

M2: Malleable Metal as a Service

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

*Northeastern University, †Boston University, ‡Red Hat Inc.,

Abstract—Existing bare-metal cloud services that provide users with physical nodes have a number of serious disadvantages over their virtual alternatives, including slow provisioning times, difficulty for users to release nodes and then reuse them to handle changes in demand, and poor tolerance to failures. We introduce M2, a bare-metal cloud service that uses network-mounted boot drives to overcome these disadvantages. We describe the architecture and implementation of M2 and compare its agility, scalability and performance to existing systems. We show that M2 can reduce provisioning time by over 50% while offering richer functionality, and comparable run time performance with respect to tools that provision images into local disks. M2 is open source and available at <https://github.com/CCI-MOC/ims>.

I. INTRODUCTION

Although virtualized cloud services can satisfy the requirements of many applications, some applications still require physical (i.e., bare-metal) nodes. Examples include performance or security sensitive applications that cannot tolerate the overhead, unpredictability, and large trusted computing base of complex virtualized cloud services [1], [2], [3], or applications that need direct and exclusive access to hardware components that are difficult to virtualize (e.g., InfiniBand [4], RAID [5], FPGAs [6], GPUs [7], etc.).

Cloud vendors have developed application-specific solutions dedicated to some of these use cases (e.g., Amazon HPC cloud [8], Amazon GPU nodes [9], Cirrascale deep learning cloud [10], etc.). However, these compartmentalized solutions lead to cloud silos, reducing the flexibility of the cloud to move resources between different users as demand warrants. Moreover, it is impossible to cover all bare-metal use cases with dedicated solutions; consider, for example, researchers that want to develop their own bare-metal operating system [11], or cloud developers that need to test software on environments identical to the eventual production environments.

The demand for bare-metal clouds has resulted in an increasing number of offerings such as IBM [12], Rackspace [13], and Internap [14]. These bare-metal cloud solutions install the tenant’s operating system and application into the server’s local disks. This installation process incurs long startup delays (tens of minutes to hours) and high networking costs to copy large disk images. Moreover, because user state is local to the server, these solutions lack rich functionality of virtual solutions including checkpointing/cloning of images, releasing and re-acquiring nodes to match demand, and fast recovery from node failures.

A number of industry and research projects have attacked the performance and functionality challenges of provisioning

bare-metal nodes [15], [16], [17], [18], [19], [20]; automating the bare-metal provisioning process, reducing the management overhead of the cloud provider, and improving the performance of copying the image to the server’s disk. For example, Omote et al. [20] proposed a lazy copy approach that copies the OS image in the background after the operating system is booted using a remote disk. While sophisticated techniques like this can reduce some of the user visible provisioning time, all these approaches end up eventually transferring the boot image to the local disk, and hence still incur overhead to copy the image and have the functionality problems discussed above.

We present M2, a provisioning tool for bare-metal clouds that addresses the challenges described above. Similar to virtualized cloud services, M2 serves user images that contain the operating system (OS) and applications from remote-mounted boot drives. M2 relies on a fast and reliable distributed storage system (CEPH [21], [22] in our implementation) for hosting images of provisioned bare-metal instances and a network isolation service (HIL [23] in our implementation) for isolating tenants in the cloud.

Key contributions of this work include:

- 1) The definition of M2 a general purpose architecture of a bare-metal cloud provisioning system that exploits remote storage¹ and allows users to:
 - rapidly release and then acquire nodes to handle fluctuation in demand,
 - rapidly recover from failed nodes by booting another node with the disk,
 - snapshot and clone disk images,
- 2) An implementation and analysis that demonstrates that:
 - it is possible to provision and deploy bare-metal systems with overheads similar to deploying virtual machines,
 - performance of the M2-provisioned servers is similar to those provisioned to local disks.²

The remainder of this paper is organized as follows. We provide related work in Section II. The design and architecture of M2 are presented in Section III and Section IV, respectively. We evaluate performance, scalability, and usability of M2 in

¹Previous provisioning systems exploited remote storage in special purpose environments, like HPC clusters, where all nodes boot the same kernel.

²As the focus of this work is to improve the provisioning time M2 only network-mounts boot drives that hosts the OS and applications. Data drives are still hosted on the local disks.

Section V. We discuss future directions for M2 in Section VI and conclude in Section VII.

II. RELATED WORK

In this section we review existing bare-metal provisioning approaches considering their fitness to support on-demand bare-metal IaaS offerings. We can broadly classify provisioning approaches into two as *diskful* and *diskless* provisioning systems, based on where the image is hosted once a bare-metal instance is provisioned.

Diskful Provisioning Systems: These systems persist the provisioning image to the local disks of the bare-metal systems. The standard provisioning tools used in many bare-metal deployments are diskful. A rich set of open source and commercial provisioning products such as Emulab [17], OpenStack Ironic [15], Crowbar [18], Canonical Metal-as-a-Service (MaaS) [16], Razor [19], and Cobbler [24] are available for automated diskful provisioning of bare-metal systems. Chandrasekar and Gibson [25] provide a comparative analysis of commonly used diskful provisioning systems.

Diskful provisioning systems can be further divided into two types. The first type of solutions automate the manual installation process of the OS and desired applications to the local disks (e.g., Foreman [26]). As they follow a step by step installation process, these solutions generally take the longest to provision. The second type of solutions copy a pre-installed image, containing the operating system and applications, onto the local disk over the network (e.g., OpenStack Ironic [27]). The size of such pre-installed images can be tens of GBs. Transferring them can overwhelm the network and persisting them to local disks still requires hundreds of seconds assuming standard HDDs are used. Both solutions have a lower bound on the time before a node is ready to use due to the need to reboot twice³, once via PXE to enter the installer, and another to boot into the freshly installed system.

When using diskful systems, repurposing a bare metal system requires formatting the local disks and then installing/copying the new system. If saving the existing disk state is desirable, the contents of the disk has to be copied away, which again requires hundreds of seconds and further increases the re-provisioning cost.

In general, diskfull systems and the automation tools that employ these provisioning systems (e.g., Ironic, MaaS, Foreman) are designed for setting up long running bare-metal systems. They consider the high startup delays tolerable assuming that the servers they provision will have long operational lifetimes. To support fast provisioning and reducing boot time of diskfull systems Omote et al. [20] propose BMCast, an OS deployment system with a special purpose de-virtualizable Virtual Machine Manager that supports OS-transparent quick startup of bare-metal instances.

Beyond the cost of transferring images to the local disk, a fundamental problem with all diskful provisioning systems is that any modifications to the image are stored on the disks

attached to the physical node allocated to a user. This means, for example, that a user cannot easily release and re-acquire nodes to match their needs, since any state on the local disk is lost when the user releases the nodes. Some of the rich functionality users take for granted in virtualized clouds is also not available in these environments. For example, a user cannot snapshot the disk of a physical node and then clone it to boot additional nodes. Perhaps most importantly, if a physical node fails, the user cannot easily start up another node from the same disk image or use the disk to diagnose the failure; any state local to that disks node is inaccessible as long as the node is down. Moreover, in diskfull systems the local disk hosted boot drives become a single point of failure. If the disk containing the boot drives fails, the bare-metal node becomes unusable until the disk is fixed or replaced and data recovery can be daunting in such cases.

