CS 524 Homework #5

1. (5 points) Explain the motivation behind the two forms of server placement (rack-mounted servers and blade servers). What is sacrificed to make a blade server more compact than a rackmounted server?

Solution:

Rack Server:

- A rack server, also called a rack-mounted server, is a computer dedicated to use as a server and designed to be installed in a framework called a rack.
- The rack contains multiple mounting slots called bays, each designed to hold a hardware unit secured in place with screws. A rack server has a low-profile enclosure, in contrast to a tower server, which is built into an upright, standalone cabinet.
- A single rack can contain multiple servers stacked one above the other, consolidating network resources and minimizing the required floor space.
- The rack server configuration also simplifies cabling among network components. In an equipment rack filled with servers, a special cooling system is necessary to prevent excessive heat buildup that would otherwise occur when many power-dissipating components are confined in a small space.

Blade Server:

- A blade server is a modular server that allows multiple servers to be housed in a smaller area. These servers are physically thin and typically only have CPUs, memory, integrated network controllers, and sometimes storage drives built in. Any video cards or other components that are needed will be facilitated by the server chassis. Which is where the blades slide into. Blade servers are often seen in large data centers. Due to their ability to fit so many servers into one single rack and their ability to provide a high processing power.
- In most cases, one large chassis such as HPE's BladeSystem will be mounted into a server rack and then multiple blade servers slide into the chassis. The chassis can then provide the power, manage networking, and more. This allows each blade server to operate more efficiently and requires fewer internal components.
- Blade servers are generally used when there is a high computing requirement with some type of Enterprise Storage System: Network Attached Storage (NAS) or a Storage Area Network (SAN). They maximize available space by providing the highest processor per RU availability. Blade Servers also provide rapid serviceability by allowing components to be swapped out without taking the machine offline. You will be able to scale to a much higher processor density using the Blade architecture. The facility will need to support a much higher thermal and electrical load per square foot.

Blade servers have many components removed to save space, minimize power consumption and other considerations, while still having all the functional components to be considered a computer. Unlike a rack-mount server, a blade server fits inside a blade enclosure, which can hold multiple blade servers, providing services such as power, cooling, networking, various interconnects and management. Together, blades and the blade enclosure form a blade system, which may itself be rack-mounted. Different blade providers have differing principles regarding what to include in the blade itself, and in the blade system as a whole.

Source:[https://whatis.techtarget.com/definition/rack-server-rack-mounted-server] [https://www.racksolutions.com/news/data-center-optimization/blade-server-vs-rack-server/] [https://en.wikipedia.org/wiki/Blade_server]

2. (5 points) Why is the use of the Ethernet technology particularly important to the data centers? [Hint: What need does the use of the Ethernet effectively eliminate?]

Solution:

- ✓ Ethernet is the primary network protocol in data centers for computer-to-computer communications. However, Ethernet is designed to be a best-effort network that may experience packet loss when the network or devices are busy.
- ✓ In IP networks, transport reliability under the end-to-end principle is the responsibility of the transport protocols, such as the Transmission Control Protocol (TCP). One area of evolution for Ethernet is to add extensions to the existing protocol suite to provide reliability without requiring the complexity of TCP.
- ✓ With the move to 10 Gbit/s and faster transmission rates, there is also a desire for finer granularity in control of bandwidth allocation and to ensure it is used more effectively. These enhancements are particularly important to make Ethernet a more viable transport for storage and server cluster traffic.
- ✓ A primary motivation is the sensitivity of Fibre Channel over Ethernet to frame loss. The higher level goal is to use a single set of Ethernet physical devices or adapters for computers to talk to a Storage Area Network, Local Area network and InfiniBand fabric.

Source: [https://en.wikipedia.org/wiki/Data_center_bridging]

3. (5 points) Explain why NAS and SAN but not DAS are readily applicable to Cloud Computing. What are the limitations of DAS? Why is DAS suitable for keeping local data (such as boot image or swap space)?

Solution:

Network Attached Storage (NAS): It is a special purpose device. It comprises hard disks, as well as management software. NAS is dedicated 100% to serve files over a network. In simple terms, NAS shares storage over a network. Once connected, you will come across

special folders named ‘Shares’ that can be accessed over the network. Multiple user logins can also be created to provide various levels of access. NAS is ideal for SMBs because it allows a cost-effective way to gain data access for multiple clients quickly at the file level. SMBs can gain from its performance, and increase their productivity. Few other advantages include easy setup and configuration compared to SAN, maximum storage resources utilization, and the easy-to-provide RAID redundancy to a large number of users.

