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)

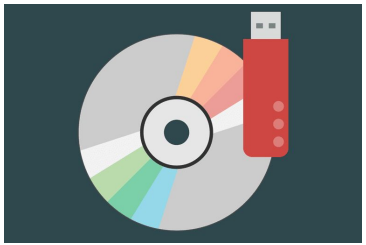
Increasing Bare Metal Cloud Offerings

- 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

aws openstack. emulab MAAS rackspace the open cloud company SOFTLAYER an IBM Company



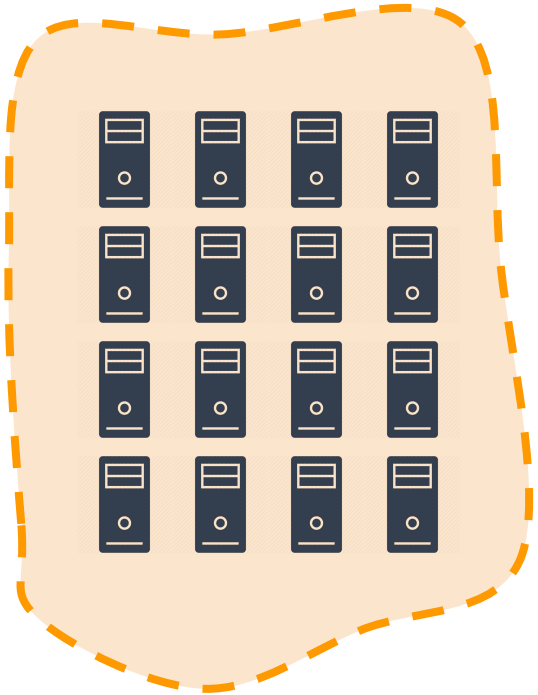
COPY

Stateful Provisioning Problems

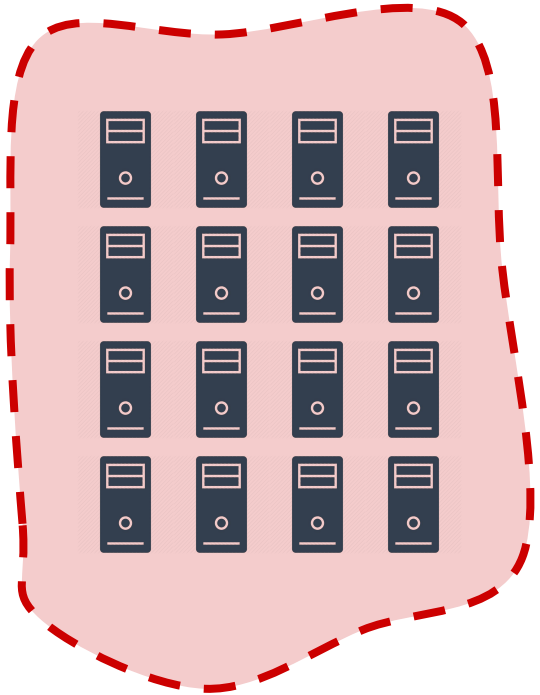
- ***Slow Provisioning***
  - Up to *Tens of Minutes* to provision
- ***Boot Storms***
  - Heavy network traffic
- ***Single Point of Failure***
  - Loss of both OS and Application
- ***Poor Reusability***
  - Saving and Restoring disk state

Poor Reusability

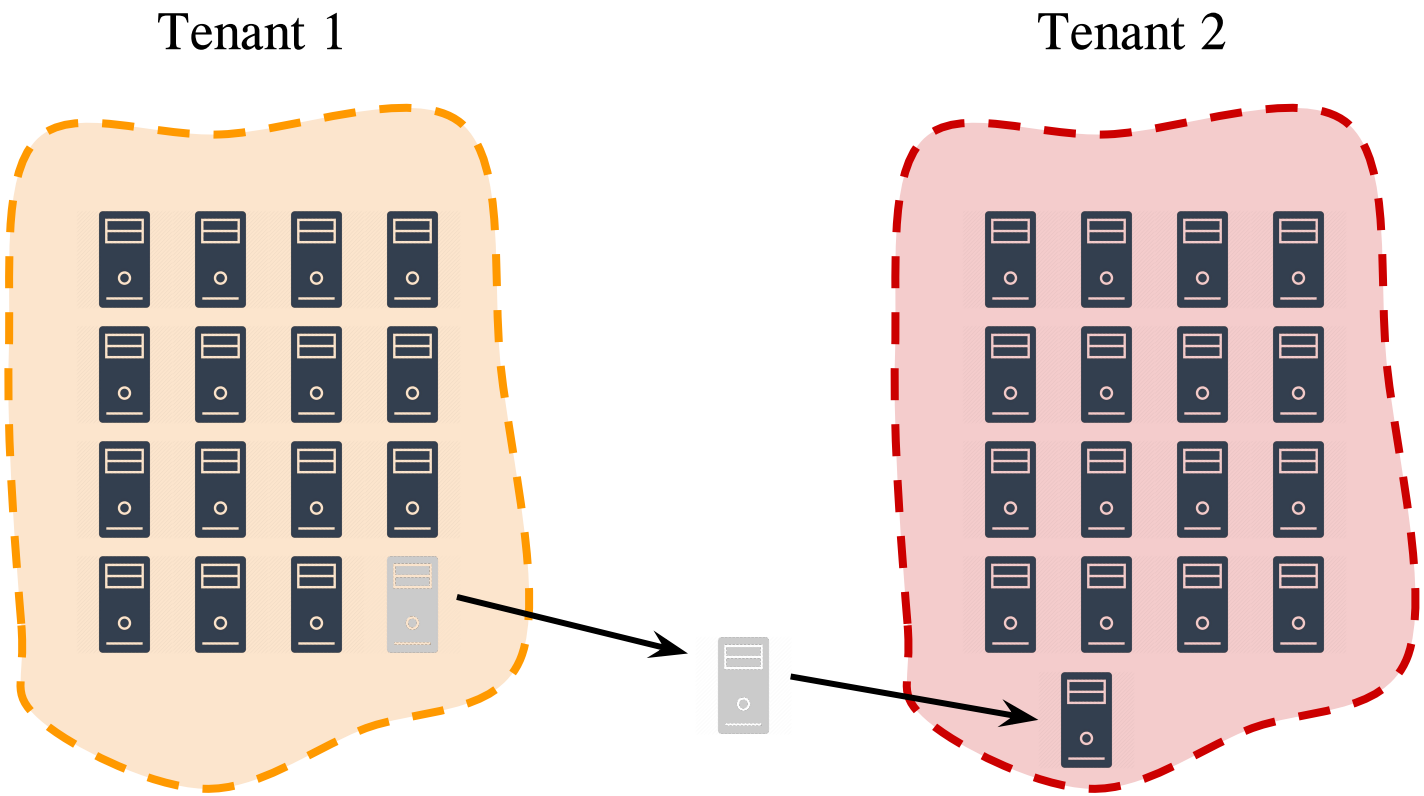
Tenant 1



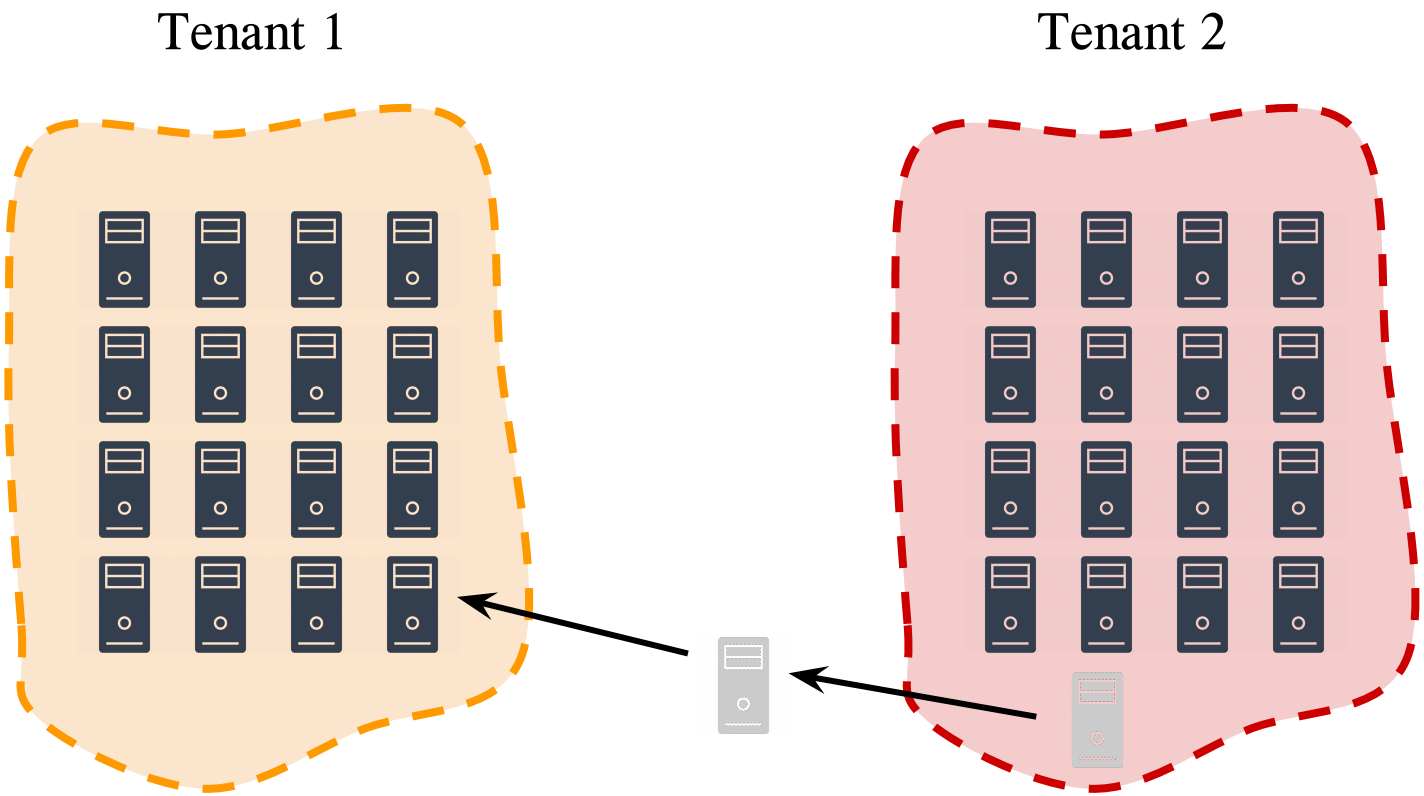
Tenant 2



Poor Reusability



Poor Reusability



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)**

Distributed Storage

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

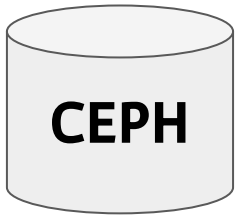
HIL

- ❏ Previously developed
- ❏ Bare Metal Allocation
- ❏ Network Allocation (layer 2)

