



Users and Developers Guide

Version 0.7.2+svn

March 17, 2015

scalaris.googlecode.com

Copyright 2007-2013 Zuse Institute Berlin.

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Contents

1. Introduction	5
1.1. Scalaris provides strong consistency and partition tolerance	5
1.2. Scientific background	6
I. Users Guide	11
2. Download and Installation	12
2.1. Requirements	12
2.2. Download	12
2.2.1. Development Branch	12
2.2.2. Releases	12
2.3. Build	13
2.3.1. Linux	13
2.3.2. Windows	13
2.3.3. Java-API	13
2.3.4. Python-API	14
2.3.5. Ruby-API	14
2.4. Installation	14
2.5. Testing the Installation	15
3. Setting up Scalaris	16
3.1. Runtime Configuration	16
3.1.1. Logging	16
3.2. Running Scalaris	17
3.2.1. Running on a local machine	17
3.2.2. Running distributed	17
3.3. Custom startup using <code>scalarisctl</code>	18
4. Using the system	19
4.1. Application Programming Interfaces (APIs)	19
4.1.1. Supported Types	20
4.1.2. Supported Operations	21
4.1.3. JSON API	28
4.1.4. Java API	32
4.2. Command Line Interfaces	33
4.2.1. Java command line interface	33
4.2.2. Python command line interface	34
4.2.3. Ruby command line interface	34
4.3. Using Scalaris from Erlang	34
4.3.1. Running a Scalaris Cluster	35
4.3.2. Transaction	39

5. Testing the system	41
5.1. Erlang unit tests	41
5.2. Java unit tests	41
5.3. Python unit tests	42
5.4. Interoperability Tests	44
6. Troubleshooting	45
6.1. Network	45
6.2. Miscellaneous	45
 II. Developers Guide	 46
7. General Hints	47
7.1. Coding Guidelines	47
7.2. Testing Your Modifications and Extensions	47
7.3. Help with Digging into the System	47
8. System Infrastructure	48
8.1. Groups of Processes	48
8.2. The Communication Layer <code>comm</code>	48
8.3. The <code>gen_component</code>	48
8.3.1. A basic <code>gen_component</code> including a message handler	49
8.3.2. How to start a <code>gen_component</code> ?	50
8.3.3. When does a <code>gen_component</code> terminate?	51
8.3.4. How to determine whether a process is a <code>gen_component</code> ?	51
8.3.5. What happens when unexpected events / messages arrive?	51
8.3.6. What if my message handler generates an exception or crashes the process?	52
8.3.7. Changing message handlers and implementing state dependent message re- sponsiveness as a state-machine	52
8.3.8. Handling several messages atomically	52
8.3.9. Halting and pausing a <code>gen_component</code>	53
8.3.10. Integration with <code>pid_groups</code> : Redirecting messages to other <code>gen_components</code>	53
8.3.11. Replying to ping messages	53
8.3.12. The debugging interface of <code>gen_component</code> : Breakpoints and step-wise exe- cution	53
8.3.13. Future use and planned extensions for <code>gen_component</code>	56
8.4. The Process' Database (<code>pdb</code>)	56
8.5. Failure Detectors (<code>fd</code>)	56
8.6. Monitoring Statistics (<code>monitor</code> , <code>rrd</code>)	57
8.7. Writing Unit Tests	58
8.7.1. Plain Unit Tests	58
8.7.2. Randomized Testing Using <code>tester</code>	58
8.7.3. Randomized Testing Using <code>proto_sched</code>	58
9. Basic Structured Overlay	59
9.1. Ring Maintenance	59
9.2. T-Man	59
9.3. Routing Tables	59
9.3.1. The routing table process (<code>rt_loop</code>)	64
9.3.2. Simple routing table (<code>rt_simple</code>)	65

9.3.3. Chord routing table (rt_chord)	69
9.4. Local Datastore	74
9.5. Cyclon	74
9.6. Vivaldi Coordinates	74
9.7. Estimated Global Information (Gossiping)	74
9.8. Load Balancing	74
9.9. Broadcast Trees	74
10. Transactions in Scalaris	75
10.1. The Paxos Module	75
10.2. Transactions using Paxos Commit	75
10.3. Applying the Tx-Modules to replicated DHTs	75
11. How a node joins the system	76
11.1. Supervisor-tree of a Scalaris node	77
11.2. Starting the sup_dht_node supervisor and general processes of a node	78
11.3. Starting the sup_dht_node_core supervisor with a peer and some paxos processes	79
11.4. Initializing a dht_node-process	80
11.5. Actually joining the ring	80
11.5.1. A single node joining an empty ring	81
11.5.2. A single node joining an existing (non-empty) ring	81
12. How data is transferred (atomically)	90
12.1. Sending data to the predecessor	91
12.1.1. Protocol	91
12.1.2. Callbacks	91
12.2. Sending data to the successor	92
12.2.1. Protocol	92
12.2.2. Callbacks	92
13. Replica Repair	93
13.1. Replica Reconciliation - rr_recon	93
13.1.1. Trivial Replica Repair	93
13.1.2. Replica Repair with Bloom Filters	94
13.1.3. Replica Repair with Merkle Trees	95
13.2. Resolve Replicas - rr_resolve	96
13.2.1. Updating a list of keys - key_upd	96
14. Directory Structure of the Source Code	97
15. Java API	98

1. Introduction

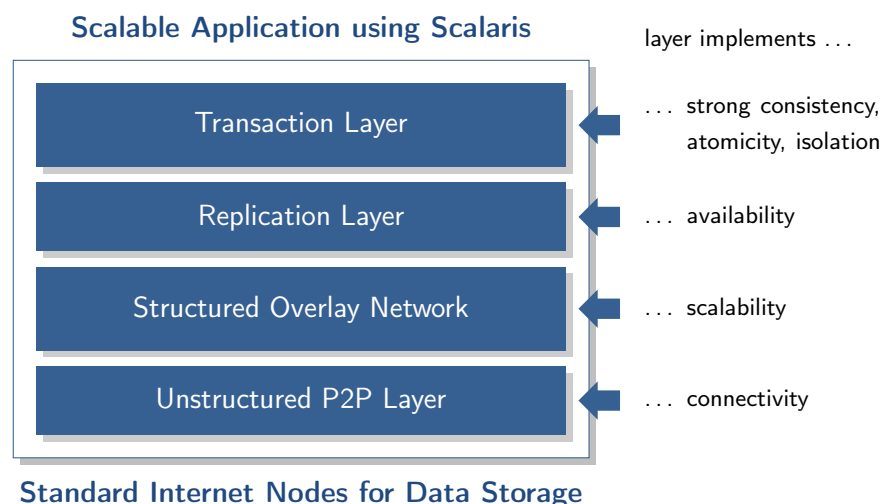
Scalaris is a scalable, transactional, distributed key-value store based on the principles of structured peer-to-peer overlay networks. It can be used as a flexible elastic data store backend to build scalable online services. Without system interruption it scales from a few PCs to thousands of servers. Servers can be added or removed on the fly without any service downtime.

Scalaris takes care of

<i>replication and fail-over</i>	for fault-tolerance
<i>self-management</i>	for low maintenance overhead
<i>automatic data partitioning</i>	for elasticity, load balancing and scalability
<i>strong consistency</i>	to ease development of applications on top of it, as inconsistencies have not to be dealt with
<i>transactions</i>	to support safe atomic updates of several data items at once

The Scalaris project was initiated and is mainly developed by [Zuse Institute Berlin \(ZIB\)](http://www.zib.de) and was partly funded by the EU projects Selfman, XtreamOS, Contrail and 4CaaS. Additional information can be found at the project homepage (<http://scalaris.googlecode.com>) and the corresponding project web page at ZIB (<http://www.zib.de/en/das/projekte/projektdetails/article/scalaris.html>).

The conceptual architecture of Scalaris consists of four layers:



1.1. Scalaris provides strong consistency and partition tolerance

In distributed computing the so called CAP theorem says that there are three desirable properties for distributed systems, but one can only have any two of them.

Strong Consistency. Any read operation has to return the result of the latest write operation on the same data item.

Availability. Items can be read and modified at any time.

Partition Tolerance. The network on which the service is running may split into several partitions which cannot communicate with each other. Later on the networks may re-join again.

For example, a service is hosted on one machine in Seattle and one machine in Berlin. This service is partition tolerant if it can tolerate that all Internet connections over the Atlantic (and Pacific) are interrupted for a few hours and then get repaired.

The goal of Scalaris is to provide strong consistency and partition tolerance. We are willing to sacrifice availability to make sure that the stored data is always consistent. I.e. when you are running Scalaris with a replication degree of four and the network splits into two partitions – one partition with three replicas and one partition with one replica – you will be able to continue to use the service only in the larger partition. All requests in the smaller partition will time out or be retried until the two networks merge again. Note, most other key-value stores tend to sacrifice consistency, which may make it hard for the application developer to detect and handle appearing inconsistencies properly.

1.2. Scientific background

Scalaris is backed by tons of research. It implements both algorithms from the literature and our own research results and combines all of them to a practical overall system. Several aspects of Scalaris were analyzed or/and developed as part of bachelor, diploma, master or PhD theses.

Scalaris in General

Publications of the Scalaris team

F. Schintke. *XtreemFS & Scalaris*. Science & Technology, pp. 54-55, 2013.

A. Reinefeld, F. Schintke, T. Schütt, S. Haridi. *A Scalable, Transactional Data Store for Future Internet Services*. Towards the Future Internet - A European Research Perspective, G. Tselentis et al. (Eds.) IOS Press, pp. 148-159, 2009.

Thorsten Schütt, Monika Moser, Stefan Plantikow, Florian Schintke, Alexander Reinefeld. *A Transactional Scalable Distributed Data Store*. 1st IEEE International Scalable Computing Challenge, co-located with CCGrid'08, 2008.

Thorsten Schütt, Florian Schintke, Alexander Reinefeld. *Scalaris: Reliable Transactional P2P Key/Value Store*. ACM SIGPLAN Erlang Workshop, 2008.

Structured Overlay Networks and Routing

The general structure of Scalaris is modelled after Chord. The Chord paper [4] describes the ring structure, the routing algorithms, and basic ring maintenance.

The main routines of our Chord node are in `src/dht_node.erl` and the join protocol is implemented in `src/dht_node_join.erl` (see also Chap. 11 on page 76). Our implementation of the routing algorithms is described in more detail in Sect. 9.3 on page 59 and the actual implementation is in `src/rt_chord.erl`. We also implemented Flexible Routing Tables according to [6] which can be found in `src/rt_frthchord.erl` and `src/rt_gfrthchord.erl`.

Publications of the Scalaris team

- Magnus Müller. *Flexible Routing Tables in a Distributed Key-Value Store*. Diploma thesis, HU-Berlin, 2013.
- Mikael Höggqvist. *Consistent Key-Based Routing in Decentralized and Reconfigurable Data Services*. Doctoral thesis, HU-Berlin, 2012.
- Philipp Borgers. *Erweiterung eines verteilten Key-Value-Stores (Riak) um einen räumlichen Index*. Bachelor thesis, FU-Berlin, 2012.
- Thorsten Schütt. *Range queries in distributed hash tables*. Doctoral thesis, 2010.
- Christian von Prollius. *Ein Peer-to-Peer System mit Bereichsabfragen in PlanetLab*. Diploma thesis, FU-Berlin, 2008.
- Jeroen Vlek. *Reducing latency: Log b routing for Chord[#]*. Bachelor thesis, Uni Amsterdam, 2008.
- Thorsten Schütt, Florian Schintke, Alexander Reinefeld. *Range Queries on structured overlay networks*. Computer Communications, 31(2), pp. 280-291, 2008.
- Thorsten Schütt, Florian Schintke, Alexander Reinefeld. *A Structured Overlay for Multi-dimensional Range Queries*. Euro-Par Conference, Luc Anne-Marie Kermarrec (Ed.)pp. 503-513, Vol.4641, LNCS, 2007.
- Alexander Reinefeld, Florian Schintke, Thorsten Schütt. *P2P Routing of Range Queries in Skewed Multidimensional Data Sets*. ZIB report ZR-07-23, 2007.
- Thorsten Schütt, Florian Schintke, Alexander Reinefeld. *Structured Overlay without Consistent Hashing*. Sixth Workshop on Global and Peer-to-Peer Computing (GP2PC'06) at Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2006), 16-19 May 2006, Singapore, p. 8, 2006.
- Thorsten Schütt, Florian Schintke, Alexander Reinefeld. *Chord[#]: Structured Overlay Network for Non-Uniform Load-Distribution*. ZIB report ZR-05-40, 2005.

Related work

- [6] Hiroya Nagao, Kazuyuki Shudo. *Flexible routing tables: Designing routing algorithms for overlays based on a total order on a routing table set*. In: Peer-to-Peer Computing, IEEE, 2011.
- P. Ganesan, B. Yang, H. Garcia-Molina. *One torus to rule them all: Multi-dimensional queries in P2P systems*. In: WebDB2004, 2004.
- Luc Onana Alima, Sameh El-Ansary, Per Brand and Seif Haridi. *DKS(N, k, f) A family of Low-Communication, Scalable and Fault-tolerant Infrastructures for P2P applications*. The 3rd International workshop on Global and P2P Computing on Large Scale Distributed Systems, (CCGRID 2003), May 2003.
- [4] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek and Hari Balakrishnan. *Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications*. ACM SIGCOMM 2001, San Deigo, CA, August 2001, pp. 149-160. http://pdos.csail.mit.edu/papers/chord:sigcomm01/chord_sigcomm.pdf

Transactions

The most interesting part is probably the transaction algorithms. The last description of the algorithms and background is in [7].

The implementation consists of the Paxos algorithm in `src/paxos` and the transaction algorithms itself in `src/transactions` (see also Chap. 10 on page 75).

Publications of the Scalaris team

[7] Florian Schintke, Alexander Reinefeld, Seif Haridi, Thorsten Schütt. *Enhanced Paxos Commit for Transactions on DHTs*. CCGRID, pp. 448-454, 2010.

Florian Schintke. *Management verteilter Daten in Grid- und Peer-to-Peer-Systemen*. Doctoral thesis, HU-Berlin, 2010.

Monika Moser, Seif Haridi, Tallat Shafaat, Thorsten Schütt, Mikael Höggqvist, Alexander Reinefeld. *Transactional DHT Algorithms*. ZIB report ZR-09-34, 2009.

Stefan Plantikow, Alexander Reinefeld, Florian Schintke. *Transactions and Concurrency Control for Peer-to-Peer-Wikis*. In: Making Grids Work, Marco Danelutto, Paraskevi Fragopoulou, Vladimir Getov (Eds.)pp. 337-349, 2008.

B. Mejías, M. Höggqvist, P. Van Roy. *Visualizing Transactional Algorithms for DHTs*. IEEE P2P Conference, 2008.

Monika Moser, Seif Haridi. *Atomic Commitment in Transactional DHTs*. Proceedings of the CoreGRID Symposium, 2007.

S. Plantikow, A. Reinefeld, F. Schintke. *Distributed Wikis on Structured Overlays*. CoreGrid Workshop on Grid Programming Models, Grid and P2P System Architecture, Grid Systems, Tools and Environments, 2007.

S. Plantikow, A. Reinefeld, F. Schintke. *Transactions for Distributed Wikis on Structured Overlays*. DSOM, Alexander Clemm, Lisandro Granville, Rolf Stadler (Eds.)pp. 256-267, Vol.4785, LNCS, 2007.

Stefan Plantikow. *Transaktionen für verteilte Wikis auf strukturierten Overlay-Netzwerken*. Diploma thesis, HU-Berlin, 2007.

Related work

Björn Kolbeck, Mikael Höggqvist, Jan Stender, Felix Hupfeld. *Please – Lease Coordination Without a Lock Server*. Intl. Parallel and Distributed Processing Symposium, pp. 978-988, 2011.

J. Gray, L. Lamport. *Consensus on transaction commit*. ACM Trans. Database Syst., 31(1):133–160, 2006.

L. Lamport. *Fast Paxos*. Distributed Computing, 19(2):79–103, 2006.

L. Lamport. *Paxos Made Simple*. SIGACT News, 32(4):51–58, December 2001.

L. Lamport. *The Part-Time Parliament*. ACM Trans. Comput. Syst., 16(2):133–169, 1998.

Ring Maintenance

We changed the ring maintenance algorithm in Scalaris. It is not the standard Chord one, but a variation of T-Man [5]. It is supposed to fix the ring structure faster. In some situations, the standard Chord algorithm is not able to fix the ring structure while T-Man can still fix it. For node sampling, our implementation relies on Cyclon [8].

The T-Man implementation can be found in `src/rm_tman.erl` and the Cyclon implementation in `src/cyclon.erl`.

Publications of the Scalaris team

Paolo Costa, Guillaume Pierre, Alexander Reinefeld, Thorsten Schütt, Maarten van Steen. *Sloppy Management of Structured P2P Services*. Proceedings of the 3rd International Workshop on Hot Topics in Autonomic Computing (HotAC III), co-located with IEEE ICAC’08, 2008.

Related work

[5] Márk Jelasity, Alberto Montresor, Ozalp Babaoglu. *T-Man: Gossip-based fast overlay topology construction*. Computer Networks (CN) 53(13):2321-2339, 2009.

[8] Spyros Voulgaris, Daniela Gavidia, Maarten van Steen. *CYCLON: Inexpensive Membership Management for Unstructured P2P Overlays*. J. Network Syst. Manage. 13(2): 2005.

Gossiping and Topology Inference

For some experiments, we implemented so called Vivaldi coordinates [2]. They can be used to estimate the network latency between arbitrary nodes.

The implementation can be found in `src/vivaldi.erl`.

For some algorithms, we use estimates of global information. These estimates are aggregated with the help of gossiping techniques [9].

The implementation can be found in `src/gossip.erl`.

Publications of the Scalaris team

Jens V. Fischer. *A Gossiping Framework for Scalaris*. Bachelor thesis, FU-Berlin, 2014.

Marie Hoffmann. *Approximate Algorithms for Distributed Systems*. Master thesis, FU-Berlin, 2012.

Thorsten Schütt, Alexander Reinefeld, Florian Schintke, Marie Hoffmann. *Gossip-based Topology Inference for Efficient Overlay Mapping on Data Centers*. Peer-to-Peer Computing, pp. 147-150, 2009.

Related work

[9] Márk Jelasity, Alberto Montresor, Ozalp Babaoglu. *Gossip-based aggregation in large dynamic networks*. ACM Trans. Comput. Syst. 23(3), 219-252 (2005).

[2] Frank Dabek, Russ Cox, Frans Kaahoe, Robert Morris. *Vivaldi: A Decentralized Network Coordinate System*. ACM SIGCOMM 2004.

Load-Balancing

Publications of the Scalaris team

Maximilian Michels. *Request-Based Load Balancing in Distributed Hash Tables*. Master thesis, FU-Berlin, 2014.

Mikael Höggqvist, Nico Kruber. *Passive/Active Load Balancing with Informed Node Placement in DHTs*. IWSOS, Thrasyvoulos Spyropoulos, Karin Hummel (Eds.)pp. 101-112, Vol.5918, Lecture Notes in Computer Science, 2009.

Nico Kruber. *DHT Load Balancing with Estimated Global Information*. Diploma thesis, HU-Berlin, 2009.

Mikael Höggqvist, Seif Haridi, Nico Kruber, Alexander Reinefeld, Thorsten Schütt. *Using Global Information for Load Balancing in DHTs*. Workshop on Decentralized Self Management for Grids, P2P, and User Communities, 2008.

Simon Rieche. *Lastbalancierung in Peer-to-Peer Systemen*. Diploma thesis, FU-Berlin, 2003.

Related work

David R. Karger, Matthias Ruhl. *Simple efficient load-balancing algorithms for peer-to-peer systems*. Theory of Computing Systems, 39(6):787–804, November 2006.

Ashwin R. Bharambe, Mukesh Agrawal, Srinivasan Seshan. *Mercury: supporting scalable multi-attribute range queries*. SIGCOMM Comput. Commun. Rev., 34(4):353–366, 2004.

Self-Management

Publications of the Scalaris team

T. Schütt, A. Reinefeld, F. Schintke, C. Hennig. *Self-Adaptation in Large-Scale Systems*. Architectures and Languages for Self-Managing Distributed Systems (SelfMan@SASO), 2009.

P. Van Roy, S. Haridi, A. Reinefeld, J.-B. Stefani, R. Yap, T. Coupaye. *Self Management for Large-Scale Distributed Systems*. Formal Methods for Components and Objects 2007 (FMCO 2007), 2008.

P. Van Roy, A. Ghodsi, S. Haridi, J.-B. Stefani, T. Coupaye, A. Reinefeld, E. Winter, R. Yap. *Self Management of Large-Scale Distributed Systems by Combining Peer-to-Peer Networks and Components*, 2005.

Other Topics

Publications of the Scalaris team

Data Placement

M. Höggqvist, S. Plantikow. *Towards Explicit Data Placement in Scalable Key/Value Stores*. Architectures and Languages for Self-Managing Distributed Systems (SelfMan@SASO), 2009.

Consistency

Tallat Shafaat, Monika Moser, Ali Ghodsi, Thorsten Schütt, Seif Haridi, Alexander Reinefeld. *Key-Based Consistency and Availability in Structured Overlay Networks*. International ICST Conference on Scalable Information Systems, 2008.

Tallat Shafaat, Monika Moser, Ali Ghodsi, Thorsten Schütt, Alexander Reinefeld. *On Consistency of Data in Structured Overlay Networks*. Coregrid Integration Workshop, 2008.

Snapshots

Stefan Keidel. *Snapshots in Scalaris*. Diploma thesis, HU-Berlin, 2012.

Replication and Replica Repair

Maik Lange. *Redundanzverwaltung in konsistenten verteilten Datenbanken*. Diploma thesis, HU-Berlin, 2012.

Part I.

Users Guide

2. Download and Installation

2.1. Requirements

For building and running Scalaris, some third-party software is required which is not included in the Scalaris sources:

- Erlang R13B01 or newer
- OpenSSL (required by Erlang's crypto module)
- GNU-like Make and autoconf (not required on Windows)

To build the Java API (and its command-line client) the following programs are also required:

- Java Development Kit 6
- Apache Ant

Before building the Java API, make sure that JAVA_HOME and ANT_HOME are set. JAVA_HOME has to point to a JDK installation, and ANT_HOME has to point to an Ant installation.

To build the Python API (and its command-line client) the following programs are also required:

- Python \geq 2.6

2.2. Download

The sources can be obtained from <http://code.google.com/p/scalaris>. RPM and DEB packages are available from <http://download.opensuse.org/repositories/home:/scalaris/> for various Linux distributions.

2.2.1. Development Branch

You find the latest development version in the svn repository:

```
# Non-members may check out a read-only working copy anonymously over HTTP.
svn checkout http://scalaris.googlecode.com/svn/trunk/ scalaris-read-only
```

2.2.2. Releases

Releases can be found under the 'Download' tab on the web-page.

2.3. Build

2.3.1. Linux

Scalaris uses autoconf for configuring the build environment and GNU Make for building the code.

```
%> ./configure
%> make
%> make docs
```

For more details read README in the main Scalaris checkout directory.

2.3.2. Windows

We are currently not supporting Scalaris on Windows. However, we have two small .bat files for building and running Scalaris nodes. It seems to work but we make no guarantees.

- Install Erlang
<http://www.erlang.org/download.html>
- Install OpenSSL (for crypto module)
<http://www.slproweb.com/products/Win32OpenSSL.html>
- Checkout Scalaris code from SVN
- adapt the path to your Erlang installation in build.bat
- start a cmd.exe
- go to the Scalaris directory
- run build.bat in the cmd window
- check that there were no errors during the compilation; warnings are fine
- go to the bin sub-directory
- adapt the path to your Erlang installation in firstnode.bat, joining_node.bat
- run firstnode.bat or one of the other start scripts in the cmd window

build.bat will generate a Emakefile if there is none yet. If you have Erlang < R13B04, you will need to adapt the Emakefile. There will be empty lines in the first three blocks ending with “}”.: add the following to these lines and try to compile again. It should work now.

```
, {d, type_forward_declarations_are_not_allowed}
, {d, forward_or_recursive_types_are_not_allowed}
```

For the most recent description please see the FAQ at <http://code.google.com/p/scalaris/wiki/FAQ>.

2.3.3. Java-API

The following commands will build the Java API for Scalaris:

```
%> make java
```

This will build `scalaris.jar`, which is the library for accessing the overlay network. Optionally, the documentation can be build:

```
%> cd java-api
%> ant doc
```

2.3.4. Python-API

The Python API for Python 2.* (at least 2.6) is located in the `python-api` directory. Files for Python 3.* can be created using `2to3` from the files in `python-api`. The following command will use `2to3` to convert the modules and place them in `python3-api`.

```
%> make python3
```

Both versions of python will compile required modules on demand when executing the scripts for the first time. However, pre-compiled modules can be created with:

```
%> make python
%> make python3
```

2.3.5. Ruby-API

The Ruby API for Ruby ≥ 1.8 is located in the `ruby-api` directory. Compilation is not necessary.

2.4. Installation

For simple tests, you do not need to install Scalaris. You can run it directly from the source directory. Note: `make install` will install Scalaris into `/usr/local` and place `scalarisctl` into `/usr/local/bin`, by default. But it is more convenient to build an RPM and install it. On openSUSE, for example, do the following:

```
export SCALARIS_SVN=http://scalaris.googlecode.com/svn/trunk
for package in main bindings; do
    mkdir -p ${package}
    cd ${package}
    svn export ${SCALARIS_SVN}/contrib/packages/${package}/checkout.sh
    ./checkout.sh
    cp * /usr/src/packages/SOURCES/
    rpmbuild -ba scalaris*.spec
    cd ..
done
```

If any additional packages are required in order to build an RPM, `rpmbuild` will print an error.

Your source and binary RPMs will be generated in `/usr/src/packages/SRPMS` and `RPMS`.

We build RPM and DEB packages for all tagged Scalaris versions as well as snapshots of svn trunk and provide them using the Open Build Service. The latest stable version is available at <http://download.opensuse.org/repositories/home:/scalaris/>. The latest svn snapshot as well as archives of previous versions are available in their respective folders below <http://download.opensuse.org/repositories/home:/scalaris/>. Packages are available for

- Fedora 16, 17,

- Mandriva 2010, 2010.1, 2011,
- openSUSE 11.4, 12.1, Factory, Tumbleweed
- SLE 10, 11, 11SP1, 11SP2,
- CentOS 5.5, 6.2,
- RHEL 5.5, 6,
- Debian 5.0, 6.0 and
- Ubuntu 10.04, 10.10, 11.04, 11.10, 12.04.

An up-to-date list of available repositories can be found at https://code.google.com/p/scalaris/wiki/FAQ#Prebuild_packages.

For those distributions which provide a recent-enough Erlang version, we build the packages using their Erlang package and recommend using the same version that came with the distribution. In this case we do not provide Erlang packages in our repository.

Exceptions are made for openSUSE-based and RHEL-based distributions as well as Debian 5.0:

- For openSUSE, we provide the package from the `devel:languages:erlang` repository.
- For RHEL-based distributions (CentOS 5, RHEL 5, RHEL 6) we included the Erlang package from the EPEL repository of RHEL 6.
- For Debian 5.0 we included the Erlang package of Ubuntu 11.04.

2.5. Testing the Installation

After installing Scalaris you can check your installation and perform some basic tests using

```
%> scalarisctl checkinstallation
```

For further details on `scalarisctl` see Section 3.3 on page 18.

3. Setting up Scalaris

Description is based on SVN revision r1810.

3.1. Runtime Configuration

Scalaris reads two configuration files from the working directory: `bin/scalaris.cfg` (mandatory) and `bin/scalaris.local.cfg` (optional). The former defines default settings and is included in the release. The latter can be created by the user to alter settings. A sample file is provided as `bin/scalaris.local.cfg.example` and needs to be altered for a distributed setup (see Section 3.2.2 on page 17). A third way to alter the configuration of Scalaris, e.g. port numbers, is to use parameters for the `scalarisctl` script (ref. Section 3.3 on page 18). The following example changes the port to 14195 and the YAWS port to 8080:

```
%> ./bin/scalarisctl -p 14194 -y 8080
```

The configuration precedence is as follows:

1. configuration parameters of `scalarisctl`
2. `bin/scalaris.local.cfg`
3. `bin/scalaris.cfg`

3.1.1. Logging

Scalaris uses the `log4erl` library (see `contrib/log4erl`) for logging status information and error messages. The log level can be configured in `bin/scalaris.cfg` for both the stdout and file logger. The default value is `warn`; only warnings, errors and severe problems are logged.

```
%% @doc Loglevel: debug < info < warn < error < fatal < none
{log_level, warn}.
{log_level_file, warn}.
```

In some cases, it might be necessary to get more complete logging information, e.g. for debugging. In Chapter 11 on page 76, we are explaining the startup process of Scalaris nodes in more detail, here the `info` level provides more detailed information.

```
%% @doc Loglevel: debug < info < warn < error < fatal < none
{log_level, info}.
{log_level_file, info}.
```


3.2. Running Scalaris

A Scalaris deployment can have a *management server* as well as *regular nodes*. The management server is optional and provides a global view on all nodes of a Scalaris deployment which contact this server, i.e. have its address specified in the `mgmt_server` configuration setting. A regular node is either the first node in a system or joins an existing system deployment.

