The optional edit part saves the TableEditor's edit state, including ECell's last position, as well as the fixed sizes of row/column panes.
- The optional audit part contains the history of commands executed in the previous editing session.

For example, 図 19.11 shows a spreadsheet with three cells, with 1874 in A1 as the ECell, the formula $2^2 * 43$ in A2, and the formula $\text{SUM}(\text{Foo})$ in A3 rendered in bold, referring to the named range Foo over A1:A2.

A	
1	1874
2	"Foo"
3	$=2^2 * 43$
	$=\text{SUM}(\text{Foo})$

図 19.11: A Spreadsheet with Three Cells

The serialized save format for the spreadsheet looks like this:

```

socialcalc:version:1.0
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary=SocialCalcSpreadsheetControlSave
--SocialCalcSpreadsheetControlSave
Content-type: text/plain; charset=UTF-8

# SocialCalc Spreadsheet Control Save
version:1.0
part:sheet
part:edit
part:audit
--SocialCalcSpreadsheetControlSave
Content-type: text/plain; charset=UTF-8

version:1.5
cell:A1:v:1874
cell:A2:vtf:n:172:2^2*43
cell:A3:vtf:n:2046:SUM(Foo):f:1
sheet:c:1:r:3
font:1:normal bold * *
name:FOO::A1\cA2
--SocialCalcSpreadsheetControlSave
Content-type: text/plain; charset=UTF-8

version:1.0
rowpane:0:1:14
colpane:0:1:16
ecline:A1

```

```
--SocialCalcSpreadsheetControlSave
Content-type: text/plain; charset=UTF-8

set A1 value n 1874
set A2 formula 2^2*43
name define Foo A1:A2
set A3 formula SUM(Foo)
--SocialCalcSpreadsheetControlSave--
```

This format is designed to be human-readable, as well as being relatively easy to generate programmatically. This makes it possible for Drupal's Sheetnode plugin to use PHP to convert between this format and other popular spreadsheet formats, such as Excel (.xls) and OpenDocument (.ods).

Now that we have a good idea about how the pieces in SocialCalc fit together, let's look at two real-world examples of extending SocialCalc.

19.6 Rich-text Editing

The first example we'll look at is enhancing SocialCalc's text cells with wiki markup to display its rich-text rendering right in the table editor (図 19.12).

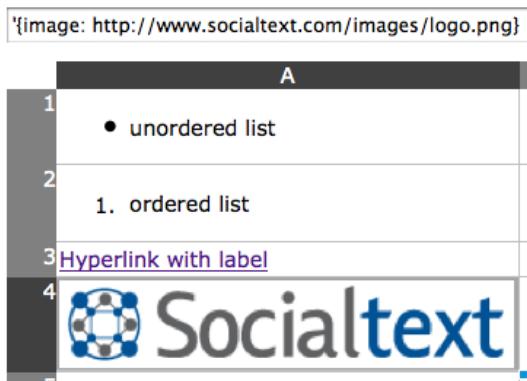


図 19.12: Rich Text Rendering in the Table Editor

We added this feature to SocialCalc right after its 1.0 release, to address the popular request of inserting images, links and text markups using a unified syntax. Since Socialtext already has an open-source wiki platform, it was natural to re-use the syntax for SocialCalc as well.

To implement this, we need a custom renderer for the `textvalueformat` of `text-wiki`, and to change the default format for text cells to use it.

What is this `textvalueformat`, you ask? Read on.

Types and Formats

In SocialCalc, each cell has a datatype and a valuetype. Data cells with text or numbers correspond to text/numeric value types, and formula cells with datatype="f" may generate either numeric or text values.

Recall that on the Render step, the Sheet object generates HTML from each of its cells. It does so by inspecting each cell's valuetype: If it begins with t, then the cell's textvalueformat attribute determines how generation is done. If it begins with n, then the nontextvalueformat attribute is used instead.

However, if the cell's textvalueformat or nontextvalueformat attribute is not defined explicitly, then a default format is looked up from its valuetype, as shown in 図 19.13.

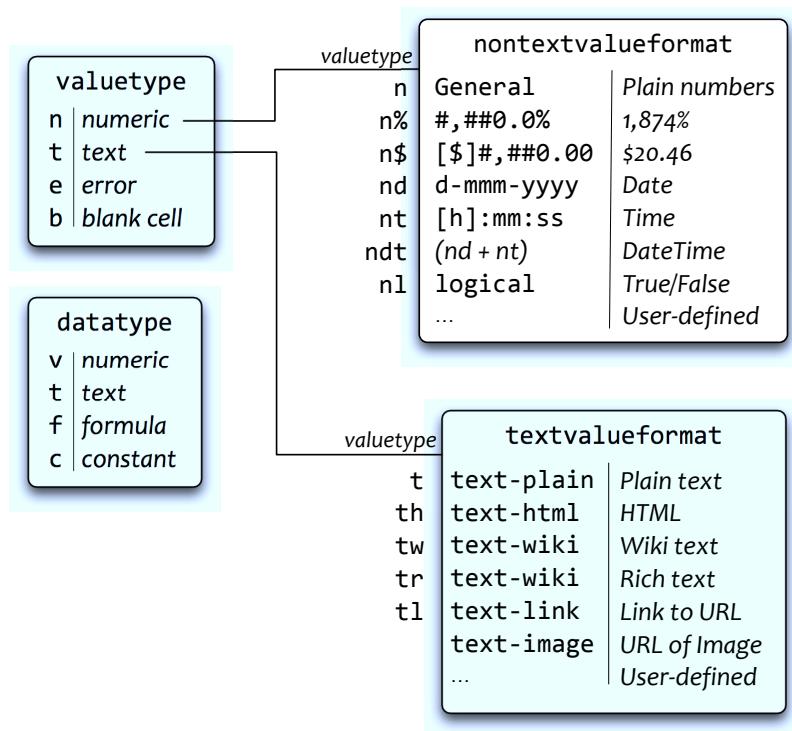


図 19.13: Value Types

Support for the text-wiki value format is coded in SocialCalc.format_text_for_display:

```
if (SocialCalc.Callbacks.expand_wiki && /^text-wiki/.test(valueformat)) {  
    // do general wiki markup  
    displayvalue = SocialCalc.Callbacks.expand_wiki(  
        displayvalue, sheetobj, linkstyle, valueformat  
    );  
}
```

Instead of inlining the wiki-to-HTML expander in `format_text_for_display`, we will define a new hook in `SocialCalc.Callbacks`. This is the recommended style throughout the SocialCalc codebase; it improves modularity by making it possible to plug in different ways of expanding wiki-text, as well as keeping compatibility with embedders that do not desire this feature.

Rendering Wikitext

Next, we'll make use of `Wikiwyg`¹, a Javascript library offering two-way conversions between wikitext and HTML.

We define the `expand_wiki` function by taking the cell's text, running it through `Wikiwyg`'s wikitext parser and its HTML emitter:

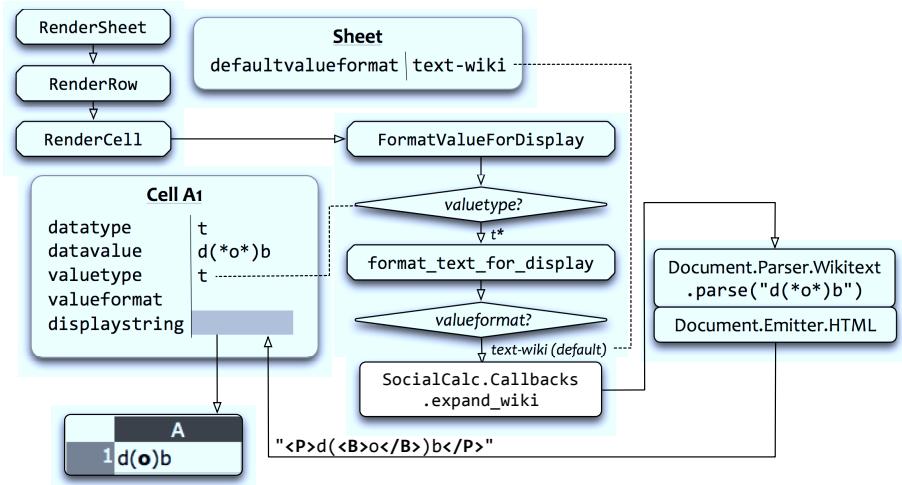
```
var parser = new Document.Parser.Wikitext();
var emitter = new Document.Emitter.HTML();
SocialCalc.Callbacks.expand_wiki = function(val) {
    // Convert val from Wikitext to HTML
    return parser.parse(val, emitter);
}
```

¹<https://github.com/audreyt/wikiwyg-js>

The final step involves scheduling the `set sheet defaulttextvalueformat text-wiki` command right after the spreadsheet initializes:

```
// We assume there's a <div id="tableeditor"/> in the DOM already
var spreadsheet = new SocialCalc.SpreadsheetControl();
spreadsheet.InitializeSpreadsheetControl("tableeditor", 0, 0, 0);
spreadsheet.ExecuteCommand('set sheet defaulttextvalueformat text-wiki');
```

Taken together, the Render step now works as shown in [図 19.14](#).



[図 19.14](#): Render Step

That's all! The enhanced SocialCalc now supports a rich set of wiki markup syntax:

```
*bold* _italic_ ‘monospace’ {{unformatted}}
> indented text
* unordered list
# ordered list
"Hyperlink with label"<http://softwaregarden.com/>
{image: http://www.socialtext.com/images/logo.png}
```

Try entering `*bold* _italic_ ‘monospace’` in A1, and you'll see it rendered as rich text ([図 19.15](#)).

19.7 Real-time Collaboration

The next example we'll explore is multi-user, real-time editing on a shared spreadsheet. This may seem complicated at first, but thanks to SocialCalc's modular design all it takes is for each on-line user to broadcast their commands to other participants.



図 19.15: Wikwyg Example

To distinguish between locally-issued commands and remote commands, we add an `isRemote` parameter to the `ScheduleSheetCommands` method:

```
SocialCalc.ScheduleSheetCommands = function(sheet, cmdstr, saveundo, isRemote) {
    if (SocialCalc.Callbacks.broadcast && !isRemote) {
        SocialCalc.Callbacks.broadcast('execute', {
            cmdstr: cmdstr, saveundo: saveundo
        });
    }
    // ...original ScheduleSheetCommands code here...
}
```

Now all we need to do is to define a suitable `SocialCalc.Callbacks.broadcast` callback function. Once it's in place, the same commands will be executed on all users connected to the same spreadsheet.

When this feature was first implemented for OLPC (One Laptop Per Child²) by SEETA's Sugar Labs³ in 2009, the `broadcast` function was built with XPCOM calls into D-Bus/Telepathy, the standard transport for OLPC/Sugar networks (see 図 19.16).

That worked reasonably well, enabling XO instances in the same Sugar network to collaborate on a common SocialCalc spreadsheet. However, it is both specific to the Mozilla/XPCOM browser platform, as well as to the D-Bus/Telepathy messaging platform.

Cross-browser Transport

To make this work across browsers and operating systems, we use the `Web::Hippie`⁴ framework, a high-level abstraction of JSON-over-WebSocket with convenient jQuery bindings, with MXHR (Multipart XML HTTP Request⁵) as the fallback transport mechanism if WebSocket is not available.

For browsers with Adobe Flash plugin installed but without native WebSocket support, we use the `web_socket.js`⁶ project's Flash emulation of WebSocket, which is often faster and more reliable than MXHR. The operation flow is shown in 図 19.17.

²<http://one.laptop.org/>

³http://seeta.in/wiki/index.php?title=Collaboration_in_SocialCalc

⁴<http://search.cpan.org/dist/Web-Hippie/>

⁵<http://about.digg.com/blog/duistream-and-mxhr>

⁶<https://github.com/gimite/web-socket-js>

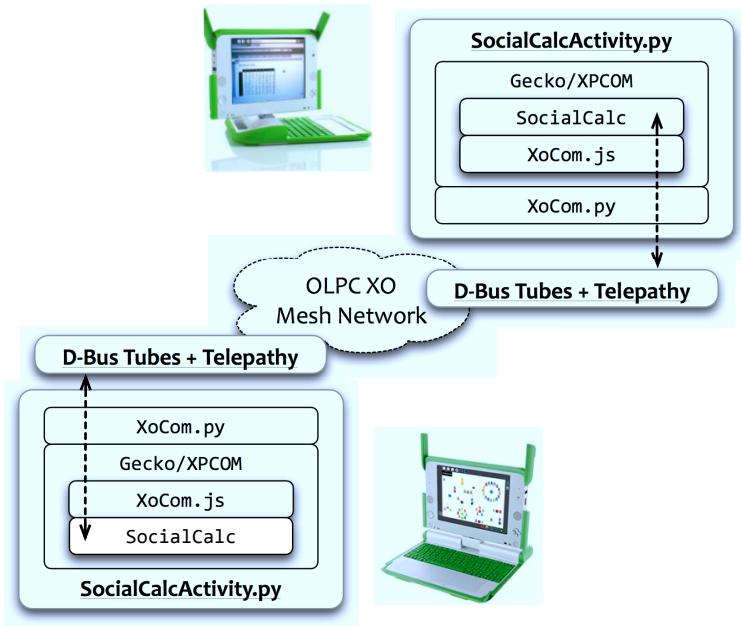


図 19.16: OLPC Implementation

The client-side `SocialCalc.Callbacks.broadcast` function is defined as:

```
var hpipe = new Hippie.Pipe();

SocialCalc.Callbacks.broadcast = function(type, data) {
    hpipe.send({ type: type, data: data });
};

$(hpipe).bind("message.execute", function (e, d) {
    var sheet = SocialCalc.CurrentSpreadsheetControl0bject.context.sheetobj;
    sheet.ScheduleSheetCommands(
        d.data.cmdstr, d.data.saveundo, true // isRemote = true
    );
    break;
});
```

Although this works quite well, there are still two remaining issues to resolve.

Conflict Resolution

The first one is a race-condition in the order of commands executed: If users A and B simultaneously perform an operation affecting the same cells, then receive and execute commands broadcast from the other user, they will end up in different states, as shown in 図 19.18.

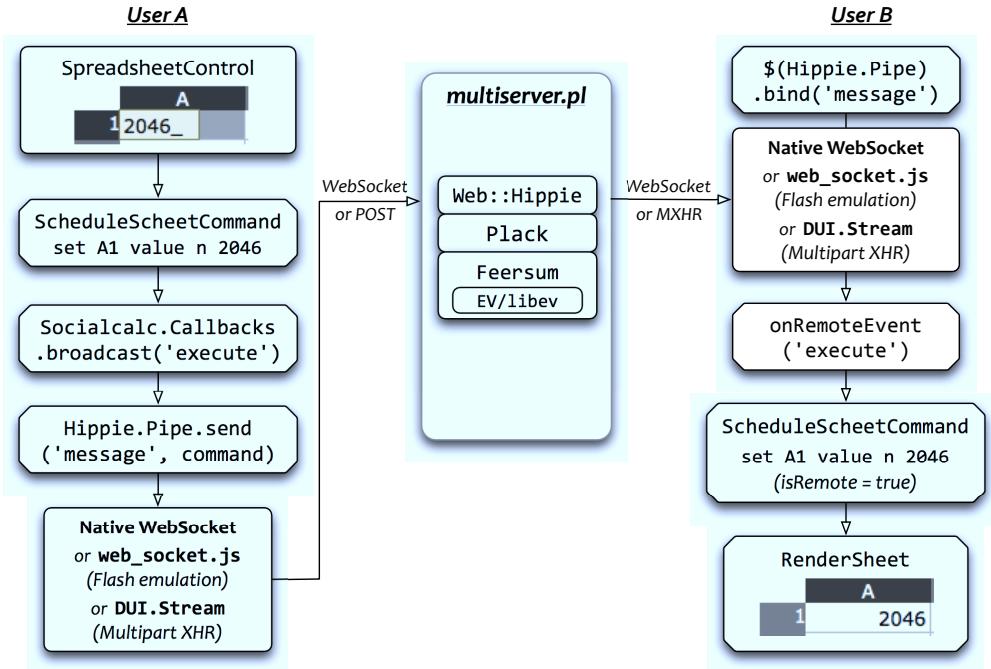


図 19.17: Cross-Browser Flow

We can resolve this with SocialCalc's built-in undo/redo mechanism, as shown in 図 19.19.

The process used to resolve the conflict is as follows. When a client broadcasts a command, it adds the command to a Pending queue. When a client receives a command, it checks the remote command against the Pending queue.

If the Pending queue is empty, then the command is simply executed as a remote action. If the remote command matches a command in the Pending queue, then the local command is removed from the queue.

Otherwise, the client checks if there are any queued commands that conflict with the received command. If there are conflicting commands, the client first Undoes those commands and marks them for later Redo. After undoing the conflicting commands (if any), the remote command is executed as usual.

When a marked-for-redo command is received from the server, the client will execute it again, then remove it from the queue.

Remote Cursors

Even with race conditions resolved, it is still suboptimal to accidentally overwrite the cell another user is currently editing. A simple improvement is for each client to broadcast its cursor position to

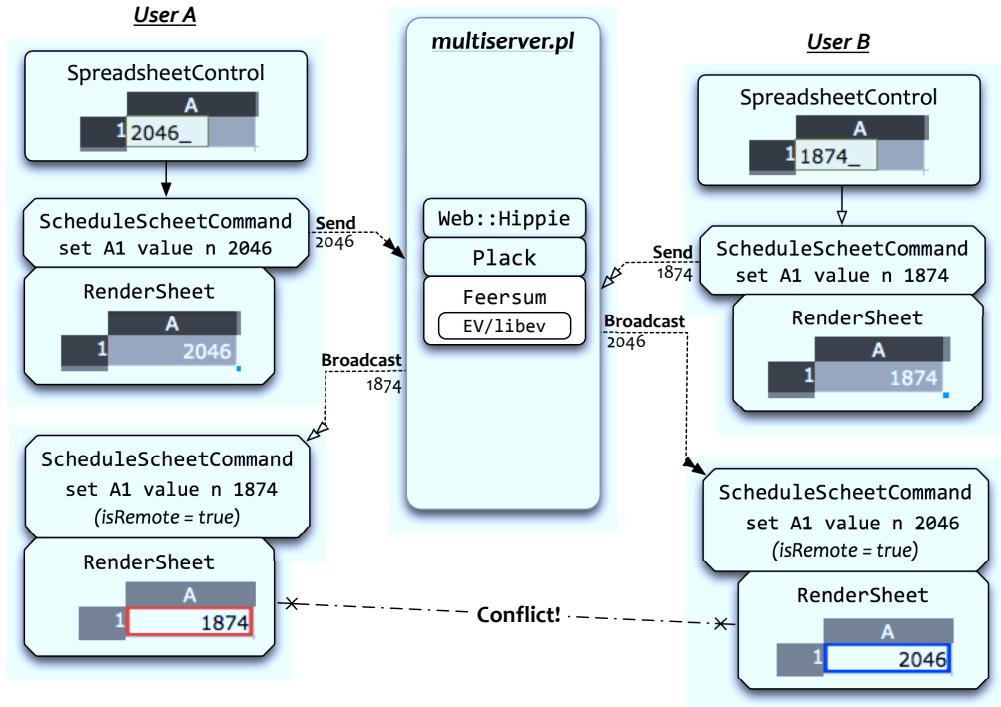


図 19.18: Race Condition Conflict

other users, so everyone can see which cells are being worked on.

To implement this idea, we add another broadcast handler to the `MoveECellCallback` event:

```
editor.MoveECellCallback.broadcast = function(e) {
    hpipe.send({
        type: 'ecell',
        data: e.ecell.coord
    });
};

$(hpipe).bind("message.ecell", function (e, d) {
    var cr = SocialCalc.coordToCr(d.data);
    var cell = SocialCalc.GetEditorCellElement(editor, cr.row, cr.col);
    // ...decorate cell with styles specific to the remote user(s) on it...
});
```

To mark cell focus in spreadsheets, it's common to use colored borders. However, a cell may already define its own border property, and since border is mono-colored, it can only represent one cursor on the same cell.

Therefore, on browsers with support for CSS3, we use the `box-shadow` property to represent multiple peer cursors in the same cell:

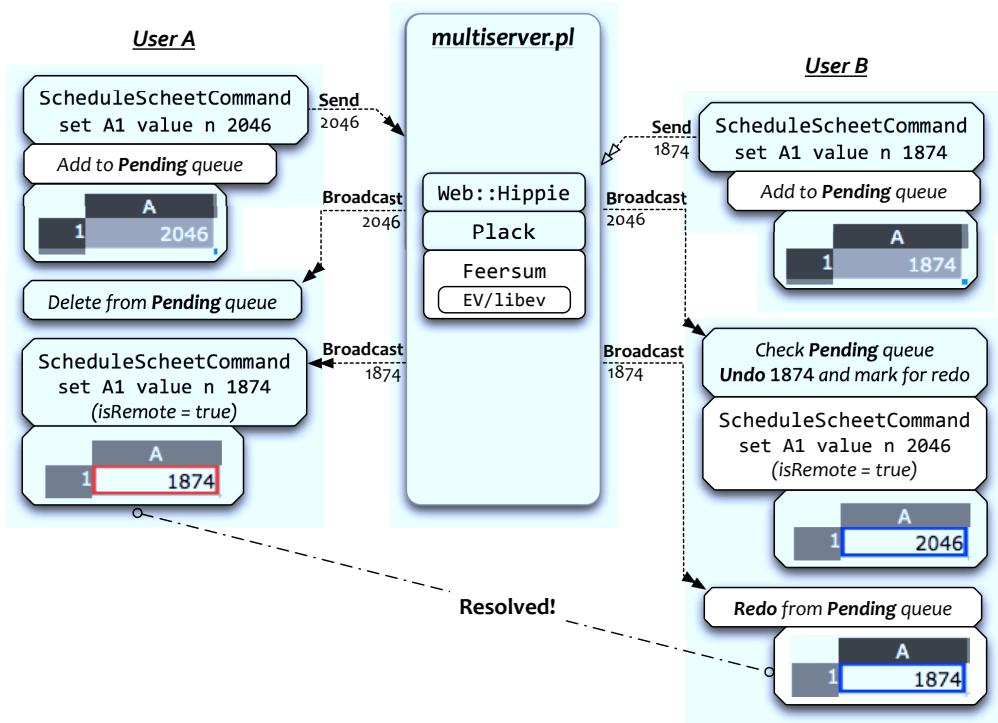


図 19.19: Race Condition Conflict Resolution

```
/* Two cursors on the same cell */
box-shadow: inset 0 0 0 4px red, inset 0 0 0 2px green;
```

図 19.20 shows how the screen would look with four people editing on the same spreadsheet.

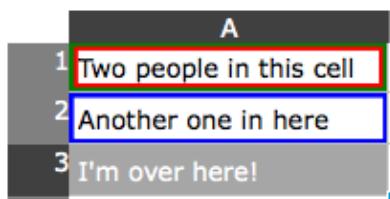


図 19.20: Four Users Editing One Spreadsheet

19.8 Lessons Learned

We delivered SocialCalc 1.0 on October 19th, 2009, the 30th anniversary of the initial release of VisiCalc. The experience of collaborating with my colleagues at Socialtext under Dan Bricklin's

guidance was very valuable to me, and I'd like to share some lessons I learned during that time.

Chief Designer with a Clear Vision

In [Bro10], Fred Brooks argues that when building complex systems, the conversation is much more direct if we focus on a coherent *design concept*, rather than derivative representations. According to Brooks, the formulation of such a coherent design concept is best kept in a single person's mind:

Since conceptual integrity is the most important attribute of a great design, and since that comes from one or a few minds working *uno animo*, the wise manager boldly entrusts each design task to a gifted chief designer.

In the case of SocialCalc, having Tracy Ruggles as our chief user-experience designer was the key for the project to converge toward a shared vision. Since the underlying SocialCalc engine was so malleable, the temptation of feature creep was very real. Tracy's ability to communicate using design sketches really helped us present features in a way that feels intuitive to users.

Wikis for Project Continuity

Before I joined the SocialCalc project, there was already over two years' worth of ongoing design and development, but I was able to catch up and start contributing in less than a week, simply due to the fact that *everything is in the wiki*. From the earliest design notes to the most up-to-date browser support matrix, the entire process was chronicled in wiki pages and SocialCalc spreadsheets.

Reading through the project's workspace brought me quickly to the same page as others, without the usual hand-holding overhead typically associated with orienting a new team member.

This would not be possible in traditional open source projects, where most conversation takes place on IRC and mailing lists and the wiki (if present) is only used for documentations and links to development resources. For a newcomer, it's much more difficult to reconstruct context from unstructured IRC logs and mail archives.

Embrace Time Zone Differences

David Heinemeier Hansson, creator of Ruby on Rails, once remarked on the benefit of distributed teams when he first joined 37signals. "The seven time zones between Copenhagen and Chicago actually meant that we got a lot done with few interruptions." With nine time zones between Taipei and Palo Alto, that was true for us during SocialCalc's development as well.

We often completed an entire Design-Development-QA feedback cycle within a 24-hour day, with each aspect taking one person's 8-hour work day in their local daytime. This asynchronous style

of collaboration compelled us to produce self-descriptive artifacts (design sketch, code and tests), which in turn greatly improved our trust in each other.

Optimize for Fun

In my 2006 keynote for the CONISLI conference [Tan06], I summarized my experience leading a distributed team implementing the Perl 6 language into a few observations. Among them, *Always have a Roadmap, Forgiveness > Permission, Remove deadlocks, Seek ideas, not consensus, and Sketch ideas with code* are particularly relevant for small distributed teams.

When developing SocialCalc, we took great care in distributing knowledge among team members with collaborative code ownership, so nobody would become a critical bottleneck.

Furthermore, we pre-emptively resolved disputes by actually coding up alternatives to explore the design space, and were not afraid of replacing fully-working prototypes when a better design arrived.

These cultural traits helped us foster a sense of anticipation and camaraderie despite the absence of face-to-face interaction, kept politics to a minimum, and made working on SocialCalc a lot of fun.

Drive Development with Story Tests

Prior to joining Socialtext, I've advocated the "interleave tests with the specification" approach, as can be seen in the Perl 6 specification⁷, where we annotate the language specification with the official test suite. However, it was Ken Pier and Matt Heusser, the QA team for SocialCalc, who really opened my eyes to how this can be taken to the next level, bringing tests to the place of *executable specification*.

In Chapter 16 of [GR09], Matt explained our story-test driven development process as follows:

The basic unit of work is a "story," which is an extremely lightweight requirements document. A story contains a brief description of a feature along with examples of what needs to happen to consider the story completed; we call these examples "acceptance tests" and describe them in plain English.

During the initial cut of the story, the product owner makes a good-faith first attempt to create acceptance tests, which are augmented by developers and testers before any developer writes a line of code.

⁷<http://perlcabal.org/syn/S02.html>

These story tests are then translated into wikitests, a table-based specification language inspired by Ward Cunningham’s FIT framework⁸, which drives automated testing frameworks such as Test::WWW::Mechanize⁹ and Test::WWW::Selenium¹⁰.

It’s hard to overstate the benefit of having story tests as a common language to express and validate requirements. It was instrumental in reducing misunderstanding, and has all but eliminated regressions from our monthly releases.

Open Source With CPAL

Last but not least, the open source model we chose for SocialCalc makes an interesting lesson in itself.

Socialtext created the Common Public Attribution License¹¹ for SocialCalc. Based on the Mozilla Public License, CPAL is designed to allow the original author to require an attribution to be displayed on the software’s user interface, and has a network-use clause that triggers share-alike provisions when derived work is hosted by a service over the network.

After its approval by both the Open Source Initiative¹² and the Free Software Foundation¹³, we’ve seen prominent sites such as Facebook¹⁴ and Reddit¹⁵ opting to release their platform’s source code under the CPAL, which is very encouraging.

Because CPAL is a “weak copyleft” license, developers can freely combine it with either free or proprietary software, and only need to release modifications to SocialCalc itself. This enabled various communities to adopt SocialCalc and made it more awesome.

There are many interesting possibilities with this open-source spreadsheet engine, and if you can find a way to embed SocialCalc into your favorite project, we’d definitely love to hear about it.

⁸<http://fit.c2.com/>

⁹<http://search.cpan.org/dist/Test-WWW-Mechanize/>

¹⁰<http://search.cpan.org/dist/Test-WWW-Selenium/>

¹¹<https://www.socialtext.net/open/?cpal>

¹²<http://opensource.org/>

¹³<http://www.fsf.org>

¹⁴<https://github.com/facebook/platform>

¹⁵<https://github.com/reddit/reddit>

Telepathy

Danielle Madeley

Telepathy¹ is a modular framework for real-time communications that handles voice, video, text, file transfer, and so on. What's unique about Telepathy is not that it abstracts the details of various instant messaging protocols, but that it provides the idea of communications as a service, in much the same way that printing is a service, available to many applications at once. To achieve this Telepathy makes extensive use of the D-Bus messaging bus and a modular design.

Communications as a service is incredibly useful, because it allows us to break communications out of a single application. This enables lots of interesting use cases: being able to see a contact's presence in your email application; start communicating with her; launching a file transfer to a contact straight from your file browser; or providing contact-to-contact collaboration within applications, known in Telepathy as *Tubes*.

Telepathy was created by Robert McQueen in 2005 and since that time has been developed and maintained by several companies and individual contributors including Collabora, the company co-founded by McQueen.

20.1 Components of the Telepathy Framework

Telepathy is modular, with each module communicating with the others via a D-Bus messaging bus. Most usually via the user's session bus. This communication is detailed in the Telepathy specification². The components of the Telepathy framework are as shown in 図 20.1:

- A Connection Manager provides the interface between Telepathy and the individual communication services. For instance, there is a Connection Manager for XMPP, one for SIP, one for

¹<http://telepathy.freedesktop.org/>, or see the developers' manual at <http://telepathy.freedesktop.org/doc/book/>

²<http://telepathy.freedesktop.org/spec/>

The D-Bus Message Bus

D-Bus is an asynchronous message bus for interprocess communication that forms the backbone of most GNU/Linux systems including the GNOME and KDE desktop environments. D-Bus is a primarily a shared bus architecture: applications connect to a bus (identified by a socket address) and can either transmit a targeted message to another application on the bus, or broadcast a signal to all bus members. Applications on the bus have a bus address, similar to an IP address, and can claim a number of well-known names, like DNS names, for example `org.freedesktop.Telepathy.AccountManager`. All processes communicate via the D-Bus daemon, which handles message passing, and name registration.

From the user's perspective, there are two buses available on every system. The system bus is a bus that allows the user to communicate with system-wide components (printers, bluetooth, hardware management, etc.) and is shared by all users on the system. The session bus is unique to that user—i.e., there is a session bus per logged-in user—and is used for the user's applications to communicate with each other. When a lot of traffic is to be transmitted over the bus, it's also possible for applications to create their own private bus, or to create a peer-to-peer, unarbitrated bus with no `dbus-daemon`.

Several libraries implement the D-Bus protocol and can communicate with the D-Bus daemon, including `libdbus`, `GDBus`, `QtDBus`, and `python-dbus`. These libraries are responsible for sending and receiving D-Bus messages, marshalling types from the language's type system into D-Bus' type format and publishing objects on the bus. Usually, the libraries also provide convenience APIs for listing connected applications and activatable applications, and requesting well-known names on the bus. At the D-Bus level, all of these are done by making method calls on an object published by `dbus-daemon` itself.

For more information on D-Bus, see <http://www.freedesktop.org/wiki/Software/dbus>.

IRC, and so on. Adding support for a new protocol to Telepathy is simply a matter of writing a new Connection Manager.

- The Account Manager service is responsible for storing the user's communications accounts and establishing a connection to each account via the appropriate Connection Manager when requested.
- The Channel Dispatcher's role is to listen for incoming channels signalled by each Connection Manager and dispatch them to clients that indicate their ability to handle that type of channel, such as text, voice, video, file transfer, tubes. The Channel Dispatcher also provides a service so that applications, most importantly applications that are not Telepathy clients, can request

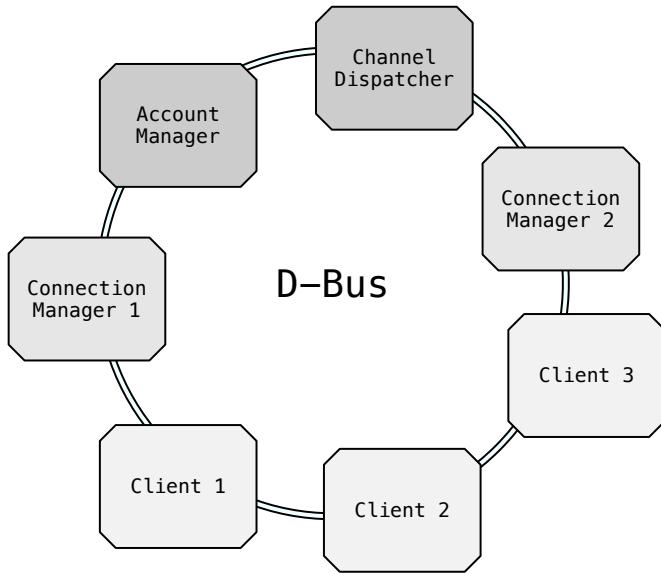


図 20.1: Example Telepathy Components

outgoing channels and have them handled locally by the appropriate client. This allows an application, such as an email application, to request a text chat with a contact, and have your IM client show a chat window.