Storage Area Networks (SAN): A SAN is a high-performance, dedicated storage network. It transfers data between storage devices and servers. It functions separately from LAN. In the SAN infrastructure, fiber channel is used to connect devices such as RAID arrays, DAS or tape libraries to servers. The main advantage of SAN is its ability to transfer large data blocks. This is very useful for bandwidth-intensive applications such as imaging, database (cloud computing, virtual environments), and transaction processing. Also, SAN offers complete reliability and 24/7 availability of data.

Limitations of DAS/Why DAS is not suitable for Cloud Computing: In DAS, the storage device is directly attached to the computer. For example, a USB-connected external hard drive. A number of cables can be used to connect DAS units, such as fiber optics, SAS, SATA, and so on. A DAS is designed to be used only by a single computer, unlike NAS and SAN that are designed to be shared resources. The main advantages of DAS is that it is high-performance, simpler to setup and configure, and typically lower-cost when compared to SAN storage. The disadvantage is that it cannot be managed over a network, and may not have the same level of redundancy as a NAS or SAN.

Source: [https://dtechconsulting.com/nas-vs-das-san/]

4. (5 points) Why is there a need for the Phy layer in the SAS architecture? How is it different from the physical layer?

Solution: SAS physical links (phys) are a set of four wires used as two differential signal pairs. One differential signal transmits in one direction, while the other differential signal transmits in the opposite direction. Data can be transmitted in both directions simultaneously. Phys are contained in SAS ports which contain one or more phys. A port is a wide port if there are more than one phy in the port. If there is only one phy in the port, it is a narrow port. A port is identified by a unique SAS worldwide name (also called SAS address).

Source: [https://www.ibm.com/support/knowledgecenter/POWER6/arebj/sasoverview.htm]

5. (10 points) List the generic file-related system calls. Why in the NFS there is no RPC invocation for the close system call? Under which circumstances other file operations may not result in an RPC invocation?

Solution:

- There are 5 different categories of system calls: process control, file manipulation, device manipulation, information maintenance, and communication. open(), write(), read(), poll(), close()
- Remote Procedure Call (RPC) is a powerful technique for constructing distributed, client-server based applications. It is based on extending the conventional local procedure calling so that the called procedure need not exist in the same address space as the calling procedure. The two processes may be on the same system, or they may be on different systems with a network connecting them.
- NFS close system call does not invocate RPC because there is no modification of file in this case and the original stateless design of servers does not keep track of past recovery.
- If a file operation is performed remotely, it may not result in RPC invocation.

Source: [http://faculty.salina.k-state.edu/tim/ossg/Introduction/sys_calls.html] [http://www.iitk.ac.in/LDP/HOWTO/SCSI-Generic-HOWTO/syscalls.html]

6. (10 points) What types of connection topologies are supported in FC-2M? Which of them is the most flexible? Why?

Solution: Three types of connection topologies are supported in FC-2M: Point-to-point, Fabric and Arbitrated loop.

- The point-to-point topology is the simplest, with a direct link between two ports (which are analogous to the SAS ports discussed earlier). It has the same effect as DAS, while supporting longer distances and working at a higher speed.
- The fabric topology is most flexible. It involves a set of ports attached to a network of interconnecting FC switches through separate physical links. The switching network (or fabric) has a 24-bit address space structured hierarchically, according to domains and areas. An attached port is assigned a unique address during the fabric login procedure (which we will discuss later). The exact address typically depends on the physical port of attachment on the fabric (or switch, to be precise).
- The fabric routes frames individually based on the destination port address in each frame header.
- The arbitrated loop topology allows three or more ports to interconnect without a fabric.

Source: [Cloud Computing: Business Trends and Technologies]

7. (5 points) How does the FCF respond to a discovery solicitation from the ENode?

Solution:

- ✓ An ENode selects a compatible FCF based on the advertisement and sends a discovery solicitation at which the capability negotiation starts.
- ✓ Upon receiving the solicitation, the FCF responds to the ENode with a solicited discovery advertisement, confirming the negotiated capabilities.
- ✓ Once receiving the solicited discovery advertisement, the ENode can proceed with setting up a virtual link to the FCF. The procedure here is similar to the fabric login procedure in FC.
- ✓ Successful completion of the login procedure results in creation of a virtual port on the ENode, a virtual port on the FCF, and a virtual link between them.

Source: [Cloud Computing: Business Trends and Technologies]

8. (5 points) Please answer the following four questions: a) What features of TCP are leveraged in iSCSI? b) Explain why these features are essential to SCSI operations. c) Why is not SCTP used in iSCSI? d) Why does iSCSI has to be deployed over an IPsec tunnel when its path traverses an untrusted network?