Diskless Provisioning Systems: These systems keep the provisioning image resident on a network accessible remote logical disk that appears as a local disk to bare-metal systems. This method of provisioning historically has been used with diskless workstations [29], [30], [31] and HPC systems [32], [33], [34] to boot multiple nodes from a single image. Furthermore, diskless provisioning is heavily used in virtualized systems [35], [36], [37]. Interestingly, diskless provisioning is not being used in cloud deployments for bare metal provisioning and there are no tools or studies that combine diskless provisioning with image management capabilities to support bare-metal provisioning and servicing of multiple images owned by multiple users. To our knowledge, M2 is the first effort in this direction.

III. M2 DESIGN

When designing M2 we first listed the set of features we wanted to have in order to support an on-demand bare-metal cloud service. These features are:

- *Rapid provisioning:* A bare-metal cloud service has to offer on-demand bare-metal servers with minimum startup overhead so that even short lived deployments with life-span of a few hours can use bare-metal servers efficiently. If servers take tens of minutes to deploy, the “effective” utilization of the cloud will decrease.
- *Rapid snapshotting, re-purposing, reprovisioning:* Abilities for quickly snapshotting the OS and applications, releasing a server when unused, and being able to quickly provision a server using a previous snapshot are critical for time-multiplexing bare-metal servers across many users. These features also enable the service to offer “elasticity” to the applications it hosts.
- *Rapid cloning:* The ability to rapidly stand up a large number of servers concurrently using the same saved image is a common request in infrastructure as a service clouds. This feature enables easy deployment of parallel/distributed applications and scalability.
- *Support for multi-tenancy:* Existing provisioning tools assume that they are available to just the administrator of the hardware and all of the hardware available in the

³Rebooting modern datacenter servers can take as long as 5 minutes [28]

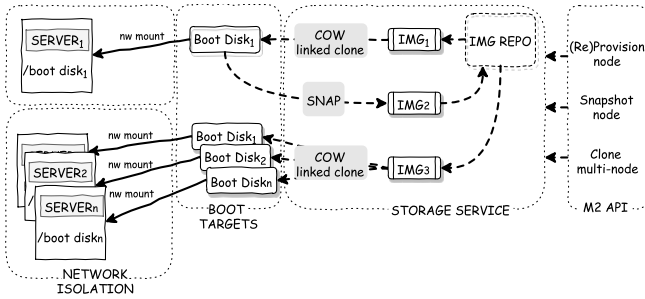


Fig. 1: M2 mockup design

system is managed by the same entity. However, in a bare-metal cloud service, the provisioning system has to ensure performance isolation and security across its users even during provisioning.

Given the above list of desirable features, we made a set of design decisions for M2. In order to offer rapid provisioning, we opted to use diskless provisioning mechanisms. Using these mechanisms M2 does not need to copy the entire image to the bare-metal server and it can save the overhead to install images and applications to local disks once a cloud image is prepared. M2 can rapidly start running applications by only fetching the necessary OS and application libraries before start-up and further required packages will be fetched on-demand as they are used. Note that the standardization of technologies such as iSCSI [38] has allowed diskless provisioning to be used with commodity servers and clients over any layer-3 network.

We also decided to network boot the nodes from images residing on a distributed storage. M2 can service the images from centralized high performance storage systems using multiple disks to improve boot time. Furthermore, many modern distributed storage systems support capabilities such as *copy-on-write (COW)*, *de-duplication* and *linked-cloning* [39], which are beneficial for capabilities such as rapid cloning and snapshotting.

We note that in diskless provisioning clients are always dependent on uninterrupted access to the centralized storage, and as the number of clients increase, the storage infrastructure has to be adequately scaled to support the increasing load. Network connectivity and availability also plays a critical role in the performance of diskless provisioning systems. However, with the advancements in faster, cheaper and redundant networking (e.g., Clos networks [40]) and storage solutions (e.g., Solid State Drives), datacenter applications increasingly lean towards disaggregating storage services to make full use of the capacity of their infrastructures [41]. Diskless provisioning approaches are very much aligned with this trend.

Figure 1 presents the conceptual design for M2 and some of the functionalities it offers. As seen in the figure, M2 stores images (user or M2 provided) in an image repository. When a provisioning API call is made for a bare-metal node with a given image, a linked-clone of that image is created, followed by network isolation of the requested bare-metal node and mounting of the clone on the bare-metal node. When the node

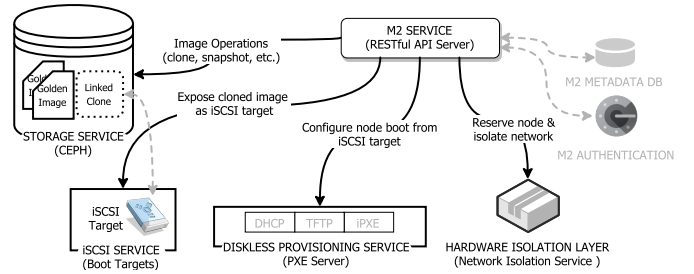


Fig. 2: M2 components and architecture

network-boots, it only fetches the parts of the image it uses, which significantly reduces the provisioning time. M2 also supports provisioning multiple nodes in parallel from a single image by simply performing parallel provisioning calls.

As seen in Figure 1, disk snapshotting API call of M2 enables users to create checkpoints/restore-points by saving the current state of the image to the repository and tagging the saved image with a unique identifier. Using linked-cloning and COW, M2 can offer rapid snapshotting.

IV. M2 ARCHITECTURE

In this section we discuss our implementation for M2. There are five major components in M2: (i) API Server, (ii) Storage Service (Ceph), (iii) iSCSI Service (TGT Server), (iv) Diskless Provisioning Service (PXE Server), and (v) Network Isolation Service. Figure 2 displays these five components. M2 follows a driver based approach and provides an abstraction for each component. This allows system administrators to replace the solution used for any of these components. For example, in the current implementation Ceph [42] is used as the storage service, but it is possible to replace it with any other storage service that supports COW like the network-based Lustre [43] or even local systems like ZFS [44] or Linux’s LVM [45].

