Lowe, M., C. Shields, D. Hancock, M. Link, C.A. Stewart and M. Pierce. Gateway Hosting at Indiana University.
In: Proceedings of the TeraGrid 2009 Conference, 22-25 June, Arlington, VA. http://hdl.handle.net/2022/13944

Gateway Hosting At Indiana University

Mike Lowe[1], Corey Shields[1], David Hancock[1], Matt Link[1], Craig Stewart[1], Marlon Pierce[1]
[1]Indiana University, USA

Abstract

The gateway hosting service at Indiana University provides science gateways and portals with hosting resources to facilitate the use of computation resources and storage within the TeraGrid. This service is designed with high availability in mind and is deployed across the Indianapolis and Bloomington campuses with redundant network, power, and storage. The service uses OpenVZ [1] to give each gateway or portal its own virtual environment while making the most efficient use of the hardware and administrative resources. OpenVZ's user beancounter quota system and fair-share scheduling for processes and I/O allows fair distribution of resource between virtual machines while allowing full utilization of the hardware. The ability to do live migration allows kernel updates without service interruption. Indiana University's research network provides multiple low latency high bandwidth connections between campuses, other TeraGrid resource providers, and the Internet at large. The service is in use by a variety of projects such as FlyBase and TeraGrid Information Services and, since the service was put into production in August 2008, there have been 5.37 hours of down time.

1 Introduction

Currently the TeraGrid has an enormous amount of computation, storage, and networking resources. Access to these resources can be difficult for users that are uncomfortable with a command line interface. As a consequence, there has been a tremendous effort to overcome the obstacles between the desktop and the supercomputer. Science Gateways [2] and portals are two promising solutions to overcoming these obstacles. An example of such endeavors is the Linked Environments for Atmospheric Discovery (LEAD) [3] project. LEAD enables undergraduate meteorology students to set up and run complicated weather simulation with a graphical interface. Nevertheless, despite the success of the LEAD and other such projects, the TeraGrid has not traditionally provided hosting resources, forcing projects to find outside resources to facilitate the use of TeraGrid resources. To support the efforts of Science Gateways and portals, Indiana University has created a facility to provide projects with hosting resources.

The primary method for projects to interact with the desktop is with a web browser; therefore, the gateway hosting project should be optimized for a web-based workload. Gateways have an interactive workload that should minimize latency for the user and be highly available. Gateways typically do not move large amounts of data themselves but may act to control third party transfers. Each gateway is run by an different group of researchers and will have its own software stack, development schedule, and maintenance schedule. Gateways may have access to hosting resources, but the hosting resources often have monetary costs and may not have suitable high availability features. To meet the needs of the gateways, a hosting service will need high availability, low latency, moderate bandwidth, and separate environments isolated from each other. A virtualized hosting solution is a natural fit as it can provide the high availability and separated, isolated user environments on a machine with sufficient capacity to ensure low latency and sufficient bandwidth.

2 Virtualization Technologies for Gateway Hosting

A number of different virtualization technologies were examined [4] to determine which one would best fit the design requirements. Among the ones considered were XEN[5], OpenVZ, VMWare [6], Linux VServer [7], and KVM [8]. The virtualization technologies fall into two basic categories: 1) ones that use a separate kernel for each virtual machine or operating system virtualization, and 2) ones that share one kernel between virtual machines or para- and hardware assisted virtualization. XEN, VMWare, and KVM, being para- or hardware assisted virtualization, all run an additional kernel for each virtual machine running. With multiple kernels running, there is duplication of functionality between kernels in the virtual machines. This duplication wastes resources because it does not help meet a design requirement or goal. Being able to use different kernels is not a design goal for this service, but maximizing performance and hardware usage is. Minimizing the administrative overhead of the service in terms of person hours is also a design goal that is not met by running more than one kernel per physical machine. Since the requirements do not call for the features of the more popular virtualization technologies and in some cases run counter to them, OpenVZ was chosen as the virtualization technology for this project.