Solution:

a) TCP is leveraged in iSCSI for the features that are essential to SCSI operations: reliable in-order delivery, automatic retransmission of unacknowledged packets, and congestion control.
b) Multiple iSCSI nodes may be reachable at the same address, and the same iSCSI node can be reached at multiple addresses. As a result, it is possible to use multiple TCP connections for a communication session between a pair of iSCSI nodes to achieve a higher throughput.
c) The Stream Control Transmission Protocol (SCTP) is similar to TCP in its support for the features essential to SCSI operations. At the time of standardization of iSCSI, however, the SCTP was considered, too new to be relied on.
d) iSCSI itself does not provide any mechanisms to protect a connection or a session. All native iSCSI communication is in the clear, subject to eavesdropping and active attacks. In an untrusted environment, iSCSI should be used along with IPsec.

Source: [Cloud Computing: Business Trends and Technologies]

9. (10 points) What is connection allegiance? Explain how iSCSI sessions are managed.

Solution:

- An iSCSI session is a set of TCP connections linking an initiator and a target. This set may grow and shrink over time, allowing us to aggregate multiple TCP connections to achieve a higher throughput.
- With the availability of multiple connections comes the problem of using them correctly in the context of carrying out I/O. It is certainly reasonable to use separate connections for control and data transfer to ensure that a connection is always available for task management. Yet such a scheme requires monitoring and coordination across multiple connections, which can even require different adaptors on the initiator or the target.
- To avoid this complexity, iSCSI employs a scheme known as connection allegiance. With this scheme, the initiator can use any connection to issue a command but must stick to the same connection for all ensuing communications.
- The iSCSI sessions need to be managed. A big part of session management is handled by the iSCSI login procedure. Successful completion of the login procedure results in a new session or adding a connection to an existing session.
- A prerequisite for the procedure is that the initiator knows the name and address of the storage device (i.e., the target) to use. One approach is to have such information pre-configured in the initiator. Then any change will require reconfiguration.

Source: [Cloud Computing: Business Trends and Technologies]

10. (10 points) Why the credential (as defined in ANSI INCITS 458-2011) itself cannot serve as a proof for access control? Give one example of a proof derived from the capability key.

Solution:

- The access control mechanism as standardized in ANSI INCITS 458-201140 is based on the notion of capability and credential.
- A capability describes the access rights of a client to an object, such as read, write, create, or delete.
- A credential is essentially a cryptographically protected tamper-proof capability, involving the keyed-Hash Message Authentication Code (HMAC)41 of a capability with a shared key. More specifically, a credential is a structure:

<apability, object storage identifier, capability key>,

where

capability key = HMAC (secret key, capability‖object storage identifier).

- At a minimum, it should be verifiable, tamper-proof, hard to forge, and safe against unauthorized use.

- A credential meets all but the last requirement; there is no in-built mechanism to bind it to the acquiring client or to the communication channel between the client and the storage device. (In contrast, a driver's license has a photograph of the driver to bind the license to the driver, although such a strong binding is not necessary for the problem at hand.) This is clearly not good, especially if the credential is subject to eavesdropping over an improperly protected storage transport. Thus, another proof scheme is in order.

Source: [Cloud Computing: Business Trends and Technologies]

11. (10 points) Describe the three approaches to the block-level virtualization. Which approach is most suitable to the needs of Cloud Computing? What are the differences between the in-band and out-of-band mechanisms of the network-based approach along with their advantages and disadvantages?

Solution: There are three approaches to block-level virtualization depending on where virtualization is done: the host, the network, or the storage device.

i. In the **host**-based approach, virtualization is handled by a volume manager, which could be part of the operating system. The volume manager is responsible for mapping native blocks into logical volumes, while keeping track of the overall storage utilization. Ideally the mapping should provide a capability to be adjusted dynamically to allow the capacity of virtual storage to grow or shrink according to the latest need of a particular application.

ii. In the **storage device**-based approach, virtualization is handled by the controller of a storage system. Because of the close proximity of the controller to physical storage, this approach tends to result in good performance.

iii. In the **network**-based approach, virtualization is handled by a special function in a storage network, which may be part of a switch. The approach is transparent to hosts and storage systems as long as they support the appropriate storage network protocols (such as FC, FCoE, or iSCSI). Depending on how control traffic and application traffic are handled, it can be further classified as in-band (symmetric) or out-of-band (asymmetric).

   a. In **in-band** approach, the virtualization function for mapping and I/O redirection is always in the path of both the control and application traffic.
   Advantages: The central point of control afforded by the in-band approach simplifies administration and support for advanced storage features such as snapshots, replication, and migration. The snapshot feature is of particular relevance to Cloud Computing. It can be applied to capture the state of a virtual machine at a certain point in time, reflecting the run-time conditions of its components (e.g., memory, disks, and network interface cards). The state information allows rolling back after applying a patch or a failure.

