7. Explain about VIRTUAL CLUSTERS

A physical cluster is a collection of servers (physical machines) interconnected by a physical network such as a LAN.

When a traditional VM is initialized, the administrator needs to manually write configuration information or specify the configuration sources.

When more VMs join a network, an inefficient configuration always causes problems with overloading or underutilization.

Amazon's Elastic Compute Cloud (EC2) is a good example of a web service that provides elastic computing power in a cloud. EC2 permits customers to create VMs and to manage user accounts over the time of their use.

Most virtualization platforms, including XenServer and VMware ESX Server, support a bridging mode which allows all domains to appear on the network as individual hosts.

By using this mode, VMs can communicate with one another freely through the virtual network interface card and configure the network automatically.

Physical versus Virtual Clusters

Virtual clusters are built with VMs installed at distributed servers from one or more physical clusters.

The VMs in a virtual cluster are interconnected logically by a virtual network across several physical networks.

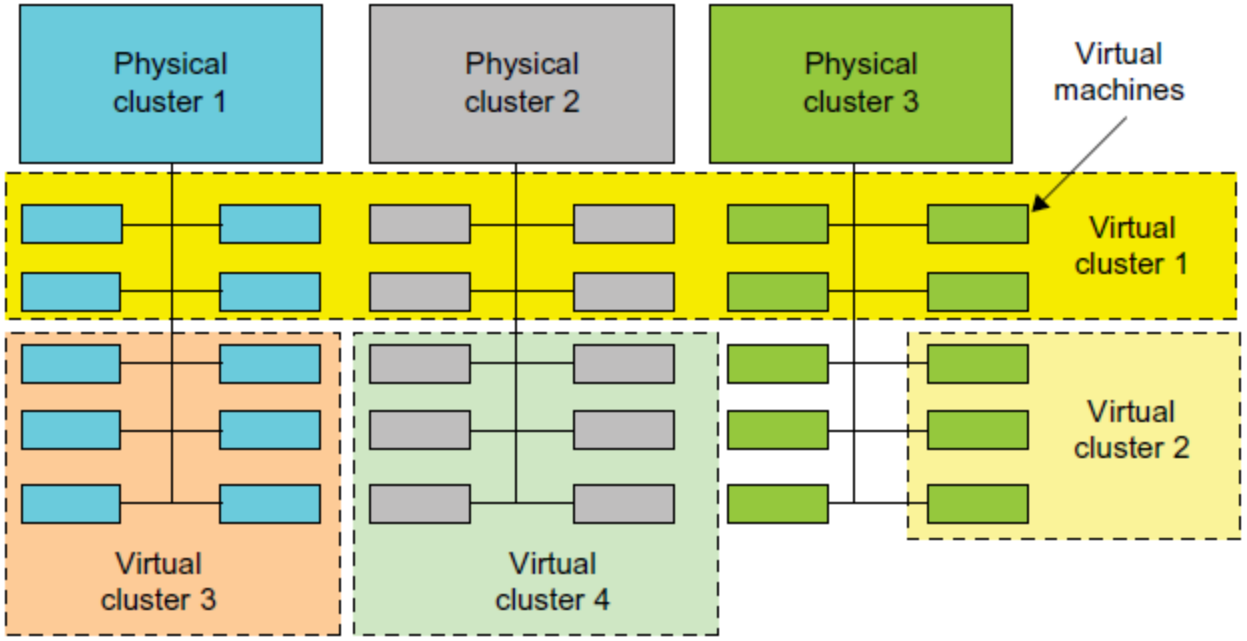


FIGURE 3.18

A cloud platform with four virtual clusters over three physical clusters shaded differently.

The provisioning of VMs to a virtual cluster is done dynamically to have the following interesting properties:

• The virtual cluster nodes can be either physical or virtual machines. Multiple VMs running with different OSes can be deployed on the same physical node.

• A VM runs with a guest OS, which is often different from the host OS, that manages the resources in the physical machine, where the VM is implemented.

• The purpose of using VMs is to consolidate multiple functionalities on the same server. This will greatly enhance server utilization and application flexibility.

VMs can be colonized (replicated) in multiple servers for the purpose of promoting distributed parallelism, fault tolerance, and disaster recovery.