OpenVZ uses a single kernel with the virtual machines placed in a change root jail. Process identifiers (PIDs) are placed in their own space, so each container has the full range of valid PIDs available to it. Networking is virtualized, with each virtual machine getting its own interface and set of firewall rules. Filesystems are mounted on the hardware node (the host system) and the desired portion bind mounted into the appropriate virtual machine. Virtual machines can make full use of any file system caching opportunities that may exist between them. There is a checkpointing facility available which makes possible live migration. All processes belonging to a container can be frozen with their open file handles including network sockets and written to disk. The file containing the frozen processes can be moved to any other machine running OpenVZ and reconstituted as a virtual machine preserving open network connections, provided the filesystems mounted by the virtual machine are identical.

OpenVZ utilizes a number of different mechanisms to ensure that virtual machines receive a minimum quality of service. Time slices are first allocated to a virtual machine based on administrator-defined values, then processes belonging to that virtual machine are scheduled by the standard Linux process scheduler. This implementation of a fair-share scheduling strategy ensures that under load each virtual machine gets all the time slices allocated to it, and if there are no eligible processes a virtual machine will yield its time slices for use by other virtual machines. Similarly the I/O scheduler is two-stage, with administrator defined priorities determining the minimum amount of the bandwidth a virtual machine is entitled to. Completely Fair Queuing (CFQ) from the standard Linux kernel is used as the second stage I/O scheduler. User bean counters are used to set limits and guarantees on various in-kernel objects and memory for virtual machines. It is possible to limit the total amount of memory used by a virtual machine's processes, the number of open TCP sockets, the number of open files, or the size of TCP buffers. There are twenty-four different user bean counters available for each virtual machine.

3 Current Design and Implementation

3.1 Hardware

The hardware consists of two nodes with 32GB of RAM, a 1TB RAID 5 array, and four AMD Opteron dual- or quad-core processors. At the time of purchase, AMD's hyper transport yielded better memory performance than Intel's front side bus designs [8]. Memory performance is critical to supporting a large number of active processes that are unlikely to share L3 cache lines. Projecting from current utilization levels, a single node of this configuration could very easily support 80 virtual machines. Power for both nodes is backed by UPS and generator, leaving ample time for migration to the other node during an extended power outage.

3.2 Networking

The research network at Indiana University spans the Bloomington and Indianapolis campuses with multiple 10 gigabit ethernet links (see Figure 1). Redundant paths to the Internet are available via dedicated links to the Indiana GigaPoP [10] and the campus networks. Connections to the TeraGrid network and the Internet2 networks are also available from the research network. Each node is connected to the research network via a 1 gigabit ethernet connection. All virtual machines running on a node share this connection. All virtual machines are allocated their own static IP address in the same subnet as the hardware nodes. Proxy ARP is used by the hardware nodes to route from the VLAN that joins the two campuses to their respective networking stacks. Within a node the traffic destined for a virtual machine is forwarded across a point to point virtual network interface to it's final destination.

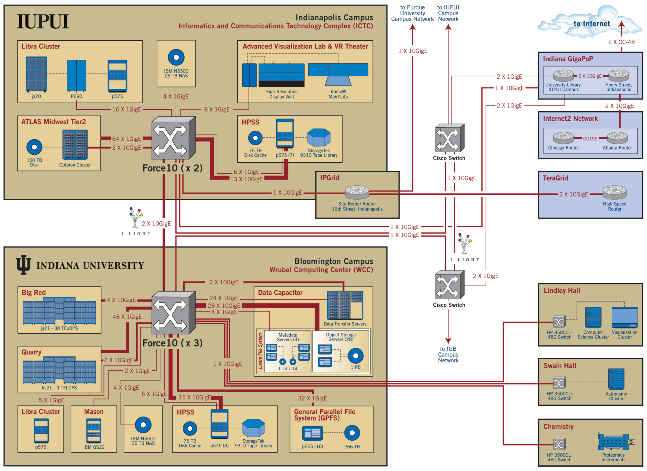

Figure 1. Indiana University's Research Networks

3.3 Operating System

Redhat Enterprise Linux 5 was selected for the operating system of the hardware nodes, with the kernel patched for OpenVZ. Further modifications to the kernel were made to enable live migration for virtual machines with NFS and Lustre file systems. To reduce the time required for live migration, virtual machines' root directories are replicated between the two hardware nodes, leaving only the memory in use by the virtual machine to be transfered during a migration. The block level replication of the underlying device is accomplished with DRBD [11] (see Figure 2). This ensures that in the event of a total failure of a hardware node, the other hardware node can resume running the virtual machines with minimal data corruption and loss.