- Telepathy clients handle or observe communications channels. They include both user interfaces like IM and VoIP clients and services such as the chat logger. Clients register themselves with the Channel Dispatcher, giving a list of channel types they wish to handle or observe.

Within the current implementation of Telepathy, the Account Manager and the Channel Dispatcher are both provided by a single process known as Mission Control.

This modular design was based on Doug McIlroy's philosophy, "Write programs that do one thing and do it well," and has several important advantages:

Robustness: a fault in one component won't crash the entire service.

Ease of development: components can be replaced within a running system without affecting others. It's possible to test a development version of one module against another known to be good.

Language independence: components can be written in any language that has a D-Bus binding.

If the best implementation of a given communications protocol is in a certain language, you are able to write your Connection Manager in that language, and still have it available to all Telepathy clients. Similarly, if you wish to develop your user interface in a certain language, you have access to all available protocols.

License independence: components can be under different software licenses that would be incompatible if everything was running as one process.

Interface independence: multiple user interfaces can be developed on top of the same Telepathy components. This allows native interfaces for desktop environments and hardware devices (e.g., GNOME, KDE, Meego, Sugar).

Security: Components run in separate address spaces and with very limited privileges. For example, a typical Connection Manager only needs access to the network and the D-Bus session bus, making it possible to use something like SELinux to limit what a component can access.

The Connection Manager manages a number of Connections, where each Connection represents a logical connection to a communications service. There is one Connection per configured account. A Connection will contain multiple Channels. Channels are the mechanism through which communications are carried out. A channel might be an IM conversation, voice or video call, file transfer or some other stateful operation. Connections and channels are discussed in detail in 20.3 節。

20.2 How Telepathy uses D-Bus

Telepathy components communicate via a D-Bus messaging bus, which is usually the user's session bus. D-Bus provides features common to many IPC systems: each service publishes objects which have a strictly namespaced object path, like `/org/freedesktop/Telepathy/AccountManager`³. Each object implements a number of interfaces. Again strictly namespaced, these have forms like `org.freedesktop.DBus.Properties` and `ofdT.Connection`. Each interface provides methods, signals and properties that you can call, listen to, or request.

The interfaces, methods, signal and properties provided by Telepathy are detailed in an XML-based D-Bus IDL that has been expanded to include more information. The specification can be parsed to generate documentation and language bindings.

Telepathy services publish a number of objects onto the bus. Mission Control publishes objects for the Account Manager and Channel Dispatcher so that their services can be accessed. Clients publish

³From here on, `/org/freedesktop/Telepathy/` and `org.freedesktop.Telepathy` will be abbreviated to `ofdT` to save space.

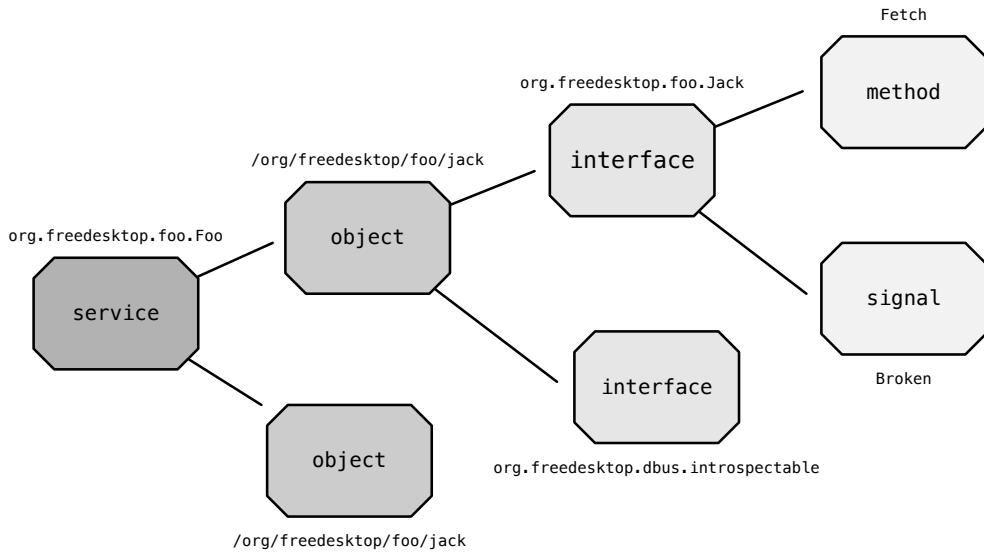


図 20.2: Conceptual Representation of Objects Published by a D-Bus Service

Publishing D-Bus Objects

Publishing D-Bus objects is handled entirely by the D-Bus library being used. In effect it is a mapping from a D-Bus object path to the software object implementing those interfaces. The paths of objects being published by a service are exposed by the optional `org.freedesktop.DBus.Introspectable` interface.

When a service receives an incoming method call with a given destination path (e.g., `/ofdT/AccountManager`), the D-Bus library is responsible for locating the software object providing that D-Bus object and then making the appropriate method call on that object.

a Client object that can be accessed by the Channel Dispatcher. Finally, Connection Managers publish a number of objects: a service object that can be used by the Account Manager to request new connections, an object per open connection, and an object per open channel.

Although D-Bus objects do not have a type (only interfaces), Telepathy simulates types several ways. The object's path tells us whether the object is a connection, channel, client, and so on, though generally you already know this when you request a proxy to it. Each object implements the base interface for that type, e.g., `ofdT.Connection` or `ofdT.Channel`. For channels this is sort of like an abstract base class. Channel objects then have a concrete class defining their channel type. Again, this is represented by a D-Bus interface. The channel type can be learned by reading the

`ChannelType` property on the `Channel` interface.

Finally, each object implements a number of optional interfaces (unsurprisingly also represented as D-Bus interfaces), which depend on the capabilities of the protocol and the Connection Manager. The interfaces available on a given object are available via the `Interfaces` property on the object's base class.

For `Connection` objects of type `ofdT.Connection`, the optional interfaces have names like `ofdT.Connection.Interface` (if the protocol has a concept of avatars), `ofdT.Connection.Interface.ContactList` (if the protocol provides a contact roster—not all do) and `ofdT.Connection.Interface.Location` (if a protocol provides geolocation information). For `Channel` objects, of type `ofdT.Channel`, the concrete classes have interface names of the form `ofdT.Channel.Type.Text`, `ofdT.Channel.Type.Call` and `ofdT.Channel.Type.FileTransfer`. Like `Connections`, optional interface have names like `ofdT.Channel.Interface.Messages` (if this channel can send and receive text messages) and `ofdT.Channel.Interface.Group` (if this channel is to a group containing multiple contacts, e.g., a multi-user chat). So, for example, a text channel implements at least the `ofdT.Channel`, `ofdT.Channel.Type.Text` and `Channel.Interface.Messages` interfaces. If it's a multi-user chat, it will also implement `ofdT.Channel.Interface.Group`.

Why an `Interfaces` Property and not D-Bus Introspection?

You might wonder why each base class implements an `Interfaces` property, instead of relying on D-Bus' introspection capabilities to tell us what interfaces are available. The answer is that different channel and connection objects may offer different interfaces to each other, depending on the capabilities of the channel or connection, but that most of the implementations of D-Bus introspection assume that all objects of the same object class will have the same interfaces. For example, in `telepathy-glib`, the D-Bus interfaces listed by D-Bus introspection are retrieved from the object interfaces a class implements, which is statically defined at compile time. We work around this by having D-Bus introspection provide data for all the interfaces that could exist on an object, and use the `Interfaces` property to indicate which ones actually do.

Although D-Bus itself provides no sanity checking that connection objects only have connection-related interfaces and so forth (since D-Bus has no concept of types, only arbitrarily named interfaces), we can use the information contained within the Telepathy specification to provide sanity checking within the Telepathy language bindings.

Why and How the Specification IDL was Expanded

The existing D-Bus specification IDL defines the names, arguments, access restrictions and D-Bus type signatures of methods, properties and signals. It provides no support for documentation, binding hints or named types.

To resolve these limitations, a new XML namespace was added to provide the required information. This namespace was designed to be generic so that it could be used by other D-Bus APIs. New elements were added to include inline documentation, rationales, introduction and deprecation versions and potential exceptions from methods.

D-Bus type signatures are the low-level type notation of what is serialized over the bus. A D-Bus type signature may look like `(ii)` (which is a structure containing two `int32`s), or it may be more complex. For example, `a{sa(usuu)}`, is a map from string to an array of structures containing `uint32`, `string`, `uint32`, `uint32` (図 20.3). These types, while descriptive of the data format, provide no semantic meaning to the information contained in the type.

In an effort to provide semantic clarity for programmers and strengthen the typing for language bindings, new elements were added to name simple types, structs, maps, enums, and flags, providing their type signature, as well as documentation. Elements were also added in order to simulate object inheritance for D-Bus objects.

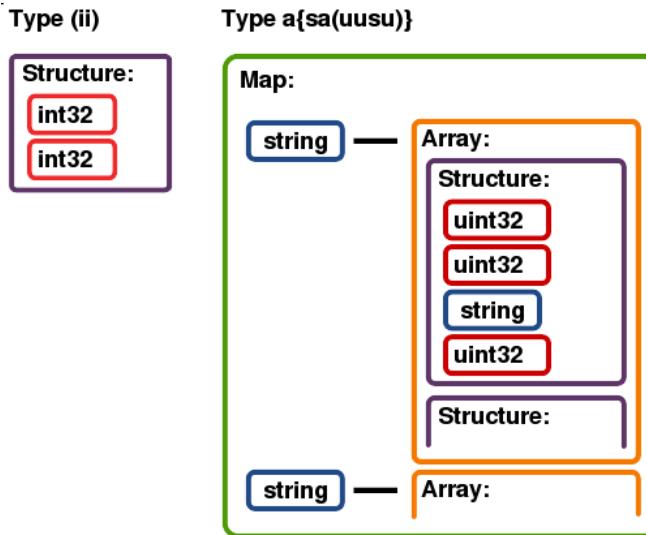


図 20.3: D-Bus Types (ii) and `a{sa(usuu)}`

Handles

Handles are used in Telepathy to represent identifiers (e.g., contacts and room names). They are an unsigned integer value assigned by the connection manager, such that the tuple (connection, handle type, handle) uniquely refers to a given contact or room.

Because different communications protocols normalize identifiers in different ways (e.g., case sensitivity, resources), handles provide a way for clients to determine if two identifiers are the same. They can request the handle for two different identifiers, and if the handle numbers match, then the identifiers refer to the same contact or room.

Identifier normalization rules are different for each protocol, so it is a mistake for clients to compare identifier strings to compare identifiers. For example, `escher@tuxedo.cat/bed` and `escher@tuxedo.cat/lit` are two instances of the same contact (`escher@tuxedo.cat`) in the XMPP protocol, and therefore have the same handle. It is possible for clients to request channels by either identifier or handle, but they should only ever use handles for comparison.

Discovering Telepathy Services

Some services, such as the Account Manager and the Channel Dispatcher, which always exist, have well known names that are defined in the Telepathy specification. However, the names of Connection Managers and clients are not well-known, and must be discovered.

There's no service in Telepathy responsible for the registration of running Connection Managers and Clients. Instead, interested parties listen on the D-Bus for the announcement of a new service. The D-Bus bus daemon will emit a signal whenever a new named D-Bus service appears on the bus. The names of Clients and Connection Managers begin with known prefixes, defined by the specification, and new names can be matched against these.

The advantage of this design is that it's completely stateless. When a Telepathy component is starting up, it can ask the bus daemon (which has a canonical list, based on its open connections) what services are currently running. For instance, if the Account Manager crashes, it can look to see what connections are running, and reassociate those with its account objects.

Connections are Services Too

As well as the Connection Managers themselves, the connections are also advertised as D-Bus services. This hypothetically allows for the Connection Manager to fork each connection off as a separate process, but to date no Connection Manager like this has been implemented. More practically, it allows all running connections to be discovered by querying the D-Bus bus daemon for all services beginning with `ofdT.Connection`.

The Channel Dispatcher also uses this method to discover Telepathy clients. These begin with the name `ofdT.Client`, e.g., `ofdT.Client.Logger`.

Reducing D-Bus Traffic

Original versions of the Telepathy specification created an excessive amount of D-Bus traffic in the form of method calls requesting information desired by lots of consumers on the bus. Later versions of the Telepathy have addressed this through a number of optimizations.

Individual method calls were replaced by D-Bus properties. The original specification included separate method calls for object properties: `GetInterfaces`, `GetChannelType`, etc. Requesting all the properties of an object required several method calls, each with its own calling overhead. By using D-Bus properties, everything can be requested at once using the standard `GetAll` method.

Furthermore, quite a number of properties on a channel are immutable for the lifetime of the channel. These include things like the channel's type, interfaces, who it's connected to and the requestor. For a file transfer channel, for example, it also includes things like the file size and its content type.

A new signal was added to herald the creation of channels (both incoming and in response to outgoing requests) that includes a hash table of the immutable properties. This can be passed directly to the channel proxy constructor (see 20.4 部), which saves interested clients from having to request this information individually.

User avatars are transmitted across the bus as byte arrays. Although Telepathy already used tokens to refer to avatars, allowing clients to know when they needed a new avatar and to save downloading unrequired avatars, each client had to individually request the avatar via a `RequestAvatar` method that returned the avatar as its reply. Thus, when the Connection Manager signalled that a contact had updated its avatar, several individual requests for the avatar would be made, requiring the avatar to be transmitted over the message bus several times.

This was resolved by adding a new method which did not return the avatar (it returns nothing). Instead, it placed the avatar in a request queue. Retrieving the avatar from the network would result in a signal, `AvatarRetrieved`, that all interested clients could listen to. This means the avatar data only needs to be transmitted over the bus once, and will be available to all the interested clients. Once the client's request was in the queue, all further client requests can be ignored until the emission of the `AvatarRetrieved`.

Whenever a large number of contacts need to be loaded (i.e., when loading the contact roster), a significant amount of information needs to be requested: their aliases, avatars, capabilities, and group memberships, and possibly their location, address, and telephone numbers. Previously in Telepathy this would require one method call per information group (most API calls, such as `GetAliases` already took a list of contacts), resulting in half a dozen or more method calls.

To solve this, the `Contacts` interface was introduced. It allowed information from multiple interfaces to be returned via a single method call. The Telepathy specification was expanded to include `Contact Attributes`: namespaced properties returned by the `GetContactAttributes` method that shadowed method calls used to retrieve contact information. A client calls `GetContactAttributes`

with a list of contacts and interfaces it is interested in, and gets back a map from contacts to a map of contact attributes to values.

A bit of code will make this clearer. The request looks like this:

```
connection[CONNECTION_INTERFACE_CONTACTS].GetContactAttributes(  
    [ 1, 2, 3 ], # contact handles  
    [ "ofdT.Connection.Interface.Aliasing",  
      "ofdT.Connection.Interface.Avatars",  
      "ofdT.Connection.Interface.ContactGroups",  
      "ofdT.Connection.Interface.Location"  
    ],  
    False # don't hold a reference to these contacts  
)
```

and the reply might look like this:

```
{ 1: { 'ofdT.Connection.Interface.Aliasing/alias': 'Harvey Cat',  
      'ofdT.Connection.Interface.Avatars/token': hex string,  
      'ofdT.Connection.Interface.Location/location': location,  
      'ofdT.Connection.Interface.ContactGroups/groups': [ 'Squid House' ],  
      'ofdT.Connection/contact-id': 'harvey@nom.cat'  
    },  
  2: { 'ofdT.Connection.Interface.Aliasing/alias': 'Escher Cat',  
      'ofdT.Connection.Interface.Avatars/token': hex string,  
      'ofdT.Connection.Interface.Location/location': location,  
      'ofdT.Connection.Interface.ContactGroups/groups': [],  
      'ofdT.Connection/contact-id': 'escher@tuxedo.cat'  
    },  
  3: { 'ofdT.Connection.Interface.Aliasing/alias': 'Cami Cat',  
      ...  
    }  
}
```

20.3 Connections, Channels and Clients

Connections

A Connection is created by the Connection Manager to establish a connection to a single protocol/account. For example, connecting to the XMPP accounts `escher@tuxedo.cat` and `cami@egg.cat` would result in two Connections, each represented by a D-Bus object. Connections are typically set up by the Account Manager, for the currently enabled accounts.

The Connection provides some mandatory functionality for managing and monitoring the connection status and for requesting channels. It can then also provide a number of optional features, depending on the features of the protocol. These are provided as optional D-Bus interfaces (as discussed in the previous section) and listed by the Connection's `Interfaces` property.

Typically Connections are managed by the Account Manager, created using the properties of the respective accounts. The Account Manager will also synchronize the user's presence for each account to its respective connection and can be asked to provide the connection path for a given account.

Channels

Channels are the mechanism through which communications are carried out. A channel is typically an IM conversation, voice or video call or file transfer, but channels are also used to provide some stateful communication with the server itself, (e.g., to search for chat rooms or contacts). Each channel is represented by a D-Bus object.

Channels are typically between two or more users, one of whom is yourself. They typically have a target identifier, which is either another contact, in the case of one-to-one communication; or a room identifier, in the case of multi-user communication (e.g., a chat room). Multi-user channels expose the Group interface, which lets you track the contacts who are currently in the channel.

Channels belong to a Connection, and are requested from the Connection Manager, usually via the Channel Dispatcher; or they are created by the Connection in response to a network event (e.g., incoming chat), and handed to the Channel Dispatcher for dispatching.

The type of channel is defined by the channel's ChannelType property. The core features, methods, properties, and signals that are needed for this channel type (e.g., sending and receiving text messages) are defined in the appropriate Channel.Type D-Bus interface, for instance Channel.Type.Text. Some channel types may implement optional additional features (e.g., encryption) which appear as additional interfaces listed by the channel's Interfaces property. An example text channel that connects the user to a multi-user chatroom might have the interfaces shown in 表 20.1.

odfT.Channel	Features common to all channels
odfT.Channel.Type.Text	The Channel Type, includes features common to text channels
odfT.Channel.Interface.Messages	Rich-text messaging
odfT.Channel.Interface.Group	List, track, invite and approve members in this channel
odfT.Channel.Interface.Room	Read and set properties such as the chatroom's subject

表 20.1: Example Text Channel

Requesting Channels, Channel Properties and Dispatching

Channels are requested using a map of properties you wish the desired channel to possess. Typically, the channel request will include the channel type, target handle type (contact or room) and

Contact List Channels: A Mistake

In the first versions of the Telepathy specification, contact lists were considered a type of channel. There were several server-defined contact lists (subscribed users, publish-to users, blocked users), that could be requested from each Connection. The members of the list were then discovered using the Group interface, like for a multi-user chat.

Originally this would allow for channel creation to occur only once the contact list had been retrieved, which takes time on some protocols. A client could request the channel whenever it liked, and it would be delivered once ready, but for users with lots of contacts this meant the request would occasionally time out. Determining the subscription/publish/blocked status of a client required checking three channels.

Contact Groups (e.g., Friends) were also exposed as channels, one channel per group. This proved extremely difficult for client developers to work with. Operations like getting the list of groups a contact was in required a significant amount of code in the client. Further, with the information only available via channels, properties such as a contact's groups or subscription state could not be published via the Contacts interface.

Both channel types have since been replaced by interfaces on the Connection itself which expose contact roster information in ways more useful to client authors, including subscription state of a contact (an enum), groups a contact is in, and contacts in a group. A signal indicates when the contact list has been prepared.

target. However, a channel request may also include properties such as the filename and filesize for file transfers, whether to initially include audio and video for calls, what existing channels to combine into a conference call, or which contact server to conduct a contact search on.

The properties in the channel request are properties defined by interfaces of the Telepathy spec, such as the ChannelType property (表 20.2). They are qualified with the namespace of the interface they come from. Properties which can be included in channel requests are marked as *requestable* in the Telepathy spec.

Property	Value
ofdT.Channel.ChannelType	ofdT.Channel.Type.Text
ofdT.Channel.TargetHandleType	Handle_Type_Contact (1)
ofdT.Channel.TargetID	escher@tuxedo.cat

表 20.2: Example Channel Requests

The more complicated example in 表 20.3 requests a file transfer channel. Notice how the requested properties are qualified by the interface from which they come. (For brevity, not all required properties are shown.)

Property	Value
ofdT.Channel.ChannelType	ofdT.Channel.Type.FileTransfer
ofdT.Channel.TargetHandleType	Handle_Type_Contact (1)
ofdT.Channel.TargetID	escher@tuxedo.cat
ofdT.Channel.Type.FileTransfer.Filename	meow.jpg
ofdT.Channel.Type.FileTransfer.ContentType	image/jpeg

表 20.3: File Transfer Channel Request

Channels can either be *created* or *ensured*. Ensuring a channel means creating it only if it does not already exist. Asking to create a channel will either result in a completely new and separate channel being created, or in an error being generated if multiple copies of such a channel cannot exist. Typically you wish to ensure text channels and calls (i.e., you only need one conversation open with a person, and in fact many protocols do not support multiple separate conversations with the same contact), and wish to create file transfers and stateful channels.

Newly created channels (requested or otherwise) are announced by a signal from the Connection. This signal includes a map of the channel's *immutable* properties. These are the properties which are guaranteed not to change throughout the channel's lifetime. Properties which are considered immutable are marked as such in the Telepathy spec, but typically include the channel's type, target handle type, target, initiator (who created the channel) and interfaces. Properties such as the channel's state are obviously not included.

Old-School Channel Requesting

Channels were originally requested simply by type, handle type and target handle. This wasn't sufficiently flexible because not all channels have a target (e.g., contact search channels), and some channels require additional information included in the initial channel request (e.g., file transfers, requesting voicemails and channels for sending SMSes). It was also discovered that two different behaviors might be desired when a channel was requested (either to create a guaranteed unique channel, or simply ensure a channel existed), and until this time the Connection had been responsible for deciding which behavior would occur. Hence, the old method was replaced by the newer, more flexible, more explicit ones.

Returning a channel's immutable properties when you create or ensure the channel makes it much faster to create a proxy object for the channel. This is information we now don't have to request. The map in 表 20.4 shows the immutable properties that might be included when we request a text channel (i.e., using the channel request in 表 20.3). Some properties (including TargetHandle and InitiatorHandle) have been excluded for brevity.

Property	Value
ofdT.Channel.ChannelType	Channel.Type.Text
ofdT.Channel.Interfaces	[Channel.Interface.Messages, Channel.Interface.Destroyable, Channel.Interface.ChatState]
ofdT.Channel.TargetHandleType	Handle_Type_Contact (1)
ofdT.Channel.TargetID	escher@tuxedo.cat
ofdT.Channel.InitiatorID	danielle.madeley@collabora.co.uk
ofdT.Channel.Requested	True
ofdT.Channel.Interface.Messages.	[text/html, text/plain]
SupportedContentTypes	

表 20.4: Example Immutable Properties Returned by a New Channel

The requesting program typically makes a request for a channel to the Channel Dispatcher, providing the account the request is for, the channel request, and optionally the name of a the desired handler (useful if the program wishes to handle the channel itself). Passing the name of an account instead of a connection means that the Channel Dispatcher can ask the Account Manager to bring an account online if required.

Once the request is complete, the Channel Dispatcher will either pass the channel to the named Handler, or locate an appropriate Handler (see below for discussion on Handlers and other clients). Making the name of the desired Handler optional makes it possible for programs that have no interest in communication channels beyond the initial request to request channels and have them handled by the best program available (e.g., launching a text chat from your email client).

The requesting program makes a channel request to the Channel Dispatcher, which in turn forwards the request to the appropriate Connection. The Connection emits the NewChannels signal which is picked up by the Channel Dispatcher, which then finds the appropriate client to handle the channel. Incoming, unrequested channels are dispatched in much the same way, with a signal from the Connection that is picked up by the Channel Dispatcher, but obviously without the initial request from a program.

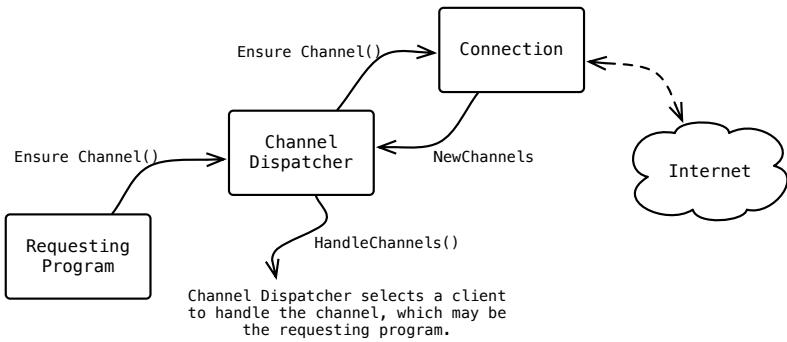


図 20.4: Channel Request and Dispatching

Clients

Clients handle or observe incoming and outgoing communications channels. A client is anything that is registered with the Channel Dispatcher. There are three types of clients (though a single client may be two, or all three, types if the developer wishes):

Observers: Observe channels without interacting with them. Observers tend to be used for chat and activity logging (e.g., incoming and outgoing VoIP calls).

Approvers: Responsible for giving users an opportunity to accept or reject an incoming channel.

Handlers: Actually interact with the channel. That might be acknowledging and sending text messages, sending or receiving a file, etc. A Handler tends to be associated with a user interface.

Clients offer D-Bus services with up to three interfaces: `Client.Observer`, `Client.Approver`, and `Client.Handler`. Each interface provides a method that the Channel Dispatcher can call to inform the client about a channel to observe, approve or handle.

The Channel Dispatcher dispatches the channel to each group of clients in turn. First, the channel is dispatched to all appropriate Observers. Once they have all returned, the channel is dispatched to all the appropriate Approvers. Once the first Approver has approved or rejected the channel, all other Approvers are informed and the channel is finally dispatched to the Handler. Channel dispatching is done in stages because Observers might need time to get set up before the Handler begins altering the channel.

Clients expose a channel filter property which is a list of filters read by the Channel Dispatcher so that it knows what sorts of channels a client is interested in. A filter must include at least the channel type, and target handle type (e.g., contact or room) that the client is interested in, but it can contain more properties. Matching is done against the channel's immutable properties, using simple equality for comparison. The filter in 表 20.5 matches all one-to-one text channels.

Clients are discoverable via D-Bus because they publish services beginning with the well-known name `ofdT.Client` (for example `ofdT.Client.Empathy.Chat`). They can also optionally install a

Property	Value
ofdT.Channel.ChannelType	Channel.Type.Text
ofdT.Channel.TargetHandleType	Handle_Type_Contact (1)

表 20.5: Example Channel Filter

file which the Channel Dispatcher will read specifying the channel filters. This allows the Channel Dispatcher to start a client if it is not already running. Having clients be discoverable in this way makes the choice of user interface configurable and changeable at any time without having to replace any other part of Telepathy.

All or Nothing

It is possible to provide a filter indicating you are interested in all channels, but in practice this is only useful as an example of observing channels. Real clients contain code that is specific to channel types.

An empty filter indicates a Handler is not interested in any channel types. However it is still possible to dispatch a channel to this handler if you do so by name. Temporary Handlers which are created on demand to handle a specific channel use such a filter.

20.4 The Role of Language Bindings

As Telepathy is a D-Bus API, and thus can driven by any programming language that supports D-Bus. Language bindings are not required for Telepathy, but they can be used to provide a convenient way to use it.

Language bindings can be split into two groups: low-level bindings that include code generated from the specification, constants, method names, etc.; and high-level bindings, which are hand-written code that makes it easier for programmers to do things using Telepathy. Examples of high-level bindings are the GLib and Qt4 bindings. Examples of low-level bindings are the Python bindings and the original libtelepathy C bindings, though the GLib and Qt4 bindings include a low-level binding.

Asynchronous Programming

Within the language bindings, all method calls that make requests over D-Bus are asynchronous: the request is made, and the reply is given in a callback. This is required because D-Bus itself is asynchronous.

Like most network and user interface programming, D-Bus requires the use of an event loop to dispatch callbacks for incoming signals and method returns. D-Bus integrates well with the GLib mainloop used by the GTK+ and Qt toolkits.

Some D-Bus language bindings (such as dbus-glib) provide a pseudo-synchronous API, where the main loop is blocked until the method reply is returned. Once upon a time this was exposed via the telepathy-glib API bindings. Unfortunately using pseudo-synchronous API turns out to be fraught with problems, and was eventually removed from telepathy-glib.

Why Pseudo-Synchronous D-Bus Calls Don't Work

The pseudo-synchronous interface offered by dbus-glib and other D-Bus bindings is implemented using a request-and-block technique. While blocking, only the D-Bus socket is polled for new I/O and any D-Bus messages that are not the response to the request are queued for later processing.

This causes several major and inescapable problems:

- The caller is blocked while waiting for the request to be answered. It (and its user interface, if any) will be completely unresponsive. If the request requires accessing the network, that takes time; if the callee has locked up, the caller will be unresponsive until the call times out.

Threading is not a solution here because threading is just another way of making your calling asynchronous. Instead you may as well make asynchronous calls where the responses come in via the existing event loop.

- Messages may be reordered. Any messages received before the watched-for reply will be placed on a queue and delivered to the client after the reply.

This causes problems in situations where a signal indicating a change of state (i.e., the object has been destroyed) is now received after the method call on that object fails (i.e., with the exception `UnknownMethod`). In this situation, it is hard to know what error to display to the user. Whereas if we receive a signal first, we can cancel pending D-Bus method calls, or ignore their responses.

- Two processes making pseudo-blocking calls on each other can deadlock, with each waiting for the other to respond to its query. This scenario can occur with processes that are both a D-Bus service and call other D-Bus services (for example, Telepathy clients). The Channel Dispatcher calls methods on clients to dispatch channels, but clients also call methods on the Channel Dispatcher to request the opening of new channels (or equally they call the Account Manager, which is part of the same process).

Method calls in the first Telepathy bindings, generated in C, simply used `typedef` callback functions. Your callback function simply had to implement the same type signature.

```
typedef void (*tp_conn_get_self_handle_reply) (
    DBusGProxy *proxy,
    guint handle,
    GError **error,
    gpointer userdata
);
```

This idea is simple, and works for C, so was continued into the next generation of bindings.

In recent years, people have developed a way to use scripting languages such as Javascript and Python, as well as a C#-like language called Vala, that use GLib/GObject-based APIs via a tool called GObject-Introspection. Unfortunately, it's extremely difficult to rebind these types of callbacks into other languages, so newer bindings are designed to take advantage of the asynchronous callback features provided by the languages and GLib.

Object Readiness

In a simple D-Bus API, such as the low-level Telepathy bindings, you can start making method calls or receive signals on a D-Bus object simply by creating a proxy object for it. It's as simple as giving an object path and interface name and getting started.

However, in Telepathy's high-level API, we want our object proxies to know what interface are available, we want common properties for the object type to be retrieved (e.g., the channel type, target, initiator), and we want to determine and track the object's state or status (e.g., the connection status).

Thus, the concept of *readiness* exists for all proxy objects. By making a method call on a proxy object, you are able to asynchronously retrieve the state for that object and be notified when state is retrieved and the object is ready for use.

Since not all clients implement, or are interested in, all the features of a given object, readiness for an object type is separated into a number of possible features. Each object implements a *core* feature, which will prepare crucial information about the object (i.e., its `Interfaces` property and basic state), plus a number of optional features for additional state, which might include extra properties or state-tracking. Specific examples of additional features you can ready on various proxies are contact info, capabilities, geolocation information, chat states (such as "Escher is typing...") and user avatars.

For example, connection object proxies have:

- a core feature which retrieves the interface and connection status;
- features to retrieve the requestable channel classes and support contact info; and
- a feature to establish a connection and return ready when connected.

The programmer requests that the object is readied, providing a list of features in which they are interested and a callback to call when all of those features are ready. If all the features are already ready, the callback can be called immediately, else the callback is called once all the information for those features is retrieved.

20.5 Robustness

One of the key advantages of Telepathy is its robustness. The components are modular, so a crash in one component should not bring down the whole system. Here are some of the features that make Telepathy robust:

- The Account Manager and Channel Dispatcher can recover their state. When Mission Control (the single process that includes the Account Manager and Channel Dispatcher) starts, it looks at the names of services currently registered on the user's session bus. Any Connections it finds that are associated with a known account are reassigned with that account (rather than a new connection being established), and running clients are queried for the list of channels they're handling.
- If a client disappears while a channel it's handling is open, the Channel Dispatcher will respawn it and reissue the channel.