Storage Service: Storage Service provides a data store for the cloud images and exposes API’s to rapidly clone and snapshot existing images. In our implementation we use Ceph as our storage solution. Ceph is an open source storage platform that implements a highly reliable and scalable object storage on a distributed cluster [42]. It exposes various interfaces for object, block and file level storage [22]. We used the block storage interface provided by Ceph aka the Reliable Autonomic Distributed Object Store Block Device (RADOS Block Device or RBD) to store the cloud images using the *librados* API. *librados* also exposes functionalities such as cloning and snapshotting to manage the RBD based cloud images. Ceph provides data store and image management capabilities like snapshotting and cloning and offers good read performance [46], which helps in achieving lower latency when M2 tries to fetch the disk blocks on-demand [47].

iSCSI and Diskless Provisioning Services: Our implementation of diskless provisioning is based on network booting bare-metal nodes from RBD based cloud images (stored in CEPH) that are exposed as iSCSI targets. We used the Linux SCSI Target Framework (TGT) [48] to expose the RBD

based cloud images as iSCSI targets. As TGT is a user-space implementation, no extra kernel code is required which improves its compatibility with modified Linux kernels. Being a user-space server also supports multi-tenancy within M2 and defense in depth by enabling the iSCSI server to be run in a Linux container [49], which can permit a single physical node to serve different tenants on different networks without exposing all the iSCSI endpoints to all tenants. TGT also provides native support for RBD based images, freeing M2 from managing additional mapping state. The current implementation of M2 does not have support for iSCSI multipathing or special iSCSI hardware, however, in the upcoming release of M2 we are working to provide support for both scenarios (see subsection VI-A).

Preboot eXecution Environment (PXE) specification provides a standardization for a client-server model for booting nodes over the network using Dynamic Host Configuration Protocol (DHCP) and Trivial File Transfer Protocol (TFTP). Although, PXE provides specifications to network boot a node from various targets (HTTP, iSCSI, AOE etc.), it is up to the NIC manufacturer to implement the support to network boot from a particular target into the NIC firmware. As mentioned earlier, M2 uses an iSCSI based approach to network boot the bare-metal nodes. To ensure that M2 can provision any bare-metal node irrespective of the NIC firmware capabilities of that node, M2 first chainloads [50] into iPXE, which eventually network boots the bare-metal nodes from the exposed iSCSI target. iPXE is an open-source implementation of network booting firmware that provides all the features mentioned in the PXE specifications [50].

The iSCSI Boot Firmware Table (iBFT) [51] gives PXE servers the ability to specify an iSCSI target to which the tenant OS should connect. iBFT makes M2 OS-agnostic, since it eliminates the need for OS-specific parsing and modification of images to configure the identity of the iSCSI server for a given node.

M2 API Server: API Server is a python based RESTful web service that controls the flow between different components of M2. Exposed APIs enable users to (de)provision nodes, clone provisioned nodes, create snapshots of provisioned nodes, (de)register users, perform various operations pertaining to images (upload, download, rename, share, list, etc.), list various resources, etc. The API server also maintains a database for various book keeping purposes such as maintaining the mapping between bare-metal nodes and cloud images, user-cloud image mappings, etc.

While most of the API calls are trivial and trigger various M2 database operations, some of them accounts to interacting with different M2 components — in particular the APIs pertaining to (re/de)-provisioning, snapshotting and cloning nodes, and image manipulation. APIs that require interacting with the storage service rely heavily on the performance of the exposed block storage management capabilities.

The provision API enables users to spawn bare-metal instances from existing cloud images hosted in the storage

service⁴. It accepts the ID defining the node to be provisioned (e.g. MAC address, NIC Number) and the ID of the cloud image to be used for provisioning as arguments. Upon receiving a provision request, the API server interacts with the Storage Service and creates a linked-clone of the cloud image (passed as the argument to the provision call) and exposes it as an iSCSI target. This is followed by the preparation of the PXE and iPXE configuration files by the Diskless Provisioning Service that will be served to the bare-metal node upon its network boot request. This is similar to how virtual machines are provisioned in different IaaS cloud offerings.

M2 exposes a snapshot API that allows users to create checkpoints/tags by saving a deep copy of the existing node state. Users can use these snapshots to revert back to any previous state in case of a failure⁵. This feature also enables users to manage different configurations of their nodes/clusters. M2 does not expose an explicit clone API. But users can clone an existing node by creating a snapshot of the current node state and provisioning a/multiple new node(s) from that snapshot.

Multi-tenancy and Allocations: For multi-tenancy it is important to segregate each M2 node based on ownership and physically using some network isolation mechanism. M2 uses Hardware Isolation Layer (HIL) [23] for allocations and to achieve multi-tenancy. HIL is a lightweight python-based layer-2 bare-metal isolation framework that orchestrate data center compute resources by controlling the networking infrastructure. It exposes an API that enables users to create isolated groups of compute resources from a hardware resource-pool. HIL is a network-switch agnostic framework that follows a driver-based model⁶. HIL is agnostic to the provisioning system running on top of it and is thus our choice for achieving network isolation (multi-tenancy) and allocations.

V. EXPERIMENTAL EVALUATION

In this section we evaluate M2's speed, scalability and performance. We start by presenting our experimental setup. Then we compare the provisioning time of M2 with that of existing provisioning solutions, present the time taken by different M2 API calls, and analyze the scalability of M2. Network overheads associated with using a diskless solutions are also provided and M2's impact on the performance of frameworks and applications is analyzed. We note that in the following experiments, only the boot drives are mounted remotely by M2 as currently M2 focuses on improving boot performance. Whenever data drives are used by applications, those drives are hosted on local disks.

A. Experimental Setup

In our experiments we used two different environments. In the first environment, each bare-metal server has two 6-core Intel Xeon E5-2630L CPUs (24 cores with hyperthreading

⁴Cloud images need to be registered and uploaded to M2 before the provision API is invoked

⁵The current implementation of M2 is limited to disk snapshots

⁶Currently HIL can manage network isolation for Cisco, Juniper, Dell and Brocade switches.

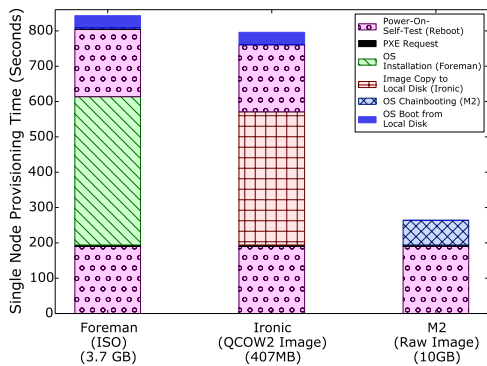


Fig. 3: Single node provisioning time comparison between M2, Foreman, and OpenStack/Ironic.⁷

enabled), 300 GB 10K SAS HDDs (two nodes had 1 TB 7.2K SATA HDDs), 128GB RAM and two Intel 82599ES 10 Gbit NICs. A four-node Fujitsu CD10000 Ceph storage cluster with four 10 Gbit external NICs and internal 40 Gbit InfiniBand interconnect is used as the storage server of M2 in this environment.

In the second environment each bare-metal server has a single Intel 8-core Xeon E5-2650 CPU (2.30GHz, 16 cores with hyperthreading enabled), 64 GB RAM, two 1.8 TB HDDs, and one 10Gbit Ethernet adapter. A ten-node Ceph cluster with a total of 90 spindles and 10GbE internal 40GbE external NICs are used as the storage server of M2 in this second environment.

For both environments, M2’s iSCSI and API servers were deployed on a virtual machine with 4 VCPUs and 4 GB RAM. The RHEL 7.1 (i.e., Centos 6.7) operating systems (OS) is installed in cases where an OS installation is performed.

