PROCEEDINGS OF SCIENCE

Co-allocation of Compute and Network resources using HARC

Jon MacLaren[*†]
E-Science NorthWest (ESNW), University of Manchester, Oxford Road, Manchester M13 9PL
E-mail: jon.maclaren@manchester.ac.uk

HARC—the Highly-Available Resource Co-allocator—is a system for reserving multiple resources in a coordinated fashion. HARC can handle multiple types of resource, and has been used to reserve time on supercomputers distributed across a nationwide testbed in the United States, together with dedicated lightpaths connecting the machines. HARC makes these multiple allocations in a single atomic step; if any resource is not available as required, then nothing is reserved. To achieve this "all or nothing" behavior, HARC treats the allocation process as a Transaction, and uses a phased commit protocol. The Paxos Commit protocol to ensure that there is no single point of failure in the system, which, if correctly deployed, has a very long Mean Time To Failure.

Here we give an overview of HARC, and explain how the current HARC Network Resource Manager (NRM) works, and is able to set-up and tear-down dedicated lightpaths.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

[*]Speaker.

[†]Special thanks to Mark Mc Keown who provided the initial design for HARC, while at the University of Manchester.

PoS(ESLEA)016

1. Motivation

The ever-improving availability of high-bandwidth, low-latency optical networks promises to enable the use of distributed scientific applications [7, 3] as a day-to-day activity, rather than simply for demonstration purposes. However, in order to enable this transition, it must also become *simple* for users to reserve *all* the resources the applications require.

The reservation of computational resources can be achieved on many supercomputers using advance reservation, now available in most commercial and research schedulers. However, distributed applications often require guaranteed levels of bandwidth between compute nodes, or between compute nodes and a visualization resource. At the network level there are switches and routers that support the bandwidth allocation over network links, and/or the configuration of dedicated end-to-end lightpaths. These low-level capabilities are sufficient to support the development of prototype middleware solutions that satisfy the requirements of these applications.

However, the development of an booking system for network resources is not a complete solution, as the user is still left with the complexity of coordinating separate booking requests for multiple computational resources with their network booking(s). Even if there is a single system available that can reserve all the required compute resources, such as Moab, or GUR (which can reserve heterogenous compute resources), this does not address the need to coordinate the scheduling of compute and network resources—a co-allocation system that can deal with multiple *types* of resources is required.

2. HARC: The Highly-Available Resource Co-allocator

HARC, the Highly-Available Resource Co-allocator [2, 8], is an open-sourced system that allows users to reserve multiple distributed resources in a single step. These resources can be of different types, e.g. supercomputer time, dedicated network connections, storage, the use of a scientific instrument, etc. Currently, HARC can be used to book High-Performance Computing resources, and lightpaths across certain GMPLS-based networks with simple topologies. The HARC Architecture is shown in Figure 1.

HARC uses a phased commit protocol to allow multiple resources to be booked in an all-or-nothing fashion (i.e. atomically). Paxos Commit [6] is used, rather than the classic 2-Phase Commit (2PC), to avoid creating a single point of failure in the system. Paxos Commit replaces 2PC's single Transaction Manager (TM) with a number of processes, or *Acceptors*, which perform the same function as the TM. The Paxos Consensus algorithm guarantees consistency, so clients can talk to any Acceptor to find the results of their requests. The overall system functions normally provided a majority of Acceptors remain in a working state. This gives a deployed system of five Acceptors a far longer Mean Time to Failure than that of any single Acceptor.

HARC is designed to be extensible, and so new types of Resource Manager can be developed without requiring changes to the Acceptor code. This differentiates HARC from other co-allocation solutions. The assumption is that the underlying resource has a scheduler capable of reserving the resource (or part thereof) for a specific user; the RM should be a small piece of code that interacts with this scheduler on the user's behalf.

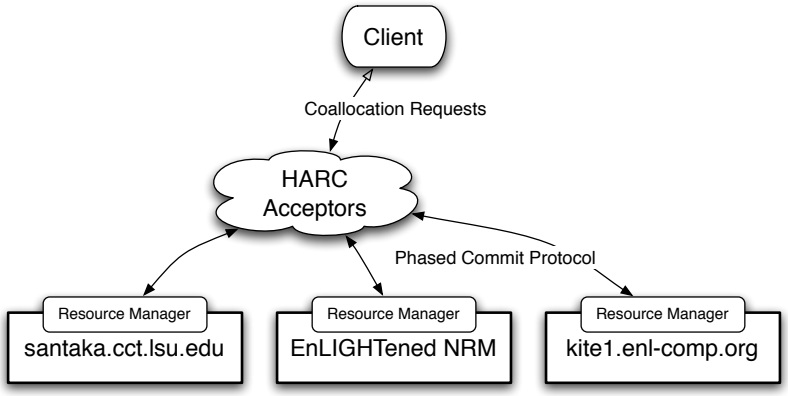

Figure 1: The HARC architecture, showing the relationship between the client, the Acceptors, and the Resource Managers (RMs).