If a client repeatedly crashes the Channel Dispatcher can attempt to launch a different client, if available, or else it will close the channel (to prevent the client repeatedly crashing on data it can't handle).

Text messages require acknowledgment before they will disappear from the list of pending messages. A client is only meant to acknowledge a message once it is sure the user has seen it (that is, displayed the message in a focused window). This way if the client crashes trying to render the message, the channel will still have the previously undisplayed message in the pending message queue.

- If a Connection crashes, the Account Manager will respawn it. Obviously the content of any stateful channels will be lost, but it will only affect the Connections running in that process and no others. Clients can monitor the state of the connections and simply re-request information like the contact roster and any stateless channels.

20.6 Extending Telepathy: Sidecars

Although the Telepathy specification tries to cover a wide range of features exported by communication protocols, some protocols are themselves extensible⁴. Telepathy's developers wanted to

⁴E.g., the Extensible Messaging and Presence Protocol (XMPP).

make it possible extend your Telepathy connections to make use of such extensions without having to extend the Telepathy specification itself. This is done through the use of *sidecars*.

Sidecars are typically implemented by plugins in a Connection Manager. Clients call a method requesting a sidecar that implements a given D-Bus interface. For example, someone's implementation of XEP-0016 privacy lists might implement an interface named `com.example.PrivacyLists`. The method then returns a D-Bus object provided by the plugin, which should implement that interface (and possibly others). The object exists alongside the main Connection object (hence the name sidecar, like on a motorcycle).

The History of Sidecars

In the early days of Telepathy, the One Laptop Per Child project needed to support custom XMPP extensions (XEPs) to share information between devices. These were added directly to Telepathy-Gabble (the XMPP Connection Manager), and exposed via undocumented interfaces on the Connection object. Eventually, with more developers wanting support for specific XEPs which have no analogue in other communications protocols, it was agreed that a more generic interface for plugins was needed.

20.7 A Brief Look Inside a Connection Manager

Most Connection Managers are written using the C/GLib language binding, and a number of high-level base classes have been developed to make writing a Connection Manager easier. As discussed previously, D-Bus objects are published from software objects that implement a number of software interfaces that map to D-Bus interfaces. Telepathy-GLib provides base objects to implement the Connection Manager, Connection and Channel objects. It also provides an interface to implement a Channel Manager. Channel Managers are factories that can be used by the BaseConnection to instantiate and manage channel objects for publishing on the bus.

The bindings also provide what are known as *mixins*. These can be added to a class to provide additional functionality, abstract the specification API and provide backwards compatibility for new and deprecated versions of an API through one mechanism. The most commonly used mixin is one that adds the D-Bus properties interface to an object. There are also mixins to implement the `ofdT.Connection.Interface.Contacts` and `ofdT.Channel.Interface.Group` interfaces and mixins making it possible to implement the old and new presence interfaces, and old and new text message interfaces via one set of methods.

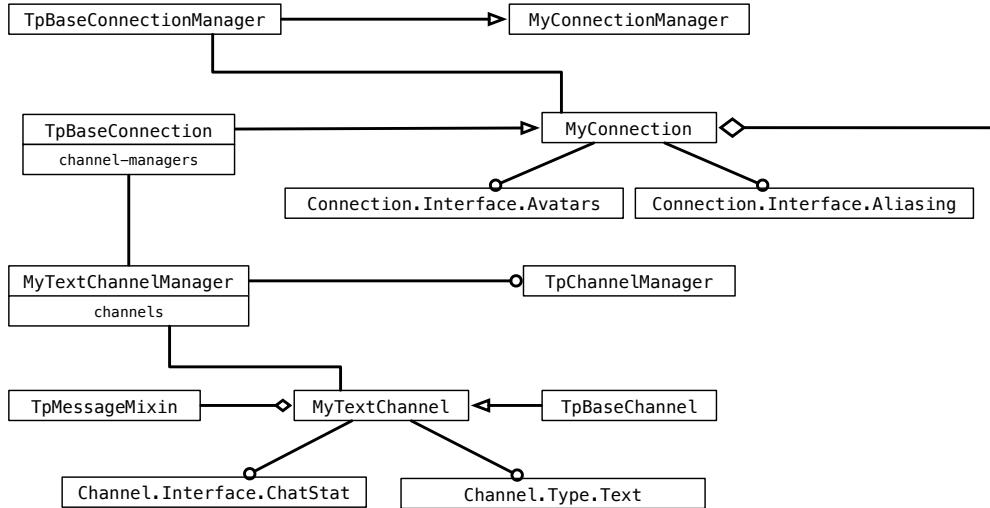


図 20.5: Example Connection Manager Architecture

Using Mixins to Solve API Mistakes

One place where mixins have been used to solve a mistake in the Telepathy specification is the **TpPresenceMixin**. The original interface exposed by Telepathy (`odfT.Connection.Interface.Presence`) was incredibly complicated, hard to implement for both Connections and Clients, and exposed functionality that was both nonexistent in most communications protocols, and very rarely used in others. The interface was replaced by a much simpler interface (`odfT.Connection.Interface.SimplePresence`), which exposed all the functionality that users cared about and had ever actually been implemented in the connection managers.

The presence mixin implements both interfaces on the Connection so that legacy clients continue to work, but only at the functionality level of the simpler interface.

20.8 Lessons Learned

Telepathy is an excellent example of how to build a modular, flexible API on top of D-Bus. It shows how you can develop an extensible, decoupled framework on top of D-Bus. One which requires no central management daemon and allows components to be restartable, without loss of data in any other component. Telepathy also shows how you can use D-Bus efficiently and effectively, minimizing the amount of traffic you transmit on the bus.

Telepathy's development has been iterative, improving its use of D-Bus as time goes on. Mistakes

were made, and lessons have been learned. Here are some of the important things we learned in designing the architecture of Telepathy:

Use D-Bus properties; don't require dozens of small D-Bus method calls to look up information.

Every method call has a round-trip time. Rather than making lots of individual calls (e.g., `GetHandle`, `GetChannelType`, `GetInterfaces`) use D-Bus properties and return all the information via a single call to `GetAll`.

Provide as much information as you can when announcing new objects. The first thing clients used to do when they learned about a new object was to request all of its properties to learn whether they were even interested in the object. By including the immutable properties of an object in the signal announcing the object, most clients can determine their interest in the object without making any method calls. Furthermore, if they are interested in the object, they do not have to bother requesting any of its immutable properties.

The `Contacts` interface allows requesting information from multiple interfaces at once. Rather than making numerous `GetAll` calls to retrieve all the information for a contact, the `Contacts` interface lets us request all the information at once, saving a number of D-Bus round trips.

Don't use abstractions that don't quite fit. Exposing the contact roster and contact groups as channels implementing the `Group` interface seemed like a good idea because it used existing abstractions rather than requiring additional interfaces. However, it made implementing clients difficult and was ultimately not suitable.

Ensure your API will meet your future needs. The original channel requesting API was very rigid, only permitting very basic channel requests. This did not meet our needs when needing to request channels that required more information. This API had to be replaced with one that had significantly more flexibility.

Thousand Parsec

Alan Laudicina and Aaron Mavrinac

A vast star empire encompasses a hundred worlds, stretching a thousand parsecs across space. Unlike some other areas of the galaxy, few warriors live here; this is an intellectual people, with a rich cultural and academic tradition. Their magnificent planets, built turn by turn around great universities of science and technology, are a beacon of light to all in this age of peace and prosperity. Starships arrive from the vast reaches of the quadrant and beyond, bearing the foremost researchers from far and wide. They come to contribute their skills to the most ambitious project ever attempted by sentient beings: the development of a decentralized computer network to connect the entire galaxy, with all its various languages, cultures, and systems of law.

Thousand Parsec is more than a video game: it is a framework, with a complete toolkit for building multiplayer, turn-based space empire strategy games. Its generic game protocol allows diverse implementations of client, server, and AI software, as well as a vast array of possible games. Though its size has made planning and execution challenging, forcing contributors to walk a thin line between excessively vertical and excessively horizontal development, it also makes it a rather interesting specimen when discussing the architecture of open source applications.

The journalist's label for the genre Thousand Parsec games inhabit is "4X"—shorthand for "explore, expand, exploit, and exterminate," the modus operandi of the player controlling an empire¹. Typically in the 4X genre of games, players will scout to reveal the map (explore), create new settlements or extend the influence of existing ones (expand), gather and use resources in areas they control (exploit), and attack and eliminate rival players (exterminate). The emphasis on economic and technological development, micromanagement, and variety of routes to supremacy yield a depth and complexity of gameplay unparalleled within the greater strategy genre.

From a player's perspective, three main components are involved in a game of Thousand Parsec.

¹Some excellent commercial examples of Thousand Parsec's inspiration include *VGA Planets* and *Stars!*, as well as the *Master of Orion*, *Galactic Civilizations*, and *Space Empires* series. For readers unfamiliar with these titles, the *Civilization* series is a popular example of the same gameplay style, albeit in a different setting. A number of real-time 4X games also exist, such as *Imperium Galactica* and *Sins of a Solar Empire*.

First, there is the client: this is the application through which the player interacts with the universe. This connects to a server over the network—communicating using the all-important protocol—to which other players’ (or, in some cases, artificial intelligence) clients are also connected. The server stores the entire game state, updating clients at the start of each turn. Players can then perform various actions and communicate them back to the server, which computes the resulting state for the next turn. The nature of the actions a player may perform is dictated by a ruleset: this in essence defines the game being played, implemented and enforced on the server side, and actualized for the player by any supporting client.

Because of the diversity of possible games, and the complexity of the architecture required to support this diversity, Thousand Parsec is an exciting project both for gamers and for developers. We hope that even the serious coder with little interest in the anatomy of game frameworks might find value in the underlying mechanics of client-server communication, dynamic configuration, meta-data handling, and layered implementation, all of which have grown rather organically toward good design over the years in quintessential open source style.

At its core, Thousand Parsec is primarily a set of standard specifications for a game protocol and other related functionality. This chapter discusses the framework mostly from this abstract viewpoint, but in many cases it is much more enlightening to refer to actual implementations. To this end, the authors have chosen the “flagship” implementations of each major component for concrete discussion.

The case model client is `tpclient-pywx`, a relatively mature wxPython-based client which at present supports the largest set of features and the latest game protocol version. This is supported by `libtpclient-py`, a Python client helper library providing caching and other functionality, and `libtpproto-py`, a Python library which implements the latest version of the Thousand Parsec protocol. For the server, `tpserver-cpp`, the mature C++ implementation supporting the latest features and protocol version, is the specimen. This server sports numerous rulesets, among which the *Missile and Torpedo Wars* milestone ruleset is exemplary for making the most extensive use of features and for being a "traditional" 4X space game.

21.1 Anatomy of a Star Empire

In order to properly introduce the things that make up a Thousand Parsec universe, it makes sense first to give a quick overview of a game. For this, we’ll examine the *Missile and Torpedo Wars* ruleset, the project’s second milestone ruleset, which makes use of most of the major features in the current mainline version of the Thousand Parsec protocol. Some terminology will be used here which will not yet be familiar; the remainder of this section will elucidate it so that the pieces all fall into place.

Missile and Torpedo Wars is an advanced ruleset in that it implements all of the methods available

in the Thousand Parsec framework. At the time of writing, it is the only ruleset to do so, and it is being quickly expanded to become a more complete and entertaining game.

Upon establishing a connection to a Thousand Parsec server, the client probes the server for a list of game entities and proceeds to download the entire catalog. This cataloger includes all of the objects, boards, messages, categories, designs, components, properties, players, and resources that make up the state of the game, all of which are covered in detail in this section. While this may seem like a lot for the client to digest at the beginning of the game—and also at the end of each turn—this information is absolutely vital for the game. Once this information has been downloaded, which generally takes on the order of a few seconds, the client now has everything it needs to plot the information onto its representation of the game universe.

When first connected to the server, a random planet is generated and assigned as the new player's "home planet", and two fleets are automatically created there. Each fleet consists of two default Scout designs, consisting of a Scout Hull with an Alpha Missile Tube. Since there is no Explosive component added, this default fleet is not yet capable of fleet-to-fleet or fleet-to-planet combat; it is, in fact, a sitting duck.

At this point, it is important for a player to begin equipping fleets with weaponry. This is achieved by creating a weapon design using a Build Weapon order, and then loading the finished product onto the target fleet through a Load Armament order. The Build Weapon order converts a planet's resources—of which each planet has amounts and proportions assigned by a random distribution—into a finished product: an explosive warhead which is planted on the creating planet's surface. The Load Armament order then transfers this completed weapon onto a waiting fleet.

Once the easily accessible surface resources of a planet are used up, it is important to obtain more through mining. Resources come in two other states: mineable and inaccessible. Using a Mine order on a planet, mineable resources may be converted over time into surface resources, which can then be used for building.

Objects

In a Thousand Parsec universe, every physical thing is an object. In fact, the universe itself is also an object. This design allows for a virtually unlimited set of elements in a game, while remaining simple for rulesets which require only a few types of objects. On top of the addition of new object types, each object can store some of its own specific information that can be sent and used via the Thousand Parsec protocol. Five basic built-in object types are currently provided by default: Universe, Galaxy, Star System, Planet, and Fleet.

The Universe is the top-level object in a Thousand Parsec game, and it is always accessible to all players. While the Universe object does not actually exert much control over the game, it does store one vastly important piece of information: the current turn number. Also known as the “year” in Thousand Parsec parlance, the turn number, naturally, increments after the completion of each turn. It is stored in an unsigned 32-bit integer, allowing for games to run until year 4,294,967,295. While not impossible in theory, the authors have not, to date, seen a game progress this far.

A Galaxy is a container for a number of proximate objects—Star Systems, Planets and Fleets—and provides no additional information. A large number of Galaxies may exist in a game, each hosting a subsection of the Universe.

Like the previous two objects, a Star System is primarily a container for lower-level objects. However, the Star System object is the first tier of object which is represented graphically by the client. These objects may contain Planets and Fleets (at least temporarily).

A Planet is a large celestial body which may be inhabited and provide resource mines, production facilities, ground-based armaments, and more. The Planet is the first tier of object which can be owned by a player; ownership of a Planet is an accomplishment not to be taken lightly, and not owning any planets is a typical condition for rulesets to proclaim a player's defeat. The Planet object has a relatively large amount of stored data, accounting for the following:

- The player ID of the Planet's owner (or -1 if not owned by any player).

- A list of the Planet’s resources, containing the resource ID (type), and the amount of surface, mineable, and inaccessible resources of this type on the Planet.

The built-in objects described above provide a good basis for many rulesets following the traditional 4X space game formula. Naturally, in keeping with good software engineering principles, object classes can be extended within rulesets. A ruleset designer thus has the ability to create new object types or store additional information in the existing object types as required by the ruleset, allowing for virtually unlimited extensibility in terms of the available physical objects in the game.

Orders

Defined by each ruleset, orders can be attached to both Fleet and Planet objects. While the core server does not ship with any default order types, these are an essential part of even the most basic game. Depending on the nature of the ruleset, orders may be used to accomplish almost any task. In the spirit of the 4X genre, there are a few standard orders which are implemented in most rulesets: these are the Move, Intercept, Build, Colonize, Mine, and Attack orders.

In order to fulfill the first imperative (explore) of 4X, one needs to be able to move about the map of the universe. This is typically achieved via a Move order appended to a Fleet object. In the flexible and extensible spirit of the Thousand Parsec framework, Move orders can be implemented differently depending on the nature of the ruleset. In *Minisec* and *Missile and Torpedo Wars*, a Move order typically takes a point in 3D space as a parameter. On the server side, the estimated time of arrival is calculated and the number of required turns is sent back to the client. The Move order also acts as a pseudo-Attack order in rulesets where teamwork is not implemented. For example, moving to a point occupied by an enemy fleet in both Minisec and Missile and Torpedo Wars is almost certain to be followed by a period of intense combat. Some rulesets supporting a Move order parameterize it differently (i.e. not using 3D points). For example, the *Risk* ruleset only allows single-turn moves to planets which are directly connected by a “wormhole”.

Typically appended to Fleet objects, the Intercept order allows an object to meet another (commonly an enemy fleet) within space. This order is similar to Move, but since two objects might be moving in different directions during the execution of a turn, it is impossible to land directly on another fleet simply using spatial coordinates, so a distinct order type is necessary. The Intercept order addresses this issue, and can be used to wipe out an enemy fleet in deep space or fend off an oncoming attack in a moment of crisis.

The Build order helps to fulfill two of the 4X imperatives—expand and exploit. The obvious means of expansion throughout the universe is to build many fleets of ships and move them far and wide. The Build order is typically appended to Planet objects and is often bound to the amount of resources that a planet contains—and how they are exploited. If a player is lucky enough to

have a home planet rich in resources, that player could gain an early advantage in the game through building.

Like the Build order, the Colonize order helps fulfill the expand and exploit imperatives. Almost always appended to Fleet objects, the Colonize order allows the player to take over an unclaimed planet. This helps to expand control over planets throughout the universe.

The Mine order embodies the exploit imperative. This order, typically appended to Planet objects and other celestial bodies, allows the player to mine for unused resources not immediately available on the surface. Doing so brings these resources to the surface, allowing them to be used subsequently to build and ultimately expand the player's grip on the universe.

Implemented in some rulesets, the Attack order allows a player to explicitly initiate combat with an enemy Fleet or Planet, fulfilling the final 4X imperative (exterminate). In team-based rulesets, the inclusion of a distinct Attack order (as opposed to simply using Move and Intercept to implicitly attack targets) is important to avoid friendly fire and to coordinate attacks.

Since the Thousand Parsec framework requires ruleset developers to define their own order types, it is possible—even encouraged—for them to think outside the box and create custom orders not found elsewhere. The ability to pack extra data into any object allows developers to do very interesting things with custom order types.

Resources

Resources are extra pieces of data that are packed into Objects in the game. Extensively used—particularly by Planet objects—resources allow for easy extension of rulesets. As with many of the design decisions in Thousand Parsec, extensibility was the driving factor in the inclusion of resources.

While resources are typically implemented by the ruleset designer, there is one resource that is in consistent use throughout the framework: the Home Planet resource, which is used to identify a player's home planet.

According to Thousand Parsec best practices, resources are typically used to represent something that can be converted into some type of object. For example, Minisec implements a Ship Parts resource, which is assigned in random quantities to each planet object in the universe. When one of these planets is colonized, you can then convert this Ship Parts resource into actual Fleets using a Build order.

Missile and Torpedo Wars makes perhaps the most extensive use of resources of any ruleset to date. It is the first ruleset where the weapons are of a dynamic nature, meaning that they can be added to a ship from a planet and also removed from a ship and added back to a planet. To account for this, the game creates a resource type for each weapon that is created in the game. This allows ships to identify a weapon type by a resource, and move them freely throughout the universe. *Missile and*

Torpedo Wars also keeps track of factories (the production capability of planets) using a Factories resource tied to each planet.

Designs

In Thousand Parsec, both weapons and ships may be composed of various components. These components are combined to form the basis of a Design—a prototype for something which can be built and used within the game. When creating a ruleset, the designer has to make an almost immediate decision: should the ruleset allow dynamic creation of weapon and ship designs, or simply use a predetermined list of designs? On the one hand, a game using pre-packaged designs will be easier to develop and balance, but on the other hand, dynamic creation of designs adds an entirely new level of complexity, challenge, and fun to the game.

User-created designs allow a game to become far more advanced. Since users must strategically design their own ships and their armaments, a stratum of variance is added to the game which can help to mitigate otherwise great advantages that might be conferred on a player based on luck (e.g., of placement) and other aspects of game strategy. These designs are governed by the rules of each component, outlined in the Thousand Parsec Component Language (TPCL, covered later in this chapter), and specific to each ruleset. The upshot is that no additional programming of functionality is necessary on the part of the developer to implement the design of weapons and ships; configuring some simple rules for each component available in the ruleset is sufficient.

Without careful planning and proper balance, the great advantage of using custom designs can become its downfall. In the later stages of a game, an inordinate amount of time can be spent designing new types of weapons and ships to build. The creation of a good user experience on the client side for design manipulation is also a challenge. Since design manipulation can be an integral part of one game, while completely irrelevant to another, the integration of a design window into clients is a significant obstacle. Thousand Parsec’s most complete client, `tpclient-pywx`, currently houses the launcher for this window in a relatively out-of-the-way place, in a sub-menu of the menu bar (which is rarely used in-game otherwise).

The Design functionality is designed to be easily accessible to ruleset developers, while allowing games to expand to virtually unlimited levels of complexity. Many of the existing rulesets allow for only predetermined designs. *Missile and Torpedo Wars*, however, allows for full weapon and ship design from a variety of components.

21.2 The Thousand Parsec Protocol

One might say that the Thousand Parsec protocol is the basis upon which everything else in the project is built. It defines the features available to ruleset writers, how servers should work, and what

clients should be able to handle. Most importantly, like an interstellar communications standard, it allows the various software components to understand one another.

The server manages the actual state and dynamics of a game according to the instructions provided by the ruleset. Each turn, a player’s client receives some of the information about the state of the game: objects and their ownership and current state, orders in progress, resource stockpiles, technological progress, messages, and everything else visible to that particular player. The player can then perform certain actions given the current state, such as issuing orders or creating designs, and send these back to the server to be processed into the computation of the next turn. All of this communication is framed in the Thousand Parsec protocol. An interesting and quite deliberate effect of this architecture is that AI clients—which are external to the server/ruleset and are the only means of providing computer players in a game—are bound by the same rules as the clients human players use, and thus cannot “cheat” by having unfair access to information or by being able to bend the rules.

The protocol specification describes a series of frames, which are hierarchical in the sense that each frame (except the Header frame) has a base frame type to which it adds its own data. There are a variety of abstract frame types which are never explicitly used, but simply exist to describe bases for concrete frames. Frames may also have a specified direction, with the intent that such frames need only be supported for sending by one side (server or client) and receiving by the other.

The Thousand Parsec protocol is designed to function either standalone over TCP/IP, or tunneled through another protocol such as HTTP. It also supports SSL encryption.

Basics

The protocol provides a few generic frames which are ubiquitous in communication between client and server. The previously mentioned Header frame simply provides a basis for all other frames via its two direct descendants, the Request and Response frames. The former is the basis for frames which initiate communication (in either direction), and the latter for frames which are prompted by these. The OK and Fail frames (both Response frames) provide the two values for Boolean logic in the exchange. A Sequence frame (also a Response) indicates to the recipient that multiple frames are to follow in response to its request.

Thousand Parsec uses numerical IDs to address things. Accordingly, a vocabulary of frames exists to push around data via these IDs. The Get With ID frame is the basic request for things with such an ID; there is also a Get With ID and Slot frame for things which are in a “slot” on a parent thing which has an ID (e.g., an order on an object). Of course, it is often necessary to obtain sequences of IDs, such as when initially populating the client’s state; this is handled using Get ID Sequence type requests and ID Sequence type responses. A common structure for requesting multiple items is a Get ID Sequence request and ID Sequence response, followed by a series of Get With ID requests and appropriate responses describing the item requested.

Players and Games

Before a client can begin interacting with a game, some formalities need to be addressed. The client must first issue a Connect frame to the server, to which the server might respond with OK or Fail—since the Connect frame includes the client’s protocol version, one reason for failure might be a version mismatch. The server can also respond with the Redirect frame, for moves or server pools. Next, the client must issue a Login frame, which identifies and possibly authenticates the player; players new to a server can first use the Create Account frame if the server allows it.

Because of the vast variability of Thousand Parsec, the client needs some way to ascertain which protocol features are supported by the server; this is accomplished via the Get Features request and Features response. Some of the features the server might respond with include:

- Availability of SSL and HTTP tunnelling (on this port or another port).
- Support for server-side component property calculation.
- Ordering of ID sequences in responses (ascending vs. descending).

Similarly, the Get Games request and sequence of Game responses informs the client about the nature of the active games on the server. A single Game frame contains the following information about a game:

- The long (descriptive) name of the game.
- A list of supported protocol versions.
- The type and version of the server.
- The name and version of the ruleset.
- A list of possible network connection configurations.
- A few optional items (number of players, number of objects, administrator details, comment, current turn number, etc.).
- The base URL for media used by the game.

It is, of course, important for a player to know who he or she is up against (or working with, as the case may be), and there is a set of frames for that. The exchange follows the common item sequence pattern with a Get Player IDs request, a List of Player IDs response, and a series of Get Player Data requests and Player Data responses. The Player Data frame contains the player’s name and race.

Turns in the game are also controlled via the protocol. When a player has finished performing actions, he or she may signal readiness for the next turn via the Finished Turn request; the next turn is computed when all players have done so. Turns also have a time limit imposed by the server, so that slow or unresponsive players cannot hold up a game; the client normally issues a Get Time Remaining request, and tracks the turn with a local timer set to the value in the server’s Time Remaining response.

Finally, Thousand Parsec supports messages for a variety of purposes: game broadcasts to all players, game notifications to a single player, player-to-player communications. These are organized into “board” containers which manage ordering and visibility; following the item sequence pattern, the exchange consists of a Get Board IDs request, a List of Board IDs response, and a series of Get Board requests and Board responses.

Once the client has information on a message board, it can issue Get Message requests to obtain messages on the board by slot (hence, Get Message uses the Get With ID and Slot base frame); the server responds with Message frames containing the message subject and body, the turn on which the message was generated, and references to any other entities mentioned in the message. In addition to the normal set of items encountered in Thousand Parsec (players, objects, and the like), there are also some special references including message priority, player actions, and order status. Naturally, the client can also add messages using the Post Message frame—a vehicle for a Messsage frame—and delete them using the Remove Message frame (based on the GetMessage frame).

Objects, Orders, and Resources

The bulk of the process of interacting with the universe is accomplished through a series of frames comprising the functionality for objects, orders, and resources.

The physical state of the universe—or at least that part of it that the player controls or has the ability to see—must be obtained upon connecting, and every turn thereafter, by the client. The client generally issues a Get Object IDs request (a Get ID Sequence), to which the server replies with a List of Object IDs response. The client can then request details about individual objects using Get Object by ID requests, which are answered with Object frames containing such details—again subject to visibility by the player—as their type, name, size, position, velocity, contained objects, applicable order types, and current orders. The protocol also provides the Get Object IDs by Position request, which allows the client to find all objects within a specified sphere of space.

The client obtains the set of possible orders following the usual item sequence pattern by issuing a Get Order Description IDs request and, for each ID in the List of Order Description IDs response, issuing a Get Order Description request and receiving a Order Description response. The implementation of the orders and order queues themselves has evolved markedly over the history of the protocol. Originally, each object had a single order queue. The client would issue an Order request (containing the order type, target object, and other information), receive an Outcome response detailing the expected result of the order, and, after completion of the order, receive a Result frame containing the actual result.

In the second version, the Order frame incorporated the contents of the Outcome frame (since, based on the order description, this did not require the server’s input), and the Result frame was re-

moved entirely. The latest version of the protocol refactored the order queue out of objects, and added the Get Order Queue IDs, List of Order Queue IDs, Get Order Queue, and Order Queue frames, which work similarly to the message and board functionality². The Get Order and Remove Order frames (both GetWithIDSlot requests) allow the client to access and remove orders on a queue, respectively. The Insert Order frame now acts as a vehicle for the Order payload; this was done to allow for another frame, Probe Order, which is used by the client in some cases to obtain information for local use.

Resource descriptions also follow the item sequence pattern: a Get Resource Description IDs request, a List of Resource Description IDs response, and a series of Get Resource Description requests and Resource Description responses.

Design Manipulation

The handling of designs in the Thousand Parsec Protocol is broken down into the manipulation of four separate sub-categories: categories, components, properties, and designs.

Categories differentiate the different design types. Two of the most commonly used design types are ships and weapons. Creating a category is simple, as it consists only of a name and description; the Category frame itself contains only these two strings. Each category is added by the ruleset to the Design Store using an Add Category request, a vehicle for the Category frame. The remainder of the management of categories is handled in the usual item sequence pattern with the Get Category IDs request and List of Category IDs response.

Components consist of the different parts and modules which comprise a design. This can be anything from the hull of a ship or missile to the tube that a missile is housed in. Components are a bit more involved than categories. A Component frame contains the following information:

- The name and description of the component.
- A list of categories to which the component belongs.
- A Requirements function, in Thousand Parsec Component Language (TPCL).
- A list of properties and their corresponding values.

Of particular note is the Requirements function associated with the component. Since components are the parts that make up a ship, weapon, or other constructed object, it is necessary to ensure that they are valid when adding them to a design. The Requirements function verifies that each component added to the design conforms to the rules of other previously added components. For example, in *Missile and Torpedo Wars*, it is impossible to hold an Alpha Missile in a ship without an Alpha Missile Tube. This verification occurs on both the client side and the server side, which is why the entire function must appear in a protocol frame, and why a concise language (TPCL, covered later in the chapter) was chosen for it.

²Actually, it's the other way around: messages and boards were derived from orders in the second version of the protocol.

```

<prop>
<CategoryIDName>Ships</CategoryIDName>
<rank value="0"/>
<name>Colonise</name>
<displayName>Can Colonise Planets</displayName>
<description>Can the ship colonise planets</description>
<tpclDisplayFunction>
  (lambda (design bits) (let ((n (apply + bits))) (cons n (if (= n 1) "Yes" "No")) ) )
</tpclDisplayFunction>
<tpclRequirementsFunction>
  (lambda (design) (cons #t ""))
</tpclRequirementsFunction>
</prop>

```

図 21.1: Example Property

All of a design's properties are communicated via Property frames. Each ruleset exposes a set of properties used within the game. These typically include things like the number of missile tubes of a certain type allowed on a ship, or the amount of armor included with a certain hull type. Like Component frames, Property frames make use of TPCL. A Property frame contains the following information:

- The (display) name and description of the property.
- A list of categories to which the property belongs.
- The name (valid TPCL identifier) of the property.
- The rank of the property.
- Calculate and Requirements functions, in Thousand Parsec Component Language (TPCL).

The rank of a property is used to distinguish a hierarchy of dependencies. In TPCL, a function may not depend on any property which has a rank less than or equal to this property. This means that if one had an Armor property of rank 1 and an Invisibility property of rank 0, then the Invisibility property could not directly depend on the Armor property. This ranking was implemented as a method of curtailing circular dependencies. The Calculate function is used to define how a property is displayed, differentiating the methods of measurement. Missile and Torpedo Wars uses XML to import game properties from a game data file. 図 21.1 shows an example property from that game data.

In this example, we have a property belonging to the Ships category, of rank 0. This property is called Colonise, and relates to the ability of a ship to colonize planets. A quick look at the TPCL Calculate function (listed here as tpclDisplayFunction) reveals that this property outputs either "Yes" or "No" depending on whether the ship in question has said capability. Adding properties in this fashion gives the ruleset designer granular control over metrics of the game and the ability to easily compare them and output them in a player-friendly format.

The actual design of ships, weapons, and other game artifacts are created and manipulated using the Design frame and related frames. In all current rulesets, these are used for building ships and weaponry using the existing pool of components and properties. Since the rules for designs are already handled in TPCL Requirements functions in both properties and components, the creation of a design is a bit simpler. A Design frame contains the following information:

- The name and description of the design.
- A list of categories to which the design belongs.
- A count of the number of instances of the design.
- The owner of the design.
- A list of component IDs and their corresponding counts.
- A list of properties and their corresponding display string.
- The feedback on the design.

This frame is a bit different from the others. Most notably, since a design is an owned item in the game, there is a relation to the owner of each design. A design also tracks the number of its instantiations with a counter.

Server Administration

A server administration protocol extension is also available, allowing for remote live control of supporting servers. The standard use case is to connect to the server via an administration client—perhaps a shell-like command interface or a GUI configuration panel—to change settings or perform other maintenance tasks. However, other, more specialized uses are possible, such as behind-the-scenes management for single-player games.

As with the game protocol described in the preceding sections, the administration client first negotiates a connection (on a port separate from the normal game port) and authenticates using Connect and Login requests. Once connected, the client can receive log messages from and issue commands to the server.

Log messages are pushed to the client via Log Message frames. These contain a severity level and text; as appropriate to the context, the client can choose to display all, some, or none of the log messages it receives.

The server may also issue a Command Update frame instructing the client to populate or update its local command set; supported commands are exposed to the client in the server's response to a Get Command Description IDs frame. Individual command descriptions must then be obtained by issuing a Get Command Description frame for each, to which the server responds with a Command Description frame.

