The Impact of Management Operations on the Virtualized Datacenter

Vijayaraghavan Soundararajan and Jennifer M. Anderson
VMware, Inc.
3401 Hillview Ave.
Palo Alto, CA 94306
1 (650) 427-1706
{ravi,jennifer}@vmware.com

ABSTRACT

Virtualization has the potential to dramatically reduce the total cost of ownership of datacenters and increase the flexibility of deployments for general-purpose workloads. If present trends continue, the datacenter of the future will be largely virtualized. The base platform in such a datacenter will consist of physical hosts that run hypervisors, and workloads will run within virtual machines on these platforms. From a system management perspective, the virtualized environment enables a number of new workflows in the datacenter. These workflows involve operations on the physical hosts themselves, such as upgrading the hypervisor, as well as operations on the virtual machines, such as reconfiguration or reverting from snapshots. While traditional datacenter design has focused on the cost vs. capability tradeoffs for the end-user applications running in the datacenter, we argue that the *management workload* from these workflows must be factored into the design of the virtualized datacenter.

In this paper, we examine data from real-world virtualized deployments to characterize common management workflows and assess their impact on resource usage in the datacenter. We show that while many end-user applications are fairly light on I/O requirements, the management workload has considerable network and disk I/O requirements. We show that the management workload scales with the increasing compute power in the datacenter. Finally, we discuss the implications of this management workload for the datacenter.

Categories and Subject Descriptors

C.4 [*Computer Systems Organization*]: Performance of Systems

General Terms

Performance, Management, Measurement, Design.

Keywords

Virtual Machine management, cloud computing, datacenter management, management workload

ISCA'10, June 19-23, 2010, Saint-Malo, France.

1. INTRODUCTION

In a virtualized datacenter, applications are run in virtual machines (VMs) on top of physical hosts. Virtualization affords significant flexibility. As a result, many enterprises, hosting providers, and cloud vendors have shifted to a virtualization-based model for running applications and providing various services like email, web serving, and payroll [25]. This flexibility allows workflows that would be much more onerous in a physical environment. For example, one key feature of virtualization platforms is the ability to move a VM between physical hosts while the VM is running. When maintenance is required in a virtualized datacenter, one can simply migrate the VMs off the server to another with no application down-time. While this capability increases productivity and allows the operation to occur regularly with ease, it comes with a cost: this live migration exerts additional load on the datacenter's network infrastructure, potentially requiring higher-performance network components than would be required strictly by the needs of the applications running in the datacenter. Moreover, virtualization may change other decisions within the datacenter. For example to achieve the best performance for virtualized workloads, it is desirable to use the latest CPUs with hardware support for virtualization. Also, since hundreds of mission-critical VMs may be running per host, purchasing more reliable yet more expensive components may be most cost-effective overall. Each of these issues factors into the cost vs. capability tradeoffs inherent in datacenter design.

In this paper, we discuss various tradeoffs in the design of a datacenter for virtualized environments. We consider in particular an often-overlooked aspect of datacenter design: the management workload. The management workload stresses parts of the system that are historically under-utilized in typical datacenters, namely, the storage and network subsystems. The flexibility that virtualization brings to datacenter operations results in new types of management workflows and places new demands on datacenter designs. The load is fundamental to the workflows and is independent of the virtualization layer being used.

The contributions of this paper are as follows. First, we introduce data from real deployments and provide insight into various workflows in the virtual datacenter. Second, we present a sensitivity analysis of how requirements change as the datacenter scales. Finally, we discuss the impact of these operations on the design of a datacenter for virtualized environments.

The outline of the paper is as follows. In Section 2, we describe the virtualized datacenter in more detail. In Section 3, we characterize the management workload in the virtualized datacenters that we have profiled. In Section 4, we examine how the management workload scales with increased compute-power in the datacenter. In Section 5, we discuss the implications for datacenter design. In Section 6, we discuss related work. Finally, we summarize this work in Section 7.

2. THE VIRTUALIZED DATACENTER

A datacenter consists of physical hosts, storage and networking infrastructure, power/cooling hardware, and a management/monitoring framework for ongoing operation. The hosts are organized into racks. Power is routed to each rack, and each rack also includes some amount of storage and network hardware.

To allow multiple hosts to share data and to simplify backup, most datacenters utilize shared storage like Storage Area Networks (SANs) [11] or Network-attached storage (NAS). SANs export a block-device interface to each host, so that the device appears to be a local hard drive even if it resides in a different rack. Common SAN protocols include FibreChannel and iSCSI. FibreChannel SANs use special-purpose cabling and switches, whereas iSCSI [20] relies on standard IP networking infrastructure. A NAS exports file-level storage and also uses standard networking infrastructure. These devices are designed for high-bandwidth, low-latency shared access by large numbers of hosts.

In addition to shared storage, a significant component of a datacenter is the networking infrastructure. There are rack switches and core infrastructure switches that are used to allow connectivity in a datacenter. Often, a single networking infrastructure is used to multiplex both storage-based traffic (NAS and iSCSI traffic) as well as standard network traffic.

Datacenters also include a management framework in order to assess the health of the components of the datacenter. An administrator has tools to configure networking and storage, power-cycle hosts, or access the console of the hosts. Once machines have been provisioned and workloads have been set up in a production datacenter, a great deal of datacenter management consists of monitoring application service levels. Administrators must also monitor the hardware to check for failures and fix failures when they occur [9].

A virtualized datacenter consists of the same hardware components as a physical datacenter. However, in a virtualized datacenter, each host contains a hypervisor layer for running virtual machines on top of the physical hosts. In addition, the management framework is extended to include management and monitoring of both VMs and hosts.

Deploying new applications or adding new users can be time-consuming in non-virtualized environments because of the need to provision physical compute, storage, or networking resources. In a virtual environment, however, the virtualization layer allows such operations to occur much more frequently.

Consider an IT administrator who must configure a test and development setup for one group, and then must replicate the setup for a different group. In a physical environment, this may occur rarely because of the complexity of setting up a group of hosts and then duplicating the setup. In a virtual environment, however, simply performing a snapshot and cloning a group of VMs is sufficient to perform the same task. As a result, virtual environments see more provisioning or reconfiguration operations than physical environments.

Another task unique to virtualization is live migration of VMs [21]. In live migration, a VM is migrated from one physical host to another while it is running. Live migration can be used for maintenance cycles, for reducing power consumption by powering-off unneeded servers, or for load balancing of VMs across the physical hosts. Live migration can use significant network bandwidth, and additional NICs may be required simply to support the virtual management operations.

We refer to the ongoing operations in a virtualized datacenter that involve the maintenance and monitoring of the physical and virtual hosts as the *management workload*. Management workloads consist of critical operations whose completion time is of paramount importance. If a set of VMs must be provisioned before the start of a business day, the administrator must be able to accurately plan how long that will take, and must ensure that the task is completed in time.

Table 1: Resource Consumption for a Set of Common Applications.
On average, applications have modest resource requirements. For disk and networking, peaks are also within modest bounds.