B. Provisioning Time Comparison

Figure 3 presents the time comparison of M2 with Foreman, and OpenStack/Ironic, two widely used provisioning systems, when we provision a single bare-metal server in our first environment with a bare RHEL 7.1 operating system. As seen in the figure, the M2 provisioning time is around five minutes. Note that firmware initialization of these bare-metal servers requires more than three minutes; hence half of the M2 provisioning time is spent in firmware initialization. Both Foreman, and OpenStack/Ironic have to go through firmware initialization phase twice. Furthermore, they have to install or network-transfer the OS to local disk, whereas M2 simply provisions the node out of a remote disk containing the operating system. Due to these advantages, M2 provisions nodes around three times faster than both Foreman and OpenStack/Ironic⁸.

⁷Bare-metal nodes were provisioned with RHEL 7.1 for all the provisioning systems. The virtual disk size of the image used for both Ironic and M2 was 10 GB. The actual disk size in the case of Ironic was 407 MB whereas for M2 it was 10 GB.

⁸In Figure 3, we do not include the time taken to prepare the provisioning target for the provisioning systems since this is a manual process for Foreman.

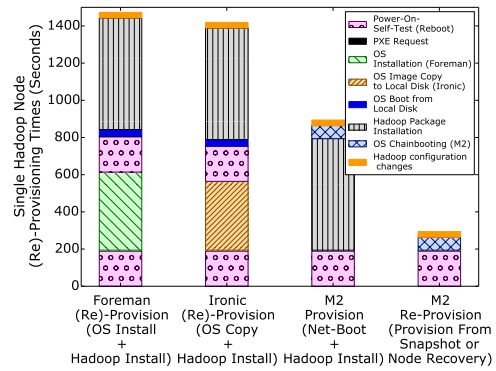


Fig. 4: Single Hadoop compute node (re-)provisioning time comparison between M2 Ironic and Foreman.

C. Provisioning Complex Frameworks

Provisioning a node for any framework such as Hadoop or SLURM is a complex and time consuming process handled in many-steps. First, the operating system is installed, and then the relevant packages for the framework is installed, which is followed by making configuration changes to the node. The first three bars of Figure 4 show the total provisioning time for a single Hadoop compute node when using Foreman, Ironic and M2. As seen in the figure, M2 can only offer ~40% improvement during this process as it is dominated by the application installation and configuration.

Even though installing and configuring frameworks such as Hadoop is a time consuming process, once a single example setup is made, M2 can leverage its snapshotting and cloning mechanism to safe-keep that example and use it for provisioning other framework nodes. As shown in the fourth bar in Figure 4, with M2, provisioning cloned images that contain desired applications and then doing a final reconfiguration is significantly faster than provisioning nodes from scratch.

D. Using M2 for Failure Recovery

In large data center deployments, node failure is a common phenomenon [52], [53], [54], [55]. Recovering from a node failure involves tedious manual operations. If the node is provisioned from the local disk (using Foreman or Ironic), the node becomes unavailable until it is fixed. In addition, if the cause of node failure was disk failure, there is a good chance that all of the user data is lost. In order to re-provision another bare-metal server using Foreman or Ironic as the same Hadoop node, it is required to re-install (or re-copy in the case of Ironic) the operating system and Hadoop packages on the server — leading to a re-provisioning time similar to the provisioning time. As shown in Figure 4, the total time to re-provision a single Hadoop compute node is the same as its provisioning time for Foreman and Ironic.

On the other hand, if this node was provisioned using M2, upon failure a new node can be re-provisioned (rebooted) using the image of the failed node that resides in Ceph. The time to re-provision the new node is significantly reduced as there is no requirement to re-install the operating system or any Hadoop packages. As shown in the last bar of Figure 4, M2

TABLE I: Time required by other M2 operations.

M2 API Call	Time (secs)
De-Provision	32.00
Snapshot	11.65
Clone Image	7.10

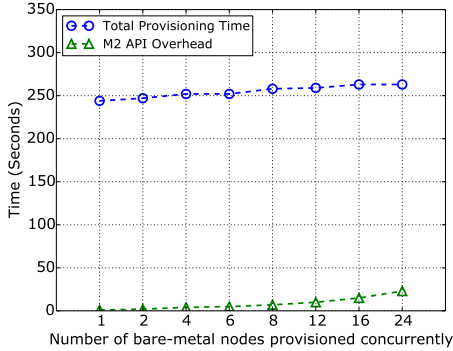


Fig. 5: M2 scalability analysis.

reduces the re-provisioning time of the nodes by up to 5 times as compared to Foreman or Ironic.

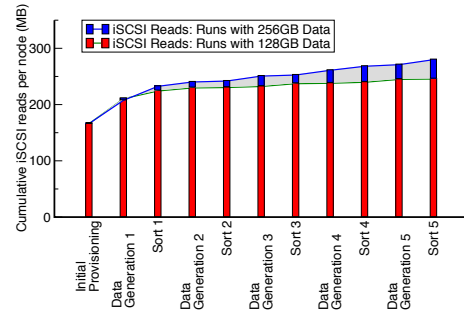
E. Operation Times of Other M2 Calls

Table I presents the time it takes to perform some of the other M2 API calls that require interaction with the storage service. The time taken by the De-Provision operation constitutes the time for those operations pertaining to HIL (detaching the provisioning network from the node), the iSCSI service (disabling the iSCSI target), the storage service (deleting the image associated with the node to be de-provisioned) and the API server orchestration. The time taken by the Snapshot and Clone Image operations are dominated by the storage operations, which include the time taken to flatten a linked clone in the case of the Snapshot operation and the time taken to create a deep copy of a cloud image in the case of the Clone Image operation.

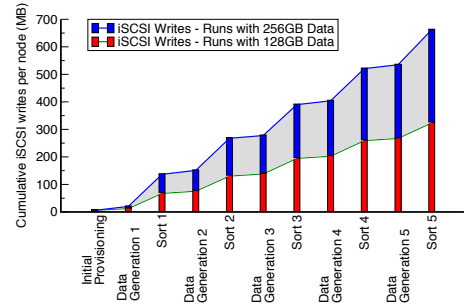
F. Scalability

In Figure 5, we show the time M2 requires for provisioning multiple nodes in parallel from our second environment. We increase the number of concurrently provisioned nodes from one to 24 and report the time it takes to provision that many nodes (upper blue circle line). As seen in the figure, provisioning 24 nodes takes only around 20 seconds longer than provisioning a single node, indicating that even with modest resource usage M2 is scalable. We observe a slight increase in the M2 overhead (lower green triangle line) since the iSCSI server VM, which has four vCPUs, has to context switch when the number of nodes increase above four.

We note here that the current M2 implementation is totally unoptimized and runs multiple M2 services on a single wimpy VM. We expected to see a significant performance degradation in our scalability analysis as requests on M2 increased but observed that not to be the case. As will be shown in the coming sections, this is due to the fact that only a tiny fraction of the provisioning image is accessed during booting



(a) Read Traffic



(b) Write Traffic

Fig. 6: Amount of read and write traffic passing through the M2 iSCSI Service hosting the boot drive during provisioning of a Hadoop node and consecutive Hadoop application runs.

and application runs and the load on the M2 services is comparatively low.