This exchange is functionally quite similar to (and, in fact, was originally based on) that of the order frames used in the main game protocol. It allows commands to be described to the user

and vetted locally to some degree, minimizing network usage. The administration protocol was conceived at a time when the game protocol was already mature; rather than starting from scratch, the developers found existing functionality in the game protocol which did almost what was needed, and added the code to the same protocol libraries.

21.3 Supporting Functionality

Server Persistence

Thousand Parsec games, like many in the turn-based strategy genre, have the potential to last for quite some time. Besides often running far longer than the circadian rhythms of the players' species, during this extended period the server process might be prematurely terminated for any number of reasons. To allow players to pick up a game where they left off, Thousand Parsec servers provide persistence by storing the entire state of the universe (or even multiple universes) in a database. This functionality is also used in a related way for saving single-player games, which will be covered in more detail later in this section.

The flagship server, `tpserver-cpp`, provides an abstract persistence interface and a modular plugin system to allow for various database back ends. At the time of writing, `tpserver-cpp` ships with modules for MySQL and SQLite.

The abstract `Persistence` class describes the functionality allowing the server to save, update, and retrieve the various elements of a game (as described in the Anatomy of a Star Empire section). The database is updated continuously from various places in the server code where the game state changes, and no matter the point at which the server is terminated or crashes, all information to that point should be recovered when the server starts again from the saved data.

Thousand Parsec Component Language

The Thousand Parsec Component Language (TPCL) exists to allow clients to create designs locally without server interaction—allowing for instant feedback about the properties, makeup, and validity of the designs. This allows the player to interactively create, for example, new classes of starship, by customizing structure, propulsion, instrumentation, defenses, armaments, and more according to available technology.

TPCL is a subset of Scheme, with a few minor changes, though close enough to the Scheme R5RS standard that any compatible interpreter can be used. Scheme was originally chosen because of its simplicity, a host of precedents for using it as an embedded language, the availability of interpreters implemented in many other languages, and, most importantly to an open source project, vast documentation both on using it and on developing interpreters for it.

Consider the following example of a Requirements function in TPCL, used by components and properties, which would be included with a ruleset on the server side and communicated to the client over the game protocol:

```
(lambda (design)
  (if (> (designType.MaxValue design) (designType.Size design))
      (if (= (designType.num-hulls design) 1)
          (cons #t "")
          (cons #f "Ship can only have one hull"))
      )
    (cons #f "This many components can't fit into this Hull")
  )
)
```

Readers familiar with Scheme will no doubt find this code easy to understand. The game (both client and server) uses it to check other component properties (MaxSize, Size, and Num-Hulls) to verify that this component can be added to a design. It first verifies that the Size of the component is within the maximum size of the design, then ensures that there are no other hulls in the design (the latter test tips us off that this is the Requirements function from a ship hull).

BattleXML

In war, every battle counts, from the short skirmish in deep space between squadrons of small lightly-armed scout craft, to the massive final clash of two flagship fleets in the sky above a capital world. On the Thousand Parsec framework, the details of combat are handled within the ruleset, and there is no explicit client-side functionality regarding combat details—typically, the player will be informed of the initiation and results of combat via messages, and the appropriate changes to the objects will take place (e.g., removal of destroyed ships). Though the player’s focus will normally be on a higher level, under rulesets with complex combat mechanics, it may prove advantageous (or, at least, entertaining) to examine the battle in more detail.

This is where BattleXML comes in. Battle data is split into two major parts: the media definition, which provides details about the graphics to be used, and the battle definition, which specifies what actually occurred during a battle. These are intended to be read by a battle viewer, of which Thousand Parsec currently has two: one in 2D and the other in 3D. Of course, since the nature of battles are entirely a feature of a ruleset, the ruleset code is responsible for actually producing BattleXML data.

The media definition is tied to the nature of the viewer, and is stored in a directory or an archive containing the XML data and any graphics or model files it references. The data itself describes what media should be used for each ship (or other object) type, its animations for actions such as firing and death, and the media and details of its weapons. File locations are assumed to be relative to the XML file itself, and cannot reference parent directories.

The battle definition is independent of the viewer and media. First, it describes a series of entities on each side at the start of the battle, with unique identifiers and information such as name, description, and type. Then, each round of the battle is described: object movement, weapons fire (with source and target), damage to objects, death of objects, and a log message. How much detail is used to describe each round of battle is dictated by the ruleset.

Metaserver

Finding a public Thousand Parsec server to play on is much like locating a lone stealth scout in deep space—a daunting prospect if one doesn’t know where to look. Fortunately, public servers can announce themselves to a metaserver, whose location, as a central hub, should ideally be well-known to players.

The current implementation is `metaserver-lite`, a PHP script, which lives at some central place like the Thousand Parsec website. Supporting servers send an HTTP request specifying the update action and containing the type, location (protocol, host, and port), ruleset, number of players, object count, administrator, and other optional information. Server listings expire after a specified timeout (by default, 10 minutes), so servers are expected to update the metaserver periodically.

The script can then, when called with no specified action, be used to embed the list of servers with details into a web site, presenting clickable URLs (typically with the `tp://` scheme name). Alternatively, the badge action presents server listings in a compact “badge” format.

Clients may issue a request to a metaserver using the get action to obtain a list of available servers. In this case, the metaserver returns one or more Game frames for each server in the list to the client. In `tpclient-pywx`, the resulting list is presented through a server browser in the initial connection window.

Single-Player Mode

Thousand Parsec is designed from the ground up to support networked multiplayer games. However, there is nothing preventing a player from firing up a local server, connecting a few AI clients, and hyperjumping into a custom single-player universe ready to be conquered. The project defines some standard metadata and functionality to support streamlining this process, making setup as easy as running a GUI wizard or double-clicking a scenario file.

At the core of this functionality is an XML DTD specifying the format for metadata regarding the capabilities and properties of each component (e.g., server, AI client, ruleset). Component packages ship with one or more such XML files, and eventually all of this metadata is aggregated into an associative array divided into two major portions: servers and AI clients. Within a server’s metadata will typically be found metadata for one or more rulesets—they are found here because even though a ruleset may be implemented for more than one server, some configuration details

may differ, so separate metadata is needed in general for each implementation. Each entry for one of these components contains the following information:

- Descriptive data, including a short (binary) name, a long (descriptive) name, and a description.
- The installed version of the component, and the earliest version whose save data is compatible with the installed version.
- The command string (if applicable) and any forced parameters passed to it.
- A set of parameters which can be specified by the player.

Forced parameters are not player-configurable and are typically options which allow the components to function appropriately for a local, single-player context. The player parameters have their own format indicating such details as the name and description, the data type, default, and range of the value, and the format string to append to the main command string.

While specialized cases are possible (e.g., preset game configurations for ruleset-specific clients), the typical process for constructing a single-player game involves selecting a set of compatible components. Selection of the client is implicit, as the player will have already launched one in order to play a game; a well-designed client follows a user-centric workflow to set up the remainder. The next natural choice to make is the ruleset, so the player is presented with a list—at this point, there is no need to bother with server details. In the event that the chosen ruleset is implemented by multiple installed servers (probably a rare condition), the player is prompted to select one; otherwise, the appropriate server is selected automatically. Next, the player is prompted to configure options for the ruleset and server, with sane defaults pulled from the metadata. Finally, if any compatible AI clients are installed, the player is prompted to configure one or more of them to play against.

With the game so configured, the client launches the local server with appropriate configuration parameters (including the ruleset, its parameters, and any parameters it adds to the server’s configuration), using the command string information from the metadata. Once it has verified that the server is running and accepting connections, perhaps using the administration protocol extension discussed previously, it launches each of the specified AI clients similarly, and verifies that they have successfully connected to the game. If all goes well, the client will then connect to the server—just as if it were connecting to an online game—and the player can begin exploring, trading, conquering, and any of a universe of other possibilities.

An alternate—and very important—use for the single-player functionality is the saving and loading of games, and, more or less equivalently, the loading of ready-to-play scenarios. In this case, the save data (probably, though not necessarily, a single file) stores the single-player game configuration data alongside the persistence data for the game itself. Provided all appropriate components in compatible versions are installed on the player’s system, launching a saved game or scenario is completely automatic. Scenarios in particular thus provide an attractive one-click entry into a game. Although Thousand Parsec does not currently have a dedicated scenario editor or a client with an

edit mode, the concept is to provide some means of crafting the persistence data outside of the normal functioning of the ruleset, and verifying its consistency and compatibility.

So far, the description of this functionality has been rather abstract. On a more concrete level, the Python client helper library, `libtpclient-py`, is currently home to the only full realization of single-player mechanics in the Thousand Parsec project. The library provides the `SinglePlayerGame` class, which upon instantiation automatically aggregates all available single-player metadata on the system (naturally, there are certain guidelines as to where the XML files should be installed on a given platform). The object can then be queried by the client for various information on the available components; servers, rulesets, AI clients, and parameters are stored as dictionaries (Python's associative arrays). Following the general game building process outlined above, a typical client might perform the following:

1. Query a list of available rulesets via `SinglePlayerGame.rulesets`, and configure the object with the chosen ruleset by setting `SinglePlayerGame.rname`.
2. Query a list of servers implementing the ruleset via `SinglePlayerGame.list_servers_with_ruleset`, prompt the user to select one if necessary, and configure the object with the chosen (or only) server by setting `SinglePlayerGame.sname`.
3. Obtain the set of parameters for the server and ruleset via `SinglePlayerGame.list_rparams` and `SinglePlayerGame.list_sparams`, respectively, and prompt the player to configure them.
4. Find available AI clients supporting the ruleset via `SinglePlayerGame.list_aiclients_with_ruleset`, and prompt the player to configure one or more of them using the parameters obtained via `SinglePlayerGame.list_aiparams`.
5. Launch the game by calling `SinglePlayerGame.start`, which will return a TCP/IP port to connect on if successful.
6. Eventually, end the game (and kill any launched server and AI client processes) by calling `SinglePlayerGame.stop`.

Thousand Parsec's flagship client, `tpclient-pywx`, presents a user-friendly wizard which follows such a procedure, initially prompting instead for a saved game or scenario file to load. The user-centric workflow developed for this wizard is an example of good design arising from the open source development process of the project: the developer initially proposed a very different process more closely aligned with how things were working under the hood, but community discussion and some collaborative development produced a result much more usable for the player.

Finally, saved games and scenarios are currently implemented in practice in `tpserver-cpp`, with supporting functionality in `libtpclient-py` and an interface in `tpclient-pywx`. This is achieved through a persistence module using SQLite, a public domain open source RDBMS which requires no external process and stores databases in a single file. The server is configured, via a forced parameter, to use the SQLite persistence module if it is available, and as usual, the database file

(living in a temporary location) is constantly updated throughout the game. When the player opts to save the game, the database file is copied to the specified location, and a special table is added to it containing the single player configuration data. It should be fairly obvious to the reader how this is subsequently loaded.

21.4 Lessons Learned

The creation and growth of the extensive Thousand Parsec framework has allowed the developers plenty of opportunity to look back and assess the design decisions that were made along the way. The original core developers (Tim Ansell and Lee Begg) built the original framework from scratch and have shared with us some suggestions on starting a similar project.

What Worked

A major key to the development of Thousand Parsec was the decision to define and build a subset of the framework, followed by the implementation. This iterative and incremental design process allowed the framework to grow organically, with new features added seamlessly. This led directly to the decision to version the Thousand Parsec protocol, which is credited with a number of major successes of the framework. Versioning the protocol allowed the framework to grow over time, enabling new methods of gameplay along the way.

When developing such an expansive framework, it is important to have a very short-term approach for goals and iterations. Short iterations, on the order of weeks for a minor release, allowed the project to move forward quickly with immediate returns along the way. Another success of the implementation was the client-server model, which allowed for the clients to be developed away from any game logic. The separation of game logic from client software was important to the overall success of Thousand Parsec.

What Didn't Work

A major downfall of the Thousand Parsec framework was the decision to use a binary protocol. As you can imagine, debugging a binary protocol is not a fun task and this has lead to many prolonged debugging sessions. We would highly recommend that nobody take this path in the future. The protocol has also grown to have too much flexibility; when creating a protocol, it is important to implement only the basic features that are required.

Our iterations have at times grown too large. When managing such a large framework on an open source development schedule, it is important to have a small subset of added features in each iteration to keep development flowing.

Conclusion

Like a construction skiff inspecting the skeletal hull of a massive prototype battleship in an orbital construction yard, we have passed over the various details of the architecture of Thousand Parsec. While the general design criteria of flexibility and extensibility have been in the minds of the developers from the very beginning, it is evident to us, looking at the history of the framework, that only an open source ecosystem, teeming with fresh ideas and points of view, could have produced the sheer volume of possibilities while remaining functional and cohesive. It is a singularly ambitious project, and as with many of its peers on the open source landscape, much remains to be done; it is our hope and expectation that over time, Thousand Parsec will continue to evolve and expand its capabilities while new and ever more complex games are developed upon it. After all, a journey of a thousand parsecs begins with a single step.

Violet

Cay Horstmann

In 2002, I wrote an undergraduate textbook on object-oriented design and patterns [Hor05]. As with so many books, this one was motivated by frustration with the canonical curriculum. Frequently, computer science students learn how to design a single class in their first programming course, and then have no further training in object-oriented design until their senior level software engineering course. In that course, students rush through a couple of weeks of UML and design patterns, which gives no more than an illusion of knowledge. My book supports a semester-long course for students with a background in Java programming and basic data structures (typically from a Java-based CS1/CS2 sequence). The book covers object-oriented design principles and design patterns in the context of familiar situations. For example, the Decorator design pattern is introduced with a Swing JScrollPane, in the hope that this example is more memorable than the canonical Java streams example.

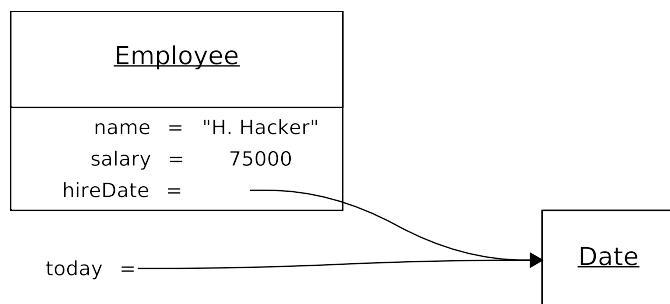


図 22.1: A Violet Object Diagram

I needed a light subset of UML for the book: class diagrams, sequence diagrams, and a variant of object diagrams that shows Java object references (図 22.1). I also wanted students to draw their own diagrams. However, commercial offerings such as Rational Rose were not only expensive but also cumbersome to learn and use [Shu05], and the open source alternatives available at the time were

too limited or buggy to be useful¹. In particular, sequence diagrams in ArgoUML were seriously broken.

I decided to try my hand at implementing the simplest editor that is (a) useful to students and (b) an example of an extensible framework that students can understand and modify. Thus, Violet was born.

22.1 Introducing Violet

Violet is a lightweight UML editor, intended for students, teachers, and authors who need to produce simple UML diagrams quickly. It is very easy to learn and use. It draws class, sequence, state, object and use-case diagrams. (Other diagram types have since been contributed.) It is open-source and cross-platform software. In its core, Violet uses a simple but flexible graph framework that takes full advantage of the Java 2D graphics API.

The Violet user interface is purposefully simple. You don't have to go through a tedious sequence of dialogs to enter attributes and methods. Instead, you just type them into a text field. With a few mouse clicks, you can quickly create attractive and useful diagrams.

Violet does not try to be an industrial-strength UML program. Here are some features that Violet does *not* have:

- Violet does not generate source code from UML diagrams or UML diagrams from source code.
- Violet does not carry out any semantic checking of models; you can use Violet to draw contradictory diagrams.
- Violet does not generate files that can be imported into other UML tools, nor can it read model files from other tools.
- Violet does not attempt to lay out diagrams automatically, except for a simple “snap to grid” facility.

(Attempting to address some of these limitations makes good student projects.)

When Violet developed a cult following of designers who wanted something more than a cocktail napkin but less than an industrial-strength UML tool, I published the code on SourceForge under the GNU General Public License. Starting in 2005, Alexandre de Pellegrin joined the project by providing an Eclipse plugin and a prettier user interface. He has since made numerous architectural changes and is now the primary maintainer of the project.

In this article, I discuss some of the original architectural choices in Violet as well as its evolution. A part of the article is focused on graph editing, but other parts—such as the use of JavaBeans properties and persistence, Java WebStart and plugin architecture—should be of general interest.

¹At the time, I was not aware of Diomidis Spinellis' admirable UMLGraph program [Spi03], in which diagrams are specified by textual declarations rather than the more common point-and-click interface.

22.2 The Graph Framework

Violet is based on a general graph editing framework that can render and edit nodes and edges of arbitrary shapes. The Violet UML editor has nodes for classes, objects, activation bars (in sequence diagrams), and so on, and edges for the various edge shapes in UML diagrams. Another instance of the graph framework might display entity-relationship diagrams or railroad diagrams.

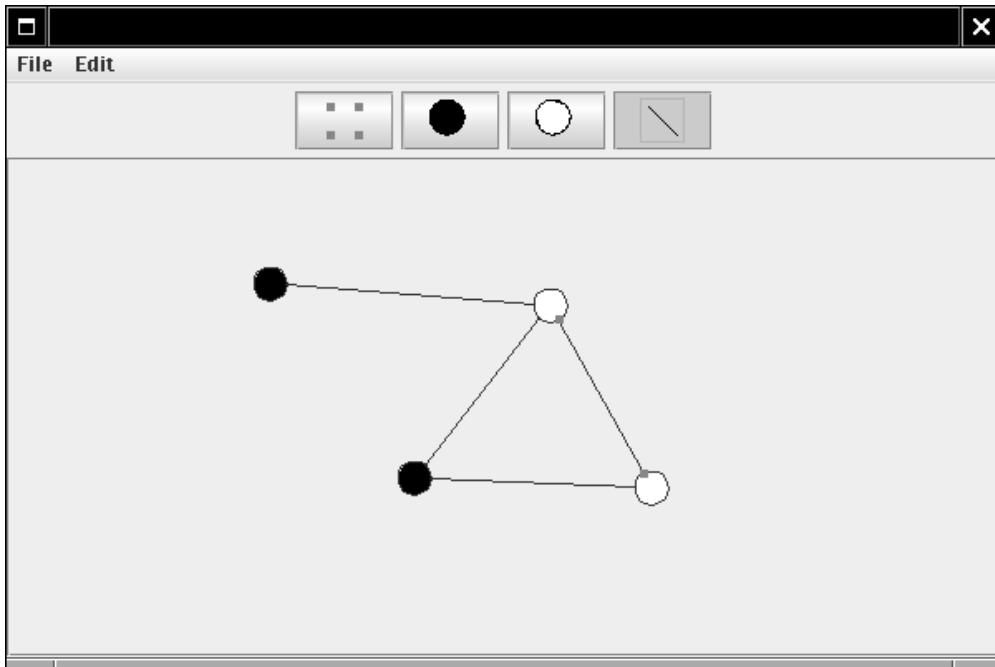


图 22.2: A Simple Instance of the Editor Framework

In order to illustrate the framework, let us consider an editor for very simple graphs, with black and white circular nodes and straight edges (图 22.2). The `SimpleGraph` class specifies prototype objects for the node and edge types, illustrating the prototype pattern:

```

public class SimpleGraph extends AbstractGraph
{
    public Node[] getNodePrototypes()
    {
        return new Node[]
        {
            new CircleNode(Color.BLACK),
            new CircleNode(Color.WHITE)
        };
    }
    public Edge[] getEdgePrototypes()
    {
        return new Edge[]
        {
            new LineEdge()
        };
    }
}

```

Prototype objects are used to draw the node and edge buttons at the top of 図 22.2. They are cloned whenever the user adds a new node or edge instance to the graph. Node and Edge are interfaces with the following key methods:

- Both interfaces have a `getShape` method that returns a Java2D Shape object of the node or edge shape.
- The Edge interface has methods that yield the nodes at the start and end of the edge.
- The `getConnectionPoint` method in the Node interface type computes an optimal attachment point on the boundary of a node (see 図 22.3).
- The `getConnectionPoints` method of the Edge interface yields the two end points of the edge. This method is needed to draw the “grabbers” that mark the currently selected edge.
- A node can have children that move together with the parent. A number of methods are provided for enumerating and managing children.

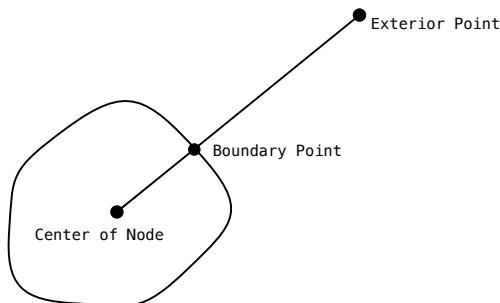


図 22.3: Finding a Connection Point on the Boundary of the Node Shape

Convenience classes `AbstractNode` and `AbstractEdge` implement a number of these methods, and classes `RectangularNode` and `SegmentedLineEdge` provide complete implementations of rectangular nodes with a title string and edges that are made up of line segments.

In the case of our simple graph editor, we would need to supply subclasses `CircleNode` and `LineEdge` that provide a `draw` method, a `contains` method, and the `getConnectionPoint` method that describes the shape of the node boundary. The code is given below, and 図 22.4 shows a class diagram of these classes (drawn, of course, with Violet).

```
public class CircleNode extends AbstractNode
{
    public CircleNode(Color aColor)
    {
        size = DEFAULT_SIZE;
        x = 0;
        y = 0;
        color = aColor;
    }

    public void draw(Graphics2D g2)
    {
        Ellipse2D circle = new Ellipse2D.Double(x, y, size, size);
        Color oldColor = g2.getColor();
        g2.setColor(color);
        g2.fill(circle);
        g2.setColor(oldColor);
        g2.draw(circle);
    }

    public boolean contains(Point2D p)
    {
        Ellipse2D circle = new Ellipse2D.Double(x, y, size, size);
        return circle.contains(p);
    }

    public Point2D getConnectionPoint(Point2D other)
    {
        double centerX = x + size / 2;
        double centerY = y + size / 2;
        double dx = other.getX() - centerX;
        double dy = other.getY() - centerY;
        double distance = Math.sqrt(dx * dx + dy * dy);
        if (distance == 0) return other;
        else return new Point2D.Double(
            centerX + dx * (size / 2) / distance,
            centerY + dy * (size / 2) / distance);
    }

    private double x, y, size, color;
    private static final int DEFAULT_SIZE = 20;
}
```

```

public class LineEdge extends AbstractEdge
{
    public void draw(Graphics2D g2)
    { g2.draw getConnectionPoints()); }

    public boolean contains(Point2D aPoint)
    {
        final double MAX_DIST = 2;
        return getConnectionPoints().ptSegDist(aPoint) < MAX_DIST;
    }
}

```

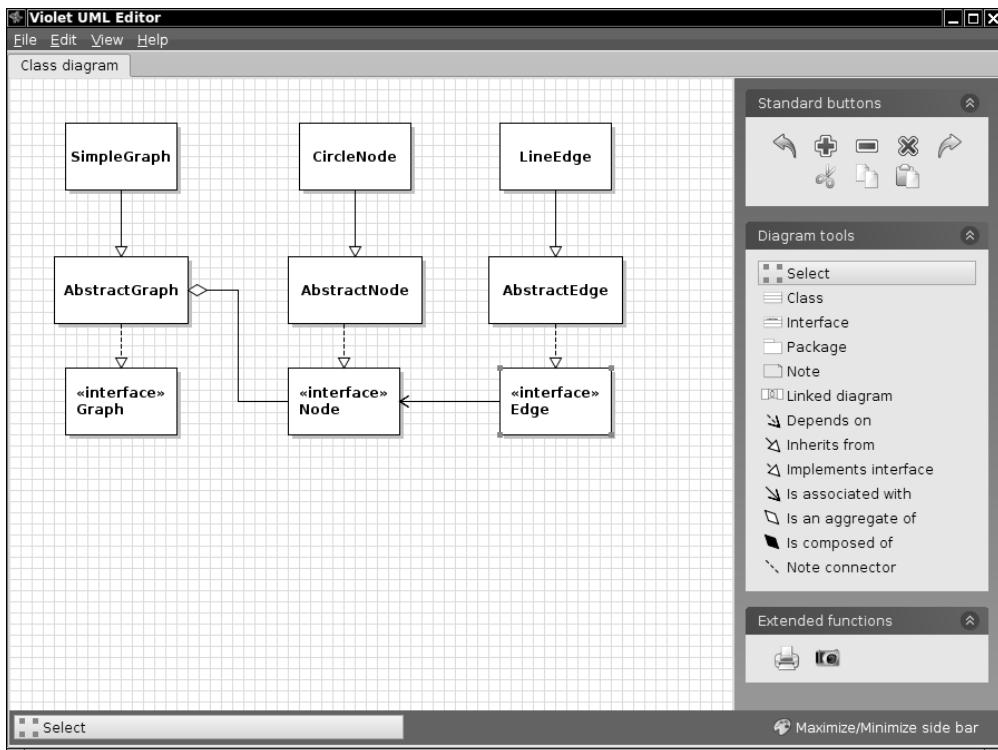


図 22.4: Class Diagram for a Simple Graph

In summary, Violet provides a simple framework for producing graph editors. To obtain an editor instance, define node and edge classes and provide methods in a graph class that yield prototype node and edge objects.

Of course, there are other graph frameworks available, such as JGraph [Ald02] and JUNG². However, those frameworks are considerably more complex, and they provide frameworks for drawing

²<http://jung.sourceforge.net>

graphs, not for applications that draw graphs.

22.3 Use of JavaBeans Properties

In the golden days of client-side Java, the JavaBeans specification was developed in order to provide portable mechanisms for editing GUI components in visual GUI builder environments. The vision was that a third-party GUI component could be dropped into any GUI builder, where its properties could be configured in the same way as the standard buttons, text components, and so on.

Java does not have native properties. Instead, JavaBeans properties can be discovered as pairs of getter and setter methods, or specified with companion BeanInfo classes. Moreover, *property editors* can be specified for visually editing property values. The JDK even contains a few basic property editors, for example for the type `java.awt.Color`.

The Violet framework makes full use of the JavaBeans specification. For example, the `CircleNode` class can expose a color property simply by providing two methods:

```
public void setColor(Color newValue)
public Color getColor()
```

No further work is necessary. The graph editor can now edit node colors of circle nodes (図 22.5).

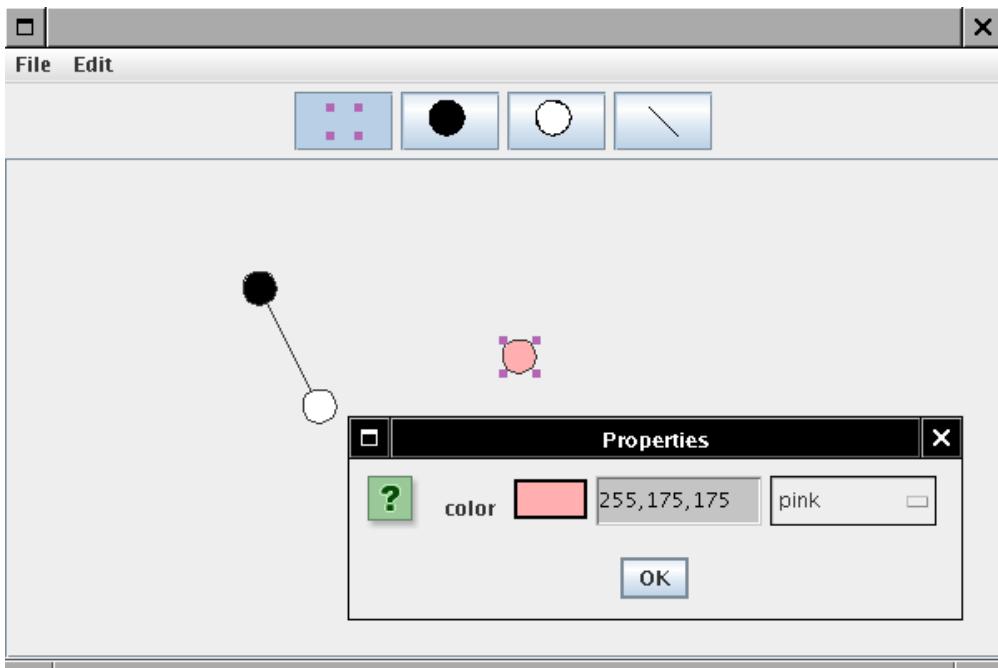


図 22.5: Editing Circle Colors with the default JavaBeans Color Editor

22.4 Long-Term Persistence

Just like any editor program, Violet must save the user's creations in a file and reload them later. I had a look at the XMI specification³ which was designed as a common interchange format for UML models. I found it cumbersome, confusing, and hard to consume. I don't think I was the only one—XMI had a reputation for poor interoperability even with the simplest models [PGL⁺05].

I considered simply using Java serialization, but it is difficult to read old versions of a serialized object whose implementation has changed over time. This problem was also anticipated by the JavaBeans architects, who developed a standard XML format for long-term persistence⁴. A Java object—in the case of Violet, the UML diagram—is serialized as a sequence of statements for constructing and modifying it. Here is an example:

```
<?xml version="1.0" encoding="UTF-8"?>
<java version="1.0" class="java.beans.XMLDecoder">
  <object class="com.horstmann.violet.ClassDiagramGraph">
    <void method="addNode">
      <object id="ClassNode0" class="com.horstmann.violet.ClassNode">
        <void property="name">...</void>
      </object>
      <object class="java.awt.geom.Point2D$Double">
        <double>200.0</double>
        <double>60.0</double>
      </object>
    </void>
    <void method="addNode">
      <object id="ClassNode1" class="com.horstmann.violet.ClassNode">
        <void property="name">...</void>
      </object>
      <object class="java.awt.geom.Point2D$Double">
        <double>200.0</double>
        <double>210.0</double>
      </object>
    </void>
    <void method="connect">
      <object class="com.horstmann.violet.ClassRelationshipEdge">
        <void property="endArrowHead">
          <object class="com.horstmann.violet.ArrowHead" field="TRIANGLE"/>
        </void>
      </object>
      <object idref="ClassNode0"/>
      <object idref="ClassNode1"/>
    </void>
  </object>
</java>
```

³<http://www.omg.org/technology/documents/formal/xmi.htm>

⁴<http://jcp.org/en/jsr/detail?id=57>

When the XMLDecoder class reads this file, it executes these statements (package names are omitted for simplicity).

```
ClassDiagramGraph obj1 = new ClassDiagramGraph();
ClassNode ClassNode0 = new ClassNode();
ClassNode0.setName(...);
obj1.addNode(ClassNode0, new Point2D.Double(200, 60));
ClassNode ClassNode1 = new ClassNode();
ClassNode1.setName(...);
obj1.addNode(ClassNode1, new Point2D.Double(200, 60));
ClassRelationshipEdge obj2 = new ClassRelationshipEdge();
obj2.setEndArrowHead(ArrowHead.TRIANGLE);
obj1.connect(obj2, ClassNode0, ClassNode1);
```

As long as the semantics of the constructors, properties, and methods has not changed, a newer version of the program can read a file that has been produced by an older version.

Producing such files is quite straightforward. The encoder automatically enumerates the properties of each object and writes setter statements for those property values that differ from the default. Most basic datatypes are handled by the Java platform; however, I had to supply special handlers for Point2D, Line2D, and Rectangle2D. Most importantly, the encoder must know that a graph can be serialized as a sequence of addNode and connect method calls:

```
encoder.setPersistenceDelegate(Graph.class, new DefaultPersistenceDelegate()
{
    protected void initialize(Class<?> type, Object oldInstance,
        Object newInstance, Encoder out)
    {
        super.initialize(type, oldInstance, newInstance, out);
        AbstractGraph g = (AbstractGraph) oldInstance;
        for (Node n : g.getNodes())
            out.writeStatement(new Statement(oldInstance, "addNode", new Object[]
            {
                n,
                n.getLocation()
            }));
        for (Edge e : g.getEdges())
            out.writeStatement(new Statement(oldInstance, "connect", new Object[]
            {
                e, e.getStart(), e.getEnd()
            }));
    }
});
```

Once the encoder has been configured, saving a graph is as simple as:

```
encoder.writeObject(graph);
```

Since the decoder simply executes statements, it requires no configuration. Graphs are simply read with:

```
Graph graph = (Graph) decoder.readObject();
```

This approach has worked exceedingly well over numerous versions of Violet, with one exception. A recent refactoring changed some package names and thereby broke backwards compatibility. One option would have been to keep the classes in the original packages, even though they no longer matched the new package structure. Instead, the maintainer provided an XML transformer for rewriting the package names when reading a legacy file.