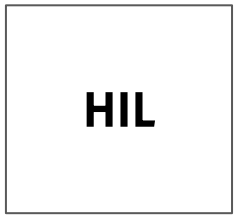
M2 Architecture Overview

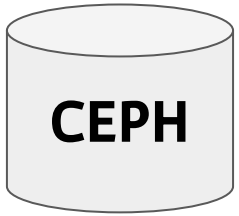
HIL

- Data Store
- Pre-Installed Images



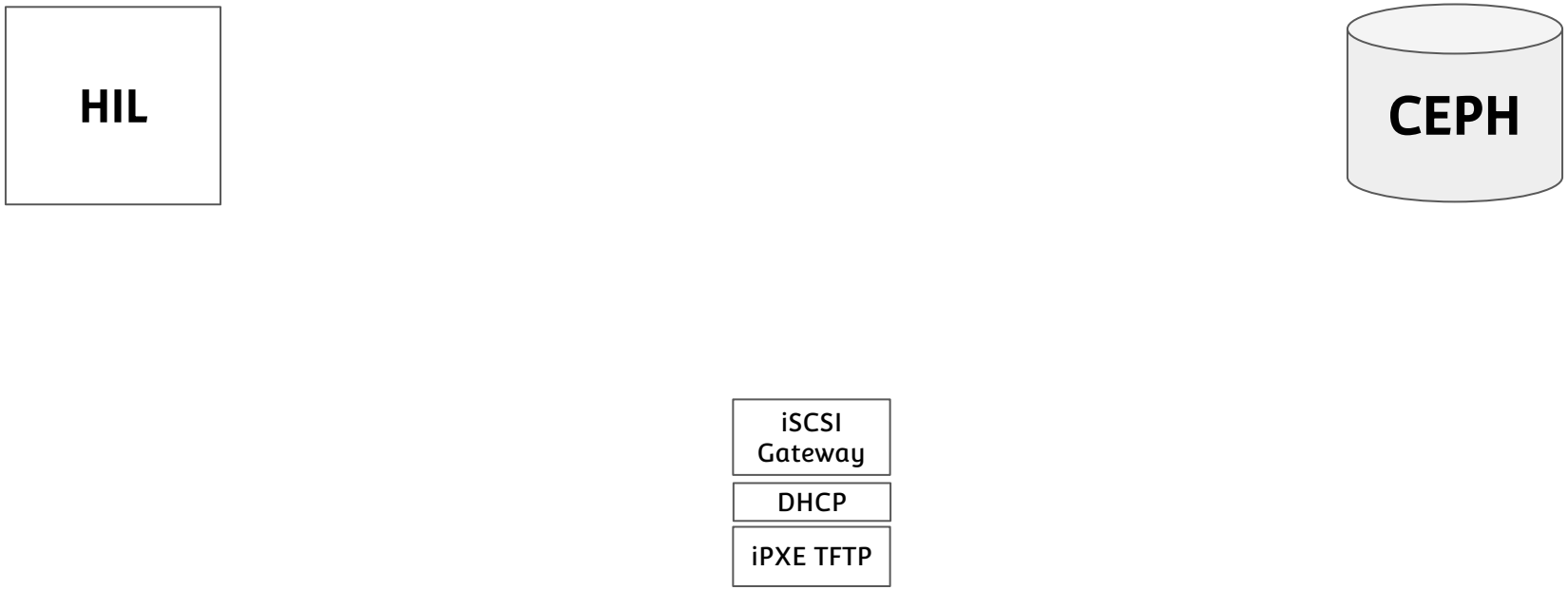
M2 Architecture Overview





- Software iSCSI Server
- TGT Software iSCSI

M2 Architecture Overview



- Diskless Booting from iSCSI target

M2 Architecture Overview

HIL

- Orchestration Engine

REST Service

iSCSI Gateway

DHCP

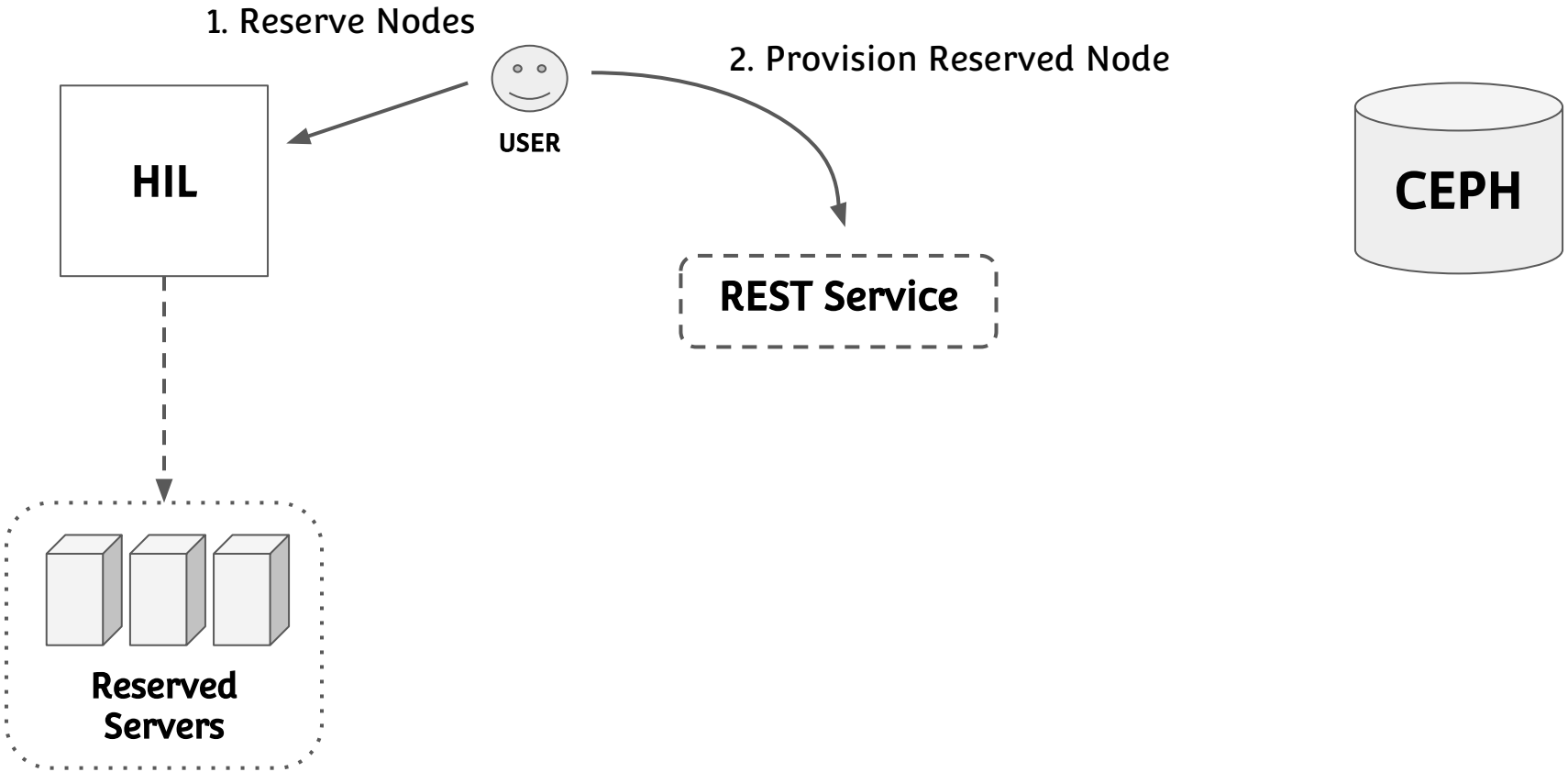
iPXE TFTP

CEPH

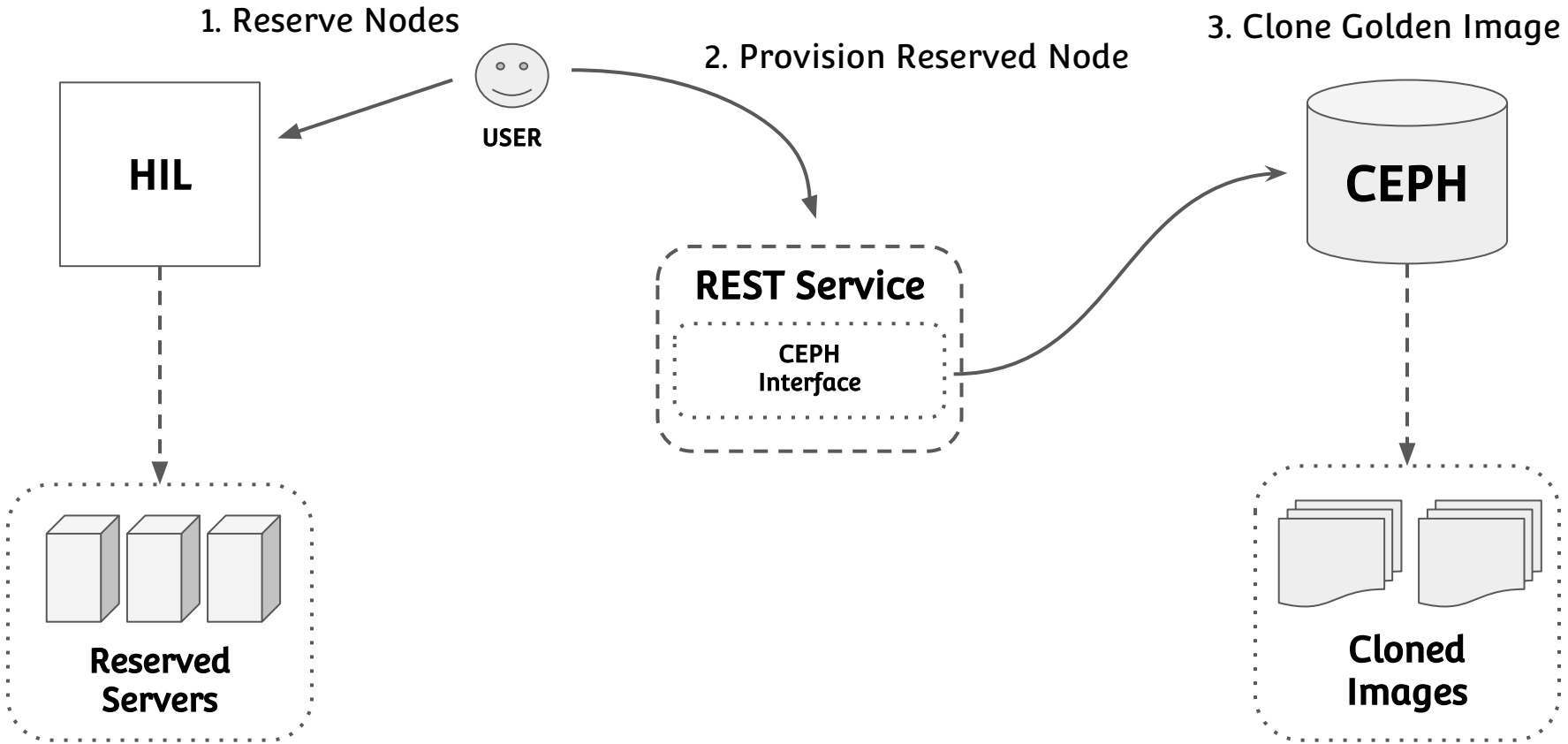
M2 Architecture Overview

M2 Architecture Overview

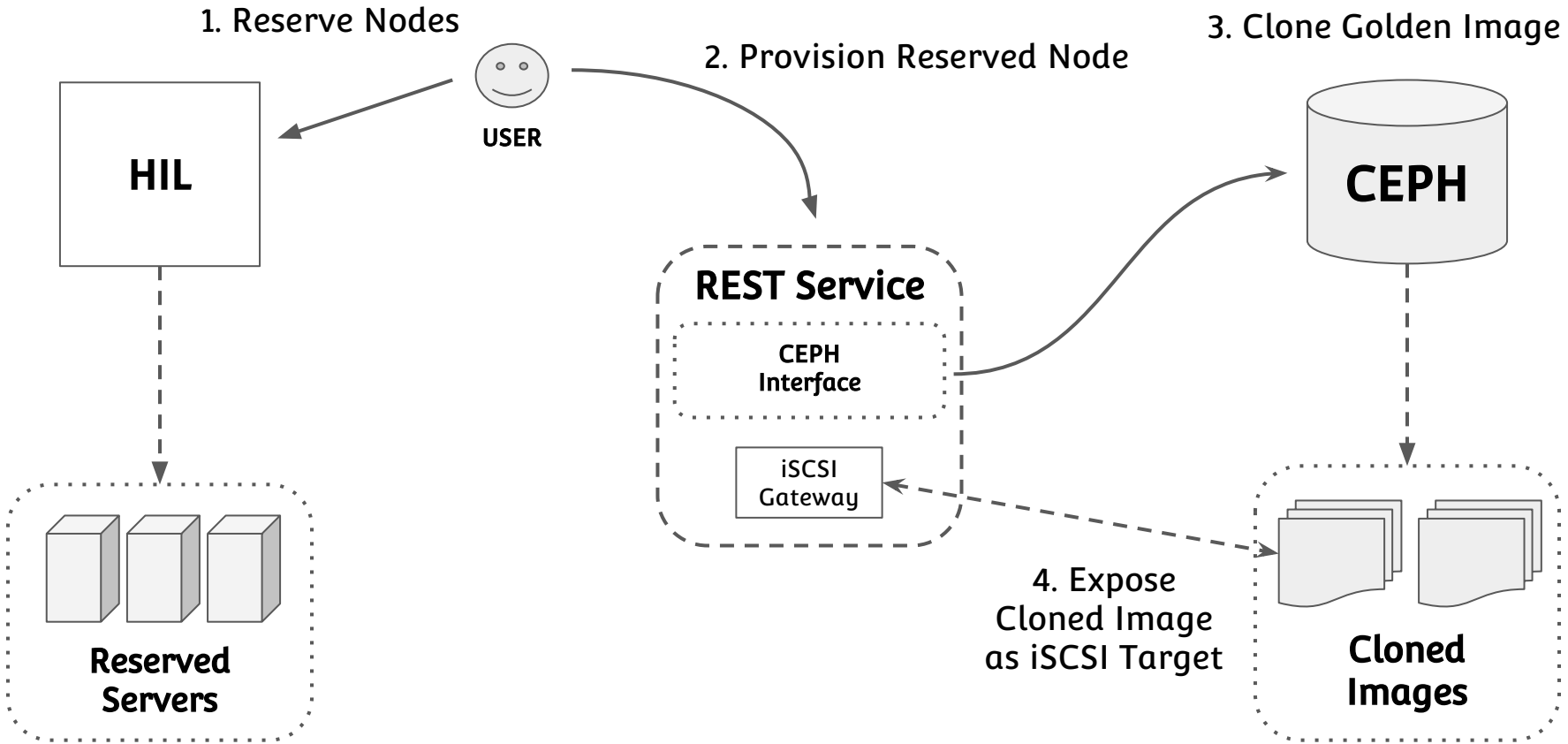
M2 Architecture Overview



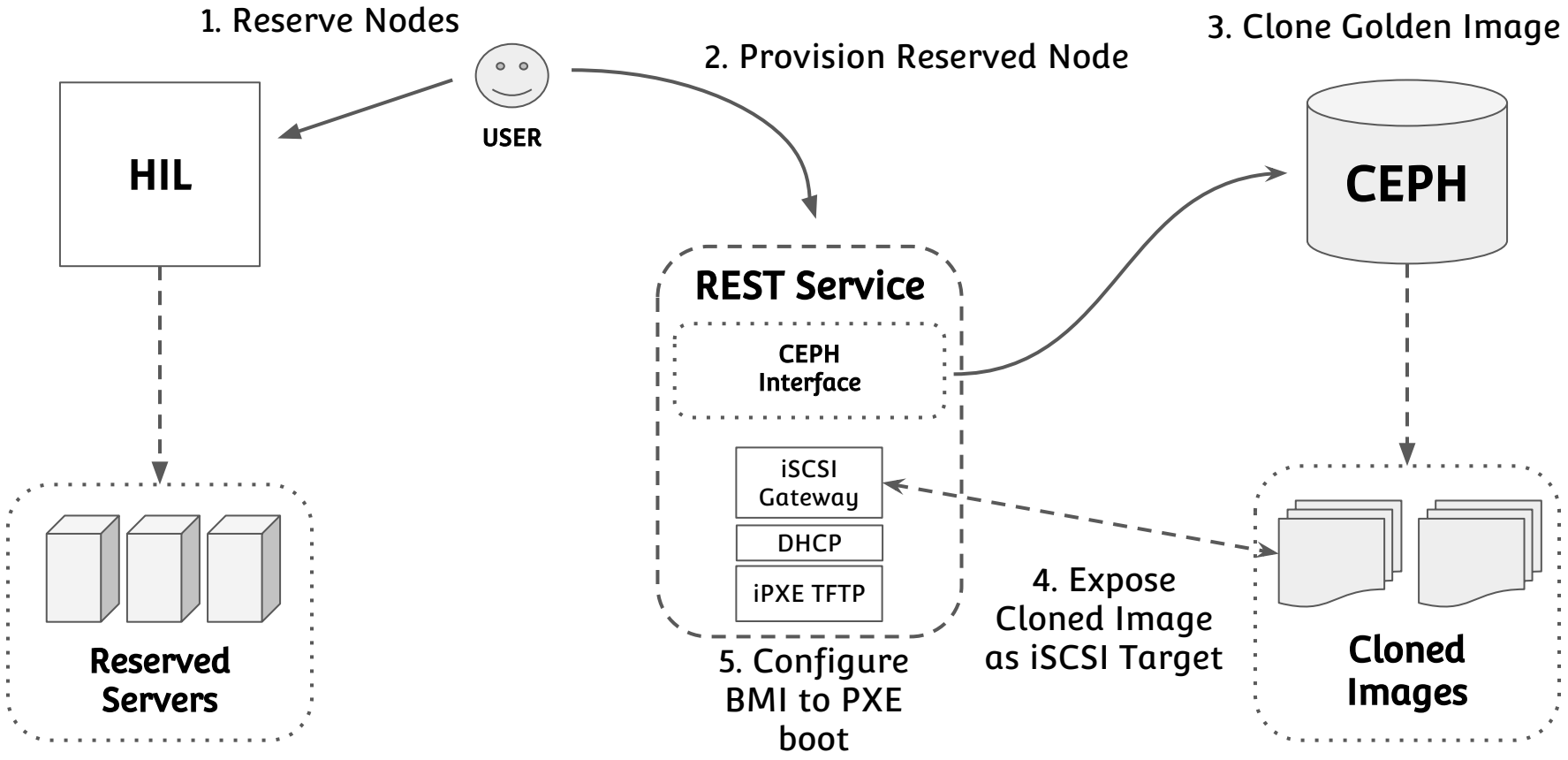
M2 Architecture Overview



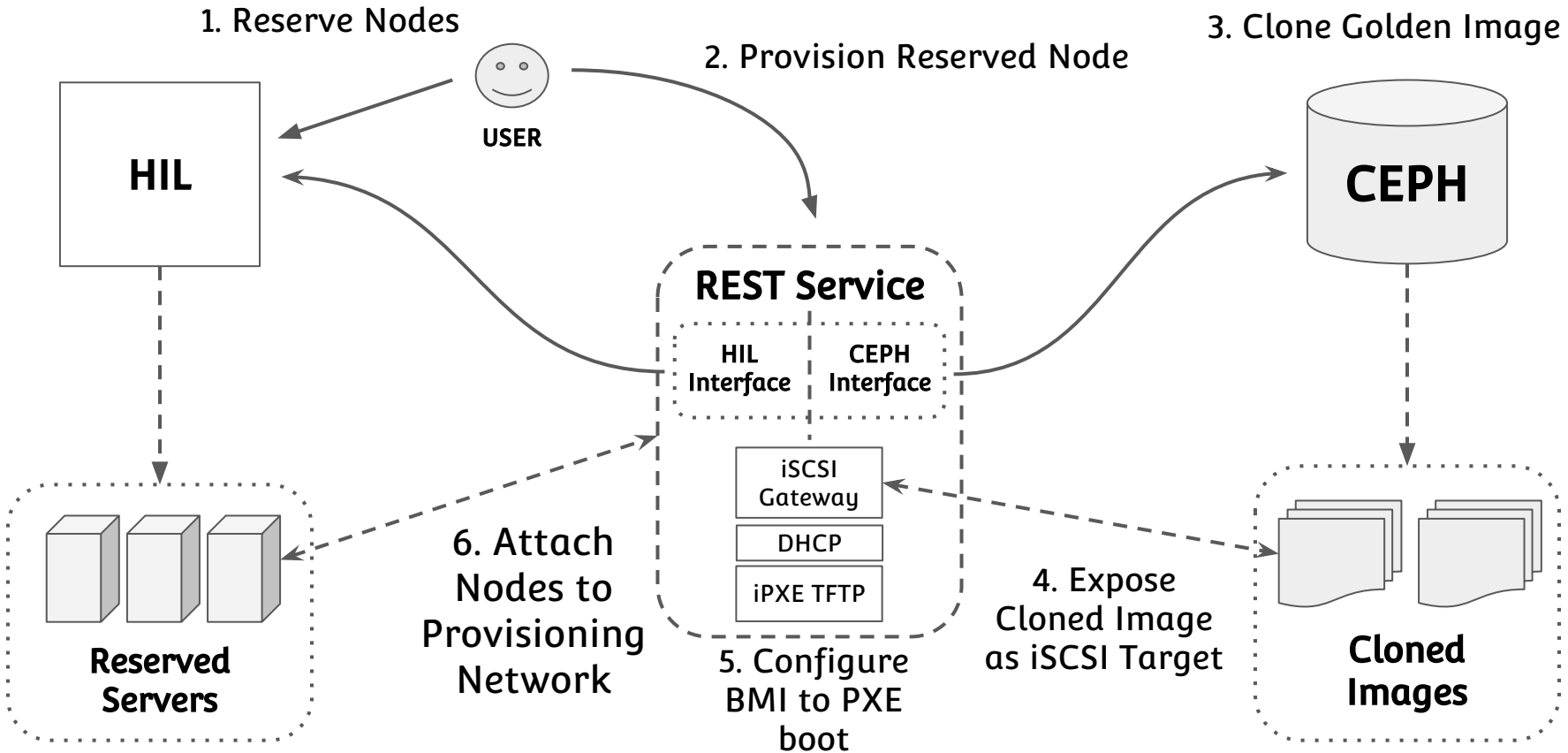
M2 Architecture Overview



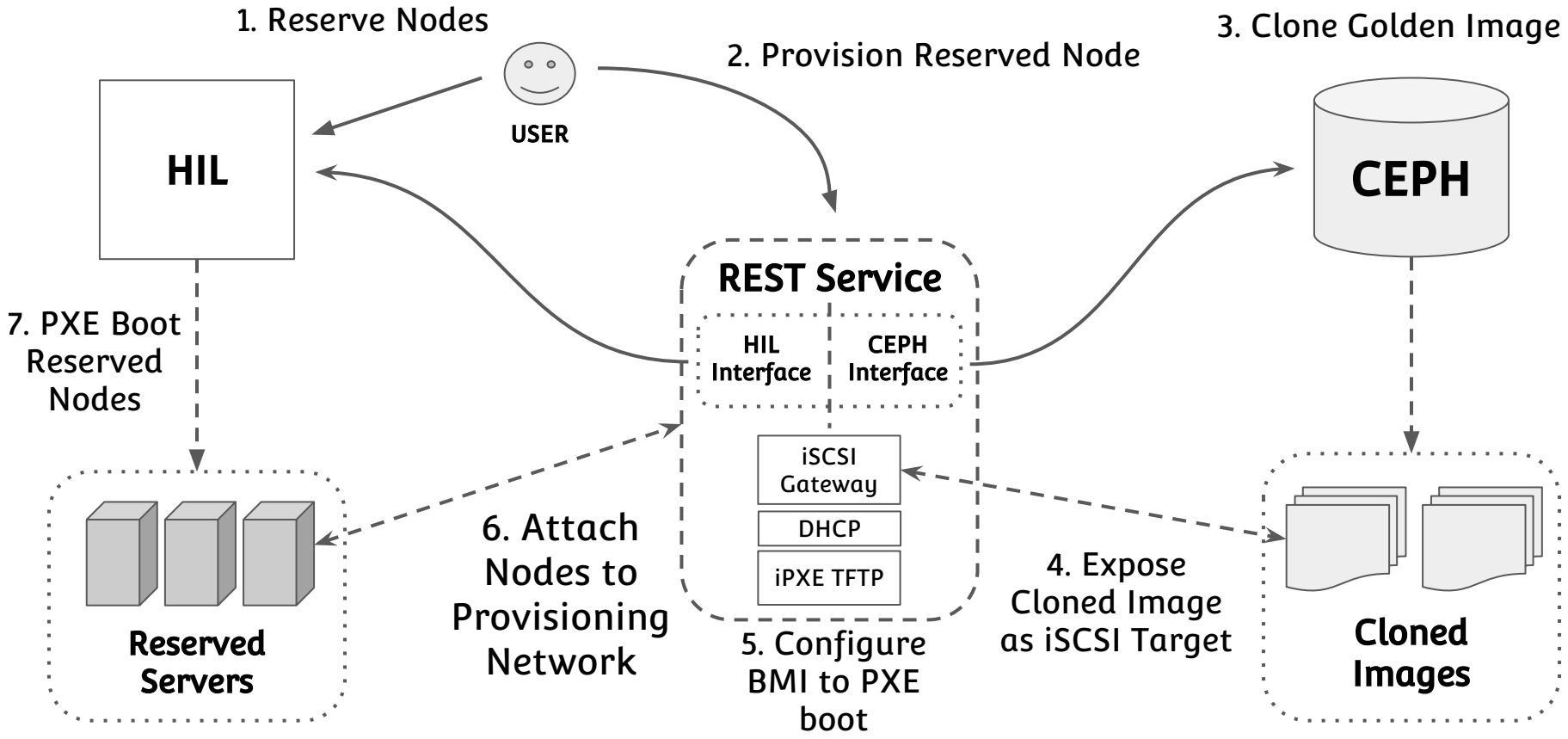
M2 Architecture Overview



M2 Architecture Overview



M2 Architecture Overview