3.2.1. Running on a local machine

Open at least two shells. In the first, inside the Scalaris directory, start the first node (`firstnode.bat` on Windows):

```
%> ./bin/firstnode.sh
```

This will start a new Scalaris deployment with a single node, including a management server. On success <http://localhost:8000> should point to the management interface page of the management server. The main page will show you the number of nodes currently in the system. A first Scalaris node should have started and the number should show 1 node. The main page will also allow you to store and retrieve key-value pairs but should not be used by applications to access Scalaris. See Section 4.1 on page 19 for application APIs.

In a second shell, you can now start a second Scalaris node. This will be a ‘regular node’:

```
%> ./bin/joining_node.sh
```

The second node will read the configuration file and use this information to contact a number of known nodes (set by the `known_hosts` configuration setting) and join the ring. It will also register itself with the management server. The number of nodes on the web page should have increased to two by now.

Optionally, a third and fourth node can be started on the same machine. In a third shell:

```
%> ./bin/joining_node.sh 2
```

In a fourth shell:

```
%> ./bin/joining_node.sh 3
```

This will add two further nodes to the deployment. The `./bin/joining_node.sh` script accepts a number as its parameter which will be added to the started node’s name, i.e. 1 will lead to a node named `node1`. The web pages at <http://localhost:8000> should show the additional nodes.

3.2.2. Running distributed

Scalaris can be installed on other machines in the same way as described in Section 2.4 on page 14. In the default configuration, nodes will look for the management server on 127.0.0.1 on port 14195. To run Scalaris distributed over several nodes, each node requires a `bin/scalaris.local.cfg` pointing to the node running the management server (if available) and containing a list of known nodes. Without a list of known nodes, a joining node will not know where to join.

In the following example, the `mgmt_server`’s location is defined as an IP address plus a TCP port and its Erlang-internal process name. If the deployment should not use a management server, replace the setting with an invalid address, e.g. ‘ null ’.

File `scalaris.local.cfg`:

```
1 % Insert the appropriate IP-addresses for your setup
2 % as comma separated integers:
3 % IP Address, Port, and label of the boot server
4 {mgmt_server, [{127,0,0,1}, 14195, mgmt_server]}.
5
6 % IP Address, Port, and label of a node which is already in the system
7 {known_hosts, [{127,0,0,1}, 14195, service_per_vm]}.
```

If you are starting the management server using `firstnode.sh`, it will listen on port 14195 and you only have to change the the IP address in the configuration file. Otherwise the other nodes will not find the management server. Calling `./bin/joining_node.sh` on a remote machine will start the node and automatically contact the configured management server.

3.3. Custom startup using `scalarisctl`

On Linux you can also use the `scalarisctl` script to start a management server and ‘regular’ nodes directly.

```
%> ./bin/scalarisctl -h
```

```
usage: scalarisctl [options] [services] <cmd>
options:
  -h            - print this help message
  -d            - daemonize
  --screen      - if daemonized, put an interactive session into screen
  -e <params>   - pass additional parameters to erl
  -n <name>     - Erlang process name (default 'node')
  -c <cookie>   - Erlang cookie to use (for distributed Erlang)
                  (default 'chocolate chip cookie')
  -p <port>     - TCP port for the Scalaris node
  -y <port>     - TCP port for the built-in webserver (YAWS)
  -k <key>      - join at the given key
  -v            - verbose
  -l <dir>      - use this logdir base directory (will create a sub-folder
                  per node)
  --dist-erl-port <port>
                  - (single) port distributed erlang listens on
  --nodes-per-vm <number>
                  - number of Scalaris nodes to start inside the VM
  -t <stype>    - select start type: first|joining|quorum|recover|nostart
services:
  -m            - global Scalaris management server
  -s            - Scalaris node (see also -f)
commands:
  checkinstallation
                  - test installation
  start         - start services (see -m and -s)
  stop          - stop a scalaris process defined by its name (see -n)
  restart       - restart a scalaris process by its name (see -n)

  list          - list locally running Erlang VMs
  debug         - connect to a running node via an Erlang shell
  dbg-check-ring <ring-size> <attempts>
                  - checks (up to) <attempts> times whether Scalaris has
                    <ring-size> nodes and the ring maintenance has settled
                    (requires a mgmt_server)
```

4. Using the system

Description is based on SVN revision r1936.

Scalaris can be used with one of the provided command line interfaces or by using one of the APIs in a custom program. The following sections will describe the APIs in general, each API in more detail and the use of our command line interfaces.

4.1. Application Programming Interfaces (APIs)

Currently we offer the following APIs:

- an *Erlang API* running on the node Scalaris is run
(functions can be called using remote connections with distributed Erlang)
- a *Java API* using Erlang's JInterface library
(connections are established using distributed Erlang)
- a generic *JSON API*
(offered by an integrated HTTP server running on each Scalaris node)
- a *Python API* for Python ≥ 2.6 using JSON to talk to Scalaris.
- a *Ruby API* for Ruby ≥ 1.8 using JSON to talk to Scalaris.

Each API contains methods for accessing functions from the three layers Scalaris is composed of. Table 4.1 shows the modules and classes of Erlang, Java, Python and Ruby and their mapping to these layers. Details about the supported operations and how to access them in each of the APIs are provided in Section 4.1.2 on page 21. A more detailed discussion about the generic JSON API including examples of JSON calls is shown in Section 4.1.3 on page 28.

	Erlang module	Java class in de.zib.scalariz	JSON file in <URL>/api/	Python / Ruby class in module scalaris
Transaction Layer	api_tx	Transaction, TransactionSingleOp	tx.yaws	Transaction, TransactionSingleOp
	api_pubsub	PubSub	pubsub.yaws	PubSub
Replication Layer	api_rdht	ReplicatedDHT	rdht.yaws	ReplicatedDHT
P2P Layer	api_dht			
	api_dht_raw		dht_raw.yaws	
	api_vm	ScalarisVM		
	api_monitor	Monitor	monitor.yaws	

Table 4.1.: Layered API structure

	Erlang	Java	JSON	Python	Ruby
boolean	<code>boolean()</code>	<code>bool</code> , <code>Boolean</code>	<code>true</code> , <code>false</code>	<code>True</code> , <code>False</code>	<code>true</code> , <code>false</code>
integer	<code>integer()</code>	<code>int</code> , <code>Integer</code> <code>long</code> , <code>Long</code> <code>BigInteger</code>	<code>int</code>	<code>int</code>	<code>Fixnum</code> , <code>Bignum</code>
float	<code>float()</code>	<code>double</code> , <code>Double</code>	<code>int frac</code> <code>int exp</code> <code>int frac exp</code>	<code>float</code>	<code>Float</code>
string	<code>string()</code>	<code>String</code>	<code>string</code>	<code>str</code>	<code>String</code>
binary	<code>binary()</code>	<code>byte[]</code>	<code>string</code> (base64-encoded)	<code>bytearray</code>	<code>String</code>
list(type)	<code>[type()]</code>	<code>List<Object></code>	<code>array</code>	<code>list</code>	<code>Array</code>
JSON	<code>json_obj()</code> *	<code>Map<String, Object></code>	<code>object</code>	<code>dict</code>	<code>Hash</code>
custom	<code>any()</code>	<code>OtpErlangObject</code>	<code>/</code>	<code>/</code>	<code>/</code>

```
*
json_obj() :: {struct, [Key::atom() | string(), Value::json_val()]}
json_val() :: string() | number() | json_obj() | {array, [any()]} | true | false | null
```

Table 4.2.: Types supported by the Sclaris APIs

4.1.1. Supported Types

Different programming languages have different types. In order for our APIs to be compatible with each other, only a subset of the available types is officially supported.

Keys are always strings. In order to avoid problems with different encodings on different systems, we suggest to only use ASCII characters.

For *values* we distinguish between *native*, *composite* and *custom* types (refer to Table 4.2 for the mapping to the language-specific types of each API).

Native types are

- boolean values
- integer numbers
- floating point numbers
- strings and
- binary objects (a number of bytes).

Composite types are

- lists of the following elements:
 - native types (*except binary objects!*),
 - composite types
- objects in JavaScript Object Notation (JSON)¹

Custom types include any Erlang term not covered by the previous types. Special care needs to be taken using custom types as they may not be accessible through every API or may be misinterpreted by an API. The use of them is discouraged.

¹see <http://json.org/>

4.1.2. Supported Operations

Most operations are available to all APIs, but some (especially convenience methods) are API- or language-specific. The following paragraphs provide a brief overview of what is available to which API. For a full reference, see the documentation of the specific API.

Transaction Layer

Read Reads the value stored at a given key using quorum read.

```
Erlang    api_tx:read(Key)
Java:     TransactionSingleOp.read(Key)
JSON:     tx.yaws/read(Key)
Python:   TransactionSingleOp.read(Key)
Ruby:     TransactionSingleOp.read(Key)
```

Write Writes a value to a given key.

```
Erlang    api_tx:write(Key, Value)
Java:     TransactionSingleOp.write(Key, Value)
JSON:     tx.yaws/write(Key, Value)
Python:   TransactionSingleOp.write(Key, Value)
Ruby:     TransactionSingleOp.write(Key, Value)
```

“Add to” & “Delete from” List Operations For the list stored at a given key, first add all elements from a given list, then remove all elements from a second given list.

```
Erlang    api_tx:add_del_on_list(Key, ToAddList, ToRemoveList)
Java:     TransactionSingleOp.addDelOnList(Key, ToAddList, ToRemoveList)
JSON:     tx.yaws/add_del_on_list(Key, ToAddList, ToRemoveList)
Python:   TransactionSingleOp.add_del_on_list(Key, ToAddList, ToRemoveList)
Ruby:     TransactionSingleOp.add_del_on_list(Key, ToAddList, ToRemoveList)
```

Add to a number Adds a given number to the number stored at a given key.

```
Erlang    api_tx:add_on_nr(Key, ToAddNumber)
Java:     TransactionSingleOp.addOnNr(Key, ToAddNumber)
JSON:     tx.yaws/add_on_nr(Key, ToAddList, ToAddNumber)
Python:   TransactionSingleOp.add_on_nr(Key, ToAddNumber)
Ruby:     TransactionSingleOp.add_on_nr(Key, ToAddNumber)
```

Atomic Test and Set Writes the given (new) value to a key if the current value is equal to the given old value.

```
Erlang    api_tx:test_and_set(Key, OldValue, NewValue)
Java:     TransactionSingleOp.testAndSet(Key, OldValue, NewValue)
JSON:     tx.yaws/add_on_nr(Key, OldValue, NewValue)
Python:   TransactionSingleOp.test_and_set(Key, OldValue, NewValue)
Ruby:     TransactionSingleOp.test_and_set(Key, OldValue, NewValue)
```

Bulk Operations Executes multiple requests, i.e. operations, where each of them will be committed.

Collecting requests and executing all of them in a single call yields better performance than executing all on their own.

```
Erlang    api_tx:req_list_commit_each(RequestList)
Java:     TransactionSingleOp.req_list(RequestList)
JSON:     tx.yaws/req_list_commit_each(RequestList)
Python:   TransactionSingleOp.req_list(RequestList)
Ruby:     TransactionSingleOp.req_list(RequestList)
```

Transaction Layer (with TLog)

Read (with TLog) Reads the value stored at a given key using quorum read as an additional part of a previous transaction or for starting a new one (*no auto-commit!*).

```
Erlang    api_tx:read(TLog, Key)
Java:     Transaction.read(Key)
JSON:     n/a - use req_list
Python:   Transaction.read(Key)
Ruby:     Transaction.read(Key)
```

Write (with TLog) Writes a value to a given key as an additional part of a previous transaction or for starting a new one (*no auto-commit!*).

```
Erlang    api_tx:write(TLog, Key, Value)
Java:     Transaction.write(Key, Value)
JSON:     n/a - use req_list
Python:   Transaction.write(Key, Value)
Ruby:     Transaction.write(Key, Value)
```

“Add to” & “Delete from” List Operations (with TLog) For the list stored at a given key, first add all elements from a given list, then remove all elements from a second given list as an additional part of a previous transaction or for starting a new one (*no auto-commit!*).

```
Erlang    api_tx:add_del_on_list(TLog, Key, ToAddList, ToRemoveList)
Java:     Transaction.addDelOnList(Key, ToAddList, ToRemoveList)
JSON:     n/a - use req_list
Python:   Transaction.add_del_on_list(Key, ToAddList, ToRemoveList)
Ruby:     Transaction.add_del_on_list(Key, ToAddList, ToRemoveList)
```

Add to a number (with TLog) Adds a given number to the number stored at a given key as an additional part of a previous transaction or for starting a new one (*no auto-commit!*).

```
Erlang    api_tx:add_on_nr(TLog, Key, ToAddNumber)
Java:     Transaction.addOnNr(Key, ToAddNumber)
JSON:     n/a - use req_list
Python:   Transaction.add_on_nr(Key, ToAddNumber)
Ruby:     Transaction.add_on_nr(Key, ToAddNumber)
```

Atomic Test and Set (with TLog) Writes the given (new) value to a key if the current value is equal to the given old value as an additional part of a previous transaction or for starting a new one (*no auto-commit!*).

```
Erlang  api_tx:test_and_set(TLog, Key, OldValue, NewValue)
Java:   Transaction.testAndSet(Key, OldValue, NewValue)
JSON:   tx.yaws/test_and_set(Key, OldValue, NewValue)
Python: Transaction.test_and_set(Key, OldValue, NewValue)
Ruby:   Transaction.test_and_set(Key, OldValue, NewValue)
```

Bulk Operations (with TLog) Executes multiple requests, i.e. operations, as an additional part of a previous transaction or for starting a new one (*no auto-commit!*). Only one commit request is allowed per call!

Collecting requests and executing all of them in a single call yields better performance than executing all on their own.

```
Erlang  api_tx:req_list(RequestList), api_tx:req_list(TLog, RequestList)
Java:   Transaction.req_list(RequestList)
JSON:   tx.yaws/req_list(RequestList), req_list(TLog, RequestList)
Python: Transaction.req_list(RequestList)
Ruby:   Transaction.req_list(RequestList)
```

Transaction Layer (Pub/Sub)

Scalaris implements a simple Publish/Subscribe system. Subscribers can subscribe URLs for some topic. If an event is published to that topic, a JSON-RPC is sent to that URL, i.e. a JSON object of the following form:

```
{
  "method": "notify",
  "params": [<topic>, <content>],
  "id"     : <number>
}
```

Publish Publishes an event under a given topic.

```
Erlang  api_pubsub:publish(Topic, Content)
Java:   PubSub.publish(Topic, Content)
JSON:   pubsub.yaws/publish(Topic, Content)
Python: PubSub.publish(Topic, Content)
Ruby:   PubSub.publish(Topic, Content)
```

Subscribe Subscribes a URL for a topic.

```
Erlang  api_pubsub:subscribe(Topic, URL)
Java:   PubSub.subscribe(Topic, URL)
JSON:   pubsub.yaws/subscribe(Topic, URL)
Python: PubSub.subscribe(Topic, URL)
Ruby:   PubSub.subscribe(Topic, URL)
```

Unsubscribe Subscribes a URL from a topic.

```
Erlang  api_pubsub:unsubscribe(Topic, URL)
Java:   PubSub.unsubscribe(Topic, URL)
JSON:   pubsub.yaws/unsubscribe(Topic, URL)
Python: PubSub.unsubscribe(Topic, URL)
Ruby:   PubSub.unsubscribe(Topic, URL)
```

Get Subscribers Gets a list of subscribed URLs for a given topic.

```
Erlang  api_pubsub:get_subscribers(Topic)
Java:   PubSub.getSubscribers(Topic)
JSON:   pubsub.yaws/get_subscribers(Topic)
Python: PubSub.get_subscribers(Topic)
Ruby:   PubSub.get_subscribers(Topic)
```

Replication Layer

Delete Tries to delete a value at a given key.

Warning: This can only be done outside the transaction layer and is thus not absolutely safe. Refer to the following thread on the mailing list: http://groups.google.com/group/scalaris/browse_thread/thread/ff1d9237e218799.

```
Erlang  api_rdht:delete(Key), api_rdht:delete(Key, Timeout)
Java:   ReplicatedDHT.delete(Key), ReplicatedDHT.delete(Key, Timeout)
JSON:   rdht.yaws/delete(Key), rdht.yaws/delete(Key, Timeout)
Python: ReplicatedDHT.delete(Key), ReplicatedDHT.delete(Key, Timeout)
Ruby:   ReplicatedDHT.delete(Key), ReplicatedDHT.delete(Key, Timeout)
```

Get Replica Keys Gets the (hashed) keys used for the replicas of a given (user) key (ref. Section [P2P Layer](#)).

```
Erlang  api_rdht:get_replica_keys(Key)
Java:   n/a
JSON:   n/a
Python: n/a
Ruby:   n/a
```

P2P Layer

Hash Key Generates the hash of a given (user) key.

```
Erlang  api_dht:hash_key(Key)
Java:   n/a
JSON:   n/a
Python: n/a
Ruby:   n/a
```

Get Replica Keys Gets the (hashed) keys used for the replicas of a given (hashed) key.

Erlang	<code>api_dht_raw:get_replica_keys(HashedKey)</code>
Java:	n/a
JSON:	n/a
Python:	n/a
Ruby:	n/a

Range Read Reads all Key-Value pairs in a given range of (hashed) keys.

Erlang	<code>api_dht_raw:range_read(StartHashedKey, EndHashedKey)</code>
Java:	n/a
JSON:	<code>dht_raw.yaws/range_read(StartHashedKey, EndHashedKey)</code>
Python:	n/a
Ruby:	n/a

P2P Layer (VM Management)

Get Scalaris Version Gets the version of Scalaris running in the requested Erlang VM.

Erlang	<code>api_vm:get_version()</code>
Java:	<code>ScalarisVM.getVersion()</code>
JSON:	n/a
Python:	n/a
Ruby:	n/a

Get Node Info Gets various information about the requested Erlang VM and the running Scalaris code, e.g. Scalaris version, erlang version, memory use, uptime.

Erlang	<code>api_vm:get_info()</code>
Java:	<code>ScalarisVM.getInfo()</code>
JSON:	n/a
Python:	n/a
Ruby:	n/a

Get Information about Different VMs Get connection info about other Erlang VMs running Scalaris nodes. Note: This info is provided by the cyclon service built into Scalaris.

Erlang	<code>api_vm:get_other_vms(MaxVMs)</code>
Java:	<code>ScalarisVM.getOtherVMs(MaxVMs)</code>
JSON:	n/a
Python:	n/a
Ruby:	n/a

Get Number of Scalaris Nodes in the VM Gets the number of Scalaris nodes running inside the Erlang VM.

Erlang	<code>api_vm:number_of_nodes()</code>
Java:	<code>ScalarisVM.getNumberOfNodes()</code>
JSON:	n/a
Python:	n/a
Ruby:	n/a

Get Scalaris Nodes Gets a list of Scalaris nodes running inside the Erlang VM.

```
Erlang    api_vm:get_nodes()
Java:     ScalarisVM.getNodes()
JSON:     n/a
Python:   n/a
Ruby:     n/a
```

Add Scalaris Nodes Starts additional Scalaris nodes inside the Erlang VM.

```
Erlang    api_vm:add_nodes(Number)
Java:     ScalarisVM.addNodes(Number)
JSON:     n/a
Python:   n/a
Ruby:     n/a
```

Shutdown Scalaris Nodes Gracefully kill some Scalaris nodes inside the Erlang VM. This will first move the data from the nodes to other nodes and then shut them down.

```
Erlang    api_vm:shutdown_node(Name),
           api_vm:shutdown_nodes(Count), api_vm:shutdown_nodes_by_name(Names)
Java:     ScalarisVM.shutdownNode(Name),
           ScalarisVM.shutdownNodes(Number), ScalarisVM.shutdownNodesByName(Names)
JSON:     n/a
Python:   n/a
Ruby:     n/a
```

Kill Scalaris Nodes Immediately kills some Scalaris nodes inside the Erlang VM.

```
Erlang    api_vm:kill_node(Name),
           api_vm:kill_nodes(Count), api_vm:kill_nodes_by_name(Names)
Java:     ScalarisVM.killNode(Name),
           ScalarisVM.killNodes(Number), ScalarisVM.killNodesByName(Names)
JSON:     n/a
Python:   n/a
Ruby:     n/a
```

Shutdown the Erlang VM Gracefully shuts down all Scalaris nodes in the Erlang VM and then exits.

```
Erlang    api_vm:shutdown_vm()
Java:     ScalarisVM.shutdownVM()
JSON:     n/a
Python:   n/a
Ruby:     n/a
```

Kill the Erlang VM Immediately kills all Scalaris nodes in the Erlang VM and then exits.

Erlang `api_vm:kill_vm()`
Java: `ScalarisVM.killVM()`
JSON: `n/a`
Python: `n/a`
Ruby: `n/a`

P2P Layer (Monitoring)

Get Node Info Gets some information about the node, e.g. Scalaris version, Erlang version, number of Scalaris nodes in the VM.

Erlang `api_monitor:get_node_info()`
Java: `Monitor.getNodeInfo()`
JSON: `monitor.yaws/get_node_info()`
Python: `n/a`
Ruby: `n/a`

Get Node Performance Gets some performance information about the node, e.g. the average latency and standard deviation of transactional operations.

Erlang `api_monitor:get_node_performance()`
Java: `Monitor.getNodePerformance()`
JSON: `monitor.yaws/get_node_performance()`
Python: `n/a`
Ruby: `n/a`

Get Service Info Gets some information about the whole Scalaris ring (may be estimated if no management server is used). Includes the overall load and the total number of nodes in the ring.

Erlang `api_monitor:get_service_info()`
Java: `Monitor.getServiceInfo()`
JSON: `monitor.yaws/get_service_info()`
Python: `n/a`
Ruby: `n/a`

Get Service Performance Gets some performance information about the whole Scalaris ring, e.g. the average latency and standard deviation of transactional operations. Both are aggregated and may be estimates.

Erlang `api_monitor:get_service_performance()`
Java: `Monitor.getServicePerformance()`
JSON: `monitor.yaws/get_service_performance()`
Python: `n/a`
Ruby: `n/a`

Convenience Methods / Classes

Connection Pool Implements a thread-safe pool of connections to Scalaris instances. Can be instantiated with a fixed maximum number of connections. Connections are either taken from a

pool of available connections or are created on demand. If finished, a connection can be put back into the pool.

Erlang	n/a
Java:	ConnectionPool
JSON:	n/a
Python:	ConnectionPool
Ruby:	n/a

Connection Policies Defines policies on how to select a node to connect to from a set of possible nodes and whether and how to automatically re-connect.

Erlang	n/a
Java:	ConnectionPolicy
JSON:	n/a
Python:	n/a
Ruby:	n/a

4.1.3. JSON API

Scalaris supports a JSON API for transactions. To minimize the necessary round trips between a client and Scalaris, it uses request lists, which contain all requests that can be done in parallel. The request list is then send to a Scalaris node with a POST message. The result contains a list of the results of the requests and - in case of a transaction - a TransLog. To add further requests to the transaction, the TransLog and another list of requests may be send to Scalaris. This process may be repeated as often as necessary. To finish the transaction, the request list can contain a 'commit' request as the last element, which triggers the validation phase of the transaction processing. Request lists are also supported for single read/write operations, i.e. every single operation is committed on its own.

The JSON-API can be accessed via the Scalaris-Web-Server running on port 8000 by default and pages under <URL>/api/. For backwards-compatibility the page <URL>/jsonrpc.yaws provides some functions otherwise provided by the different pages under <URL>/api/ but beware that this may be removed in future. Other examples include <http://localhost:8000/api/tx.yaws>. See Table 4.1 on page 19 for a mapping of the layers to the different pages. Requests are issued by sending a JSON object with header "Content-type"="application/json" to this URL. The result will then be returned as a JSON object with the same content type. The following table shows how both objects look like:

Request

```
{
  "jsonrpc": "2.0",
  "method"  : "<method>",
  "params"  : [<params>],
  "id"      : <number>
}
```

Result

```
{
  "result" : <result_object>,
  "id"     : <number>
}
```

The id in the request can be an arbitrary number which identifies the request and is returned in the result. The following operations (shown as <method>(<params>)) are currently supported (the given result is the <result_object> mentioned above):

generic, e.g. for testing - <URL>/api/*.yaws

- `nop(Value)` - no operation, result:

```
"ok"
```

single operations, e.g. `read/write` - <URL>/api/tx.yaws:

- `req_list_commit_each(<req_list_ce>)` - commit each request in the list, result:

```
{["status": "ok"] or {"status": "ok", "value": <json_value>} or  
 {"status": "fail", "reason": "timeout" or "abort" or "not_found" or  
   "not_a_list" or "not_a_number"} or  
 {"status": "fail", "reason": "key_changed", "value": <json_value>}]}
```

- `read(<key>)` - read the value at key, result:

```
{"status": "ok", "value": <json_value>} or  
{"status": "fail", "reason": "timeout" or "not_found"}
```

- `write(<key>, <json_value>)` - write value (inside `json_value`) to key, result:

```
{"status": "ok"} or  
{"status": "fail", "reason": "timeout" or "abort"}
```

- `add_del_on_list(<key>, ToAdd, ToRemove)` - adding to / removing from a list (for the list at key adds all values in the ToAdd list and then removes all values in the ToRemove list; if there is no value at key, uses an empty list - both value lists are `<value>`), result:

```
{"status": "ok"} or  
{"status": "fail", "reason": "timeout" or "abort" or "not_a_list"}
```

- `add_on_nr(<key>, <value>)` - adding to a number (adds value to the number at key - both values must be numbers), result:

```
{"status": "ok"} or  
{"status": "fail", "reason": "timeout" or "abort" or "not_a_number"}
```

- `test_and_set(<key>, OldValue, NewValue)` - atomic test-and-set (write NewValue to key if the current value is OldValue - both values are `<json_value>`), result:

```
{"status": "ok"} or  
{"status": "fail", "reason": "timeout" or "abort" or "not_found"} or  
{"status": "fail", "reason": "key_changed", "value": <json_value>}
```

transactions - <URL>/api/tx.yaws:

- `req_list(<req_list>)` - process a list of requests, result:

```
{"tlog": <tlog>,  
 "results": [{"status": "ok"} or {"status": "ok", "value": <json_value>} or  
 {"status": "fail", "reason": "timeout" or "abort" or "not_found" or  
   "not_a_list" or "not_a_number"} or  
 {"status": "fail", "reason": "key_changed", "value": <json_value>}]}
```

- `req_list(<tlog>, <req_list>)` - process a list of requests with a previous translog, result:

```
{"tlog": <tlog>,  
 "results": [{"status": "ok"} or {"status": "ok", "value": <json_value>} or
```

```

{"status": "fail", "reason": "timeout" or "abort" or "not_found" or
                             "not_a_list" or "not_a_number"} or
{"status": "fail", "reason": "key_changed", "value": <json_value>}}

```

replication layer functions - <URL>/api/rdht.yaws:

- delete(<key>) - delete the value at key, default timeout 2s, result:

```

{"ok": <number>, "results": ["ok" or "locks_set" or "undef"]} or
{"failure": "timeout", "ok": <number>, "results": ["ok" or "locks_set" or "undef"]}

```

- delete(<key>, Timeout) - delete the value at key with a timeout of Timeout Milliseconds, result:

```

{"ok": <number>, "results": ["ok" or "locks_set" or "undef"]} or
{"failure": "timeout", "ok": <number>, "results": ["ok" or "locks_set" or "undef"]}

```

raw DHT functions - <URL>/api/dht_raw.yaws:

- range_read(From, To) - read a range of (raw) keys, result:

```

{"status": "ok" or "timeout",
 "value": [{ "key": <key>, "value": <json_value>, "version": <version> }]}

```

publish/subscribe - <URL>/api/pubsub.yaws:

- publish(Topic, Content) - publish Content to Topic (<key>), result:

```

{"status": "ok"}

```

- subscribe(Topic, URL) - subscribe URL to Topic (<key>), result:

```

{"status": "ok"} or
{"status": "fail", "reason": "timeout" or "abort"}

```

- unsubscribe(Topic, URL) - unsubscribe URL from Topic (<key>), result:

```

{"status": "ok"} or
{"status": "fail", "reason": "timeout" or "abort" or "not_found"}

```

- get_subscribers(Topic) - get subscribers of Topic (<key>), result:

```

[<urls>]

```

monitor - <URL>/api/monitor.yaws:

- get_node_info() - gets some information about the node, result:

```

{"status": "ok" or "timeout",
 "value": [{"scalaris_version": <version_string>,
            "erlang_version": <version_string>,
            "dht_nodes": <number>}]}

```

- get_node_performance() - gets some performance information about the node, result:

```

{"status": "ok" or "timeout",
 "value": [{"latency_avg": <perf_data>, "latency_stddev": <perf_data>}]}

```

- `get_service_info()` - gets some information about the Scalaris ring, result:

```
{ "status": "ok" or "timeout",
  "value": [ { "total_load": <number>, "nodes": <number> } ] }
```

- `get_service_performance()` - gets some performance information about the Scalaris ring, result:

```
{ "status": "ok" or "timeout",
  "value": [ { "latency_avg": <perf_data>, "latency_stddev": <perf_data> } ] }
```

Note:

```
<json_value> = { "type": "as_is" or "as_bin", "value": <value> }
<operation> = { "read": <key> } or { "write", { <key>: <json_value> } } or
              { "add_del_on_list": { "key": <key>, "add": [ <value> ], "del": [ <value> ] } } or
              { "add_on_nr": { <key>: <value> } } or
              { "test_and_set": { "key": <key>, "old": <json_value>, "new": <json_value> } }
<req_list_ce> = [ <operation> ]
<req_list> = [ <operation> or { "commit", _ } ]
<perf_data> = { <number>: <perf_val>, ... }
```

The <value> inside <json_value> is either a base64-encoded string representing a binary object (type = "as_bin") or the value itself (type = "as_is").