22.5 Java WebStart

Java WebStart is a technology for launching an application from a web browser. The deployer posts a JNLP file that triggers a helper application in the browser which downloads and runs the Java program. The application can be digitally signed, in which case the user must accept the certificate, or it can be unsigned, in which case the program runs in a sandbox that is slightly more permissive than the applet sandbox.

I do not think that end users can or should be trusted to judge the validity of a digital certificate and its security implications. One of the strengths of the Java platform is its security, and I feel it is important to play to that strength.

The Java WebStart sandbox is sufficiently powerful to enable users to carry out useful work, including loading and saving files and printing. These operations are handled securely and conveniently from the user perspective. The user is alerted that the application wants to access the local filesystem and then chooses the file to be read or written. The application merely receives a stream object, without having an opportunity to peek at the filesystem during the file selection process.

It is annoying that the developer must write custom code to interact with a `FileOpenService` and a `FileSaveService` when the application is running under WebStart, and it is even more annoying that there is no WebStart API call to find out whether the application was launched by WebStart.

Similarly, saving user preferences must be implemented in two ways: using the Java preferences API when the application runs normally, or using the WebStart preferences service when the application is under WebStart. Printing, on the other hand, is entirely transparent to the application programmer.

Violet provides simple abstraction layers over these services to simplify the lot of the application programmer. For example, here is how to open a file:

```
FileService service = FileService.getInstance(initialDirectory);
// detects whether we run under WebStart
FileService.Open open = fileService.open(defaultDirectory, defaultName,
extensionFilter);
InputStream in = open.getInputStream();
String title = open.getName();
```

The `FileService.Open` interface is implemented by two classes: a wrapper over `JFileChooser` or the JNLP `FileOpenService`.

No such convenience is a part of the JNLP API itself, but that API has received little love over its lifetime and has been widely ignored. Most projects simply use a self-signed certificate for their WebStart application, which gives users no security. This is a shame—open source developers should embrace the JNLP sandbox as a risk-free way to try out a project.

22.6 Java 2D

Violet makes intensive use of the Java2D library, one of the lesser known gems in the Java API. Every node and edge has a method `getShape` that yields a `java.awt.Shape`, the common interface of all Java2D shapes. This interface is implemented by rectangles, circles, paths, and their unions, intersections, and differences. The `GeneralPath` class is useful for making shapes that are composed of arbitrary line and quadratic/cubic curve segments, such as straight and curved arrows.

To appreciate the flexibility of the Java2D API, consider the following code for drawing a shadow in the `AbstractNode.draw` method:

```
Shape shape = getShape();
if (shape == null) return;
g2.translate(SHADOW_GAP, SHADOW_GAP);
g2.setColor(SHADOW_COLOR);
g2.fill(shape);
g2.translate(-SHADOW_GAP, -SHADOW_GAP);
g2.setColor(BACKGROUND_COLOR);
g2.fill(shape);
```

A few lines of code produce a shadow for any shape, even shapes that a developer may add at a later point.

Of course, Violet saves bitmap images in any format that the `javax.imageio` package supports; that is, GIF, PNG, JPEG, and so on. When my publisher asked me for vector images, I noted another advantage of the Java 2D library. When you print to a PostScript printer, the Java2D operations are translated into PostScript vector drawing operations. If you print to a file, the result can be consumed by a program such as `ps2eps` and then imported into Adobe Illustrator or Inkscape. Here is the code, where `comp` is the Swing component whose `paintComponent` method paints the graph:

```
DocFlavor flavor = DocFlavor.SERVICE_FORMATTED.PRINTABLE;
String mimeType = "application/postscript";
StreamPrintServiceFactory[] factories;
StreamPrintServiceFactory.lookupStreamPrintServiceFactories(flavor, mimeType);
FileOutputStream out = new FileOutputStream(fileName);
PrintService service = factories[0].getPrintService(out);
SimpleDoc doc = new SimpleDoc(new Printable() {
    public int print(Graphics g, PageFormat pf, int page) {
```

```

        if (page >= 1) return Printable.NO_SUCH_PAGE;
        else {
            double sf1 = pf.getImageableWidth() / (comp.getWidth() + 1);
            double sf2 = pf.getImageableHeight() / (comp.getHeight() + 1);
            double s = Math.min(sf1, sf2);
            Graphics2D g2 = (Graphics2D) g;
            g2.translate((pf.getWidth() - pf.getImageableWidth()) / 2,
                         (pf.getHeight() - pf.getImageableHeight()) / 2);
            g2.scale(s, s);

            comp.paint(g);
            return Printable.PAGE_EXISTS;
        }
    }
}, flavor, null);
DocPrintJob job = service.createPrintJob();
PrintRequestAttributeSet attributes = new HashPrintRequestAttributeSet();
job.print(doc, attributes);

```

At the beginning, I was concerned that there might be a performance penalty when using general shapes, but that has proven not to be the case. Clipping works well enough that only those shape operations that are required for updating the current viewport are actually executed.

22.7 No Swing Application Framework

Most GUI frameworks have some notion of an application that manages a set of documents that deals with menus, toolbars, status bars, etc. However, this was never a part of the Java API. JSR 296⁵ was supposed to supply a basic framework for Swing applications, but it is currently inactive. Thus, a Swing application author has two choices: reinvent a good number of wheels or base itself on a third party framework. At the time that Violet was written, the primary choices for an application framework were the Eclipse and NetBeans platform, both of which seemed too heavyweight at the time. (Nowadays, there are more choices, among them JSR 296 forks such as GUTS⁶.) Thus, Violet had to reinvent mechanisms for handling menus and internal frames.

In Violet, you specify menu items in property files, like this:

```

file.save.text=Save
file.save.mnemonic=S
file.save.accelerator=ctrl S
file.save.icon=/icons/16x16/save.png

```

A utility method creates the menu item from the prefix (here `file.save`). The suffixes `.text`, `.mnemonic`, and so on, are what nowadays would be called “convention over configuration”. Using

⁵<http://jcp.org/en/jsr/detail?id=296>

⁶<http://kenai.com/projects/guts>

resource files for describing these settings is obviously far superior to setting up menus with API calls because it allows for easy localization. I reused that mechanism in another open source project, the GridWorld environment for high school computer science education⁷.

An application such as Violet allows users to open multiple “documents”, each containing a graph. When Violet was first written, the multiple document interface (MDI) was still commonly used. With MDI, the main frame has a menu bar, and each view of a document is displayed in an internal frame with a title but no menu bar. Each internal frame is contained in the main frame and can be resized or minimized by the user. There are operations for cascading and tiling windows.

Many developers disliked MDI, and so this style user interface has gone out of fashion. For a while, a single document interface (SDI), in which an application displays multiple top level frames, was considered superior, presumably because those frames can be manipulated with the standard window management tools of the host operating system. When it became clear that having lots of top level windows isn’t so great after all, tabbed interfaces started to appear, in which multiple documents are again contained in a single frame, but now all displayed at full size and selectable with tabs. This does not allow users to compare two documents side by side, but seems to have won out.

The original version of Violet used an MDI interface. The Java API has a internal frames feature, but I had to add support for tiling and cascading. Alexandre switched to a tabbed interface, which is somewhat better-supported by the Java API. It would be desirable to have an application framework where the document display policy was transparent to the developer and perhaps selectable by the user.

Alexandre also added support for sidebars, a status bar, a welcome panel, and a splash screen. All this should ideally be a part of a Swing application framework.

22.8 Undo/Redo

Implementing multiple undo/redo seems like a daunting task, but the Swing undo package ([Top00], Chapter 9) gives good architectural guidance. An `UndoManager` manages a stack of `UndoableEdit` objects. Each of them has an `undo` method that undoes the effect of the edit operation, and a `redo` method that undoes the undo (that is, carries out the original edit operation). A `CompoundEdit` is a sequence of `UndoableEdit` operations that should be undone or redone in their entirety. You are encouraged to define small, atomic edit operations (such as adding or removing a single edge or node in the case of a graph) that are grouped into compound edits as necessary.

A challenge is to define a small set of atomic operations, each of which can be undone easily. In Violet, they are:

- adding or removing a node or edge

⁷<http://horstmann.com/gridworld>

- attaching or detaching a node's child
- moving a node
- changing a property of a node or edge

Each of these operations has an obvious undo. For example the undo of adding a node is the node's removal. The undo of moving a node is to move it by the opposite vector.

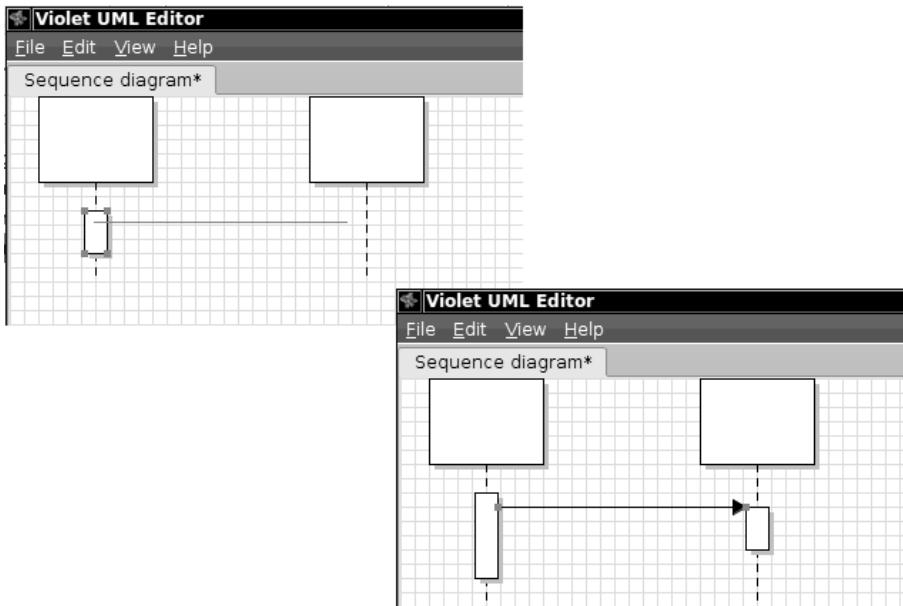


図 22.6: An Undo Operation Must Undo Structural Changes in the Model

Note that these atomic operations are *not* the same as the actions in the user interface or the methods of the Graph interface that the user interface actions invoke. For example, consider the sequence diagram in 図 22.6, and suppose the user drags the mouse from the activation bar to the lifeline on the right. When the mouse button is released, the method:

```
public boolean addEdgeAtPoints(Edge e, Point2D p1, Point2D p2)
```

is invoked. That method adds an edge, but it may also carry out other operations, as specified by the participating Edge and Node subclasses. In this case, an activation bar will be added to the lifeline on the right. Undoing the operation needs to remove that activation bar as well. Thus, the *model* (in our case, the graph) needs to record the structural changes that need to be undone. It is not enough to collect controller operations.

As envisioned by the Swing undo package, the graph, node, and edge classes should send `UndoableEditEvent` notifications to an `UndoManager` whenever a structural edit occurs. Violet has a more general design where the graph itself manages listeners for the following interface:

```

public interface GraphModificationListener
{
    void nodeAdded(Graph g, Node n);
    void nodeRemoved(Graph g, Node n);
    void nodeMoved(Graph g, Node n, double dx, double dy);
    void childAttached(Graph g, int index, Node p, Node c);
    void childDetached(Graph g, int index, Node p, Node c);
    void edgeAdded(Graph g, Edge e);
    void edgeRemoved(Graph g, Edge e);
    void propertyChangedOnNodeOrEdge(Graph g, PropertyChangeEvent event);
}

```

The framework installs a listener into each graph that is a bridge to the undo manager. For supporting undo, adding generic listener support to the model is overdesigned—the graph operations could directly interact with the undo manager. However, I also wanted to support an experimental collaborative editing feature.

If you want to support undo/redo in your application, think carefully about the atomic operations in your model (and not your user interface). In the model, fire events when a structural change happens, and allow the Swing undo manager to collect and group these events.

22.9 Plugin Architecture

For a programmer familiar with 2D graphics, it is not difficult to add a new diagram type to Violet. For example, the activity diagrams were contributed by a third party. When I needed to create railroad diagrams and ER diagrams, I found it faster to write Violet extensions instead of fussing with Visio or Dia. (Each diagram type took a day to implement.)

These implementations do not require knowledge of the full Violet framework. Only the graph, node, and edge interfaces and convenience implementations are needed. In order to make it easier for contributors to decouple themselves from the evolution of the framework, I designed a simple plugin architecture.

Of course, many programs have a plugin architecture, many quite elaborate. When someone suggested that Violet should support OSGi, I shuddered and instead implemented the simplest thing that works.

Contributors simply produce a JAR file with their graph, node, and edge implementations and drop it into a `plugins` directory. When Violet starts, it loads those plugins, using the Java ServiceLoader class. That class was designed to load services such as JDBC drivers. A ServiceLoader loads JAR files that promise to provide a class implementing a given interface (in our case, the `Graph` interface.)

Each JAR file must have a subdirectory `META-INF/services` containing a file whose name is the fully qualified classname of the interface (such as `com.horstmann.violet.Graph`), and that contains the names of all implementing classes, one per line. The ServiceLoader constructs a class loader for the plugin directory, and loads all plugins:

```
ServiceLoader<Graph> graphLoader = ServiceLoader.load(Graph.class, classLoader);
for (Graph g : graphLoader) // ServiceLoader<Graph> implements Iterable<Graph>
    registerGraph(g);
```

This is a simple but useful facility of standard Java that you might find valuable for your own projects.

22.10 Conclusion

Like so many open source projects, Violet was born of an unmet need—to draw simple UML diagrams with a minimum of fuss. Violet was made possible by the amazing breadth of the Java SE platform, and it draws from a diverse set of technologies that are a part of that platform. In this article, I described how Violet makes use of Java Beans, Long-Term Persistence, Java Web Start, Java 2D, Swing Undo/Redo, and the service loader facility. These technologies are not always as well understood as the basics of Java and Swing, but they can greatly simplify the architecture of a desktop application. They allowed me, as the initial sole developer, to produce a successful application in a few months of part-time work. Relying on these standard mechanisms also made it easier for others to improve on Violet and to extract pieces of it into their own projects.

VisTrails

Juliana Freire, David Koop, Emanuele Santos,
Carlos Scheidegger, Claudio Silva, and Huy T. Vo

VisTrails¹ is an open-source system that supports data exploration and visualization. It includes and substantially extends useful features of scientific workflow and visualization systems. Like scientific workflow systems such as Kepler and Taverna, VisTrails allows the specification of computational processes which integrate existing applications, loosely-coupled resources, and libraries according to a set of rules. Like visualization systems such as AVS and ParaView, VisTrails makes advanced scientific and information visualization techniques available to users, allowing them to explore and compare different visual representations of their data. As a result, users can create complex workflows that encompass important steps of scientific discovery, from data gathering and manipulation to complex analyses and visualizations, all integrated in one system.

A distinguishing feature of VisTrails is its provenance infrastructure [FSC⁺06]. VisTrails captures and maintains a detailed history of the steps followed and data derived in the course of an exploratory task. Workflows have traditionally been used to automate repetitive tasks, but in applications that are exploratory in nature, such as data analysis and visualization, very little is repeated—change is the norm. As a user generates and evaluates hypotheses about their data, a series of different, but related, workflows are created as they are adjusted iteratively.

VisTrails was designed to manage these rapidly-evolving workflows: it maintains provenance of data products (e.g., visualizations, plots), of the workflows that derive these products, and their executions. The system also provides annotation capabilities so users can enrich the automatically-captured provenance.

Besides enabling reproducible results, VisTrails leverages provenance information through a series of operations and intuitive user interfaces that help users to collaboratively analyze data. Notably, the system supports reflective reasoning by storing temporary results, allowing users to examine the actions that led to a result and to follow chains of reasoning backward and forward. Users

¹<http://www.vistrails.org>

can navigate workflow versions in an intuitive way, undo changes without losing results, visually compare multiple workflows and show their results side-by-side in a visualization spreadsheet.

VisTrails addresses important usability issues that have hampered a wider adoption of workflow and visualization systems. To cater to a broader set of users, including many who do not have programming expertise, it provides a series of operations and user interfaces that simplify workflow design and use [FSC⁺06], including the ability to create and refine workflows by analogy, to query workflows by example, and to suggest workflow completions as users interactively construct their workflows using a recommendation system [SVK⁺07]. We have also developed a new framework that allows the creation of custom applications that can be more easily deployed to (non-expert) end users.

The extensibility of VisTrails comes from an infrastructure that makes it simple for users to integrate tools and libraries, as well as to quickly prototype new functions. This has been instrumental in enabling the use of the system in a wide range of application areas, including environmental sciences, psychiatry, astronomy, cosmology, high-energy physics, quantum physics, and molecular modeling.

To keep the system open-source and free for all, we have built VisTrails using only free, open-source packages. VisTrails is written in Python and uses Qt as its GUI toolkit (through PyQt Python bindings). Because of the broad range of users and applications, we have designed the system from the ground up with portability in mind. VisTrails runs on Windows, Mac and Linux.

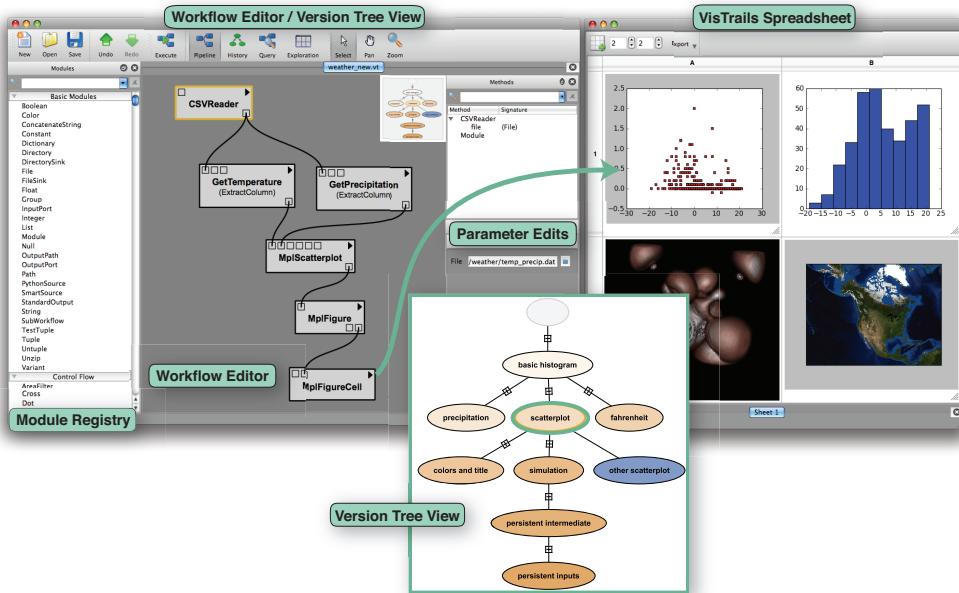


图 23.1: Components of the VisTrails User Interface

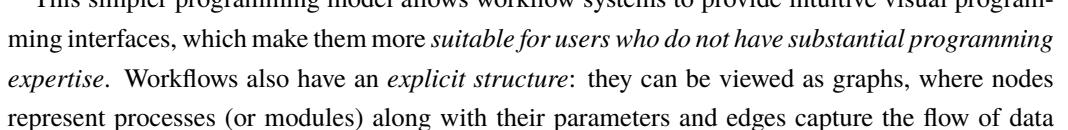
23.1 System Overview

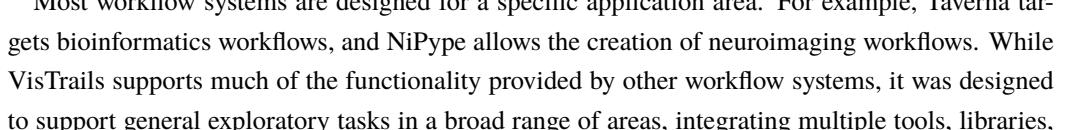
Data exploration is an inherently creative process that requires users to locate relevant data, to integrate and visualize this data, to collaborate with peers while exploring different solutions, and to disseminate results. Given the size of data and complexity of analyses that are common in scientific exploration, tools are needed that better support creativity.

There are two basic requirements for these tools that go hand in hand. First, it is important to be able to specify the exploration processes using formal descriptions, which ideally, are executable. Second, to reproduce the results of these processes as well as reason about the different steps followed to solve a problem, these tools must have the ability to systematically capture provenance. VisTrails was designed with these requirements in mind.

Workflows and Workflow-Based Systems

Workflow systems support the creation of pipelines (workflows) that combine multiple tools. As such, they enable the automation of repetitive tasks and result reproducibility. Workflows are rapidly replacing primitive shell scripts in a wide range of tasks, as evidenced by a number of workflow-based applications, both commercial (e.g., Apple’s Mac OS X Automator and Yahoo! Pipes) and academic (e.g., NiPype, Kepler, and Taverna).

Workflows have a number of advantages compared to scripts and programs written in high-level languages. They provide a simple programming model whereby a sequence of tasks is composed by connecting the outputs of one task to the inputs of another.  23.1 shows a workflow which reads a CSV file that contains weather observations and creates a scatter plot of the values.

This simpler programming model allows workflow systems to provide intuitive visual programming interfaces, which make them more *suitable for users who do not have substantial programming expertise*. Workflows also have an *explicit structure*: they can be viewed as graphs, where nodes represent processes (or modules) along with their parameters and edges capture the flow of data between the processes. In the example of , the module CSVReader takes as a parameter a filename (/weather/temp_precip.dat), reads the file, and feeds its contents into the modules GetTemperature and GetPrecipitation, which in turn send the temperature and precipitation values to a matplotlib function that generates a scatter plot.

Most workflow systems are designed for a specific application area. For example, Taverna targets bioinformatics workflows, and NiPype allows the creation of neuroimaging workflows. While VisTrails supports much of the functionality provided by other workflow systems, it was designed to support general exploratory tasks in a broad range of areas, integrating multiple tools, libraries, and services.

Data and Workflow Provenance

The importance of keeping provenance information for results (and data products) is well recognized in the scientific community. The provenance (also referred to as the audit trail, lineage, and pedigree) of a data product contains information about the process and data used to derive the data product. Provenance provides important documentation that is key to preserving the data, to determining the data's quality and authorship, and to reproducing as well as validating the results [FKSS08].

An important component of provenance is information about *causality*, i.e., a description of a process (sequence of steps) which, together with input data and parameters, caused the creation of a data product. Thus, the structure of provenance mirrors the structure of the workflow (or set of workflows) used to derive a given result set.

In fact, a catalyst for the widespread use of workflow systems in science has been that they can be easily used to automatically capture provenance. While early workflow systems have been *extended* to capture provenance, VisTrails was *designed* to support provenance.

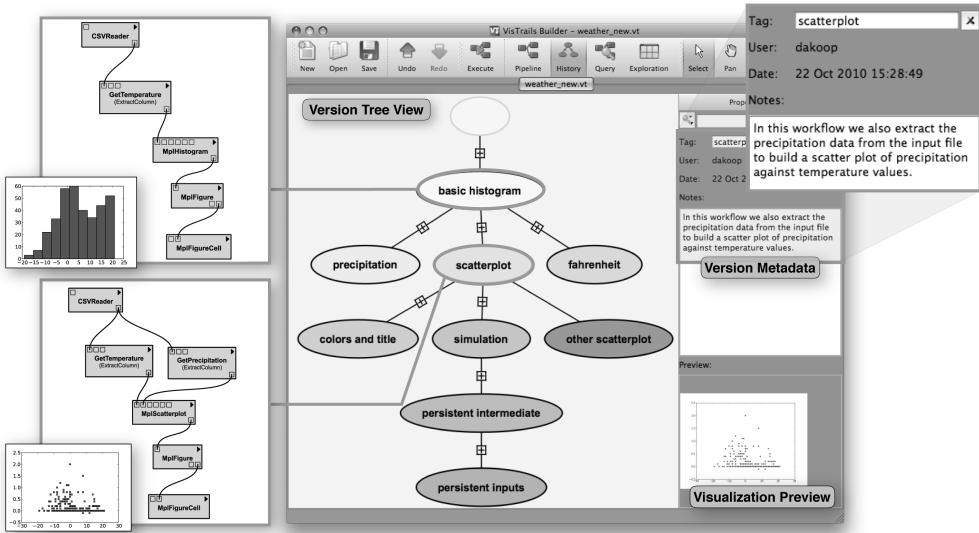


図 23.2: Provenance of Exploration Enhanced by Annotations

User Interface and Basic Functionality

The different user interface components of the system are illustrated in 図 23.1 and 図 23.2. Users create and edit workflows using the Workflow Editor. To build the workflow graphs, users can drag modules from the Module Registry and drop them into the Workflow Editor canvas. VisTrails

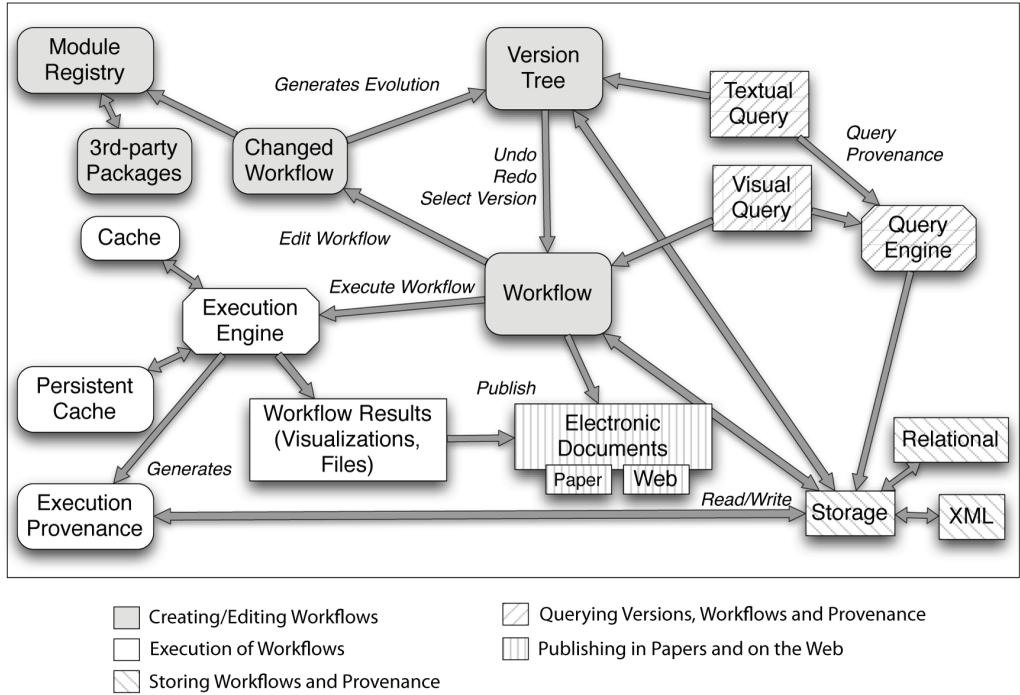


図 23.3: VisTrails Architecture

provides a series of built-in modules, and users can also add their own (see 23.3 節 for details). When a module is selected, VisTrails displays its parameters (in the Parameter Edits area) where the user can set and modify their values.

As a workflow specification is refined, the system captures the changes and presents them to the user in the Version Tree View described below. Users may interact with the workflows and their results in the VisTrails Spreadsheet. Each cell in the spreadsheet represents a view that corresponds to a workflow instance. In 図 23.1, the results of the workflow shown in the Workflow Editor are displayed on the top-left cell of the spreadsheet. Users can directly modify the parameters of a workflow as well as synchronize parameters across different cells in the spreadsheet.

The Version Tree View helps users to navigate through the different workflow versions. As shown in 図 23.2, by clicking on a node in the version tree, users can view a workflow, its associated result (Visualization Preview), and metadata. Some of the metadata is automatically captured, e.g., the id of the user who created a particular workflow and the creation date, but users may also provide additional metadata, including a tag to identify the workflow and a written description.

23.2 Project History

Initial versions of VisTrails were written in Java and C++ [BCC⁺05]. The C++ version was distributed to a few early adopters, whose feedback was instrumental in shaping our requirements for the system.

Having observed a trend in the increase of the number of Python-based libraries and tools in multiple scientific communities, we opted to use Python as the basis for VisTrails. Python is quickly becoming a universal modern glue language for scientific software. Many libraries written in different languages such as Fortran, C, and C++ use Python bindings as a way to provide scripting capabilities. Since VisTrails aims to facilitate the orchestration of many different software libraries in workflows, a pure Python implementation makes this much easier. In particular, Python has dynamic code loading features similar to the ones seen in LISP environments, while having a much bigger developer community, and an extremely rich standard library. Late in 2005, we started the development of the current system using Python/PyQt/Qt. This choice has greatly simplified extensions to the system, in particular, the addition of new modules and packages.

A beta version of the VisTrails system was first released in January 2007. Since then, the system has been downloaded over twenty-five thousand times.

23.3 Inside VisTrails

The internal components that support the user-interface functionality described above are depicted in the high-level architecture of VisTrails, shown in 図 23.3. Workflow execution is controlled by the Execution Engine, which keeps track of invoked operations and their respective parameters and captures the provenance of workflow execution (Execution Provenance). As part of the execution, VisTrails also allows the caching of intermediate results both in memory and on disk. As we discuss in 23.3 節, only new combinations of modules and parameters are re-run, and these are executed by invoking the appropriate functions from the underlying libraries (e.g., matplotlib). Workflow results, connected to their provenance, can then be included in electronic documents (23.4 節).

Information about changes to workflows is captured in a Version Tree, which can be persisted using different storage back ends, including an XML file store in a local directory and a relational database. VisTrails also provides a query engine that allows users to explore the provenance information.

We note that, although VisTrails was designed as an interactive tool, it can also be used in server mode. Once workflows are created, they can be executed by a VisTrails server. This feature is useful in a number of scenarios, including the creation of Web-based interfaces that allows users to interact with workflows and the ability to run workflows in high-performance computing environments.

The Version Tree: Change-Based Provenance

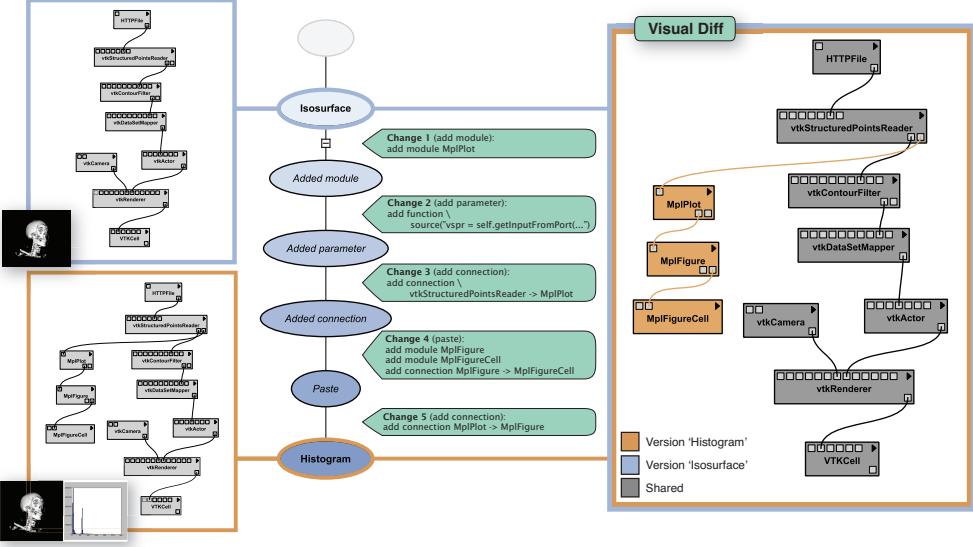


图 23.4: Change-Based Provenance Model

A new concept we introduced with VisTrails is the notion of provenance of workflow evolution [FSC⁺06]. In contrast to previous workflow and workflow-based visualization systems, which maintain provenance only for derived data products, VisTrails treats the workflows as first-class data items and also captures their provenance. The availability of workflow-evolution provenance supports reflective reasoning. Users can explore multiple chains of reasoning without losing any results, and because the system stores intermediate results, users can reason about and make inferences from this information. It also enables a series of operations which simplify exploratory processes. For example, users can easily navigate through the space of workflows created for a given task, visually compare the workflows and their results (see 图 23.4), and explore (large) parameter spaces. In addition, users can query the provenance information and learn by example.

The workflow evolution is captured using the change-based provenance model. As illustrated in 图 23.4, VisTrails stores the operations or changes that are applied to workflows (e.g., the addition of a module, the modification of a parameter, etc.), akin to a database transaction log. This information is modeled as a tree, where each node corresponds to a workflow version, and an edge between a parent and a child node represents the change applied to the parent to obtain the child. We use the terms version tree and vistrail (short for *visual trail*) interchangeably to refer to this tree. Note that the change-based model uniformly captures both changes to parameter values and to workflow definitions. This sequence of changes is sufficient to determine the provenance of data products and it also captures information about how a workflow evolves over time. The model is both simple

and compact—it uses substantially less space than the alternative of storing multiple *versions* of a workflow.