M2 Interfaces

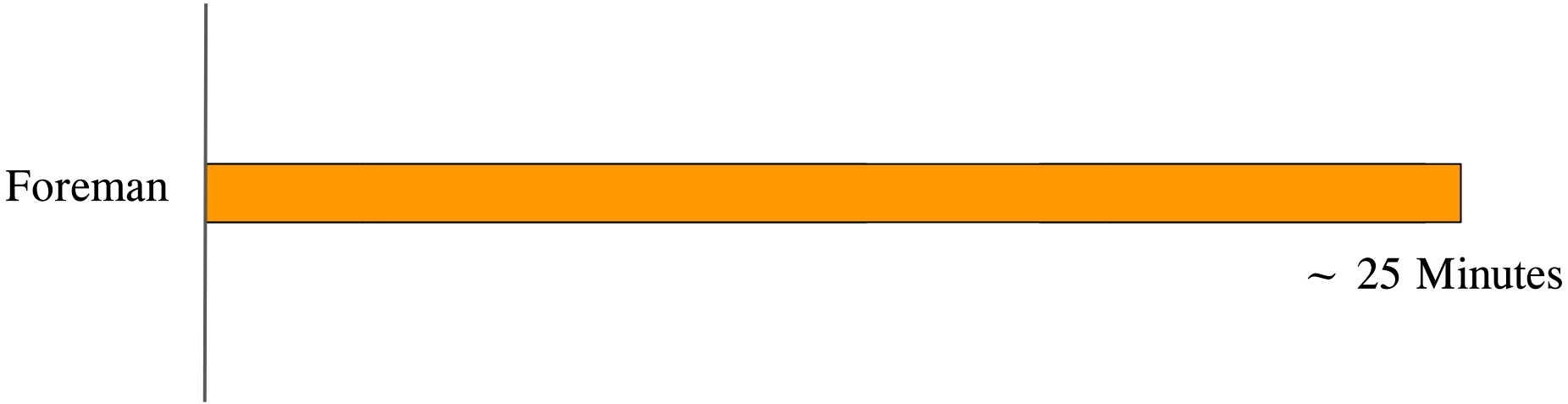
- ❏ Provision, Deprovision
- ❏ Create/Remove Snapshot
- ❏ List/Delete Images and Snapshots

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

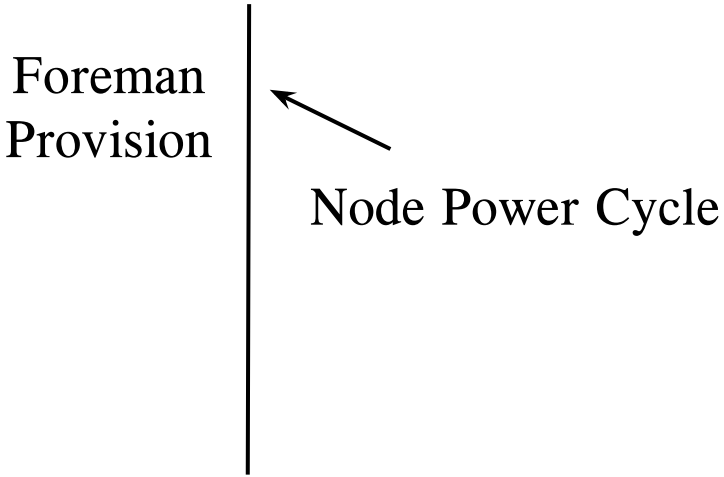
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



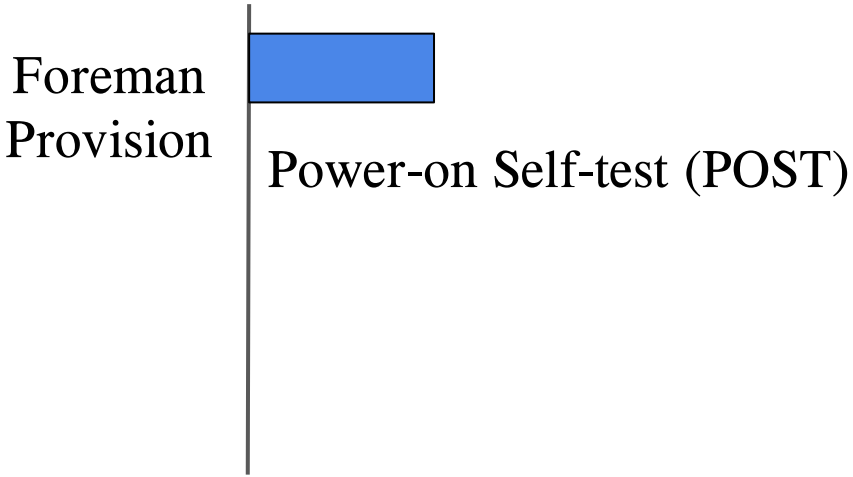
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



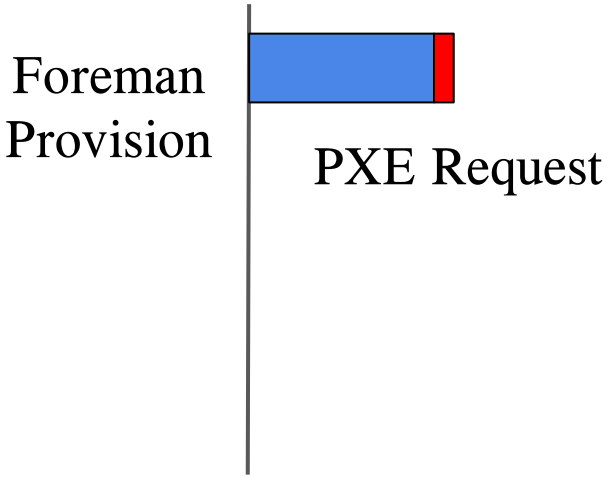
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



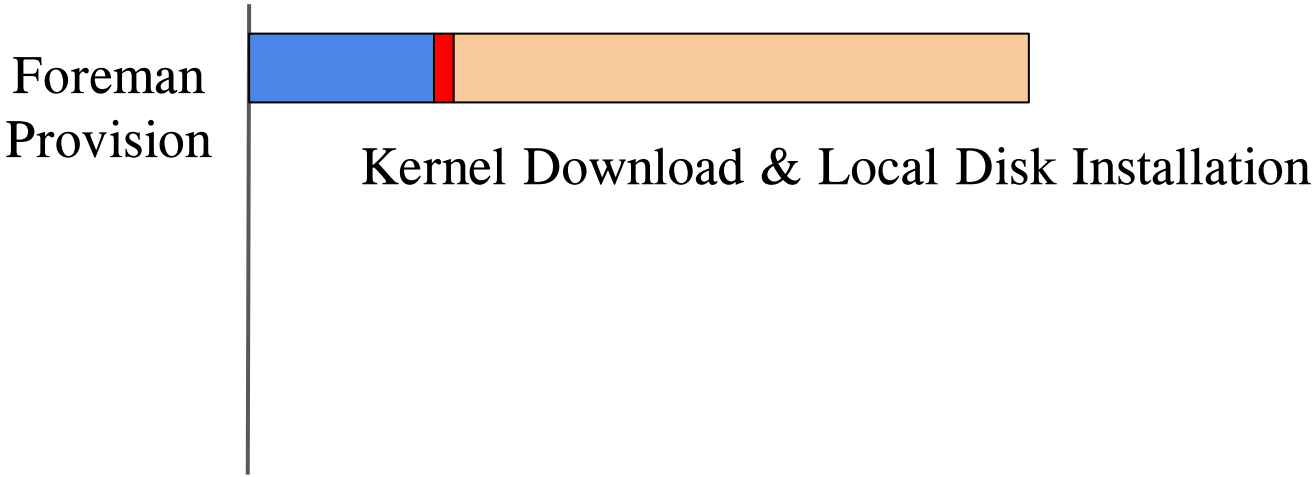
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



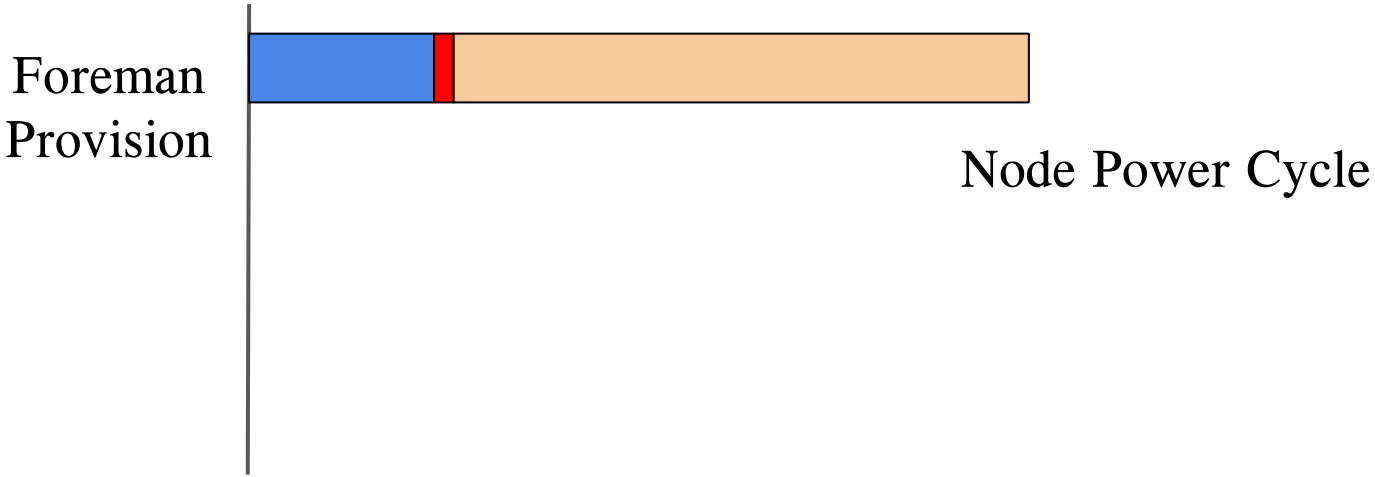
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



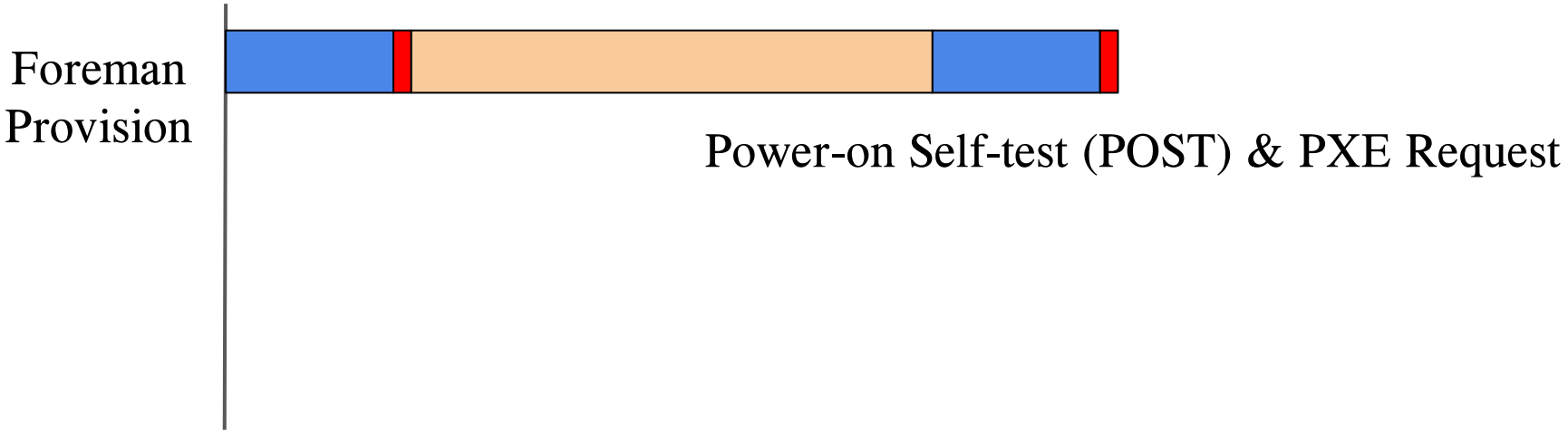
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



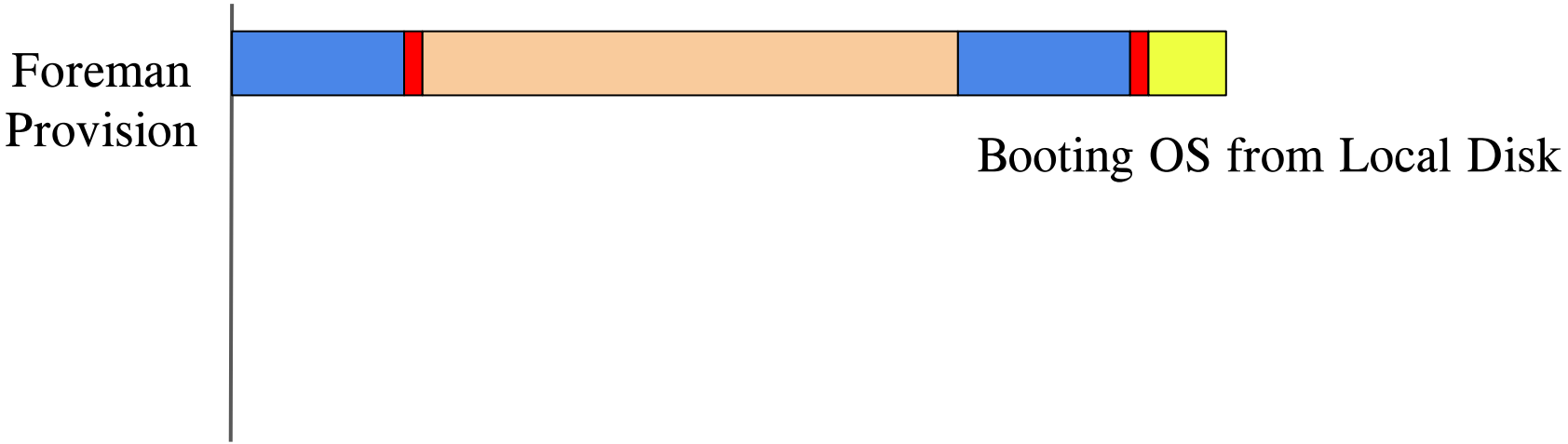
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