JSON-Example

The following example illustrates the message flow:

Client

Make a transaction, that sets two keys →

```
{ "jsonrpc": "2.0",
  "method": "req_list",
  "params": [
    [ { "write": { "keyA": { "type": "as_is", "value": "valueA" } } },
      { "write": { "keyB": { "type": "as_is", "value": "valueB" } } },
      { "commit": "" } ]
  ],
  "id": 0
}
```

Scalaris node

←

Scalaris sends results back

```
{ "error": null,
  "result": {
    "results": [ { "status": "ok" }, { "status": "ok" }, { "status": "ok" } ],
    "tlog": <TLOG> // this is the translog for further operations!
  },
  "id": 0
}
```

In a second transaction: Read the two keys →

```
{ "jsonrpc": "2.0",
  "method": "req_list",
  "params": [
    [ { "read": "keyA" },
      { "read": "keyB" } ]
  ],
  "id": 0
}
```

←

Scalaris sends results back

```
{
  "error": null,
  "result": {
    "results": [
      { "status": "ok", "value": { "type": "as_is", "value": "valueA" } },
      { "status": "ok", "value": { "type": "as_is", "value": "valueB" } }
    ],
    "tlog": <TLOG>
  },
  "id": 0
}
```

Calculate something with the read values →
 and make further requests, here a write
 and the commit for the whole transaction. Also include the latest translog we

```
{
  "jsonrpc": "2.0",
  "method": "req_list",
  "params": [
    <TLOG>,
    [ { "write": { "keyA": { "type": "as_is", "value": "valueA2" } } },
      { "commit": "" } ]
  ],
  "id": 0
}
```

←

Scalaris sends results back

```
{
  "error": null,
  "result": {
    "results": [ { "status": "ok" }, { "status": "ok" } ],
    "tlog": <TLOG>
  },
  "id": 0
}
```

Examples of how to use the JSON API are the Python and Ruby API which use JSON to communicate with Scalaris.

4.1.4. Java API

The `scalaris.jar` provides a Java command line client as well as a library for Java programs to access Scalaris. The library provides several classes:

- `TransactionSingleOp` provides methods for reading and writing values.
- `Transaction` provides methods for reading and writing values in transactions.
- `PubSub` provides methods for a simple topic-based pub/sub implementation on top of Scalaris.
- `ReplicatedDHT` provides low-level methods for accessing the replicated DHT of Scalaris.

For details regarding the API we refer the reader to the Javadoc:

```
%> cd java-api
%> ant doc
%> firefox doc/index.html
```


4.2. Command Line Interfaces

4.2.1. Java command line interface

As mentioned above, the `scalaris.jar` file contains a small command line interface client. For convenience, we provide a wrapper script called `scalaris` which sets up the Java environment:

```
%> ./java-api/scalaris --noconfig --help
```

```
../java-api/scalaris [script options] [options]
Script Options:
  --help, -h          print this message and scalaris help
  --noconfig          suppress sourcing of config files in $HOME/.scalaris/
                      and ${prefix}/etc/scalaris/
  --execdebug        print scalaris exec line generated by this
                      launch script
  --noerl            do not ask erlang for its (local) host name

usage: scalaris [Options]
  -h,--help          print this message
  -v,--verbose       print verbose information,
                      e.g. the properties read
                      gets the local host's name as
                      known to Java (for debugging
                      purposes)
  -lh,--localhost    run selected mini
                      benchmark(s) [1|...|18|all]
                      (default: all benchmarks, 500
                      operations, 10 threads per
                      Scalaris node)
  -b,--minibench <[ops]> <[tpn]> <[benchs]>
                      print monitoring information
                      read an item
                      write an item
                      atomic test and set, i.e.
                      write <key> to <new> if the
                      current value is <old>
                      delete an item (default
                      timeout: 2000ms)
                      WARNING: This function can
                      lead to inconsistent data
                      (e.g. deleted items can
                      re-appear). Also when
                      re-creating an item the
                      version before the delete can
                      re-appear.
  -m,--monitor <node>
  -r,--read <key>
  -w,--write <key> <value>
  --test-and-set <key> <old> <new>
  -d,--delete <key> <[timeout]>
                      publish a new message for the
                      given topic
                      subscribe to a topic
                      unsubscribe from a topic
                      get subscribers of a topic
                      starts a service exposing
                      Scalaris monitoring values
                      via JMX
  -p,--publish <topic> <message>
  -s,--subscribe <topic> <url>
  -u,--unsubscribe <topic> <url>
  -g,--getsubscribers <topic>
  -jmx,--jmxservice <node>
```

read, write, delete and similar operations can be used to read, write and delete from/to the overlay, respectively. `getsubscribers`, `publish`, and `subscribe` are the PubSub functions. The others provide debugging and testing functionality.

```
%> ./java-api/scalaris -write foo bar
write(foo, bar)
%> ./java-api/scalaris -read foo
read(foo) == bar
```

Per default, the scalaris script tries to connect to a management server at localhost. You can change the node it connects to (and further connection properties) by adapting the values defined in `java-api/scalaris.properties`.

4.2.2. Python command line interface

```
%> ./python-api/scalaris --help
```

```
usage: ../python-api/scalaris_client.py [Options]
-r,--read <key>
    read an item
-w,--write <key> <value>
    write an item
--test-and-set <key> <old_value> <new_value>
    atomic test and set, i.e. write <key> to
    <new_value> if the current value is <old_value>
-d,--delete <key> [<timeout>]
    delete an item (default timeout: 2000ms)
    WARNING: This function can lead to inconsistent
    data (e.g. deleted items can re-appear).
    Also if an item is re-created, the version
    before the delete can re-appear.
-p,--publish <topic> <message>
    publish a new message for the given topic
-s,--subscribe <topic> <url>
    subscribe to a topic
-g,--getsubscribers <topic>
    get subscribers of a topic
-u,--unsubscribe <topic> <url>
    unsubscribe from a topic
-h,--help
    print this message
-b,--minibench [<ops> [<threads_per_node> [<benchmarks>]]]
    run selected mini benchmark(s)
    [1|...|9|all] (default: all benchmarks, 500
    operations each, 10 threads per Scalaris node)
```

4.2.3. Ruby command line interface

```
%> ./ruby-api/scalaris --help
```

```
Usage: scalaris_client [options]
-r, --read KEY          read key KEY
-w, --write KEY,VALUE   write key KEY to VALUE
--test-and-set KEY,OLDVALUE,NEWVALUE
    write key KEY to NEWVALUE if the current value is OLDVALUE
--add-del-on-list KEY,TOADD,TOREMOVE
    add and remove elements from the value at key KEY
--add-on-nr KEY,VALUE   add VALUE to the value at key KEY
-h, --help             Show this message
```

4.3. Using Scalaris from Erlang

In this section, we will describe how to use Scalaris with two small examples. After having build Scalaris as described in 2, Scalaris can be run from the source directory directly.

4.3.1. Running a Scalaris Cluster

In this example, we will set up a simple Scalaris cluster consisting of up to five nodes running on a single computer.

Adapt the configuration. The first step is to adapt the configuration to your needs. We use the sample local configuration from 3.1, copy it to `bin/scalaris.local.cfg` and add a number of different known hosts. Note that the management server will run on the same port as the first node started in the example, hence we adapt its port as well.

```
{listen_ip, {127,0,0,1}}.  
{mgmt_server, {{127,0,0,1},14195,mgmt_server}}.  
{known_hosts, [{127,0,0,1},14195, service_per_vm},  
                {{127,0,0,1},14196, service_per_vm},  
                {{127,0,0,1},14197, service_per_vm},  
                {{127,0,0,1},14198, service_per_vm}  
                % Although we will be using 5 nodes later, only 4 are added as known nodes.  
                ]}.  
}]}
```

Bootstrapping. In a shell (from now on called S1), start the first node ("premier"):

```
./bin/scalarisctl -m -n premier@127.0.0.1 -p 14195 -y 8000 -s -f start
```

The `-m` and `-f` options instruct `scalarisctl` to start the management server and the first_node (see Section 3.3 on page 18 for further details on `scalarisctl`). Note that the command above will produce some output about unknown nodes. This is expected, as some nodes defined in the configuration file above are not started yet.

After you run the above command and no further error occurred, you can query the locally available nodes using `scalarisctl`. Enter into a new shell (called MS):

```
./bin/scalarisctl list  
epmd: up and running on port 4369 with data:  
name premier at port 47235
```

Scalaris also contains a webserver. You can access it by pointing your browser to <http://127.0.0.1:8000> (or the respective IP address of the node). With the above example, you can see the first node ("premier") and its management role.

Adding Nodes. We will now add four additional nodes to the cluster. Use a new shell (S2 to S5) for each of the following commands. Each newly added node is a "real" Scalaris node and could run on another physical computer than the other nodes.

```
./bin/scalarisctl -n second@127.0.0.1 -p 14196 -y 8001 -s start  
./bin/scalarisctl -n n3@127.0.0.1 -p 14197 -y 8002 -s start  
./bin/scalarisctl -n n4@127.0.0.1 -p 14198 -y 8003 -s start  
./bin/scalarisctl -n n5@127.0.0.1 -p 14199 -y 8004 -s start
```

Note that the last added nodes should not report a node as not reachable.

The management server should now report that the nodes have indeed joined Scalaris successfully. Query `scalarisctl`:

```
./bin/scalarisctl list
epmd: up and running on port 4369 with data:
name n5 at port 47801
name n4 at port 54614
name n3 at port 41710
name second at port 44329
name premier at port 44862
```

The actual output might differ, as the port numbers are assigned by the operating system.

Each node offers a web console. Point your browser to any url for <http://127.0.0.1:8001> to <http://127.0.0.1:8004>. Observe that all nodes claim the cluster ring to consist of 5 nodes.

The web interface of node premier differs from the other interfaces. This is due to the fact that the management server is running on this node, adding additional information to the web interface.

Entering Data Using the Web Interface. A node's web interface can be used to query and enter data into Scalaris. To try this, point your browser to <http://127.0.0.1:8000> (or any of the other nodes) and use the provided HTML form.

1. Lookup key hello. This will return `{fail,not_found}`
2. Add new keys k1 and k2 with values v1 and v2, respectively. Then, lookup that key on the current and one of the other nodes. This should return `{ok,"v1"}` and `{ok, "v2"}` on both nodes.
3. Update the key k1 by adding it on any node with value v1updated.
4. Update the key k2 by adding it on any node with value v2updated. Lookup the key again and you should receive `{ok, v2updated}`

Simulating Node Failure. To simulate a node failure, we will simply stop n4 using `scalarisctl`:

```
./bin/scalarisctl -n n4@127.0.0.1 stop
```

Other nodes will notice the crash of n4. By querying the available nodes in the shell MS again, you will now see only 4 nodes.

Although the node n4 left the system, the data in the system is still consistent. Try to query the keys you added above. You should receive the values for each.

We will start a new node with the name n4 again:

```
./bin/scalarisctl -n n4@127.0.0.1 -p 14198 -y 8003 -s start
```

The node list (again, query `scalarisctl` in shell MS) will report n4 as alive again. You can still lookup the keys from above and should also receive the same result for the queries.

After running the above, we went from a five-node cluster to a 4-node cluster and back to a five-node cluster without any data loss due to a leaving node. The system was not unavailable for users and would have served any user requests without violating the data consistency or availability.

Controlling Scalaris Using the Erlang Shell. The calls to `scalarisctl` above which started a new Scalaris node ended within an Erlang shell. Each of those shells can be used to control a local Scalaris node and issue queries to the distributed database. Enter shell S1 and hit <return> to see the Erlang shell prompt. Now, enter the following commands and check that the output is similar

to the one provided here. You can stop the Erlang shell using `quit()`., which then also stops the corresponding `Scalaris` node.

```
(premier@127.0.0.1)1> api_tx:read("k0").
{fail,not_found}
(premier@127.0.0.1)2> api_tx:read("k1").
{ok,"v1updated"}
(premier@127.0.0.1)3> api_tx:read("k2").
{ok,"v2updated"}
(premier@127.0.0.1)4> api_tx:read(<<"k1">>).
{ok,"v1updated"}
(premier@127.0.0.1)5> api_tx:read(<<"k2">>).
{ok,"v2updated"}
(premier@127.0.0.1)6> api_tx:write(<<"k3">>,<<"v3">>).
{ok}
(premier@127.0.0.1)7> api_tx:read(<<"k3">>).
{ok,<<"v3">>}
(premier@127.0.0.1)8> api_tx:read("k3").
{ok,<<"v3">>}
(premier@127.0.0.1)9> api_tx:write(<<"k4">>,{1,2,3,four}).
{ok}
(premier@127.0.0.1)10> api_tx:read("k4").
{ok,{1,2,3,four}}
```

Attaching a Client to `Scalaris`. Now we will connect a true client to our 5 nodes `Scalaris` cluster. This client will not be a `Scalaris` node itself and thus represents a user application interacting with `Scalaris`.

We use a new shell to run an Erlang shell to do remote API calls to the server nodes.

```
erl -name client@127.0.0.1 -hidden -setcookie 'chocolate chip cookie'
```

The requests to `Scalaris` will be done using `rpc:call/4`. A production system would have some more sophisticated client side module, dispatching requests automatically to server nodes, for example.

```
(client@127.0.0.1)1> net_adm:ping('n3@127.0.0.1').
pong
(client@127.0.0.1)2> rpc:call('n3@127.0.0.1', api_tx, read, [<<"k0">>]).
{fail,not_found}
(client@127.0.0.1)3> rpc:call('n3@127.0.0.1', api_tx, read, [<<"k4">>]).
{ok,{1,2,3,four}}
(client@127.0.0.1)4> rpc:call('n4@127.0.0.1', api_tx, read, [<<"k4">>]).
{ok,{1,2,3,four}}
(client@127.0.0.1)5> rpc:call('n5@127.0.0.1', api_tx, write, [<<"num5">>,55]).
{ok}
(client@127.0.0.1)6> rpc:call('n3@127.0.0.1', api_tx, read, [<<"num5">>]).
{ok,55}
(client@127.0.0.1)7> rpc:call('n2@127.0.0.1', api_tx, add_on_nr, [<<"num5">>,2]).
{badrpc,nodedown}
(client@127.0.0.1)8> rpc:call('second@127.0.0.1', api_tx, add_on_nr, [<<"num5">>,2]).
{ok}
(client@127.0.0.1)9> rpc:call('n3@127.0.0.1', api_tx, read, [<<"num5">>]).
{ok,57}
(client@127.0.0.1)10> rpc:call('n4@127.0.0.1', api_tx, test_and_set, [<<"num5">>,57,59]).
{ok}
(client@127.0.0.1)11> rpc:call('n5@127.0.0.1', api_tx, read, [<<"num5">>]).
{ok,59}
(client@127.0.0.1)12> rpc:call('n4@127.0.0.1', api_tx, test_and_set, [<<"num5">>,57,55]).
{fail,{key_changed,59}}
(client@127.0.0.1)13> rpc:call('n3@127.0.0.1', api_tx, read, [<<"num5">>]).
{ok,59}
(client@127.0.0.1)14> rpc:call('n5@127.0.0.1', api_tx, test_and_set,
[<<"k2">>,"v2updated",<<"v2updatedTWICE">>]).
{ok}
```

```
(client@127.0.0.1)15> rpc:call('n4@127.0.0.1', api_tx, read, [<<"k2">>]).
{ok,<<"v2updatedTWICE">>}}
(client@127.0.0.1)16> rpc:call('n3@127.0.0.1', api_tx, add_on_nr, [<<"num5">>,-4]).
{ok}
(client@127.0.0.1)17> rpc:call('n4@127.0.0.1', api_tx, read, [<<"num5">>]).
{ok,55}
(client@127.0.0.1)18> q().
ok
```

To show that the above calls actually worked with Scalaris, connect another client to the cluster and read updates made by the first:

```
erl -name clientagain@127.0.0.1 -hidden -setcookie 'chocolate chip cookie'
```

```
(clientagain@127.0.0.1)1> net_adm:ping('n5@127.0.0.1').
pong
(clientagain@127.0.0.1)2> rpc:call('n4@127.0.0.1', api_tx, read, [<<"k0">>]).
{fail,not_found}
(clientagain@127.0.0.1)3> rpc:call('n4@127.0.0.1', api_tx, read, [<<"k1">>]).
{ok,"v1updated"}
(clientagain@127.0.0.1)4> rpc:call('n3@127.0.0.1', api_tx, read, [<<"k2">>]).
{ok,<<"v2updatedTWICE">>}}
(clientagain@127.0.0.1)5> rpc:call('second@127.0.0.1', api_tx, read, [<<"num5">>]).
{ok,55}
```

Shutting Down Scalaris. Firstly, we list the available nodes using `scalarisctl` using the shell MS.

```
./bin/scalarisctl list
epmd: up and running on port 4369 with data:
name n4 at port 52504
name n5 at port 47801
name n3 at port 41710
name second at port 44329
name premier at port 44862
```

Secondly, we shut down each of the nodes:

```
./bin/scalarisctl -n second@127.0.0.1 stop
'second@127.0.0.1'
./bin/scalarisctl -n n3@127.0.0.1 stop
'n3@127.0.0.1'
./bin/scalarisctl -n n4@127.0.0.1 stop
'n4@127.0.0.1'
./bin/scalarisctl -n n5@127.0.0.1 stop
'n5@127.0.0.1'
```

Only the first node remains:

```
./bin/scalarisctl list
epmd: up and running on port 4369 with data:
name premier at port 44862

./bin/scalarisctl -n premier@127.0.0.1 stop
'premier@127.0.0.1'
./bin/scalarisctl list
epmd: up and running on port 4369 with data:
(nothing)
```

The Scalaris API offers more transactional operations than just single-key read and write. The next part of this section will describe how to build transaction logs for atomic operations and how

Scalaris handles conflicts in concurrently running transactions. See the module `api_tx` for more functions to access the data layer of Scalaris.

4.3.2. Transaction

In this section, we will describe how to build transactions using `api_tx:req_list(Tlog, List)` on the client side.

The setup is similar to the five nodes cluster in the previous section. To simplify the example all API calls are typed inside the Erlang shells of nodes `n4` and `n5`.

Consider two concurrent transactions A and B. A is a long-running operation, whereas B is only a short transaction. In the example, A starts before B and B ends before A. B is "timely" nested in A and disturbs A.

Single Read Operations. We first issue two read operations on nodes `n4`, `n5` to see that we are working on the same state for key `k1`:

```
(n4@127.0.0.1)10> api_tx:read(<<"k1">>).  
{ok,<<"v1">>}  
(n5@127.0.0.1)17> api_tx:read(<<"k1">>).  
{ok,<<"v1">>}
```

Create Transaction Logs and Add Operations. Now, we create two transaction logs for the transactions and add the operations which are to be run atomically. A will be created on node `n5`, B on `n4`:

```
(n5@127.0.0.1)18> T5longA0 = api_tx:new_tlog().  
[]  
(n5@127.0.0.1)19> {T5longA1, R5longA1} = api_tx:req_list(T5longA0, [{read, <<"k1">>}]).  
[{[76,<<"k1">>,1,75,'$empty'],[{ok,<<"v1">>}]}]  
(n4@127.0.0.1)11> T4shortB0 = api_tx:new_tlog().  
[]  
(n4@127.0.0.1)12> {T4shortB1, R4shortB1} = api_tx:req_list(T4shortB0, [{read, <<"k1">>}]).  
[{[76,<<"k1">>,1,75,'$empty'],[{ok,<<"v1">>}]}]  
(n4@127.0.0.1)13> {T4shortB2, R4shortB2} = api_tx:req_list(T4shortB1,  
                                                         [{write, <<"k1">>, <<"v1Bshort">>}]).  
[{[77,<<"k1">>,1,75, <<131,109,0,0,0,8,118,49,66,115,104,111,114,116>>}],  
 [ok]}]  
(n4@127.0.0.1)14> {T4shortB3, R4shortB3} = api_tx:req_list(T4shortB2, [{read, <<"k1">>}]).  
[{[77,<<"k1">>,1,75, <<131,109,0,0,0,8,118,49,66,115,104,111,114,116>>}],  
 [ok,<<"v1Bshort">>}]}]
```

To finish the transaction log for B, we add `{commit}`. This operation should return an `ok`:

```
(n4@127.0.0.1)15> {T4shortB4, R4shortB4} = api_tx:req_list(T4shortB3, [{commit}]).  
[[],[{ok}]]  
(n4@127.0.0.1)16> [R4shortB1,R4shortB2,R4shortB3,R4shortB4].  
[{[ok,<<"v1">>}],[{ok}],[{ok,<<"v1Bshort">>}],[{ok}]]
```

This concludes the creation of B. Now we will try to commit the long running transaction A after reading the key `k1` again. This and further attempts to write the key will fail, as the transaction B wrote this key since A started.

```
(n5@127.0.0.1)20> {T5longA2, R5longA2} = api_tx:req_list(T5longA1, [{read, <<"k1">>}]).  
[{[76,<<"k1">>,2,{fail,abort},' $empty'],  
 [ok,<<"v1Bshort">>}]} % <-- SEE #### FAIL and ABORT ####
```

```

(n5@127.0.0.1)21> {T5longA3, R5longA3} = api_tx:req_list(T5longA2,
[{write, <<"k1">>, <<"v1Along">>}]).
[{76, <<"k1">>, 2, {fail, abort}, '$empty'}, [{ok}]]
(n5@127.0.0.1)22> {T5longA4, R5longA4} = api_tx:req_list(T5longA3, [{read, <<"k1">>}]).
[{76, <<"k1">>, 2, {fail, abort}, '$empty'},
[{ok, <<"v1Bshort">>}]]
(n5@127.0.0.1)23> {T5longA5, R5longA5} = api_tx:req_list(T5longA4, [{commit}]).
[[], [{fail, abort, [<<"k1">>]}]] % <-- SEE #### FAIL and ABORT ####
(n4@127.0.0.1)17> api_tx:read(<<"k1">>).
{ok, <<"v1Bshort">>}
(n5@127.0.0.1)24> api_tx:read(<<"k1">>).
{ok, <<"v1Bshort">>}

```

As expected, the first coherent commit B constructed on n4 has won.

Note that in a real system, operations in `api_tx:req_list(Tlog, List)` should be grouped together with a trailing `{commit}` as far as possible. The individual separation of all reads, writes and commits was done here on purpose to study the transactional behaviour.

5. Testing the system

Description is based on SVN revision r1618.

5.1. Erlang unit tests

There are some unit tests in the `test` directory which test `Scalaris` itself (the Erlang code). You can call them by running `make test` in the main directory. The results are stored in a local `index.html` file.

The tests are implemented with the `common-test` package from the Erlang system. For running the tests we rely on `run_test`, which is part of the `common-test` package, but (on `erlang < R14`) is not installed by default. `configure` will check whether `run_test` is available. If it is not installed, it will show a warning and a short description of how to install the missing file.

Note: for the unit tests, we are setting up and shutting down several overlay networks. During the shut down phase, the runtime environment will print extensive error messages. These error messages do not indicate that tests failed! Running the complete test suite takes about 10-20 minutes, depending on your machine.

If the test suite is interrupted before finishing, the results may not have been linked into the `index.html` file. They are however stored in the `ct_run.ct@...` directory.