Application	CPU Usage (%)		Avg. Memory Usage	Disk Bandwidth		Network Bandwidth	
	Average	Peak		Average	Peak	Average	Peak
Apache	5.7	32	4 GB	243 KBps	1.5 MBps	1.9 KBps	94.0 KBps
Cognos 8	1.1	3	2 GB	151 KBps	1.2 MBps	0.6 KBps	5.8 KBps
DB2	3.4	34	6 GB	498 KBps	31.0 MBps	5.2 KBps	170.0 KBps
Oracle	5.3	96	2 GB	286 KBps	19.0 MBps	5.0 KBps	665.0 KBps
SAS	7.2	90	2 GB	290 KBps	7.8 MBps	0.5 KBps	4.4 KBps
Siebel	5.8	27	8 GB	392 KBps	5.3 MBps	55.0 KBps	1.3 MBps
SQL 2000	5.3	64	2 GB	309 KBps	16.0 MBps	32.0 KBps	1.4 MBps
Windows 2003	3.0	70	4 GB	325 KBps	35.0 MBps	13.0 KBps	1.6 MBps

An important distinction between the management workload and the application workload is the type of resources used. From profiling numerous customer environments [27], we see that many applications have modest CPU, memory, network, and disk requirements, as shown in Table 1. These are average values over a wide range of servers. In addition, many organizations have some key applications (like databases) that can consume large amounts of network and disk bandwidth. The datacenter designer often explicitly over-provisions for these servers. Hot spots, when they occur, typically result in spikes in CPU usage, and disk and network remain modest. For example, the applications use at most ~500KBps disk bandwidth on average, with peaks up to 35MBps. The network requirements are at most 55 KBps on average, with peaks of a maximum 1.6MBps, well below the bandwidth available in commodity 1Gbps network cards. Despite the modest storage and networking requirements of many applications, administrators often order servers with multiple storage adapters (HBAs) and NICs. These resources are often used for management operations like backups or for lights-out management [13]. This indicates that on a per-host level, server administrators must factor in the resource usage of management operations. With virtualization, the number of applications consolidated onto a single host increases, and that the same sorts of factors start to apply at the rack and datacenter level.

3. MANAGEMENT OPERATIONS IN VIRTUALIZED DATACENTERS

We start with a question: what management operations do administrators perform in virtualized environments? To help answer this question, we collected detailed profiles of management workload activity from 17 enterprise datacenters running VMware's virtualization software [26].

Table 2 lists the most common operations found in the datacenters we profiled. Each datacenter is a different size, so the operations occur at different frequencies per datacenter. For each operation, we chose a single site and computed the average number of operations per day by taking the total number of operations performed and dividing by the number of days the management server was running. To illustrate the burstiness of management operations, we also examined the profiles to find the maximum number of times the operation was performed on a given day. Note that this peak number of operations is sometimes several orders of magnitude more than the average.

Many of the operations in Table 2 are self-explanatory, but for completeness, we define them here. *VM reconfigure* is reconfiguring the hardware for a VM (e.g., adding a NIC or adding a disk). *VM powerOn* and *powerOff* are powering on and off VMs, respectively, and *VM reset* is soft-resetting a VM (the equivalent of hitting the reset switch on a physical host). *Automated Live Migration* is load balancing performed in an automated manner [30] by moving VMs between hosts while the VMs are powered on. Live migration requires shared storage between the source and destination hosts. *Patch Install* involves installing a patch on a physical host (e.g., updating the hypervisor) or installing a patch in a VM (e.g., updating the guest OS with the latest security fixes). A *Create Snapshot* operation checkpoints the state of a VM. This allows a user to perform operations on the VM and then rollback to a known state in case of failure. A common use for snapshots is when installing the latest version of an application. The user snaphots the VM and then installs the software. If the installation succeeds, the snapshot can be removed and the user simply continues from the current state. If the software is buggy or crashes, however, the user can *revert* the snapshot, restoring the VM to the checkpointed state without the software installed. *Committing a snapshot* means writing to disk all of changes that occurred since the VM snapshot was taken, and removing the snapshot file. Finally, *VM clone* creates a replica of a powered-off VM. This is useful when duplicating a configuration, as mentioned earlier: for example, when a new employee joins a company, the standard desktop VM image can quickly be deployed to the employee's computer.

Table 2: Common Management Operations in Virtualized Datacenters. The average and peak number of operations per day at a given site are shown. Differences between average and peak frequency demonstrate the burstiness of management activity.

Operation	Average Number Per Day at Various Sites	Peak Per Day at Various Sites
VM reconfigure	2.3	699
Automated Live Migration	51	3156
VM powerOn	90	1576
VM powerOff	35	1535
VM reset	4.6	176
Patch Install	5.3	250
Create Snapshot	4.8	56
Snapshot Revert	7.0	101
Snapshot Commit	13	19
VM Clone	6.0	44

The customer data shows that even in production virtualized environments, there is some amount of ongoing management traffic. VM-related operations dominate the list, which makes sense since the number of VMs is typically much larger than the number of hosts. The data also illustrates some of the differences between expected management operations in a physical datacenter vs. a virtualized datacenter. For example, in some environments, VMs are powered on an average of 90 times in a day (and sometimes as often as 1500 times per day), and VMs are migrated over 50 times per day. Such activity is not found in datacenters without virtualization, as migrating applications that run directly on physical servers is much more involved than migrating VMs. In addition, physical hosts are seldom powered-on or powered-off except in case of failure or periodic maintenance.

3.1 Virtualized Datacenter Workflows

This aggregate data in Table 2 clearly shows significant day-to-day management activity and also illustrates the burstiness of such activity. The next step is to analyze the management workflows involved so that we can understand the context for these operations. In the subsequent sections, we explore these workflows in more detail.

3.1.1 Periodic Snapshot/Revert

Consider Figure 1, which shows the number of management operations (tasks) performed per day for a deployment over the period of one year. These management tasks consist of operations listed in Table 2 as well as other host- and VM-specific operations. We can see a periodic background load among some ambient management activity due to automated load balancing.

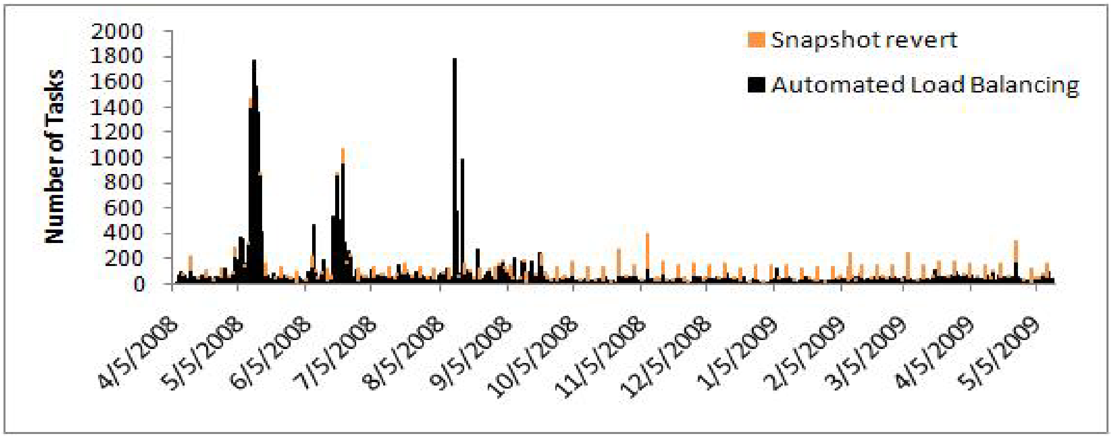

Figure 1: Management Traffic Demonstrating Snapshot/Revert over a One-Year Period. There are periodic spikes in snapshot/revert operations over the course of a year, mixed in with automated load balancing operations.