• The size (number of nodes) of a virtual cluster can grow or shrink dynamically, similar to the way an overlay network varies in size in a peer-to-peer (P2P) network.

• The failure of any physical nodes may disable some VMs installed on the failing nodes. But the failure of VMs will not pull down the host system.

Virtual cluster are created based on **application partitioning or customization.**

There are common installations for most users or applications, such as operating systems or user-level programming libraries.

These software packages can be preinstalled as templates (called template VMs). With these templates, users can build their own software stacks.

New OS instances can be copied from the template VM. User-specific components such as programming libraries and applications can be installed to those instances.

As a large number of VM images might be present, the most important thing is to determine how to store those images in the system efficiently.

Each VM can be installed on a remote server or replicated on multiple servers belonging to the same or different physical clusters.

The boundary of a virtual cluster can change as VM nodes are added, removed, or migrated dynamically over time.

Since system virtualization has been widely used, it is necessary to effectively manage VMs running on a mass of physical computing nodes (also called virtual clusters) and consequently build a high-performance virtualized computing environment.

This involves virtual cluster deployment, monitoring and management over large-scale clusters, as well as resource scheduling, load balancing, server consolidation, fault tolerance, and other techniques. The different node colors in Figure 3.18 refer to different virtual clusters.

In a virtual cluster system, it is quite important to store the large number of VM images efficiently.

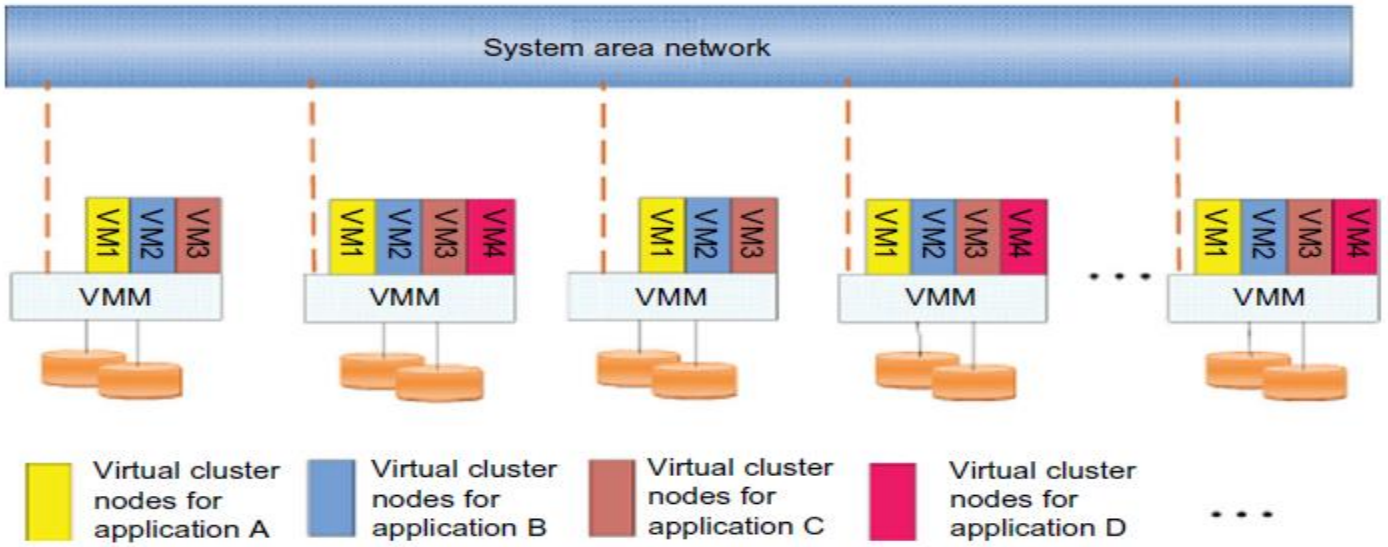


FIGURE 3.19

The concept of a virtual cluster based on application partitioning.