5.2. Java unit tests

The Java unit tests can be run by executing `make java-test` in the main directory. This will start a `Scalaris` node with the default ports and test all Java functions part of the Java API. A typical run will look like the following:

```
%> make java-test
[...]
tools.test:
[junit] Running de.zib.tools.PropertyLoaderTest
[junit] Testsuite: de.zib.tools.PropertyLoaderTest
[junit] Tests run: 3, Failures: 0, Errors: 0, Time elapsed: 0.113 sec
[junit] Tests run: 3, Failures: 0, Errors: 0, Time elapsed: 0.113 sec
[junit]
[junit] ----- Standard Output -----
[junit] Working Directory = <scalarisdir>/java-api/classes
[junit] -----
[...]
scalaris.test:
[junit] Running de.zib.sclaris.ConnectionTest
[junit] Testsuite: de.zib.sclaris.ConnectionTest
[junit] Tests run: 7, Failures: 0, Errors: 0, Time elapsed: 0.366 sec
[junit] Tests run: 7, Failures: 0, Errors: 0, Time elapsed: 0.366 sec
[junit]
[junit] Running de.zib.sclaris.DefaultConnectionPolicyTest
[junit] Testsuite: de.zib.sclaris.DefaultConnectionPolicyTest
[junit] Tests run: 12, Failures: 0, Errors: 0, Time elapsed: 0.314 sec
```

```

[junit] Tests run: 12, Failures: 0, Errors: 0, Time elapsed: 0.314 sec
[junit]
[junit] Running de.zib.scalarish.PeerNodeTest
[junit] Testsuite: de.zib.scalarish.PeerNodeTest
[junit] Tests run: 5, Failures: 0, Errors: 0, Time elapsed: 0.077 sec
[junit] Tests run: 5, Failures: 0, Errors: 0, Time elapsed: 0.077 sec
[junit]
[junit] Running de.zib.scalarish.PubSubTest
[junit] Testsuite: de.zib.scalarish.PubSubTest
[junit] Tests run: 33, Failures: 0, Errors: 0, Time elapsed: 4.105 sec
[junit] Tests run: 33, Failures: 0, Errors: 0, Time elapsed: 4.105 sec
[junit]
[junit] ----- Standard Error -----
[junit] 2011-03-25 15:07:04.412: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:04.558: INFO::Started SelectChannelConnector@127.0.0.1:59235
[junit] 2011-03-25 15:07:05.632: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:05.635: INFO::Started SelectChannelConnector@127.0.0.1:41335
[junit] 2011-03-25 15:07:05.635: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:05.643: INFO::Started SelectChannelConnector@127.0.0.1:38552
[junit] 2011-03-25 15:07:05.643: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:05.646: INFO::Started SelectChannelConnector@127.0.0.1:34704
[junit] 2011-03-25 15:07:06.864: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:06.864: INFO::Started SelectChannelConnector@127.0.0.1:57898
[junit] 2011-03-25 15:07:06.864: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:06.865: INFO::Started SelectChannelConnector@127.0.0.1:47949
[junit] 2011-03-25 15:07:06.865: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:06.866: INFO::Started SelectChannelConnector@127.0.0.1:53886
[junit] 2011-03-25 15:07:07.090: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:07.093: INFO::Started SelectChannelConnector@127.0.0.1:33141
[junit] 2011-03-25 15:07:07.094: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:07.096: INFO::Started SelectChannelConnector@127.0.0.1:39119
[junit] 2011-03-25 15:07:07.096: INFO::jetty-7.3.0.v20110203
[junit] 2011-03-25 15:07:07.097: INFO::Started SelectChannelConnector@127.0.0.1:41603
[junit] -----
[junit] Running de.zib.scalarish.ReplicatedDHTTest
[junit] Testsuite: de.zib.scalarish.ReplicatedDHTTest
[junit] Tests run: 6, Failures: 0, Errors: 0, Time elapsed: 0.732 sec
[junit] Tests run: 6, Failures: 0, Errors: 0, Time elapsed: 0.732 sec
[junit]
[junit] Running de.zib.scalarish.TransactionSingleOpTest
[junit] Testsuite: de.zib.scalarish.TransactionSingleOpTest
[junit] Tests run: 28, Failures: 0, Errors: 0, Time elapsed: 0.632 sec
[junit] Tests run: 28, Failures: 0, Errors: 0, Time elapsed: 0.632 sec
[junit]
[junit] Running de.zib.scalarish.TransactionTest
[junit] Testsuite: de.zib.scalarish.TransactionTest
[junit] Tests run: 18, Failures: 0, Errors: 0, Time elapsed: 0.782 sec
[junit] Tests run: 18, Failures: 0, Errors: 0, Time elapsed: 0.782 sec
[junit]

test:

BUILD SUCCESSFUL
Total time: 10 seconds
'jtest_boot@csr-pc9.zib.de'

```

5.3. Python unit tests

The Python unit tests can be run by executing `make python-test` in the main directory. This will start a Scalaris node with the default ports and test all Python functions part of the Python API. A typical run will look like the following:

```

%> make python-test
[...]
testDoubleClose (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testRead_NotConnected (TransactionSingleOpTest.TestTransactionSingleOp) ... ok

```

```

testRead_NotFound (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetList1 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetList2 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetList_NotConnected (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetList_NotFound (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetString1 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetString2 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetString_NotConnected (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTestAndSetString_NotFound (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTransactionSingleOp1 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testTransactionSingleOp2 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testWriteList1 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testWriteList2 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testWriteList_NotConnected (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testWriteString1 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testWriteString2 (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testWriteString_NotConnected (TransactionSingleOpTest.TestTransactionSingleOp) ... ok
testAbort_Empty (TransactionTest.TestTransaction) ... ok
testAbort_NotConnected (TransactionTest.TestTransaction) ... ok
testCommit_Empty (TransactionTest.TestTransaction) ... ok
testCommit_NotConnected (TransactionTest.TestTransaction) ... ok
testDoubleClose (TransactionTest.TestTransaction) ... ok
testRead_NotConnected (TransactionTest.TestTransaction) ... ok
testRead_NotFound (TransactionTest.TestTransaction) ... ok
testTransaction1 (TransactionTest.TestTransaction) ... ok
testTransaction3 (TransactionTest.TestTransaction) ... ok
testWriteList1 (TransactionTest.TestTransaction) ... ok
testWriteString (TransactionTest.TestTransaction) ... ok
testWriteString_NotConnected (TransactionTest.TestTransaction) ... ok
testWriteString_NotFound (TransactionTest.TestTransaction) ... ok
testDelete1 (ReplicatedDHTTest.TestReplicatedDHT) ... ok
testDelete2 (ReplicatedDHTTest.TestReplicatedDHT) ... ok
testDelete_notExistingKey (ReplicatedDHTTest.TestReplicatedDHT) ... ok
testDoubleClose (ReplicatedDHTTest.TestReplicatedDHT) ... ok
testReplicatedDHT1 (ReplicatedDHTTest.TestReplicatedDHT) ... ok
testReplicatedDHT2 (ReplicatedDHTTest.TestReplicatedDHT) ... ok
testDoubleClose (PubSubTest.TestPubSub) ... ok
testGetSubscribersOtp_NotConnected (PubSubTest.TestPubSub) ... ok
testGetSubscribers_NotExistingTopic (PubSubTest.TestPubSub) ... ok
testPubSub1 (PubSubTest.TestPubSub) ... ok
testPubSub2 (PubSubTest.TestPubSub) ... ok
testPublish1 (PubSubTest.TestPubSub) ... ok
testPublish2 (PubSubTest.TestPubSub) ... ok
testPublish_NotConnected (PubSubTest.TestPubSub) ... ok
testSubscribe1 (PubSubTest.TestPubSub) ... ok
testSubscribe2 (PubSubTest.TestPubSub) ... ok
testSubscribe_NotConnected (PubSubTest.TestPubSub) ... ok
testSubscription1 (PubSubTest.TestPubSub) ... ok
testSubscription2 (PubSubTest.TestPubSub) ... ok
testSubscription3 (PubSubTest.TestPubSub) ... ok
testSubscription4 (PubSubTest.TestPubSub) ... ok
testUnsubscribe1 (PubSubTest.TestPubSub) ... ok
testUnsubscribe2 (PubSubTest.TestPubSub) ... ok
testUnsubscribe_NotConnected (PubSubTest.TestPubSub) ... ok
testUnsubscribe_NotExistingTopic (PubSubTest.TestPubSub) ... ok
testUnsubscribe_NotExistingUrl (PubSubTest.TestPubSub) ... ok

```

```

-----
Ran 58 tests in 12.317s

```

```
OK
```

```
'jtest_boot@csr-pc9.zib.de'
```

5.4. Interoperability Tests

In order to check whether the common types described in Section 4.1 on page 19 are fully supported by the APIs and yield to the appropriate types in another API, we implemented some interoperability tests. They can be run by executing `make interop-test` in the main directory. This will start a Scalaris node with the default ports, write test data using both the Java and the Python APIs and let each API read the data it wrote itself as well as the data the other API read. On success it will print

```
%> make interop-test
[...]  
all tests successful
```

6. Troubleshooting

Description is based on SVN revision r1618.

6.1. Network

Scalaris uses a couple of TCP ports for communication. It does not use UDP at the moment.

	HTTP Server	Inter-node communication
default (see bin/scalaris.cfg)	8000	14195–14198
first node (bin/firstnode.sh)	8000	14195
joining node 1 (bin/joining_node.sh)	8001	14196
other joining nodes (bin/joining_node.sh <ID>)	8000 + <ID>	14195 + <ID>
standalone mgmt server (bin/mgmt-server.sh)	7999	14194

Please make sure that at least 14195 and 14196 are not blocked by firewalls in order to be able to start at least one first and one joining node on each machine..

6.2. Miscellaneous

For up-to-date information about frequently asked questions and troubleshooting, please refer to our FAQs at <https://code.google.com/p/scalaris/wiki/FAQ> and our mailing list at <http://groups.google.com/group/scalaris>.

Part II.

Developers Guide

7. General Hints

7.1. Coding Guidelines

- Keep the code short
- Use `gen_component` to implement additional processes
- Don't use `receive` by yourself (Exception: to implement single threaded user API calls (`cs_api`, `yaws_calls`, etc))
- Don't use `erlang:now/0`, `erlang:send_after/3`, `receive after` etc. in performance critical code, consider using `msg_delay` instead.
- Don't use `timer:tc/3` as it catches exceptions. Use `util:tc/3` instead.

7.2. Testing Your Modifications and Extensions

- Run the testsuites using `make test`
- Run the java api test using `make java-test` (Scalaris output will be printed if a test fails; if you want to see it during the tests, start a `bin/firstnode.sh` and run the tests by `cd java; ant test`)
- Run the Ruby client by starting Scalaris and running `cd contrib; ./jsonrpc.rb`

7.3. Help with Digging into the System

- use `ets:i/0,1` to get details on the local state of some processes
- consider changing `pdb.erl` to use `ets` instead of `erlang:put/get`
- Have a look at `strace -f -p PID` of beam process
- Get message statistics via the Web-interface
- enable/disable tracing for certain modules
- Trace messages using the `trace_mpath` module
- Use `etop` and look at the total memory size and atoms generated
- send processes `sleep` or `kill` messages to test certain behaviour (see `gen_component.erl`)
- use `admin:number_of_nodes()`.
- use `admin:check_ring()`.

8. System Infrastructure

8.1. Groups of Processes

- What is it? How to distinguish from Erlangs internal named processes?
- Joining a process group
- Why do we do this... (managing several independent nodes inside a single Erlang VM for testing)

8.2. The Communication Layer `comm`

- in general
- format of messages (tuples)
- use messages with cookies (server and client side)
- What is a message tag?

8.3. The `gen_component`

Description is based on SVN revision r2675.

The generic component model implemented by `gen_component` allows to add some common functionality to all the components that build up the Scalaris system. It supports:

event-handlers: message handling with a similar syntax as used in [3].

FIFO order of messages: components cannot be inadvertently locked as we do not use selective receive statements in the code.

sleep and halt: for testing components can sleep or be halted.

debugging, breakpoints, stepwise execution: to debug components execution can be steered via breakpoints, step-wise execution and continuation based on arriving events and user defined component state conditions.

basic profiling,

state dependent message handlers: depending on its state, different message handlers can be used and switched during runtime. Thereby a kind of state-machine based message handling is supported.

prepared for `pid_groups`: allows to send events to named processes inside the same group as the actual component itself (`send_to_group_member`) when just holding a reference to any group member, and

unit-testing of event-handlers: as message handling is separated from the main loop of the component, the handling of individual messages and thereby performed state manipulation can easily be tested in unit-tests by directly calling message handlers.

In Scalaris all Erlang processes should be implemented as `gen_component`. The only exception are functions interfacing to the client, where a transition from asynchronous to synchronous request handling is necessary and that are executed in the context of a client's process or a process that behaves as a proxy for a client (`cs_api`).

8.3.1. A basic `gen_component` including a message handler

To implement a `gen_component`, the component has to provide the `gen_component` behaviour:

File `gen_component.erl`:

```

115 -ifdef(have_callback_support).
116 -callback init(Args::term()) -> user_state().
117 -else.
118 -spec behaviour_info(atom()) -> [{atom(), arity()}] | undefined.
119 behaviour_info(callbacks) ->
120     [
121         {init, 1} %% initialize component
122         %% note: can use arbitrary on-handler, but by default on/2 is used:
123         %% {on, 2} %% handle a single message
124         %% on(Msg, UserState) -> NewUserState | unknown_event | kill
125     ];
126 behaviour_info(_Other) -> undefined.
127 -endif.

```

This is illustrated by the following example:

File `msg_delay.erl`:

```

92 %% initialize: return initial state.
93 -spec init([]) -> state().
94 init([]) ->
95     ?TRACE("msg_delay:init for pid group ~p~n", [pid_groups:my_groupname()]),
96     %% For easier debugging, use a named table (generates an atom)
97     %% TableName = erlang:list_to_atom(pid_groups:my_groupname() ++ "_msg_delay"),
98     %% TimeTable = pdb:new(TableName, [set, protected, named_table]),
99     %% use random table name provided by ets to *not* generate an atom
100     TimeTable = pdb:new(?MODULE, [set]),
101     comm:send_local(self(), {msg_delay_periodic}),
102     _State = {TimeTable, _Round = 0}.
103
104 -spec on(message(), state()) -> state().
105 on({msg_delay_req, Seconds, Dest, Msg} = _FullMsg,
106     {TimeTable, Counter} = State) ->
107     ?TRACE("msg_delay:on(~.0p, ~.0p)~n", [_FullMsg, State]),
108     Future = trunc(Counter + Seconds),
109     EMsg = case erlang:get(trace_mpath) of
110         undefined -> Msg;
111         PState -> trace_mpath:epidemic_reply_msg(PState, comm:this(), Dest, Msg)
112     end,
113     case pdb:get(Future, TimeTable) of
114         undefined ->
115             pdb:set({Future, [{Dest, EMsg}]}, TimeTable);
116         {_, MsgQueue} ->
117             pdb:set({Future, [{Dest, EMsg} | MsgQueue]}, TimeTable)
118     end,
119     State;
120
121 %% periodic trigger
122 on({msg_delay_periodic} = Trigger, {TimeTable, Counter} = _State) ->
123     ?TRACE("msg_delay:on(~.0p, ~.0p)~n", [Trigger, State]),
124     % triggers are not allowed to be infected!
125     ?DBG_ASSERT2(not trace_mpath:infected(), trigger_infected),
126     _ = case pdb:take(Counter, TimeTable) of
127         undefined -> ok;
128         {_, MsgQueue} ->

```

```

129     _ = [ case Msg of
130         {'$gen_component', trace_mpath, PState, _From, _To, OrigMsg} ->
131             case element(2, PState) of %% element 2 is the logger
132                 {proto_sched, _} ->
133                     log:log("msg_delay: proto_sched not ready for delayed messages, so
134                         %% these messages should not be
135                         %% accepted to the database of
136                         %% msg_delay anyway, but instead
137                         %% should be immediately delivered
138                         %% to proto_sched for the 'delayed'
139                         %% messages pool (which is not
140                         %% implemented yet) (see send_local,
141                         %% send_local_as_client)
142                         %% erlang:throw(redirect_proto_sched_msgs_at_submission_please)
143                     ok;
144                 _ ->
145                     trace_mpath:start(PState),
146                     comm:send_local(Dest, OrigMsg),
147                     trace_mpath:stop()
148             end;
149         _ ->
150             ?DBG_ASSERT2(not trace_mpath:infected(), infected_with_uninfected_msg),
151             comm:send_local(Dest, Msg)
152         end || {Dest, Msg} <- MsgQueue ]
153     end,
154     _ = comm:send_local_after(1000, self(), Trigger),
155     {TimeTable, Counter + 1};
156
157 on({web_debug_info, Requestor}, {TimeTable, Counter} = State) ->
158     KeyValueList =
159         [{"queued messages (in 0-10s, messages):", ""}] |
160         [begin
161             Future = trunc(Counter + Seconds),
162             Queue = case pdb:get(Future, TimeTable) of
163                 undefined -> none;
164                 {_, Q} -> Q
165             end,
166             {webhelpers:safe_html_string("~p", [Seconds]),
167              webhelpers:safe_html_string("~p", [Queue])}
168         end || Seconds <- lists:seq(0, 10)]],
169     comm:send_local(Requestor, {web_debug_info_reply, KeyValueList}),
170     State.

```

`your_gen_component:init/1` is called during start-up of a `gen_component` and should return the initial state to be used for this `gen_component`. Later, the current state of the component can be retrieved using `gen_component:get_state/1`.

To react on messages / events, a message handler is used. The default message handler is given to `gen_component:start_link/4` as well as `gen_component:start/4` or `gen_component:start/5`. It can be changed by calling `gen_component:change_handler/2` (see Section 8.3.7). When an event / message for the component arrives, this handler is called with the event itself and the current state of the component. In the handler, the state of the component may be adjusted depending upon the event. The handler itself may trigger new events / messages for itself or other components and has finally to return the updated state of the component or the atoms `unknown_event` or `kill`. It must neither call `receive` nor `timer:sleep/1` nor `erlang:exit/1`.

8.3.2. How to start a `gen_component`?

A `gen_component` can be started using one of:

```

gen_component:start(Module, Handler, Args, GenCOptions = [])
gen_component:start_link(Module, Handler, Args, GenCOptions = [])

```

Module: the name of the module your component is implemented in

Handler: the initial message handler

Args: List of parameters passed to `Module:init/1` for initialization

GenCOptions: optional parameter. List of options for `gen_component`

`{pid_groups_join_as, ProcessGroup, ProcessName}`: registers the new process with the given process group (also called instanceid) and name using `pid_groups`.

`{erlang_register, ProcessName}`: registers the process as a named Erlang process.

`{wait_for_init}`: wait for `Module:init/1` to return before returning to the caller.

These functions are compatible to the Erlang/OTP supervisors. They spawn a new process for the component which itself calls `Module:init/1` with the given Args to initialize the component. `Module:init/1` should return the initial state for your component. For each message sent to this component, the default message handler `Module:on(Message, State)` will be called, which should react on the message and return the updated state of your component.

`gen_component:start()` and `gen_component:start_link()` return the pid of the spawned process as `{ok, Pid}`.

8.3.3. When does a `gen_component` terminate?

A `gen_component` can be stopped using:

`gen_component:kill(Pid)` or by returning `kill` from the current message handler.

8.3.4. How to determine whether a process is a `gen_component`?

A `gen_component` can be detected by:

`gen_component:is_gen_component(Pid)`, which returns a boolean.

8.3.5. What happens when unexpected events / messages arrive?

Your message handler (default is `your_gen_component:on/2`) should return `unknown_event` in the final clause (`your_gen_component:on(_, _)`). `gen_component` then will nicely report on the unhandled message, the component's name, its state and currently active message handler, as shown in the following example:

```
# bin/boot.sh
[...]  
(boot@localhost)10> pid_groups ! {no_message}.  
{no_message}  
[error] unknown message: {no_message} in Module: pid_groups and  
handler on in State null  
(boot@localhost)11>
```

The `pid_groups` (see Section 8.1) is a `gen_component` which registers itself as named Erlang process with the `gen_component` option `erlang_register` and therefore can be addressed by its name in the Erlang shell. We send it a `{no_message}` and `gen_component` reports on the unhandled message. The `pid_groups` module itself continues to run and waits for further messages.

8.3.6. What if my message handler generates an exception or crashes the process?

`gen_component` catches exceptions generated by message handlers and reports them with a stack trace, the message, that generated the exception, and the current state of the component.

If a message handler terminates the process via `erlang:exit/1`, this is out of the responsibility scope of `gen_component`. As usual in Erlang, all linked processes will be informed. If for example `gen_component:start_link/2` or `/3` was used for starting the `gen_component`, the spawning process will be informed, which may be an Erlang supervisor process taking further actions.

8.3.7. Changing message handlers and implementing state dependent message responsiveness as a state-machine

Sometimes it is beneficial to handle messages depending on the state of a component. One possibility to express this is implementing different clauses depending on the state variable, another is introducing case clauses inside message handlers to distinguish between current states. Both approaches may become tedious, error prone, and may result in confusing source code.

Sometimes the use of several different message handlers for different states of the component leads to clearer arranged code, especially if the set of handled messages changes from state to state. For example, if we have a component with an initialization phase and a production phase afterwards, we can handle in the first message handler messages relevant during the initialization phase and simply queue all other requests for later processing using a common default clause.

When initialization is done, we handle the queued user requests and switch to the message handler for the production phase. The message handler for the initialization phase does not need to know about messages occurring during production phase and the message handler for the production phase does not need to care about messages used during initialization. Both handlers can be made independent and may be extended later on without any adjustments to the other.

One can also use this scheme to implement complex state-machines by changing the message handler from state to state.

To switch the message handler `gen_component:change_handler(State, new_handler)` is called as the last operation after a message in the active message handler was handled, so that the return value of `gen_component:change_handler/2` is propagated to `gen_component`. The new handler is given as an atom, which is the name of the 2-ary function in your component module to be called.

Starting with non-default message handler.

It is also possible to change the message handler right from the start in your `your_gen_component:init/1` to avoid the default message handler `your_gen_component:on/2`. Just create your initial state as usual and call `gen_component:change_handler(State, my_handler)` as the final call in your `your_gen_component:init/1`. We prepared `gen_component:change_handler/2` to return `State` itself, so this will work properly.

8.3.8. Handling several messages atomically

The message handler is called for each message separately. Such a single call is atomic, i.e. the component does not perform any other action until the called message handler finishes. Sometimes, it is necessary to execute two or more calls to the message handler atomically (without other

interleaving messages). For example if a message A contains another message B as payload, it may be necessary to handle A and B directly one after the other without interference of other messages. So, after handling A you want to call your message handler with B.

In most cases, you could just do so by calculating the new state as result of handling message A first and then calling the message handler with message B and the new state by yourself.

It is safer to use `gen_component:post_op(2)` in such cases: When *B* contains a special message, which is usually handled by the `gen_component` module itself (like `send_to_group_member`, `kill`, `sleep`), the direct call to the message handler would not achieve the expected result. By calling `gen_component:post_op(B, NewState)` to return the new state after handling message A, message B will be handled directly after the current message A.

8.3.9. Halting and pausing a `gen_component`

Using `gen_component:kill(Pid)` and `gen_component:sleep(Pid, Time)` components can be terminated or paused.

8.3.10. Integration with `pid_groups`: Redirecting messages to other `gen_components`

Each `gen_component` by itself is prepared to support `comm:send_to_group_member/3` which forwards messages inside a group of processes registered via `pid_groups` (see Section 8.1) by their name. So, if you hold a `Pid` of one member of a process group, you can send messages to other members of this group, if you know their registered Erlang name. You do not necessarily have to know their individual `Pid`.

In consequence, no `gen_component` can individually handle messages of the form `{send_to_group_member, _, _}` as such messages are consumed by `gen_component` itself.

8.3.11. Replying to ping messages

Each `gen_component` replies automatically to `{ping, Pid}` requests with a `{pong}` send to the given `Pid`. Such messages are generated, for example, by `vivaldi_latency` which is used by our `vivaldi` module.

In consequence, no `gen_component` can individually handle messages of the form: `{ping, _}` as such messages are consumed by `gen_component` itself.

8.3.12. The debugging interface of `gen_component`: Breakpoints and step-wise execution

We equipped `gen_component` with a debugging interface, which especially is beneficial, when testing the interplay between several `gen_components`. It supports breakpoints (bp) which can pause the `gen_component` depending on the arriving messages or depending on user defined conditions. If a breakpoint is reached, the execution can be continued step-wise (message by message) or until the next breakpoint is reached.

We use it in our unit tests to steer protocol interleavings and to perform tests using random protocol interleavings between several processes (see `paxos_SUITE`). It allows also to reproduce given protocol interleavings for better testing.

Managing breakpoints.

Breakpoints are managed by the following functions:

`gen_component:bp_set(Pid, MsgTag, BPName)`: For the component running under `Pid` a breakpoint `BPName` is set. It is reached, when a message with a message tag `MsgTag` is next to be handled by the component (See `comm:get_msg_tag/1` and Section 8.2 for more information on message tags). The `BPName` is used as a reference for this breakpoint, for example to delete it later.

`gen_component:bp_set_cond(Pid, Cond, BPName)`: The same as `gen_component:bp_set/3` but a user defined condition implemented in `{Module, Function, Params = 2}` = `Cond` is checked by calling `Module:Function(Message, State)` to decide whether a breakpoint is reached or not. `Message` is the next message to be handled by the component and `State` is the current state of the component. `Module:Function/2` should return a boolean.

`gen_component:bp_del(Pid, BPName)`: The breakpoint `BPName` is deleted. If the component is in this breakpoint, it will not be released by this call. This has to be done separately by `gen_component:bp_cont/1`. But the deleted breakpoint will no longer be considered for newly entering a breakpoint.

`gen_component:bp_barrier(Pid)`: Delay all further handling of breakpoint requests until a breakpoint is actually entered.

Note, that the following call sequence may not catch the breakpoint at all, as during the sleep the component not necessarily consumes a ping message and the set breakpoint 'sample_bp' may already be deleted before a ping message arrives.

```
gen_component:bp_set(Pid, ping, sample_bp),
timer:sleep(10),
gen_component:bp_del(Pid, sample_bp),
gen_component:bp_cont(Pid).
```

To overcome this, `gen_component:bp_barrier/1` can be used:

```
gen_component:bp_set(Pid, ping, sample_bp),
gen_component:bp_barrier(Pid),
%% After the bp_barrier request, following breakpoint requests
%% will not be handled before a breakpoint is actually entered.
%% The gen_component itself is still active and handles messages as usual
%% until it enters a breakpoint.
gen_component:bp_del(Pid, sample_bp),
% Delete the breakpoint after it was entered once (ensured by bp_barrier).
% Release the gen_component from the breakpoint and continue.
gen_component:bp_cont(Pid).
```

None of the calls in the sample listing above is blocking. It just schedules all the operations, including the `bp_barrier`, for the `gen_component` and immediately finishes. The actual events of entering and continuing the breakpoint in the `gen_component` happens independently later on, when the next ping message arrives.

Managing execution.

The execution of a `gen_component` can be managed by the following functions:

`gen_component:bp_step(Pid)`: This is the only blocking breakpoint function. It waits until the `gen_component` is in a breakpoint and has handled a single message. It returns the module,

the active message handler, and the handled message as a tuple `{Module, On, Message}`. This function does not actually finish the breakpoint, but just lets a single message pass through. For further messages, no breakpoint condition has to be valid, the original breakpoint is still active. To leave a breakpoint, use `gen_component:bp_cont/1`.

`gen_component:bp_cont(Pid)`: Leaves a breakpoint. `gen_component` runs as usual until the next breakpoint is reached.

If no further breakpoints should be entered after continuation, you should delete the registered breakpoint using `gen_component:bp_del/2` before continuing the execution with `gen_component:bp_cont/1`. To ensure, that the breakpoint is entered at least once, `gen_component:bp_barrier/1` should be used before deleting the breakpoint (see the example above). Otherwise it could happen, that the delete request arrives at your `gen_component` before it was actually triggered. The following continuation request would then unintentional apply to an unrelated breakpoint that may be entered later on.

`gen_component:runnable(Pid)`: Returns whether a `gen_component` has messages to handle and is runnable. If you know, that a `gen_component` is in a breakpoint, you can use this to check, whether a `gen_component:bp_step/1` or `gen_component:bp_cont/1` is applicable to the component.

Tracing handled messages – getting a message interleaving protocol.

We use the debugging interface of `gen_component` to test protocols with random interleaving. First we start all the components involved, set breakpoints on the initialization messages for a new Paxos consensus and then start a single Paxos instance on all of them. The outcome of the Paxos consensus is a `learner_decide` message. So, in `paxos_SUITE:step_until_decide/3` we look for runnable processes and select randomly one of them to perform a single step until the protocol finishes with a decision.

File `paxos_SUITE.erl`:

```

234 -spec prop_rnd_interleave(1..4, 4..16, {pos_integer(), pos_integer(), pos_integer()})
235     -> true.
236 prop_rnd_interleave(NumProposers, NumAcceptors, Seed) ->
237     ct:pal("Called with: paxos_SUITE:prop_rnd_interleave(~p, ~p, ~p).~n",
238         [NumProposers, NumAcceptors, Seed]),
239     Majority = NumAcceptors div 2 + 1,
240     {Proposers, Acceptors, Learners} =
241         make(NumProposers, NumAcceptors, 1, "rnd_interleave"),
242     %% set bp on all processes
243     _ = [ gen_component:bp_set(comm:make_local(X), ?proposer_initialize, bp)
244         || X <- Proposers ],
245     _ = [ gen_component:bp_set(comm:make_local(X), acceptor_initialize, bp)
246         || X <- Acceptors ],
247     _ = [ gen_component:bp_set(comm:make_local(X), learner_initialize, bp)
248         || X <- Learners ],
249     %% start paxos instances
250     _ = [ proposer:start_paxosid(X, paxidrndinterl, Acceptors,
251         proposal, Majority, NumProposers, Y)
252         || {X,Y} <- lists:zip(Proposers, lists:seq(1, NumProposers)) ],
253     _ = [ acceptor:start_paxosid(X, paxidrndinterl, Learners)
254         || X <- Acceptors ],
255     _ = [ learner:start_paxosid(X, paxidrndinterl, Majority,
256         comm:this(), cpaxidrndinterl)
257         || X <- Learners ],
258     %% randomly step through protocol
259     OldSeed = random:seed(Seed),
260     Steps = step_until_decide(Proposers ++ Acceptors ++ Learners, cpaxidrndinterl, 0),
261     ct:pal("Needed ~p steps~n", [Steps]),
262     _ = case OldSeed of

```



```

263         undefined -> ok;
264         _ -> random:seed(OldSeed)
265     end,
266     _ = [ gen_component:kill(comm:make_local(X))
267           || X <- lists:flatten([Proposers, Acceptors, Learners]) ],
268     true.
269
270 step_until_decide(Processes, PaxId, SumSteps) ->
271     %% io:format("Step ~p~n", [SumSteps]),
272     Runnable = [ X || X <- Processes, gen_component:runnable(comm:make_local(X)) ],
273     case Runnable of
274     [] ->
275         ct:pal("No runnable processes of ~p~n", [length(Processes)]),
276         timer:sleep(5), step_until_decide(Processes, PaxId, SumSteps);
277     _ ->
278         Num = random:uniform(length(Runnable)),
279         _ = gen_component:bp_step(comm:make_local(lists:nth(Num, Runnable))),
280         receive
281             {learner_decide, cpaxidrndinterl, _, _Res} = _Any ->
282                 %% io:format("Received ~p~n", [_Any]),
283                 SumSteps
284         after 0 -> step_until_decide(Processes, PaxId, SumSteps + 1)
285         end
286     end.

```

To get a message interleaving protocol, we either can output the results of each `gen_component:bp_step/1` call together with the `Pid` we selected for stepping, or alter the definition of the macro `TRACE_BP_STEPS` in `gen_component`, when we execute all `gen_components` locally in the same Erlang virtual machine.

File `gen_component.erl`:

```

40 %-define(TRACE_BP_STEPS(X,Y), io:format(X,Y)).      %% output on console
41 %-define(TRACE_BP_STEPS(X,Y), log:pal(X,Y)).        %% output even if called by unittest
42 %-define(TRACE_BP_STEPS(X,Y), io:format(user,X,Y)). %% clean output even if called by unittest
43 -define(TRACE_BP_STEPS(X,Y), ok).