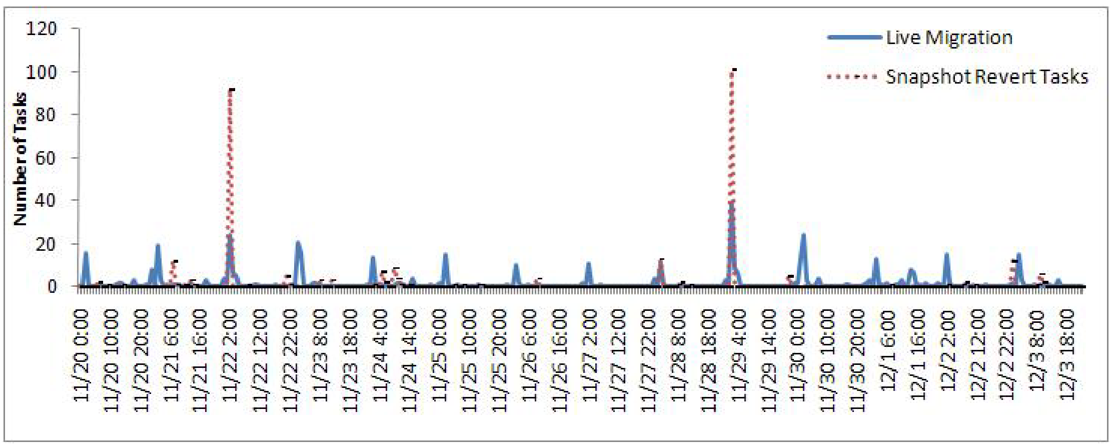

Figure 2: Management Traffic of Snapshot/Revert over a Two-Week Period. Snapshot/revert operations occur once per week to undo the changes during the week. Live migrations occur periodically to balance load due to application resource needs.

Looking more carefully at a selected two-week period in Figure 2, we see that there are small spikes daily around 2am and large spikes each Saturday at 2am (11/22 and 11/29). The tasks are subdivided into live migration tasks and snapshot revert tasks. In Figure 2, the weekly spikes at 2am are snapshot revert tasks, needed because this deployment is used for training purposes. This customer teaches a class. At the beginning of the week, the VM is snapshotted. Throughout the week, the students modify the state of the VM during the class. At the end of the week, the customer reverts the VMs back to the pre-class state, in preparation for the next week's class. This revert of snapshots occurs during a weekly maintenance window in one large burst.

The presence of automated migrations at 2am also indicates some nightly maintenance, most likely backups. These migrations sometimes happen at exactly the same time as the revert operations. Such overlapping of various management tasks can increase the burstiness of the management traffic, further complicating the resource usage issues in the management workload.

3.1.2 Host and VM Patching

Figure 3 illustrates another common management activity: patch management. We see spikes of "patch installation" routines occurring roughly once per month. For this customer, patching is a periodic maintenance event across the entire infrastructure. Once per month, the customer brings down various hosts and VMs and applies the latest patches. Because this occurs across the entire infrastructure, the total number of patch installs is very bursty, occurring 250 at a time for one day per month. On days with patch updates, the number of management tasks is significantly larger than on non-patch days.

3.1.3 After-Hours Maintenance

The previous examples have shown weekly and monthly spikes in management activity. Figure 4 shows a workflow with finer-grain activity, including nearly all of the operations listed in Table 2 plus some additional tasks.

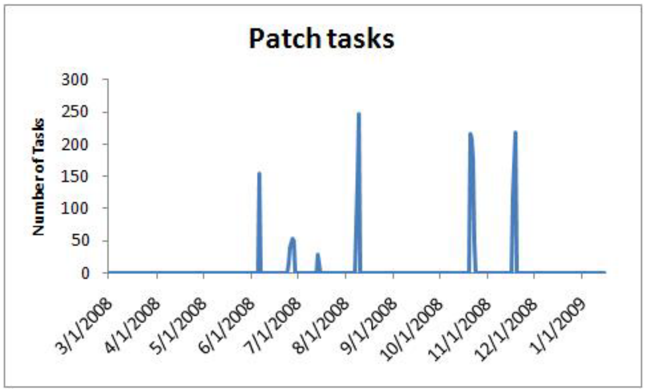

Figure 3: Patching Tasks during a 10-Month Period. Patching occurs approximately once per month, causing a burst in network and disk activity.

We see periodic spikes of activity after 6pm each night for a given two-week period. The majority of these tasks are powering on VMs and committing snapshots. In this case, an IT department was doing a rollout of new VMs and phasing out old ones. Numerous snapshots were taken so that new software could be tested and verified. Once verified, these changes could be committed. The maintenance window was after 6pm each night. At this time, the VM snapshots were committed, and other VMs were powered on and tested in the new environment. This *after-hours maintenance* workflow is fairly common in both physical datacenters and virtualized datacenters, though the operation mix is different between the two. For example, a virtualized datacenter will see snapshot/commit operations, whereas a physical datacenter may see nightly backups of various shared storage devices.

3.1.4 Boot Storm

One of the increasingly common uses of virtualization is "Virtual Desktop Infrastructure" (VDI), in which virtual desktops are hosted on physical machines in a datacenter and are accessed remotely through thin clients. VDI deployments are susceptible to so-called "boot-storms," in which all VMs are turned on at the start of the day as employees come to work. As a result, bursts of hundreds of power operations to VMs can occur daily. Because VDI deployments consist of large numbers of similar virtual machines, they are subject to other types of "storms" as well. For example, anti-virus storms can happen when the virtual machines perform nightly virus scans. Figure 5 shows the number of tasks vs. time in a VDI deployment. Note the burst in tasks around 9am, in which over 2500 tasks occur within 1 hour. Both of these examples (anti-virus storm and boot storm) further illustrate the bursty nature of management traffic.

3.1.5 Automated Live Migration

One of the big advantages of virtualization is automated resource management. An application is contained within a VM, and as the resource needs of the application grow or shrink, the VM can be migrated to other hosts to balance the load. Often, the balancing is done within a specified set of hosts (for example, within the same rack), and high-speed links are used between those hosts. VMware provides a facility known as *Distributed Resource Scheduling* (DRS) [30] for automated load balancing. We contrast these automated tasks with non-automated tasks, such as an administrator manually powering off a VM in order to add a new device or to reconfigure the virtual network connected to the VM.

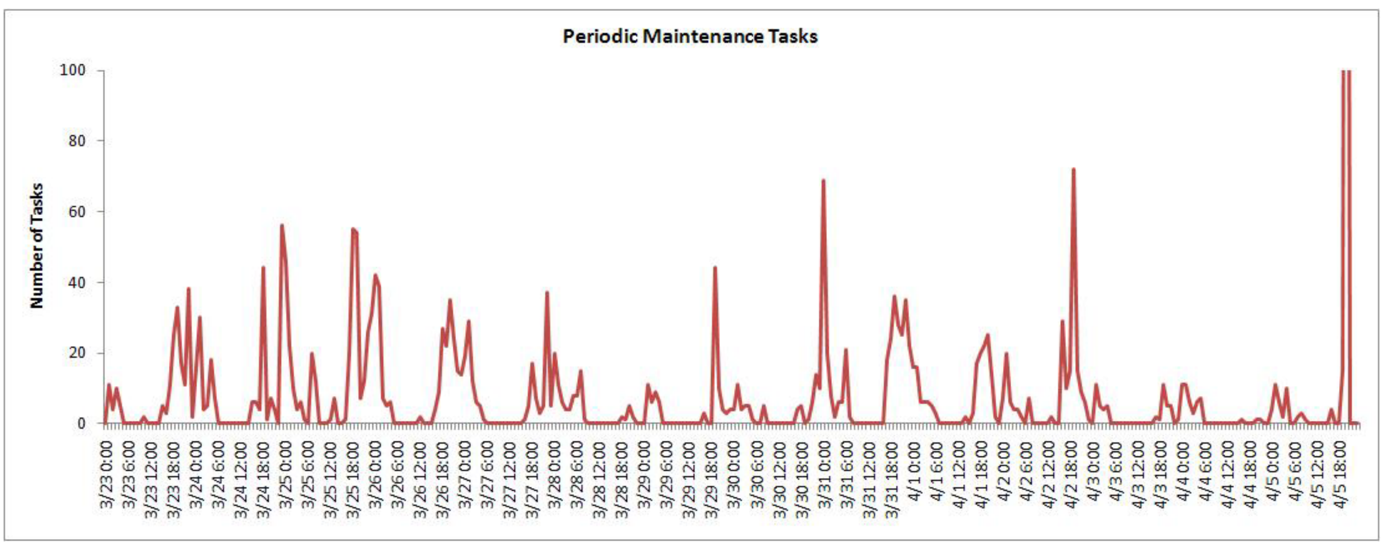

Figure 4: Management Activity Demonstrating After-Hours Maintenance over a Two-Week Period. The management activity at this site occurs after hours during the nightly maintenance window and is significantly higher than the activity during the day.