G. M2 Network Traffic Analysis

Figure 6 shows the per-node cumulative read and write traffic passing through the M2 iSCSI Service during initial provisioning of a bare-metal Hadoop node and then over five consecutive “data generation and sort” jobs performed over the same node. “Data generation and sort” jobs of 128 GB and 256 GB are performed. Bare-metal nodes from the first environment are used during these experiments. The size of the image containing the operating system and the Hadoop packages was 8GB. Only the boot drive of the server is mounted remotely and the data drives are hosted on the local disk of the node in these experiments.

Figure 6a shows that only ~ 170 MB of the ~ 8 GB image is read over the network during initial provisioning. Furthermore, both read and write curves flatten after repeated runs, demonstrating that (even with the 256GB case where the total data handled is substantially larger than the system memory) the file cache is effective at caching the boot drive. After initial boot and application start-up, the sustained read bandwidth incurred is around 3KB/s; effectively negligible.

Figure 6b shows the writes to the network-mounted storage; in contrast to the read case, log writes continue throughout the experiment, at an average rate of approximately 14 KB/s. On further examination, these writes target paths such as `/var/log`, `/hadoop/log`, and `/var/run`. (Note that in our deployments, `/tmp`

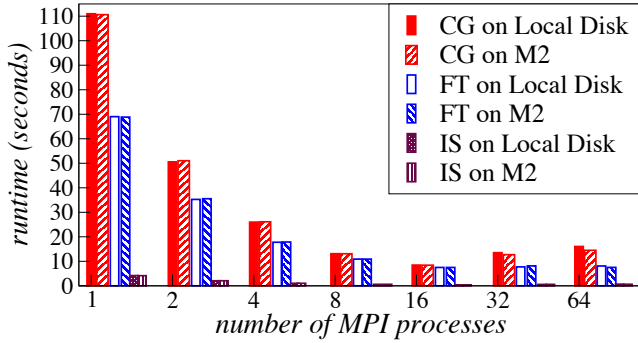


Fig. 7: M2 and local-disk runtime performance comparison of HPC applications (Conjugate-Gradient (CG), Fourier Transform (FT), and Integer Sort (IS) benchmarks from the NAS suite [56]).

and /swap are configured to reside on the local disk of servers.) Most of these writes are log file updates made by Hadoop. Although they could be directed to local storage, we did not do so due to their utility for debugging and negligible impact on the data rate.

H. Performance of M2 Provisioned Systems

In this section we examine the impact of M2 on the performance of applications and frameworks that generally run on bare-metal nodes. To this end, we compared the performance of these applications and frameworks when they run on top of M2-provisioned systems (network-mounted) and systems provisioned via Foreman (installed from a local disk).

1) *HPC Applications Runtime Performance:* In Figure 7, we compare the runtime of HPC applications (FT, CG, IS) from the NAS Parallel Benchmarks (NPB) [56] running on Foreman-provisioned (installed from local disk) and M2 provisioned (network-mounted) clusters. We ran these benchmarks in our second environment. NPB is a set of programs designed to evaluate the performance of parallel supercomputers. Three benchmarks (i.e., FT, IS and CG) with distinct behaviors were used to evaluate the system. IS performs random memory access, CG has an irregular memory access and communication pattern, and FT does frequent all-to-all communications. We used class B of the MPI version of NPB. Each benchmark was compiled to run with 2^n processes where $n \in \{1, \dots, 8\}$. Each build was executed using OpenMPI on local and remote installations.

As shown in Figure 7, almost equal execution times were noted in the case of both M2 and Foreman, resulting in same height bars. The results indicate that M2 and diskless provisioning has no additional overheads when executing CPU- or memory-intensive HPC jobs. Note that it is already known that HPC applications perform well with remote boot drives as such solutions are frequently employed in Beowulf clusters and supercomputers. Hence good performance of M2 is expected in this scenario.

2) *Hadoop Runtime Performance:* To measure M2’s performance under high network and disk I/O usage we tested its

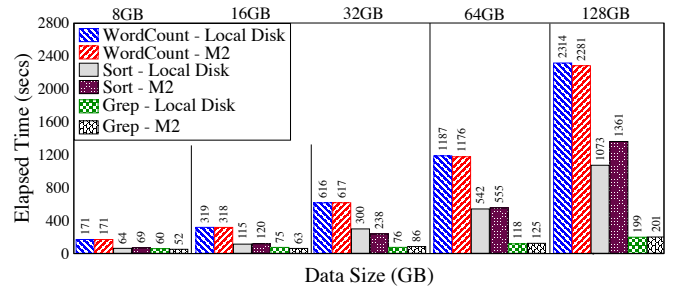


Fig. 8: M2 and local-disk runtime performance comparison of standard Hadoop benchmarks (WordCount, Sort, Grep).

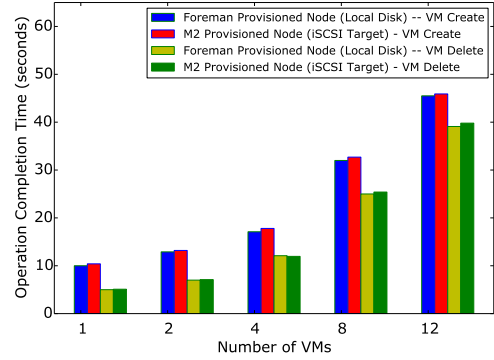


Fig. 9: OpenStack operation performance comparison between Foreman and M2-provisioned nodes.

performance when it runs Hadoop jobs. We performed a series of experiments on an 8-node Hadoop cluster as we varied the data set size between 8GB, 16GB, 32GB, 64GB and 128GB. We used the first environment for these experiments. Figure 8 compares the runtime of standard Hadoop benchmarks (Sort, Grep, WordCount) running on clusters installed from local disks and from network-mounted clusters. In both cases, data disks hosting the Hadoop Distributed File System reside on local disks of the servers. Reported numbers are the average of five runs. We observe that deviations among runs on the same configuration are negligible.

As shown in Figure 8, the difference in runtime performances of local-disk installed and M2 provisioned systems are negligible, with the exception of the Sort experiments for 32GB data and 128GB data. We hypothesize that this exception may be caused by the non-deterministic behavior of random sorting benchmarks. The “good” performance of M2 justifies our hypothesis that even for applications that create a significant amount of network traffic and disk I/O, the performance of the application does not get adversely impacted by remote mounting the boot drive.