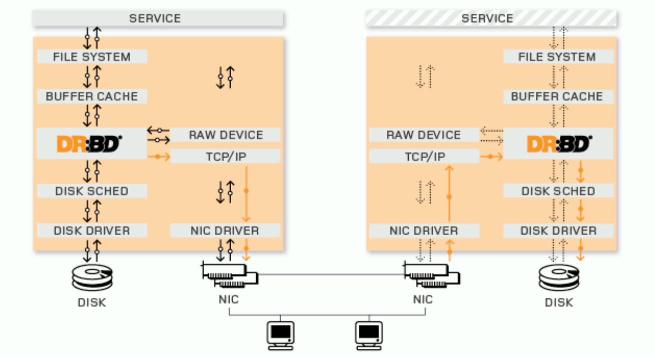

Figure 2. Logical Diagram of DRBD Disk Mirroring

Virtual machines can run any Linux distribution that supports a version 2.6 kernel. CentOS 4, CentOS 5, Redhat Enterprise Linux 4, Redhat Enterprise Linux 5, Gentoo, Unbuntu, Fedora, and Debian are a few of the supported Linux distributions. Any machine run by any service provider can be backed up with tar and converted to a virtual machine with very little effort.

Virtual machines can mount network file systems available to other resources at IU. This includes the home directories, Data Capacitor a lustre based filesystem, and Data Capacitor-WAN a wide area lustre based filesystem [12]. These filesystems are mounted on the hardware node, then they are bind mounted into the virtual machine's filesystem. Using this methodology allows full user separation in a simple and transparent way.

4 Current Users

There are currently 15 projects that have access to the IU Gateway Hosting Service. Some of the most active ones are GridChem, FlyBase [13], TeraGrid Information Services, TeraGrid GIG Data Movement Portal, OGCE, QuakeSim and FloodGrid. FlyBase, for example, is a searchable database of the *Drosophila* genome and is classified as a TeraGrid data collection. It is important to node that the benefit of this service is not only confined to Science Gateways, FlyBase illustrates the utility for data collections that are accessed via the web. QuakeSim [14] is a NASA project for modeling earthquake fault systems.

Projects wanting to start using the Gateway Hosting Service will need to consult with the system administrators at IU after requesting an allocation in the TeraGrid User Portal. This will ensure that the service meets their needs and that proper provisioning and capacity planning are done. Once a virtual machine is allocated, best efforts will be made to ensure that the virtual machine is always available unless the user requests it and that there will always be reasonably sufficient capacity in reserve to expand the resources available to any virtual machine.

5 Conclusion and Future Work

The Gateway Hosting Service at Indiana University is unique among all other TeraGrid resources in that it provides hosting services for TeraGrid projects. The distributed and highly available nature of the service has led to having only 5.37 hours of down time since August 2008. The use of OpenVZ minimizes administrative overhead while maximizing hardware utilization. Network filesystems can be made available to users from within their virtual machines. Future plans include expanding the hardware in terms of both disk space and memory size and will be executed when utilization levels call for it. When complete, the new data center currently under construction on IU's Bloomington campus will have larger backup power facilities, further reducing the chance of outages. The move to the new data center is scheduled for August 2009.

References

[1] OpenVZ, http://www.openvz.org

[2] TeraGrid Science Gateways, http://www.TeraGrid.org/gateways/

[3] Linked Environment for Atmospheric Discovery, http://www.leadproject.org

[4] Comparison of platform virtual machines, http://en.wikipedia.org/wiki/Comparison_of_platform_virtual_machines

[5] XEN, http://xen.org

[6] VMWare, http://www.vmware.com

[7] Linux VServer, http://linux-vserver.org

[8] KVM, http://www.linux-kvm.org

[9] P Conway, B Hughes, The AMD Opteron Northbridge Architecture, IEEE Micro, Vol. 27, Issue 2, Pages 10-21, 2007.

[10] Indiana GigaPOP, http://indiana.gigapop.net

[11] DRBD, http://www.drbd.org

[12] Wide Area Filesystem Performance using Lustre on the TeraGrid, http://datacapacitor.researchtechnologies.uits.iu.edu/dcpapers.shtml

[13] FlyBase, http://flybase.org

[14] QuakeSim, http://quakesim.jpl.nasa.gov/