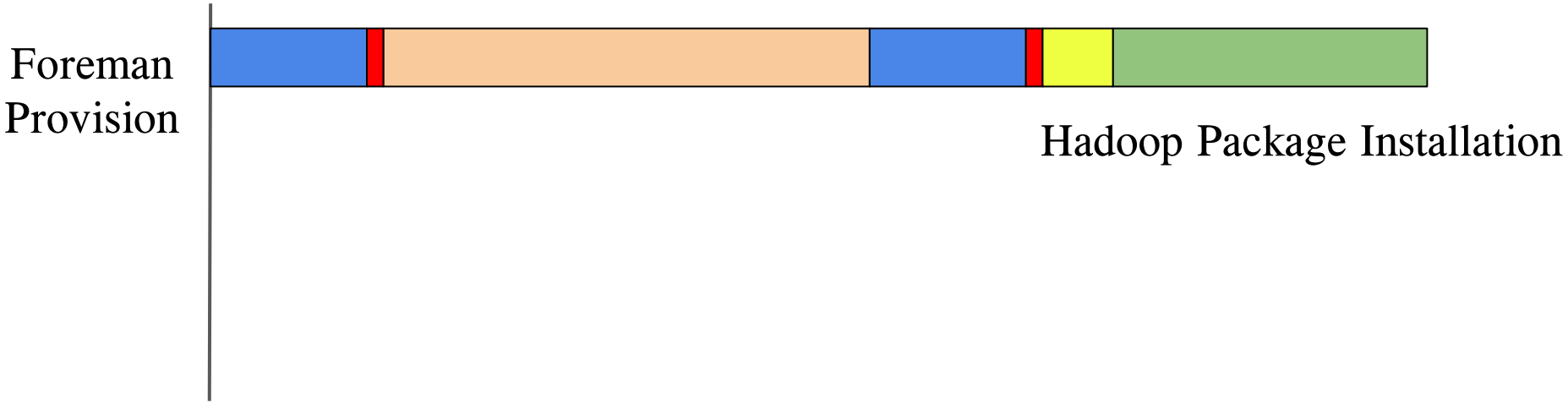
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



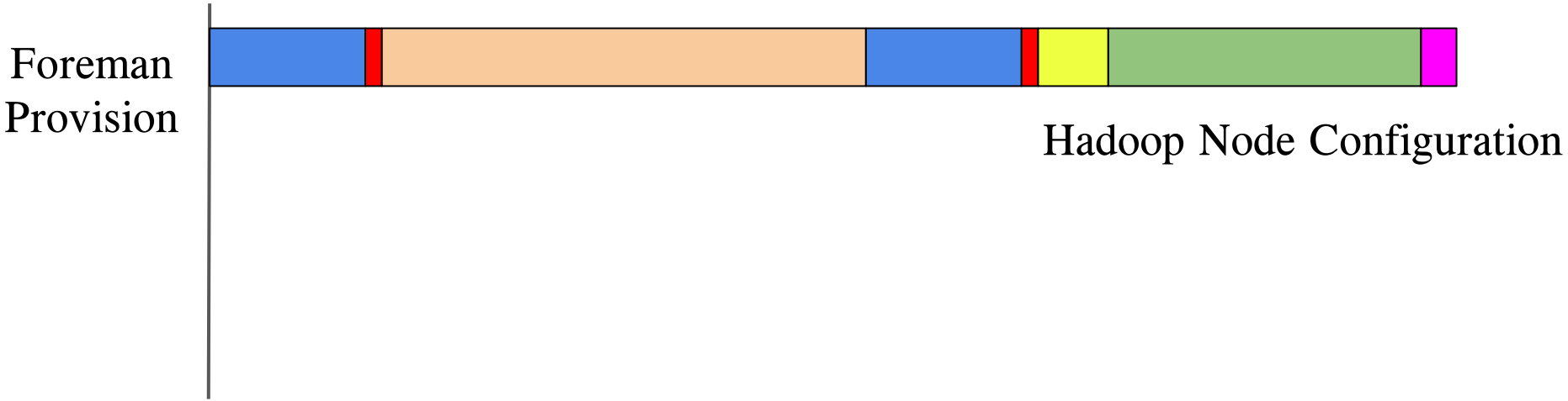
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



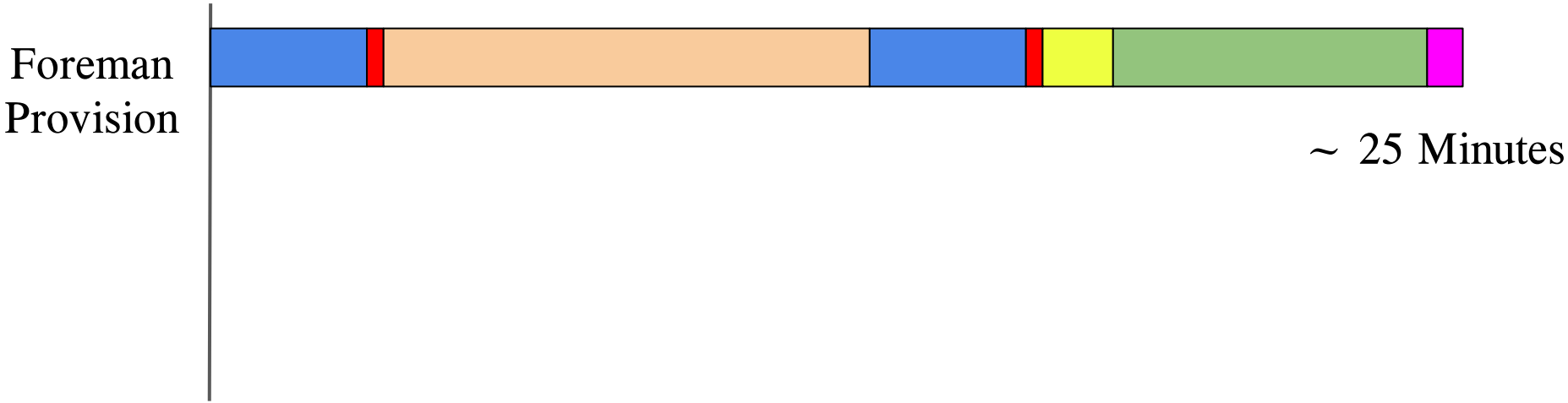
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



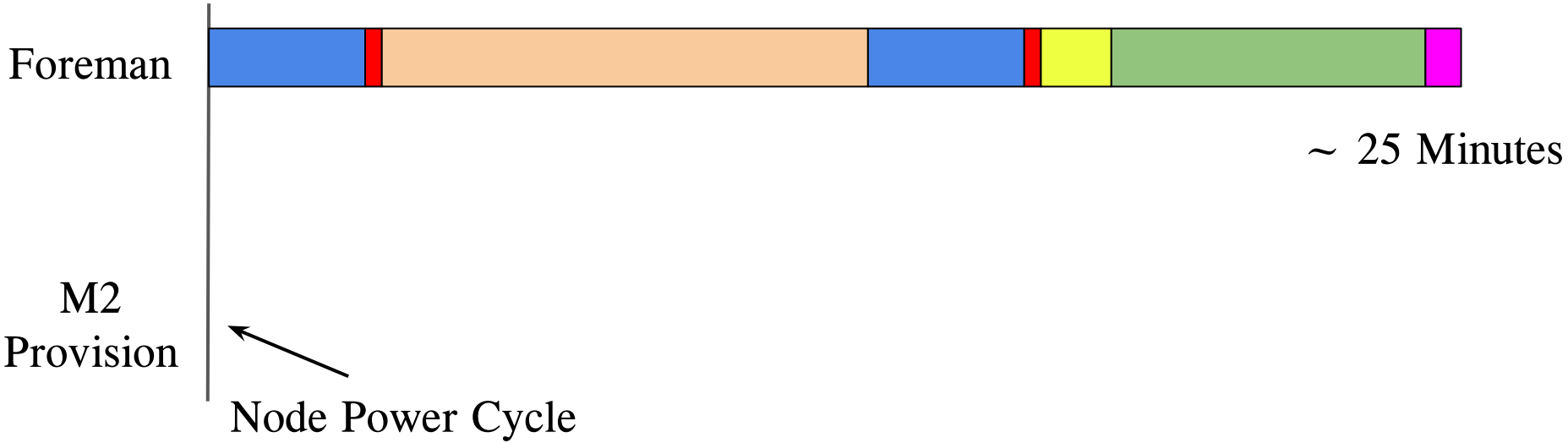
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



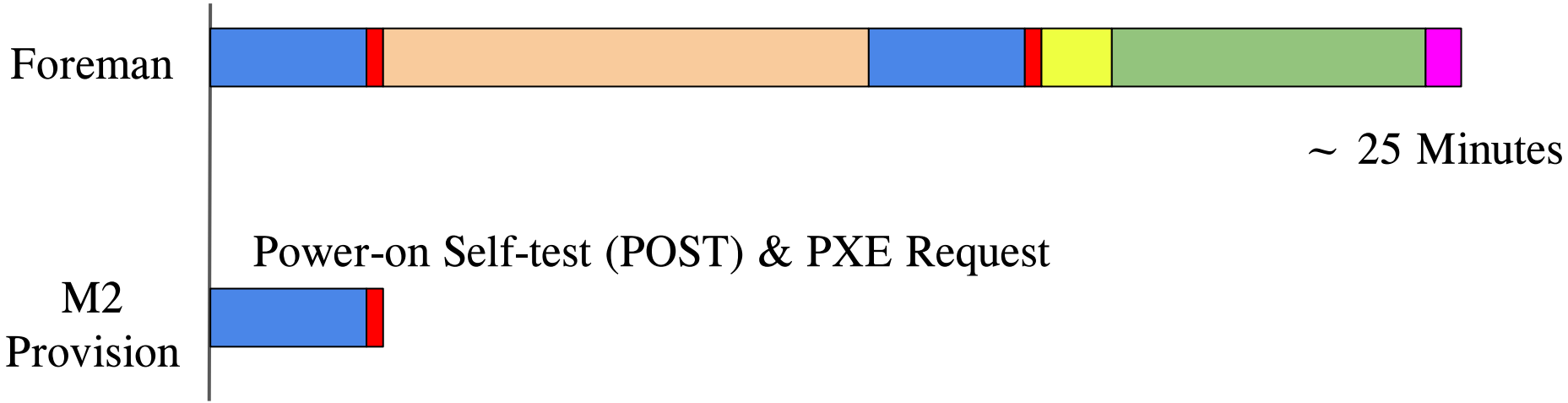
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



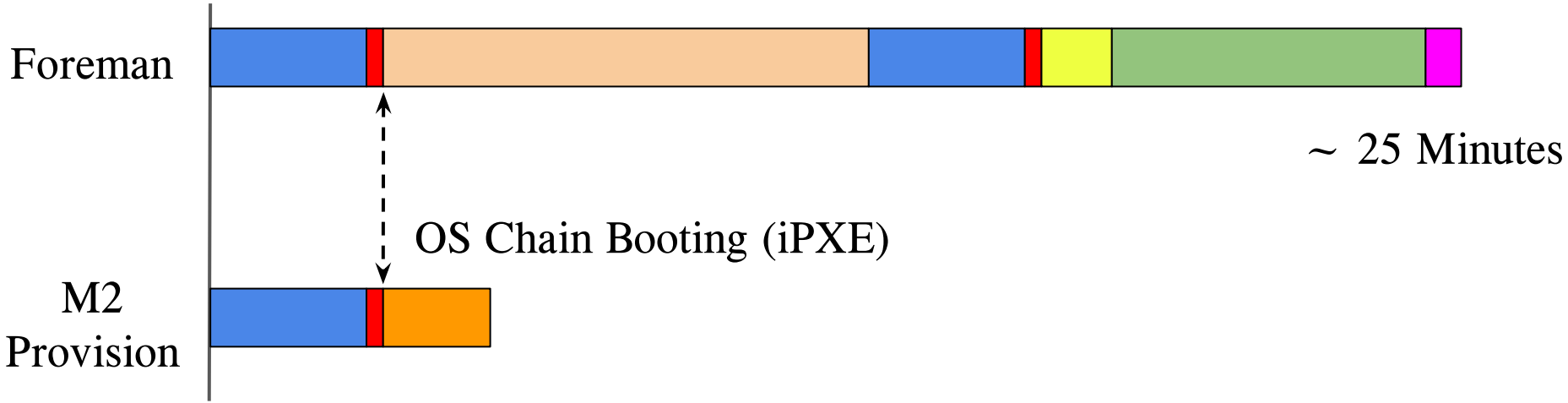
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



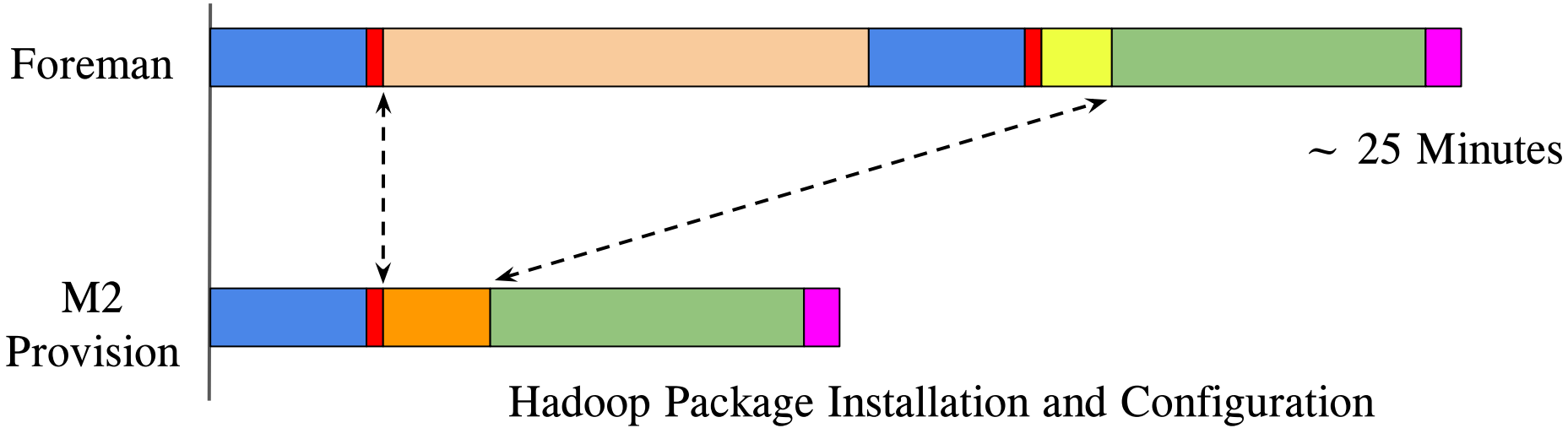
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



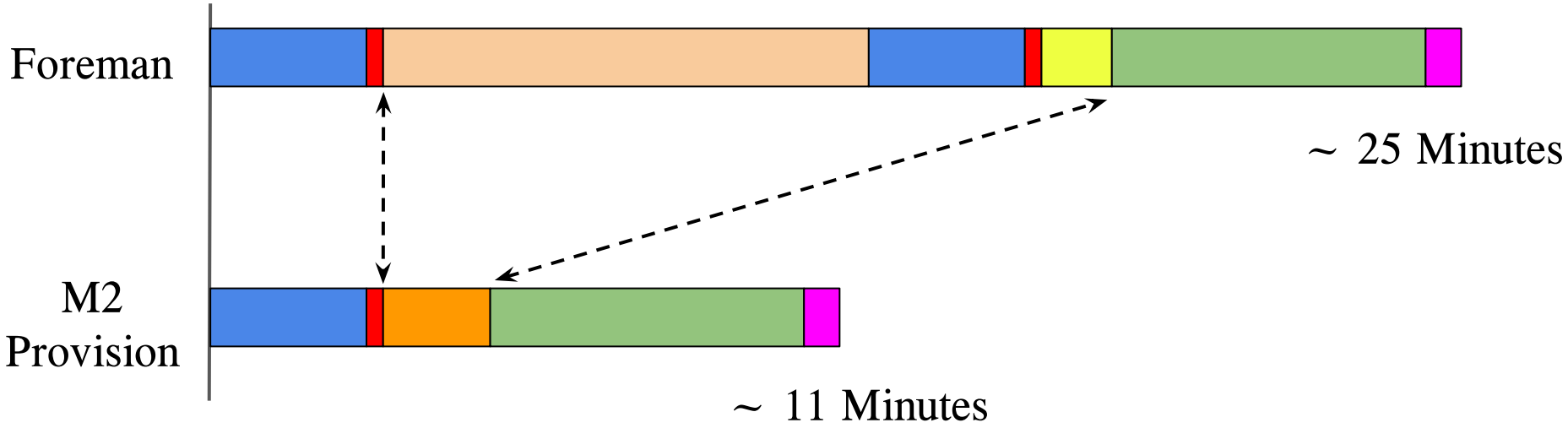
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