Fast Deployment and Effective Scheduling in clusters

The system should have the capability of fast deployment. Here, deployment means two things:

to construct and distribute software stacks (OS, libraries, applications) to a physical node inside clusters as fast as possible, and to quickly switch runtime environments from one user's virtual cluster to another user's virtual cluster.

If one user finishes using his system, the corresponding virtual cluster should shut down or suspend quickly to save the resources to run other VMs for other users.

The live migration of VMs allows workloads of one node to transfer to another node. However, it does not guarantee that VMs can randomly migrate among themselves.

In fact, the potential overhead caused by live migrations of VMs may have serious negative effects on cluster utilization, throughput, and QoS issues.

Therefore, the challenge is to determine how to design migration strategies to implement green computing without influencing the performance of clusters.

Another advantage of virtualization is load balancing of applications in a virtual cluster. Load balancing can be achieved using the load index and frequency of user logins.

The automatic scale-up and scale-down mechanism of a virtual cluster can be implemented based on this model.

Consequently, we can increase the resource utilization of nodes and shorten the response time of systems.

Mapping VMs onto the most appropriate physical node should promote performance.

Dynamically adjusting loads among nodes by live migration of VMs is desired, when the loads on cluster nodes become quite unbalanced.

High-Performance Virtual Storage in clusters

The template VM can be distributed to several physical hosts in the cluster to customize the VMs. In addition, existing software packages reduce the time for customization as well as switching virtual environments.

- It is important to efficiently manage the disk spaces occupied by template software packages.

- Some storage architecture design can be applied to reduce duplicated blocks in a distributed file system of virtual clusters.

- Hash values are used to compare the contents of data blocks. Users have their own profiles which store the identification of the data blocks for corresponding VMs in a user-specific virtual cluster.

- New blocks are created when users modify the corresponding data.
- Newly created blocks are identified in the users' profiles.

What are the various ways to manage a virtual cluster.

- First, you can use a guest-based manager, by which the cluster manager resides on a guest system. In this case, multiple VMs form a virtual cluster. For example, openMosix is an open source Linux cluster running different guest systems on top of the Xen hypervisor. Another example is Sun's cluster Oasis, an experimental Solaris cluster of VMs supported by a VMware VMM.

- Second, you can build a cluster manager on the host systems. The host-based manager supervises the guest systems and can restart the guest system on another physical machine. A good example is the VMware HA system that can restart a guest system after failure.

- These two cluster management systems are either guest-only or host-only, but they do not mix

- A third way to manage a virtual cluster is to use an independent cluster manager on both the host and guest systems. This will make infrastructure management more complex.

- Finally, you can use an integrated cluster on the guest and host systems. This means the manager must be designed to distinguish between virtualized resources and physical resources.

- Basically, there are four steps to deploy a group of VMs onto a target cluster: preparing the disk image, configuring the VMs, choosing the destination nodes, and executing the VM deployment command on every host.

- Many systems use templates to simplify the disk image preparation process. A template is a disk image that includes a preinstalled operating system with or without certain application software.

- Users choose a proper template according to their requirements and make a duplicate of it as their own disk image.

- Templates could implement the COW (Copy on Write) format. A new COW backup file is very small and easy to create and transfer.

- Therefore, it definitely reduces disk space consumption. In addition, VM deployment time is much shorter than that of copying the whole raw image file.

- The deployment principle is to fulfill the VM requirement and to balance workloads among the whole host network

8 .Live VM Migration Steps and Performance Effects

In a cluster built with mixed nodes of host and guest systems, the normal method of operation is to run everything on the physical machine.

When a VM fails, its role could be replaced by another VM on a different node, as long as they both run with the same guest OS. In other words, a physical node can fail over to a VM on another host.

This is different from physical-to-physical failover in a traditional physical cluster.

The advantage is enhanced failover flexibility. The potential drawback is that a VM must stop playing its role if its residing host node fails. However, this problem can be mitigated with VM life migration.

The process of life migration of a VM from host A to host B. The migration copies the VM state file from the storage area to the host machine.

Live migration means moving a VM from one physical node to another while keeping its OS environment and applications unbroken.

- This capability is being increasingly utilized in today's enterprise environments to provide efficient online system maintenance, reconfiguration, load balancing, and proactive fault tolerance.