```

8.3.13. Future use and planned extensions for `gen_component`

`gen_component` could be further extended. For example it could support hot-code upgrade or could be used to implement algorithms that have to be run across several components of *Scalaris* like snapshot algorithms or similar extensions.

8.4. The Process' Database (pdb)

- How to use it and how to switch from `erlang:put/set` to `ets` and implied limitations.

8.5. Failure Detectors (fd)

- uses Erlang monitors locally
- is independent of component load
- uses heartbeats between Erlang virtual machines
- uses a single proxy heartbeat server per Erlang virtual machine, which itself uses Erlang monitors to monitor locally

- uses dynamic timeouts to implement an eventually perfect failure detector.

8.6. Monitoring Statistics (`monitor`, `rrd`)

Description is based on SVN revision r6684.

The `monitor` module offers several methods to gather meaningful statistics using the `rrd()` data type defined in `rrd`.

`rrd()` records work with time slots, i.e. a fixed slot length is given at creation and items which should be inserted will be either put into the current slot, or a new slot will be created. Each data item thus needs a time stamp associated with it. It must not be a real time, but can also be a virtual time stamp.

The `rrd` module thus offers two different APIs: one with transparent time handling, e.g. `rrd:create/3`, `rrd:add_now/2`, and one with manual time handling, e.g. `rrd:create/4`, `rrd:add/3`.

To allow different evaluations of the stored data, the following types of data are supported:

- `gauge`: only stores the newest value of a time slot, e.g. for thermometers,
- `counter`: sums up all values inside a time slot,
- `timing`: records time spans and stores values to easily calculate e.g. the sum, the standard deviation, the number of events, the min and max,
- `timing_with_hist`: similar to `timing` but also records a more detailed (approximated) histogram of the data,
- `event`: records each event (including its time stamp) inside a time slot in a list (this should be rarely used as the amount of data stored may be very big).
- `histogram, N`: records values in an approximative histogram of size `N`
- `histogram_rt, N, BaseKey`: histogram of size `N` which operates on the key space of the DHT. `BaseKey` is the key with the largest distance to all keys in the histogram.

The `monitor` offers functions to conveniently store and retrieve such values. It is also started as a process in each `dht_node` and `basic_services` group as well as inside each `clients_group`. This process ultimately stores the whole `rrd()` structure. There are three paradigms how values can be stored:

1. Values are gathered in the process that is generating the values. Inside this process, the `rrd()` is stored in the erlang dictionary. Whenever a new time slot is started, the values will be reported to the monitor process of the gathering process' group.
2. Values are gathered in the process that is generating the values. Inside this process, the `rrd()` is handled manually. After changing the `rrd()`, a manual check for reporting needs to be issued using `monitor:check_report/4`.
3. Values are immediately send to the monitor process where it undergoes the same procedures until it is finally stored and available to other processes. This is especially useful if the process generating the values does not live long or does not regularly create new data, e.g. the client.

The following example illustrates the first mode, i.e. gathering data in the generating process. It has been taken from the `cyclon` module which uses a `counter` data type:

```
% initialise the monitor with an empty rrd() using a 60s monitoring interval
monitor:proc_set_value(?MODULE, 'shuffle', rrd:create(60 * 1000000, 3, counter)),
```

```
% update the value by adding one
monitor:proc_set_value(?MODULE, 'shuffle', fun(Old) -> rrd:add_now(1, Old) end),
% check regularly whether to report the data to the monitor:
monitor:proc_check_timeslot(?MODULE, 'shuffle')
```

The first two parameters of `monitor:proc_set_value/3` define the name of a monitored value, the module's name and a unique key. The second can be either an `rrd()` or an update fun. The `monitor:proc_check_timeslot/3` function can be used if your module does not regularly create new data. In this case, the monitor process would not have the latest data for others to retrieve. This function forces a check and creates the new time slot if needed (thus reporting the data).

This is how forwarding works (taken from `api_tx`):

```
monitor:client_monitor_set_value(
  ?MODULE, 'req_list',
  fun(Old) ->
    Old2 = case Old of
      _ -> Old
    end,
    % 10s monitoring interval, only keep newest in the client process
    undefined -> rrd:create(10 * 1000000, 1, {timing, ms});
    _ -> Old
  end,
  rrd:add_now(TimeInUs / 1000, Old2)
end),
```

As in this case there is no safe way of initialising the value, it is more useful to provide an update fun to `monitor:client_monitor_set_value/3`. This function is only useful for the client processes as it reports to the monitor in the `clients_group` (recall that client processes do not belong to any group). All other processes should use `monitor:monitor_set_value/3` with the same semantics.

8.7. Writing Unit Tests

8.7.1. Plain Unit Tests

8.7.2. Randomized Testing Using `tester`

8.7.3. Randomized Testing Using `proto_sched`

9. Basic Structured Overlay

9.1. Ring Maintenance

9.2. T-Man

9.3. Routing Tables

Description is based on SVN revision r4005.

Each node of the ring can perform searches in the overlay.

A search is done by a lookup in the overlay, but there are several other demands for communication between peers. Scalaris provides a general interface to route a message to the (other) peer, which is currently responsible for a given key.

File `api_dht_raw.erl`:

```
34 -spec unreliable_lookup(Key::?RT:key(), Msg::comm:message()) -> ok.
35 unreliable_lookup(Key, Msg) ->
36     comm:send_local(pid_groups:find_a(dht_node),
37         {?lookup_aux, Key, 0, Msg}).
38
39 -spec unreliable_get_key(Key::?RT:key()) -> ok.
40 unreliable_get_key(Key) ->
41     unreliable_lookup(Key, {?get_key, comm:this(), noid, Key}).
42
43 -spec unreliable_get_key(CollectorPid::comm:myid(),
44     ReqId::{rdht_req_id, pos_integer()},
45     Key::?RT:key()) -> ok.
46 unreliable_get_key(CollectorPid, ReqId, Key) ->
47     unreliable_lookup(Key, {?get_key, CollectorPid, ReqId, Key}).
```

The message `Msg` could be a `get_key` which retrieves content from the responsible node or a `get_node` message, which returns a pointer to the node.

All currently supported messages are listed in the file `dht_node.erl`.

The message routing is implemented in `dht_node_lookup.erl`

File `dht_node_lookup.erl`:

```
52 %% @doc Find the node responsible for Key and send him the message Msg.
53 -spec lookup_aux(State::dht_node_state:state(), Key::intervals:key(),
54     Hops::non_neg_integer(), Msg::comm:message()) -> ok.
55 lookup_aux(State, Key, Hops, Msg) ->
56     case config:read(leases) of
57     true ->
58         lookup_aux_leases(State, Key, Hops, Msg);
59     - ->
60         lookup_aux_chord(State, Key, Hops, Msg)
61     end.
62
63 -spec lookup_aux_chord(State::dht_node_state:state(), Key::intervals:key(),
```

```

64         Hops::non_neg_integer(), Msg::comm:message()) -> ok.
65 lookup_aux_chord(State, Key, Hops, Msg) ->
66     WrappedMsg = ?RT:wrap_message(Key, Msg, State, Hops),
67     case ?RT:next_hop(State, Key) of
68     {succ, P} -> % found node -> terminate
69         comm:send(P, {?lookup_fin, Key, ?HOPS_TO_DATA(Hops + 1), WrappedMsg}, [{shepherd, self()}]);
70     {other, P} ->
71         comm:send(P, {?lookup_aux, Key, Hops + 1, WrappedMsg}, [{shepherd, self()}])
72     end.
73
74 -spec lookup_aux_leases(State::dht_node_state:state(), Key::interval:key(),
75     Hops::non_neg_integer(), Msg::comm:message()) -> ok.
76 lookup_aux_leases(State, Key, Hops, Msg) ->
77     case leases:is_responsible(State, Key) of
78     true ->
79         comm:send_local(dht_node_state:get(State, monitor_proc),
80             {lookup_hops, Hops}),
81         DHTNode = pid_groups:find_a(dht_node),
82         %log:log("aux -> fin: ~p ~p~n", [self(), DHTNode]),
83         comm:send_local(DHTNode,
84             {?lookup_fin, Key, ?HOPS_TO_DATA(Hops + 1), Msg});
85     maybe ->
86         DHTNode = pid_groups:find_a(dht_node),
87         %log:log("aux -> fin: ~p ~p~n", [self(), DHTNode]),
88         comm:send_local(DHTNode,
89             {?lookup_fin, Key, ?HOPS_TO_DATA(Hops + 1), Msg});
90     false ->
91         MyRange = dht_node_state:get(State, my_range),
92         case intervals:in(Key, MyRange) of
93         true ->
94             % @doc we are here because leases and rm disagree
95             % over responsibility. One cause for this case can
96             % be join/sliding. Our successor still has the
97             % lease for our range. But rm already believes
98             % that we are responsible for our range. The
99             % solution is to forward the lookup to our
100             % successor instead of asking rt.
101             Succ = node:pidX(dht_node_state:get(State, succ)),
102             WrappedMsg = ?RT:wrap_message(Key, Msg, State, Hops),
103             comm:send(Succ, {?lookup_aux, Key, Hops + 1, WrappedMsg}, [{shepherd, self()}]);
104         false ->
105             WrappedMsg = ?RT:wrap_message(Key, Msg, State, Hops),
106             P = element(2, ?RT:next_hop(State, Key)),
107             comm:send(P, {?lookup_aux, Key, Hops + 1, WrappedMsg}, [{shepherd, self()}])
108         end
109     end.
110
111 %% @doc Find the node responsible for Key and send him the message Msg.
112 -spec lookup_fin(State::dht_node_state:state(), Key::interval:key(),
113     Data::data(), Msg::comm:message()) -> dht_node_state:state().
114 lookup_fin(State, Key, Hops, Msg) ->
115     case config:read(leases) of
116     true ->
117         lookup_fin_leases(State, Key, Hops, Msg);
118     _ ->
119         lookup_fin_chord(State, Key, Hops, Msg)
120     end.
121
122 -spec lookup_fin_chord(State::dht_node_state:state(), Key::interval:key(),
123     Data::data(), Msg::comm:message()) -> dht_node_state:state().
124 lookup_fin_chord(State, Key, Data, Msg) ->
125     MsgFwd = dht_node_state:get(State, msg_fwd),
126     FwdList = [P || {I, P} <- MsgFwd, intervals:in(Key, I)],
127     Hops = ?HOPS_FROM_DATA(Data),
128     case FwdList of
129     [] ->
130         case dht_node_state:is_db_responsible__no_msg_fwd_check(Key, State) of
131         true ->
132             %comm:send_local(dht_node_state:get(State, monitor_proc),
133                 {lookup_hops, Hops}),
134             %Unwrap = ?RT:unwrap_message(Msg, State),

```

```

135         %gen_component:post_op(Unwrap, State);
136         deliver(State, Msg, false, Hops);
137     false ->
138         % do not warn if
139         % a) received lookup_fin due to a msg_fwd while sliding and
140         % before the other node removed the message forward or
141         % b) our pred is not be aware of our ID change yet (after
142         % moving data to our successor) yet
143         SlidePred = dht_node_state:get(State, slide_pred),
144         SlideSucc = dht_node_state:get(State, slide_succ),
145         Neighbors = dht_node_state:get(State, neighbors),
146         InSlideIntervalFun =
147             fun(SlideOp) ->
148                 slide_op:is_slide(SlideOp) andalso
149                 slide_op:get_sendORreceive(SlideOp) == 'send' andalso
150                 intervals:in(Key, slide_op:get_interval(SlideOp))
151             end,
152         case lists:any(InSlideIntervalFun, [SlidePred, SlideSucc]) orelse
153             intervals:in(Key, nodelist:succ_range(Neighbors)) of
154             true -> ok;
155             false ->
156                 DBRange = dht_node_state:get(State, db_range),
157                 DBRange2 = [begin
158                     case intervals:is_continuous(Interval) of
159                     true -> {intervals:get_bounds(Interval), Id};
160                     _ -> {Interval, Id}
161                     end
162                     end || {Interval, Id} <- DBRange],
163                 log:log(warn,
164                     "[ ~.0p ] Routing is damaged (~p)!! Trying again...~n"
165                     " myrange:~p~n db_range:~p~n msgfwd:~p~n Key:~p~n"
166                     " pred:~.4p~n node:~.4p~n succ:~.4p",
167                     [self(), Data, intervals:get_bounds(nodelist:node_range(Neighbors),
168                     DBRange2, MsgFwd, Key, nodelist:pred(Neighbors),
169                     nodelist:node(Neighbors), nodelist:succ(Neighbors))]
170                 end,
171                 lookup_aux(State, Key, Hops, Msg),
172                 State
173             end;
174         [Pid] -> comm:send(Pid, {?lookup_fin, Key, ?HOPS_TO_DATA(Hops + 1), Msg}),
175         State
176     end.
177
178 -spec lookup_fin_leases(State::dht_node_state:state(), Key::intervals:key(),
179     Data::data(), Msg::comm:message()) -> dht_node_state:state().
180 lookup_fin_leases(State, Key, Data, Msg) ->
181     Hops = ?HOPS_FROM_DATA(Data),
182     case leases:is_responsible(State, Key) of
183     true ->
184         deliver(State, Msg, true, Hops);
185     maybe ->
186         deliver(State, Msg, false, Hops);
187     false ->
188         log:log("lookup_fin fail: ~p", [self()]),
189         lookup_aux(State, Key, Hops, Msg),
190         State
191     end.

```

Each node is responsible for a certain key interval. The function `intervals:in/2` is used to decide, whether the key is between the current node and its successor. If that is the case, the final step is delivers a `lookup_fin` message to the local node. Otherwise, the message is forwarded to the next nearest known peer (listed in the routing table) determined by `?RT:next_hop/2`.

`rt_beh.erl` is a generic interface for routing tables. It can be compared to interfaces in Java. In Erlang interfaces can be defined using a so called 'behaviour'. The files `rt_simple` and `rt_chord` implement the behaviour 'rt_beh'.

The macro `?RT` is used to select the current implementation of routing tables. It is defined in

include/scalaris.hrl.

File scalaris.hrl:

```
27 %%The RT macro determines which kind of routingtable is used. Uncomment the
28 %%one that is desired.
29
30 %%Standard Chord routingtable
31 -define(RT, rt_chord).
32 % first valid key:
33 -define(MINUS_INFINITY, 0).
34 -define(MINUS_INFINITY_TYPE, 0).
35 % first invalid key:
36 -define(PLUS_INFINITY, 16#10000000000000000000000000000000).
37 -define(PLUS_INFINITY_TYPE, 16#10000000000000000000000000000000).
38
39 %%Simple routingtable
40 %-define(RT, rt_simple).
41
42 %% Flexible Routing Tables
43 %% Standard flexible routingtable
44 %-define(RT, rt_frtchord).
45 %% Grouped Flexible Routing Table
46 %-define(RT, rt_gfrtchord).
```

The functions, that have to be implemented for a routing mechanism are defined in the following file:

File rt_beh.erl:

```
28 -ifdef(have_callback_support).
29 -include("scalaris.hrl").
30 -include("client_types.hrl").
31 -type rt() :: term().
32 -type external_rt() :: term().
33 -type key() :: term().
34
35 -callback empty_ext(nodelist:neighborhood()) -> external_rt().
36 -callback init(nodelist:neighborhood()) -> rt().
37 -callback hash_key(client_key() | binary()) -> key().
38 -callback get_random_node_id() -> key().
39 -callback next_hop(dht_node_state:state(), key()) -> {succ | other, comm:mypid()}.
40
41 -callback init_stabilize(nodelist:neighborhood(), rt()) -> rt().
42 -callback update(OldRT::rt(), OldNeighbors::nodelist:neighborhood(),
43                 NewNeighbors::nodelist:neighborhood())
44                 -> {trigger_rebuild, rt()} | {ok, rt()}.
45 -callback filter_dead_node(rt(), DeadPid::comm:mypid(), Reason::fd:reason()) -> rt().
46
47 -callback to_pid_list(rt()) -> [comm:mypid()].
48 -callback get_size(rt() | external_rt()) -> non_neg_integer().
49 -callback get_replica_keys(key()) -> [key()].
50 -callback get_key_segment(key()) -> pos_integer().
51
52 -callback n() -> number().
53 -callback get_range(Begin::key(), End::key() | ?PLUS_INFINITY_TYPE) -> number().
54 -callback get_split_key(Begin::key(), End::key() | ?PLUS_INFINITY_TYPE,
55                         SplitFraction::{Num::number(), Denom::pos_integer()})
56                         -> key() | ?PLUS_INFINITY_TYPE.
57 -callback get_split_keys(Begin::key(), End::key() | ?PLUS_INFINITY_TYPE,
58                          Parts::pos_integer()) -> [key()].
59 -callback get_random_in_interval(intervals:simple_interval2()) -> key().
60
61 -callback dump(RT::rt()) -> KeyValueType::{[Index::string(), Node::string()]}.
62
63 -callback to_list(dht_node_state:state()) -> nodelist:snodelist().
64 -callback export_rt_to_dht_node(rt(), Neighbors::nodelist:neighborhood()) -> external_rt().
65 -callback handle_custom_message(comm:message(), rt_loop:state_active()) -> rt_loop:state_active() | u
66
67 -callback check(OldRT::rt(), NewRT::rt(), Neighbors::nodelist:neighborhood(),
```

```

68         ReportToFD::boolean() -> ok.
69 -callback check(OldRT::rt(), NewRT::rt(), OldNeighbors::nodelist:neighborhood(),
70               NewNeighbors::nodelist:neighborhood(), ReportToFD::boolean() -> ok.
71
72 -callback check_config() -> boolean().
73 -callback wrap_message(Key::key(), Msg::comm:message(), State::dht_node_state:state(), Hops::non_neg_
74 -callback unwrap_message(Msg::comm:message(), State::dht_node_state:state()) ->
75     comm:message().
76
77 -else.
78 -spec behaviour_info(atom()) -> [{atom(), arity()}] | undefined.
79 behaviour_info(callbacks) ->
80 [
81     % create a default routing table
82     {empty_ext, 1},
83     % initialize a routing table
84     {init, 1},
85     % mapping: key space -> identifier space
86     {hash_key, 1}, {get_random_node_id, 0},
87     % routing
88     {next_hop, 2},
89     % trigger for new stabilization round
90     {init_stabilize, 2},
91     % adapt RT to changed neighborhood
92     {update, 3},
93     % dead nodes filtering
94     {filter_dead_node, 3},
95     % statistics
96     {to_pid_list, 1}, {get_size, 1},
97     % gets all (replicated) keys for a given (hashed) key
98     % (for symmetric replication)
99     {get_replica_keys, 1},
100    % get the segment of the ring a key belongs to (1-4)
101    {get_key_segment, 1},
102    % address space size, range and split key
103    % (may all throw 'throw:not_supported' if unsupported by the RT)
104    {n, 0}, {get_range, 2}, {get_split_key, 3},
105    % get a random key within the requested interval
106    {get_random_in_interval, 1},
107    % for debugging and web interface
108    {dump, 1},
109    % for bulkowner
110    {to_list, 1},
111    % convert from internal representation to version for dht_node
112    {export_rt_to_dht_node, 2},
113    % handle messages specific to a certain routing-table implementation
114    {handle_custom_message, 2},
115    % common methods
116    {check, 4}, {check, 5},
117    {check_config, 0},
118    % wrap and unwrap lookup messages
119    {wrap_message, 4},
120    {unwrap_message, 2}
121 ];
122 behaviour_info(_Other) ->
123     undefined.
124 -endif.

```

empty/1 gets a successor and generates an empty routing table for use inside the routing table implementation. The data structure of the routing table is undefined. It can be a list, a tree, a matrix ...

empty_ext/1 similarly creates an empty external routing table for use by the dht_node. This process might not need all the information a routing table implementation requires and can thus work with less data.

hash_key/1 gets a key and maps it into the overlay's identifier space.

get_random_node_id/0 returns a random node id from the overlay's identifier space. This is used for example when a new node joins the system.

next_hop/2 gets a dht_node's state (including the external routing table representation) and a key and returns the node, that should be contacted next when searching for the key, i.e. the known node nearest to the id.

init_stabilize/2 is called periodically to rebuild the routing table. The parameters are the identifier of the node, its successor and the old (internal) routing table state. This method may send messages to the routing_table process which need to be handled by the handle_custom_message/handler since they are implementation-specific.

update/7 is called when the node's ID, predecessor and/or successor changes. It updates the (internal) routing table with the (new) information.

filter_dead_node/2 is called by the failure detector and tells the routing table about dead nodes. This function gets the (internal) routing table and a node to remove from it. A new routing table state is returned.

to_pid_list/1 get the PIDs of all (internal) routing table entries.

get_size/1 get the (internal or external) routing table's size.

get_replica_keys/1 Returns for a given (hashed) Key the (hashed) keys of its replicas. This used for implementing symmetric replication.

n/0 gets the number of available keys. An implementation may throw `throw:not_supported` if the operation is unsupported by the routing table.

dump/1 dump the (internal) routing table state for debugging, e.g. by using the web interface. Returns a list of `{Index, Node_as_String}` tuples which may just as well be empty.

to_list/1 convert the (external) representation of the routing table inside a given `dht_node_state` to a sorted list of known nodes from the routing table, i.e. first=succ, second=next known node on the ring, ... This is used by bulk-operations to create a broadcast tree.

export_rt_to_dht_node/2 convert the internal routing table state to an external state. Gets the internal state and the node's neighborhood for doing so.

handle_custom_message/2 handle messages specific to the routing table implementation. `rt_loop` will forward unknown messages to this function.

check/5, check/6 check for routing table changes and send an updated (external) routing table to the `dht_node` process.

check_config/0 check that all required configuration parameters exist and satisfy certain restrictions.

wrap_message/1 wraps a message send via a `dht_node_lookup:lookup_aux/4`.

unwrap_message/2 unwraps a message send via `dht_node_lookup:lookup_aux/4` previously wrapped by `wrap_message/1`.

9.3.1. The routing table process (rt_loop)

The `rt_loop` module implements the process for all routing tables. It processes messages and calls the appropriate methods in the specific routing table implementations.

File `rt_loop.erl`:

```

40 -opaque(state_active() :: {Neighbors      :: nodelist:neighborhood(),
41                               RTState      :: ?RT:rt()}).
42 -type(state_inactive() :: {inactive,
43                               MessageQueue :: msg_queue:msg_queue()}).
44 %% -type(state() :: state_active() | state_inactive()).

```

If initialized, the node's id, its predecessor, successor and the routing table state of the selected implementation (the macro `RT` refers to).

File `rt_loop.erl`:


```

162 % Message handler to manage the trigger
163 on_active({trigger_rt}, State) ->
164     msg_delay:send_trigger(get_base_interval(), {trigger_rt}),
165     gen_component:post_op({periodic_rt_rebuild}, State);
166
167 % Actual periodic rebuilding of the RT
168 on_active({periodic_rt_rebuild}, {Neighbors, OldRT}) ->
169     % start periodic stabilization
170     % log:log(debug, "[ RT ] stabilize"),
171     NewRT = ?RT:init_stabilize(Neighbors, OldRT),
172     ?RT:check(OldRT, NewRT, Neighbors, true),
173     {Neighbors, NewRT};

```

Periodically (see `pointer_base_stabilization_interval` config parameter) a trigger message is sent to the `rt_loop` process that starts the periodic stabilization implemented by each routing table.

File `rt_loop.erl`:

```

147 % update routing table with changed ID, pred and/or succ
148 on_active({update_rt, OldNeighbors, NewNeighbors}, {_Neighbors, OldRT}) ->
149     case ?RT:update(OldRT, OldNeighbors, NewNeighbors) of
150         {trigger_rebuild, NewRT} ->
151             ?RT:check(OldRT, NewRT, OldNeighbors, NewNeighbors, true),
152             % trigger immediate rebuild
153             gen_component:post_op({periodic_rt_rebuild}, {NewNeighbors, NewRT})
154         ;
155         {ok, NewRT} ->
156             ?RT:check(OldRT, NewRT, OldNeighbors, NewNeighbors, true),
157             {NewNeighbors, NewRT}
158     end;

```

Every time a node's neighborhood changes, the `dht_node` sends an `update_rt` message to the routing table which will call `?RT:update/7` that decides whether the routing table should be rebuild. If so, it will stop any waiting trigger and schedule an immediate (periodic) stabilization.

9.3.2. Simple routing table (`rt_simple`)

One implementation of a routing table is the `rt_simple`, which routes via the successor. Note that this is inefficient as it needs a linear number of hops to reach its goal. A more robust implementation, would use a successor list. This implementation is also not very efficient in the presence of churn.

Data types

First, the data structure of the routing table is defined:

File `rt_simple.erl`:

```

26 -type key() :: 0..16#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF. % 128 bit numbers
27 -opaque rt() :: Succ::node:node_type().
28 -type external_rt() :: Succ::node:node_type(). %% @todo: make opaque
29 -type custom_message() :: none().

```

The routing table only consists of a node (the successor). Keys in the overlay are identified by integers ≥ 0 .

A simple rm_beh behaviour

File rt_simple.erl:

```
41 %% @doc Creates an "empty" routing table containing the successor.
42 -spec empty(nodelist:neighborhood()) -> rt().
43 empty(Neighbors) -> nodelist:succ(Neighbors).
```

File rt_simple.erl:

```
272 -spec empty_ext(nodelist:neighborhood()) -> external_rt().
273 empty_ext(Neighbors) -> empty(Neighbors).
```

The empty routing table (internal or external) consists of the successor.

File rt_simple.erl:

Keys are hashed using MD5 and have a length of 128 bits.

File rt_simple.erl:

```
59 %% @doc Generates a random node id, i.e. a random 128-bit number.
60 -spec get_random_node_id() -> key().
61 get_random_node_id() ->
62     case config:read(key_creator) of
63         random -> hash_key(randoms:getRandomString());
64         random_with_bit_mask ->
65             {Mask1, Mask2} = config:read(key_creator_bitmask),
66             (hash_key(randoms:getRandomString()) band Mask2) bor Mask1
67     end.
```

Random node id generation uses the helpers provided by the randoms module.

File rt_simple.erl:

```
277 %% @doc Returns the next hop to contact for a lookup.
278 -spec next_hop(dht_node_state:state(), key()) -> {succ | other, comm:mypid()}.
279 next_hop(State, Key) ->
280     Neighbors = dht_node_state:get(State, neighbors),
281     case intervals:in(Key, nodelist:succ_range(Neighbors)) of
282         true -> {succ, node:pidX(nodelist:succ(Neighbors))};
283         _ -> {other, node:pidX(nodelist:succ(Neighbors))}
284     end.
```

Next hop is always the successor.

File rt_simple.erl:

```
75 %% @doc Triggered by a new stabilization round, renews the routing table.
76 -spec init_stabilize(nodelist:neighborhood(), rt()) -> rt().
77 init_stabilize(Neighbors, _RT) -> empty(Neighbors).
```

init_stabilize/2 resets its routing table to the current successor.

File rt_simple.erl:

```
81 %% @doc Updates the routing table due to a changed node ID, pred and/or succ.
82 -spec update(OldRT::rt(), OldNeighbors::nodelist:neighborhood(),
83             NewNeighbors::nodelist:neighborhood()) -> {ok, rt()}.
84 update(_OldRT, _OldNeighbors, NewNeighbors) ->
85     {ok, nodelist:succ(NewNeighbors)}.
```

update/7 updates the routing table with the new successor.

File rt_simple.erl:

```
89 %% @doc Removes dead nodes from the routing table (rely on periodic
90 %%     stabilization here).
91 -spec filter_dead_node(rt(), DeadPid::comm:mypid(), Reason::fd:reason()) -> rt().
92 filter_dead_node(RT, _DeadPid, _Reason) -> RT.
```

filter_dead_node/2 does nothing, as only the successor is listed in the routing table and that is reset periodically in init_stabilize/2.

File rt_simple.erl:

```
96 %% @doc Returns the pids of the routing table entries.
97 -spec to_pid_list(rt()) -> [comm:mypid()].
98 to_pid_list(Succ) -> [node:pidX(Succ)].
```

to_pid_list/1 returns the pid of the successor.

File rt_simple.erl:

```
102 %% @doc Returns the size of the routing table.
103 -spec get_size(rt() | external_rt()) -> non_neg_integer().
104 get_size(_RT) -> 1.
```

The size of the routing table is always 1.

File rt_simple.erl:

```
192 %% @doc Returns the replicas of the given key.
193 -spec get_replica_keys(key()) -> [key()].
194 get_replica_keys(Key) ->
195     [Key,
196      Key bxor 16#40000000000000000000000000000000,
197      Key bxor 16#80000000000000000000000000000000,
198      Key bxor 16#C0000000000000000000000000000000
199     ].
```

This get_replica_keys/1 implements symmetric replication.