There are a number of benefits that come from the use of this model.  23.4 shows the visual difference functionality that VisTrails provides for comparing two workflows. Although the workflows are represented as graphs, using the change-based model, comparing two workflows becomes very simple: it suffices to navigate the version tree and identify the series of actions required to transform one workflow into the other.

Another important benefit of the change-based provenance model is that the underlying version tree can serve as a mechanism to support collaboration. Because designing workflows is a notoriously difficult task, it often requires multiple users to collaborate. Not only does the version tree provide an intuitive way to visualize the contribution of different users (e.g., by coloring nodes according to the user who created the corresponding workflow), but the monotonicity of the model allows for simple algorithms for synchronizing changes performed by multiple users.

Provenance information can be easily captured while a workflow is being executed. Once the execution completes, it is also important to maintain *strong* links between a data product and its provenance, i.e., the workflow, parameters and input files used to derive the data product. When data files or provenance are moved or modified, it can be difficult to find the data associated with the provenance or to find the provenance associated with the data. VisTrails provides a persistent storage mechanism that manages input, intermediate, and output data files, strengthening the links between provenance and data. This mechanism provides better support for reproducibility because it ensures the data referenced in provenance information can be readily (and correctly) located. Another important benefit of such management is that it allows caching of intermediate data which can then be shared with other users.

Workflow Execution and Caching

The execution engine in VisTrails was designed to allow the integration of new and existing tools and libraries. We tried to accommodate different styles commonly used for wrapping third-party scientific visualization and computation software. In particular, VisTrails can be integrated with application libraries that exist either as pre-compiled binaries that are executed on a shell and use files as input/outputs, or as C++/Java/Python class libraries that pass internal objects as input/output.

VisTrails adopts a dataflow execution model, where each module performs a computation and the data produced by a module flows through the connections that exist between modules. Modules are executed in a bottom-up fashion; each input is generated on-demand by recursively executing upstream modules (we say module A is *upstream* of B when there is a sequence of connections that goes from A to B). The intermediate data is temporarily stored either in memory (as a Python object) or on disk (wrapped by a Python object that contains information on accessing the data).

To allow users to add their own functionality to VisTrails, we built an extensible package system (see 23.3 節). Packages allow users to include their own or third-party modules in VisTrails workflows. A package developer must identify a set of computational modules and for each, identify the input and output ports as well as define the computation. For existing libraries, a compute method needs to specify the translation from input ports to parameters for the existing function and the mapping from result values to output ports.

In exploratory tasks, similar workflows, which share common sub-structures, are often executed in close succession. To improve the efficiency of workflow execution, VisTrails caches intermediate results to minimize recomputation. Because we reuse previous execution results, we implicitly assume that cacheable modules are functional: given the same inputs, modules will produce the same outputs. This requirement imposes definite behavior restrictions on classes, but we believe they are reasonable.

There are, however, obvious situations where this behavior is unattainable. For example, a module that uploads a file to a remote server or saves a file to disk has a significant side effect while its output is relatively unimportant. Other modules might use randomization, and their non-determinism might be desirable; such modules can be flagged as non-cacheable. However, some modules that are not naturally functional can be converted; a function that writes data to two files might be wrapped to output the contents of the files.

Data Serialization and Storage

One of the key components of any system supporting provenance is the serialization and storage of data. VisTrails originally stored data in XML via simple `fromXML` and `toXML` methods embedded in its internal objects (e.g., the version tree, each module). To support the evolution of the schema of these objects, these functions encoded any translation between schema versions as well. As the project progressed, our user base grew, and we decided to support different serializations, including relational stores. In addition, as schema objects evolved, we needed to maintain better infrastructure for common data management concerns like versioning schemas, translating between versions, and supporting entity relationships. To do so, we added a new database (db) layer.

The db layer is composed of three core components: the domain objects, the service logic, and the persistence methods. The domain and persistence components are versioned so that each schema version has its own set of classes. This way, we maintain code to read each version of the schema. There are also classes that define translations for objects from one schema version to those of another. The service classes provide methods to interface with data and deal with detection and translation of schema versions.

Because writing much of this code is tedious and repetitive, we use templates and a meta-schema to define both the object layout (and any in-memory indices) and the serialization code. The meta-schema is written in XML, and is extensible in that serializations other than the default XML and

relational mappings VisTrails defines can be added. This is similar to object-relational mappings and frameworks like Hibernate² and SQLObject³, but adds some special routines to automate tasks like re-mapping identifiers and translating objects from one schema version to the next. In addition, we can also use the same meta-schema to generate serialization code for many languages. After originally writing meta-Python, where the domain and persistence code was generated by running Python code with variables obtained from the meta-schema, we have recently migrated to Mako templates⁴.

Automatic translation is key for users that need to migrate their data to newer versions of the system. Our design adds hooks to make this translation slightly less painful for developers. Because we maintain a copy of code for each version, the translation code just needs to map one version to another. At the root level, we define a map to identify how any version can be transformed to any other. For distant versions, this usually involves a chain through multiple intermediate versions. Initially, this was a forward-only map, meaning new versions could not be translated to old versions, but reverse mappings have been added for more-recent schema mappings.

Each object has an `update_version` method that takes a different version of an object and returns the current version. By default, it does a recursive translation where each object is upgraded by mapping fields of the old object to those in a new version. This mapping defaults to copying each field to one with the same name, but it is possible to define a method to "override" the default behavior for any field. An override is a method that takes the old object and returns a new version. Because most changes to the schema only affect a small number of fields, the default mappings cover most cases, but the overrides provide a flexible means for defining local changes.

Extensibility Through Packages and Python

The first prototype of VisTrails had a fixed set of modules. It was an ideal environment to develop basic ideas about the VisTrails version tree and the caching of multiple execution runs, but it severely limited long-term utility.

We see VisTrails as infrastructure for computational science, and that means, literally, that the system should provide scaffolding for other tools and processes to be developed. An essential requirement of this scenario is extensibility. A typical way to achieve this involves defining a target language and writing an appropriate interpreter. This is appealing because of the intimate control it offers over execution. This appeal is amplified in light of our caching requirements. However, implementing a full-fledged programming language is a large endeavor that has never been our primary goal. More importantly, forcing users who are just trying to use VisTrails to learn an entirely new language was out of the question.

²<http://www.hibernate.org>

³<http://www.sqlobject.org>

⁴<http://www.makotemplates.org>

We wanted a system which made it easy for a user to add custom functionality. At the same time, we needed the system to be powerful enough to express fairly complicated pieces of software. As an example, VisTrails supports the VTK visualization library⁵. VTK contains about 1000 classes, which change depending on compilation, configuration, and operating system. Since it seems counterproductive and ultimately hopeless to write different code paths for all these cases, we decided it was necessary to dynamically determine the set of VisTrails modules provided by any given package, and VTK naturally became our model target for a complex package.

Computational science was one of the areas we originally targeted, and at the time we designed the system, Python was becoming popular as "glue code" among these scientists. By specifying the behavior of user-defined VisTrails modules using Python itself, we would all but eliminate a large barrier for adoption. As it turns out, Python offers a nice infrastructure for dynamically-defined classes and reflection. Almost every definition in Python has an equivalent form as a first-class expression. The two important reflection features of Python for our package system are:

- Python classes can be defined dynamically via function calls to the type callable. The return value is a representation of a class that can be used in exactly the same way that a typically-defined Python class can.
- Python modules can be imported via function calls to `__import__`, and the resulting value behaves in the same way as the identifier in a standard `import` statement. The path from which these modules come from can also be specified at runtime.

Using Python as our target has a few disadvantages, of course. First of all, this dynamic nature of Python means that while we would like to ensure some things like type safety of VisTrails packages, this is in general not possible. More importantly, some of the requirements for VisTrails modules, notably the ones regarding referential transparency (more on that later) cannot be enforced in Python. Still, we believe that it is worthwhile to restrict the allowed constructs in Python via cultural mechanisms, and with this caveat, Python is an extremely attractive language for software extensibility.

VisTrails Packages and Bundles

A VisTrails package encapsulates a set of modules. Its most common representation in disk is the same representation as a Python package (in a possibly unfortunate naming clash). A Python package consists of a set of Python files which define Python values such as functions and classes. A VisTrails package is a Python package that respects a particular interface. It has files that define specific functions and variables. In its simplest form, a VisTrails package should be a directory containing two files: `__init__.py` and `init.py`.

⁵<http://www.vtk.org>

The first file `__init__.py` is a requirement of Python packages, and should only contain a few definitions which should be constant. Although there is no way to guarantee that this is the case, VisTrails packages failing to obey this are considered buggy. The values defined in the file include a globally unique identifier for the package which is used to distinguish modules when workflows are serialized, and package versions (package versions become important when handling workflow and package upgrades, see 23.4 節). This file can also include functions called `package_dependencies` and `package_requirements`. Since we allow VisTrails modules to subclass from other VisTrails modules beside the root `Module` class, it is conceivable for one VisTrails package to extend the behavior of another, and so one package needs to be initialized before another. These inter-package dependencies are specified by `package_dependencies`. The `package_requirements` function, on the other hand, specifies system-level library requirements which VisTrails, in some cases, can try to automatically satisfy, through its bundle abstraction.

A bundle is a system-level package that VisTrails manages via system-specific tools such as Red-Hat’s RPM or Ubuntu’s APT. When these properties are satisfied, VisTrails can determine the package properties by directly importing the Python module and accessing the appropriate variables.

The second file, `init.py`, contains the entry points for all the actual VisTrails module definitions. The most important feature of this file is the definition of two functions, `initialize` and `finalize`. The `initialize` function is called when a package is enabled, after all the dependent packages have themselves been enabled. It performs setup tasks for all of the modules in a package. The `finalize` function, on the other hand, is usually used to release runtime resources (for example, temporary files created by the package can be cleaned up).

Each VisTrails module is represented in a package by one Python class. To register this class in VisTrails, a package developer calls the `add_module` function once for each VisTrails module. These VisTrails modules can be arbitrary Python classes, but they must respect a few requirements. The first of these is that each must be a subclass of a basic Python class defined by VisTrails called, perhaps boringly, `Module`. VisTrails modules can use multiple inheritance, but only one of the classes should be a VisTrails module—no diamond hierarchies in the VisTrails module tree are allowed. Multiple inheritance becomes useful in particular to define class mix-ins: simple behaviors encoded by parent classes which can be composed together to create more complicated behaviors.

The set of available ports determine the interface of a VisTrails module, and so impact not only the display of these modules but also their connectivity to other modules. These ports, then, must be explicitly described to the VisTrails infrastructure. This can be done either by making appropriate calls to `add_input_port` and `add_output_port` during the call to `initialize`, or by specifying the per-class lists `_input_ports` and `_output_ports` for each VisTrails module.

Each module specifies the computation to be performed by overriding the `compute` method. Data is passed between modules through ports, and accessed through the `get_input_from_port` and `set_result` methods. In traditional dataflow environments, execution order is specified on-demand

by the data requests. In our case, the execution order is specified by the topological sorting of the workflow modules. Since the caching algorithm requires an acyclic graph, we schedule the execution in reverse topological sorted order, so the calls to these functions do not trigger executions of upstream modules. We made this decision deliberately: it makes it simpler to consider the behavior of each module separately from all the others, which makes our caching strategy simpler and more robust.

As a general guideline, VisTrails modules should refrain from using functions with side-effects during the evaluation of the `compute` method. As discussed in 23.3 節, this requirement makes caching of partial workflow runs possible: if a module respects this property, then its behavior is a function of the outputs of upstream modules. Every acyclic subgraph then only needs to be computed once, and the results can be reused.

Passing Data as Modules

One peculiar feature of VisTrails modules and their communication is that the data that is passed between VisTrails modules are themselves VisTrails modules. In VisTrails, there is a single hierarchy for module and data classes. For example, a module can provide *itself* as an output of a computation (and, in fact, every module provides a default "self" output port). The main disadvantage is the loss of conceptual separation between computation and data that is sometimes seen in dataflow-based architectures. There are, however, two big advantages. The first is that this closely mimics the object type systems of Java and C++, and the choice was not accidental: it was very important for us to support automatic wrapping of large class libraries such as VTK. These libraries allow objects to produce other objects as computational results, making a wrapping that distinguishes between computation and data more complicated.

The second advantage this decision brings is that defining constant values and user-settable parameters in workflows becomes easier and more uniformly integrated with the rest of the system. Consider, for example, a workflow that loads a file from a location on the Web specified by a constant. This is currently specified by a GUI in which the URL can be specified as a parameter (see the Parameter Edits area in 図 23.1). A natural modification of this workflow is to use it to fetch a URL that is *computed* somewhere upstream. We would like the rest of the workflow to change as little as possible. By assuming modules can output themselves, we can simply connect a string with the right value to the port corresponding to the parameter. Since the output of a constant evaluates to itself, the behavior is exactly the same as if the value had actually been specified as a constant.

There are other considerations involved in designing constants. Each constant type has a different ideal GUI interface for specifying values. For example, in VisTrails, a file constant module provides a file chooser dialog; a Boolean value is specified by a checkbox; a color value has a color picker native to each operating system. To achieve this generality, a developer must subclass a custom

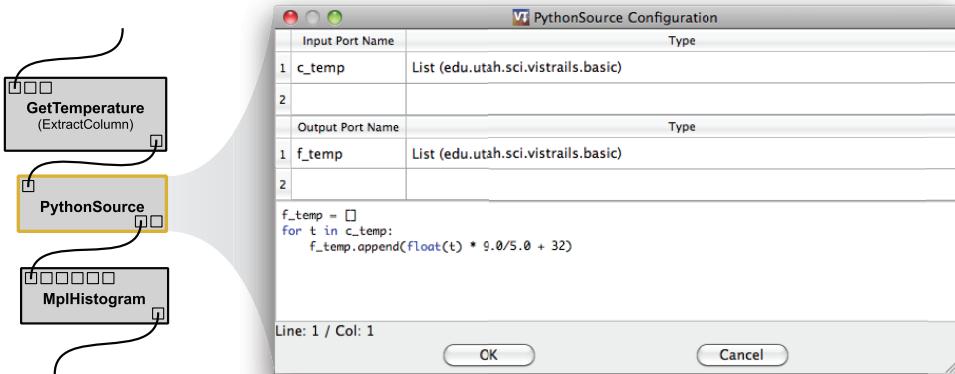


図 23.5: Prototyping New Functionality with the PythonSource Module

constant from the Constant base class and provide overrides which define an appropriate GUI widget and a string representation (so that arbitrary constants can be serialized to disk).

We note that, for simple prototyping tasks, VisTrails provides a built-in PythonSource module. A PythonSource module can be used to directly insert scripts into a workflow. The configuration window for PythonSource (see 図 23.5) allows multiple input and output ports to be specified along with the Python code that is to be executed.

23.4 Components and Features

As discussed above, VisTrails provides a set of functionalities and user interfaces that simplify the creation and execution of exploratory computational tasks. Below, we describe some of these. We also briefly discuss how VisTrails is being used as the basis for an infrastructure that supports the creation of provenance-rich publications. For a more comprehensive description of VisTrails and its features, see VisTrails' online documentation⁶.

Visual Spreadsheet

VisTrails allows users to explore and compare results from multiple workflows using the Visual Spreadsheet (see 図 23.6). The spreadsheet is a VisTrails package with its own interface composed of sheets and cells. Each sheet contains a set of cells and has a customizable layout. A cell contains the visual representation of a result produced by a workflow, and can be customized to display diverse types of data.

⁶<http://www.vistrails.org/usersguide>



図 23.6: The Visual Spreadsheet

To display a cell on the spreadsheet, a workflow must contain a module that is derived from the base `SpreadsheetCell` module. Each `SpreadsheetCell` module corresponds to a cell in the spreadsheet, so one workflow can generate multiple cells. The `compute` method of the `SpreadsheetCell` module handles the communication between the Execution Engine (図 23.3) and the spreadsheet. During execution, the spreadsheet creates a cell according to its type on-demand by taking advantage of Python’s dynamic class instantiation. Thus, custom visual representations can be achieved by creating a subclass of `SpreadsheetCell` and having its `compute` method send a custom cell type to the spreadsheet. For example, the workflow in 図 23.1, `MplFigureCell` is a `SpreadsheetCell` module designed to display images created by `matplotlib`.

Since the spreadsheet uses `PyQt` as its GUI back end, custom cell widgets must be subclassed

from PyQt's `QWidget`. They must also define the `updateContents` method, which is invoked by the spreadsheet to update the widget when new data arrives. Each cell widget may optionally define a custom toolbar by implementing the `toolbar` method; it will be displayed in the spreadsheet toolbar area when the cell is selected.

图 23.6 shows the spreadsheet when a VTK cell is selected, in this case, the toolbar provides specific widgets to export PDF images, save camera positions back to the workflow, and create animations. The spreadsheet package defines a customizable `QCellWidget`, which provides common features such as history replay (animation) and multi-touch events forwarding. This can be used in place of `QWidget` for faster development of new cell types.

Even though the spreadsheet only accepts PyQt widgets as cell types, it is possible to integrate widgets written with other GUI toolkits. To do so, the widget must export its elements to the native platform, and PyQt can then be used to grab it. We use this approach for the `VTKCell` widget because the actual widget is written in C++. At run-time, the `VTKCell` grabs the window id, a Win32, X11, or Cocoa/Carbon handle depending on the system, and maps it to the spreadsheet canvas.

Like cells, sheets may also be customized. By default, each sheet lives in a tabbed view and has a tabular layout. However, any sheet can be undocked from the spreadsheet window, allowing multiple sheets to be visible at once. It is also possible to create a different sheet layout by subclassing the `StandardWidgetSheet`, also a PyQt widget. The `StandardWidgetSheet` manages cell layouts as well as interactions with the spreadsheet in editing mode. In editing mode, users can manipulate the cell layout and perform advanced actions on the cells, rather than interacting with cell contents. Such actions include applying analogies (see 23.4 節) and creating new workflow versions from parameter explorations.

Visual Differences and Analogies

As we designed VisTrails, we wanted to enable the *use* of provenance information in addition to its capture. First, we wanted users to see the exact differences between versions, but we then realized that a more helpful feature was being able to apply these differences to other workflows. Both of these tasks are possible because VisTrails tracks the evolution of workflows.

Because the version tree captures all of the changes and we can invert each action, we can find a complete sequence of actions that transform one version to another. Note that some changes will cancel each other out, making it possible to compress this sequence. For example, the addition of a module that was later deleted need not be examined when computing the difference. Finally, we have some heuristics to further simplify the sequence: when the same module occurs in both workflows but was added through separate actions, we cancel the adds and deletes.

From the set of changes, we can create a visual representation that shows similar and different modules, connections, and parameters. This is illustrated in 图 23.4. Modules and connections that appear in both workflows are colored gray, and those appearing in only one are colored according

to the workflow they appear in. Matching modules with different parameters are shaded a lighter gray and a user can inspect the parameter differences for a specific module in a table that shows the values in each workflow.

The analogy operation allows users to take these differences and apply them to other workflows. If a user has made a set of changes to an existing workflow (e.g., changing the resolution and file format of an output image), he can apply the same changes to other workflows via an analogy. To do so, the user selects a source and a target workflow, which delimits the set of desired changes, as well as the workflow they wish to apply the analogy to. VisTrails computes the difference between the first two workflows as a template, and then determines how to remap this difference in order to apply it to the third workflow. Because it is possible to apply differences to workflows that do not exactly match the starting workflow, we need a soft matching that allows correspondences between similar modules. With this matching, we can remap the difference so the sequence of changes can be applied to the selected workflow [SVK⁺07]. The method is not foolproof and may generate new workflows that are not exactly what was desired. In such cases, a user may try to fix any introduced mistakes, or go back to the previous version and apply the changes manually.

To compute the soft matching used in analogies, we want to balance local matches (identical or very similar modules) with the overall workflow structure. Note that the computation of even the identical matching is inefficient due to the hardness of subgraph isomorphism, so we need to employ a heuristic. In short, if two somewhat-similar modules in the two workflows share similar neighbors, we might conclude that these two modules function similarly and should be matched as well. More formally, we construct a product graph where each node is a possible pairing of modules in the original workflows and an edge denotes shared connections. Then, we run steps diffusing the scores at each node across the edges to neighboring nodes. This is a Markov process similar to Google’s PageRank, and will eventually converge leaving a set of scores that now includes some global information. From these scores, we can determine the best matching, using a threshold to leave very dissimilar modules unpaired.

Querying Provenance

The provenance captured by VisTrails includes a set of workflows, each with its own structure, metadata, and execution logs. It is important that users can access and explore these data. VisTrails provides both text-based and visual (WYSIWYG) query interfaces. For information like tags, annotations, and dates, a user can use keyword search with optional markup. For example, look for all workflows with the keyword `plot` that were created by user `:~dakoop`. However, queries for specific subgraphs of a workflow are more easily represented through a visual, query-by-example interface, where users can either build the query from scratch or copy and modify an existing piece of a pipeline.

In designing this query-by-example interface, we kept most of the code from the existing Workflow Editor, with a few changes to parameter construction. For parameters, it is often useful to search for ranges or keywords rather than exact values. Thus, we added modifiers to the parameter value fields; when a user adds or edits a parameter value, they may choose to select one of these modifiers which default to exact matches. In addition to visual query construction, query results are shown visually. Matching versions are highlighted in the version tree, and any selected workflow is displayed with the matching portion highlighted. The user can exit query results mode by initiating another query or clicking a reset button.

Persistent Data

VisTrails saves the provenance of how results were derived and the specification of each step. However, reproducing a workflow run can be difficult if the data needed by the workflow is no longer available. In addition, for long-running workflows, it may be useful to store intermediate data as a persistent cache across sessions in order to avoid recomputation.

Many workflow systems store filesystem paths to data as provenance, but this approach is problematic. A user might rename a file, move the workflow to another system without copying the data, or change the data contents. In any of these cases, storing the path as provenance is not sufficient. Hashing the data and storing the hash as provenance helps to determine whether the data might have changed, but does not help one locate the data if it exists. To solve this problem, we created the Persistence Package, a VisTrails package that uses version control infrastructure to store data that can be referenced from provenance. Currently we use Git to manage the data, although other systems could easily be employed.

We use universally unique identifiers (UUIDs) to identify data, and commit hashes from git to reference versions. If the data changes from one execution to another, a new version is checked in to the repository. Thus, the (uuid, version) tuple is a compound identifier to retrieve the data in any state. In addition, we store the hash of the data as well as the signature of the upstream portion of the workflow that generated it (if it is not an input). This allows one to link data that might be identified differently as well as reuse data when the same computation is run again.

The main concern when designing this package was the way users were able to select and retrieve their data. Also, we wished to keep all data in the same repository, regardless of whether it is used as input, output, or intermediate data (an output of one workflow might be used as the input of another). There are two main modes a user might employ to identify data: choosing to create a new reference or using an existing one. Note that after the first execution, a new reference will become an existing one as it has been persisted during execution; a user may later choose to create another reference if they wish but this is a rare case. Because a user often wishes to always use the latest version of data, a reference identified without a specific version will default to the latest version.

Recall that before executing a module, we recursively update all of its inputs. A persistent data module will not update its inputs if the upstream computations have already been run. To determine this, we check the signature of the upstream subworkflow against the persistent repository and retrieve the precomputed data if the signature exists. In addition, we record the data identifiers and versions as provenance so that a specific execution can be reproduced.

Upgrades

With provenance at the core of VisTrails, the ability to upgrade old workflows so they will run with new versions of packages is a key concern. Because packages can be created by third-parties, we need both the infrastructure for upgrading workflows as well as the hooks for package developers to specify the upgrade paths. The core action involved in workflow upgrades is the replacement of one module with a new version. Note that this action is complicated because we must replace all of the connections and parameters from the old module. In addition, upgrades may need to reconfigure, reassign, or rename these parameters or connections for a module, e.g., when the module interface changes.

Each package (together with its associated modules) is tagged by a version, and if that version changes, we assume that the modules in that package may have changed. Note that some, or even most, may not have changed, but without doing our own code analysis, we cannot check this. We, however, attempt to automatically upgrade any module whose interface has not changed. To do this, we try replacing the module with the new version and throw an exception if it does not work. When developers have changed the interface of a module or renamed a module, we allow them to specify these changes explicitly. To make this more manageable, we have created a `remap_module` method that allows developers to define only the places where the default upgrade behavior needs to be modified. For example, a developer that renamed an input port ‘file’ to ‘value’ can specify that specific remapping so when the new module is created, any connections to ‘file’ in the old module will now connect to ‘value’. Here is an example of an upgrade path for a built-in VisTrails module:

```
def handle_module_upgrade_request(controller, module_id, pipeline):
    module_remap = {'GetItemsFromDirectory':
                    [(None, '1.6', 'Directory',
                      {'dst_port_remap':
                         {'dir': 'value'},
                         'src_port_remap':
                           {'itemlist': 'itemList'}},
                      )],
                    }
    return UpgradeWorkflowHandler.remap_module(controller, module_id, pipeline,
                                                module_remap)
```

This piece of code upgrades workflows that use the old `GetItemsFromDirectory` (any version up to 1.6) module to use the `Directory` module instead. It maps the `dir` port from the old module

to value and the `itemlist` port to `itemList`.

Any upgrade creates a new version in the version tree so that executions before and after upgrades can be differentiated and compared. It is possible that the upgrades change the execution of the workflow (e.g., if a bug is fixed by a package developer), and we need to track this as provenance information. Note that in older vistrails, it may be necessary to upgrade every version in the tree. In order to reduce clutter, we only upgrade versions that a user has navigated to. In addition, we provide a preference that allows a user to delay the persistence of any upgrade until the workflow is modified or executed; if a user just views that version, there is no need to persist the upgrade.

Sharing and Publishing Provenance-Rich Results

While reproducibility is the cornerstone of the scientific method, current publications that describe computational experiments often fail to provide enough information to enable the results to be repeated or generalized. Recently, there has been a renewed interest in the publication of reproducible results. A major roadblock to the more widespread adoption of this practice is the fact that it is hard to create a bundle that includes all of the components (e.g., data, code, parameter settings) needed to reproduce a result as well as verify that result.

By capturing detailed provenance, and through many of the features described above, VisTrails simplifies this process for computational experiments that are carried out within the system. However, mechanisms are needed to both link documents to and share the provenance information.

We have developed VisTrails packages that enable results present in papers to be linked to their provenance, like a deep caption. Using the LaTeX package we developed, users can include figures that link to VisTrails workflows. The following LaTeX code will generate a figure that contains a workflow result:

```
\begin{figure}[t]
\centering
\vistrail[wfid=119,buildalways=false]{width=0.9\linewidth}
}
\caption{Visualizing a binary star system simulation. This is an image
that was generated by embedding a workflow directly in the text.}
\label{fig:astrophysics}
\end{figure}
```

When the document is compiled using pdflatex, the `\vistrail` command will invoke a Python script with the parameters received, which sends an XML-RPC message to a VisTrails server to execute the workflow with id 119. This same Python script downloads the results of the workflow from the server and includes them in the resulting PDF document by generating hyperlinked LaTeX `\includegraphics` commands using the specified layout options (`width=0.9\linewidth`).

It is also possible to include VisTrails results into Web pages, wikis, Word documents and PowerPoint presentations. The linking between Microsoft PowerPoint and VisTrails was done through the

Component Object Model (COM) and Object Linking and Embedding (OLE) interface. In order for an object to interact with PowerPoint, at least the `IObject`, `IDataObject` and `IPersistStorage` interface of COM must be implemented. As we use the `QAxAggregated` class of Qt, which is an abstraction for implementing COM interfaces, to build our OLE object, both `IDataObject` and `IPersistStorage` are automatically handled by Qt. Thus, we only need to implement the `IObject` interface. The most important call in this interface is `DoVerb`. It lets VisTrails react to certain actions from PowerPoint, such as object activation. In our implementation, when the VisTrails object is activated, we load the VisTrails application and allow users to open, interact with and select a pipeline that they want to insert. After they close VisTrails, the pipeline result will be shown in PowerPoint. Pipeline information is also stored with the OLE object.

To enable users to freely share their results together with the associated provenance, we have created crowdLabs.⁷ crowdLabs is a social Web site that integrates a set of usable tools and a scalable infrastructure to provide an environment for scientists to collaboratively analyze and visualize data. crowdLabs is tightly integrated with VisTrails. If a user wants to share any results derived in VisTrails, she can connect to the crowdLabs server directly from VisTrails to upload the information. Once the information is uploaded, users can interact with and execute the workflows through a Web browser—these workflows are executed by a VisTrails server that powers crowdLabs. For more details on how VisTrails is used to create reproducible publications, see <http://www.vistrails.org>.

23.5 Lessons Learned

Luckily, back in 2004 when we started thinking about building a data exploration and visualization system that supported provenance, we never envisioned how challenging it would be, or how long it would take to get to the point we are at now. If we had, we probably would never have started.

Early on, one strategy that worked well was quickly prototyping new features and showing them to a select set of users. The initial feedback and the encouragement we received from these users was instrumental in driving the project forward. It would have been impossible to design VisTrails without user feedback. If there is one aspect of the project that we would like to highlight is that most features in the system were designed as direct response to user feedback. However, it is worthy to note that many times what a user asks for is not the best solution for his/her need—being responsive to users does not necessarily mean doing exactly what they ask for. Time and again, we have had to design and re-design features to make sure they would be useful and properly integrated in the system.

Given our user-centric approach, one might expect that every feature we have developed would be heavily used. Unfortunately this has not been the case. Sometimes the reason for this is that

⁷<http://www.crowdlabs.org>

the feature is highly "unusual", since it is not found in other tools. For instance, analogies and even the version tree are not concepts that most users are familiar with, and it takes a while for them to get comfortable with them. Another important issue is documentation, or lack thereof. As with many other open source projects, we have been much better at developing new features than at documenting the existing ones. This lag in documentation leads not only to the underutilization of useful features, but also to many questions on our mailing lists.

One of the challenges of using a system like VisTrails is that it is very general. Despite our best efforts to improve usability, VisTrails is a complex tool and requires a steep learning curve for some users. We believe that over time, with improved documentation, further refinements to the system, and more application- and domain-specific examples, the adoption bar for any given field will get lower. Also, as the concept of provenance becomes more widespread, it will be easier for users to understand the philosophy that we have adopted in developing VisTrails.

Acknowledgments

We would like to thank all the talented developers that contributed to VisTrails: Erik Anderson, Louis Bavoil, Clifton Brooks, Jason Callahan, Steve Callahan, Lorena Carlo, Lauro Lins, Tommy Ellkvist, Phillip Mates, Daniel Rees, and Nathan Smith. Special thanks to Antonio Baptista who was instrumental in helping us develop the vision for the project; and Matthias Troyer, whose collaboration has helped us to improve the system, and in particular has provided much of the impetus for the development and release of the provenance-rich publication functionality. The research and development of the VisTrails system has been funded by the National Science Foundation under grants IIS 1050422, IIS-0905385, IIS 0844572, ATM-0835821, IIS-0844546, IIS-0746500, CNS-0751152, IIS-0713637, OCE-0424602, IIS-0534628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, CCF-0528201, CNS-0551724, the Department of Energy SciDAC (VACET and SDM centers), and IBM Faculty Awards.

VTK

Berk Geveci and Will Schroeder

The Visualization Toolkit (VTK) is a widely used software system for data processing and visualization. It is used in scientific computing, medical image analysis, computational geometry, rendering, image processing and informatics. In this chapter we provide a brief overview of VTK, including some of the basic design patterns that make it a successful system.

To really understand a software system it is essential to not only understand what problem it solves, but also the particular culture in which it emerged. In the case of VTK, the software was ostensibly developed as a 3D visualization system for scientific data. But the cultural context in which it emerged adds a significant back story to the endeavor, and helps explains why the software was designed and deployed as it was.

At the time VTK was conceived and written, its initial authors (Will Schroeder, Ken Martin, Bill Lorensen) were researchers at GE Corporate R&D. We were heavily invested in a precursor system known as LYMB which was a Smalltalk-like environment implemented in the C programming language. While this was a great system for its time, as researchers we were consistently frustrated by two major barriers when trying to promote our work: 1) IP issues and 2) non-standard, proprietary software. IP issues were a problem because trying to distribute the software outside of GE was nearly impossible once the corporate lawyers became involved. Second, even if we were deploying the software inside of GE, many of our customers balked at learning a proprietary, non-standard system since the effort to master it did not transition with an employee once she left the company, and it did not have the widespread support of a standard tool set. Thus in the end the primary motivation for VTK was to develop an open standard, or *collaboration platform* through which we could easily transition technology to our customers. Thus choosing an open source license for VTK was probably the most important design decision that we made.