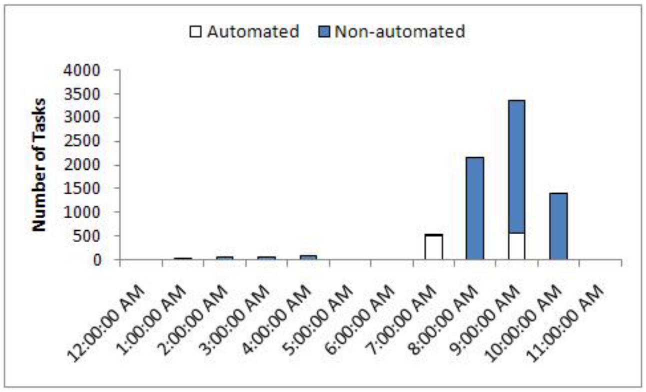

Figure 5: VDI Boot-storm Activity in a Test and Development Environment. Non-automated tasks include user-initiated booting of VMs, while automated tasks include live migration. There is a spike in VM booting at between 8am and 9am, when many users come in to work and power-on their VMs.

Figure 6 shows automated and non-automated tasks for one deployment over a two-week period. The automated tasks were largely live migrations. There are two notable regions. One is a spike in non-automated tasks on 7/31. In this case, a number of host reconfiguration operations were occurring, in preparation for a vast movement of VMs. This movement occurred a few days later on 8/4. In the background, over the entire two-week period we consistently see live migrations in response to application load. We find that a large number of enterprise-class customers enabled automated resource management in their infrastructures, providing an ambient level of management traffic that is specific to virtualized datacenters.

3.2 Multi-tenant Management Issues

All of the previous use cases have considered data from individual datacenters that hosted servers from single corporations. A multi-tenant environment is one in which multiple customers (or departments) share access to the same physical resources. For example, perhaps a provider is hosting Coke and Pepsi in the same virtualized datacenter. The hosting provider must guarantee performance isolation even in the face of unpredictable management traffic. For example, the hosting provider may end up with combinations of all of the workflows that we have seen. Consider Figure 2, with periodic peaks at 2am each Saturday morning and a correspondingly large amount of disk traffic at this time. Suppose this 2am traffic were occurring in Chicago while a VDI deployment from Europe experienced a 9am boot storm. This 9am spike would occur at the same time as the 2am spike in Chicago, and the combination could cause extreme stress on the storage subsystem.

Next, consider the case in which the previous traffic is combined with the nightly patching from Figure 3 or the periodic maintenance in Figure 4, or even the automated load balancing in Figure 6**.** All of these conflicting needs can stress the datacenter more than the application traffic itself, and various design questions arise. For example, shared storage is essential, but should hosts with conflicting management operations be placed in the same rack? How do we place VMs to avoid spikes that occur because of time zone issues? How do we provision the storage and networking to accommodate these bursts? How do we schedule management operations among customers with different demands? These issues demonstrate that considering the management workload is essential in a virtualized datacenter.

4. SCALING OF MANAGEMENT WORKLOAD

As shown in Table 1 many enterprise workloads have light IO requirements. In contrast, a management workload puts a non-trivial load on the IO subsystem. In this section, we discuss the management workload in more detail and perform a sensitivity analysis to see how the management workload scales with increasing compute density.

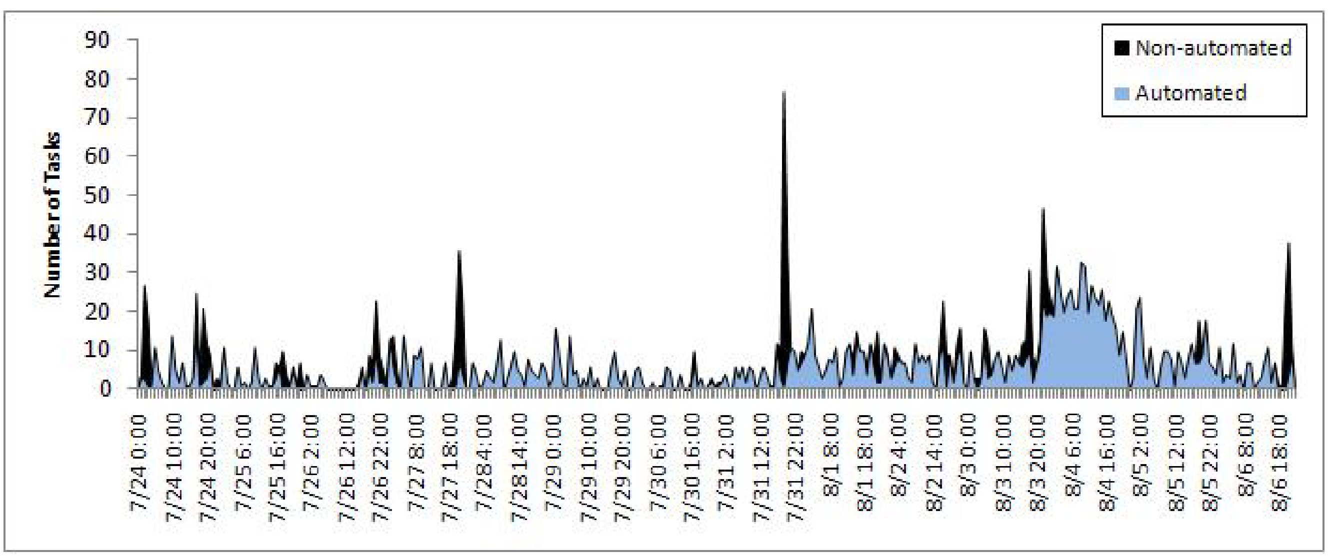

Figure 6: Automated Management Operations over a Two-Week Period. This customer site exhibits a large amount of automated operations (live migrations). The spike in non-automated activity on 7/31 reflects a number of manual host reconfigurations in preparation for a vast movement of VMs. This movement occurred in an automated way starting on 8/3.

4.1 Common Management Workload Patterns

Our customer data shows some common management workload patterns, each exerting a different background load in a virtualized datacenter. In Table 3, we summarize these patterns and indicate whether the load scales with the number of hosts or number of VMs, whether the pattern impacts disk or network, and whether the pattern impacts CPU or memory usage. We discuss this resource usage in more detail in the following sections.

The industry is tending toward larger number of cores per processor socket [3][15]. As the number of cores per socket increases, we will see corresponding increases in the numbers of cores per server, memory per server, and the number of VMs that can run on a server. With greater VM densities, we will see an increase in the management workload for a given datacenter size. In the next few sections, we analyze the resource usage under these various loads as we increase the number of hosts and VMs. The changing resource usage can impact various aspects of datacenter design.

4.2 Datacenter and Server Configurations

For our sensitivity analysis, we try to choose configurations that represent valid design points in the next few years. We focus on the number of hosts in a rack and the compute density in the rack. For the purposes of this study, we assume that the memory in the server is sufficient to meet the VM consolidation demands.

Racks in a datacenter come with a variety of power specifications that support a varying number of hosts. We assume configurations ranging from 10 hosts in a rack, representing the lower end of the power spectrum, to a very dense configuration of 40 hosts in a rack. For our default server building blocks, we use quad-socket servers with 8 cores per socket, for a total of 32 cores per host. Further, we apply a ratio of VMs to cores of 4:1, giving us 128 VMs per server. This assumption is reasonable based on statistics from customer data. Today, 1-way and 2-way VMs are the most common configurations, with only a small number of 4-way VMs. We summarize the various configurations in Table 4.

