LA-UR- 09-05103

Approved for public release; distribution is unlimited.

Title:	Booting Over Infiniband With Perceus Cluster Management
Author(s):	Matthew Dosanjh, INST-OFF William Pickett, IINST-OFF Graham Van Heule, INST-OFF
Intended for:	Academic Distribution

Los Alamos
NATIONAL LABORATORY
EST 1943

Form 836 (7/06)

Abstracts

Booting Over Infiniband With Perceus Cluster Management

Matthew Dosanjh, UNM
William Pickett, NMT
Graham Van Heule, MTU

Abstract: 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 processes; 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.

To configure the network to boot over IB the IB cards must be upgraded to use the gPXE protocol, after which the cluster management software must be configured to recognize and work with the IB cards allowing for a diskless boot. The potential implications include the evaluation of scalability in a large cluster, such as Coyote. As IB has higher bandwidth than Ethernet, clusters would gain more computing time by decreasing boot time. This also leads to potential research of multicast booting over IB.

Booting Over Infiniband With Perceus Cluster Management

PRESENTED BY

Matthew Dosanjh – UNM

William Pickett – NMT

Graham Van Heule – MTU

On 8/3/2009

Outline

- Motivation
- Goals
- What We Did
- Issues Faced
- Future Research
- Conclusions

Motivation

- **Coyote**
  - Has no Ethernet network
  - Uses two stage boot
    - Stage 1 is a small kernel loaded from local flash memory
    - Stage 2 is downloaded by stage one over Infiniband
  - Local flash memory will eventually deteriorate
- **There exist two solutions**
  - Purchase and install an expensive Ethernet network
  - Configure the cluster to grab the stage one image over Infiniband.

Our Project

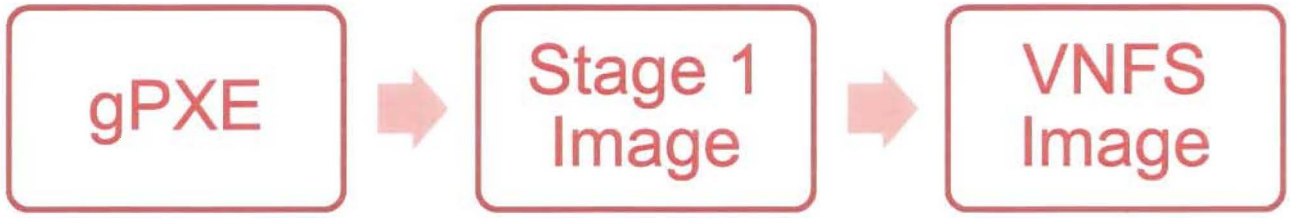
- Our goal is to get this cluster to boot over Infiniband to determine if it is feasible to do it to a larger cluster in a production environment
- Perceus – cluster management software
- DHCP – Dynamic Host Configuration Protocal
- Infiniband – high bandwidth, low latency network fabric

Outline

- Motivation
- Goals
- What We Did
- Issues Faced
- Future Research
- Conclusions

Steps On The Road To Completion

- Created Perceus VNFS image with Infiniband drivers
- Burned gPXE into Infiniband card firmware
- Added Infiniband drivers to stage 1 image
- Patched DHCP to recognize the 32 digit MAC address of Infiniband
- Patched Perceus to accept Infiniband MAC addresses



Issues Encountered

- **DHCP doesn't have support for Infiniband at it's current version**
  - When patched for Infiniband DHCP doesn't send the correct MAC address

  Ethernet MAC: 00:01:02:03:04:05

  Infiniband MAC: 00:01:02:03:04:05:06:07:08:09:10:11:12:13:14:15:16:17:18:19:20

- **The default initramfs doesn't contain Infiniband drivers**
  - Kernel is not by default configured to handle Infiniband
- **Large lack of documentation for Perceus' Infiniband capabilities**

Outline

- Motivation
- Goals
- What We Did
- Issues Faced
- Future Research
- Conclusions

Ideas For Future Research

- 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 comparison of the boot speed of Ethernet and Infiniband

Conclusions

- **We have successfully booted over Infiniband**
  - However we still have issues getting a unique hardware identifier
  - It currently can only boot one node.
- **Further research would be required for large scale deployment**

Questions