- It provides desirable features to satisfy requirements for computing resources in modern computing systems, including server consolidation, performance isolation, and ease of management.

- Traditional migration suspends VMs before the transportation and then resumes them at the end of the process.

- By importing the precopy mechanism, a VM could be live migrated without stopping the VM and keep the applications running during the migration

- Live migration is a key feature of system virtualization technologies.

- Here, we focus on VM migration within a cluster environment where a network-accessible storage system, such as storage area network (SAN) or network attached storage (NAS), is employed.

- Only memory and CPU status needs to be transferred from the source node to the target node.
- Live migration techniques mainly use the precopy approach, which first transfers all memory pages, and then only copies modified pages during the last round iteratively.
- The VM service downtime is expected to be minimal by using iterative copy operations.
- When applications' writable working set becomes small, the VM is suspended and only the CPU state and dirty pages in the last round are sent out to the destination.

VMs can be live-migrated from one physical machine to another; in case of failure, one VM can be replaced by another VM. Virtual clusters can be applied in computational grids, cloud platforms, and high-performance computing (HPC) systems.

The major attraction of this scenario is that virtual clustering provides dynamic resources that can be quickly put together upon user demand or after a node failure.

Motivation is to design a live VM migration scheme with negligible downtime, the lowest network bandwidth consumption possible, and a reasonable total migration time.

Furthermore, we should ensure that the migration will not disrupt other active services residing in the same host through resource contention (e.g., CPU, network bandwidth).

A VM can be in one of the following four states.

- An inactive state is defined by the virtualization platform, under which the VM is not enabled.
- An active state refers to a VM that has been instantiated at the virtualization platform to perform a real task.
- A paused state corresponds to a VM that has been instantiated but disabled to process a task or paused in a waiting state.
- A VM enters the suspended state if its machine file and virtual resources are stored back to the disk.