4.3 Workload Investigation

In this section we examine each of our workloads in turn. We discuss their scaling behavior and analyze their expected resource usage under the configurations listed in Table 4.

Table 3: Workflows and their Resource Usage in Virtualized Datacenters.

Workload Pattern	Host or VM scaling?	Disk or Network impact?	Impact on CPU/Memory usage?
Periodic Snapshot/revert	VMs	Disk	Memory
Host/VM Patching	Both	Both	Both
Boot Storm	VMs	Disk	Both
After-hours Maintenance	Both	Both	Both
Automated Load Balancing	VMs	Network	Both

Table 4: Datacenter and Server Configurations in Sensitivity Analysis.

Configuration	Number of hosts in rack	Cores in rack	VMs in rack
Small	10	320	1280
Medium	20	640	2560
Large	30	960	3840
Future	40	1280	5120

4.3.1 Periodic Snapshot/Revert

The first workload we consider is the periodic reversion of snapshots. To revert a snapshot, a host must read data from disk for each VM being reverted. This data is the memory state of the VM at the time of the snapshot, and its size is correlated with the size of the VM's memory. Reverting a large number of snapshots simultaneously can thus cause a large amount of disk traffic. If a host has 128 VMs, and each VM has a snapshot that is being reverted, then the host must support the disk traffic required for these 128 reverts. Given a time window for reverting snapshots of 30 minutes and 1GB VM memory size, then the host must support 1GB/30minutes = 555 KBps for each revert operation. Hosts running 128 VMs must therefore support (128 × 555 KBps) = 70 MBps of additional disk traffic for the half hour window. Comparing to the data in Table 1, we see that management traffic alone can exceed the disk traffic as the highest peaks for end-user applications. Figure **7** shows the overall rack storage bandwidth requirements for our configurations for the snapshot/revert management workload. We assume that the maintenance window for snapshot/revert is either 30 minutes or 1 hour, similar to what we have seen in practice.

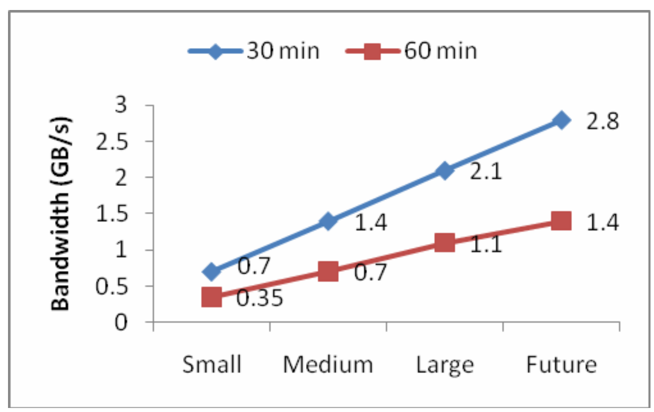

Figure 7: Per-Rack Storage Bandwidth Requirements for Snapshot/Revert Management Workload, 30-minute and 60-minute Maintenance Windows. For a 30-minute maintenance window and the 'Future' configuration with VMs of 1 GB memory size, the entire rack must support up to 2.8GBps of storage bandwidth above and beyond typical application needs.

The rack disk bandwidth requirement at the largest configuration and with the smallest revert window is 2.8 GBps. At first glance, this seems modest for the entire rack, and a typical high-end storage-array network (SAN) [11] can handle this bandwidth. However, 70 MBps must be sustained per host, which suggests that each host must have a 1 Gbps SAN HBA and be connected to a SAN switch with 1 Gbps ports in order to satisfy this particular load. The choice of SAN switch may limit the number of servers that can go in the rack. With the smallest configuration of 10 hosts and 128 VMs/host and a 1-hour revert window, the per-host bandwidth is halved to 35 MBps per host. This bandwidth requirement suggests a less-expensive SAN and less-expensive SAN switch (or perhaps even high-end local storage). Aside from the storage requirements, reverting a large number of snapshots at once can also cause memory pressure and CPU load on the server, in addition to the CPU and memory required to run applications.

In summary, the snapshot/revert workload stresses the per-host storage bandwidth. The system administrator and datacenter designer must work together to plan out the requirements: varying the maintenance window and limiting the number of VMs that are reverted at one time can help reduce the storage bandwidth requirements as well as the CPU and memory requirements. Choosing to place the storage within the rack or outside the rack will impact the choice of storage and storage switch as well as the inter-rack uplink storage bandwidth requirements.

4.3.2 *Patching VMs and Hosts*

The second workload we consider is patch management. Typically, patch management in a virtualized datacenter involves downloading a number of patches from a patch server to each of the patched hosts and VMs. In VMware's Virtual Update Manager [34], patching a host involves downloading a set of patches and rebooting the host until the patches are fully installed. Patching a VM requires the VM to remotely mount an ISO image that is located on the patch server. VM and host patches can be up to hundreds of MBs per month [34]. We assume 20 patches of 30 MB each, for a total patch size per VM of 600MB. Because the number of VMs is two orders of magnitude larger than the number of hosts, we will focus on VM patching.

Assuming the patch server resides outside of the rack, the remote mounting of the ISO image by the VMs requires network traffic from outside the rack to the VMs inside the rack. The length of the maintenance window and the number of VMs in the rack are needed to determine the amount of bandwidth required to handle this patching sequence. In Figure 8, we list the network bandwidth required from the patch server to the rack for our configurations, assuming maintenance windows from 1 hour to 8 hours. We assume that up to 48 VMs can be patched concurrently [34]. Since the number of VMs ranges from 1280 per rack to 5120 per rack (see Table 4), we make the simplifying assumption that it will take (1280/48) = 27 iterations for patching in the smallest configuration, (2560/48) = 54 iterations in the Medium configuration, 80 in the Large configuration, and 107 iterations in the Future configuration: during each iteration, we patch 48 VMs. Each iteration must be able to accommodate the bandwidth requirements for patching 48 VMs concurrently, and all iterations must complete in the maintenance window. For example, in the Large configuration, there are 80 iterations. For a 1-hour maintenance window, we have (1 hour/80 iterations) = 45s per iteration. Since each patch is 600 MB, (48 × 600 MB) = 28800 MB must be downloaded into a rack from the patch server. In a 45s iteration, this means the bandwidth into the rack must be at least (28800 MB/45s) = 640 MBps = 5.1 Gbps.

As Figure 8 shows, the bandwidth into the rack for this management operation alone ranges from 0.22 Gbps to 6.8 Gbps depending on the configuration. The choice of maintenance window is important: 0.22 Gbps into the rack suggests a commodity switch with 1 Gbps uplinks, while 6.8 Gbps suggests a more expensive switch with a higher-bandwidth uplink, or that the patch server should reside within the rack. Moreover, consider the disk bandwidth requirements at the patch server. For a patch server to provide 0.22 Gbps = 27 MBps is no problem: a local disk is fine. However, consider the patch server for the Future scenario and a 1-hour maintenance window. This patch server would need to provide 6.8 Gbps = 853 MBps. Clearly, this is impractical, and multiple patch servers are required in this case. It is also possible to optimize this by caching patches that apply to multiple VMs, avoiding a separate download per VM.