3. Reserving Network Connections using HARC

How best to reserve network connectivity in advance is still a research topic. When the EnLIGHTened Computing project [1] started, there were deployed reservation systems such as the G-lambda project's GNS-WSI2 [5] and EGEE's BAR [9]. However, the project chose to implement a new, simple, timetable-based system, which was embedded in a HARC Resource Manager; this component is referred to as the HARC Network Resource Manager (NRM). There is a single HARC NRM for the entire testbed (centralized).

A schematic for the EnLIGHTened Computing testbed is shown in Figure 2. At the heart of the network are three Calient Diamondwave PXCs (UO1 in Chicago, UO2 in Raleigh, and UO3 in Baton Rouge) and a single Diamondwave PX (UO4 in Caltech). The software on the switches supports GMPLS, and connections across the testbed can be initiated by sending a TL1 command to a switch at either end of the connection. A dedicated lightpath can be set-up between any two entites at the edge of the cloud. These are either routers (UR2, UR3) or compute nodes (RA1, VC1, CH1); the router X1U is a special case, used to connect through to the Japanese JGN II network. All links in the network are 10 Gigabit Ethernet (10 GE), except for the connections to Japan, which are Gigabit Ethernet.

The NRM accepts requests for network connections on a first-come, first-served basis. Requests specify the start and end points of the connection (using the three letter acronyms shown on Figure 2), as well as the required bandwidth, and also the desired setup and teardown times. The example in Figure 3 would be used to request a lightpath between two supercomputers at MCNC in Raleigh and CCT in Baton Rouge.

Typically, GMPLS chooses the best path through the network when a path is set up, dynamically avoiding non-functioning components. However, when scheduling links in advance, it is important for the scheduler to be in control of the routes that each lightpath uses, to ensure that all paths follows the schedule. In the current EnLIGHTened testbed network, for any two endpoints, there is only a single possible path through the network. Even though this is the case, the NRM does specify the full path to the switches during the provisioning process, in an Explicit Route Object (ERO), which is sent as part of the TL1 command.

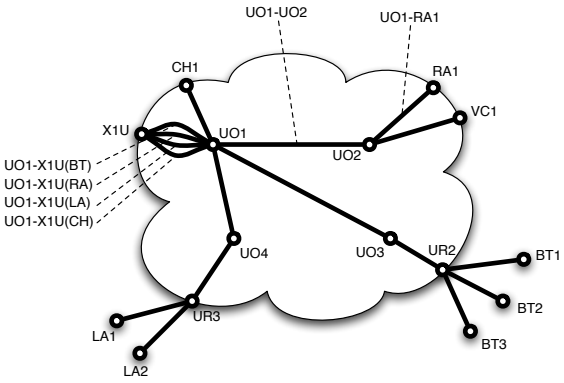

Figure 2: A simplified schematic of the EnLIGHTenend testbed network.

```
<Schedule><TimeSpecification><Exact>
   <StartTime>2007-04-25T21:00:00Z</StartTime>
   <EndTime>2007-04-25T22:00:00Z</EndTime>
</Exact></TimeSpecification></Schedule>
<Work>
   <Path>
      <From>RA1</From><To>BT2</To>
      <BandwidthMbs>10240</BandwidthMbs>
   </Path>
</Work>
```

Figure 3: XML Snippet from a HARC NRM Message.

3.1 The Future of the NRM

The current HARC NRM needs to be split into two components: a pure scheduling component with a service interface, and a much smaller HARC NRM component, which simply becomes an interface between the HARC Acceptors and the network scheduling service.[1] This is consistent with the other HARC Resource Managers that have been developed, as explained in Section 2.

The internals of the current scheduling code are also very simple. Although the network topology has not been hardcoded into the service in any way (all configuration is obtained from a set of files), there is still an assumption that given two endpoints, there is a single path through the network. Soon the EnLIGHTened testbed will be extended with a Calient Diamondwave PXC in Kansas City, creating additional paths between most endpoints; additional code will be required to deal with this correctly.

