M2: Malleable Metal as a Service

Apoorve Mohan*, Ata Turk†, Ravi S. Gudimetla‡,
Sahil Tikale†, Jason Hennessey†, Ugur Kaynar†,
Gene Cooperman*, Peter Desnoyers*, and Orran Krieger†

*Northeastern University, †Boston University, and ‡Red Hat Inc. (Massachusetts Open Cloud)
Increasing Bare Metal Cloud Offerings

- **Performance and Security** sensitive applications

- Application that require **accelerators**:
  - FPGA’s, Infiniband, GPUs, etc.

- Setting up frameworks to provide different services:
  - OpenStack, Hadoop, Slurm, etc.

- AWS Bare Metal, IBM Softlayer/Bluemix, Rackspace, Internap, etc.
Existing Bare Metal Offerings Provision to Local Disk - Stateful

- Over the network from an ISO or a Pre-installed image
Stateful Provisioning Problems

- **Slow Provisioning**
  - Up to *Tens of Minutes* to provision

- **Boot Storms**
  - Heavy network traffic

- **Single Point of Failure**
  - Loss of both OS and Application

- **Poor Reusability**
  - Saving and Restoring disk state
Poor Reusability

Tenant 1

Tenant 2
Poor Reusability

Tenant 1

Tenant 2
Poor Reusability

Tenant 1

Tenant 2
What’s the Solution?

- **Slow Provisioning**
  - Up to *Tens of Minutes* to provision

- **Boot Storms**
  - Heavy network traffic

- **Single Point of Failure**
  - Loss of both OS and Application

- **Poor Reusability**
  - Saving and Restoring disk state non-trivial
Why Not Provision Bare Metal like Virtual Machines?

- Over the network from a *pre-installed virtual disk (boot drive)*
How Can Netboot Solve These Problems?

- **Slow Provisioning**
  - Up to *Tens of Minutes* to provision

- **Boot Storms**
  - Heavy network traffic

- **Single Point of Failure**
  - Loss of both OS and Application

- **Poor Reusability**
  - Saving and Restoring disk state non-trivial
How Can Netboot Solve These Problems?

- **Slow Provisioning**
  - Up to *Tens of Minutes* to Provision

- **Boot Storms**
  - Heavy network traffic

- **Single Point of Failure**
  - Loss of both OS and Application

- **Poor Reusability**
  - Saving and Restoring disk state non-trivial

- Only copy what you need
How Can Netboot Solve These Problems?

- **Slow Provisioning**
  - Up to *Tens of Minutes* to provision

- **Boot Storms**
  - Heavy network traffic

- **Single Point of Failure**
  - Loss of both OS and Application

- **Poor Reusability**
  - Saving and Restoring disk state non-trivial

- Only copy what you need
- Multiple NICs and Distributed File System
How Can Netboot Solve These Problems?

- **Slow Provisioning**
  - Up to *Tens of Minutes* to provision

- **Boot Storms**
  - Heavy network traffic

- **Single Point of Failure**
  - Loss of both OS and Application

- **Poor Reusability**
  - Saving and Restoring disk state non-trivial

- Only copy what you need
- Multiple NICs and Distributed File System
- Reboot from a saved image
Wait, but what about application performance?

- Won’t there be overhead due to constant access to the boot drive over the network?
Wait, but what about application performance?

- Won’t there be overhead due to constant access to the boot drive over the network?

With TenGigabitEthernet and Fast and Reliable Distributed Storage, is this really a problem?

- Separate Communication/Data and Provisioning Networks.

Also, how big of a performance issue is it to have remote boot drives?