3) *OpenStack Operations Performance:* Cloud management systems such as OpenStack that offer virtualized services are also generally deployed on bare-metal servers. In this experiment, we set up OpenStack-based clouds to run on top of M2-provisioned and Foreman-provisioned systems in our second environment. We measured the performance of the two virtual machine operations, namely VM create and VM

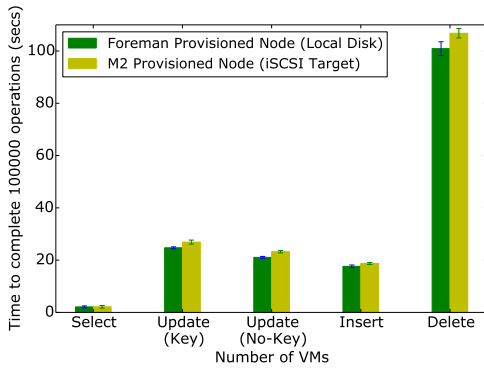


Fig. 10: MariaDB operation latency comparison between Foreman and M2-provisioned nodes.

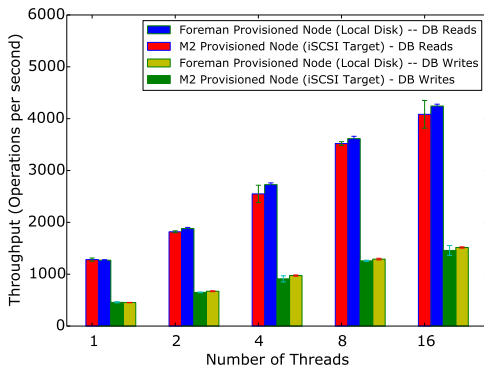


Fig. 11: MySQL read/write throughput comparison between Foreman and M2-provisioned nodes.

delete, in these two setups. We used the Rally benchmarking tool [57] for these experiments, varying the number of parallel operation requests issued between 1 and 12. As shown in Figure 9, negligible performance degradation was observed for both creation and deletion operations between Foreman and M2-provisioned nodes.

4) *Latency and Throughput of Database Operations:* Due to their stringent performance requirements database systems are commonly deployed over bare-metal servers. To test if M2-provisioned servers can provide satisfactory performance while running database applications we compared the latency and throughput of various database operations when running commonly used databases on nodes provisioned via M2 versus Foreman. Note that again, the data disks holding the actual database data is hosted on local disks in both cases.

Figure 10 compares the latency of database operations when running the popular MariaDB database on a node provisioned via M2 versus Foreman. This experiment was performed using the Sysbench benchmarking tool [58] in our second environment. Multi-threaded Online Transaction Processing (OLTP) tests for Select, Update, Insert and Delete operations were performed on the default “sbtest” [59] table generated by Sysbench with 1 million rows with InnoDB as the storage engine for MariaDB. The number of select operations executed during the test was 100,000, whereas 10,000 operations were executed for each of update, insert and delete operations. The update operation test had two versions — updating an indexed

column (Key) and updating a non-indexed column (No-key). For each test, the number of threads was fixed at 4.

As seen in Figure 10, there is a negligible impact on the latency of the select operation for MariaDB when running on a M2 provisioned system. In contrast, in the case of update, insert and delete operations, we observe $\sim 4\%$ degradation in the case of M2 provisioned nodes.

Figure 11 compares the MariaDB read and write throughput when it runs on a node provisioned via M2 and Foreman. This experiment was also performed using Sysbench. In this experiment, we measured the total number of random reads and random writes performed in 300 seconds — varying the number of threads. The throughput of both random reads and random writes in the case of M2-provisioned nodes was either at par with their Foreman-provisioned counterparts or saw a degradation less than 5%. These experiments were also executed in our second environment.

Results in Figure 10 and Figure 11 indicate that the impact of remote-mounting the boot drive to database performance is less than 5%. This is potentially due to the excessive system memory use of the databases. Considering the additional benefits M2 offers such as easy fault recovery and easy backup, we believe many deployments will find this additional impact tolerable.

VI. FUTURE WORK

A. Scaling

We believe that the evaluation shows the performance of M2 to be sufficient at moderate scales and that the boot disk is not even necessarily a major factor for ongoing application performance. When M2 does need to scale up, at least three strategies could be mixed and matched to do so.

iSCSI multipath: One of the advantages of selecting iSCSI as the gateway protocol is that many iSCSI clients support multipathing support. This means that clients can distribute queries across a number of iSCSI endpoints for both performance and redundancy. When paired with a highly-scalable backend filesystem like Ceph or Lustre, this would mean that M2 implementations could separately scale the iSCSI Service and the backend Storage Service.

Caching: Without breaking the consistency model, where the iSCSI Service is just a gateway and the backend filesystem is the coherence point, disks could be segmented into read-only and read-write components. For example, the /usr filesystem could be a read-only disk that is updated infrequently and /var could be on a read-write partition. Segmenting the disk this way would allow iSCSI Service to cache the read-only partitions locally, reducing load on the Storage Service overall and especially offloading the Storage Service in case of read-only hotspots.

Custom hardware components: Several vendors, like NetApp, Dell and EMC, sell high performance storage infrastructures that manage scaling and redundancy, and which expose that storage via an iSCSI endpoint. Due to the modular design of M2 these commercial solutions could be employed as the storage service to improve scalability.

B. Improved security

Threats against M2 can come from: the publicly-facing API service, the M2-provider itself or within an M2-serviced network; the cloud provider is trusted. For threats against the API, M2 relies on pluggable authentication modules. In the case where tenants do not trust the cloud provider, they can maintain some confidentiality and integrity using encryption such as Linux LUKS [60] or Windows Bitlocker [61], though additional protections might be needed to prevent side channel analysis [62] or block-level replay attacks.

In the current implementation, hostile nodes active on the network M2 is managing can pose a threat due to the potential for accessing the boot disks of other nodes on the same network. This is in part due to the lack of strong identity inherent in several of the protocols M2 uses. For example, one node could pretend to be another node booting, by spoofing its MAC address; a node could intercept the active iSCSI connection of another, since modern iSCSI implementations do not authenticate or encrypt every packet. Transport Layer Security (TLS) [63] or IPsec [64] could be applied to reduce this risk, but there still exists the bootstrapping problem: how does M2 differentiate between a real node, and a compromised node that is faking its MAC or IP address?

To address this issue we are working on a solution that relies on Trusted Platform Modules [65], which are tamper-resistant, discrete chips contained in some bare-metal nodes that give the node a cryptographic identity via a local public/private keypair. A security-sensitive tenant could use the TPM as part of a protocol to grant cryptographically-guarded access to M2 resources using attestation systems like [66], [67]. Such systems could also provide a defense against corrupted firmware. We are working on this at present [68], though details of this work is out of scope for this paper.

C. Transition between physical and virtual

One interesting thought we had was: if M2's goal is providing VM-like management capabilities for bare metal images, would it be possible to transition nodes between physical and virtual nodes? This could be especially helpful for using the right amount of resources in HPC, dev/test or staging, where a lighter VM could be used for development and then a physical node for performance or predictability. Such a system could function by taking a Ceph RBD instance in use by a VM (a common choice in OpenStack clusters), then using M2 to export that to a physical node via PXE and iSCSI. A preliminary test within our M2 environment was able to do just this using a standard Linux distribution as the image. We are working to generalize this solution as a standard feature in the next M2 release.