The NRM also needs to be able to cope with both planned and unplanned downtime of parts of the network, and—where possible—should ensure that users are not permitted to schedule lightpaths for times when the network is not going to be available. This will involve some level of integration between the NRM and relevant monitoring software.

4. Conclusions

There are two deployments of HARC in use today: the EnLIGHTened testbed in the United States; and a second on NorthWest Grid,[2] a regional Grid in England. A trial deployment is planned for TeraGrid,[3] and HARC is being evaluated for deployment on the UK National Grid Service.[4] An alternate Network Resource Manager that interfaces to the ESLEA Circuit Reservation Software [4] is also being considered. This would allow HARC to be used to co-allocate parts of the UK Lite network.

The prototype HARC Network Resource Manager component, described in Section 3, has been used to schedule some of the optical network connections being used to broadcast Thomas

[1] The G-lambda project's GNS-WSI2 [5] interface is currently being evaluated for its suitability for this task.

[2] http://www.nw-grid.ac.uk/

[3] http://www.teragrid.org/

[4] http://www.ngs.ac.uk/

Sterling's HPC Class from Louisiana State University.[5] Previously, HARC was used in the high-profile EnLIGHTened/G-lambda experiments at GLIF 2006 and SC'06, where compute resources across the US and Japan were co-allocated together with end-to-end optical network connections.[6]

Although these early successes are encouraging, if the advance scheduling of lightpaths is to become a production activity, then the network scheduling service(s) need to be properly integrated with the other control/management plane software to ensure that these activities do not interfere with the pre-scheduled lightpaths (and vice-versa).

Acknowledgements

The implementation of HARC took place while the author was employed at the Center of Computation & Technology at Louisiana State University. During this time, the work was supported in part by the National Science Foundation "EnLIGHTened Computing" project [1], NSF Award #0509465.

References

[1] EnLIGHTened Computing: Highly-dynamic Applications Driving Adaptive Grid Resources [Online]. http://www.enlightenedcomputing.org.

[2] HARC: The Highly-Available Resource Co-allocator [Online]. http://www.cct.lsu.edu/∼maclaren/HARC.

[3] R. J. Blake, P. V. Coveney, P. Clarke, and S. M. Pickles. The teragyroid experiment–supercomputing 2003. *Scientific Computing*, 13(1):1–17, 2005.

[4] A. C. Davenhall, P. E. L. Clarke, N. Pezzi, and L. Liang. The ESLEA Circuit Reservation Software. In *Proceedings of "Lighting the Blue Touchpaper for UK e-Science: closing conference of ESLEA Project"*. PoS(ESLEA)015, 2007.

[5] G-lambda Project. Grid Network Service / Web Services Interface, version 2. http://www.g-lambda.net/wordpress/?page_id=19.

[6] J. Gray and L. Lamport. Consensus on transaction commit. *ACM TODS*, 31(1):130–160, March 2006.

[7] A. Hutanu, G. Allen, S. D. Beck, P. Holub, H. Kaiser, A. Kulshrestha, M. Liška, J. MacLaren, L. Matyska, R. Paruchuri, S. Prohaska, E. Seidel, B. Ullmer, and S. Venkataraman. Distributed and collaborative visualization of large data sets using high-speed networks. *Future Generation Computer Systems. The International Journal of Grid Computing: Theory, Methods and Applications*, 22(8):1004–1010, 2006.

[8] J. MacLaren, M. M. Keown, and S. Pickles. Co-Allocation, Fault Tolerance and Grid Computing. In *Proceedings of the UK e-Science All Hands Meeting 2006*, pages 155–162, 2006.

[9] C. Palansuriya, M. Büchli, K. Kavoussanakis, A. Patil, C. Tziouvaras, A. Trew, A. Simpson, and R. Baxter. End-to-End Bandwidth Allocation and Reservation for Grid applications. In *Proceedings of BROADNETS 2006*. http://www.x-cd.com/BroadNets06CD/pdfs/87.pdf, October 2006.

[5] This class is the First Distance Learning Course ever offered in Hi-Def Video. Participating locations include other sites in Louisiana, and Masaryk University the Czech Republic. See http://www.cct.lsu.edu/news/news/201

[6] See http://www.gridtoday.com/grid/884756.html