MARK PARSONS AND PAUL GRAHAM

Is the Grid Delivering for Business?

The concept of the Grid has been with us for over five years now. Those of us who work in the grid domain have eagerly anticipated the uptake of grid technology by the business world. Sometimes it seems as if we are no closer to delivering on this goal than we were years ago. We do believe, however, that the advent of Web Services based Grid technology, such as the Globus Toolkit™ Version 3, has placed us much closer to delivering a standards-based, ubiquitous middleware solution that businesses can easily deploy to help them transform their business processes.

EPCC, the supercomputing center at the University of Edinburgh in Scotland, has a long history of working with businesses to introduce them to new technologies and demonstrate the benefits that those technologies can deliver. In this article, we examine a deployment of the Globus Toolkit and EPCC's data access and integration services (OGSA-DAI) to solve a data-mining problem in a large UK bus company.

We'll start by discussing some of the issues surrounding the Grid and its deployment in a business context.

Building a Business Grid

Over the past few years, most people in the IT sector will have noticed a slew of articles on subjects such as Autonomic Computing, Adaptable Computing, Grid Computing, Utility Computing, and On Demand Computing. Although each of these developments fits into a specific niche, they are all part of a spectrum of new solutions to traditional IT challenges. Each of them involves two components: first, a new way of thinking about business processes, and second, new technologies to support the delivery of these solutions.

While powerful, the uptake of these solutions by the general business community has been patchy, with many companies steering clear until the technologies become more mainstream. This process is not being helped by IT vendors, each of whom has a subtly different message. With little to do after the dot-com bubble burst, it seems at times as if their marketing departments have spent too much time inventing new names for these technologies. There is, however, one common thread that has emerged: *Web Services* — a term with which most readers will almost certainly be familiar.

Put very simply, Web Services are a standardized way for applications to talk to each other and to describe what they're expecting to hear in that conversation. They allow applications to communicate by sending documents written in XML to each other using a communications protocol called SOAP. The documents are received by the Web Service via an interface that follows the rules of the Web Services Description Language. This type of communication means that Web Services written on one vendor's hardware and software can talk to Web Services on another vendor's with complete transparency. In IT terms, this step is a major move forward and one of the main reasons that service-oriented architectures are being seriously considered by many businesses.

Many of the solutions from IT vendors are now being labeled as Grid technology, although in many cases they simply represent a re-labeling from previous Cluster Computing offerings. This situation is certainly not true of the Globus Toolkit Version 3 (GT3), used in the work described later in this article. GT3 is the first web services based version of the toolkit and it seeks to provide the *software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources*. This version of GT3 presents a vision of the Grid to which the majority of the Grid community subscribes.

In the scientific community, several forces have been driving us toward Grid Computing. These include the relentless increase in microprocessor performance, the availability and reliability of modern high-performance global networks, and the explosion of data quantities. Increasingly, business is coming up against these issues — particularly with regard to data management in widely distributed organizations. Businesses may not have the sheer data volume issues yet, but they certainly do have very complex data access and integration issues.

One of the criticisms often made against the Grid is that it's technology driven and over-hyped by IT vendors. While there is certainly some truth in these criticisms, at EPCC we have always believed the best way to counter them is to work with companies to demonstrate the benefits of the technology. In the past year we've done just this with First Group in the FirstDIG project, which we explain below.

On to the Buses

First is a major company in the transport sector, operating world-

wide and with over 10,000 vehicles in the UK alone. They are the UK's largest operator of bus services. In our project First was represented by First Bus South Yorkshire.

In running their services, First collects many different kinds of operational data, ranging from revenue and fuel consumption to customer contact and scheduling. These data are collected in different ways, such as manual entry, ticket machines, or even GPS. The data are also stored in many different databases and systems. This situation is where First hoped to use Grid technologies. The bus industry is incredibly competitive in the UK and any opportunity to get a competitive edge is of great interest. Coupled with this fact are new technologies, such as electronic ticketing, which allow the companies to know where and when regular customers get on a bus. Simple information such as this, gathered systematically, has never been available to the industry before.

The heterogeneous nature of First's data repositories can be traced back to reasons such as earlier acquisitions of other companies and the incremental construction of IT systems. This condition does not cause a problem for the day-to-day running and querying of the databases, but First was interested in using data mining to leverage non-obvious and valuable information from their data.

One of the techniques of data mining involves correlating and comparing the data from various sources in order to develop this non-obvious information. However, computing these correlations is always difficult when the databases involved are on disparate operating systems, different database management systems, have different internal representations of the data, and are even geographically distributed between various sections of the company. Incidentally, this issue is common to many organizations.