File rt_simple.erl:

```
108 %% @doc Returns the size of the address space.
109 -spec n() -> 16#10000000000000000000000000000000.
110 n() -> 16#10000000000000000000000000000000.
```

There are 2^{128} available keys.

File rt_simple.erl:

```
207 %% @doc Dumps the RT state for output in the web interface.
208 -spec dump(RT::rt()) -> KeyValueType::[{Index::string(), Node::string()}].
209 dump(Succ) -> [{"0", webhelpers:safe_html_string("~p", [Succ])}].
```

dump/1 lists the successor.

File rt_simple.erl:

```
295 %% @doc Converts the (external) representation of the routing table to a list
296 %%     in the order of the fingers, i.e. first=succ, second=shortest finger,
297 %%     third=next longer finger,...
298 -spec to_list(dht_node_state:state()) -> nodelist:nodelist().
299 to_list(State) -> [dht_node_state:get(State, succ)].
```

to_list/1 lists the successor from the external routing table state.

File rt_simple.erl:

```
288 %% @doc Converts the internal RT to the external RT used by the dht_node. Both
289 %%     are the same here.
290 -spec export_rt_to_dht_node(rt(), Neighbors::nodelist:neighborhood()) -> external_rt().
291 export_rt_to_dht_node(RT, _Neighbors) -> RT.
```

export_rt_to_dht_node/2 states that the external routing table is the same as the internal table.

File rt_simple.erl:

```
229 %% @doc There are no custom messages here.
230 -spec handle_custom_message
231       (custom_message() | any(), rt_loop:state_active()) -> unknown_event.
232 handle_custom_message(_Message, _State) -> unknown_event.
```

Custom messages could be send from a routing table process on one node to the routing table process on another node and are independent from any other implementation.

File rt_simple.hrl:

```
236 %% @doc Notifies the dht_node and failure detector if the routing table changed.
237 %%     Provided for convenience (see check/5).
238 -spec check(OldRT::rt(), NewRT::rt(), Neighbors::nodelist:neighborhood(),
239           ReportToFD::boolean()) -> ok.
240 check(OldRT, NewRT, Neighbors, ReportToFD) ->
241     check(OldRT, NewRT, Neighbors, Neighbors, ReportToFD).
242
243 %% @doc Notifies the dht_node if the (external) routing table changed.
244 %%     Also updates the failure detector if ReportToFD is set.
245 %%     Note: the external routing table only changes the internal RT has
246 %%     changed.
247 -spec check(OldRT::rt(), NewRT::rt(), OldNeighbors::nodelist:neighborhood(),
248           NewNeighbors::nodelist:neighborhood(), ReportToFD::boolean()) -> ok.
249 check(OldRT, NewRT, _OldNeighbors, NewNeighbors, ReportToFD) ->
250     case OldRT == NewRT of
251         true -> ok;
252         _ ->
253             Pid = pid_groups:get_my(dht_node),
254             RT_ext = export_rt_to_dht_node(NewRT, NewNeighbors),
255             comm:send_local(Pid, {rt_update, RT_ext}),
256             % update failure detector:
257             case ReportToFD of
258                 true ->
259                     NewPids = to_pid_list(NewRT),
260                     OldPids = to_pid_list(OldRT),
261                     fd:update_subscriptions(OldPids, NewPids);
262                 _ -> ok
263             end
264     end.
```

Checks whether the routing table changed and in this case sends the dht_node an updated (external) routing table state. Optionally the failure detector is updated. This may not be necessary, e.g. if check is called after a crashed node has been reported by the failure detector (the failure detector already unsubscribes the node in this case).

File rt_simple.hrl:

```
236 %% @doc Notifies the dht_node and failure detector if the routing table changed.
237 %%     Provided for convenience (see check/5).
238 -spec check(OldRT::rt(), NewRT::rt(), Neighbors::nodelist:neighborhood(),
239           ReportToFD::boolean()) -> ok.
240 check(OldRT, NewRT, Neighbors, ReportToFD) ->
241     check(OldRT, NewRT, Neighbors, Neighbors, ReportToFD).
```

```

242
243 %% @doc Notifies the dht_node if the (external) routing table changed.
244 %%      Also updates the failure detector if ReportToFD is set.
245 %%      Note: the external routing table only changes the internal RT has
246 %%      changed.
247 -spec check(OldRT::rt(), NewRT::rt(), OldNeighbors::nodelist:neighborhood(),
248           NewNeighbors::nodelist:neighborhood(), ReportToFD::boolean()) -> ok.
249 check(OldRT, NewRT, _OldNeighbors, NewNeighbors, ReportToFD) ->
250     case OldRT == NewRT of
251         true -> ok;
252         _ ->
253             Pid = pid_groups:get_my(dht_node),
254             RT_ext = export_rt_to_dht_node(NewRT, NewNeighbors),
255             comm:send_local(Pid, {rt_update, RT_ext}),
256             % update failure detector:
257             case ReportToFD of
258                 true ->
259                     NewPids = to_pid_list(NewRT),
260                     OldPids = to_pid_list(OldRT),
261                     fd:update_subscriptions(OldPids, NewPids);
262                 _ -> ok
263             end
264     end.

```

File `rt_simple.erl`:

```

303 %% @doc Wrap lookup messages. This is a noop in rt_simple.
304 -spec wrap_message(Key::key(), Msg::comm:message(), State::dht_node_state:state(),
305                  Hops::non_neg_integer()) -> comm:message().
306 wrap_message(_Key, Msg, _State, _Hops) -> Msg.

```

Wraps a message send via `dht_node_lookup:lookup/4` if needed. This routing algorithm does not need callbacks when finishing the lookup, so it does not need to wrap the message.

File `rt_simple.erl`:

```

310 %% @doc Unwrap lookup messages. This is a noop in rt_simple.
311 -spec unwrap_message(Msg::comm:message(), State::dht_node_state:state()) -> comm:message().
312 unwrap_message(Msg, _State) -> Msg.

```

Unwraps a message previously wrapped with `rt_simple:wrap_message/1`. As that function does not wrap messages, `rt_simple:unwrap_message/2` doesn't have to do anything as well.

9.3.3. Chord routing table (`rt_chord`)

The file `rt_chord.erl` implements Chord's routing.

Data types

File `rt_chord.erl`:

```

26 -type key() :: 0..16#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF. % 128 bit numbers
27 -type index() :: {pos_integer(), non_neg_integer()}.
28 -opaque rt() :: gb_trees:tree(index(), Node::node:node_type()).
29 -type external_rt() :: gb_trees:tree(NodeId::key(), Node::node:node_type()). %% @todo: make opaque
30 -type custom_message() ::
31     {rt_get_node, Source_PID::comm:mypid(), Index::index()} |
32     {rt_get_node_response, Index::index(), Node::node:node_type()}.

```

The routing table is a `gb_tree`. Identifiers in the ring are integers. Note that in Erlang integer can be of arbitrary precision. For Chord, the identifiers are in $[0, 2^{128})$, i.e. 128-bit strings.

The `rm_beh` behaviour for Chord (excerpt)

File `rt_chord.erl`:

```
48 %% @doc Creates an empty routing table.
49 -spec empty(nodelist:neighborhood()) -> rt().
50 empty(_Neighbors) -> gb_trees:empty().
```

File `rt_chord.erl`:

```
435 -spec empty_ext(nodelist:neighborhood()) -> external_rt().
436 empty_ext(_Neighbors) -> gb_trees:empty().
```

`empty/1` returns an empty `gb_tree`, same for `empty_ext/1`.

`rt_chord:hash_key/1`, `rt_chord:get_random_node_id/0`, `rt_chord:get_replica_keys/1` and `rt_chord:n/0` are implemented like their counterparts in `rt_simple.erl`.

File `rt_chord.erl`:

```
440 %% @doc Returns the next hop to contact for a lookup.
441 %%     If the routing table has less entries than the rt_size_use_neighbors
442 %%     config parameter, the neighborhood is also searched in order to find a
443 %%     proper next hop.
444 %%     Note, that this code will be called from the dht_node process and
445 %%     it will thus have an external_rt!
446 -spec next_hop(dht_node_state:state(), key()) -> {succ | other, comm:mypid()}.
447 next_hop(State, Id) ->
448     Neighbors = dht_node_state:get(State, neighbors),
449     case intervals:in(Id, nodelist:succ_range(Neighbors)) of
450     true ->
451         {succ, node:pidX(nodelist:succ(Neighbors))};
452     _ ->
453         % check routing table:
454         RT = dht_node_state:get(State, rt),
455         RTSize = get_size(RT),
456         NodeRT = case util:gb_trees_largest_smaller_than(Id, RT) of
457             {value, _Key, N} ->
458                 N;
459             nil when RTSize == 0 ->
460                 nodelist:succ(Neighbors);
461             nil -> % forward to largest finger
462                 {_Key, N} = gb_trees:largest(RT),
463                 N
464         end,
465         FinalNode =
466             case RTSize < config:read(rt_size_use_neighbors) of
467             false -> NodeRT;
468             _ ->
469                 % check neighborhood:
470                 nodelist:largest_smaller_than(Neighbors, Id, NodeRT)
471             end,
472         {other, node:pidX(FinalNode)}
473     end.
```

If the (external) routing table contains at least one item, the next hop is retrieved from the `gb_tree`. It will be the node with the largest id that is smaller than the id we are looking for. If the routing table is empty, the successor is chosen. However, if we haven't found the key in our routing table, the next hop will be our largest finger, i.e. entry.

File `rt_chord.erl`:

```
81 %% @doc Starts the stabilization routine.
82 -spec init_stabilize(nodelist:neighborhood(), rt()) -> rt().
83 init_stabilize(Neighbors, RT) ->
```

```

84      % calculate the longest finger
85      case first_index(Neighbors) of
86          null -> ok;
87          {Key, Index} ->
88              % trigger a lookup for Key
89              api_dht_raw:unreliable_lookup(
90                  Key, {?send_to_group_member, routing_table,
91                      {rt_get_node, comm:this(), Index}})
92      end,
93      RT.

```

The routing table stabilization is triggered for the first index and then runs asynchronously, as we do not want to block the `rt_loop` to perform other request while recalculating the routing table.

We have to find the node responsible for the calculated finger and therefore perform a lookup for the node with a `rt_get_node` message, including a reference to ourselves as the reply-to address and the index to be set.

The lookup performs an overlay routing by passing the message until the responsible node is found. There, the message is delivered to the `routing_table` process. The remote node sends the requested information back directly. It includes a reference to itself in a `rt_get_node_response` message. Both messages are handled by `rt_chord:handle_custom_message/2`:

File `rt_chord.erl`:

```

365 %% @doc Chord reacts on 'rt_get_node_response' messages in response to its
366 %%      'rt_get_node' messages.
367 -spec handle_custom_message(custom_message(), rt_loop:state_active()) ->
368                                     rt_loop:state_active() | unknown_event.
369 handle_custom_message({rt_get_node, Source_PID, Index}, State) ->
370     MyNode = nodelist:node(rt_loop:get_neighb(State)),
371     comm:send(Source_PID, {rt_get_node_response, Index, MyNode}, ?SEND_OPTIONS),
372     State;
373 handle_custom_message({rt_get_node_response, Index, Node}, State) ->
374     OldRT = rt_loop:get_rt(State),
375     Neighbors = rt_loop:get_neighb(State),
376     case stabilize(Neighbors, OldRT, Index, Node) of
377         {NewRT, true} ->
378             check_do_update(OldRT, NewRT, rt_loop:get_neighb(State), true);
379         {NewRT, false} -> ok
380     end,
381     rt_loop:set_rt(State, NewRT);
382 handle_custom_message(_Message, _State) ->
383     unknown_event.

```

File `rt_chord.erl`:

```

223 %% @doc Updates one entry in the routing table and triggers the next update.
224 %%      Changed indicates whether a new node was inserted (the RT structure may
225 %%      change independently from this indicator!).
226 -spec stabilize(Neighbors::nodelist:neighborhood(), OldRT::rt(), Index::index(),
227               Node::node:node_type()) -> {NewRT::rt(), Changed::boolean()}.
228 stabilize(Neighbors, RT, Index, Node) ->
229     MyId = nodelist:nodeid(Neighbors),
230     Succ = nodelist:succ(Neighbors),
231     case (node:id(Succ) /= node:id(Node)) % reached succ?
232         andalso (not intervals:in( % there should be nothing shorter
233             node:id(Node), % than succ
234             nodelist:succ_range(Neighbors))) of
235         true ->
236             NextIndex = next_index(Index),
237             NextKey = calculateKey(MyId, NextIndex),
238             CurrentKey = calculateKey(MyId, Index),
239             case CurrentKey /= NextKey of
240                 true ->
241                     Msg = {rt_get_node, comm:this(), NextIndex},

```

```

242         api_dht_raw:unreliable_lookup(
243             NextKey, {?send_to_group_member, routing_table, Msg});
244         - -> ok
245     end,
246     Changed = (Index == first_index() orelse
247         (gb_trees:lookup(prev_index(Index), RT) /= {value, Node})),
248     {gb_trees:enter(Index, Node, RT), Changed};
249 - ->
250     {RT, false}
251 end.

```

stabilize/5 assigns the received routing table entry and triggers the routing table stabilization for the the next shorter entry using the same mechanisms as described above.

If the shortest finger is the successor, then filling the routing table is stopped, as no further new entries would occur. It is not necessary, that Index reaches 1 to make that happen. If less than 2^{128} nodes participate in the system, it may happen earlier.

File `rt_chord.erl`:

```

255 %% @doc Updates the routing table due to a changed node ID, pred and/or succ.
256 -spec update(OldRT::rt(), OldNeighbors::odelist:neighborhood(),
257     NewNeighbors::odelist:neighborhood())
258     -> {ok | trigger_rebuild, rt()}.
259 update(OldRT, OldNeighbors, NewNeighbors) ->
260     NewPred = oodelist:pred(NewNeighbors),
261     OldSucc = oodelist:succ(OldNeighbors),
262     NewSucc = oodelist:succ(NewNeighbors),
263     NewNodeId = oodelist:nodeid(NewNeighbors),
264     % only re-build if a new successor occurs or the new node ID is not between
265     % Pred and Succ any more (which should not happen since this must come from
266     % a slide!)
267     % -> if not rebuilding, update the node IDs though
268     case node:same_process(OldSucc, NewSucc) andalso
269         intervals:in(NewNodeId, node:mk_interval_between_nodes(NewPred, NewSucc)) of
270     true ->
271         NewRT = gb_trees:map(
272             fun(_K, N) ->
273                 case node:same_process(N, NewPred) of
274                 true ->
275                     node:newer(N, NewPred);
276                 false ->
277                     case node:same_process(N, NewSucc) of
278                     true -> node:newer(N, NewSucc);
279                     false -> N
280                 end
281                 end
282             end, OldRT),
283     {ok, NewRT};
284 false ->
285     % to be on the safe side ...
286     {trigger_rebuild, empty(NewNeighbors)}
287 end.

```

Tells the `rt_loop` process to rebuild the routing table starting with an empty (internal) routing table state.

File `rt_chord.erl`:

```

97 %% @doc Removes dead nodes from the routing table.
98 -spec filter_dead_node(rt(), DeadPid::comm:my_pid(), Reason::fd:reason()) -> rt().
99 filter_dead_node(RT, DeadPid, _Reason) ->
100     DeadIndices = [Index || {Index, Node} <- gb_trees:to_list(RT),
101         node:same_process(Node, DeadPid)],
102     lists:foldl(fun(Index, Tree) -> gb_trees:delete(Index, Tree) end,
103         RT, DeadIndices).

```


filter_dead_node removes dead entries from the gb_tree.

File rt_chord.erl:

```
477 -spec export_rt_to_dht_node(rt(), Neighbors::nodelist:neighborhood()) -> external_rt().
478 export_rt_to_dht_node(RT, Neighbors) ->
479     Id = nodelist:nodeid(Neighbors),
480     Pred = nodelist:pred(Neighbors),
481     Succ = nodelist:succ(Neighbors),
482     % always include the pred and succ in the external representation
483     % note: we are subscribed at the RM for changes to these nodes
484     Tree = gb_trees:enter(node:id(Succ), Succ,
485         gb_trees:enter(node:id(Pred), Pred, gb_trees:empty())),
486     util:gb_trees_foldl(fun (_K, V, Acc) ->
487         % only store the ring id and the according node structure
488         case node:id(V) == Id of
489             true -> Acc;
490             false -> gb_trees:enter(node:id(V), V, Acc)
491         end
492     end, Tree, RT).
```

export_rt_to_dht_node converts the internal gb_tree structure based on indices into the external representation optimised for look-ups, i.e. a gb_tree with node ids and the nodes themselves.

File rt_chord.hrl:

```
387 %% @doc Notifies the dht_node and failure detector if the routing table changed.
388 %% Provided for convenience (see check/5).
389 -spec check(OldRT::rt(), NewRT::rt(), Neighbors::nodelist:neighborhood(),
390     ReportToFD::boolean()) -> ok.
391 check(OldRT, OldRT, _Neighbors, _ReportToFD) ->
392     ok;
393 check(OldRT, NewRT, Neighbors, ReportToFD) ->
394     check_do_update(OldRT, NewRT, Neighbors, ReportToFD).
395
396 %% @doc Notifies the dht_node if the (external) routing table changed.
397 %% Also updates the failure detector if ReportToFD is set.
398 %% Note: the external routing table also changes if the Pred or Succ
399 %% change.
400 -spec check(OldRT::rt(), NewRT::rt(), OldNeighbors::nodelist:neighborhood(),
401     NewNeighbors::nodelist:neighborhood(), ReportToFD::boolean()) -> ok.
402 check(OldRT, NewRT, OldNeighbors, NewNeighbors, ReportToFD) ->
403     case nodelist:pred(OldNeighbors) == nodelist:pred(NewNeighbors) andalso
404         nodelist:succ(OldNeighbors) == nodelist:succ(NewNeighbors) andalso
405         OldRT == NewRT of
406         true -> ok;
407         _ -> check_do_update(OldRT, NewRT, NewNeighbors, ReportToFD)
408     end.
409
410 %% @doc Helper for check/4 and check/5.
411 -spec check_do_update(OldRT::rt(), NewRT::rt(), NewNeighbors::nodelist:neighborhood(),
412     ReportToFD::boolean()) -> ok.
413 check_do_update(OldRT, NewRT, NewNeighbors, ReportToFD) ->
414     Pid = pid_groups:get_my(dht_node),
415     RT_ext = export_rt_to_dht_node(NewRT, NewNeighbors),
416     case Pid of
417         failed -> ok;
418         _ -> comm:send_local(Pid, {rt_update, RT_ext})
419     end,
420     % update failure detector:
421     case ReportToFD of
422         true ->
423             NewPids = to_pid_list(NewRT),
424             OldPids = to_pid_list(OldRT),
425             fd:update_subscriptions(OldPids, NewPids);
426         _ -> ok
427     end.
```

Checks whether the routing table changed and in this case sends the `dht_node` an updated (external) routing table state. Optionally the failure detector is updated. This may not be necessary, e.g. if `check` is called after a crashed node has been reported by the failure detector (the failure detector already unsubscribes the node in this case).

File `rt_chord.erl`:

```
506 %% @doc Wrap lookup messages. This is a noop in Chord.
507 -spec wrap_message(Key::key(), Msg::comm:message(), State::dht_node_state:state(),
508                   Hops::non_neg_integer()) -> comm:message().
509 wrap_message(_Key, Msg, _State, _Hops) -> Msg.
```

Wraps a message send via `dht_node_lookup:lookup/4` if needed. This routing algorithm does not need callbacks when finishing the lookup, so it does not need to wrap the message.

File `rt_chord.erl`:

```
513 %% @doc Unwrap lookup messages. This is a noop in Chord.
514 -spec unwrap_message(Msg::comm:message(), State::dht_node_state:state()) -> comm:message().
515 unwrap_message(Msg, _State) -> Msg.
```

Unwraps a message previously wrapped with `rt_chord:wrap_message/1`. As that function does not wrap messages, `rt_chord:unwrap_message/2` doesn't have to do anything as well.

9.4. Local Datastore

9.5. Cyclon

9.6. Vivaldi Coordinates

9.7. Estimated Global Information (Gossiping)

9.8. Load Balancing

9.9. Broadcast Trees

10. Transactions in Scalaris

10.1. The Paxos Module

10.2. Transactions using Paxos Commit

10.3. Applying the Tx-Modules to replicated DHTs

Introduces transaction processing on top of a Overlay

11. How a node joins the system

Description is based on SVN revision r1370.

After starting a new Scalaris-System as described in Section 3.2.1 on page 17, ten additional local nodes can be started by typing `api_vm:add_nodes(10)` in the Erlang-Shell that is opened during startup ¹.

```
scalaris> ./bin/firstnode.sh
[...]
```

```
(firstnode@csr-pc9)1> api_vm:add_nodes(10)
```

In the following we will trace what this function does in order to add additional nodes to the system. The function `api_vm:add_nodes(pos_integer())` is defined as follows.

File `api_vm.erl`:

```
65 %% @doc Adds Number Scalaris nodes to this VM.
66 -spec add_nodes(non_neg_integer()) -> {[pid_groups:groupname()], [{error, term()}]}.
67 add_nodes(Number) when is_integer(Number) andalso Number >= 0 ->
68     Result = {Ok, _Failed} = admin:add_nodes(Number),
69     % at least wait for the successful nodes to have joined, i.e. left the join phases
70     util:wait_for(
71         fun() ->
72             DhtModule = config:read(dht_node),
73             NotReady = [Name || Name <- Ok,
74                             not DhtModule:is_alive(
75                                 gen_component:get_state(
76                                     pid_groups:pid_of(Name, dht_node)))]],
77             [] == NotReady
78         end),
79     Result.
```

It uses the `admin:add_nodes/1` function to actually add the given number of nodes and then waits for all nodes to successfully complete their join phases.

File `admin.erl`:

```
45 % @doc add new Scalaris nodes on the local node
46 -spec add_node_at_id(?RT:key()) -> pid_groups:groupname() | {error, term()}.
47 add_node_at_id(Id) ->
48     add_node([{{dht_node, id}, Id}, {skip_psv_lb}, {add_node}]).
49
50 -spec add_node([tuple()]) -> pid_groups:groupname() | {error, term()}.
51 add_node(Options) ->
52     DhtNodeId = randoms:getRandomString(),
53     Group = pid_groups:new("dht_node_"),
54     Desc = sup:supervisor_desc(
55         DhtNodeId, sup_dht_node, start_link,
56         [{Group,
57           [{my_sup_dht_node_id, DhtNodeId}, {add_node} | Options]}]),
58     Sup = erlang:whereis(main_sup),
59     case sup:start_sup_as_child([" +"], Sup, Desc) of
60         {ok, _Child, Group} ->
61             DhtNodePid = pid_groups:pid_of(Group, dht_node),
```

¹Increase the log level to info to get more detailed startup logs. See Section 3.1.1 on page 16

```

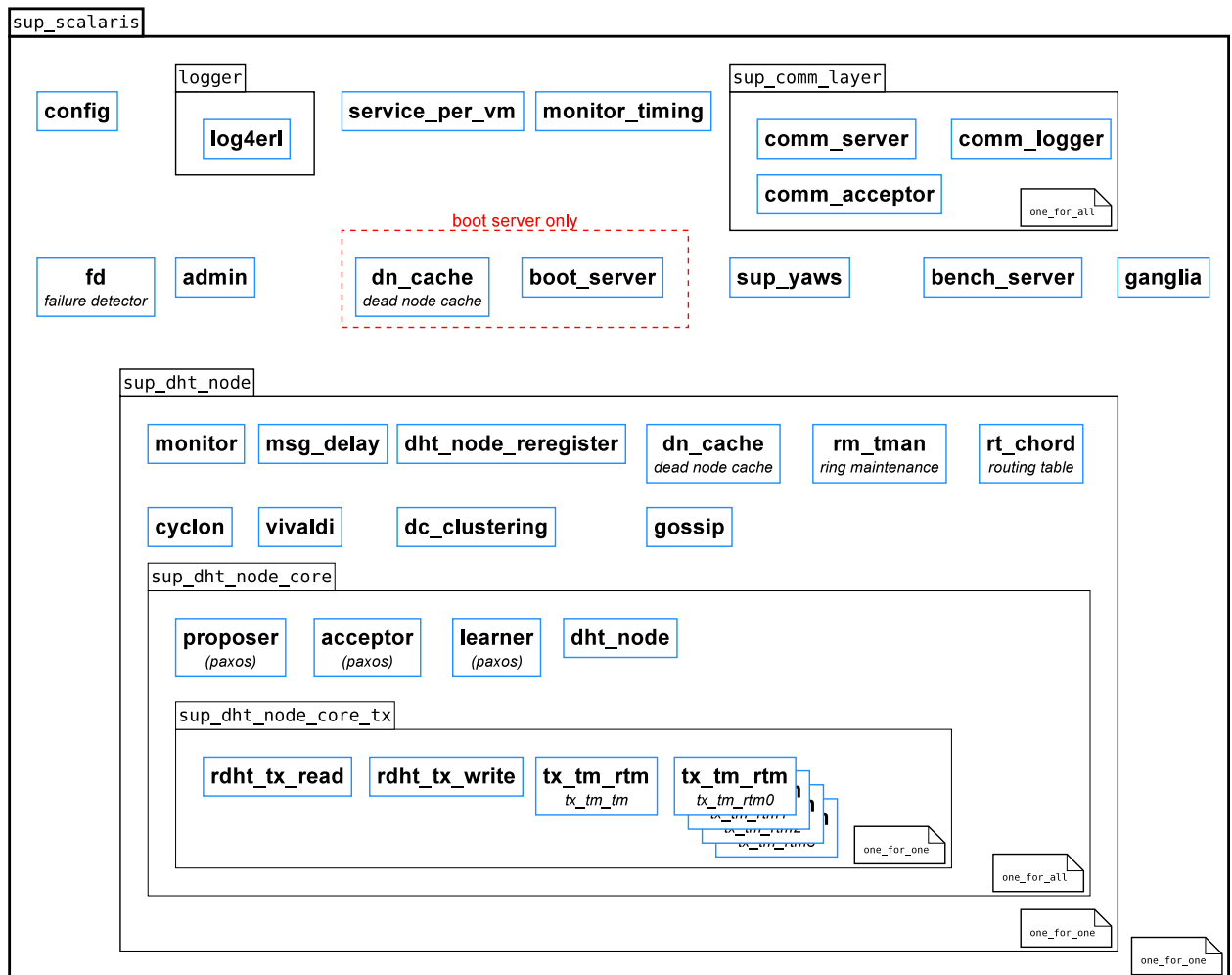
62         comm:send_local(DhtNodePid, {join, start}),
63         Group;
64     {error, already_present}      -> add_node(Options); % try again, different Id
65     {error, {already_started, _}} -> add_node(Options); % try again, different Id
66     {error, _Error} = X          -> X
67 end.
68
69 -spec add_nodes(non_neg_integer()) -> {[pid_groups:groupname()], [{error, term()}]}.
70 add_nodes(0) -> {[], []];
71 add_nodes(Count) ->
72     Results = [add_node([]) || _X <- lists:seq(1, Count)],
73     lists:partition(fun(E) -> not is_tuple(E) end, Results).

```

The function `admin:add_nodes/1` calls `admin:add_node([])` `Count` times. This function starts a new child with the given options for the main supervisor `main_sup`. In particular, it sets a random ID that is passed to the new node as its suggested ID to join at. To actually perform the start, the function `sup_dht_node:start_link/1` is called by the Erlang supervisor mechanism. For more details on the OTP supervisor mechanism see Chapter 18 of the Erlang book [1] or the online documentation at <http://www.erlang.org/doc/man/supervisor.html>.

11.1. Supervisor-tree of a Scalaris node

When a new Erlang VM with a Scalaris node is started, a `sup_scalaris` supervisor is started that creates further workers and supervisors according to the following scheme (processes starting order: left to right, top to bottom):



When new nodes are started using `admin:add_node/1`, only new `sup_dht_node` supervisors are started.

11.2. Starting the `sup_dht_node` supervisor and general processes of a node

Starting supervisors is a two step process: a call to `supervisor:start_link/2,3`, e.g. from a custom supervisor's own `start_link` method, will start the supervisor process. It will then call `Module:init/1` to find out about the restart strategy, maximum restart frequency and child processes. Note that `supervisor:start_link/2,3` will not return until `Module:init/1` has returned and all child processes have been started.

Let's have a look at `sup_dht_node:init/1`, the 'DHT node supervisor'.

File `sup_dht_node.erl`:

```

42 -spec init([pid_groups:groupname(), tuple()])
43     -> {ok, {one_for_one, MaxRetries::pos_integer(),
44             PeriodInSeconds::pos_integer()}, []}.
45 init([DHTNodeGroup, _Options] = X) ->
46     pid_groups:join_as(DHTNodeGroup, ?MODULE),
47     mgmt_server:connect(),
48     supspec(X).