Disadvantages: Virtualization function could become a bottleneck and a single point of failure. Caching and clustering are common techniques to mitigate these problems. The performance of other virtual machines on the same host may suffer when the snapshot of a virtual machine is being taken.

b. In **out-of-band** approach, the virtualization function is in the path of the control traffic but not the application traffic. The virtualization function directs the application traffic.
Advantages: In comparison with the in-band approach, the approach results in better performance since the application traffic can go straight to the destination without incurring any processing delay in the virtualization function. But this approach does not lend itself to supporting advanced storage features. More important, it imposes an additional requirement on the host to distinguish the control and application traffic and route the traffic appropriately. As a result, the host needs to add a virtualization adaptor, which, incidentally, may also support caching of both metadata and application data to improve performance.
Disadvantage: Per-host caching, however, faces the challenging problem of keeping the distributed cache consistent.

Source: [Cloud Computing: Business Trends and Technologies]

12. (5 points) Explain the difference (in terms of their capabilities) between the NOR flash- and NAND flash solid state drives.

Solution: NOR flash has its basic construct properties that resemble those of a NOR gate. NOR flash is fast (at least faster than hard disk), and it can be randomly addressed to a given byte. Its storage density is limited however.

NAND flash has its basic construct properties similar to those of a NAND gate. NAND flash, however, allows random access only in units that are larger than a byte. The NAND flash has made a splash in consumer electronics, and it is used much more widely than NOR flash—in digital cameras, portable music players, and smart phones.

Source: [Cloud Computing: Business Trends and Technologies]

13. (5 points) What are the three limitations that stand in the way of deploying the NAND flash solid state drives in the Cloud?

Solution: To be deployed in the Cloud, the solid-state drives must overcome three limitations inherent to NAND flash:

i. A write operation over the existing content requires that this content be erased first. (This makes write operations much slower than read operations.)
ii. Erase operations are done on a block basis, while write operations on a page basis47;
iii. Memory cells wear out after a limited number of write–erase cycles.

Given the limitations, directly updating the contents of a page in place will cause high latency because of the need to read, erase, and reprogram the entire block. Obviously, this is not desirable, which gives rise to the practice of relocate-on-write (or out-of-place write).

Source: [Cloud Computing: Business Trends and Technologies]

14. (10 points) Explain the mechanism of consistent hashing used in Memcached servers.

Solution:

- Depending on the size of DRAM available on a server, caching the workload data may need more than one server. In this case, the hash table is distributed across multiple servers, which form a cluster with aggregated DRAM. Memcached servers, by design, are neither aware of one another nor coordinated centrally. It is the job of a client to select what server to use, and the client (armed with the knowledge of the servers in use) does so based on the key of the data item to be cached.
- A naïve scheme might be as: $s=H(k) \bmod n$; where H(k) is a hashing function, k the key, n the number of server, and s the server label, which is assigned the remainder of the division of H(k) over n. The scheme works as long as n is constant, but it will most likely yield a different server when the number of servers grows or shrinks dynamically—as is typically the case in Cloud Computing. As a result, cache misses abound, application performance degrades, and all servers in the latest cluster have to be updated.
- Since this is undesirable, and so another scheme is in order.
- Memcached implementations usually employ variants of consistent hashing to minimize the updates required as the server pool changes and maximize the chance of having the same server for a given key.
- The basic algorithm of consistent hashing can be outlined as follows:
    - Map the range of a hash function to a circle, with the largest value wrapping around to the smallest value in a clockwise fashion;
    - Assign a value (i.e., a point on the circle) to each server in the pool as its identifier49; and
    - To cache a data item of key k, select the server whose identifier is equal to or larger than H(k).

- An immediate result of consistent hashing is that a departure or an arrival of a server only affects its immediate neighbors. In other words, when a new server p joins the pool, certain keys that were previously assigned to the original p's successor will now be reassigned to server p, while other servers are not affected.
- Similarly, when an old server p leaves the pool, the keys previously assigned to it will now be reassigned to p's successor while other servers are not affected.
- The basic algorithm allows the server pool to scale effectively and provides a sound foundation for further enhancements.

Source: [Cloud Computing: Business Trends and Technologies]