Booting Over Infiniband With Perceus Cluster Management

Summary

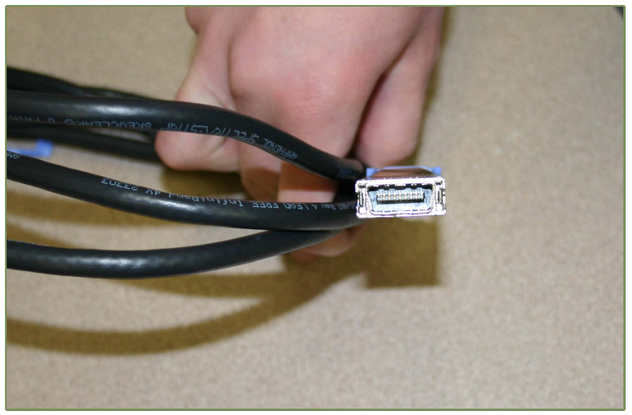
Two main network fabrics are used in large diskless HPC clusters: Ethernet is typically used for cluster management tasks such as booting and IB is typically used for fast data communication. Configuring a cluster of diskless nodes to boot over IB fabric using Perceus could help eliminate the need for Ethernet in clusters, reducing costs and reducing the number of parts. The motivation behind this project is a situation currently facing the Coyote super computer. It is wired exclusively with IB and uses a two stage boot process; it loads a small kernel from flash memory and proceeds to download the rest through IB. Those who manage the cluster would prefer to move away from flash memory, leaving only two viable options: purchase and install an expensive Ethernet network, or configure the computers to fully boot over IB.

What Is Perceus?

Perceus is a suite of cluster management tools that allow a cluster to boot over a network. It does this by sending an image to each node, which is then loaded into ram. This allows for the nodes to operate with out hard disk drives.

What Is Infiniband?

Infiniband is a high speed network fabric. It is used in high performance computing as it offers 2.5 times the bandwidth of Ethernet and has considerably lower latency. The issue with Infiniband is that it is still developing as a technology; the are no definite standards other than an abstract set of commands called verbs. This also means that Infiniband cards are not designed to net boot. This is why Ethernet networks are still included in today's clusters.



What We Have Done

- Created a Perceus image with Infiniband drivers.
- Flashed Infiniband cards to include support for gPXE booting.
- Added Infiniband drivers to Stage 1 image.
- Patched the DHCP source code to recognize the 32 digit MAC addresses.

What We Plan To Do

We have run into an issue where Infiniband isn't configured after loading the stage one image. We are currently trying to solve this problem in a couple different manners;

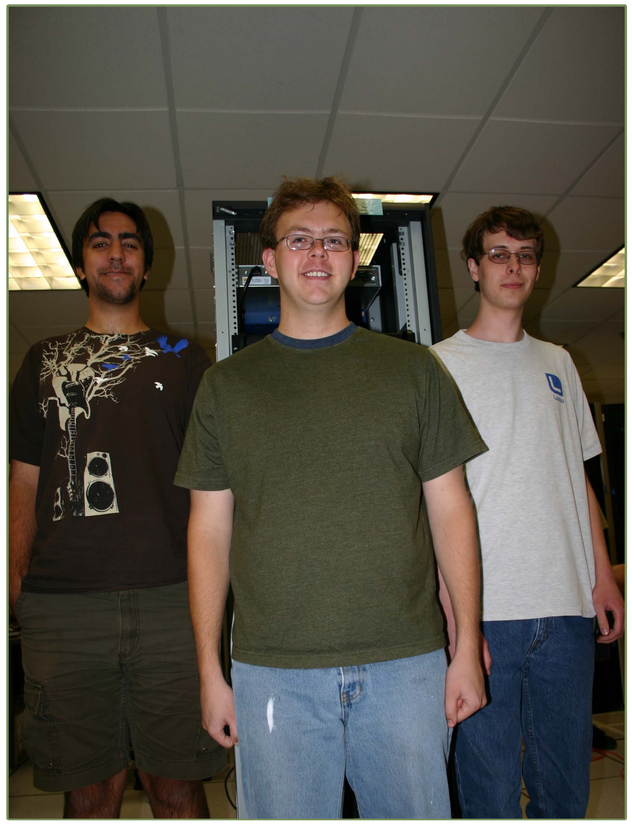
- Replacing the busybox with an older one patched to deal with Infiniband.
- Switching to an older version of Perceus and patching it.
- Fixing the source code of the new busybox.

What is the Significance?

This project was inspired by an issue facing the coyote super computer; it was built without an Ethernet network booting the Stage 1 kernel off of flash memory. This has led to concern over the lifespan of the flash memory several ideas have been discussed to eliminate the need for the flash memory. One is to alter the system to boot stage 1 over Infiniband.

In general, booting over Infiniband would allow the clusters to operate without an Ethernet network . This lowers cost, reduces the number of parts per node, removes some of the points of failure and would increase the mean time between failures.

As Infiniband has higher bandwidth than Ethernet which could lead to a decrease in boot time for clusters. This would add to the valuable up time.



Ideas For Future Research

Through our work on this project we have come up with several ideas for future research as an extensions to this project.

- Multicast boot over Infiniband may be a quick and efficient solution for a larger cluster.
- Using iSCSI rather than NFS when booting over Infiniband.
- Bottleneck research, doing quantitative analysis of the boot speed of Ethernet and Infiniband.

Team Members

Matthew Dosanjh	University of New Mexico
William Pickett	New Mexico Tech
Graham Van Heule	Michigan Tech

Mentors

Andrew Shewmaker	HPC-5
Ben McClelland	HPC-3
Andree Jacobson	University of New Mexico

Los Alamos NATIONAL LABORATORY EST. 1943

National Security Education Center
A CONSORTIUM OF LANL INSTITUTES

ISTI
INFORMATION SCIENCE & TECHNOLOGY INSTITUTE

New Mexico Consortium's IAS
Institute for Advanced Studies at Los Alamos National Laboratory

LA-UR 09-04881