```

The return value of the `init/1` function specifies the child processes of the supervisor and how to start them. Here, we define a list of processes to be observed by a `one_for_one` supervisor. The processes are: `Monitor`, `Delayer`, `Reregister`, `DeadNodeCache`, `RingMaintenance`, `RoutingTable`, `Cyclon`, `Vivaldi`, `DC_Clustering`, `Gossip` and a `SupDHTNodeCore_AND` process in this order.

The term `{one_for_one, 10, 1}` specifies that the supervisor should try 10 times to restart each process before giving up. `one_for_one` supervision means, that if a single process stops, only that process is restarted. The other processes run independently.

When the `sup_dht_node:init/1` is finished the supervisor module starts all the defined processes by calling the functions that were defined in the returned list.

For a join of a new node, we are only interested in the starting of the `SupDHTNodeCore_AND` process here. At that point in time, all other defined processes are already started and running.

11.3. Starting the `sup_dht_node_core` supervisor with a peer and some paxos processes

Like any other supervisor the `sup_dht_node_core` supervisor calls its `sup_dht_node_core:init/1` function:

File `sup_dht_node_core.erl`:

```
41 -spec init([pid_groups:groupname(), Options::tuple()]) ->
42     {ok, {one_for_all, MaxRetries::pos_integer(),
43           PeriodInSeconds::pos_integer(),
44           [ProcessDescr::supervisor:child_spec()]} }.
45 init([DHTNodeGroup, _Options] = X) ->
46     pid_groups:join_as(DHTNodeGroup, ?MODULE),
47     supspec(X).
```

It defines five processes, that have to be observed using a `one_for_all`-supervisor, which means, that if one fails, all have to be restarted. The `dht_node` module implements the main component of a full `Scalaris` node which glues together all the other processes. Its `dht_node:start_link/2` function will get the following parameters: (a) the processes' group that is used with the `pid_groups` module and (b) a list of options for the `dht_node`. The process group name was calculated a bit earlier in the code. *Exercise: Try to find where.*

File `dht_node.erl`:

```
529 %% @doc spawns a scalaris node, called by the scalaris supervisor process
530 -spec start_link(pid_groups:groupname(), tuple()) -> {ok, pid()}.
531 start_link(DHTNodeGroup, Options) ->
532     gen_component:start_link(?MODULE, fun ?MODULE:on/2, Options,
533                               [{pid_groups_join_as, DHTNodeGroup, dht_node},
534                               {wait_for_init},
535                               {spawn_opts, [{fullsweep_after, 0},
536                                             {min_heap_size, 131071}]}]).
```

Like many other modules, the `dht_node` module implements the `gen_component` behaviour. This behaviour was developed by us to enable us to write code which is similar in syntax and semantics to the examples in [3]. Similar to the supervisor behaviour, a module implementing this behaviour has to provide an `init/1` function, but here it is used to initialize the state of the component. This function is described in the next section.

Note: `?MODULE` is a predefined Erlang macro, which expands to the module name, the code belongs to (here: `dht_node`).

11.4. Initializing a dht_node-process

File dht_node.erl:

```
481 %% @doc joins this node in the ring and calls the main loop
482 -spec init(Options::[tuple()])
483     -> dht_node_state:state() |
484     {'$gen_component', [{on_handler, Handler::gen_component:handler()}], State::dht_node_join:
485 init(Options) ->
486     {my_sup_dht_node_id, MySupDhtNode} = lists:keyfind(my_sup_dht_node_id, 1, Options),
487     erlang:put(my_sup_dht_node_id, MySupDhtNode),
488     % start trigger here to prevent infection when tracing e.g. node joins
489     % (otherwise the trigger would be started at the end of the join and thus
490     % be infected forever)
491     % NOTE: any trigger started here, needs an exception for queuing messages
492     %       in dht_node_join to prevent infection with msg_queue:send/1!
493     rm_loop:init_first(),
494     dht_node_move:send_trigger(),
495
496     Recover = config:read(start_type) ==> recover,
497     case {is_first(Options), config:read(leases), Recover, is_add_nodes(Options)} of
498     {_, true, true, false} ->
499         % we are recovering
500         dht_node_join_recover:join(Options);
501     {true, true, false, _} ->
502         msg_delay:send_trigger(1, {l_on_cseq, renew_leases}),
503         Id = l_on_cseq:id(intervals:all()),
504         TmpState = dht_node_join:join_as_first(Id, 0, Options),
505         %% we have to inject the first lease by hand, as otherwise
506         %% no routing will work.
507         l_on_cseq:add_first_lease_to_db(Id, TmpState);
508     {false, true, _, true} ->
509         msg_delay:send_trigger(1, {l_on_cseq, renew_leases}),
510         % get my ID (if set, otherwise chose a random ID):
511         Id = case lists:keyfind({dht_node, id}, 1, Options) of
512             {{dht_node, id}, IdX} -> IdX;
513             _ -> ?RT:get_random_node_id()
514         end,
515         dht_node_join:join_as_other(Id, 0, Options);
516     {IsFirst, _, _, _} ->
517         % get my ID (if set, otherwise chose a random ID):
518         Id = case lists:keyfind({dht_node, id}, 1, Options) of
519             {{dht_node, id}, IdX} -> IdX;
520             _ -> ?RT:get_random_node_id()
521         end,
522         if IsFirst -> dht_node_join:join_as_first(Id, 0, Options);
523         true      -> dht_node_join:join_as_other(Id, 0, Options)
524         end
525     end.
```

The `gen_component` behaviour registers the `dht_node` in the process dictionary. Formerly, the process had to do this itself, but we moved this code into the behaviour. If an ID was given to `dht_node:init/1` function as a `{{dht_node, id}, KEY}` tuple, the given `Id` will be used. Otherwise a random key is generated. Depending on whether the node is the first inside a VM marked as first or not, the according function in `dht_node_join` is called. Also the pid of the node's supervisor is kept for future reference.

11.5. Actually joining the ring

After retrieving its identifier, the node starts the join protocol which processes the appropriate messages calling `dht_node_join:process_join_state(Message, State)`. On the existing node, join messages will be processed by `dht_node_join:process_join_msg(Message, State)`.

11.5.1. A single node joining an empty ring

File `dht_node_join.erl`:

```
108 -spec join_as_first(Id::?RT:key(), IdVersion::non_neg_integer(), Options::[tuple()])
109     -> dht_node_state:state().
110 join_as_first(Id, IdVersion, _Options) ->
111     comm:init_and_wait_for_valid_IP(),
112     log:log(info, "[ Node ~w ] joining as first: (~.0p, ~.0p)",
113         [self(), Id, IdVersion]),
114     Me = node:new(comm:this(), Id, IdVersion),
115     % join complete, State is the first "State"
116     finish_join(Me, Me, Me, db_dht:new(), msg_queue:new(), []).
```

If the ring is empty, the joining node will be the only node in the ring and will thus be responsible for the whole key space. It will trigger all known nodes to initialize the comm layer and then finish the join. `dht_node_join:finish_join/5` just creates a new state for a Scalaris node consisting of the given parameters (the node as itself, its predecessor and successor, an empty database and the queued messages that arrived during the join). It then activates all dependent processes and creates a routing table from this information.

The `dht_node_state:state()` type is defined in

File `dht_node_state.erl`:

```
82 -record(state, {rt           = ?required(state, rt)           :: ?RT:external_rt(),
83                 rm_state    = ?required(state, rm_state)    :: rm_loop:state(),
84                 join_time   = ?required(state, join_time)   :: erlang_timestamp(),
85                 db          = ?required(state, db)           :: db_dht:db(),
86                 tx_tp_db    = ?required(state, tx_tp_db)     :: any(),
87                 proposer    = ?required(state, proposer)    :: pid(),
88                 % slide with pred (must not overlap with 'slide with succ!'):
89                 slide_pred  = null :: slide_op:slide_op() | null,
90                 % slide with succ (must not overlap with 'slide with pred!'):
91                 slide_succ  = null :: slide_op:slide_op() | null,
92                 % additional range to respond to during a move:
93                 db_range    = [] :: [{interval:interval(), slide_op:id()}],
94                 monitor_proc = ?required(state, monitor_proc) :: pid(),
95                 prbr_kv_db  = ?required(state, prbr_kv_db)   :: prbr:state(),
96                 txid_db1    = ?required(state, txid_db1)    :: prbr:state(),
97                 txid_db2    = ?required(state, txid_db2)    :: prbr:state(),
98                 txid_db3    = ?required(state, txid_db3)    :: prbr:state(),
99                 txid_db4    = ?required(state, txid_db4)    :: prbr:state(),
100                 lease_db1   = ?required(state, lease_db1)   :: prbr:state(),
101                 lease_db2   = ?required(state, lease_db2)   :: prbr:state(),
102                 lease_db3   = ?required(state, lease_db3)   :: prbr:state(),
103                 lease_db4   = ?required(state, lease_db4)   :: prbr:state(),
104                 lease_list  = ?required(state, lease_list)  :: lease_list:lease_list(),
105                 snapshot_state = null :: snapshot_state:snapshot_state() | null,
106                 mr_state    = ?required(state, mr_state)    :: orddict:orddict(),
107                 mr_master_state = ?required(state, mr_master_state) :: orddict:orddict()
108             }).
109 -opaque state() :: #state{}
```

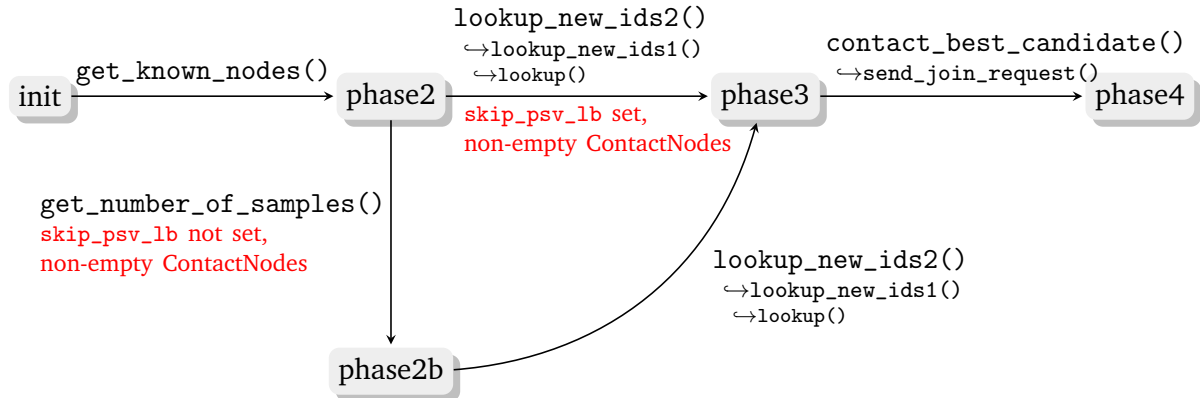
11.5.2. A single node joining an existing (non-empty) ring

If a node joins an existing ring, its join protocol will step through the following four phases:

- **phase2** finding nodes to contact with the help of the configured `known_hosts`
- **phase2b** getting the number of Ids to sample (may be skipped)
- **phase3** lookup nodes responsible for all sampled Ids

- **phase4** joining a selected node and setting up item movements

The following figure shows a (non-exhaustive) overview of the transitions between the phases in the normal case. We will go through these step by step and discuss what happens if errors occur.



At first all nodes set in the `known_hosts` configuration parameter are contacted. Their responses are then handled in phase 2. In order to separate the join state from the ordinary `dht_node` state, the `gen_component` is instructed to use the `dht_node:on_join/2` message handler which delegates every message to `dht_node_join:process_join_state/2`.

File `dht_node_join.erl`:

```

120 -spec join_as_other(Id::?RT:key(), IdVersion::non_neg_integer(), Options::[tuple()])
121     -> {'$gen_component', [{on_handler, Handler::gen_component:handler()}],
122         State::{join, phase2(), msg_queue:msg_queue()}}.
123 join_as_other(Id, IdVersion, Options) ->
124     comm:init_and_wait_for_valid_IP(),
125     log:log(info, "[ Node ~w ] joining, trying ID: (~.0p, ~.0p)",
126         [self(), Id, IdVersion]),
127     JoinUUID = uid:get_pids_uid(),
128     gen_component:change_handler(
129         {join, {phase1, JoinUUID, Options, IdVersion, [], [Id], []},
130             msg_queue:new()}),
131     fun ?MODULE:process_join_state/2).

```

Phase 2 and 2b

Phase 2 collects all `dht_node` processes inside the contacted VMs. It therefore mainly processes `get_dht_nodes_response` messages and integrates all received nodes into the list of available connections. The next step depends on whether the `{skip_psv_lb}` option for skipping any passive load balancing algorithm has been given to the `dht_node` or not. If it is present, the node will only use the ID that has been initially passed to `dht_node_join:join_as_other/3`, issue a lookup for the responsible node and move to phase 3. Otherwise, the passive load balancing's `lb_psv_*:-get_number_of_samples/1` method will be called asking for the number of IDs to sample. Its answer will be processed in phase 2b.

`get_dht_nodes_response` messages arriving in phase 2b or later will be processed anyway and received `dht_node` processes will be integrated into the connections. These phases' operations will not be interrupted and nothing else is changed though.

File `dht_node_join.erl`:

```

167 % in phase 2 add the nodes and do lookups with them / get number of samples
168 process_join_state({get_dht_nodes_response, Nodes} = _Msg,
169     {join, JoinState, QueuedMessages})

```

```

170 when element(1, JoinState) == phase2 ->
171     ?TRACE_JOIN1(_Msg, JoinState),
172     JoinOptions = get_join_options(JoinState),
173     %% additional nodes required when firstnode jumps and he's the only known host
174     DhtNodes = Nodes ++ proplists:get_value(bootstrap_nodes, JoinOptions, []),
175     Connections = [{null, Node} || Node <- DhtNodes, Node /= comm:this()],
176     JoinState1 = add_connections(Connections, JoinState, back),
177     NewJoinState = phase2_next_step(JoinState1, Connections),
178     ?TRACE_JOIN_STATE(NewJoinState),
179     {join, NewJoinState, QueuedMessages};
180
181 % in all other phases, just add the provided nodes:
182 process_join_state({get_dht_nodes_response, Nodes} = _Msg,
183     {join, JoinState, QueuedMessages})
184 when element(1, JoinState) == phase2b orelse
185     element(1, JoinState) == phase3 orelse
186     element(1, JoinState) == phase4 ->
187     ?TRACE_JOIN1(_Msg, JoinState),
188     Connections = [{null, Node} || Node <- Nodes, Node /= comm:this()],
189     JoinState1 = add_connections(Connections, JoinState, back),
190     ?TRACE_JOIN_STATE(NewJoinState1),
191     {join, JoinState1, QueuedMessages};

```

Phase 2b will handle `get_number_of_samples` messages from the passive load balance algorithm. Once received, new (unique) IDs will be sampled randomly so that the total number of join candidates (selected IDs together with fully processed candidates from further phases) is at least as high as the given number of samples. Afterwards, lookups will be created for all previous IDs as well as the new ones and the node will move to phase 3.

File `dht_node_join.erl`:

```

217 % note: although this message was send in phase2b, also accept message in
218 % phase2, e.g. messages arriving from previous calls
219 process_join_state({join, get_number_of_samples, Samples, Conn} = _Msg,
220     {join, JoinState, QueuedMessages})
221 when element(1, JoinState) == phase2 orelse
222     element(1, JoinState) == phase2b ->
223     ?TRACE_JOIN1(_Msg, JoinState),
224     % prefer node that send get_number_of_samples as first contact node
225     JoinState1 = reset_connection(Conn, JoinState),
226     % (re-)issue lookups for all existing IDs and
227     % create additional samples, if required
228     NewJoinState = lookup_new_ids2(Samples, JoinState1),
229     ?TRACE_JOIN_STATE(NewJoinState),
230     {join, NewJoinState, QueuedMessages};
231
232 % ignore message arriving in other phases:
233 process_join_state({join, get_number_of_samples, _Samples, Conn} = _Msg,
234     {join, JoinState, QueuedMessages}) ->
235     ?TRACE_JOIN1(_Msg, JoinState),
236     NewJoinState = reset_connection(Conn, JoinState),
237     ?TRACE_JOIN_STATE(NewJoinState),
238     {join, NewJoinState, QueuedMessages};

```

Lookups will make Scalaris find the node currently responsible for a given ID and send a request to simulate a join to this node, i.e. a `get_candidate` message. Note that during such an operation, the joining node would become the existing node's predecessor. The simulation will be delegated to the passive load balance algorithm the joining node requested, as set by the `join_lb_psv` configuration parameter.

File `dht_node_join.erl`:

```

574 process_join_msg({join, get_candidate, Source_PID, Key, LbPsv, Conn} = _Msg, State) ->
575     ?TRACE1(_Msg, State),
576     call_lb_psv(LbPsv, create_join, [State, Key, Source_PID, Conn]);

```

Phase 3

The result of the simulation will be send in a `get_candidate_response` message and will be processed in phase 3 of the joining node. It will be integrated into the list of processed candidates. If there are no more IDs left to process, the best among them will be contacted. Otherwise further `get_candidate_response` messages will be awaited. Such messages will also be processed in the other phases where the candidate will be simply added to the list.

File `dht_node_join.erl`:

```
270 process_join_state({join, get_candidate_response, OrigJoinId, Candidate, Conn} = _Msg,
271                   {join, JoinState, QueuedMessages})
272   when element(1, JoinState) == phase3 ->
273     ?TRACE_JOIN1(_Msg, JoinState),
274     JoinState0 = reset_connection(Conn, JoinState),
275     JoinState1 = remove_join_id(OrigJoinId, JoinState0),
276     JoinState2 = integrate_candidate(Candidate, JoinState1, front),
277     NewJoinState =
278       case get_join_ids(JoinState2) of
279         [] -> % no more join ids to look up -> join with the best:
280           contact_best_candidate(JoinState2);
281         [_|_] -> % still some unprocessed join ids -> wait
282           JoinState2
283       end,
284     ?TRACE_JOIN_STATE(NewJoinState),
285     {join, NewJoinState, QueuedMessages};
286
287 % In phase 2 or 2b, also add the candidate but do not continue.
288 % In phase 4, add the candidate to the end of the candidates as they are sorted
289 % and the join with the first has already started (use this candidate as backup
290 % if the join fails). Do not start a new join.
291 process_join_state({join, get_candidate_response, OrigJoinId, Candidate, Conn} = _Msg,
292                   {join, JoinState, QueuedMessages})
293   when element(1, JoinState) == phase2 orelse
294     element(1, JoinState) == phase2b orelse
295     element(1, JoinState) == phase4 ->
296     ?TRACE_JOIN1(_Msg, JoinState),
297     JoinState0 = reset_connection(Conn, JoinState),
298     JoinState1 = remove_join_id(OrigJoinId, JoinState0),
299     JoinState2 = case get_phase(JoinState1) of
300                   phase4 -> integrate_candidate(Candidate, JoinState1, back);
301                   _      -> integrate_candidate(Candidate, JoinState1, front)
302                 end,
303     ?TRACE_JOIN_STATE(JoinState2),
304     {join, JoinState2, QueuedMessages};
```

If `dht_node_join:contact_best_candidate/1` is called and candidates are available (there should be at this stage!), it will sort the candidates by using the passive load balance algorithm, send a `join_request` message and continue with phase 4.

File `dht_node_join.erl`:

```
876 %% @doc Contacts the best candidate among all stored candidates and sends a
877 %%      join_request (Timeouts = 0).
878 -spec contact_best_candidate(JoinState::phase_2_4())
879       -> phase2() | phase2b() | phase4().
880 contact_best_candidate(JoinState) ->
881   JoinState1 = sort_candidates(JoinState),
882   send_join_request(JoinState1, 0).
```

File `dht_node_join.erl`:

```
886 %% @doc Sends a join request to the first candidate. Timeouts is the number of
887 %%      join_request_timeout messages previously received.
888 %%      PreCond: the id has been set to the ID to join at and has been updated
```

```

889 %%                in JoinState.
890 -spec send_join_request(JoinState::phase_2_4(), Timeouts::non_neg_integer())
891     -> phase2() | phase2b() | phase4().
892 send_join_request(JoinState, Timeouts) ->
893     case get_candidates(JoinState) of
894     [] -> % no candidates -> start over (can happen, e.g. when join candidates are busy):
895         start_over(JoinState);
896     [BestCand | _] ->
897         Id = node_details:get(lb_op:get(BestCand, n1_new), new_key),
898         IdVersion = get_id_version(JoinState),
899         NewSucc = node_details:get(lb_op:get(BestCand, n1succ_new), node),
900         Me = node:new(comm:this(), Id, IdVersion),
901         CandId = lb_op:get(BestCand, id),
902         MyMTE = case dht_node_move:use_incremental_slides() of
903             true -> dht_node_move:get_max_transport_entries();
904             false -> unknown
905         end,
906         Msg = {join, join_request, Me, CandId, MyMTE},
907         ?TRACE_SEND(node:pidX(NewSucc), Msg),
908         comm:send(node:pidX(NewSucc), Msg),
909         msg_delay:send_local(
910             get_join_request_timeout() div 1000, self(),
911             {join, join_request_timeout, Timeouts, CandId, get_join_uuid(JoinState)}),
912         set_phase(phase4, JoinState)
913     end.

```

The join_request message will be received by the existing node which will set up a slide operation with the new node. If it is not responsible for the key (anymore), it will deny the request and reply with a {join, join_response, not_responsible, Node} message. If it is responsible for the ID and is not participating in a slide with its current predecessor, it will set up a slide with the joining node:

File dht_node_join.erl:

```

580 process_join_msg({join, join_request, NewPred, CandId, MaxTransportEntries} = _Msg, State)
581     when (not is_atom(NewPred)) -> % avoid confusion with not_responsible message
582     ?TRACE1(_Msg, State),
583     TargetId = node:id(NewPred),
584     JoinType = {join, 'send'},
585     MyNode = dht_node_state:get(State, node),
586     Command = dht_node_move:check_setup_slide_not_found(
587         State, JoinType, MyNode, NewPred, TargetId),
588     case Command of
589     {ok, JoinType} ->
590         MoveFullId = uid:get_global_uid(),
591         State1 = dht_node_move:exec_setup_slide_not_found(
592             Command, State, MoveFullId, NewPred, TargetId, join,
593             MaxTransportEntries, null, nomsg, {none}, false),
594         % set up slide, now send join_response:
595         MyOldPred = dht_node_state:get(State1, pred),
596         % no need to tell the ring maintenance -> the other node will trigger an update
597         % also this is better in case the other node dies during the join
598         %% rm_loop:notify_new_pred(comm:this(), NewPred),
599         SlideOp = dht_node_state:get(State1, slide_pred),
600         Msg = {join, join_response, MyNode, MyOldPred, MoveFullId, CandId,
601             slide_op:get_target_id(SlideOp), slide_op:get_next_op(SlideOp)},
602         dht_node_move:send(node:pidX(NewPred), Msg, MoveFullId),
603         State1;
604     {abort, ongoing_slide, JoinType} ->
605         ?TRACE("[ ~.0p ]~n rejecting join_request from ~.0p due to a running slide~n",
606             [self(), NewPred]),
607         ?TRACE_SEND(node:pidX(NewPred), {join, join_response, busy, CandId}),
608         comm:send(node:pidX(NewPred), {join, join_response, busy, CandId}),
609         State;
610     {abort, _Reason, JoinType} -> % all other errors:
611         ?TRACE("~p", [Command]),
612         ?TRACE_SEND(node:pidX(NewPred),
613             {join, join_response, not_responsible, CandId}),
614         comm:send(node:pidX(NewPred),

```

```

615         {join, join_response, not_responsible, CandId}},
616         State
617     end;

```

Phase 4

The joining node will receive the `join_response` message in phase 4 of the join protocol. If everything is ok, it will notify its ring maintenance process that it enters the ring, start all required processes and join the slide operation set up by the existing node in order to receive some of its data.

If the join candidate's node is not responsible for the candidate's ID anymore or the candidate's ID already exists, the next candidate is contacted until no further candidates are available and the join protocol starts over using `dht_node_join:start_over/1`.

Note that the `join_response` message will actually be processed in any phase. Therefore, if messages arrive late, the join can be processed immediately and the rest of the join protocol does not need to be executed again.

File `dht_node_join.erl`:

```

343 process_join_state({join, join_response, Reason, CandId} = _Msg,
344                   {join, JoinState, QueuedMessages} = State)
345   when element(1, JoinState) == phase4 andalso
346     (Reason == not_responsible orelse Reason == busy) ->
347     ?TRACE_JOIN1(_Msg, JoinState),
348     % the node we contacted is not responsible for the selected key anymore
349     % -> try the next candidate, if the message is related to the current candidate
350     case get_candidates(JoinState) of
351     [] -> % no candidates -> should not happen in phase4!
352         log:log(error, "[ Node ~w ] empty candidate list in join phase 4, "
353                      "starting over", [self()]),
354         NewJoinState = start_over(JoinState),
355         ?TRACE_JOIN_STATE(NewJoinState),
356         {join, NewJoinState, QueuedMessages};
357     [Candidate | _Rest] ->
358         case lb_op:get(Candidate, id) == CandId of
359         false -> State; % unrelated/old message
360         _ ->
361             if Reason == not_responsible ->
362                 log:log(info,
363                      "[ Node ~w ] node contacted for join is not "
364                      "responsible for the selected ID (anymore), "
365                      "trying next candidate",
366                      [self()]);
367             Reason == busy ->
368                 log:log(info,
369                      "[ Node ~w ] node contacted for join is busy, "
370                      "trying next candidate",
371                      [self()])
372             end,
373             NewJoinState = try_next_candidate(JoinState),
374             ?TRACE_JOIN_STATE(NewJoinState),
375             {join, NewJoinState, QueuedMessages}
376         end
377     end;
378
379 % in other phases remove the candidate from the list (if it still exists):
380 process_join_state({join, join_response, Reason, CandId} = _Msg,
381                   {join, JoinState, QueuedMessages})
382   when (Reason == not_responsible orelse Reason == busy) ->
383     ?TRACE_JOIN1(_Msg, JoinState),
384     {join, remove_candidate(CandId, JoinState), QueuedMessages};
385
386 % note: accept (delayed) join_response messages in any phase

```

```

387 process_join_state({join, join_response, Succ, Pred, MoveId, CandId, TargetId, NextOp} = _Msg,
388                    {join, JoinState, QueuedMessages} = State) ->
389     ?TRACE_JOIN1(_Msg, JoinState),
390     % only act on related messages, i.e. messages from the current candidate
391     Phase = get_phase(JoinState),
392     State1 = case get_candidates(JoinState) of
393     [] when Phase == phase4 ->
394         % no candidates -> should not happen in phase4!
395         log:log(error, "[ Node ~w ] empty candidate list in join phase 4, "
396                     "starting over", [self()]),
397         reject_join_response(Succ, Pred, MoveId, CandId),
398         NewJoinState = start_over(JoinState),
399         ?TRACE_JOIN_STATE(NewJoinState),
400         {join, NewJoinState, QueuedMessages};
401     [] ->
402         % in all other phases, ignore the delayed join_response if no
403         % candidates exist
404         reject_join_response(Succ, Pred, MoveId, CandId),
405         State;
406     [Candidate | _Rest] ->
407         CandidateNode = node_details:get(lb_op:get(Candidate, n1succ_new), node),
408         CandidateNodeSame = node:same_process(CandidateNode, Succ),
409         case lb_op:get(Candidate, id) == CandId of
410         false ->
411             % ignore old/unrelated message
412             log:log(warn, "[ Node ~w ] ignoring old or unrelated "
413                     "join_response message", [self()]),
414             reject_join_response(Succ, Pred, MoveId, CandId),
415             State;
416         _ when not CandidateNodeSame ->
417             % id is correct but the node is not (should never happen!)
418             log:log(error, "[ Node ~w ] got join_response but the node "
419                     "changed, trying next candidate", [self()]),
420             reject_join_response(Succ, Pred, MoveId, CandId),
421             NewJoinState = try_next_candidate(JoinState),
422             ?TRACE_JOIN_STATE(NewJoinState),
423             {join, NewJoinState, QueuedMessages};
424         _ ->
425             MyId = TargetId,
426             MyIdVersion = get_id_version(JoinState),
427             case MyId == node:id(Succ) orelse MyId == node:id(Pred) of
428             true ->
429                 log:log(warn, "[ Node ~w ] chosen ID already exists, "
430                         "trying next candidate", [self()]),
431                 reject_join_response(Succ, Pred, MoveId, CandId),
432                 % note: can not keep Id, even if skip_psv_lb is set
433                 JoinState1 = remove_candidate_front(JoinState),
434                 NewJoinState = contact_best_candidate(JoinState1),
435                 ?TRACE_JOIN_STATE(NewJoinState),
436                 {join, NewJoinState, QueuedMessages};
437             _ ->
438                 ?TRACE("[ ~.0p ]~n joined MyId:~.0p, MyIdVersion:~.0p~n "
439                         "Succ: ~.0p~n Pred: ~.0p~n",
440                         [self(), MyId, MyIdVersion, Succ, Pred]),
441                 Me = node:new(comm:this(), MyId, MyIdVersion),
442                 log:log(info, "[ Node ~w ] joined between ~w and ~w",
443                         [self(), Pred, Succ]),
444                 rm_loop:notify_new_succ(node:pidX(Pred), Me),
445                 rm_loop:notify_new_pred(node:pidX(Succ), Me),
446
447                 JoinOptions = get_join_options(JoinState),
448
449                 finish_join_and_slide(Me, Pred, Succ, db_dht:new(),
450                                     QueuedMessages, MoveId, NextOp, JoinOptions)
451             end
452         end
453     end,
454     State1;

```


File dht_node_join.erl:

```

948 %% @doc Finishes the join and sends all queued messages.
949 -spec finish_join(Me::node:node_type(), Pred::node:node_type(),
950                 Succ::node:node_type(), DB::db_dht:db(),
951                 QueuedMessages::msg_queue:msg_queue(),
952                 JoinOptions::tuple())
953     -> dht_node_state:state().
954 finish_join(Me, Pred, Succ, DB, QueuedMessages, JoinOptions) ->
955     %% get old rt loop subscription table (if available)
956     MoveState = proplists:get_value(move_state, JoinOptions, []),
957     OldSubscrTable = proplists:get_value(subscr_table, MoveState, null),
958     RMState = rm_loop:init(Me, Pred, Succ, OldSubscrTable),
959     Neighbors = rm_loop:get_neighbors(RMState),
960     %% wait for the ring maintenance to initialize and tell us its table ID
961     rt_loop:activate(Neighbors),
962     if MoveState == [] ->
963         dc_clustering:activate(),
964         gossip:activate(Neighbors);
965     true -> ok
966 end,
967 dht_node_reregister:activate(),
968 msg_queue:send(QueuedMessages),
969 NewRT_ext = ?RT:empty_ext(Neighbors),
970 service_per_vm:register_dht_node(node:pidX(Me)),
971 dht_node_state:new(NewRT_ext, RMState, DB).
972
973 -spec reject_join_response(Succ::node:node_type(), Pred::node:node_type(),
974                          MoveFullId::slide_op:id(), CandId::lb_op:id()) -> ok.
975 reject_join_response(Succ, _Pred, MoveId, _CandId) ->
976     %% similar to dht_node_move:abort_slide/9 - keep message in sync!
977     Msg = {move, slide_abort, pred, MoveId, ongoing_slide},
978     ?TRACE_SEND(node:pidX(Succ), Msg),
979     dht_node_move:send_no_slide(node:pidX(Succ), Msg, 0).
980
981 %% @doc Finishes the join by setting up a slide operation to get the data from
982 %% the other node and sends all queued messages.
983 -spec finish_join_and_slide(Me::node:node_type(), Pred::node:node_type(),
984                          Succ::node:node_type(), DB::db_dht:db(),
985                          QueuedMessages::msg_queue:msg_queue(),
986                          MoveId::slide_op:id(), NextOp::slide_op:next_op(),
987                          JoinOptions::tuple())
988     -> {'$gen_component', [{on_handler, Handler::gen_component:handler()}]},
989     State::dht_node_state:state().
990 finish_join_and_slide(Me, Pred, Succ, DB, QueuedMessages, MoveId, NextOp, JoinOptions) ->
991     State = finish_join(Me, Pred, Succ, DB, QueuedMessages, JoinOptions),
992     {SourcePid, Tag} =
993         case lists:keyfind(jump, 1, JoinOptions) of
994             {jump, JumpTag, Pid} -> {Pid, JumpTag};
995             _ -> {null, join}
996         end,
997     State1 = dht_node_move:exec_setup_slide_not_found(
998         {ok, {join, 'rcv'}}, State, MoveId, Succ, node:id(Me), Tag,
999         unknown, SourcePid, nomsg, NextOp, false),
1000     gen_component:change_handler(State1, fun dht_node:on/2).