FIGURE ONE A First Bus on the job

Over the past decade a typical solution has been to build a data warehouse where all of the data in the disparate databases is captured and then analyzed. Although a very powerful solution, many organizations have discovered that building and populating data warehouses is much harder than first envisaged. Issues range from a perceived desire to retain ownership of data by those who have collected it to issues of data cleanliness and updating of the warehouse from disparate data sources.

A New Way to See Your Data

EPCC's solution was to provide a mechanism so that First could access their data via a single interface in a uniform manner, protecting the user from the differences between, and the location of, the underlying databases. We like to think of this as virtual data warehousing.

> Web services are a standardized way for applications to talk to each other and to describe what they're expecting to hear in that conversation

This goal was achieved by *Grid-enabling* the databases via the OGSA-DAI (Open Grid Services Architecture — Data Access and Integration) software. This software, developed in collaboration with IBM, Oracle and the Universities of Manchester and Newcastle, and which is built on top of GT3, is Grid middleware that exposes the databases as Grid Services, thus enabling access from other machines in a secure manner.

Two of First's databases were identified as suitable candidates via a preliminary data mining process. EPCC then deployed the OGSA-DAI software for these databases. The software incorporated a GUI front-end to the Grid Services making access to the data easy and straightforward for the user. Functionality was included in the GUI to enable a simple `JOIN` SQL query to be performed thereby allowing data from the different databases to be combined for data mining analysis. Although the demonstration was simple the power of its approach was not lost on Darren Unwin, First South Yorkshire's Divisional IT Manager, who said, "The results of this exercise will revolutionize the way we do things in the bus industry."

Taking the Bus to OASIS

Anyone who has been engaged in Grid research over the past few years will know that we have all learned an enormous amount from the early Web Services based Grids built around GT3. They will also know that since January 2004, a proposal has existed to replace the underlying technology inside GT3 known as OGSI with a standards-based Web Services solution known as the Web Services Resource Framework (WS-RF). WS-RF is now gradually making its way through the OASIS standards process and an early version of WS-RF now exists

See Grid, page 54

Grid, from page 45
that will be released as part of GT4 in the next few months.

Although the move from OGSI to WS-RF has necessitated some changes to the way OGSA-DAI operates, the fundamentals are the same, and we are committed to moving our technology to WS-RF as and when the various standards settle down. At the same time, OGSA-DAI is built on the GGF DAIS proposals and we will continue to follow these specifications as we enhance the product.

Conclusion

In many ways, we are about to enter the most exciting part of the Grid story. Many of the early technologies are maturing and IT vendors are beginning to release mainstream products based on them. The future of Grid research is in the development of higher-level services exactly like those demonstrated successfully to First.

Only by working closely with business to solve real business problems will we reach our goal of a ubiquitous distributed computing solution that cures headaches rather than creates them.

The OGSA-DAI and the FirstDIG projects were funded by the United Kingdom's Department of Trade & Industry and the Engineering & Physical Sciences Research Council through the e-Science Grid Core Programme.

Globus Toolkit is a registered trademark held by the University of Chicago. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38 and under Contract DE-AC03-76SF0098 with the University of California; by the National Science Foundation; by the NASA Information Power Grid program; and by IBM.

Mark Parsons and Paul Graham are senior staff at EPCC, Europe's premier supercomputing center at The University of Edinburgh (www.epcc.ed.ac.uk). EPCC is a founding member of the Globus Alliance.

Resources

The Globus Toolkit
- www.globus.org/toolkit

OGSA-DAI
- www.ogsa-dai.org.uk

FirstDIG
- www.epcc.ed.ac.uk/firstdig

Advertisers' Index

The Advertisers' Index lists each company's Web address and advertisement page. To advertise in ClusterWorld Magazine, please contact adsales@clusterworld.com for a media kit containing an editorial schedule, rate card, and ad close dates.

ClusterWorld Magazine is published monthly by InfoStrada LLC at 330 Townsend St. Suite 112, San Francisco, CA 94107. The U.S. subscription rate is $39.95 for 12 issues. In Canada and Mexico, a one-year subscription is $89.95 U.S. In all other countries, the annual rate is $119.95 US. Non-U.S. subscriptions must be pre-paid in U.S. funds drawn on a U.S. bank.

POSTMASTER: Send address changes to ClusterWorld Magazine, P.O. Box 592, Mt. Morris, IL 61054-0592