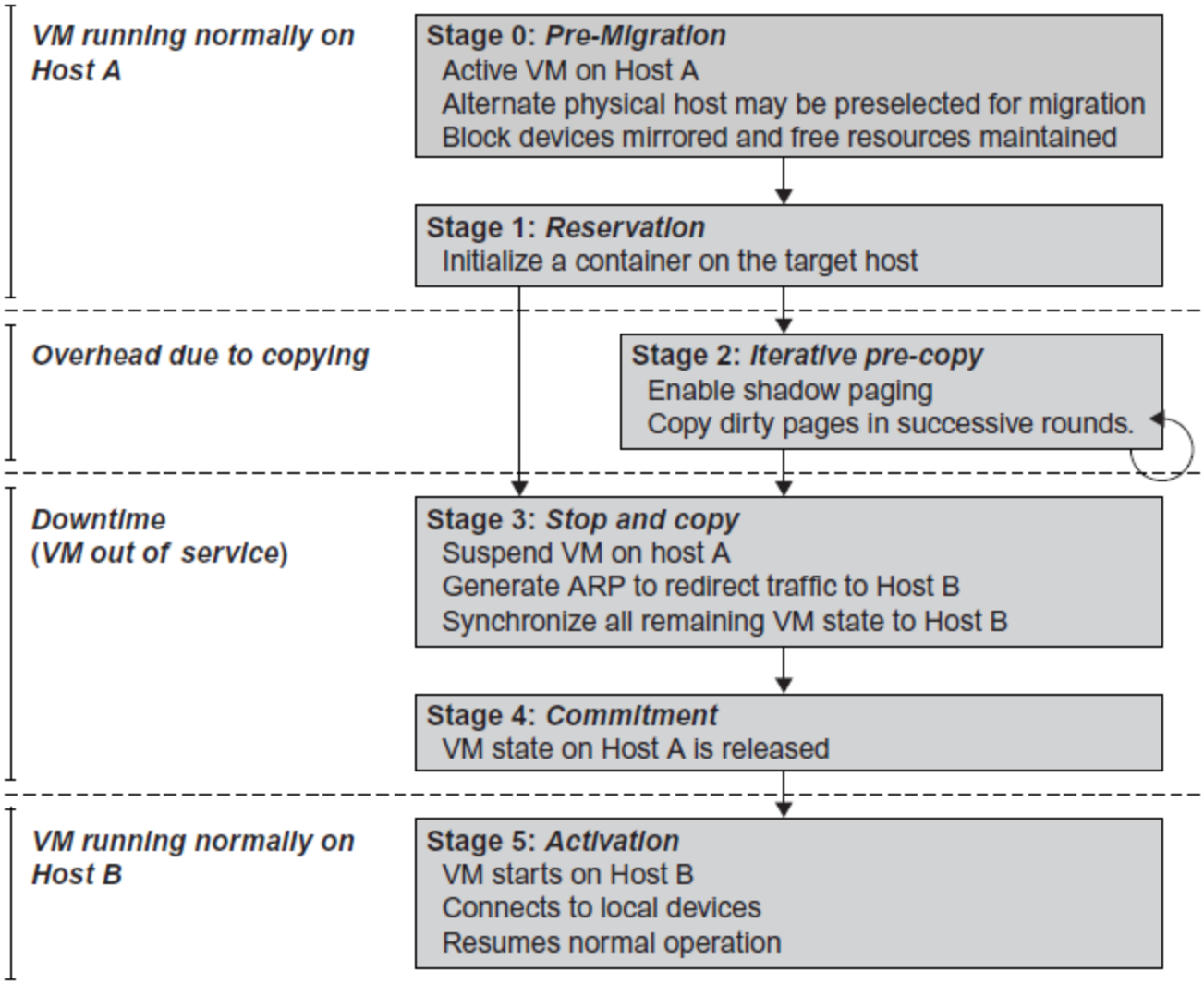


FIGURE 3.20

Live migration process of a VM from one host to another.

Steps 0 and 1: Start migration. This step makes preparations for the migration, including determining the migrating VM and the destination host. Although users could manually make a VM migrate to an appointed host, in most circumstances, the migration is automatically started by strategies such as load balancing and server consolidation.

Steps 2: Transfer memory. Since the whole execution state of the VM is stored in memory, sending the VM's memory to the destination node ensures continuity of the service provided by the VM.

All of the memory data is transferred in the first round, and then the migration controller recopies the memory data which is changed in the last round. These steps keep iterating until the dirty portion of the memory is small enough to handle the final copy. Although pre-copying memory is performed iteratively, the execution of programs is not obviously interrupted.

Step 3: Suspend the VM and copy the last portion of the data. The migrating VM's execution is suspended when the last round's memory data is transferred. Other non memory data such as CPU and network states should be sent as well. During this step, the VM is stopped and its applications will no longer run. This "service unavailable" time is called the "downtime" of migration, which should be as short as possible so that it can be negligible to users.

Steps 4 and 5: Commit and activate the new host. After all the needed data is copied, on the destination host, the VM reloads the states and recovers the execution of programs in it, and the service provided by this VM continues.

Then the network connection is redirected to the new VM and the dependency to the source host is cleared. The whole migration process finishes by removing the original VM from the source host.

Memory Migration

Moving the memory instance of a VM from one physical host to another can be approached in any number of ways.

- Memory migration can be in a range of hundreds of megabytes to a few gigabytes in a typical system today, and it needs to be done in an efficient manner.

- The Internet Suspend-Resume (ISR) technique exploits temporal locality as memory states are likely to have considerable overlap in the suspended and the resumed instances of a VM.

- Temporal locality refers to the fact that the memory states differ only by the amount of work done since a VM was last suspended before being initiated for migration.

- To exploit temporal locality, each file in the file system is represented as a tree of small subfiles.

- A copy of this tree exists in both the suspended and resumed VM instances. The advantage of using a tree-based representation of files is that the caching ensures the transmission of only those files which have been changed

File System Migration

To support VM migration, a system must provide each VM with a consistent, location-independent view of the file system that is available on all hosts.

A simple way to achieve this is to provide each VM with its own virtual disk which the file system is mapped to and transport the contents of this virtual disk along with the other states of the VM.

However, due to the current trend of high capacity disks, migration of the contents of an entire disk over a network is not a viable solution.

Another way is to have a global file system across all machines where a VM could be located.

This way removes the need to copy files from one machine to another because all files are network accessible.