```

The macro ?RT maps to the configured routing algorithm. It is defined in include/scalaris.hrl. For further details on the routing see Chapter 9.3 on page 59.

Timeouts and other errors

The following table summarizes the timeout messages send during the join protocol on the joining node. It shows in which of the phases each of the messages is processed and describes (in short) what actions are taken. All of these messages are influenced by their respective config parameters, e.g. join_timeout parameter in the config files defines an overall timeout for the whole join operation. If it takes longer than join_timeout ms, a {join, timeout} will be send and processed

as given in this table.

	known_hosts_ _timeout	get_number_of_ _samples_ _timeout	lookup_ _timeout	join_request_ _timeout	timeout
phase2	get known nodes from configured VMs	ignore	ignore	ignore	
phase2b	ignore	remove contact node, re-start join → phase 2 or 2b	ignore	ignore	
phase3	ignore	ignore	remove contact node, lookup remaining IDs → phase 2 or 3	ignore	
phase3b	ignore	ignore	ignore	ignore	
phase4	ignore	ignore	ignore	timeouts < 3? ² → contact candidate otherwise: remove candidate no candidates left? → phase 2 or 2b otherwise: → contact next one → phase 3b or 4	re-start join → phase 2 or 2b

On the existing node, there is only one timeout message which is part of the join protocol: the `join_response_timeout`. It will be send when a slide operation is set up and if the timeout hits before the next message exchange, it will increase the slide operation's number of timeouts. The slide will be aborted if at least `join_response_timeouts` timeouts have been received. This parameter is set in the config file.

Misc. (all phases)

Note that join-related messages arriving in other phases than those handling them will be ignored. Any other messages during a `dht_node`'s join will be queued and re-send when the join is complete.

²set by the `join_request_timeouts` config parameter

12. How data is transferred (atomically)

Description is based on SVN revision r4750.

A data transfer from a node to one of its (two) neighbours is also called a *slide*. A slide operation is defined in the `slide_op` module, the protocol is mainly implemented in `dht_node_move`. Parts of the slide are dependent on the ring maintenance implementation and are split off into modules implementing the `slide_beh` behaviour.

Though the protocols are mainly symmetric, we distinguish between sending data to the predecessor and sending data to the successor, respectively. In the following protocol visualisations, arrows denote message exchanges, pseudo-code for operations that are being executed is put at the side of each time bar. Functions in green are those implemented in the `slide_beh` behaviour, if annotated with an arrow pointing to itself, this callback is asynchronous. During the protocol, the slide operation goes through several phases which are shown in black boxes.

In general, a slide consists of three steps:

1. set up slide
2. send data & start recording changes, i.e. delta
3. send delta & transfer responsibility

The latter two may be repeated to execute incremental slides which further reduce periods of unavailability. During this period, no node is responsible for the range to transfer and messages are thus delayed until the receiving node gains responsibility.

12.1. Sending data to the predecessor

12.1.1. Protocol



12.1.2. Callbacks

	slide_chord	slide_leases
← prepare_rcv_data	<i>nothing to do</i>	<i>nothing to do</i>
→ prepare_send_data	add DB range	<i>nothing to do</i>
← update_rcv_data	set MSG forward, change my ID	<i>nothing to do</i>
→ prepare_send_delta	wait until pred up-to-date, then: remove DB range	split own lease into two ranges, locally disable lease sent to pred
← finish_delta	remove MSG forward	<i>nothing to do</i>
→ finish_delta_ack	<i>nothing to do</i>	hand over the lease to pred, notify pred of owner change

12.2. Sending data to the successor

12.2.1. Protocol



12.2.2. Callbacks

	slide_chord	slide_leases
→ prepare_rcv_data	set MSG forward	<i>nothing to do</i>
← prepare_send_data	add DB range, change my ID	<i>nothing to do</i>
→ update_rcv_data	<i>nothing to do</i>	<i>nothing to do</i>
← prepare_send_delta	remove DB range	split own lease into two ranges, locally disable lease sent to succ
→ finish_delta	remove MSG forward, add DB range, wait until pred up-to-date then: remove DB range	<i>nothing to do</i>
← finish_delta_ack	<i>nothing to do</i>	hand over the lease to succ, notify succ of owner change

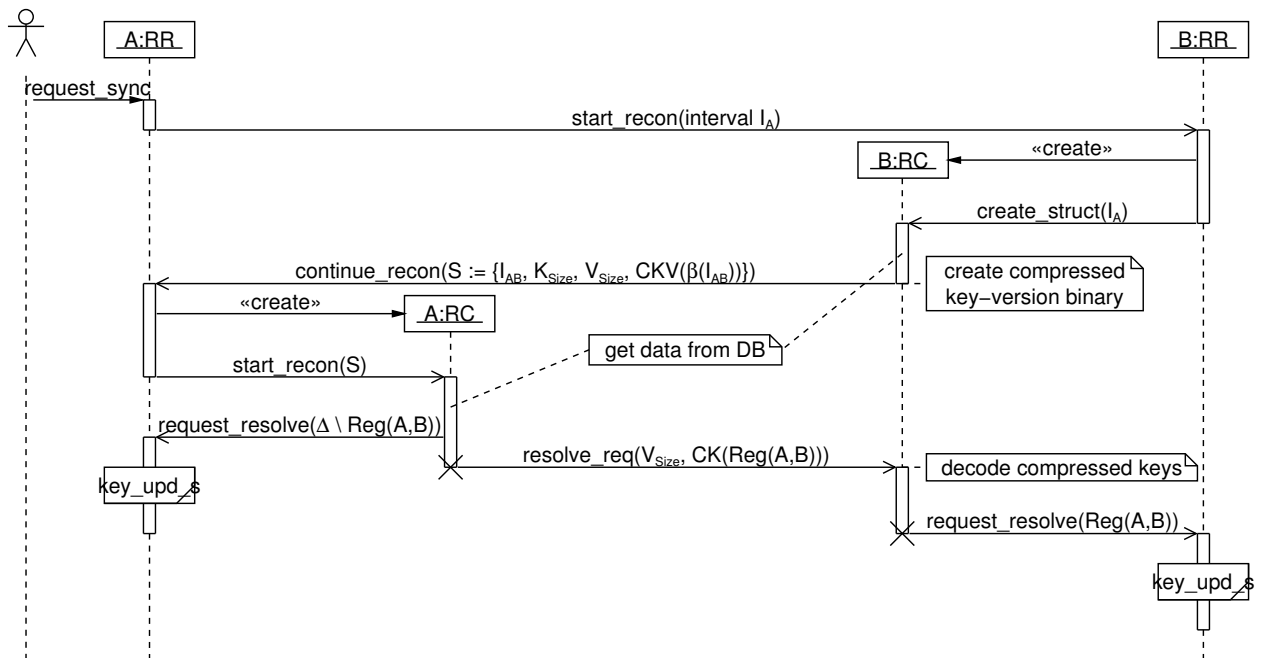
13. Replica Repair

Description is based on SVN revision r6345.

13.1. Replica Reconciliation - rr_recon

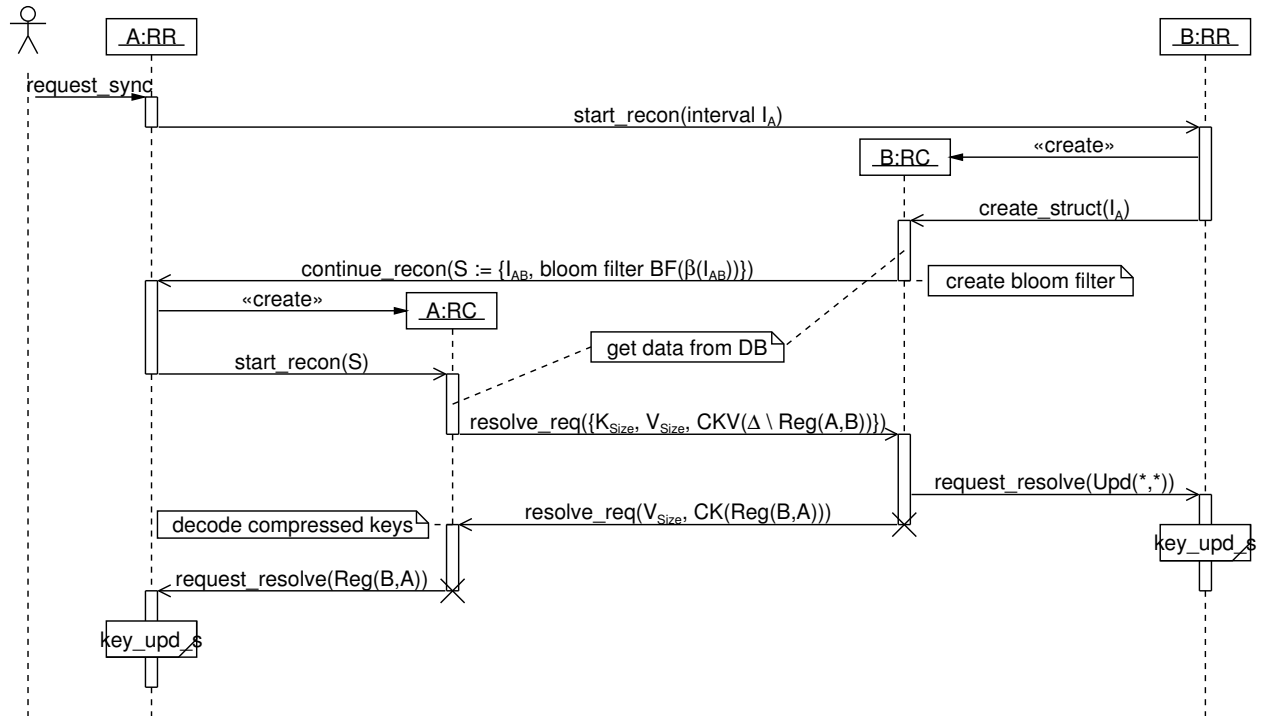
13.1.1. Trivial Replica Repair

Protocol



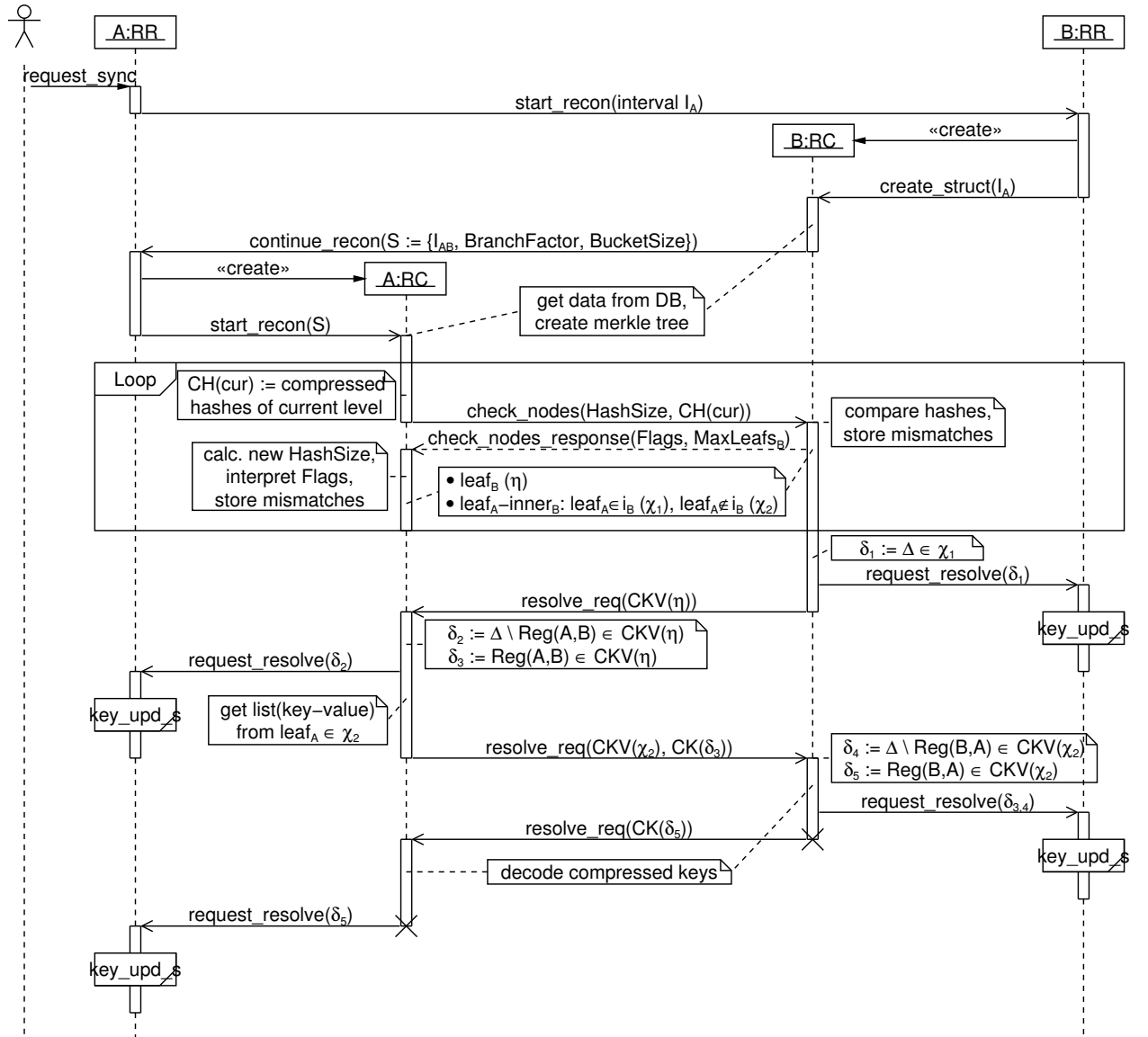
13.1.2. Replica Repair with Bloom Filters

Protocol



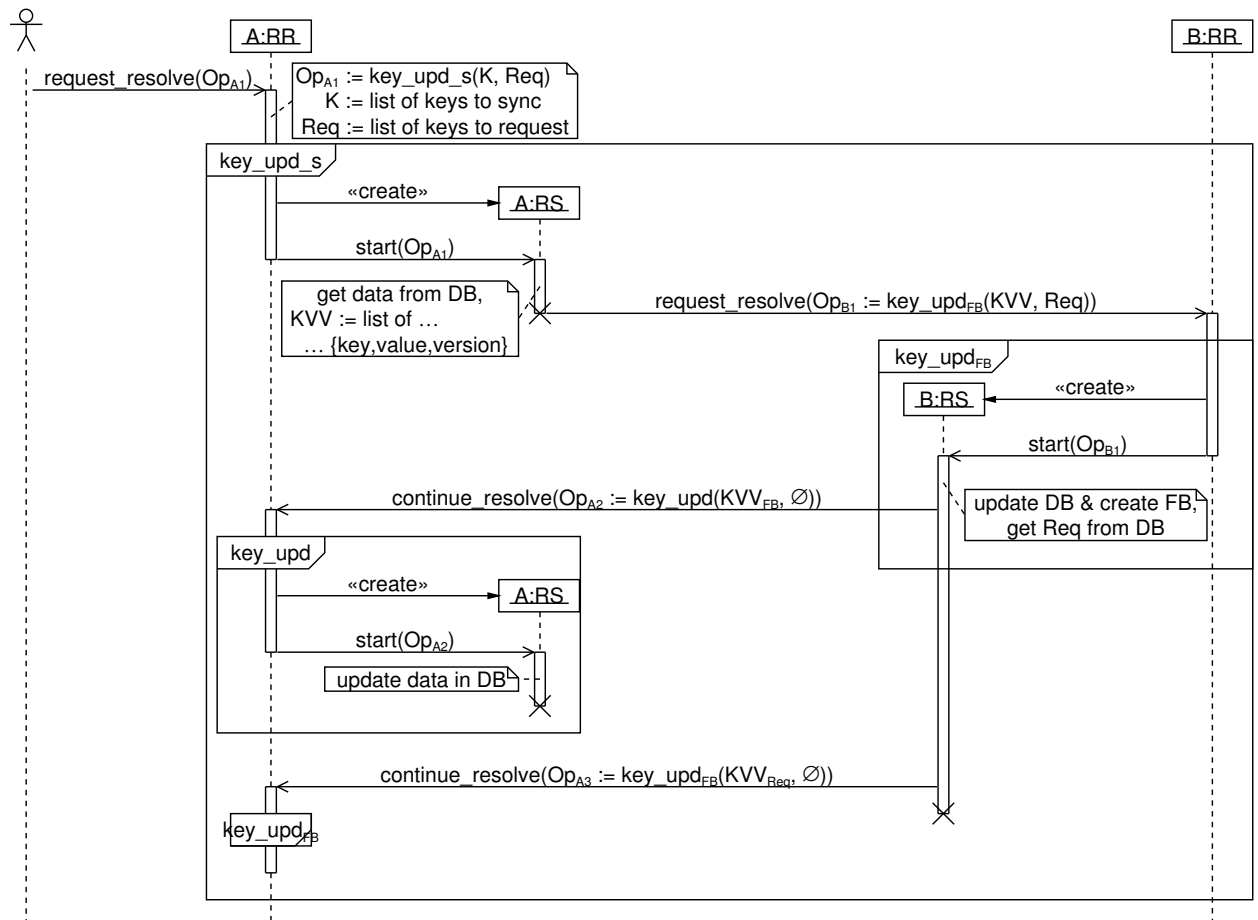
13.1.3. Replica Repair with Merkle Trees

Protocol



13.2. Resolve Replicas - rr_resolve

13.2.1. Updating a list of keys - key_upd



14. Directory Structure of the Source Code

The directory tree of Scalaris is structured as follows:

<code>bin</code>	contains shell scripts needed to work with Scalaris (e.g. start the management server, start a node, ...)
<code>contrib</code>	necessary third party packages (yaws and log4erl)
<code>doc</code>	generated Erlang documentation
<code>docroot</code>	root directory of the node's webserver
<code>ebin</code>	the compiled Erlang code (beam files)
<code>java-api</code>	a Java API to Scalaris
<code>python-api</code>	a Python 2 API to Scalaris
<code>python3-api</code>	a Python 3 API to Scalaris
<code>ruby-api</code>	a Ruby API to Scalaris
<code>log</code>	log files
<code>src</code>	contains the Scalaris source code
<code>include</code>	contains macros for the source code
<code>test</code>	unit tests for Scalaris
<code>user-dev-guide</code>	contains the sources for this document

15. Java API

For the Java API documentation, we refer the reader to the documentation generated by javadoc or doxygen. The following commands create the documentation:

```
%> cd java-api  
%> ant doc  
%> doxygen
```

The documentation can then be found in `java-api/doc/index.html` (javadoc) and `java-api/doc-doxygen/html/index.html` (doxygen).

The API is divided into four classes:

- `de.zib.scalarisc.Transaction` for (multiple) operations inside a transaction
- `de.zib.scalarisc.TransactionSingleOp` for single transactional operations
- `de.zib.scalarisc.ReplicatedDHT` for non-transactional (inconsistent) access to the replicated DHT items, e.g. deleting items
- `de.zib.scalarisc.PubSub` for topic-based publish/subscribe operations

Bibliography

- [1] Joe Armstrong. *Programming Erlang: Software for a Concurrent World*. Pragmatic Programmers, ISBN: 978-1-9343560-0-5, July 2007
- [2] Frank Dabek, Russ Cox, Frans Kaashoek, Robert Morris. *Vivaldi: A Decentralized Network Coordinate System*. ACM SIGCOMM 2004.
- [3] Rachid Guerraoui and Luis Rodrigues. *Introduction to Reliable Distributed Programming*. Springer-Verlag, 2006.
- [4] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek and Hari Balakrishnan. *Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications*. ACM SIGCOMM 2001, San Deigo, CA, August 2001, pp. 149-160. http://pdos.csail.mit.edu/papers/chord:sigcomm01/chord_sigcomm.pdf
- [5] Márk Jelasity, Alberto Montresor, Ozalp Babaoglu. *T-Man: Gossip-based fast overlay topology construction*. Computer Networks (CN) 53(13):2321-2339, 2009.
- [6] Hiroya Nagao, Kazuyuki Shudo. *Flexible routing tables: Designing routing algorithms for overlays based on a total order on a routing table set*. In: Peer-to-Peer Computing, IEEE, 2011.
- [7] F. Schintke, A. Reinefeld, S. Haridi, T. Schütt. *Enhanced Paxos Commit for Transactions on DHTs*. 10th IEEE/ACM Int. Conf. on Cluster, Cloud and Grid Computing, pp. 448-454, May 2010.
- [8] Spyros Voulgaris, Daniela Gavidia, Maarten van Steen. *CYCLON: Inexpensive Membership Management for Unstructured P2P Overlays*. J. Network Syst. Manage. 13(2): 2005.
- [9] Márk Jelasity, Alberto Montresor, Ozalp Babaoglu. *Gossip-based aggregation in large dynamic networks*. ACM Trans. Comput. Syst. 23(3), 219-252 (2005).

Index

?RT
 next_hop, 61
 update, 65

admin
 add_node, 77, 78
 add_nodes, 76, 77

api_tx, 58

api_vm
 add_nodes, 76

comm, 3, 48, 48
 get_msg_tag, 54
 send_to_group_member, 53

cs_api, 49

cyclon, 25, 57

dht_node, 63–65, 68, 74, 79, 82
 init, 80
 on_join, 82

dht_node_join, 80
 contact_best_candidate, 84, 84
 finish_join, 81, 87
 finish_join_and_slide, 87
 join_as_other, 82
 process_join_msg, 80
 process_join_state, 80, 82
 send_join_request, 84
 start_over, 86

dht_node_move, 90

dht_node_state
 state, 81

erlang
 exit, 50, 52
 now, 47
 send_after, 47

ets
 i, 47

fd, 56

gen_component, 3, 47, 48, 48–56
 bp_barrier, 54, 55
 bp_cont, 54, 55
 bp_del, 54, 55
 bp_set, 54
 bp_set_cond, 54
 bp_step, 54–56
 change_handler, 50, 52, 52
 get_state, 50
 is_gen_component, 51
 kill, 51, 53
 post_op, 53
 runnable, 55
 sleep, 53
 start, 50, 51
 start_link, 50–52

intervals
 in, 61

lb_psv_*
 get_number_of_samples, 82

monitor, 57, 57
 check_report, 57
 client_monitor_set_value, 58
 monitor_set_value, 58
 proc_check_timeslot, 58
 proc_set_value, 58

msg_delay, 47

paxos_SUITE, 53
 step_until_decide, 55

pdb, 56

pid_groups, 3, 48, 48, 51, 53, 79

randoms, 66

rm_beh, 66, 70

routing_table, 71

rr_recon, 4, 93

rr_resolve, 4, 96

rrd, 57, 57
 add, 57
 add_now, 57
 create, 57

rt_beh, 59

- check, 64
- check_config, 64
- dump, 64
- empty, 63
- empty_ext, 63
- export_rt_to_dht_node, 64
- filter_dead_node, 64
- get_random_node_id, 63
- get_replica_keys, 64
- get_size, 64
- handle_custom_message, 64
- hash_key, 63
- init_stabilize, 64
- n, 64
- next_hop, 63
- to_list, 64
- to_pid_list, 64
- unwrap_message, 64
- update, 64
- wrap_message, 64
- rt_chord, 69
 - empty, 70
 - empty_ext, 70
 - export_rt_to_dht_node, 73
 - filter_dead_node, 72
 - get_random_node_id, 70
 - get_replica_keys, 70
 - handle_custom_message, 71, 71
 - hash_key, 70
 - init_stabilize, 70
 - n, 70
 - next_hop, 70
 - stabilize, 71
 - unwrap_message, 74
 - update, 72
 - wrap_message, 74
- rt_loop, 64, 64, 72
- rt_simple, 65
 - dump, 67
 - empty, 66
 - empty_ext, 66
 - export_rt_to_dht_node, 68
 - filter_dead_node, 67
 - get_random_node_id, 66
 - get_replica_keys, 67
 - get_size, 67
 - handle_custom_message, 68
 - hash_key, 66
 - init_stabilize, 66
 - n, 67
 - next_hop, 66
 - to_list, 67
 - to_pid_list, 67
 - unwrap_message, 69
 - update, 66
 - wrap_message, 69
- slide_beh, 90
- slide_op, 90
- sup_dht_node
 - init, 78, 79
 - start_link, 77
- sup_dht_node_core, 79
- sup_scalaris, 77
- supervisor
 - start_link, 78
- timer
 - sleep, 50
 - tc, 47
- trace_mpath, 47
- util
 - tc, 47
- vivaldi, 53
- vivaldi_latency, 53
- your_gen_component
 - init, 50, 52
 - on, 51, 52