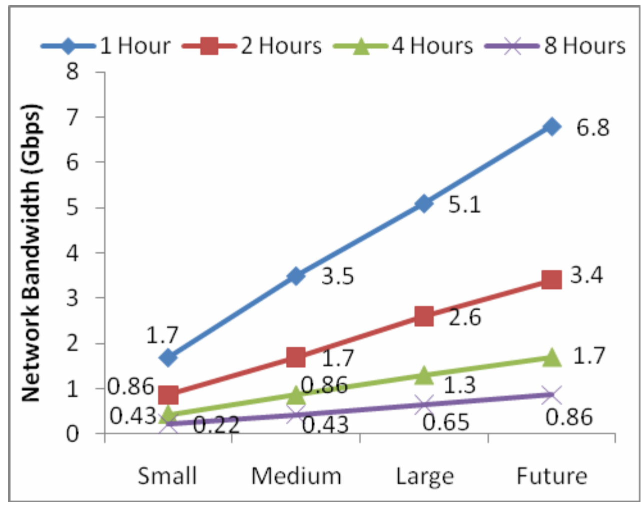

Figure 8: Per-Rack Network Bandwidth Requirements for Patch Maintenance Cycles Ranging from 1 Hour to 8 hours. We assume 48 VMs can be patched concurrently. The Future configuration requires 6.8Gbps of additional network bandwidth per rack to complete within a one-hour maintenance window.

To summarize, for this patch management workload, the bandwidth required into the rack is non-trivial, ranging from 0.22 Gbps to 6.8 Gbps, and depends on a variety of factors, including the number of hosts in the rack, the number of VMs per host, the size of the patches, and the length of the maintenance window. The datacenter administrator and designer have many choices to make, ranging from the location of the patch server (inside or outside the rack), the type of storage for the patch server, the type of switch to handle the management traffic, and the length of the maintenance window, and these tradeoffs can seriously impact the end-user.

4.3.3 *Boot Storm*

In the boot-storm workload, a large number of VMs are powered-on at once. When booting a VM, data is read from disk until the VM is able to run an OS and complete the boot sequence. In our setups, about 200 MB of data is read from disk before a VM can boot. Thus, if 500 VMs were powered-on at once, then this is an additional 500 × 200 MB = 100 GB of disk traffic that the rack must be able to handle. Moreover, booting such a large number of VMs at once stresses the memory system on each host and saturates the CPU. Windows operating systems zero out all of memory before start-up, so each host writes a significant amount of memory for each VM that is booted. This can be optimized a bit in the hypervisor by using a shared zero pages.

In Figure 9, we show the per-rack storage bandwidth requirements for various configurations for the boot-storm workload, assuming 200MB must be read for each VM. We look at boot-storm-windows of 15 minutes and 30 minutes, based on the performance requirements from typical customer deployments.

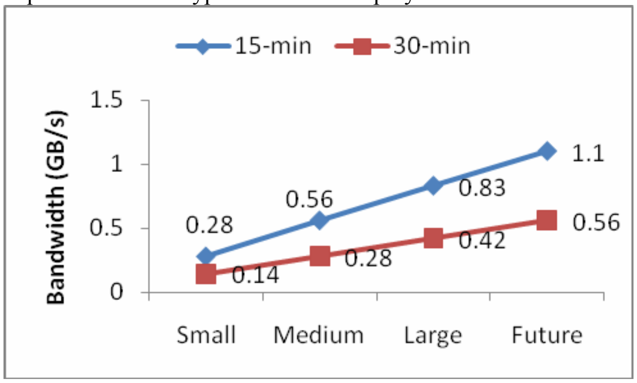

Figure 9: Per-Rack Storage Bandwidth Requirements, Boot-Storm Workload, 15-minute and 30-minute Boot-Storm Windows. A 15-minute window for completing boot-storms for thousands of VMs can incur over 1GBps of storage traffic.

The per-rack storage bandwidth required for the largest configuration and the smallest boot-storm window is 1.1 GBps, which may suggest a high-end SAN device for the entire rack. However, the per-host bandwidth in this case is (128 × 200 MB/15 minutes) = 28.4 MBps. This per-host bandwidth can be satisfied even with a simple local SCSI disk, though reliability concerns may suggest a backup strategy involving SAN or software redundancy [6][9]. Another option is a network-attached storage (NAS), which also would require the rack network switch to satisfy 1.1 GBps, or nearly 10 Gbps. This may require more than a simple commodity switch with four 1 Gbps uplinks.

4.3.4 After-Hours Maintenance

In some environments, we see a large number of management operations after hours, often in settings where rolling upgrades or rolling backups are taking place. These sorts of workloads can cause spikes in disk writes commensurate to the sizes of the VMs being backed up, which can be on the order of GBs per VM. The data we've collected suggests that after-hours maintenance tends to involve small numbers of concurrent operations. The most common operations we have seen include reconfiguring VMs, powering on VMs, cloning VMs, and creating snapshots or committing snapshots of VMs.

The number of concurrent operations varies with the size of the virtualized environment, and the load exerted on the system depends on the type of operation. Reconfiguring a VM is a fast operation with small resource requirements, so we will focus on the other operations instead. As noted before, power operations require loading approximately 200 MB per VM. Creating a snapshot of a running VM requires writing the contents of memory to disk and is proportional to the size of the VM's memory. Committing a snapshot requires merging the contents of a change log to disk, requiring reads from the change log and then writes to disk. In our experience, a rule of thumb for Windows VMs is that the amount of data read (and therefore the amount of data written to disk) is at least the size of the paging file in Windows. For a 2 GB VM, therefore, we would need to read 2 GB from the change log and write those 2 GB to disk.

In the environments we studied, the number of operations per VM is correlated with the size of the environment. In addition, maintenance operations tend to be bursty: they are initiated at once and allowed to complete before more are performed. During maintenance windows we see 20 operations per hour on average for the deployments we examined. Table 5 shows the minimum rack bandwidth required to do 20 or 40 concurrent operations in 30 minutes or 1 hour. To compute this, we take the number of concurrent operations, multiply by the size of data that must be transferred, and then divide by the time. For example, to clone 20 16 GB VMs, we must read up to (20 × 16 GB) = 320 GB and then write up to 320 GB, all in 1 hour, taking 0.18 GBps.

As the table suggests, the types of periodic maintenance operations that occur can impact the storage and bandwidth requirements. If periodic maintenance is a small number of power operations, then a simple 1 Gbps network is more than sufficient for the rack. If, however, a large number of clones must be done in a short amount of time, as in the 40-clone case for 16 GB VMs, then 0.36 GBps is required. If this clone is occurring via NAS, this requires more than 1 Gbps for the management network from that rack.

Table 5: Per-Rack Storage/Network Bandwidth Required for Concurrent Operations during a Maintenance Window.

Operation	**30-minute window**		**1-hour window**	
	# of Concurrent Operations		**# of Concurrent Operations**	
	20	**40**	**20**	**40**
Power on (200MB to boot)	2.2 MBps	4.4 MBps	1.1 MBps	2.2 MBps
Create Snapshot (2GB)	0.022 GBps	0.044 GBps	0.011 GBps	0.022 GBps
Commit Snapshot (2GB)	0.044 GBps	0.088 GBps	0.022 GBps	0.044 GBps
Clone VM (16 GB)	0.360 GBps	0.720 GBps	0.180 GBps	0.360 GBps

4.3.5 Automated Load Balancing

Automated load balancing heavily leverages live migration. Because this is a CPU-intensive operation, the CPU usage of the live-migration agent is typically capped to avoid saturating the CPU and slowing down applications running on the same host. Each live migration must transfer approximately the size of the VM's memory (not counting zero pages) between hosts. Most virtualized deployments utilize an entirely different network and separate NIC per host to accommodate live migration traffic [28].

In the environments we studied, the rate of automated live migration varied quite a bit, from 0 to 80 per hour. In Table 6, we show the intra-rack bandwidth required for varying rates of live migration per hour and varying amounts of data transferred. For example, for 20 live migrations per hour, with each live migration transferring 512 MB, the rack bandwidth required is (20 × 512MB/1 hour) = 23 Mbps.

As the table suggests, the bandwidth demands within the rack are quite modest. However, each host essentially requires a gigabit NIC to be devoted exclusively to live migration traffic.

Table 6: Rack Network Bandwidth Required for Live Migration.