In smart copying, the VMM exploits spatial locality. Typically, people often move between the same small number of locations, such as their home and office.

In these conditions, it is possible to transmit only the difference between the two file systems at suspending and resuming locations.

This technique significantly reduces the amount of actual physical data that has to be moved.

Network Migration

- A migrating VM should maintain all open network connections without relying on forwarding mechanisms on the original host or on support from mobility or redirection mechanisms.

- To enable remote systems to locate and communicate with a VM, each VM must be assigned a virtual IP address known to other entities. This address can be distinct from the IP address of the host machine where the VM is currently located.

- Each VM can also have its own distinct virtual MAC address.
- The VMM maintains a mapping of the virtual IP and MAC addresses to their corresponding VMs.

- In general, a migrating VM includes all the protocol states and carries its IP address with it.

- If the source and destination machines of a VM migration are typically connected to a single switched LAN, an unsolicited ARP reply from the migrating host is provided advertising that the IP has moved to a new location.

- This solves the open network connection problem by reconfiguring all the peers to send future packets to a new location

9.VIRTUALIZATION FOR DATA-CENTER AUTOMATION

Data centers have grown rapidly in recent years, and all major IT companies are pouring their resources into building new data centers.

- Data Centers are centralized repositories of information.
- These include server farms and networking equipment that stores, processes, and distributes huge volumes of data for clients.
- Data centers can offer services like data warehousing, data insights, data storage, etc.

Company	Headquarters	Founded In	# of Data Centers	Markets Served	Services
Equinix	Redwood City, CA, US	1998	202 (12 more to come)	24 countries	5
Digital Realty	San Francisco, CA, US	2004	214	14 countries	3
China Telecom	Beijing, China	2002	456	>10 countries	6
NTT Communications	Tokyo, Japan	1999	48	17 countries	9
Telehouse/KDDI	London, UK /Tokyo, Japan	1988/1953	40	12 countries	4

Data-center automation means that huge volumes of hardware, software, and database resources in these data centers can be allocated dynamically to millions of Internet users simultaneously, with guaranteed QoS and cost-effectiveness

- Virtualization is moving towards enhancing mobility, reducing planned downtime (for maintenance), and increasing the number of virtual clients, high availability (HA), backup services, workload balancing, and further increases in client bases.

In data centers, a large number of heterogeneous workloads can run on servers at various times. These workloads can be roughly divided into two categories: chatty workloads and non-interactive workloads.

Chatty workloads may burst at some point and return to a silent state at some other point. A web video service is an example of this, whereby a lot of people use it at night and few people use it during the day.

Non-interactive workloads do not require people's efforts to make progress after they are submitted. High-performance computing is a typical example of this. At various stages, the requirements for resources of these workloads are dramatically different. However, to guarantee that a workload will always be able to cope with all demand levels, the workload is statically allocated enough resources so that peak demand is satisfied.

- However, to guarantee that a workload will always be able to cope with all demand levels, the workload is statically allocated enough resources so that peak demand is satisfied.
- it is common that most servers in data centers are underutilized.
- A large amount of hardware, space, power, and management cost of these servers is wasted.

Server Consolidation in Data Centers

- Server consolidation is an approach to improve the low utility ratio of hardware resources by reducing the number of physical servers.

Among several server consolidation techniques such as
- centralized and physical consolidation,
- virtualization-based server consolidation is the most powerful.

- Data centers need to optimize their resource management.

- In general, the use of VMs increases resource management complexity.

- This causes a challenge in terms of how to improve resource utilization as well as guarantee QoS in data centers.

Server virtualization has the following side effects:

- Consolidation enhances hardware utilization. Many underutilized servers are consolidated into fewer servers to enhance resource utilization. Consolidation also facilitates backup services and disaster recovery.
- This approach enables more agile provisioning and deployment of resources. In a virtual environment, the images of the guest OSes and their applications are readily cloned and reused.
- The total cost of ownership is reduced. In this sense, server virtualization causes deferred purchases of new servers, a smaller data-center footprint, lower maintenance costs, and lower power, cooling, and cabling requirements.
- This approach improves availability and business continuity. The crash of a guest OS has no effect on the host OS or any other guest OS. It becomes easier to transfer a VM from one server to another, because virtual servers are unaware of the underlying hardware.