- In cloud, data already coming over the network.
M2: Malleable Metal as a Service

A Multi-tenant Bare Metal Cloud Service
M2 Architecture Overview

- Previously developed
- Bare Metal Allocation
- Network Allocation (layer 2)
M2 Architecture Overview

- Data Store
- Pre-Installed Images
M2 Architecture Overview

- Software iSCSI Server
- TGT Software iSCSI

iSCSI Gateway
M2 Architecture Overview

- Diskless Booting from iSCSI target
M2 Architecture Overview

- HIL
- Orchestration Engine
- CEPH
- REST Service
- iSCSI Gateway
- DHCP
- iPXE TFTP
M2 Architecture Overview

HIL

USER

REST Service

CEPH
M2 Architecture Overview

1. Reserve Nodes

HIL

USER

Reserved Servers

REST Service

CEPH
M2 Architecture Overview

1. Reserve Nodes
   - USER
   - HIL
   - REST Service
   - Reserved Servers

2. Provision Reserved Node
   - CEPH
M2 Architecture Overview

1. Reserve Nodes
2. Provision Reserved Node
3. Clone Golden Image

HIL

Reserved Servers

USER

REST Service
CEPH Interface

CEPH

Cloned Images
M2 Architecture Overview

1. Reserve Nodes
2. Provision Reserved Node
3. Clone Golden Image
4. Expose Cloned Image as iSCSI Target

Reserved Servers

HIL

USER

CEPH

REST Service

CEPH Interface

iSCSI Gateway

Cloned Images
M2 Architecture Overview

1. Reserve Nodes
2. Provision Reserved Node
3. Clone Golden Image

HIL

Reserved Servers

USER

REST Service

CEPH Interface

iSCSI Gateway
DHCP
iPXE TFTP

4. Expose Cloned Image as iSCSI Target
5. Configure BMI to PXE boot
1. Reserve Nodes

2. Provision Reserved Node

3. Clone Golden Image

4. Expose Cloned Image as iSCSI Target

5. Configure BMI to PXE boot

6. Attach Nodes to Provisioning Network

M2 Architecture Overview

HIL

Reserved Servers

CEPH

Cloned Images

REST Service

HIL Interface

CEPH Interface

iSCSI Gateway

DHCP

iPXE TFTP

USER
M2 Architecture Overview

1. Reserve Nodes
2. Provision Reserved Node
3. Clone Golden Image
4. Expose Cloned Image as iSCSI Target
5. Configure BMI to PXE boot
6. Attach Nodes to Provisioning Network
7. PXE Boot Reserved Nodes

REST Service

- HIL Interface
- CEPH Interface

- iSCSI Gateway
- DHCP
- iPXE TFTP

HIL

Reserved Servers

CEPH

Cloned Images
M2 Interfaces

- Provision, Deprovision
- Create/Remove Snapshot
- List/Delete Images and Snapshots
Evaluation Environment

- 8 nodes with
  - 128 GB RAM
  - 24 HT-cores
  - Single 10 Gbps NIC (communication and boot drive)

- Ceph cluster
  - 10 nodes (Infiniband interconnect)
  - 90 OSD’s
  - 40Gbps outlink

- Software iSCSI and M2 all running in a VM
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 25 Minutes
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Foreman Provision

Node Power Cycle
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

- Foreman Provision
- Power-on Self-test (POST)
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Foreman Provision

PXE Request
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Foreman Provision

Kernel Download & Local Disk Installation
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Power-on Self-test (POST) & PXE Request
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Foreman Provision

Booting OS from Local Disk
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 25 Minutes
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 25 Minutes
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 25 Minutes

Foreman

Power-on Self-test (POST) & PXE Request

M2 Provision
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 25 Minutes

Foreman

OS Chain Booting (iPXE)

M2 Provision
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 25 Minutes

Hadoop Package Installation and Configuration
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Foreman

~ 25 Minutes

M2 Provision

~ 11 Minutes
Poor Reusability

Tenant 1

Tenant 2
Poor Reusability

Tenant 1

Tenant 2
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

Foreman Provision

Foreman Re-Pro

~ 25 Minutes
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

- M2 Provision
  - ~ 11 Minutes

- M2 Re-Pro
  - ~ 5 Minutes 30 Seconds

- Hadoop Package Installation overhead removed ( ).
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

~ 5 Minutes 30 Seconds

~ 25 Minutes

~ 5X
BMI Reduces Provisioning/Re-Provisioning Times.

POST ( ) dominates BMI provisioning time.