The final choice of a non-reciprocal, permissive license (i.e., BSD not GPL) in hindsight was an exemplary decision made by the authors because it ultimately enabled the service and consulting based business that became Kitware. At the time we made the decision we were mostly interested

in reduced barriers to collaborating with academics, research labs, and commercial entities. We have since discovered that reciprocal licenses are avoided by many organizations because of the potential havoc they can wreak. In fact we would argue that reciprocal licenses do much to slow the acceptance of open source software, but that is an argument for another time. The point here is: one of the major design decisions to make relative to any software system is the choice of copyright license. It's important to review the goals of the project and then address IP issues appropriately.

24.1 What Is VTK?

VTK was initially conceived as a scientific data visualization system. Many people outside of the field naively consider visualization a particular type of geometric rendering: examining virtual objects and interacting with them. While this is indeed part of visualization, in general data visualization includes the whole process of transforming data into sensory input, typically images, but also includes tactile, auditory, and other forms. The data forms not only consist of geometric and topological constructs, including such abstractions as meshes or complex spatial decompositions, but attributes to the core structure such as scalars (e.g., temperature or pressure), vectors (e.g., velocity), tensors (e.g., stress and strain) plus rendering attributes such as surface normals and texture coordinate.

Note that data representing spatial-temporal information is generally considered part of scientific visualization. However there are more abstract data forms such as marketing demographics, web pages, documents and other information that can only be represented through abstract (i.e., non-spatial temporal) relationships such as unstructured documents, tables, graphs, and trees. These abstract data are typically addressed by methods from information visualization. With the help of the community, VTK is now capable of both scientific and information visualization.

As a visualization system, the role of VTK is to take data in these forms and ultimately transform them into forms comprehensible by the human sensory apparatus. Thus one of the core requirements of VTK is its ability to create data flow pipelines that are capable of ingesting, processing, representing and ultimately rendering data. Hence the toolkit is necessarily architected as a flexible system and its design reflects this on many levels. For example, we purposely designed VTK as a toolkit with many interchangeable components that can be combined to process a wide variety of data.

24.2 Architectural Features

Before getting too far into the specific architectural features of VTK, there are high-level concepts that have significant impact on developing and using the system. One of these is VTK's hybrid wrapper facility. This facility automatically generates language bindings to Python, Java, and

Tcl from VTK's C++ implementation (additional languages could be and have been added). Most high-powered developers will work in C++. User and application developers may use C++ but often the interpreted languages mentioned above are preferred. This hybrid compiled/interpreted environment combines the best of both worlds: high performance compute-intensive algorithms and flexibility when prototyping or developing applications. In fact this approach to multi-language computing has found favor with many in the scientific computing community and they often use VTK as a template for developing their own software.

In terms of software process, VTK has adopted CMake to control the build; CDash/CTest for testing; and CPack for cross-platform deployment. Indeed VTK can be compiled on almost any computer including supercomputers which are often notoriously primitive development environments. In addition, web pages, wiki, mailing lists (user and developer), documentation generation facilities (i.e., Doxygen) and a bug tracker (Mantis) round out the development tools.

Core Features

As VTK is an object-oriented system, the access of class and instance data members is carefully controlled in VTK. In general, all data members are either protected or private. Access to them is through Set and Get methods, with special variations for Boolean data, modal data, strings and vectors. Many of these methods are actually created by inserting macros into the class header files. So for example:

```
vtkSetMacro(Tolerance,double);
vtkGetMacro(Tolerance,double);
```

become on expansion:

```
virtual void SetTolerance(double);
virtual double GetTolerance();
```

There are many reasons for using these macros beyond simply code clarity. In VTK there are important data members controlling debugging, updating an object's modified time (MTIME), and properly managing reference counting. These macros correctly manipulate these data and their use is highly recommended. For example, a particularly pernicious bug in VTK occurs when the object's MTIME is not managed properly. In this case code may not execute when it should, or may execute too often.

One of the strengths of VTK is its relatively simplistic means of representing and managing data. Typically various data arrays of particular types (e.g., vtkFloatArray) are used to represent contiguous pieces of information. For example, a list of three XYZ points would be represented with a vtkFloatArray of nine entries (x,y,z, x,y,z, etc.) There is the notion of a tuple in these arrays, so a 3D point is a 3-tuple, whereas a symmetric 3×3 tensor matrix is represented by a 6-tuple (where

symmetry space savings are possible). This design was adopted purposely because in scientific computing it is common to interface with systems manipulating arrays (e.g., Fortran) and it is much more efficient to allocate and deallocate memory in large contiguous chunks. Further, communication, serializing and performing IO is generally much more efficient with contiguous data. These core data arrays (of various types) represent much of the data in VTK and have a variety of convenience methods for inserting and accessing information, including methods for fast access, and methods that automatically allocate memory as needed when adding more data. Data arrays are subclasses of the `vtkDataArray` abstract class meaning that generic virtual methods can be used to simplify coding. However, for higher performance static, templated functions are used which switch based on type, with subsequent, direct access into the contiguous data arrays.

In general C++ templates are not visible in the public class API; although templates are used widely for performance reasons. This goes for STL as well: we typically employ the PIMPL¹ design pattern to hide the complexities of a template implementation from the user or application developer. This has served us particularly well when it comes to wrapping the code into interpreted code as described previously. Avoiding the complexity of the templates in the public API means that the VTK implementation, from the application developer point of view, is mostly free of the complexities of data type selection. Of course under the hood the code execution is driven by the data type which is typically determined at run time when the data is accessed.

Some users wonder why VTK uses reference counting for memory management versus a more user-friendly approach such as garbage collection. The basic answer is that VTK needs complete control over when data is deleted, because the data sizes can be huge. For example, a volume of byte data $1000 \times 1000 \times 1000$ in size is a gigabyte in size. It is not a good idea to leave such data lying around while the garbage collector decides whether or not it is time to release it. In VTK most classes (subclasses of `vtkObject`) have the built-in capability for reference counting. Every object contains a reference count that is initialized to one when the object is instantiated. Every time a use of the object is registered, the reference count is increased by one. Similarly, when a use of the object is unregistered (or equivalently the object is deleted) the reference count is reduced by one. Eventually the object's reference count is reduced to zero, at which point it self destructs. A typical example looks like the following:

```
vtkCamera *camera = vtkCamera::New();      //reference count is 1
camera->Register(this);                  //reference count is 2
camera->Unregister(this);                //reference count is 1
renderer->SetActiveCamera(camera);       //reference count is 2
renderer->Delete();                     //ref count is 1 when renderer is deleted
camera->Delete();                       //camera self destructs
```

There is another important reason why reference counting is important to VTK—it provides the ability to efficiently copy data. For example, imagine a data object D1 that consists of a number

¹http://en.wikipedia.org/wiki/Opaque_pointer.

of data arrays: points, polygons, colors, scalars and texture coordinates. Now imagine processing this data to generate a new data object D2 which is the same as the first plus the addition of vector data (located on the points). One wasteful approach is to completely (deep) copy D1 to create D2, and then add the new vector data array to D2. Alternatively, we create an empty D2 and then pass the arrays from D1 to D2 (shallow copy), using reference counting to keep track of data ownership, finally adding the new vector array to D2. The latter approach avoids copying data which, as we have argued previously, is essential to a good visualization system. As we will see later in this chapter, the data processing pipeline performs this type of operation routinely, i.e., copying data from the input of an algorithm to the output, hence reference counting is essential to VTK.

Of course there are some notorious problems with reference counting. Occasionally reference cycles can exist, with objects in the cycle referring to each other in a mutually supportive configuration. In this case, intelligent intervention is required, or in the case of VTK, the special facility implemented in `vtkGarbageCollector` is used to manage objects which are involved in cycles. When such a class is identified (this is anticipated during development), the class registers itself with the garbage collector and overloads its own `Register` and `UnRegister` methods. Then a subsequent object deletion (or `unregister`) method performs a topological analysis on the local reference counting network, searching for detached islands of mutually referencing objects. These are then deleted by the garbage collector.

Most instantiation in VTK is performed through an object factory implemented as a static class member. The typical syntax appears as follows:

```
vtkLight *a = vtkLight::New();
```

What is important to recognize here is what is actually instantiated may not be a `vtkLight`, it could be a subclass of `vtkLight` (e.g., `vtkOpenGLLight`). There are a variety of motivations for the object factory, the most important being application portability and device independence. For example, in the above we are creating a light in a rendered scene. In a particular application on a particular platform, `vtkLight::New` may result in an OpenGL light, however on different platforms there is potential for other rendering libraries or methods for creating a light in the graphics system. Exactly what derived class to instantiate is a function of run-time system information. In the early days of VTK there were a myriad of options including gl, PHIGS, Starbase, XGL, and OpenGL. While most of these have now vanished, new approaches have appeared including DirectX and GPU-based approaches. Over time, an application written with VTK has not had to change as developers have derived new device specific subclasses to `vtkLight` and other rendering classes to support evolving technology. Another important use of the object factory is to enable the run-time replacement of performance-enhanced variations. For example, a `vtkImageFFT` may be replaced with a class that accesses special-purpose hardware or a numerics library.

Representing Data

One of the strengths of VTK is its ability to represent complex forms of data. These data forms range from simple tables to complex structures such as finite element meshes. All of these data forms are subclasses of `vtkDataObject` as shown in [図 24.1](#) (note this is a partial inheritance diagram of the many data object classes).

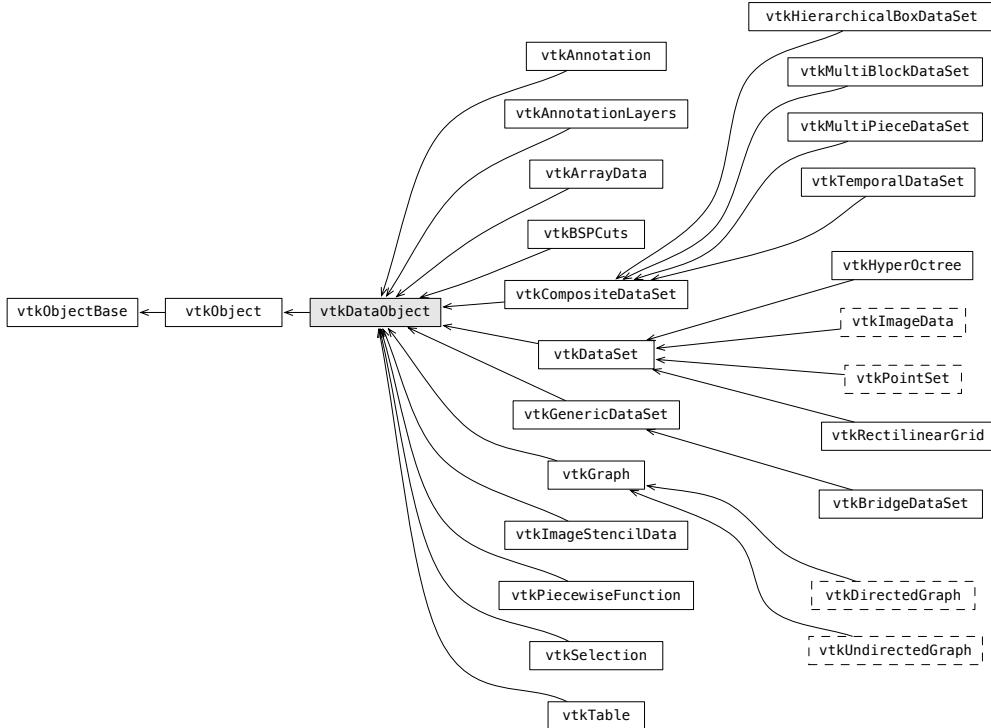


図 24.1: Data Object Classes

One of the most important characteristics of `vtkDataObject` is that it can be processed in a visualization pipeline (next subsection). Of the many classes shown, there are just a handful that are typically used in most real world applications. `vtkDataSet` and derived classes are used for scientific visualization ([図 24.2](#)). For example, `vtkPolyData` is used to represent polygonal meshes; `vtkUnstructuredGrid` to represent meshes, and `vtkImageData` represents 2D and 3D pixel and voxel data.

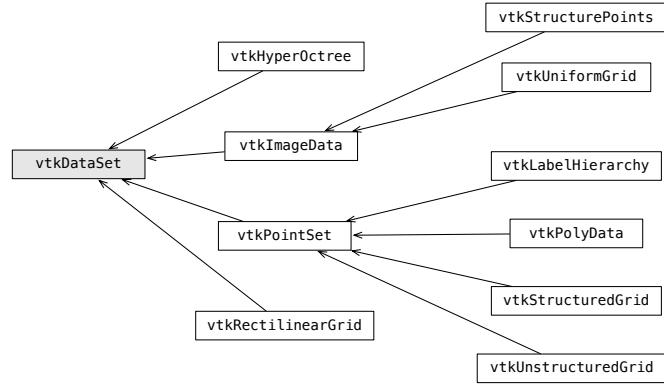


図 24.2: Data Set Classes

Pipeline Architecture

VTK consists of several major subsystems. Probably the subsystem most associated with visualization packages is the data flow/pipeline architecture. In concept, the pipeline architecture consists of three basic classes of objects: objects to represent data (the `vtkDataObjects` discussed above), objects to process, transform, filter or map data objects from one form into another (`vtkAlgorithm`); and objects to execute a pipeline (`vtkExecutive`) which controls a connected graph of interleaved data and process objects (i.e., the pipeline). 図 24.3 depicts a typical pipeline.

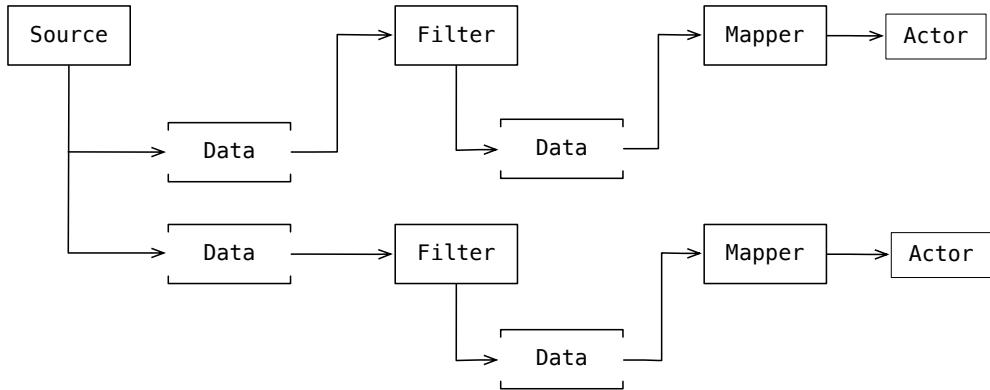


図 24.3: Typical Pipeline

While conceptually simple, actually implementing the pipeline architecture is challenging. One reason is that the representation of data can be complex. For example, some datasets consist of hierarchies or grouping of data, so executing across the data requires non-trivial iteration or recursion. To compound matters, parallel processing (whether using shared-memory or scalable, distributed

approaches) require partitioning data into pieces, where pieces may be required to overlap in order to consistently compute boundary information such as derivatives.

The algorithm objects also introduce their own special complexity. Some algorithms may take multiple inputs and/or produce multiple outputs of different types. Some can operate locally on data (e.g., compute the center of a cell) while others require global information, for example to compute a histogram. In all cases, the algorithms treat their inputs as immutable, algorithms only read their input in order to produce their output. This is because data may be available as input to multiple algorithms, and it is not a good idea for one algorithm to trample on the input of another.

Finally the executive can be complicated depending on the particulars of the execution strategy. In some cases we may wish to cache intermediate results between filters. This minimizes the amount of recomputation that must be performed if something in the pipeline changes. On the other hand, visualization data sets can be huge, in which case we may wish to release data when it is no longer needed for computation. Finally, there are complex execution strategies, such as multi-resolution processing of data, which require the pipeline to operate in iterative fashion.

To demonstrate some of these concepts and further explain the pipeline architecture, consider the following C++ example:

```
vtkPExodusIIReader *reader = vtkPExodusIIReader::New();
reader->SetFileName("exampleFile.exo");

vtkContourFilter *cont = vtkContourFilter::New();
cont->SetInputConnection(reader->GetOutputPort());
cont->SetNumberOfContours(1);
cont->SetValue(0, 200);

vtkQuadricDecimation *deci = vtkQuadricDecimation::New();
deci->SetInputConnection(cont->GetOutputPort());
deci->SetTargetReduction( 0.75 );

vtkXMLPolyDataWriter *writer = vtkXMLPolyDataWriter::New();
writer->SetInputConnection(deci->GetOutputPort());
writer->SetFileName("outputFile.vtp");
writer->Write();
```

In this example, a reader object reads a large unstructured grid (or mesh) data file. The next filter generates an isosurface from the mesh. The `vtkQuadricDecimation` filter reduces the size of the isosurface, which is a polygonal dataset, by decimating it (i.e., reducing the number of triangles representing the isocontour). Finally after decimation the new, reduced data file is written back to disk. The actual pipeline execution occurs when the `Write` method is invoked by the writer (i.e., upon demand for the data).

As this example demonstrates, VTK's pipeline execution mechanism is demand driven. When a sink such as a writer or a mapper (a data rendering object) needs data, it asks its input. If the input filter already has the appropriate data, it simply returns the execution control to the sink. However,

if the input does not have the appropriate data, it needs to compute it. Consequently, it must first ask its input for data. This process will continue upstream along the pipeline until a filter or source that has "appropriate data" or the beginning of the pipeline is reached, at which point the filters will execute in correct order and the data will flow to the point in the pipeline at which it was requested.

Here we should expand on what "appropriate data" means. By default, after a VTK source or filter executes, its output is cached by the pipeline in order to avoid unnecessary executions in the future. This is done to minimize computation and/or I/O at the cost of memory, and is configurable behavior. The pipeline caches not only the data objects but also the metadata about the conditions under which these data objects were generated. This metadata includes a time stamp (i.e., ComputeTime) that captures when the data object was computed. So in the simplest case, the "appropriate data" is one that was computed after all of the pipeline objects upstream from it were modified. It is easier to demonstrate this behavior by considering the following examples. Let's add the following to the end of the previous VTK program:

```
vtkXMLPolyDataWriter *writer2 = vtkXMLPolyDataWriter::New();
writer2->SetInputConnection(deci->GetOutputPort());
writer2->SetFileName("outputFile2.vtp");
writer2->Write();
```

As explained previously, the first `writer->Write` call causes the execution of the entire pipeline. When `writer2->Write()` is called, the pipeline will realize that the cached output of the decimation filter is up to date when it compares the time stamp of the cache with the modification time of the decimation filter, the contour filter and the reader. Therefore, the data request does not have to propagate past `writer2`. Now, let's consider the following change.

```
cont->SetValue(0, 400);

vtkXMLPolyDataWriter *writer2 = vtkXMLPolyDataWriter::New();
writer2->SetInputConnection(deci->GetOutputPort());
writer2->SetFileName("outputFile2.vtp");
writer2->Write();
```

Now the pipeline executive will realize that the contour filter was modified after the outputs of the contour and decimation filters were last executed. Thus, the cache for these two filters are stale and they have to be re-executed. However, since the reader was not modified prior to the contour filter its cache is valid and hence the reader does not have to re-execute.

The scenario described here is the simplest example of a demand-driven pipeline. VTK's pipeline is much more sophisticated. When a filter or a sink requires data, it can provide additional information to request specific data subsets. For example, a filter can perform out-of-core analysis by streaming pieces of data. Let's change our previous example to demonstrate.

```
vtkXMLPolyDataWriter *writer = vtkXMLPolyDataWriter::New();
writer->SetInputConnection(deci->GetOutputPort());
```

```

writer->SetNumberOfPieces(2);

writer->SetWritePiece(0);
writer->SetFileName("outputFile0.vtp");
writer->Write();

writer->SetWritePiece(1);
writer->SetFileName("outputFile1.vtp");
writer->Write();

```

Here the writer asks the upstream pipeline to load and process data in two pieces each of which are streamed independently. You may have noticed that the simple execution logic described previously will not work here. By this logic when the `Write` function is called for the second time, the pipeline should not re-execute because nothing upstream changed. Thus to address this more complex case, the executives have additional logic to handle piece requests such as this. VTK's pipeline execution actually consists of multiple passes. The computation of the data objects is actually the last pass. The pass before then is a request pass. This is where sinks and filters can tell upstream what they want from the forthcoming computation. In the example above, the writer will notify its input that it wants piece 0 of 2. This request will actually propagate all the way to the reader. When the pipeline executes, the reader will then know that it needs to read a subset of the data. Furthermore, information about which piece the cached data corresponds to is stored in the metadata for the object. The next time a filter asks for data from its input, this metadata will be compared with the current request. Thus in this example the pipeline will re-execute in order to process a different piece request.

There are several more types of request that a filter can make. These include requests for a particular time step, a particular structured extent or the number of ghost layers (i.e., boundary layers for computing neighborhood information). Furthermore, during the request pass, each filter is allowed to modify requests from downstream. For example, a filter that is not able to stream (e.g., the streamline filter) can ignore the piece request and ask for the whole data.

Rendering Subsystem

At first glance VTK has a simple object-oriented rendering model with classes corresponding to the components that make up a 3D scene. For example, `vtkActors` are objects that are rendered by a `vtkRenderer` in conjunction with a `vtkCamera`, with possibly multiple `vtkRenderers` existing in a `vtkRenderWindow`. The scene is illuminated by one or more `vtkLights`. The position of each `vtkActor` is controlled by a `vtkTransform`, and the appearance of an actor is specified through a `vtkProperty`. Finally, the geometric representation of an actor is defined by a `vtkMapper`. Mappers play an important role in VTK, they serve to terminate the data processing pipeline, as well as interface to the rendering system. Consider this example where we decimate data and write the result to a file, and then visualize and interact with the result by using a mapper:

```

vtkOBJReader *reader = vtkOBJReader::New();
reader->SetFileName("exampleFile.obj");

vtkTriangleFilter *tri = vtkTriangleFilter::New();
tri->SetInputConnection(reader->GetOutputPort());

vtkQuadricDecimation *deci = vtkQuadricDecimation::New();
deci->SetInputConnection(tri->GetOutputPort());
deci->SetTargetReduction( 0.75 );

vtkPolyDataMapper *mapper = vtkPolyDataMapper::New();
mapper->SetInputConnection(deci->GetOutputPort());

vtkActor *actor = vtkActor::New();
actor->SetMapper(mapper);

vtkRenderer *renderer = vtkRenderer::New();
renderer->AddActor(actor);

vtkRenderWindow *renWin = vtkRenderWindow::New();
renWin->AddRenderer(renderer);

vtkRenderWindowInteractor *interactor = vtkRenderWindowInteractor::New();
interactor->SetRenderWindow(renWin);

renWin->Render();

```

Here a single actor, renderer and render window are created with the addition of a mapper that connects the pipeline to the rendering system. Also note the addition of a `vtkRenderWindowInteractor`, instances of which capture mouse and keyboard events and translate them into camera manipulations or other actions. This translation process is defined via a `vtkInteractorStyle` (more on this below). By default many instances and data values are set behind the scenes. For example, an identity transform is constructed, as well as a single default (head) light and property.

Over time this object model has become more sophisticated. Much of the complexity has come from developing derived classes that specialize on an aspect of the rendering process. `vtkActors` are now specializations of `vtkProp` (like a prop found on stage), and there are a whole slew of these props for rendering 2D overlay graphics and text, specialized 3D objects, and even for supporting advanced rendering techniques such as volume rendering or GPU implementations (see 図 24.4).

Similarly, as the data model supported by VTK has grown, so have the various mappers that interface the data to the rendering system. Another area of significant extension is the transformation hierarchy. What was originally a simple linear 4×4 transformation matrix, has become a powerful hierarchy that supports non-linear transformations including thin-plate spline transformation. For example, the original `vtkPolyDataMapper` had device-specific subclasses (e.g., `vtkOpenGLPolyDataMapper`). In recent years it has been replaced with a sophisticated graphics pipeline referred to as the “painter” pipeline illustrated in 図 24.4.

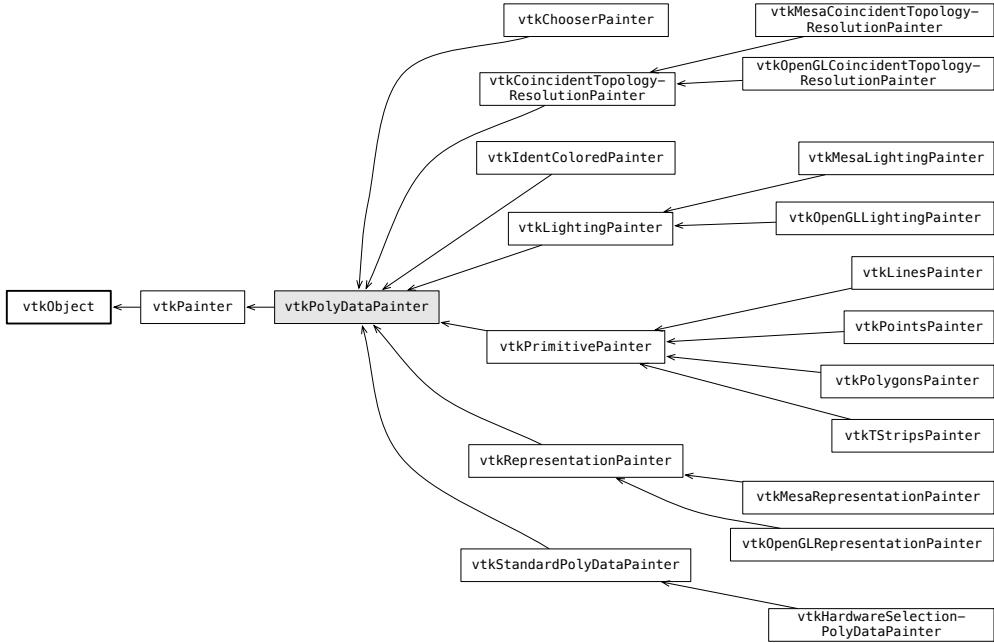


図 24.4: Display Classes

The painter design supports a variety of techniques for rendering data that can be combined to provide special rendering effects. This capability greatly surpasses the simple `vtkPolyDataMapper` that was initially implemented in 1994.

Another important aspect of a visualization system is the selection subsystem. In VTK there is a hierarchy of "pickers", roughly categorized into objects that select `vtkProps` based on hardware-based methods versus software methods (e.g., ray-casting); as well as objects that provide different levels of information after a pick operations. For example, some pickers provide only a location in XYZ world space without indicating which `vtkProp` they have selected; others provide not only the selected `vtkProp` but a particular point or cell that make up the mesh defining the prop geometry.

Events and Interaction

Interacting with data is an essential part of visualization. In VTK this occurs in a variety of ways. At its simplest level, users can observe events and respond appropriately through commands (the command/observer design pattern). All subclasses of `vtkObject` maintain a list of observers which register themselves with the object. During registration, the observers indicate which particular event(s) they are interested in, with the addition of an associated command that is invoked if and when the event occurs. To see how this works, consider the following example in which a filter

(here a polygon decimation filter) has an observer which watches for the three events StartEvent, ProgressEvent, and EndEvent. These events are invoked when the filter begins to execute, periodically during execution, and then on completion of execution. In the following the vtkCommand class has an Execute method that prints out the appropriate information relative to the time it takes to execute the algorithm:

```
class vtkProgressCommand : public vtkCommand
{
public:
    static vtkProgressCommand *New() { return new vtkProgressCommand; }
    virtual void Execute(vtkObject *caller, unsigned long, void *callData)
    {
        double progress = *(static_cast<double*>(callData));
        std::cout << "Progress at " << progress << std::endl;
    }
};

vtkCommand* pobserver = vtkProgressCommand::New();

vtkDecimatePro *deci = vtkDecimatePro::New();
deci->SetInputConnection( byu->GetOutputPort() );
deci->SetTargetReduction( 0.75 );
deci->AddObserver( vtkCommand::ProgressEvent, pobserver );
```

While this is a primitive form of interaction, it is a foundational element to many applications that use VTK. For example, the simple code above can be easily converted to display and manage a GUI progress bar. This Command/Observer subsystem is also central to the 3D widgets in VTK, which are sophisticated interaction objects for querying, manipulating and editing data and are described below.

Referring to the example above, it is important to note that events in VTK are predefined, but there is a back door for user-defined events. The class vtkCommand defines the set of enumerated events (e.g., vtkCommand::ProgressEvent in the above example) as well as a user event. The UserEvent, which is simply an integral value, is typically used as a starting offset value into a set of application user-defined events. So for example vtkCommand::UserEvent+100 may refer to a specific event outside the set of VTK defined events.

From the user's perspective, a VTK widget appears as an actor in a scene except that the user can interact with it by manipulating handles or other geometric features (the handle manipulation and geometric feature manipulation is based on the picking functionality described earlier.) The interaction with this widget is fairly intuitive: a user grabs the spherical handles and moves them, or grabs the line and moves it. Behind the scenes, however, events are emitted (e.g., InteractionEvent) and a properly programmed application can observe these events, and then take the appropriate action. For example they often trigger on the vtkCommand::InteractionEvent as follows:

```
vtkLW2Callback *myCallback = vtkLW2Callback::New();
```

```

myCallback->PolyData = seeds;      // streamlines seed points, updated on interaction
myCallback->Actor = streamline;   // streamline actor, made visible on interaction

vtkLineWidget2 *lineWidget = vtkLineWidget2::New();
lineWidget->SetInteractor(iren);
lineWidget->SetRepresentation(rep);
lineWidget->AddObserver(vtkCommand::InteractionEvent,myCallback);

```

VTK widgets are actually constructed using two objects: a subclass of `vtkInteractorObserver` and a subclass of `vtkProp`. The `vtkInteractorObserver` simply observes user interaction in the render window (i.e., mouse and keyboard events) and processes them. The subclasses of `vtkProp` (i.e., actors) are simply manipulated by the `vtkInteractorObserver`. Typically such manipulation consists of modifying the `vtkProp`'s geometry including highlighting handles, changing cursor appearance, and/or transforming data. Of course, the particulars of the widgets require that subclasses are written to control the nuances of widget behavior, and there are more than 50 different widgets currently in the system.

Summary of Libraries

VTK is a large software toolkit. Currently the system consists of approximately 1.5 million lines of code (including comments but not including automatically generated wrapper software), and approximately 1000 C++ classes. To manage the complexity of the system and reduce build and link times the system has been partitioned into dozens of subdirectories. 表 24.1 lists these subdirectories, with a brief summary describing what capabilities the library provides.

24.3 Looking Back/Looking Forward

VTK has been an enormously successful system. While the first line of code was written in 1993, at the time of this writing VTK is still growing strong and if anything the pace of development is increasing.² In this section we talk about some lessons learned and future challenges.

Managing Growth

One of the most surprising aspects to the VTK adventure has been the project's longevity. The pace of development is due to several major reasons:

- New algorithms and capabilities continue to be added. For example, the informatics subsystem (Titan, primarily developed by Sandia National Labs and Kitware) is a recent significant

²See the latest VTK code analysis at <http://www.ohloh.net/p/vtk/analyses/latest>.

Common	core VTK classes
Filtering	classes used to manage pipeline dataflow
Rendering	rendering, picking, image viewing, and interaction
VolumeRendering	volume rendering techniques
Graphics	3D geometry processing
GenericFiltering	non-linear 3D geometry processing
Imaging	imaging pipeline
Hybrid	classes requiring both graphics and imaging functionality
Widgets	sophisticated interaction
IO	VTK input and output
Infovis	information visualization
Parallel	parallel processing (controllers and communicators)
Wrapping	support for Tcl, Python, and Java wrapping
Examples	extensive, well-documented examples

表 24.1: VTK Subdirectories

addition. Additional charting and rendering classes are also being added, as well as capabilities for new scientific dataset types. Another important addition were the 3D interaction widgets. Finally, the on-going evolution of GPU-based rendering and data processing is driving new capabilities in VTK.

- The growing exposure and use of VTK is a self-perpetuating process that adds even more users and developers to the community. For example, ParaView is the most popular scientific visualization application built on VTK and is highly regarded in the high-performance computing community. 3D Slicer is a major biomedical computing platform that is largely built on VTK and received millions of dollars per year in funding.
- VTK’s development process continues to evolve. In recent years the software process tools CMake, CDash, CTest, and CPack have been integrated into the VTK build environment. More recently, the VTK code repository has moved to Git and a more sophisticated work flow. These improvements ensure that VTK remains on the leading edge of software development in the scientific computing community.

While growth is exciting, validates the creation of the software system, and bodes well for the future of VTK, it can be extremely difficult to manage well. As a result, the near term future of VTK focuses more on managing the growth of the community as well as the software. Several steps have been taken in this regard.