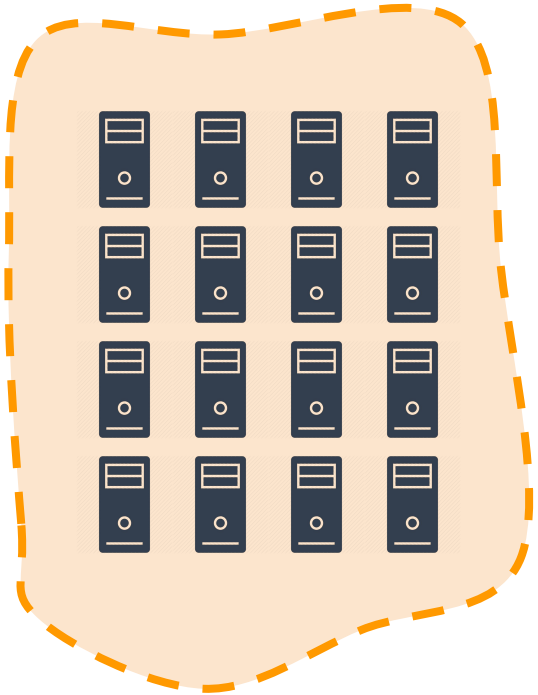
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)

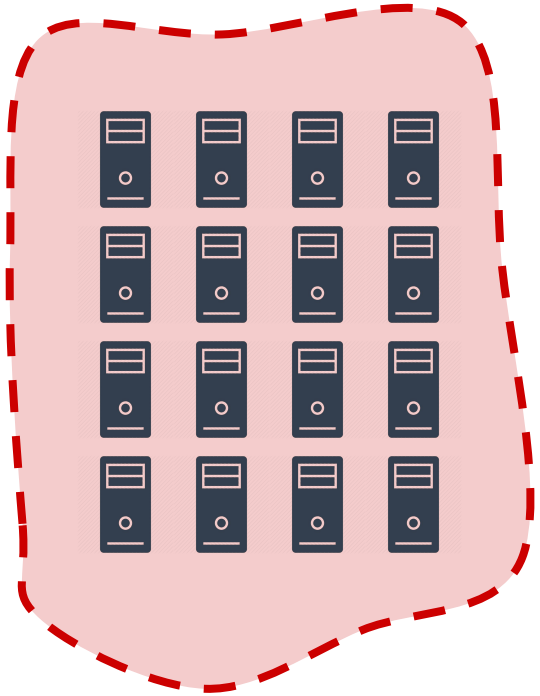


Poor Reusability

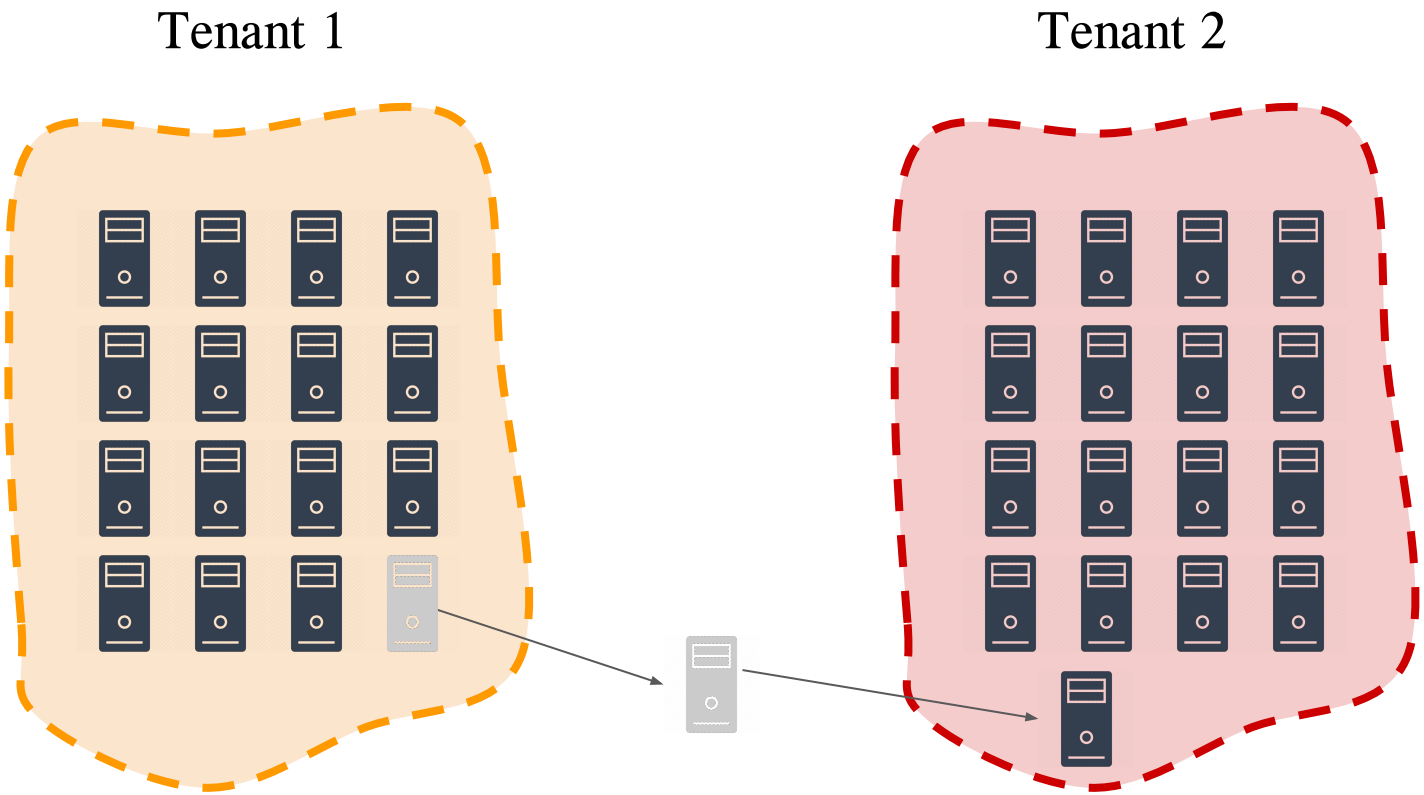
Tenant 1



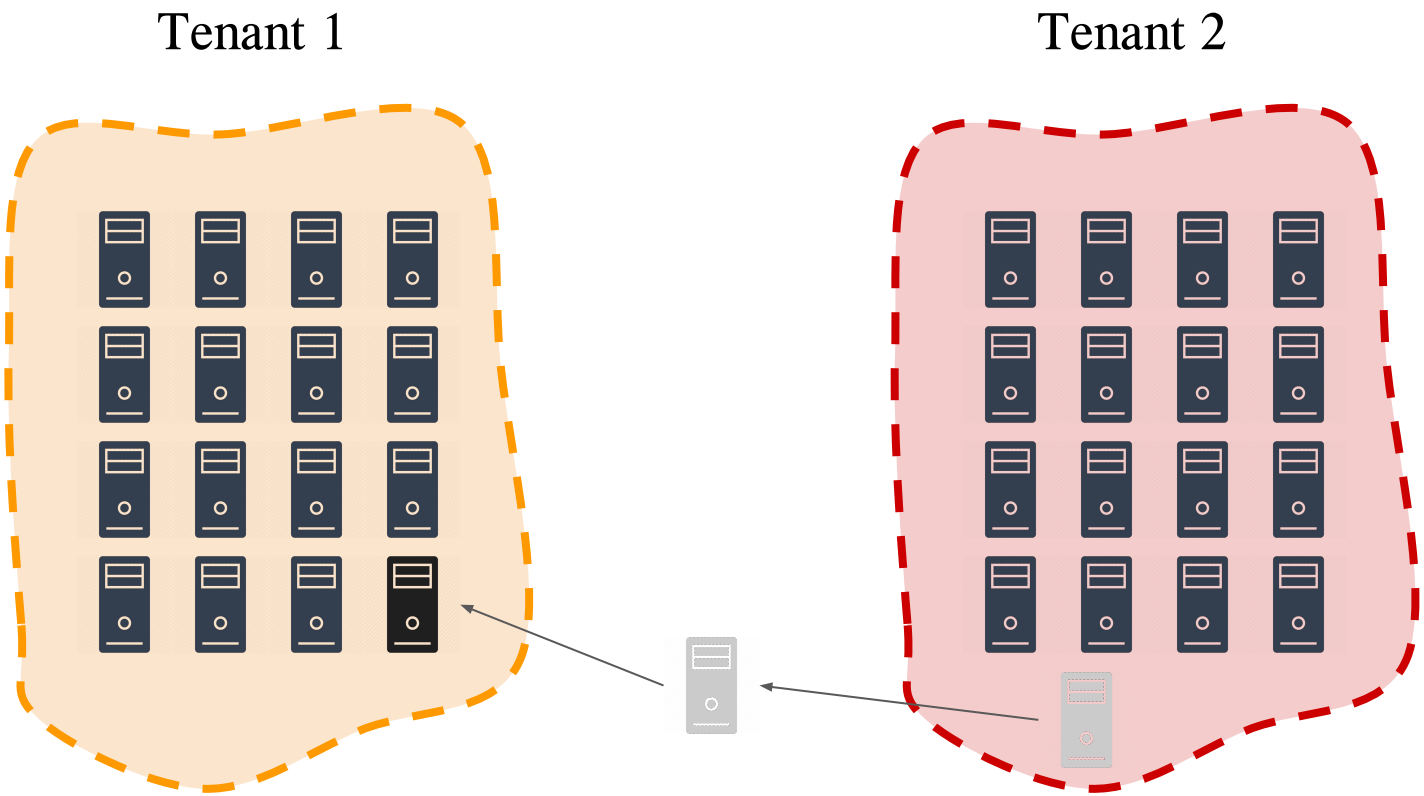
Tenant 2



Poor Reusability

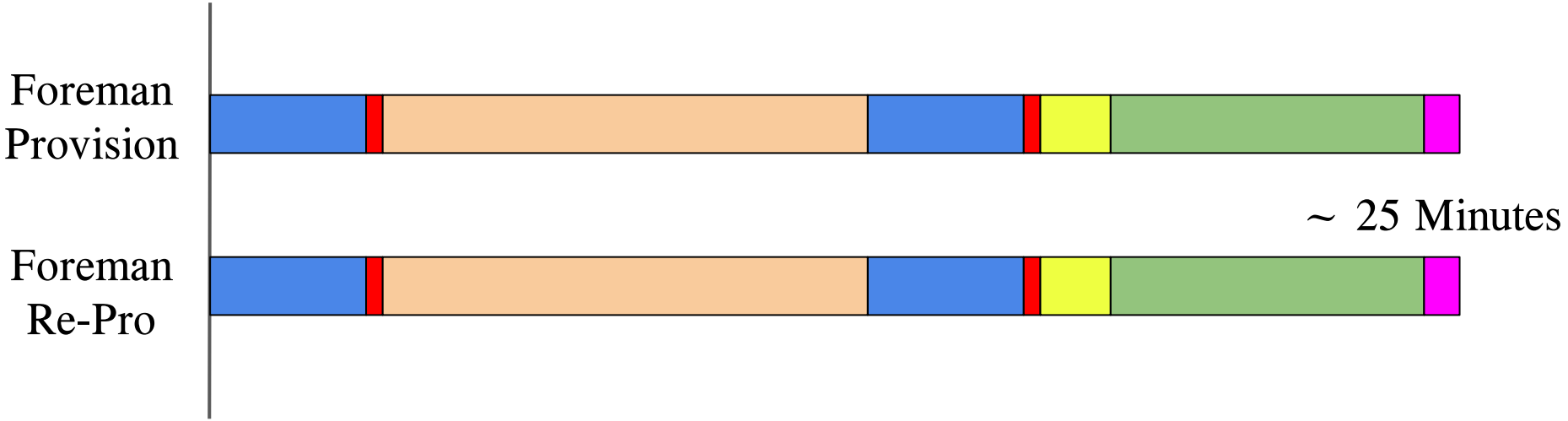


Poor Reusability



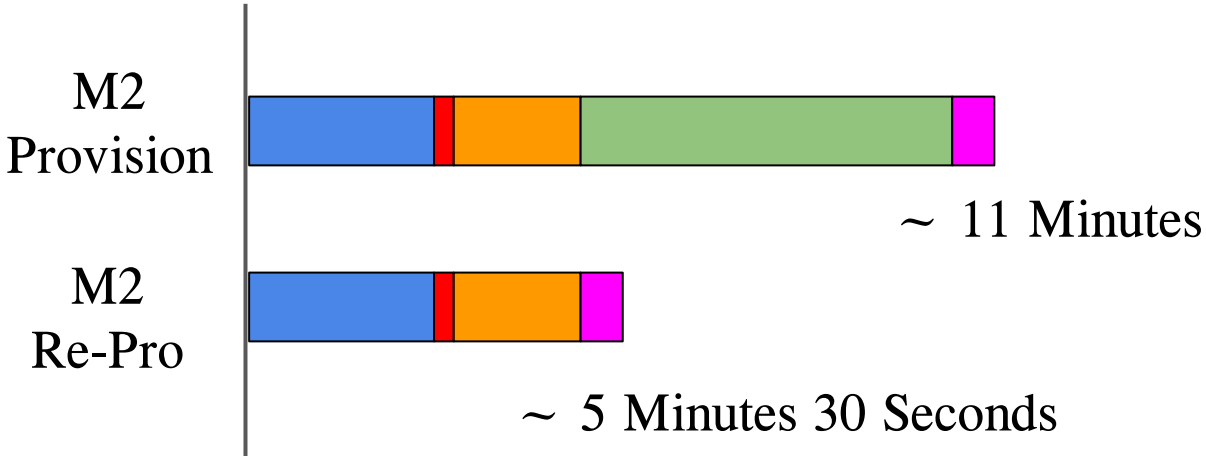
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



Provisioning/Re-Provisioning Times Comparison

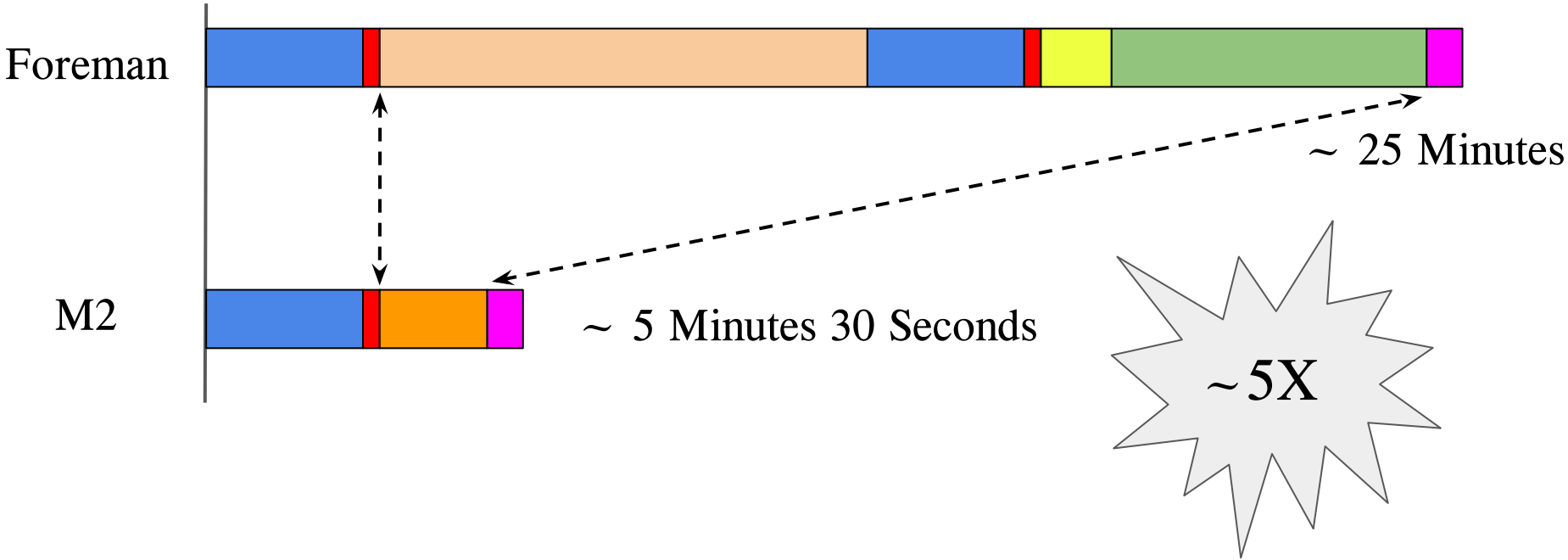
(Single Hadoop Node)



- Hadoop Package Installation overhead removed ( ).