Provisioning/Re-Provisioning Times Comparison (Single Hadoop Node)

- Foreman: ~ 25 Minutes
- M2: ~ 5 Minutes 30 Seconds

- BMI Reduces Provisioning/Re-Provisioning Times.
- POST ( ) dominates BMI provisioning time.
Provisioning/Re-Provisioning Times Comparison
(Single Hadoop Node)

- Without POST ( ) Re-Provision time ~2 Minutes.
- ~2 Minutes to spin up Virtual machines in AWS.
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB

~ 170 MB of 10GB Image (< 2% of the Boot Drive is accessed)
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB

~ 3KB/s after initial provisioning
Cumulative Read Traffic Analysis (24 Hours - Hadoop)

Case (ii) : Data > RAM
- RAM: 128 GB
- Data: 256 GB
Cumulative Write Traffic Analysis (24 Hours - Hadoop)

Case (i) : Data == RAM
- RAM: 128 GB
- Data: 128 GB

Mostly Hadoop Logs (~ 14KB/s)
Cumulative Write Traffic Analysis (24 Hours - Hadoop)

Case (ii) : Data > RAM
- RAM: 128 GB
- Data: 256 GB
Network traffic to boot drive is irrelevant!

- Resilient to boot storm
NAS Parallel Benchmark Performance (HPC)

Negligible Performance Impact

![Performance Impact Graph]

- CG on Local Disk
- CG on M2
- FT on Local Disk
- FT on M2
- IS on Local Disk
- IS on M2

Runtime (secs)

Number of MPI Ranks
Hadoop Performance (Big Data)

Negligible Performance Impact

![Diagram showing performance impact with varying data set sizes (8GB, 16GB, 32GB, 64GB, 128GB) for different operations: WordCount, Sort, Grep, with two types of storage: Local Disk and M2. The diagram compares the elapsed time in seconds for each operation and data set size.]
- Performance Impact Due to Remote Boot Drive is Negligible!
M2 Stress Test - Concurrent Provisioning

- **M2 VM Configuration**
  - 4 vCPUS
  - 4 GB RAM
  - TGT Software iSCSI
  - No iSCSI Caching

- **Provisioning Time (secs)**
- **Concurrently Provisioned Servers**

- **Total Provisioning Time**
- **M2 API Overhead**
More About M2

- M2 core is ~3500 lines of new python code (i.e. excluding HIL)

- Used at Massachusetts Open Cloud since Fall 2016

- Used in production on dedicated 16 node cluster (with 6 hour lease)
  - OS Research groups
  - Security researchers need access to TPM
  - Experimental installation OpenStack
  - Research groups that need deterministic performance (no virtualization)

- Performance re-provisioning critical for these use cases
Conclusion: M2 is a new bare-metal cloud service

- Overcomes the problems of stateful provisioning systems
  - Rapid provisioning
  - Resilient to boot storms
  - No single points of failure
  - Rapid re-provisioning (Reusability)
- Negligible performance impact


In Progress

- User transparent checkpointing of the node memory state.
- Rapid attestation of bare metal systems before re-assigning them.
- Using rapid multiplexing frameworks to improve datacenter utilization.
Acknowledgements

- **Current and previous core M2 team members:** Naved Ansari, Sourabh Bollapragada, Daniel Finn, Paul Grosu, Sirushti Murugesan and Pranay Surana.

- **Other Mass Open Cloud (MOC) team members:** Chris Hill, Laura Kamfonik, Rajul Kumar, Radoslav Nikiforov Milanov, Piyanai Saowarattitada, and Rahul Sharma.

- MassTech Collaborative Research Matching Grant Program, National Science Foundation awards ACI-1440788, 1347525, 1149232 and 1414119.

- Massachusetts Open Cloud commercial partners Brocade, Cisco, Intel, Lenovo, Red Hat, and Two Sigma.
Thank You

M2 is an open source projects.

We welcome you to contribute, use and provide us with feedback and suggestions to improve it.

https://github.com/CCI-MOC/ims
https://info.massopencloud.org/blog/bare-metal-imaging/
Questions?

1. Reserve Nodes
2. Provision Reserved Node
3. Clone Golden Image
4. Expose Cloned Image as iSCSI Target
5. Configure BMI to PXE boot
6. Attach Nodes to Provisioning Network
7. PXE Boot Reserved Nodes

- REST Service
  - HIL Interface
  - CEPH Interface
  - iSCSI Gateway
  - DHCP
  - TFTP iPXE

- Cloned Images
- Reserved Servers