Live migration rate (per hour)	Amount of memory transferred per live migration					
	512 MB	1024 MB	2048 MB	4096 MB	8192 MB	16384 MB
20	23 Mbps	46 Mbps	91 Mbps	182 Mbps	364 Mbps	728 Mbps
40	46 Mbps	91 Mbps	182 Mbps	364 Mbps	728 Mbps	1.5 Gbps
60	68 Mbps	137 Mbps	273 Mbps	546 Mbps	1.1 Gbps	2.2 Gbps
80	91 Mbps	182 Mbps	364 Mbps	728 Mbps	1.5 Gbps	2.9 Gbps

Consider the 2.9 Gbps case. If this rate of live migrations is seen with 10 hosts in a rack, then each host must be able to sustain 290 Mbps. Moreover, if VMs are located across racks for fault-tolerance purposes, the traffic demands will change. For example, if 80 concurrent live migrations of 16 GB VMs must occur at once, and the transfer is from one rack to another, then the rack-level switch must have multiple 1 Gbps uplinks. For traffic purely within the rack, a switch that uses 1 Gbps ports can likely be oversubscribed and still satisfy the bandwidth demands easily.

4.4 Summary

Table 7 summarizes the results of our study, showing the workload, the per-rack bandwidth required, and the resource (storage or network) that is the primary resource used. Of course, if NAS devices are used instead of SAN devices, then the bandwidth tradeoffs shift from storage to networking.

Table 7: Per-Rack Storage and Network Requirements for Management Workloads.

Workflow	Rack Bandwidth Requirement	Resource tradeoff
Periodic Snapshot/revert	0.35 GBps – 2.8 GBps	Storage
Host/VM Patching	0.22 Gbps – 6.8 Gbps	Network
Boot storm	0.14 GBps – 1.1 GBps	Storage
After-hours maintenance	2.2 MBps – 0.18 GBps	Storage/network
Automated Load Balancing	23 Mbps – 2.9 Gbps	Network

It is important to note that the basic mechanisms required to perform each of these workflows is expected to be similar regardless of virtualization platform. For example, live migration requires transferring memory over the network regardless of which hypervisor is used, and reverting a snapshot will require some method of reading memory for the VM from disk. Consequently, we expect that these results are applicable across any virtualized environments, and are not confined to the one used in the environments we profiled.

One natural question to ponder is what will happen as storage and network technology continues to improve. 10 Gbps is becoming the standard for networking and high-end storage arrays can support over 100 GBps in aggregate data bandwidth. Moreover, technologies like Cisco UCS [8] and InfiniBand [14] offer significant network and storage bandwidth through unified storage/network fabrics. In addition, as SSDs [16] continue to improve, they will be used to improve aggregate storage bandwidth. Although the per-host and per-rack bandwidth will improve, compute-density will also increase, and the management workload will scale as a result. Though the absolute numbers for management traffic are not staggering (2.9 Gbps for rack network bandwidth and 1 GBps for rack storage bandwidth), they do induce substantial traffic for critical operations that must fit within maintenance windows. Moreover, there is a direct relationship between cost and capability of components and the speed with which management operations can be completed. The datacenter designer must be aware of this tradeoff.

5. IMPLICATIONS FOR ARCHITECTS

Our study provides two major results for the computer architecture community: the definition of a management workload for virtualized datacenters, and a set of implications for datacenter design. We discuss these in more detail below.

5.1 Definition of the Management Workload

This paper describes a collection of common workflows in the virtualized datacenter and quantifies their impact. We bring attention to the fact that management operations comprise a unique *management workload*. This is a key step towards formulating a general-purpose management workload that architects can use for studying virtualized datacenters.

The virtualization management workload has different characteristics from the individual applications, such as databases and mail servers, running in the datacenter. Furthermore, the virtualized management workload also differs from parallelized high-performance computing applications and internet-scale applications like Google Search.

The virtualization management workload primarily stresses the disk and network subsystems in the datacenter. It is bursty by nature, and has hard latency requirements. Moreover, the bursts of management activity are not controlled by individual administrators, since a given datacenter may host virtual machines from multiple customers in multiple geographies each with different maintenance windows. Management operations are not typically checkpointed (as are large-scale parallel applications). Failures are not tolerated because users could lose all access to their machines, so efficient methods of fault containment, isolation, and recovery are necessary.

5.2 Virtualized Datacenter Design

This characterization of management operations in a virtualized datacenter also yields several implications for computer architects. The datacenter architect must factor in a “management budget” when constructing a datacenter. The choice of appropriate compute elements, racks, storage, networking, and power management is important to ensure good price-performance within a given power budget.

Starting at the processor level, there are some design tradeoffs for architects to consider. Virtualization deployments are benefiting to a great extent from the latest advances in processor design.

Applications running in virtual machines see improved performance with virtualization extensions. Both virtualized applications and management operations benefit from increased core counts (due to their concurrent nature), and larger caches. However, employing the latest processors both costs more money and consumes more power than older-generation processors or processors without virtualization extensions (like mobile processors). In addition, virtualized datacenters are designed to run legacy enterprise workloads, which expect reliable, server-grade hardware and do not tolerate frequent failures.

At the system level, the storage and network requirements for management operations (such as live migration or cloning a virtual machine) suggest a high-speed I/O bus coupled with multiple adapters. For example, multiple network adapters may be required for management operations (like live migration, fault-tolerance, or high-availability) in addition to the networking demands of the applications. A convergent fabric [17] can be a solution as long as the individual storage/network requirements do not overwhelm the convergent adapter on the node or the I/O bus.

Going beyond the individual server, our results show that the datacenter infrastructure and rack architecture are also impacted by management workloads. While there are benefits to maximizing the density of compute servers in a rack, the higher densities result in greater network and storage requirements for the management workload. This design motivates a hierarchical network/storage interconnect with high-bandwidth storage/networking within a rack feeding into global I/O interconnect. Now the scheduling of these management operations becomes important: it may be useful to consider the management required for a given virtual machine when placing it in a particular rack. For example, if a virtual machine is likely to be migrated frequently, then it will benefit from high intra-rack bandwidth rather than high inter-rack bandwidth. The choice of overall storage is important as well: NFS may be less expensive and sufficient for certain workloads, but the additional bandwidth and reduced latency of a SAN may be required for management workloads, since they stress both storage and networking. Increased numbers of low-cost storage elements throughout the datacenter may be more efficient than using fewer numbers of large storage arrays. For situations like patch management, a local patch store per rack is critical, so designing for rack-local storage and high intra-rack bandwidth is key. Solutions like RamClouds [22] can help reduce critical-path bandwidth requirements for networking and storage, but still require sophisticated solutions for persisting data in a timely manner.

Finally, we consider the issue of power management. Various solutions exist for dynamically varying the power delivery in a datacenter [1][23] or for rebalancing workloads to optimize power consumption [29]. Because the management workload potentially operates around the clock, the power budget must take it into account. For example, it is possible that distributing storage and networking throughout the datacenter may be more efficient than consolidating the storage/networking because the power is distributed more uniformly throughout the datacenter. More evaluation is needed before any conclusions can be drawn.

6. RELATED WORK

To the best of our knowledge, this is the first work that focuses on virtualized datacenters and specifically considers the impact of the management workload on virtualized datacenter operations.

While our work was done in the context of enterprise workloads running within a virtualized datacenter, many of our findings can be applied to cloud computing infrastructures such as Amazon EC2 [2] and Terremark [25] Such infrastructures are often virtualized. As multiple customers use a given set of datacenters, each user's management needs comprises a workload by itself, and the collection of management workloads is analogous to a cloud-computing workload.