VII. CONCLUSION

In this work we proposed M2 a system that brings the attractive image management capabilities (such as fast snapshotting, cloning, rapid provisioning etc.) of virtualized solutions to bare-metal systems. M2 makes use of remote-mounted boot drives to host user images containing the operating system

and applications and exploits advancements in disaggregated storage and networking technologies to offer high performance. Our analysis show that M2 provisioned systems and frameworks perform as well as local-disk-based systems. We also show that rapid provisioning and snapshotting capabilities of M2 unleash additional features and capabilities such as elasticity and support for fast transition among different frameworks for datacenter administrators.

VIII. ACKNOWLEDGMENT

We gratefully acknowledge Sourabh Bollapragada, Naved Ansari, Daniel Finn, Sirushti Murugesan and Paul Grosu for their significant contributions in development and documentation of M2. Also, Piyanai Saowarattitada, Chris Hill, Radoslav Nikiforov Milanov, Laura Kamfonik, Rahul Sharma, Rajul Kumar, and Sourabh Bollapragada for their assistance in the evaluations.

Partial support for this work was provided by the MassTech Collaborative Research Matching Grant Program, National Science Foundation awards ACI-1440788, 1347525, 1149232 and 1414119 as well as the several commercial partners of the Massachusetts Open Cloud which include Brocade, Cisco, Intel, Lenovo, Red Hat, and Two Sigma.

REFERENCES

- [1] ZDNet, "Facebook: Virtualisation does not scale," <http://www.zdnet.com/article/facebook-virtualisation-does-not-scale/>, 2011.
- [2] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 6, pp. 931–945, Jun. 2011.
- [3] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 199–212.
- [4] Softlayer, "SoftLayer to outflank rivals with bare metal, InfiniBand, and Power8," 2014. [Online]. Available: <https://www.enterprisetech.com/2014/07/30/softlayer-outflank-rivals-bare-metal-infiniband-power8/>
- [5] S. K. Barker and P. Shenoy, "Empirical evaluation of latency-sensitive application performance in the cloud," in *Proceedings of the First Annual ACM SIGMM Conference on Multimedia Systems*, ser. MMSys '10. New York, NY, USA: ACM, 2010, pp. 35–46.
- [6] S. M. Trimberger and J. J. Moore, "FPGA security: Motivations, features, and applications," *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1248–1265, Aug 2014.
- [7] C. Maurice, C. Neumann, O. Heen, and A. Francillon, *Confidentiality Issues on a GPU in a Virtualized Environment*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 119–135.
- [8] AWS, "Power of HPC on AWS," 2017. [Online]. Available: <https://aws.amazon.com/hpc/>
- [9] —, "Amazon EC2 elastic GPUs," 2017. [Online]. Available: <https://aws.amazon.com/hpc/>
- [10] Cirrascale, "Cloud services for deep learning," 2017. [Online]. Available: <http://www.cirrascale.com/cloud/>
- [11] D. Schatzberg, J. Cadden, H. Dong, O. Krieger, and J. Appavoo, "Ebbri: A framework for building per-application library operating systems." in *OSDI*, 2016, pp. 671–688.
- [12] Softlayer, "Big data solutions," <http://www.softlayer.com/big-data>, 2015.
- [13] Rackspace, "Rackspace cloud big data OnMetal," <http://go.rackspace.com/baremetalbigdata/>, 2015.
- [14] Internap, "Bare-Metal AgileSERVER," <http://www.internap.com/bare-metal/>, 2015.
- [15] Openstack, "Ironic," <http://docs.openstack.org/developer/ironic/deploy/user-guide.html>, 2015.
- [16] Canonical, "Metal as a service (MAAS)," <http://maas.ubuntu.com/docs/>, 2015.

