An Experiment on Bare-Metal BigData Provisioning

Ata Turk, Ravi S. Gudimetla, Emine Ugur Kaynar, Jason Hennessey, Sahil Tikale, Peter Desnoyers, Orran Krieger
BigData Analytics on the Cloud

- BigData deployments are moving to the cloud
  - On-demand usage (Cost), Elasticity, Agility, Simplicity, …
  - Virtualized IaaS solutions: Amazon EMR, Azure HDInsight, …

- Virtualization drawbacks
  - Overhead, unpredictability, security concerns, device functionality, …
  - Bare-metal cloud solutions: IBM, Rackspace, and Internap, …
Bare-Metal BigData
Cloud Solutions

• Bare-Metal cloud provisioning
  • Automated provisioning: Ironic, MaaS, …
  • Image copy to local disk => long waits => loss of agility & elasticity

• OS streaming*, Lazy copy & de-virtualization**

• What about network booting?
  • incur an ongoing unacceptable overhead during runtime

• Large parts of the HPC community has been doing it for the last 20 years.

• Virtualized IaaS is doing it all the time.

• Why not bare-metal cloud?
Network-Mounted BigData System

- Clients access kernel and init ramdisk via PXE
- Mount OS & BigData apps from a remote iSCSI volume
- Use local disk for ephemeral storage (HDFS, /swap, /tmp, ...)

Diagram:

1. PXE Boot
2. Mount OS & BigData apps
3. Local Disk for HDFS

iSCSI Gateway

DHCP Server

TFTP

Central image store
Bare-Metal BigData Provisioning Prototype

Bare-Metal BigData Provisioning System

HIL

Provisioning VM

iSCSI Gateway

CEPH

6
Bare-Metal BigData Provisioning Prototype

**Hardware Isolation Layer:**
A service to allocate bare-metal nodes out of a shared pool and isolate network.
Bare-Metal BigData Provisioning Prototype

**CEPH**: Central image store hosting user images with BigData applications
Bare-Metal BigData Provisioning Prototype

**Provisioning VM:**
Gateway between isolated servers and image store
Bare-Metal BigData Provisioning Prototype

1: Isolate network and reserve servers

Reserved Servers

Bare-Metal BigData Provisioning System
Bare-Metal BigData Provisioning Prototype
Bare-Metal BigData Provisioning Prototype

Bare-Metal BigData Provisioning System

HIL

Provisioning VM

iSCSI Gateway

Reserved Servers

CEPH

Cloned Images

3: chroot images for per-node config
Bare-Metal BigData Provisioning Prototype

Bare-Metal BigData Provisioning System

- HIL
- Provisioning VM
- iSCSI Gateway
- CEPH

Reserved Servers

Cloned Images

4: PXE-boot servers using cloned images
Provisioning Time

Elapsed Time (Secs)

- Post Config
- BigData App Installation
- OS Reboot
- OS Installation
- OS Boot (inc. kernel+initrd download)
- DHCP Request
- Firmware initialization

Local disk installation: 1200 seconds
iSCSI boot: 200 seconds
 Provisioning Time Scaling

![Bar chart showing provisioning time scaling for 2 Node, 4 Node, and 8 Node configurations. The chart indicates the elapsed time in seconds for Bigdata Post Script, Booting, Ceph Cloning, and Haas Initialization.

- **Bigdata Post Script**
- **Booting**
- **Ceph Cloning**
- **Haas Initialization**

The elapsed time increases with the number of nodes, reflecting the overhead and complexity of provisioning and initializing nodes in a distributed system.]
Read Traffic over Boot Drive

Cumulative iSCSI reads per node (MB)

- iSCSI Reads: Runs with 256GB Data
- iSCSI Reads: Runs with 128GB Data
Read Traffic over Boot Drive

Cumulative iSCSI reads per node (MB)

~170MB / 8GB Boot Image => 2%
Read Traffic over Boot Drive

Cumulative iSCSI reads per node (MB)

- iSCSI Reads: Runs with 256GB Data
- iSCSI Reads: Runs with 128GB Data

3KB/s read after initial boot
Write Traffic over Boot Drive

Cumulative iSCSI writes per node (MB)

- iSCSI Writes - Runs with 256GB Data
- iSCSI Writes - Runs with 128GB Data
Write Traffic over Boot Drive

Cumulative iSCSI writes per node (MB)

- **iSCSI Writes - Runs with 256GB Data**
- **iSCSI Writes - Runs with 128GB Data**

14KB/s write
Runtime Performance of Network-Mounted Boot Drive
Runtime Performance of Network-Mounted Boot Drive

The chart illustrates the elapsed time (in seconds) for various operations (WordCount, Sort, Grep) on data of different sizes (8GB, 16GB, 32GB, 64GB, 128GB) when using local disk and iSCSI mounted drives. The data shows that iSCSI mounted drives generally have a longer elapsed time compared to local disks, especially for larger data sizes.
Runtime Performance of Network-Mounted Boot Drive

![Bar chart showing elapsed time for different data sizes and operations (WordCount, Sort, Grep) on local disk and iSCSI mounted drive. The chart compares 8GB, 16GB, 32GB, 64GB, and 128GB data sizes.](chart)
Take-aways

- Network booting the OS for bare-metal BigData
  - uses only a fraction of boot disk during start-up
  - improves provisioning time with no runtime degradation
    - provisioning time < 5 mins, boot disk reads: ~3KB/s, writes: ~14KB/s
- Enormous effort on bare-metal provisioning on local disks may be unnecessary, especially for BigData deployments
- We are building a new Bare Metal Imaging Service using remote network boot mechanisms
  - enable capabilities available on virtualized platforms (e.g. snapshotting, cloning, …) to bare metal cloud solutions
Questions
Provisioning Time

Provisioning Time

Elapsed Time (Secs)

Local disk installation

iSCSI boot

Emulab*

Ironic*