Our work focuses on the management aspects of large-scale systems, but another important component is monitoring for errors or assessing system-wide performance. One of the challenges of monitoring is collecting large amounts of data and then correlating this data to events in the system, like network overloads or drive failures. Examples of such monitoring and reporting systems include Ganglia [19], Chukwa [7], and the Google System Health Infrastructure [24]. As compute densities increase, efficient monitoring for virtualized datacenters becomes more important. The management operations themselves require monitoring to determine whether they have completed or failed due to some system failure.

Some of our conclusions about diversity of hosts and storage subsystem and the need for adequately provisioning networks may differ from the design principles of warehouse-scale computers (WSCs) used by Google [6], Amazon [2] and Yahoo [35] to run their Internet services. One reason is that WSCs run different workloads from the datacenters we are considering. For example, most WSCs run internet-scale applications that have built-in mechanisms for dealing with hardware failures. In contrast, the datacenters we consider run enterprise applications, such as Exchange or SQL Server, that do not have built-in fault-resiliency. They typically rely on external (to the application) high-availability or fault-tolerance solutions. Such applications expect that the hardware is generally reliable and failures are uncommon. In such cases, more expensive hardware with more robust reliability guarantees is essential, and using cheaper components that can fail with impunity may not be an option. Essentially, the difference is between cloud-scale applications (like those that run in WSCs) vs. applications running in a cloud (like Exchange).

Previous work in cloud computing [5] has focused on similar workloads as WSCs: namely, cloud-scale applications (like MapReduce [10] or Hadoop [4]) vs. applications running in a cloud (like various instances of Microsoft Exchange). These cloud-scale applications have bursty behavior similar to our management workload, so similar techniques for balancing load or provisioning resources may be applicable to the management workload.

7. CONCLUSIONS

The datacenter of the future will make heavy use of virtualized infrastructure. Virtualization allows a more efficient use of hardware through server consolidation and also enables a range of new workflows that can reduce total cost of ownership and improve administrator productivity in the datacenter. While easing the day-to-day tasks of an administrator, however, virtualization also introduces a number of new considerations into the datacenter. In particular, the management operations constitute a workload over and above the actual applications that are running in VMs in the datacenter. Understanding of this management workload is critical for proper datacenter design.

In this paper we have examined data from a number of virtualized deployments and studied the management operations in each of those deployments. We have provided a characterization of the operations into various management workloads and studied their resource usage. We demonstrate five different management workloads present in the virtualized datacenter: snapshot/revert, VM patching, boot storm, after-hours maintenance, and automated live migration. The complexity of this management workload is likely to increase as more value-added virtualization features like high-availability [31], disaster recovery [32], and storage load balancing [12] become even more commonplace. We have found that in general, management operations have significant storage and networking requirements over and above that of most traditional user applications: over 1 GBps additional rack bandwidth required for management storage traffic and almost 3 Gbps additional rack bandwidth required for management network traffic. We find that burstiness is the rule, rather than the exception, in the virtualized management workload, and planning for this burstiness is essential for efficient datacenter operations: in particular, phasing management operations and choosing appropriate hardware can reduce the length of maintenance cycles from hours to minutes and minimize perturbations to the application load in the system. Finally, we show that the management workload scales as the computation power of servers increases and becomes even more critical with the increasing proliferation of massively multi-core processors.

8. ACKNOWLEDGMENTS

Our thanks to Jeremy Rodriguez, Devaki Kulkarni, Sunil Satnur, John Liang, Xuwen Yu, Sam McBride, Warakorn Thapanangkun, Bruce Herndon, Shicong Meng, Carl Waldspurger, Ben Verghese, Dan Scales, and Julie Brodeur for helpful comments on previous drafts of this paper. We would also like to thank our customers for graciously letting us use data collected from their environments.

9. REFERENCES

[1] Ahmad, F., et al. Joint Optimization of Idle and Cooling Power in Data Centers While Maintaining Response Time. In Proceedings of ASPLOS XV (Pittsburgh, PA, March 13-17, 2010). 243-256.

[2] Amazon EC2. http://aws.amazon.com/ec2.

[3] AMD Magny-Cours. http://www.amd.com/us/products/server/processors/6000-series-platforms/Pages/6000-series-platform.aspx

[4] Apache Hadoop. http://hadoop.apache.org/

[5] Armbrust, M., et al. Above the Clouds: A Berkeley View of Cloud Computing. University of California at Berkeley Technical Report No. UCB/EECS-2009-28.

[6] Barroso, L., et al. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. Morgan and Claypool Publishers. 2009.

[7] Boulon, J., et al. Chukwa, a large-scale monitoring system. In Proceedings of CCA '08. Oct 2008

[8] Cisco. Unified Computing System. http://www.cisco.com/en/US/netsol/ns944/#~in_depth

[9] Dean, J. Designs, Lessons, and Advice from Building Large Distributed Systems. Keynote from LADIS 2009 (Big Sky, Montana, October 10-11, 2009).

[10] Dean, J., et al. MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, vol. 51, no. 1 (2008), pp. 107-113.

[11] EMC. EMC Symmetrix DMX-3 950 specifications. http://www.emc.com/collateral/hardware/specification-sheet/c1153-dmx3-950-ss.pdf

[12] Gulati, A., et al. BASIL: Automated IO Load Balancing across Storage Devices. In Proceedings of USENIX FAST ’10 (San Jose, CA, February 23-26, 2010). 169-182.

[13] HP. Lights-out management. http://h18000.www1.hp.com/products/servers/management/ilo/

[14] InfiniBand Trade Association. http://www.infinibandta.org/

[15] Intel Nehalem. http://www.intel.com/p/en_US/products/server/processor/xeon7000

[16] Intel. Solid State Drives. http://www.intel.com/design/flash/nand/mainstream/index.htm

[17] Leigh, K., et al, Fabric Convergence Implications on Systems Architecture. In Proceedings of HPCA ’08 (Salt Lake City, UT, February 16-20, 2008). 15-26.

[18] Lim, K., et al, Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments. In Proceedings of ISCA 2008 (Beijing, China, June 21-25, 2008). 315-236.

[19] Massie, M.,et al. The Ganglia Distributed Monitoring System: Design, Implementation, and Experience. In Parallel Computing Volume 30, Issue 7, pp 817- 840, 2004.

[20] NAS Devices. http://www.nasdevices.com

[21] Nelson, M., et al. Fast Transparent Migration for Virtual Machines. In Proceedings of USENIX ’05 (Anaheim, CA, April 10-15, 2005). 391-394.

[22] Ousterhout, J., et al. The Case for RamClouds: Scalable High-Performance Storage Entirely in DRAM. In SIGOPS OSR. December, 2009.

[23] Pelley, S., et al. Power Routing: Dynamic Power Provisioning in the Data Center. In Proceedings of ASPLOS XV (Pittsburgh, PA, March 13-17, 2010). 231-242.

[24] Pinheiro, E., et al. Failure Trends in a Large Disk Drive Population. In Proceedings of FAST ’07 (San Jose, CA, February 13-16, 2007). 17-28.

[25] Terremark. http://vcloudexpress.terremark.com/

[26] VMware. http://www.vmware.com

[27] VMware. Capacity Planner. http://www.vmware.com/products/capacity-planner/

[28] VMware. Communities Forum. http://communities.vmware.com/thread/67695

[29] VMware. VMware Distributed Power Management Concepts and Use. http://www.vmware.com/files/pdf/DPM.pdf

[30] VMware. VMware DRS. http://www.vmware.com/products/drs/

[31] VMware. VMware High-Availability. http://www.vmware.com/products/high-availability

[32] VMware. VMware Site Recovery Manager. http://www.vmware.com/products/site-recovery-manager

[33] VMware. VirtualCenter 1.0 Support. http://www.vmware.com/support/vc/doc/c11clonevm.html

[34] VMware. VMware Update Manager Performance and Best Practices. http://www.vmware.com/pdf/vum_1.0_performance.pdf

[35] Yahoo! http://www.yahoo.com