- [17] D. Anderson, M. Hibler, L. Stoller, T. Stack, and J. Lepreau, "Automatic online validation of network configuration in the Emulab network testbed," in *IEEE International Conference on Autonomic Computing, 2006. ICAC '06*, Jun. 2006, pp. 134–142.
- [18] OpenCrowbar, "The Crowbar project," <https://opencrowbar.github.io, 2015>.
- [19] Puppetlabs, "Provisioning with Razor," https://docs.puppetlabs.com/pe/latest/razor_intro.html, 2015.
- [20] Y. Omote, T. Shinagawa, and K. Kato, "Improving agility and elasticity in bare-metal clouds," in *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '15. New York, NY, USA: ACM, 2015, pp. 145–159.
- [21] S. A. Weil, S. A. Brandt, E. L. Miller, D. D. E. Long, and C. Maltzahn, "Ceph: A scalable, high-performance distributed file system," in *Proceedings of the 7th Symposium on Operating Systems Design and Implementation*, ser. OSDI '06. Berkeley, CA, USA: USENIX Association, 2006, pp. 307–320. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1298455.1298485>
- [22] —, "Ceph storage," 2017. [Online]. Available: <http://ceph.com/ceph-storage/>
- [23] J. Hennessey, S. Tikale, A. Turk, E. U. Kaynar, C. Hill, P. Desnoyers, and O. Krieger, "HIL: Designing an exokernel for the data center," in *Proceedings of the Seventh ACM Symposium on Cloud Computing*, ser. SoCC '16. New York, NY, USA: ACM, 2016, pp. 155–168. [Online]. Available: <https://doi.acm.org/10.1145/2987550.2987588>
- [24] Cobbler, "Cobbler," <https://cobbler.github.io, 2015>.
- [25] A. Chandrasekar and G. Gibson, "A comparative study of baremetal provisioning frameworks," Parallel Data Laboratory, Carnegie Mellon University, Tech. Rep. CMU-PDL-14-109, 2014.
- [26] Foreman, "Foreman provisioning and configuration system," <http://theforeman.org>.
- [27] D. Van der Veen *et al.*, "Openstack Ironic Wiki," <https://wiki.openstack.org/wiki/Ironic>.
- [28] vmware, "Hidden benefits of virtualisation reboot time and the impact on server availability and regular operations," <https://blogs.vmware.com/tam/2013/05/hidden-benefits-of-virtualisation-reboot-time-and-the-impact-on-server-availability-and-regular-operations.html, 2017>.
- [29] Y. Klimenko, "Technique for reliable network booting of an operating system to a client computer," Oct. 26 1999, uS Patent 5,974,547. [Online]. Available: <https://www.google.com/patents/US5974547>
- [30] D. Sposato, "Method and apparatus for remotely booting a client computer from a network by emulating remote boot chips," Oct. 8 2002, uS Patent 6,463,530. [Online]. Available: <https://www.google.com/patents/US6463530>
- [31] C. Haun, C. Prouse, J. Sokol, and P. Resch, "Providing a reliable operating system for clients of a net-booted environment," Jun. 15 2004, uS Patent 6,751,658. [Online]. Available: <https://www.google.com/patents/US6751658>
- [32] K. Salah, R. Al-Shaikh, and M. Sindi, "Towards green computing using diskless high performance clusters," in *Network and Service Management (CNSM), 2011 7th International Conference on*. IEEE, 2011, pp. 1–4.
- [33] B. Guler, M. Hussain, T. Leng, and V. Mashayekhi, "The advantages of diskless hpc clusters using nas," *Technical Report Dell Power Solutions, 2002*.
- [34] D. Daly, J. H. Choi, J. E. Moreira, and A. Waterland, "Base operating system provisioning and bringup for a commercial supercomputer," in *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*. IEEE, 2007, pp. 1–7.
- [35] R. Lewis, "Virtual disk image system with local cache disk for iscsi communications," Aug. 2 2005, uS Patent 6,925,533. [Online]. Available: <https://www.google.com/patents/US6925533>
- [36] B. Cully, G. Lefebvre, D. Meyer, M. Feeley, N. Hutchinson, and A. Warfield, "Remus: High availability via asynchronous virtual machine replication," in *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*. San Francisco, 2008, pp. 161–174.
- [37] M. Nelson, B.-H. Lim, G. Hutchins *et al.*, "Fast transparent migration for virtual machines," in *USENIX Annual technical conference, general track, 2005*, pp. 391–394.
- [38] M. Chadalapaka, J. Satran, K. Meth, and D. Black, "Internet Small Computer System Interface (iSCSI) Protocol (Consolidated)," RFC 7143 (Proposed Standard), RFC Editor, Fremont, CA, USA, pp. 1–295, Apr. 2014. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc7143.txt>
- [39] VMware.com, "VMware Workstation 5.0 Understanding Clones," https://www.vmware.com/support/ws5/doc/ws_clone_overview.html, 2017.
- [40] S. Hogg, "Clos networks: What's old is new again," Jan 2014. [Online]. Available: <https://www.networkworld.com/article/2226122/cisco-subnet-clos-networks-what-s-old-is-new-again.html>
- [41] A. Klimovic, C. Kozyrakis, E. Thereska, B. John, and S. Kumar, "Flash storage disaggregation," in *Proceedings of the Eleventh European Conference on Computer Systems*. ACM, 2016, p. 29.
- [42] S. A. Weil, *Ceph: reliable, scalable, and high-performance distributed storage*, 2007, vol. 68, no. 11.
- [43] "About the lustre file system." [Online]. Available: <http://lustre.org/about/>
- [44] J. Bonwick and B. Moore, "Zfs: The last word in file systems," 2007. [Online]. Available: https://wiki.illumos.org/download/attachments/1146951/zfs_last.pdf
- [45] D. Teigland and H. Mauelshagen, "Volume managers in Linux," in *USENIX Annual Technical Conference, FREENIX Track, 2001*, pp. 185–197. [Online]. Available: http://static.usenix.org/legacy/events/usenix01/freenix01/full_papers/teigland/teigland_html/
- [46] OpenStack, "Ceph IO Performance." [Online]. Available: https://docs.openstack.org/performance-docs/latest/test_results/ceph_testing/index.html#ceph-rbd-performance-results-50-osd
- [47] V. Inc, "iSCSI Performance Depends on Storage Performance." [Online]. Available: <https://docs.vmware.com/en/VMware-vSphere/6.0/com.vmware.vsphere.storage.doc/GUID-548C8064-23DB-44EB-8FFC-BFEF5D39DA3A.html>
- [48] F. Tomonori and M. Christie, "tgt: Framework for storage target drivers," in *Linux Symposium, 2006*.
- [49] D. Merkel, "Docker: Lightweight Linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, Mar. 2014. [Online]. Available: <https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment>
- [50] mcb30, "iPXE: Open Source Boot Firmware," 2015. [Online]. Available: <http://ipxe.org>
- [51] Microsoft, "About iSCSI Boot," 2009. [Online]. Available: [https://technet.microsoft.com/en-us/library/ee619722\(v=ws.10\).aspx](https://technet.microsoft.com/en-us/library/ee619722(v=ws.10).aspx)
- [52] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "A large-scale study of flash memory failures in the field," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 1. ACM, 2015, pp. 177–190.
- [53] —, "Revisiting memory errors in large-scale production data centers: Analysis and modeling of new trends from the field," in *Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on*. IEEE, 2015, pp. 415–426.
- [54] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 8, no. 3, pp. 1–154, 2013.
- [55] B. Schroeder and G. Gibson, "A large-scale study of failures in high-performance computing systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 4, pp. 337–350, 2010.
- [56] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, L. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. Schreiber *et al.*, "The nas parallel benchmarks," *The International Journal of Supercomputing Applications*, vol. 5, no. 3, pp. 63–73, 1991.
- [57] "Rally Benchmarking Tool for OpenStack." [Online]. Available: <https://wiki.openstack.org/wiki/Rally>
- [58] A. Kopytov, "Sysbench: a system performance benchmark," URL: <http://sysbench.sourceforge.net, 2004>.
- [59] —, "Sysbench manual," *MySQL AB*, 2012.
- [60] C. Fruhwirth, "New methods in hard disk encryption," 07 2005. [Online]. Available: <http://clemens.endorphin.org/nmihde-A4-ds.pdf>
- [61] "Bitlocker drive encryption overview." [Online]. Available: [https://technet.microsoft.com/en-us/library/cc732774\(v=ws.11\).aspx](https://technet.microsoft.com/en-us/library/cc732774(v=ws.11).aspx)
- [62] E. Stefanov and E. Shi, "ObliviStore: High performance oblivious cloud storage," in *2013 IEEE Symposium on Security and Privacy*. Berkeley, CA, USA: IEEE Computer Society Press, May 19–22, 2013, pp. 253–267.
- [63] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2," RFC 5246 (Proposed Standard), RFC Editor, Fremont, CA, USA, pp. 1–104, Aug. 2008, updated by RFCs 5746, 5878, 6176, 7465, 7507, 7568, 7627, 7685, 7905, 7919. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc5246.txt>

- [64] S. Kent and K. Seo, "Security Architecture for the Internet Protocol," RFC 4301 (Proposed Standard), RFC Editor, Fremont, CA, USA, pp. 1–101, Dec. 2005, updated by RFCs 6040, 7619. [Online]. Available: <https://www.rfc-editor.org/rfc/rfc4301.txt>
- [65] T. C. Group, "Trusted Platform Module (TPM)," [Online]. Available: <https://trustedcomputinggroup.org/work-groups/trusted-platform-module/>
- [66] N. Schear, P. T. Cable, II, T. M. Moyer, B. Richard, and R. Rudd, "Bootstrapping and maintaining trust in the cloud," in *Proceedings of the 32Nd Annual Conference on Computer Security Applications*, ser. ACSAC '16. New York, NY, USA: ACM, 2016, pp. 65–77.
- [67] N. Santos, R. Rodrigues, K. P. Gummadi, and S. Saroiu, "Policy-sealed data: A new abstraction for building trusted cloud services," in *In USENIX Security*, 2012.
- [68] M. O. Cloud, "Secure cloud." [Online]. Available: <https://massopen.cloud/blog/secure-cloud/>