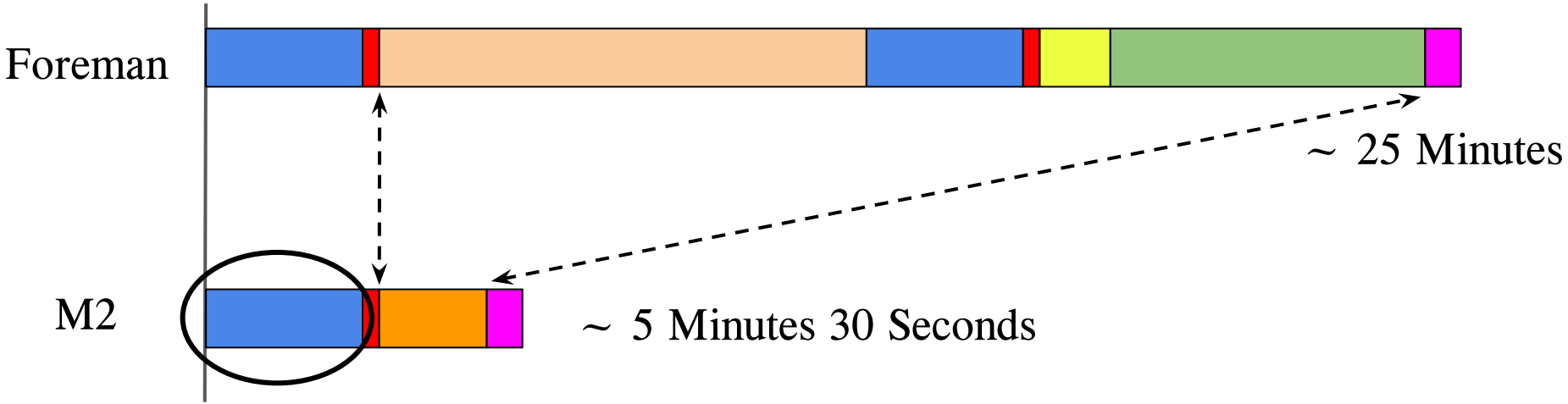
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)



Provisioning/Re-Provisioning Times Comparison

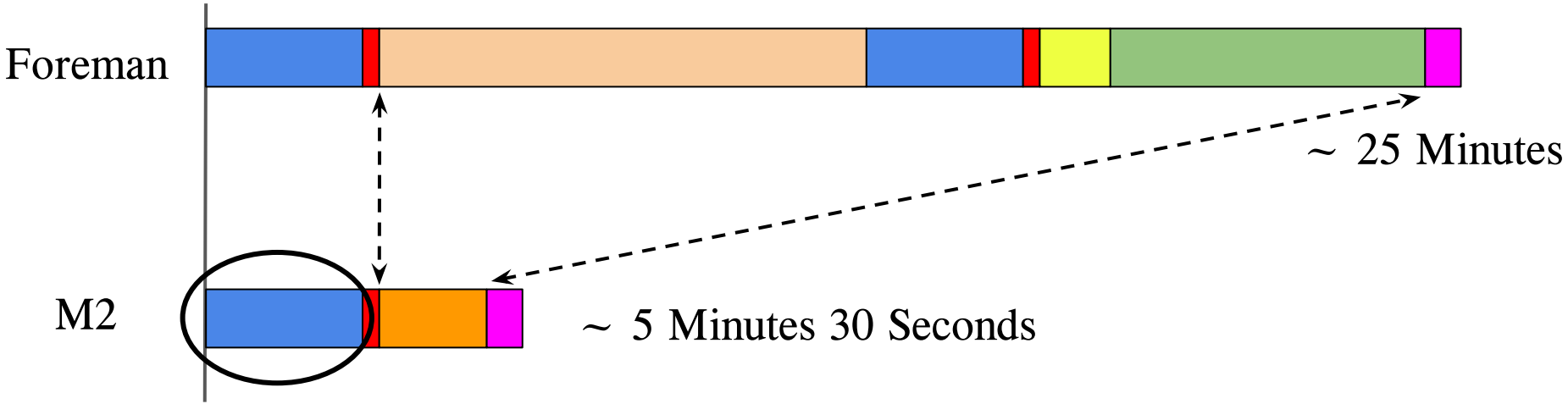
(Single Hadoop Node)



- BMI Reduces Provisioning/Re-Provisioning Times.
- POST ( ) dominates BMI provisioning time.

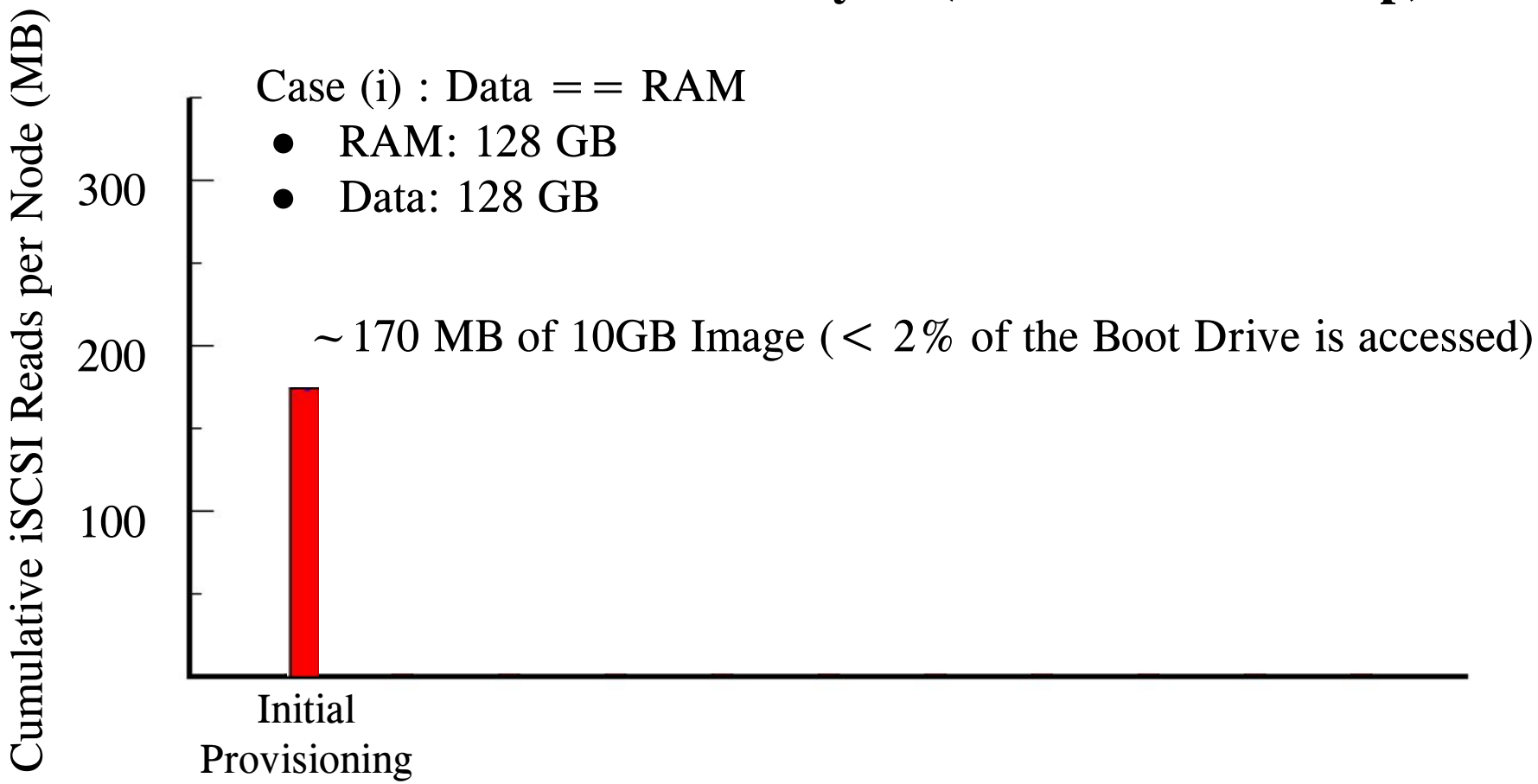
Provisioning/Re-Provisioning Times Comparison

(Single Hadoop Node)

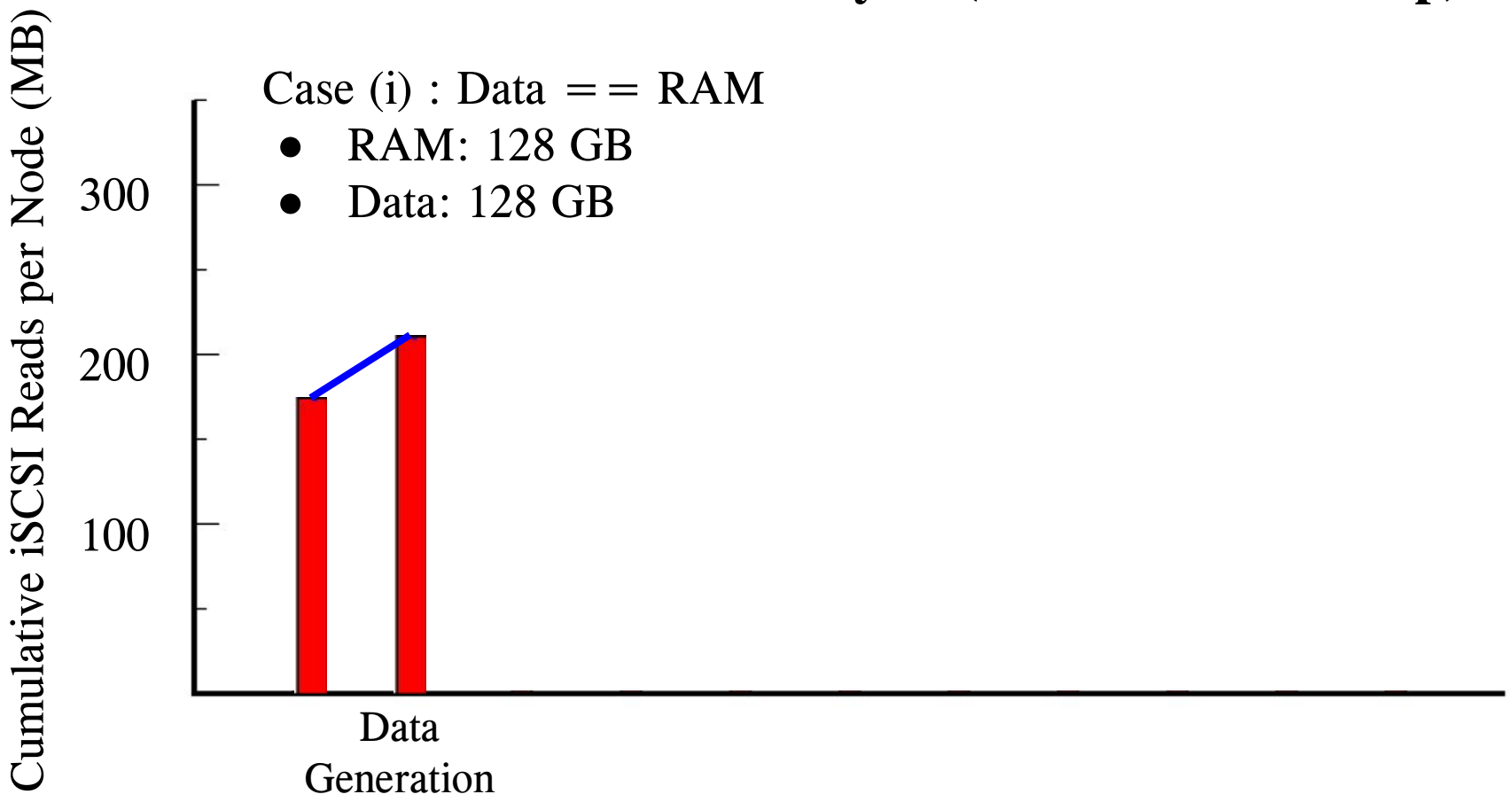


- Without POST ( ) Re-Provision time ~2 Minutes.
- ~2 Minutes to spin up Virtual machines in AWS.

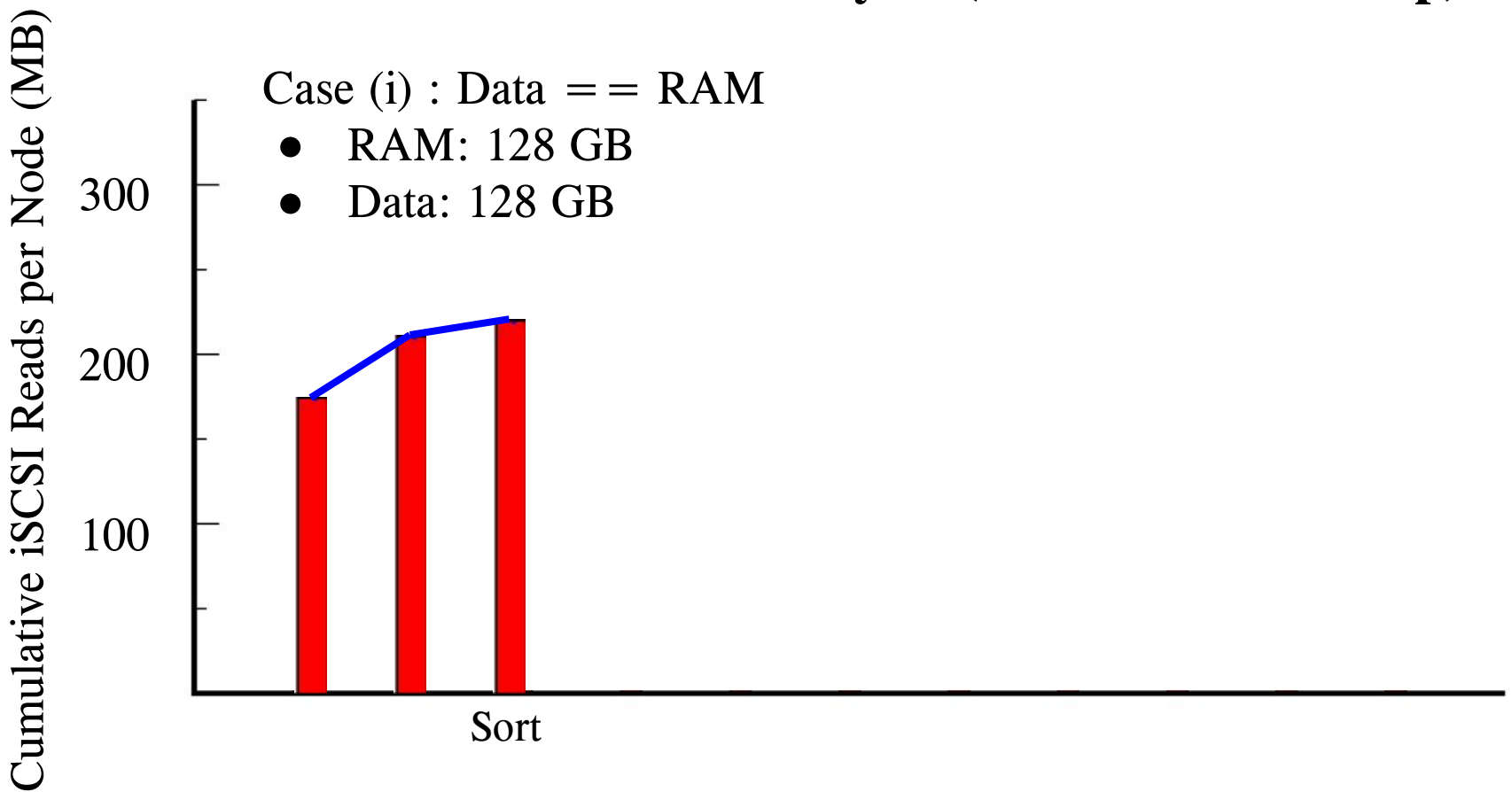
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



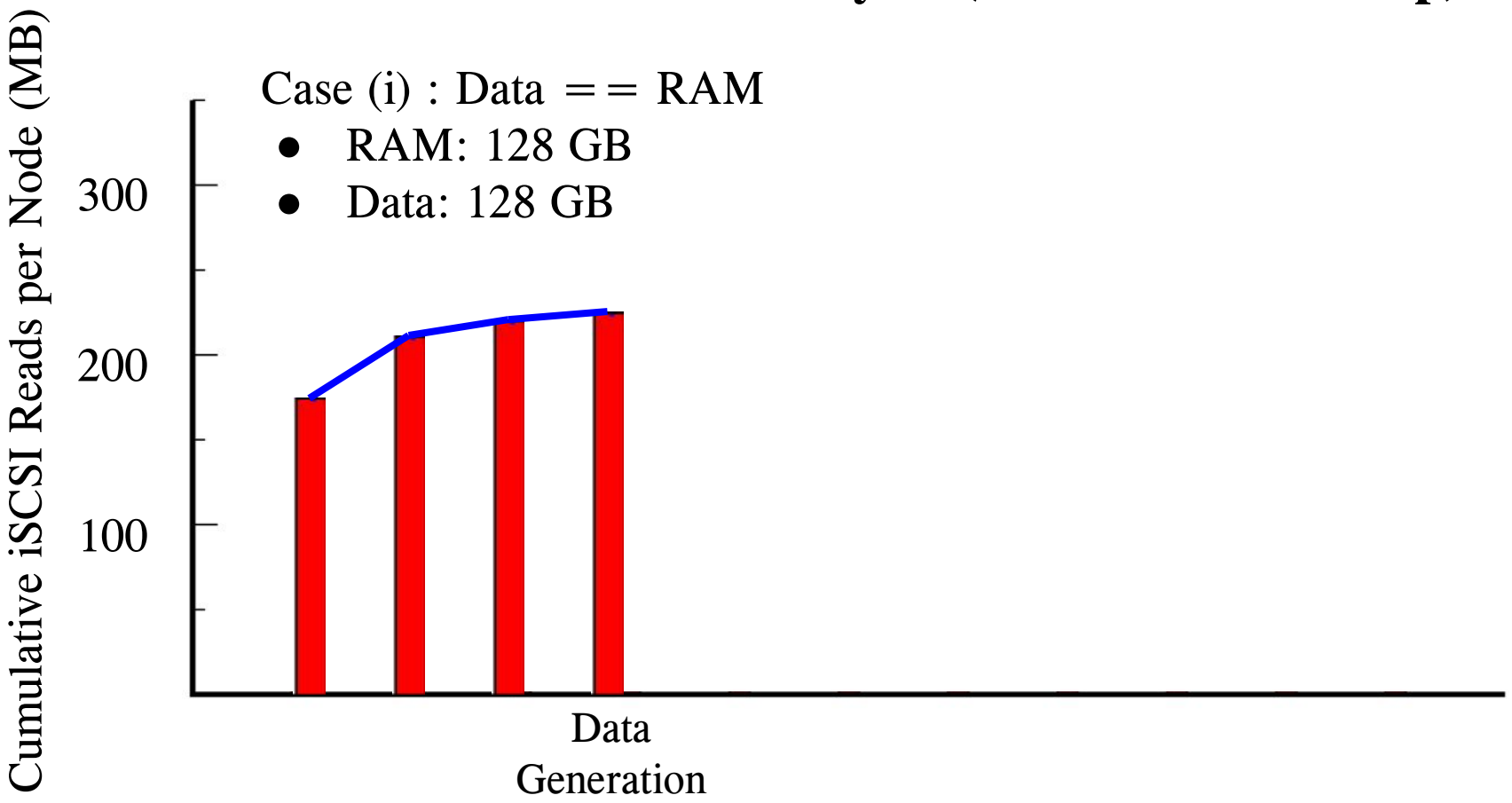
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