- To automate data-center operations, one must consider resource scheduling, architectural support, power management, automatic or autonomic resource management, performance of analytical models, and so on.

- In virtualized data centers, an efficient, on-demand, fine-grained scheduler is one of the key factors to improve resource utilization.

- Scheduling and reallocations can be done in a wide range of levels in a set of data centers. The levels match at least at the VM level, server level, and data-center level.

- Ideally, scheduling and resource reallocations should be done at all levels.

- Dynamic CPU allocation is based on VM utilization and application-level QoS metrics. One method considers both CPU and memory flowing as well as automatically adjusting resource overhead based on varying workloads in hosted services.

- Another scheme uses a two-level resource management system to handle the complexity involved. A local controller at the VM level and a global controller at the server level are designed.
- They implement autonomic resource allocation via the interaction of the local and global controllers. Multicore and virtualization are two cutting techniques
- that can enhance each other.

- One can also consider a VM-aware power budgeting scheme using multiple managers integrated to achieve better power management.
- The power budgeting policies cannot ignore the heterogeneity problems. Consequently, one must address the trade-off of power saving and data-center performance.

Virtual Storage Management

- In system virtualization, virtual storage includes the storage managed by VMMs and guest OSes.

- Generally, the data stored in this environment can be classified into two categories: VM images and application data.

- The VM images are special to the virtual environment, while application data includes all other data which is the same as the data in traditional OS environments

The most important aspects of system virtualization are encapsulation and isolation

- Traditional operating systems and applications running on them can be encapsulated in VMs.
- System virtualization allows multiple VMs to run on a physical machine and the VMs are completely isolated.

- To achieve encapsulation and isolation, both the system software and the hardware platform, such as CPUs and chipsets, are rapidly updated.

- However, storage is lagging. The storage systems become the main bottleneck of VM deployment.

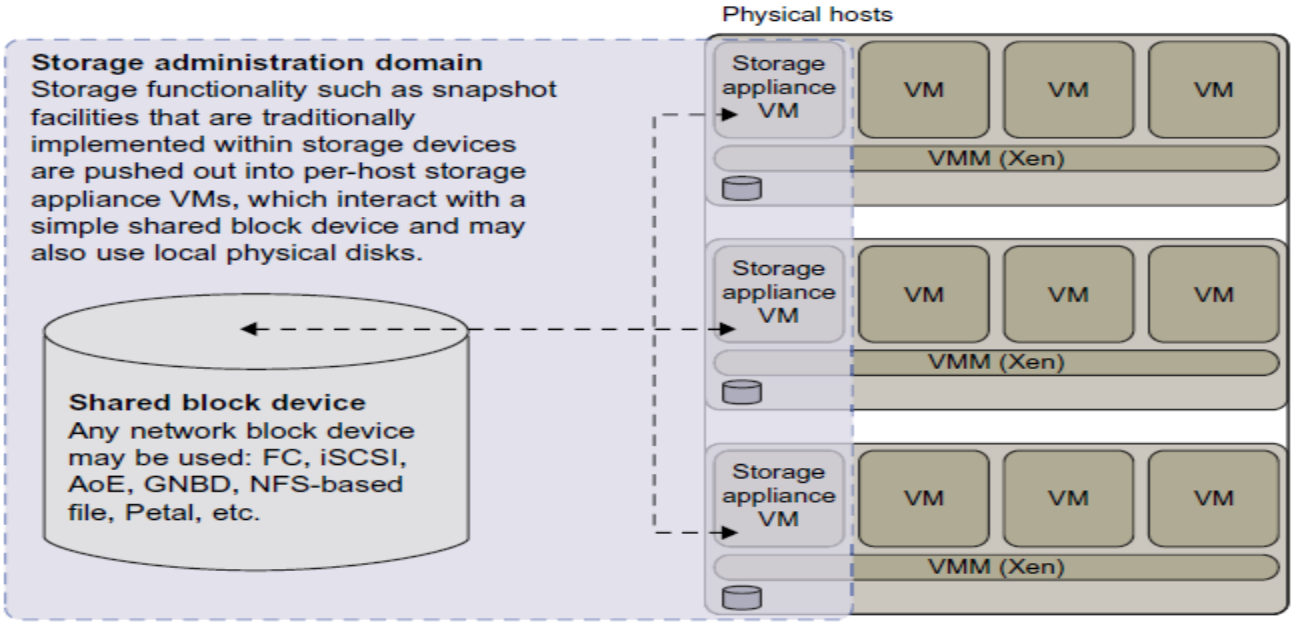
  - On the one hand, storage management of the guest OS performs as though it is operating in a real hard disk while the guest OSes cannot access the hard disk directly.