First, formalized management structures are being created. An Architecture Review Board has been created to guide the development of the community and technology, focusing on high-level,

strategic issues. The VTK community is also establishing a recognized team of Topic Leads to guide the technical development of particular VTK subsystems.

Next, there are plans to modularize the toolkit further, partially in response to workflow capabilities introduced by git, but also to recognize that users and developers typically want to work with small subsystems of the toolkit, and do not want to build and link against the entire package. Further, to support the growing community, it's important that contributions of new functionality and subsystems are supported, even if they are not necessarily part of the core of the toolkit. By creating a loose, modularized collection of modules it is possible to accommodate the large number of contributions on the periphery while maintaining core stability.

Technology Additions

Besides the software process, there are many technological innovations in the development pipeline.

- Co-processing is a capability where the visualization engine is integrated into the simulation code, and periodically generates data extracts for visualization. This technology greatly reduces the need to output large amounts of complete solution data.
- The data processing pipeline in VTK is still too complex. Methods are under way to simplify and refactor this subsystem.
- The ability to directly interact with data is increasingly popular with users. While VTK has a large suite of widgets, many more interaction techniques are emerging including touch-screen-based and 3D methods. Interaction will continue its development at a rapid pace.
- Computational chemistry is increasing in importance to materials designers and engineers. The ability to visualize and interact with chemistry data is being added to VTK.
- The rendering system in VTK has been criticized for being too complex, making it difficult to derive new classes or support new rendering technology. In addition, VTK does not directly support the notion of a scene graph, again something that many users have requested.
- Finally new forms of data are constantly emerging. For example, in the medical field hierarchical volumetric datasets of varying resolution (e.g., confocal microscopy with local magnification).

Open Science

Finally Kitware and more generally the VTK community are committed to Open Science. Pragmatically this is a way of saying we will promulgate open data, open publication, and open source—the features necessary to ensure that we are creating reproducible scientific systems. While VTK has long been distributed as an open source and open data system, the documentation process has been lacking. While there are decent books [Kit10, SML06] there have been a variety of ad hoc

ways to collect technical publications including new source code contributions. We are improving the situation by developing new publishing mechanisms like the *VTK Journal*³ that enable of articles consisting of documentation, source code, data, and valid test images. The journal also enables automated reviews of the code (using VTK's quality software testing process) as well as human reviews of the submission.

Lessons Learned

While VTK has been successful there are many things we didn't do right:

Design Modularity: We did a good job choosing the modularity of our classes. For example, we didn't do something as silly as creating an object per pixel, rather we created the higher-level `vtkImageClass` that under the hood treats data arrays of pixel data. However in some cases we made our classes too high level and too complex, in many instances we've had to refactor them into smaller pieces, and are continuing this process. One prime example is the data processing pipeline. Initially, the pipeline was implemented implicitly through interaction of the data and algorithm objects. We eventually realized that we had to create an explicit pipeline executive object to coordinate the interaction between data and algorithms, and to implement different data processing strategies.

Missed Key Concepts: One of our biggest regrets is not making widespread use of C++ iterators. In many cases the traversal of data in VTK is akin to the scientific programming language Fortran. The additional flexibility of iterators would have been a significant benefit to the system. For example, it is very advantageous to process a local region of data, or only data satisfying some iteration criterion.

Design Issues: Of course there is a long list of design decisions that are not optimal. We have struggled with the data execution pipeline, having gone through multiple generations each time making the design better. The rendering system too is complex and hard to derive from. Another challenge resulted from the initial conception of VTK: we saw it as a read-only visualization system for viewing data. However, current customers often want it to be capable of editing data, which requires significantly different data structures.

One of the great things about an open source system like VTK is that many of these mistakes can and will be rectified over time. We have an active, capable development community that is improving the system every day and we expect this to continue into the foreseeable future.

³<http://www.midasjournal.org/?journal=35>

Battle for Wesnoth

Richard Shimooka and David White

Programming tends to be considered a straightforward problem solving activity; a developer has a requirement and codes a solution. Beauty is often judged on the technical implementation's elegance or effectiveness; this book is replete with excellent examples. Yet beyond its immediate computing functions, code can have a profound effect on people's lives. It can inspire people to participate and create new content. Unfortunately, serious barriers exist that prevent individuals from participating in a project.

Most programming languages require significant technical expertise to utilize, which is out of reach for many. In addition, enhancing the accessibility of code is technically difficult and is not necessary for many programs. It rarely translates into neat coding scripts or clever programming solutions. Achieving accessibility requires considerable forethought in project and program design, which often runs counter-intuitive to normal programming standards. Moreover most projects rely upon an established staff of skilled professionals that are expected to operate at a reasonably high level. They do not require additional programming resources. Thus, code accessibility becomes an afterthought, if considered at all.

Our project, the Battle for Wesnoth, attempted to address this issue from its origins. The program is a turn-based fantasy strategy game, produced in an open source model based on a GPL2 license. It has been a moderate success, with over four million downloads at the time of this writing. While this is an impressive metric, we believe the real beauty of our project is the development model that allowed a band of volunteers from widely different skill levels to interact in a productive way.

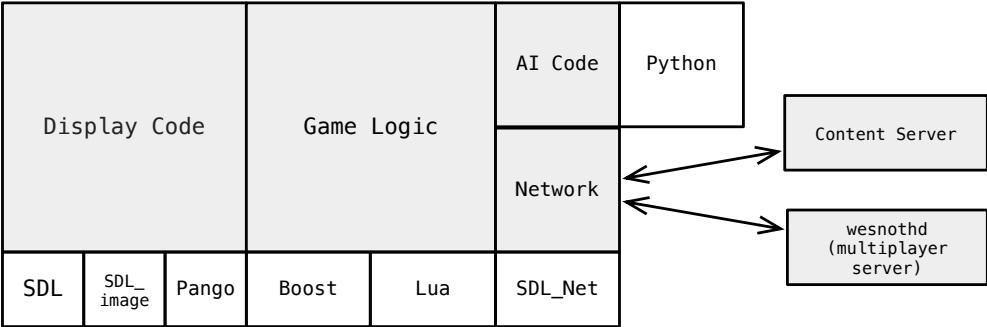
Enhancing accessibility was not a vague objective set by developers, it was viewed as essential for the project's survival. Wesnoth's open source approach meant that the project could not immediately expect large numbers of highly skilled developers. Making the project accessible to a wide a number of contributors, with varying skill levels, would ensure its long-term viability.

Our developers attempted to lay the foundations for broadening accessibility right from its earliest iteration. This would have undeniable consequences for all aspect of the programming architecture.

Major decisions were made largely with this objective in mind. This chapter will provide an in-depth examination of our program with a focus on the efforts to increase accessibility.

The first part of this chapter offers a general overview of the project’s programming, covering its language, dependencies and architecture. The second part will focus on Wesnoth’s unique data storage language, known as Wesnoth Markup Language (WML). It will explain the specific functions of WML, with a particular emphasis on its effects on in-game units. The next section covers multiplayer implementation and external programs. The chapter will end with some concluding observations on our structure and the challenges of broadening participation.

25.1 Project Overview

Wesnoth’s core engine is written in C++, totalling around 200,000 lines at the time of this publication. This represents the core game engine, approximately half of the code base without any content. The program also allows in game content to be defined in a unique data language known as Wesnoth Markup Language (WML). The game ships with another 250,000 lines of WML code. The proportion has shifted over the project’s existence. As the program matured, game content that was hardcoded in C++ has increasingly been rewritten so that WML can be used to define its operation.  FIG 25.1 gives a rough picture of the program’s architecture; green areas are maintained by Wesnoth developers, while white areas are external dependencies.

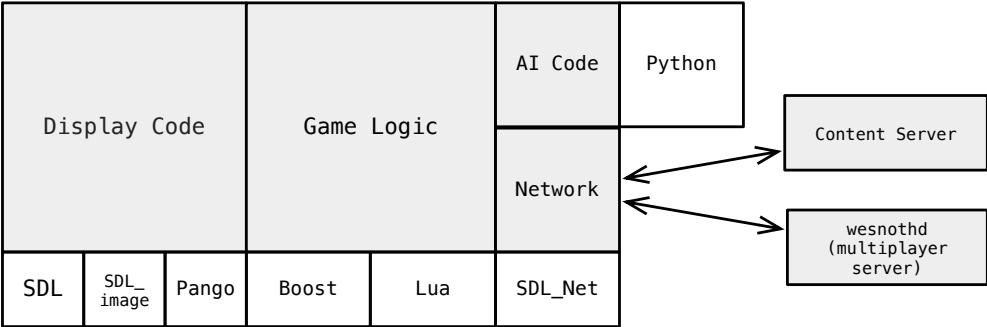


FIG 25.1: Program Architecture

Overall, the project attempts to minimize dependencies in most cases so as to maximize the portability of the application. This has the added benefit of reducing the program’s complexity, and decreases the need for developers to learn the nuances of a large number of third party APIs. At the same time, the prudent use of some dependencies can actually achieve the same effect. For example, Wesnoth uses the Simple Directmedia Layer (SDL) for video, I/O and event handling. It was chosen because it is easy to use and provides a common I/O interface across many platforms. This allows it to be portable to a wide array of platforms, rather than the alternative of coding to specific APIs on

different platforms. This comes at a price however; it is harder to take advantage of some platform specific features. SDL also has an accompanying family of libraries that are used by Wesnoth for various purposes:

- `SDL_Mixer` for audio and sound
- `SDL_Image` for loading PNG and other image formats
- `SDL_Net` for network I/O

Additionally, Wesnoth uses several other libraries:

- Boost for a variety of advanced C++ features
- Pango with Cairo for internationalized fonts
- zlib for compression
- Python and Lua for scripting support
- GNU gettext for internationalization

Throughout Wesnoth's engine, the use of WML objects—that is, string dictionaries with child nodes—is fairly ubiquitous. Many objects can be constructed from a WML node, and also serialize themselves to a WML node. Some parts of the engine keep data in this WML dictionary based format, interpreting it directly rather than parsing it into a C++ data structure.

Wesnoth utilizes several important subsystems, most of which are as self-contained as possible. This segmented structure has advantages for accessibility. An interested party can easily work a code in a specific area and introduce changes without damaging the rest of the program. The major subdivisions include:

- A WML parser with preprocessor
- Basic I/O modules that abstract underlying libraries and system calls—a video module, a sound module, a network module
- A GUI module containing widget implementations for buttons, lists, menus, etc.
- A display module for rendering the game board, units, animations, and so forth
- An AI module
- A pathfinding module that includes many utility functions for dealing with a hexagonal gaming board
- A map generation module for generating different kinds of random maps

There are also different modules for controlling different parts of the game flow:

- The titlescreen module, for controlling display of the title screen.
- The storyline module, for showing cut-scene sequences.
- The lobby module, for displaying and allowing setup of games on the multiplayer server.
- The “play game” module that controls the main gameplay.

The “play game” module and the main display module are the largest within Wesnoth. Their purpose is the least well defined, as their function is ever-changing and thus difficult to have a clear specification for. Consequently, the modules has often been in danger of suffering from the Blob anti-pattern over the program’s history—i.e., becoming huge dominant segments without well-defined behaviors. The code in the display and play game modules are regularly reviewed to see if any of it can be separated into a module of its own.

There are also ancillary features that are part of the overall project, but are separate from the main program. This includes a multiplayer server that facilitates multiplayer network games, as well as a content server that allows users to upload their content to a common server and share it with others. Both are written in C++.

25.2 Wesnoth Markup Language

As an extensible game engine, Wesnoth uses a simple data language to store and load all game data. Although XML was considered initially, we decided that we wanted something a little more friendly to non-technical users, and a little more relaxed with regard to use of visual data. We therefore developed our own data language, called Wesnoth Markup Language (WML). It was designed with the least technical of users in mind: the hope was that even users who find Python or HTML intimidating would be able to make sense of a WML file. All Wesnoth game data is stored in WML, including unit definitions, campaigns, scenarios, GUI definitions, and other game logic configuration.

WML shares the same basic attributes as XML: elements and attributes, though it doesn’t support text within elements. WML attributes are represented simply as a dictionary mapping strings to strings, with the program logic responsible for interpretation of attributes. A simple example of WML is a trimmed definition for the Elvish Fighter unit within the game:

```
[unit_type]
    id=Elvish Fighter
    name=_ "Elvish Fighter"
    race=elf
    image="units/elves-wood/fighter.png"
    profile="portraits/elves/fighter.png"
    hitpoints=33
    movement_type=woodland
    movement=5
    experience=40
    level=1
    alignment=neutral
    advances_to=Elvish Captain,Elvish Hero
    cost=14
    usage=fighter
    {LESS_NIMBLE_ELF}
```

```

[attack]
  name=sword
  description=_"sword"
  icon=attacks/sword-elven.png
  type=blade
  range=melee
  damage=5
  number=4
[/attack]
[/unit_type]

```

Since internationalization is important in Wesnoth, WML does have direct support for it: attribute values which have an underscore prefix are translatable. Any translatable string is converted using GNU gettext to the translated version of the string when the WML is parsed.

Rather than have many different WML documents, Wesnoth opts for the approach of all main game data being presented to the game engine in just a single document. This allows for a single global variable to hold the document, and when the game is loaded all unit definitions, for instance, are loaded by looking for elements with the name `unit_type` within a `units` element.

Though all data is stored in a single conceptual WML document, it would be unwieldy to have it all in a single file. Wesnoth therefore supports a preprocessor that is run over all WML before parsing. This preprocessor allows one file to include the contents of another file, or an entire directory. For instance:

```
{gui/default/window/}
```

will include all the .cfg files within `gui/default/window/`.

Since WML can become very verbose, the preprocessor also allows macros to be defined to condense things. For instance, the `{LESS_NIMBLE_ELF}` invocation in the definition of the Elvish Fighter is a call to a macro that makes certain elf units less nimble under certain conditions, such as when they are stationed in a forest:

```
#define LESS_NIMBLE_ELF
  [defense]
    forest=40
  [/defense]
#endif
```

This design has the advantage of making the engine agnostic to how the WML document is broken up into files. It is the responsibility of WML authors to decide how to structure and divide all game data into different files and directories.

When the game engine loads the WML document, it also defines some preprocessor symbols according to various game settings. For instance, a Wesnoth campaign can define different difficulty settings, with each difficulty setting resulting in a different preprocessor symbol being defined. As an example, a common way to vary difficulty is by varying the amount of resources given to an opponent (represented by gold). To facilitate this, there is a WML macro defined like this:

```

#define GOLD EASY_AMOUNT NORMAL_AMOUNT HARD_AMOUNT
#ifndef EASY
    gold={EASY_AMOUNT}
#endif
#ifndef NORMAL
    gold={NORMAL_AMOUNT}
#endif
#ifndef HARD
    gold={HARD_AMOUNT}
#endif
#endif

```

This macro can be invoked using, for instance, {GOLD 50 100 200} within the definition of an opponent to define how much gold the opponent has based on the difficulty level.

Since the WML is processed conditionally, if any of the symbols provided to the WML document change during execution of the Wesnoth engine, the entire WML document must be re-loaded and processed. For instance, when the user starts the game, the WML document is loaded and available campaigns among other things are loaded. But then, if the user chooses to start a campaign and chooses a certain difficulty level—easy for instance—then the entire document will have to be re-loaded with EASY defined.

This design is convenient in that a single document contains all game data, and that symbols can allow easy configuration of the WML document. However, as a successful project, more and more content is available for Wesnoth, including much downloadable content—all of which ends up inserted into the core document tree—which means the WML document is many megabytes in size. This has become a performance issue for Wesnoth: Loading the document may take up to a minute on some computers, causing delays in-game any time the document needs to be reloaded. Additionally, it uses a substantial amount of memory. Some measures are used to counter this: when a campaign is loaded, it has a symbol unique to that campaign defined in the preprocessor. This means that any content specific to that campaign can be #ifdefed to only be used when that campaign is needed.

Additionally, Wesnoth uses a caching system to cache the fully preprocessed version of the WML document for a given set of key definitions. Naturally this caching system must inspect the timestamp of all WML files so that if any have changed, the cached document is regenerated.

25.3 Units in Wesnoth

The protagonists of Wesnoth are its units. An Elvish Fighter and an Elvish Shaman might battle against a Troll Warrior and an Orcish Grunt. All units share the same basic behavior, but many have special abilities that alter the normal flow of gameplay. For example, a troll regenerates some of its health every turn, an Elvish shaman slows its opponents with an entangling root, and a Wose is invisible in a forest.

What is the best way to represent this in an engine? It is tempting to make a base unit class in C++, with different types of units derived from it. For instance, a `wose_unit` class could derive from `unit`, and `unit` could have a virtual function, `bool is_invisible() const`, which returns false, which the `wose_unit` overrides, returning true if the unit happens to be in forest.

Such an approach would work reasonably well for a game with a limited set of rules. Unfortunately Wesnoth is quite a large game and such an approach is not easily extendable. If a person wanted to add a new type of unit under this approach, it would require the addition of a new C++ class to the game. Additionally, it does not allow different characteristics to be combined well: what if you had a unit that regenerated, could slow enemies with a net, and was invisible in a forest? You would have to write an entirely new class that duplicates code in the other classes.

Wesnoth's unit system doesn't use inheritance at all to accomplish this task. Instead, it uses a `unit` class to represent instances of units, and a `unit_type` class, which represents the immutable characteristics that all units of a certain type share. The `unit` class has a reference to the type of object that it is. All the possible `unit_type` objects are stored in a globally held dictionary that is loaded when the main WML document is loaded.

A unit type has a list of all the abilities that that unit has. For instance, a Troll has the “regeneration” ability that makes it heal life every turn. A Saurian Skirmisher has the “skirmisher” ability that allows it to move through enemy lines. Recognition of these abilities is built into the engine—for instance, the pathfinding algorithms will check if a unit has the “skirmisher” flag set to see if it can move freely past enemy lines. This approach allows an individual to add new units, which have any combination of abilities made by the engine, by only editing WML. Of course, it doesn't allow adding completely new abilities and unit behavior without modifying the engine.

Additionally, each unit in Wesnoth may have any number of ways to attack. For instance, an Elvish Archer has a long-range bow attack and also a short-range sword attack. Each deals different damage amounts and characteristics. To represent an attack, there is an `attack_type` class, with every `unit_type` instance having a list of possible `attack_types`.

To give each unit more character, Wesnoth has a feature known as traits. Upon recruitment, most units are assigned two traits at random from a predefined list. For instance, a strong unit does more damage with its melee attacks, while an intelligent unit needs less experience before it “levels up.” Also, it is possible for units to acquire equipment during the game that make them more powerful. For instance, there might be a sword a unit can pick up that makes their attacks do more damage. To implement traits and equipment Wesnoth allows modifications on units, which are WML-defined alterations to a unit's statistics. The modification can even be applied to certain types of attacks. For instance, the strong trait gives strong units more damage when attacking in melee, but not when using a ranged strike.

Allowing completely configurable unit behavior with WML would be an admirable goal, so it is instructional to consider why Wesnoth has never achieved such a goal. WML would need to be much more flexible than it is if it were to allow arbitrary unit behavior. Rather than being a data-

oriented language, WML would have to be extended into a full-fledged programming language and that would be intimidating for many aspiring contributors.

Additionally, the Wesnoth AI, which is developed in C++, recognizes the abilities present in the game. It takes into account regeneration, invisibility, and so forth, and attempts to maneuver its units to take best advantage of these different abilities. Even if a unit ability could be created using WML, it would be difficult to make the AI sophisticated enough to recognize this ability to take advantage of it. Implementing an ability but not having it accounted for by the AI would not be a very satisfying implementation. Similarly, implementing an ability in WML and then having to modify the AI in C++ to account for the ability would be awkward. Thus, having units definable in WML, but having abilities hard-wired into the engine is considered a reasonable compromise that works best for Wesnoth's specific requirements.

25.4 Wesnoth's Multiplayer Implementation

The Wesnoth multiplayer implementation uses a simple-as-possible approach to implementing multiplayer in Wesnoth. It attempts to mitigate the possibility of malicious attacks on the server, but doesn't make a serious attempt to prevent cheating. Any movement that is made in a Wesnoth game—moving of a unit, attacking an enemy, recruiting a unit, and so forth—can be saved as a WML node. For instance, a command to move a unit might be saved into WML like this:

```
[move]
  x="11,11,10,9,8,7"
  y="6,7,7,8,8,9"
[/move]
```

This shows the path that a unit follows as a result of a player's commands. The game then has a facility to execute any such WML command given to it. This is very useful because it means that a complete replay can be saved, by storing the initial state of the game and then all subsequent commands. Being able to replay games is useful both for players to observe each other playing, as well as to help in certain kinds of bug reports.

We decided that the community would try to focus on friendly, casual games for the network multiplayer implementation of Wesnoth. Rather than fight a technical battle against anti-social crackers trying to compromise cheat prevention systems, the project would simply not try hard to prevent cheating. An analysis of other multiplayer games indicated that competitive ranking systems were a key source of anti-social behavior. Deliberately preventing such functions on the server greatly reduced the motivation for individuals to cheat. Moreover the moderators try to encourage a positive gaming community where individuals develop personal rapport with other players and play with them. This placed a greater emphasis on relationships rather than competition. The outcome of these efforts has been deemed successful, as thus far efforts to maliciously hack the game have been largely isolated.

Wesnoth's multiplayer implementation consists of a typical client-server infrastructure. A server, known as wesnothd, accepts connections from the Wesnoth client, and sends the client a summary of available games. Wesnoth will display a 'lobby' to the player who can choose to join a game or create a new game for others to join. Once players are in a game and the game starts, each instance of Wesnoth will generate WML commands describing the actions the player makes. These commands are sent to the server, and then the server relays them on to all the other clients in the game. The server will thus act as a very thin, simple relay. The replay system is used on the other clients to execute the WML commands. Since Wesnoth is a turn-based game, TCP/IP is used for all network communication.

This system also allows observers to easily watch a game. An observer can join a game in-progress, in which case the server will send the WML representing the initial state of the game, followed by a history of all commands that have been carried out since the start of the game. This allows new observers to get up to speed on the state of the game. They can see a history of the game, although it does take time for the observer to get to the game's current position—the history of commands can be fast forwarded but it still takes time. The alternative would be to have one of the clients generate a snapshot of the game's current state as WML and send it to the new observer; however this approach would burden clients with overhead based on observers, and could facilitate denial-of-service attacks by having many observers join a game.

Of course, since Wesnoth clients do not share any kind of game state with each other, only sending commands, it is important that they agree on the rules of the game. The server is segmented by version, with only players using the same version of the game able to interact. Players are immediately alerted if their client's game becomes out of sync with others. This also is a useful system to prevent cheating. Although it is rather easy for a player to cheat by modifying their client, any difference between versions will immediately be identified to players where it can be dealt with.

25.5 Conclusion

We believe that the beauty of the Battle for Wesnoth as a program is how it made coding accessible to a wide variety of individuals. To achieve this aim, the project often made compromises that do not look elegant whatsoever in the code. It should be noted that many of the project's more talented programmers frown upon WML for its inefficient syntax. Yet this compromise enabled one of the project's greatest successes. Today Wesnoth can boast of hundreds of user-made campaigns and eras, created mostly by users with little or no programming experience. Furthermore it has inspired a number of people to take up programming as a profession, using the project as a learning tool. Those are tangible accomplishments that few programs can equal.

One of the key lessons a reader should take away from Wesnoth's efforts is to consider the challenges faced by lesser skilled programmers. It requires an awareness of what blocks contributors

from actually performing coding and developing their skills. For example an individual might want to contribute to the program but does not possess any programming skills. Dedicated technological editors like `emacs` or `vim` possess a significant learning curve that might prove daunting for such an individual. Thus WML was designed to allow a simple text editor to open up its files, giving anybody the tools to contribute.

However, increasing a code base's accessibility is not a simple objective to achieve. There are no hard and fast rules for increasing code's accessibility. Rather it requires a balance between different considerations, which can have negative consequences that the community must be aware of. This is apparent in how the program dealt with dependencies. In some cases, dependencies can actually increase barriers to participation, while in others they can allow people to contribute more easily. Every issue must be considered on a case-by-case basis.

We should also be careful not to overstate some of Wesnoth's successes. The project enjoyed some advantages that are not easily replicated by other programs. Making code accessible to a wider public is partly a result of the program's setting. As an open source program, Wesnoth had several advantages in this regard. Legally the GNU license allows someone to open up an existing file, understand how it works and makes changes. Individuals are encouraged to experiment, learn and share within this culture, which might not be appropriate for other programs. Nevertheless we hope that there are certain elements that might prove useful for all developers and help them in their effort to find beauty in coding.

関連図書

- [AF94] Rick Adams and Donnalyn Frey. *!%:: A Directory of Electronic Mail Addressing & Networks*. O'Reilly Media, Sebastopol, CA, fourth edition, 1994.
- [Ald02] Gaudenz Alder. *The JGraph Swing Component*. PhD thesis, ETH Zurich, 2002.
- [BCC⁺05] Louis Bavoil, Steve Callahan, Patricia Crossno, Juliana Freire, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. VisTrails: Enabling Interactive Multiple-View Visualizations. In *Proceedings of IEEE Visualization*, pages 135–142, 2005.
- [Bro10] Frederick P. Brooks, Jr. *The Design of Design: Essays from a Computer Scientist*. Pearson Education, 2010.
- [CDG⁺06] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: a distributed storage system for structured data. In *Proc. 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06)*. USENIX Association, 2006.
- [CIRT00] P. H. Carns, W. B. Ligon III, R. B. Ross, and R. Thakur. PVFS: A Parallel File System for Linux Clusters. *Proc. 4th Annual Linux Showcase and Conference*, pages 317—–327, 2000.
- [Com79] Douglas Comer. Ubiquitous B-Tree. *ACM Computing Surveys*, 11:121–137, June 1979.
- [CRS⁺08] Brian F. Cooper, Raghu Ramakrishnan, Utkarsh Srivastava, Adam Silberstein, Philip Bohannon, Hans Arno Jacobsen, Nick Puz, Daniel Weaver, and Ramana Yerneni. PNUTS: Yahoo!'s hosted data serving platform. *PVLDB*, 1(2):1277–1288, 2008.
- [DG04] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Proc. Sixth Symposium on Operating System Design and Implementation*, 2004.

- [DHJ⁺07] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: Amazon’s Highly Available Key-Value Store. In *SOSP’07: Proceedings of Twenty-First ACM SIGOPS Symposium on Operating Systems Principles*, pages 205–220, 2007.
- [FKSS08] Juliana Freire, David Koop, Emanuele Santos, and Cláudio T. Silva. Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.
- [FSC⁺06] Juliana Freire, Cláudio T. Silva, Steve Callahan, Emanuele Santos, Carlos E. Scheidegger, and Huy T. Vo. Managing Rapidly-Evolving Scientific Workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10–18. Springer Verlag, 2006.
- [GGL03] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google File System. *Proc. ACM Symposium on Operating Systems Principles*, pages 29—43, 2003.
- [GL02] Seth Gilbert and Nancy Lynch. Brewer’s Conjecture and the Feasibility of Consistent Available Partition-Tolerant Web Services. *ACM SIGACT News*, 33(2), 2002.
- [GLPT76] Jim Gray, Raymond Lorie, Gianfranco Putzolu, and Irving Traiger. Granularity of Locks and Degrees of Consistency in a Shared Data Base. *Proc. 1st International Conference on Very Large Data Bases*, pages 365–394, 1976.
- [GR09] Adam Goucher and Tim Riley, editors. *Beautiful Testing*. O’Reilly, 2009.
- [Gra81] Jim Gray. The Transaction Concept: Virtues and Limitations. *Proc. Seventh International Conference on Very Large Data Bases*, pages 144–154, 1981.
- [Hor05] Cay Horstmann. *Object-Oriented Design and Patterns*. Wiley, 2 edition, 2005.
- [HR83] Theo Haerder and Andreas Reuter. Principles of Transaction-Oriented Database Recovery. *ACM Computing Surveys*, 15, December 1983.
- [Kit10] Kitware. *VTK User’s Guide*. Kitware, Inc., 11 edition, 2010.
- [Knu74] Donald E. Knuth. Structured Programming with go to Statements. *ACM Computing Surveys*, 6(4), 1974.
- [LA04] Chris Lattner and Vikram Adve. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proc. 2004 International Symposium on Code Generation and Optimization (CGO’04)*, Mar 2004.

- [LCWB⁺11] H. Andrés Lagar-Cavilla, Joseph A. Whitney, Roy Bryant, Philip Patchin, Michael Brudno, Eyal de Lara, Stephen M. Rumble, M. Satyanarayanan, and Adin Scannell. SnowFlock: Virtual Machine Cloning as a First-Class Cloud Primitive. *ACM Transactions on Computer Systems*, 19(1), 2011.
- [Mac06] Matt Mackall. Towards a Better SCM: Revlog and Mercurial. 2006 Ottawa Linux Symposium, 2006.
- [MQ09] Marshall Kirk McKusick and Sean Quinlan. GFS: Evolution on Fast-forward. *ACM Queue*, 7(7), 2009.
- [PGL⁺05] Anna Persson, Henrik Gustavsson, Brian Lings, Björn Lundell, Anders Mattson, and Ulf Årlig. OSS Tools in a Heterogeneous Environment for Embedded Systems Modelling: an Analysis of Adoptions of XMI. *SIGSOFT Software Engineering Notes*, 30(4), 2005.
- [PPT⁺93] Rob Pike, Dave Presotto, Ken Thompson, Howard Trickey, and Phil Winterbottom. The Use of Name Spaces in Plan 9. *Operating Systems Review*, 27(2):72–76, 1993.
- [Rad94] Sanjay Radia. Naming Policies in the Spring System. *Proc. 1st IEEE Workshop on Services in Distributed and Networked Environments*, pages 164–171, 1994.
- [RP93] Sanjay Radia and Jan Pachl. The Per-Process View of Naming and Remote Execution. *IEEE Parallel and Distributed Technology*, 1(3):71–80, 1993.
- [Shu05] Rose Shumba. Usability of Rational Rose and Visio in a Software Engineering Course. *SIGCSE Bulletin*, 37(2), 2005.
- [Shv10] Konstantin V. Shvachko. HDFS Scalability: The limits to growth. *login:*, 35(2), 2010.
- [SML06] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. Kitware, Inc., 4 edition, 2006.
- [SO92] Margo Seltzer and Michael Olson. LIBTP: Portable, Modular Transactions for UNIX. *Proc. 1992 Winter USENIX Conference*, pages 9–26, January 1992.
- [Spi03] Diomidis Spinellis. On the Declarative Specification of Models. *IEEE Software*, 20(2), 2003.
- [SVK⁺07] Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, and Cláudio T. Silva. Querying and Creating Visualizations by Analogy. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1560–1567, 2007.

- [SY91] Margo Seltzer and Ozan Yigit. A New Hashing Package for UNIX. *Proc. 1991 Winter USENIX Conference*, pages 173–184, January 1991.
- [Tan06] Audrey Tang. –Ofun: Optimizing for Fun. <http://www.slideshare.net/autang/ofun-optimizing-for-fun>, 2006.
- [Top00] Kim Topley. *Core Swing: Advanced Programming*. Prentice-Hall, 2000.

これもきっと



O'REILLY

「このツールを使えばソフトウェア開発がより改善できるよ!」「このテクノロジーを使えば…!」「このプラクティスを使えば…!」…こんな主張があふれかえっている。しかし、その中で真実はどれくらいあるだろうか?中には、単なる希望的観測に過ぎないものもあるんじゃないかな? *Making Software* は、トップレベルの研究者や実務者たちが、ソフトウェア開発の世界での実証に基づくさまざまな発見をまとめたものだ。こんな疑問に対する答えが書かれている。

- 優秀なプログラマーは凡人の 10 倍の生産性がある?
- テストファーストを採用すれば、よりよいコードをより高速に書けるようになる?
- コードメトリクスで、ソフトウェアのバグの数を予想できる?
- デザインパターンって、実際のところどうなの?
- ペアプログラミングって、どんな効果がある?
- 地理的に離れていることと、組織体系上で離れていること。どちらのほうが影響が大きい?

The Architecture of Open Source Applications と同様、*Making Software* の収益もアムネスティ・インターナショナルに寄付される。

Making Software: What Really Works, and Why We Believe It

edited by Andy Oram and Greg Wilson

O'Reilly Media, 2010, 978-0596808327

<http://oreilly.com/catalog/9780596808303>

奥付

表紙の画像は Chris Denison の *48 Free Street Mural* in Portland, Maine からのもので、撮影者は Peter Dutton である。

表紙のフォントは Caroline Hadilaksono による Junction である。テキストのフォントは T_EX Gyre Termes で見出しのフォントは T_EX Gyre Heros、どちらも Bogusław Jackowski と Janusz M. Nowacki の作品だ。コードのフォントは Raph Levien による Inconsolata である。