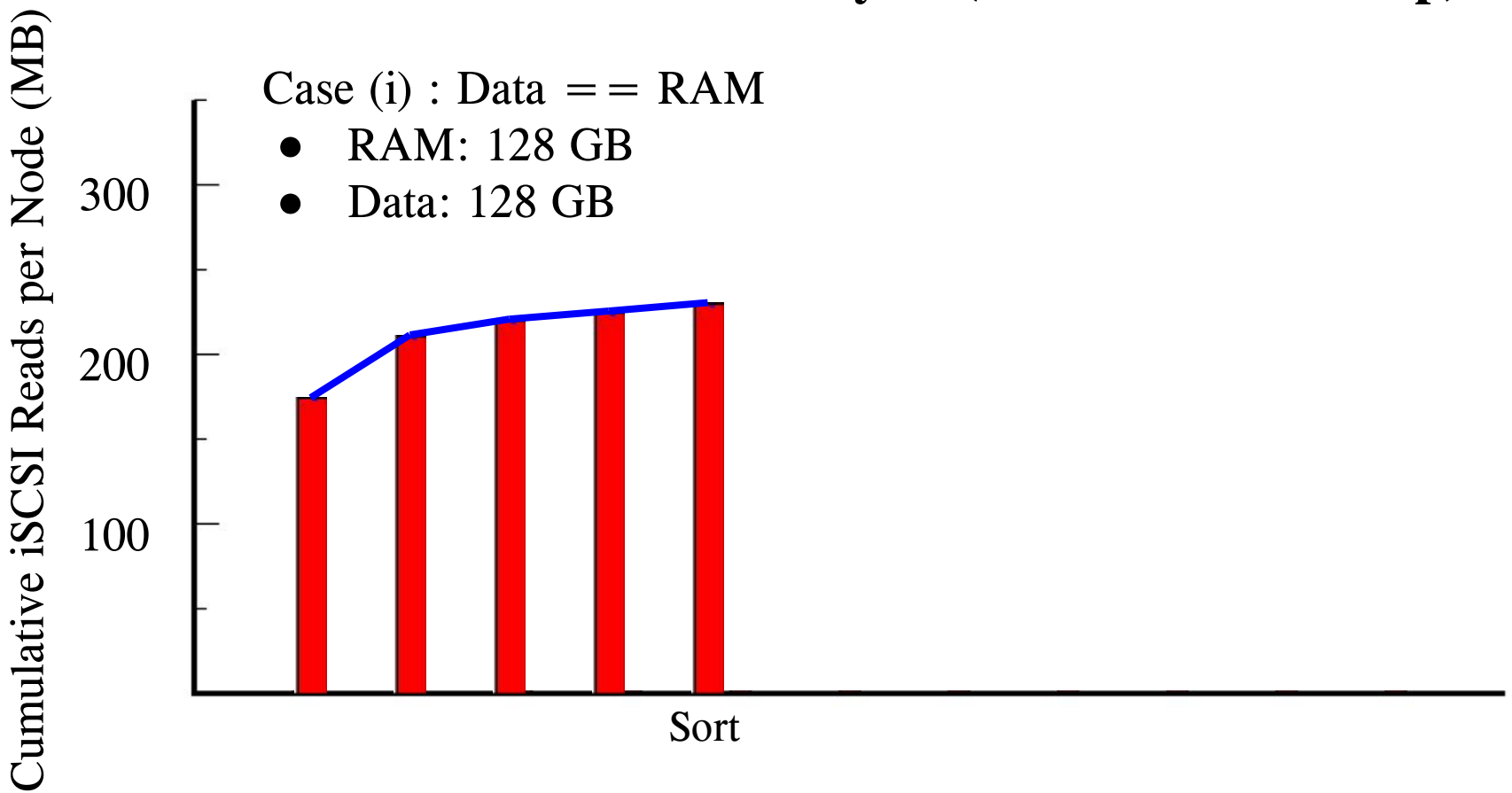
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



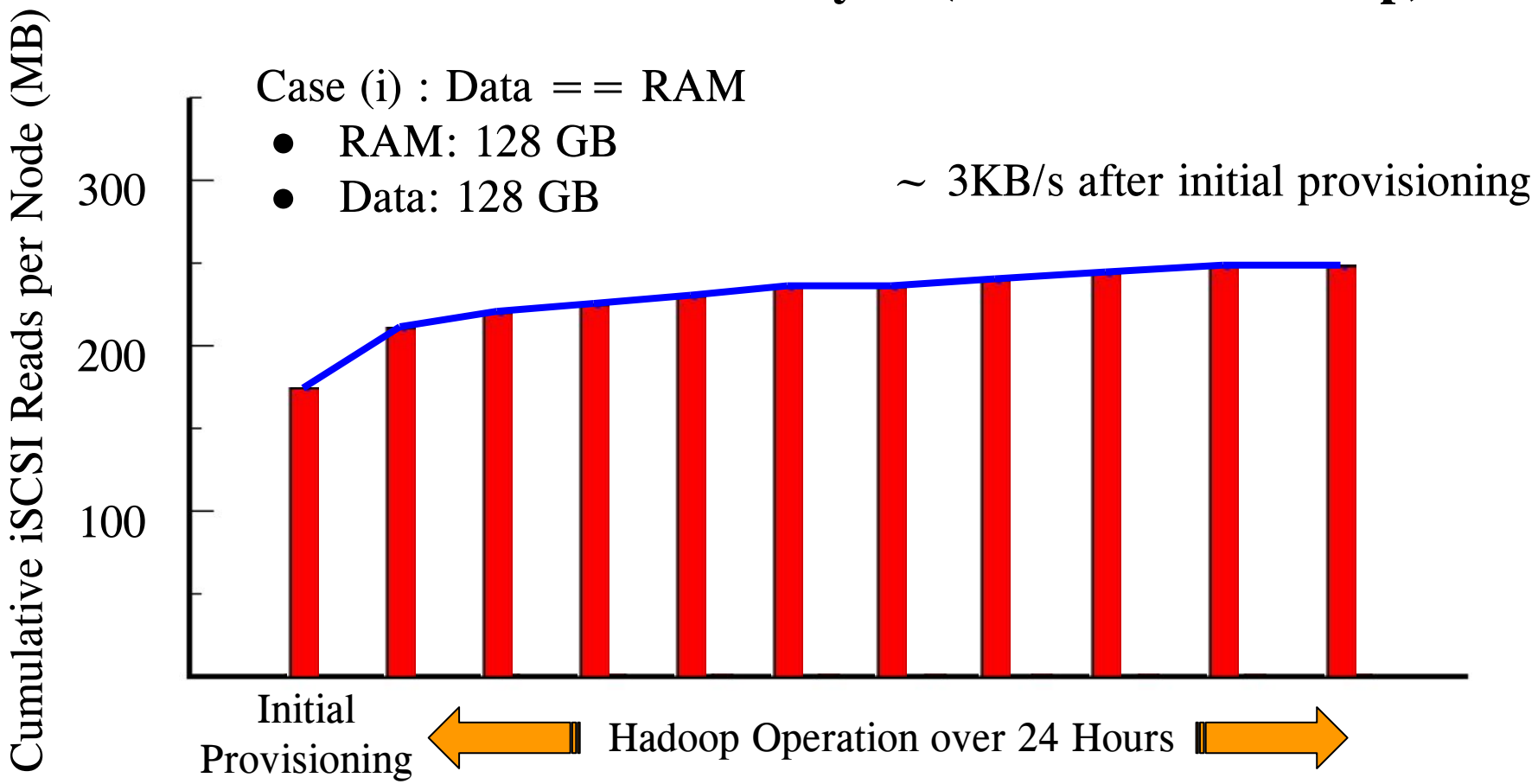
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



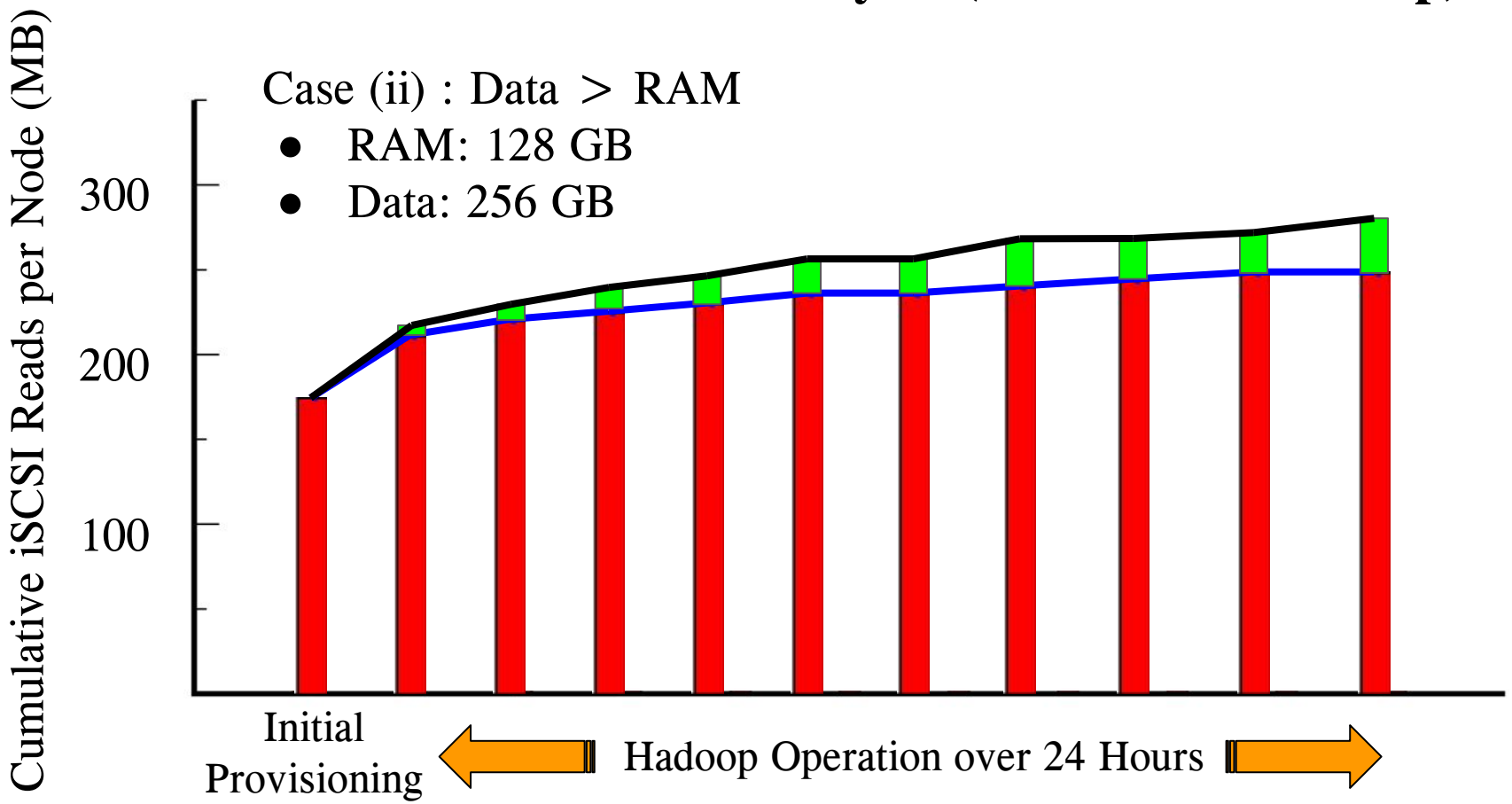
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



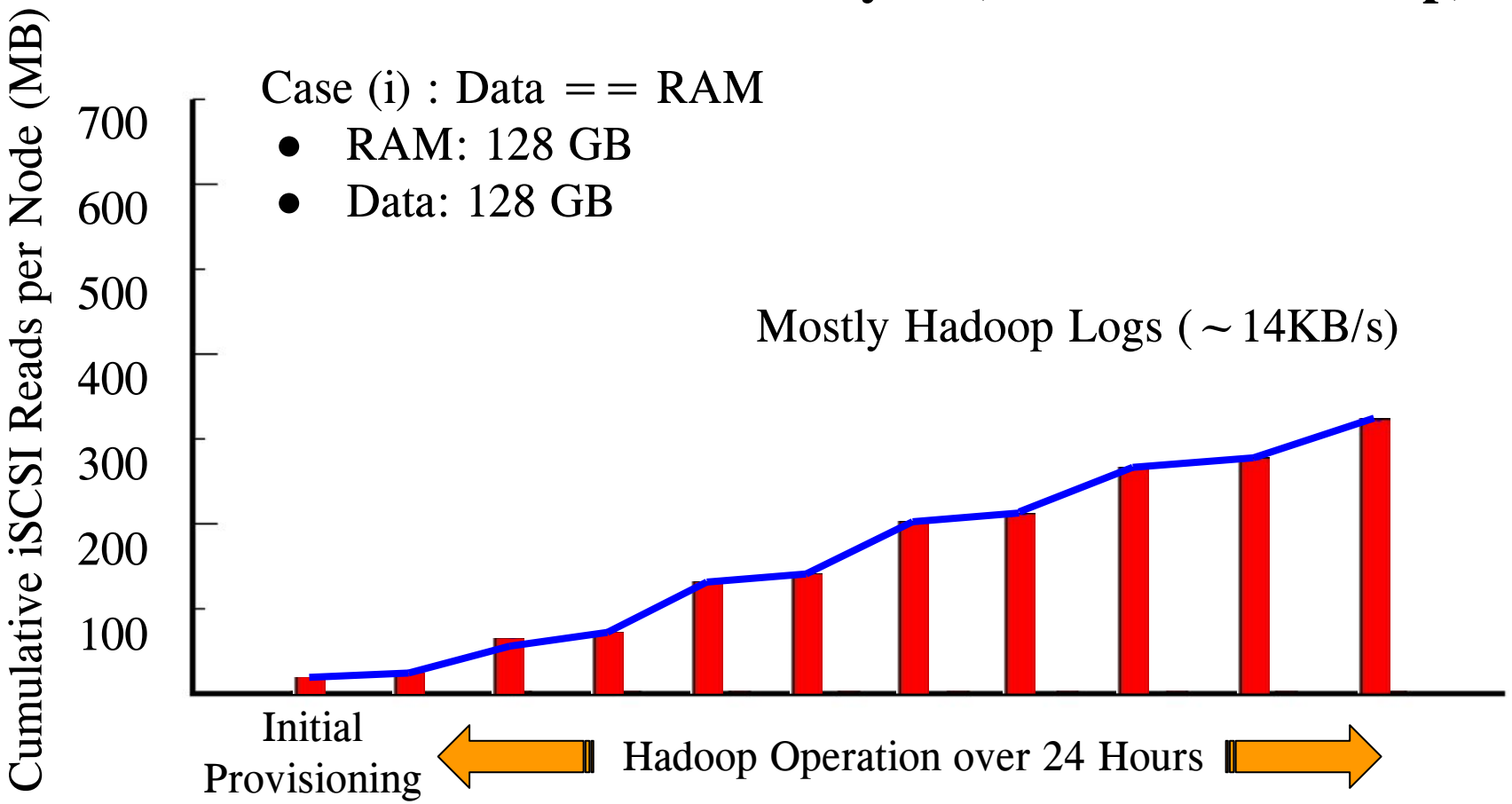
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