  - On the other hand, many guest OSes contest the hard disk when many VMs are running on a single physical machine.

  - Therefore, storage management of the underlying VMM is much more complex than that of guest OSes (traditional OSes). The problem is more at Data Centres where hundreds of VM are running on the servers.

  - Parallax is a distributed storage system customized for virtualization environments. Content Addressable Storage (CAS) is a solution to reduce the total size of VM images, and therefore supports a large set of VM-based systems in data centers.

- Parallax designs a novel architecture in which storage features that have traditionally been implemented directly on high-end storage arrays and switchers are relocated into a federation of storage VMs.
- These storage VMs share the same physical hosts as the VMs that they serve. Figure provides an overview of the Parallax system architecture.
- It supports all popular system virtualization techniques, such as paravirtualization and full virtualization.
- For each physical machine, Parallax customizes a special storage appliance VM. The storage appliance VM acts as a block virtualization layer between individual VMs and the physical storage device. It provides a virtual disk for each VM on the same physical machine



Cloud OS for Virtualized Data Centers

Data centers must be virtualized to serve as cloud providers.

- There are four virtual infrastructure (VI) managers and OSes.
- These VI managers and OSes are specially tailored for virtualizing data centers which often own a large number of servers in clusters.

- Nimbus, Eucalyptus, and OpenNebula are all open source software available to the general public.
- Only vSphere 4 is a proprietary OS for cloud resource virtualization and management over data centers.

- These VI managers are used to create VMs and aggregate them into virtual clusters as elastic resources.
- Nimbus and Eucalyptus support essentially virtual networks. OpenNebula has additional features to provision dynamic resources and make advance reservations. All three public VI managers apply Xen and KVM for virtualization.

- vSphere 4 uses the hypervisors ESX and ESXi from VMware. Only vSphere 4 supports virtual storage in addition to virtual networking and data protection.

Trust Management in Virtualized Data Centers

A VMM provides a layer of software between the operating systems and system hardware to create one or more VMs on a single physical platform.

- A VM entirely encapsulates the state of the guest operating system running inside it. Encapsulated machine state can be copied and shared over the network and removed like a normal file, which proposes a challenge to VM security.

- In general, a VMM can provide secure isolation and a VM accesses hardware resources through the control of the VMM, so the VMM is the base of the security of a virtual system.
- Normally, one VM is taken as a management VM to have some privileges such as creating, suspending, resuming, or deleting a VM.

VM-Based Intrusion Detection

- Intrusions are unauthorized access to a certain computer from local or network users and intrusion detection is used to recognize the unauthorized access.
- An intrusion detection system (IDS) is built on operating systems, and is based on the characteristics of intrusion actions. A typical IDS can be classified as a host-based IDS (HIDS) or a network-based IDS (NIDS), depending on the data source.

- A HIDS can be implemented on the monitored system. When the monitored system is attacked by hackers, the HIDS also faces the risk of being attacked.
- A NIDS is based on the flow of network traffic which can't detect fake actions.

- Virtualization-based intrusion detection can isolate guest VMs on the same hardware platform. Even some VMs can be invaded successfully; they never influence other VMs, which is similar to the way in which a NIDS operates. Furthermore, a VMM monitors and audits access requests for hardware and system software
- The VM-based IDS contains a policy engine and a policy module. The policy framework can monitor events in different guest VMs by operating system interface library and PTrace indicates trace to secure policy of monitored host. It's difficult to predict and prevent all intrusions without delay. Therefore, an analysis of the intrusion action is extremely important after an intrusion occurs.