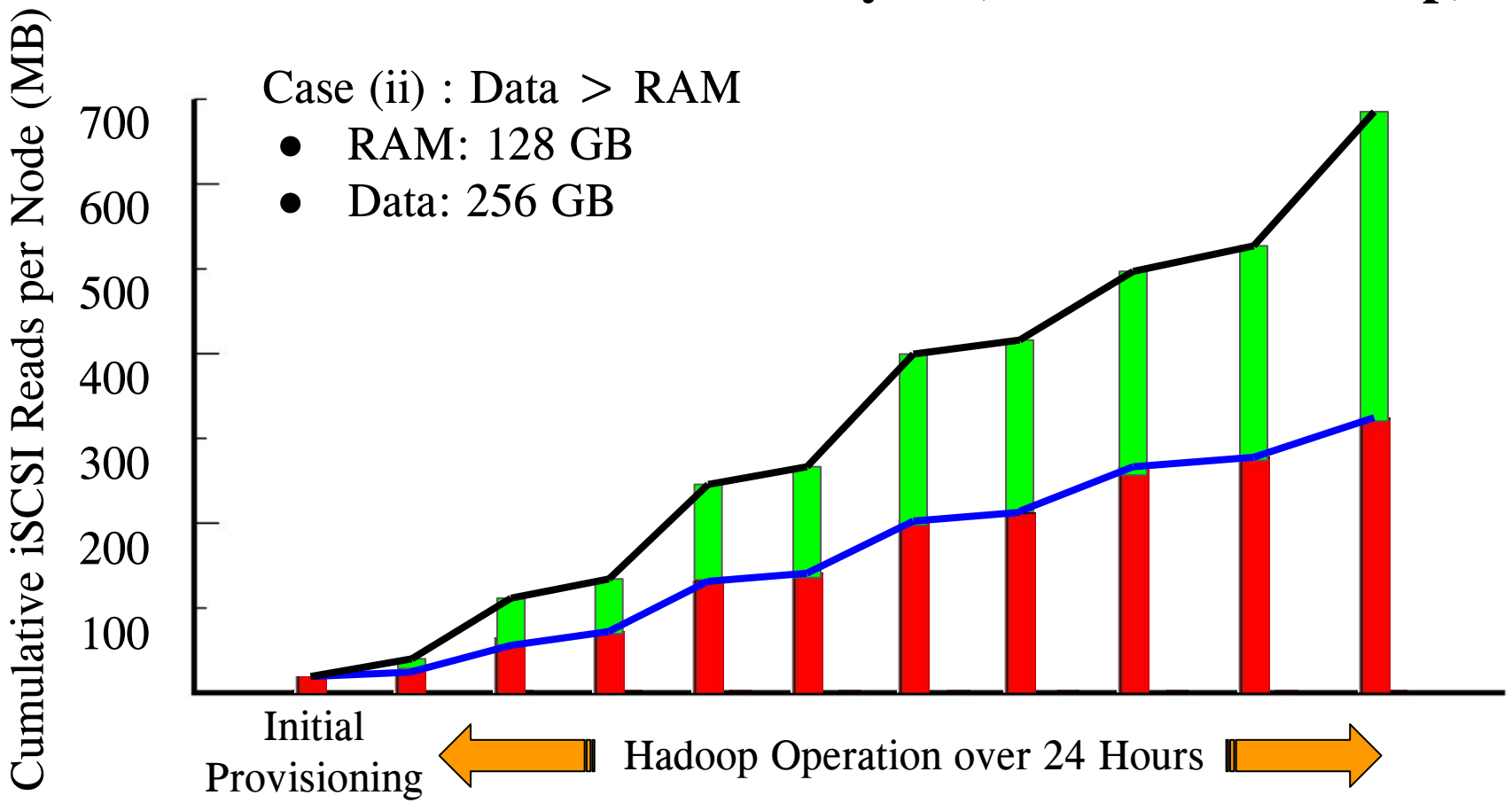
Cumulative Read Traffic Analysis (24 Hours - Hadoop)



Cumulative Write Traffic Analysis (24 Hours - Hadoop)



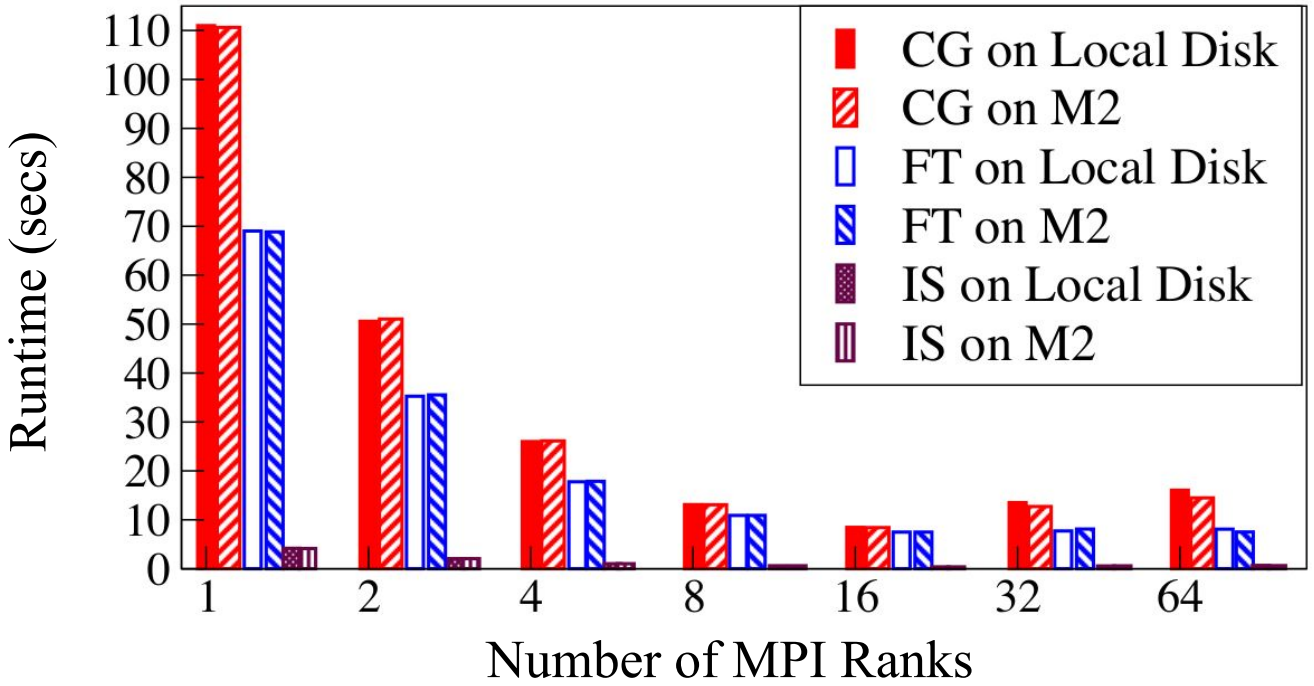
Cumulative Write Traffic Analysis (24 Hours - Hadoop)



- ❏ Network traffic to boot drive is irrelevant!
  - Resilient to boot storm

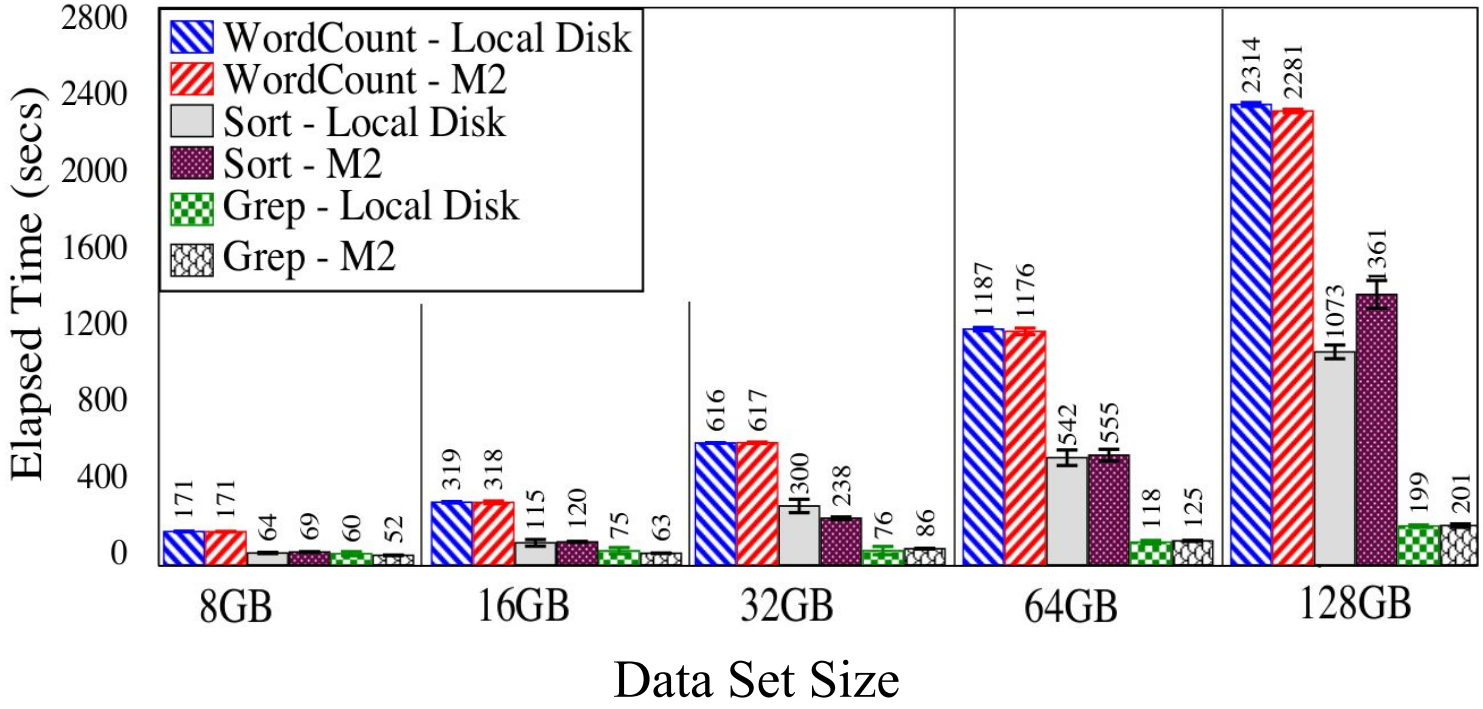
NAS Parallel Benchmark Performance (HPC)

Negligible Performance Impact



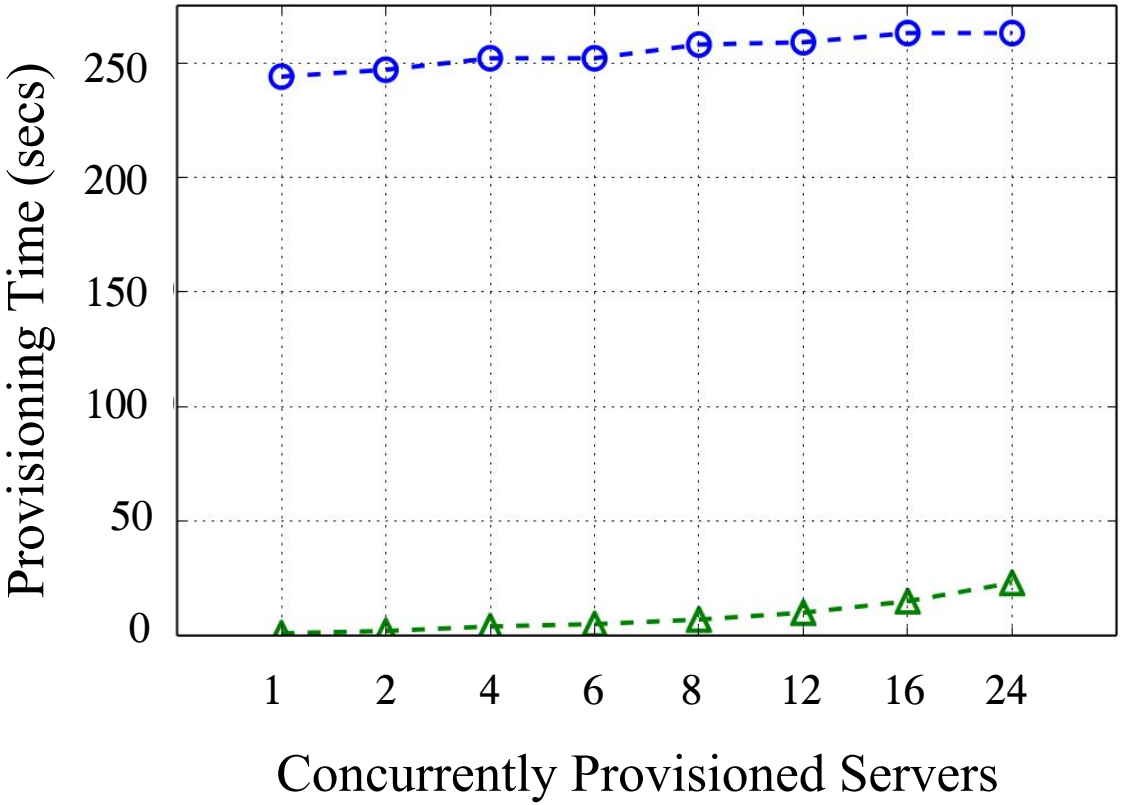
Hadoop Performance (Big Data)

Negligible Performance Impact



- Performance Impact Due to Remote Boot Drive is Negligible!

M2 Stress Test - Concurrent Provisioning



Total Provisioning Time

M2 API Overhead

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

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 provisio 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.

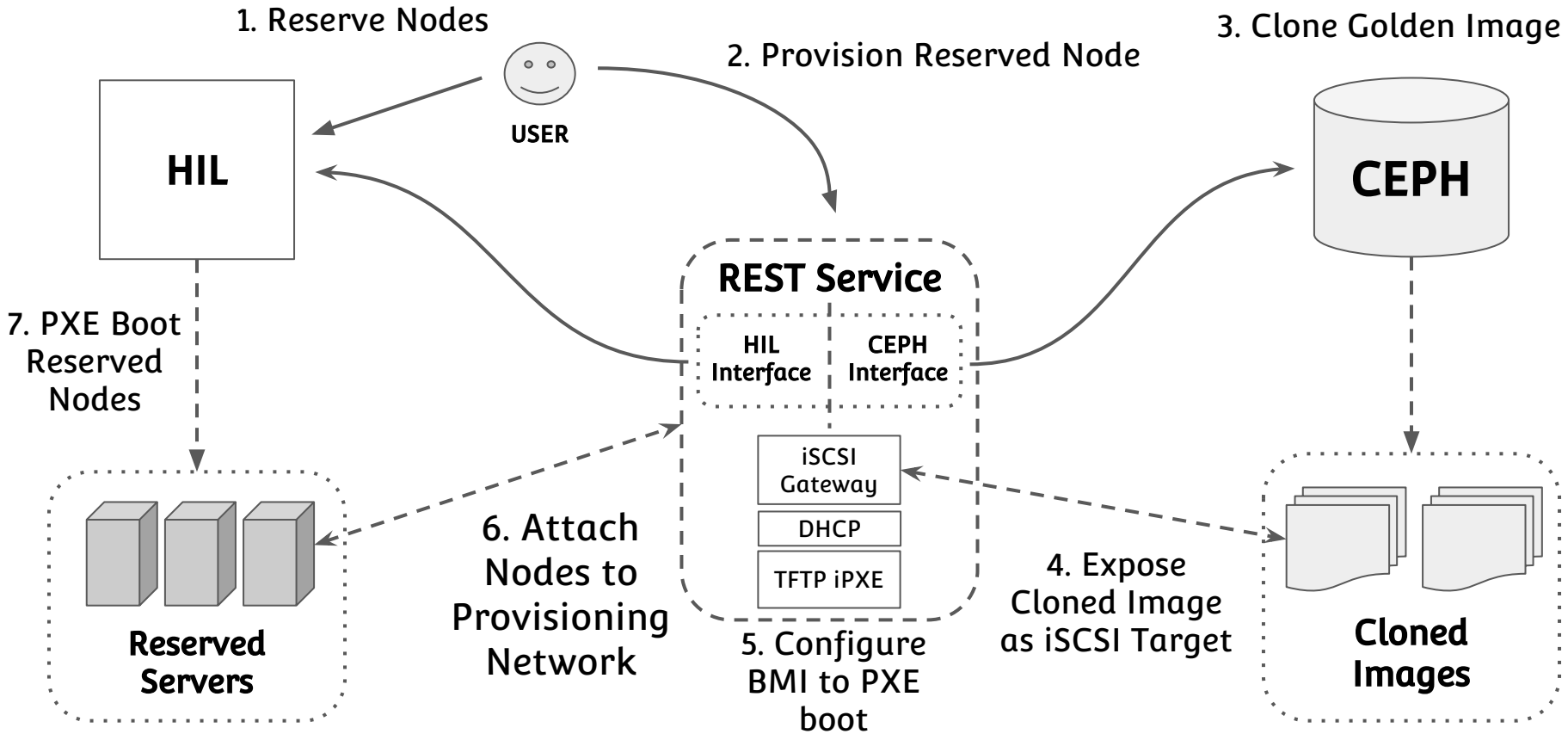
Thank You

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
USER
2. Provision Reserved Node
3. Clone Golden Image
HIL
CEPH
REST Service
HIL Interface
CEPH Interface
7. PXE Boot Reserved Nodes
iSCSI Gateway
DHCP
TFTP iPXE
6. Attach Nodes to Provisioning Network
5. Configure BMI to PXE boot
4. Expose Cloned Image as iSCSI Target
Reserved Servers
Cloned Images