Homework 5: CS 524

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 rack-mounted server?**

ANS.

The motivation behind the two different commonly used server placement as can be understood below are based upon:

- Space Constraints
- Processing Power
- Cooling, Energy and other infrastructure
- Availability, Maintenance and Flexibility

The Blade servers are more densely packed than Rack server as it sacrifices the cooling and other components like video cards to server chassis for each unit and thus allows it to be thinner only packing CPU, memory, networking and storage. It can thus allow it to be more compact than a rack mounted server by using centralized systems.

Rack Server:

A rack server, or rack-mounted server, is any server that is built specifically to be mounted within a server rack. Rack servers are a general-purpose machine that can be configured to support a wide range of requirements. It can be secured into the rack using mounting screws or rails, depending on the design. The height, or the amount of rack units the system might take up, can vary quite a bit. Depending on what is required from the system. Larger servers allow for additional CPUs, memory, or other components. The servers themselves are mounted one on top of the other within a rack. To help minimize the amount of space used.

Rack Server Pros:

• Self-contained: Each rack server has everything necessary to run as a stand-alone or networked system: its own power source, CPU, and memory. This enables rack servers to run intensive computing operations.

• Efficiency: Rack-mounted servers and other computing devices make highly efficient use of limited data center space. Rack servers can be easily expanded with additional memory, storage, and processors. And it's physically simple to hot-swap rack servers if admins have shared or clustered the server data for redundancy.

• Cost-effective: Smaller deployments offer management and energy efficiency at lower cost.

Rack Server Cons

• Power usage: Densely populated racks require more cooling units, which raises energy costs. Large numbers of rack servers will raise energy needs overall.

• Maintenance: Dense racks require more troubleshooting and management time.

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. 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).

Blade Server Pros

• Low energy spend: Instead of powering and cooling multiple servers in separate racks, the chassis supplies power to multiple blade servers. This reduces energy spend.

• Processing Power: Blade servers provide high processing power while taking up minimal space.

• Multi-Purpose: They can host primary operating systems and hypervisors, databases, applications, web services, and other enterprise-level processes and applications.

• Availability: The blade server environment simplifies centralized monitoring and maintenance, load balancing, and clustered failover. Hot swapping also helps to increase system availability.

Blade Server Cons

• Upfront costs: Over time, operating expenses are reasonable thanks to simplified management interfaces and lower energy usage. However, initial capital, deployment, and configuration costs can be high.

• Energy costs: High density blade servers require advanced climate control. Heating, cooling, and ventilation are all necessary expenditures in order to maintain blade server performance.

(reference: https://www.racksolutions.com/news/data-center-optimization/blade-server-vs-rack-server/, https://www.serverwatch.com/hardware/blade-servers-vs-rack-servers/)

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?]**

ANS.

The servers of a data center need to be interconnected, and they need to connect to the outside world as well. As the number of servers increases, more cables have to fit into a given space. Top-of-Rack (ToR) and End-of-Row (EoR) are two approaches to connectivity resulting in different cabling options. Both ToR and EoR switches are typically implemented using Ethernet technology. Ethernet technology is particularly important to data centers because of its potential to eliminate employing separate transport mechanisms (e.g., FC) for storage and interprocessor traffic.

(References: Textbook: Cloud Computing: Business Trends and Technologies)

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)?**

ANS.

In terms of how it is connected to servers, storage may be classified as Direct-Attached Storage (DAS), Network-Attached Storage (NAS), and Storage Area Network (SAN). Networked-attached storage provides connectivity to the virtual server through a TCP/IP connection and storage access is provided at the file level. Storage area networks provide connectivity to the virtual server using either the Fibre Channel (FC) or iSCSI protocols. NAS is the better option if you require the ability to clone large numbers of virtual hosts. However, SAN will be a better fit if you require the highest levels of availability and resiliency.

DAS, as the term implies, is directly attached to a processor through a point-to-point link (The dominant technology in this case is the hard-disk drive). In contrast, NAS and SAN reside across a network. This network is purpose-built for, and dedicated to, storage traffic in the case of SAN.

NAS and SAN are readily applicable to Cloud Computing, but DAS has a limitation. An essential feature of Cloud Computing is flexible allocation of virtual machines based on, among other factors, resource availability and geographical location. In the DAS case, when a virtual machine moves to a new physical host, the associated storage needs to move to the same host, too, which is likely to result in consuming both much bandwidth and much time.

Directly attached storage(DAS) is faster than any other storage methods. This is because it does not involve any overhead of data transfer over the network (all data transfer occurs on a dedicated connection between the server and the storage device. Thus DAS is suitable for keeping local data (such as boot image or swap space).

(References: Textbook: Cloud Computing: Business Trends and Technologies, https://www.slashroot.in/san-vs-nas-difference-between-storage-area-network-and-network-attached-storage)

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

ANS.

Physical layer deals with the physical and electrical characteristics of cables, connectors, and transceivers.

Phy layer deals with line coding, out-of-band signals, and other preparations (e.g., speed negotiation) necessary for serial transmission. The name of the layer reflects the logical construct phy that represents a transceiver (consisting of a transmitter and a receiver) on a device. A phy has an 8-bit identifier that is unique within a device. The identifier is assigned by a management function. Its value is an integer equal to or greater than zero and less than the number of phys on the device.

(References: Textbook: Cloud Computing: Business Trends and Technologies)

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

ANS.

The NFS proper is client/server based. The NFS client initiates requests through the NFS protocol, which relies on the Remote Procedure Call (RPC). The NFS server only responds to requests, taking no actions on its own. The use of RPC hides the network-related details.

The close system call does not result in an RPC invocation. There are two reasons for this. First, the NFS protocol does not have the close routine because of the original stateless design of servers (which do not keep track of past requests) to facilitate crash discovery. Second, in this case there is no file modification.

A remote file operation, even if it has an RPC counterpart, does not necessarily result in an RPC invocation. No such invocation is needed when the information is stored in the client cache, which reduces the number of remote procedure calls and improves performance. Nevertheless, caching makes it difficult to maintain file consistency.

(References: Textbook: Cloud Computing: Business Trends and Technologies)

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

ANS.

F C-2M is concerned with end-to-end connectivity, addressing, and path selection. Three types of connection are supported:

a. point-to-point
b. fabric, and
c. 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. On the loop, only two ports can communicate with each other at any given time through arbitration.

In all three types of topology, communication may be simplex, full-duplex, or half-duplex; and a port may be on an HBA, a storage device controller, a hub, or a switch.

(References: Textbook: Cloud Computing: Business Trends and Technologies)

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

ANS.

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 a creation of virtual port on the ENode, a virtual port on the FCF and a virtual link between them.

(References: Textbook: 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?

ANS.

a. A primary purpose of the TCP protocol is so that OS can ensure that any dropped or lost packets are handled by the TCP/IP stack in the operating system. The application can simply hand off the data to the network driver which will guarantee the delivery of the packet.
   - TCP is reliable protocol.
   - TCP ensures that the data reaches intended destination in the same order it was sent.
   - TCP is connection oriented. TCP requires that connection between two remote points be established before sending actual data.
   - TCP provides error-checking and recovery mechanism.
   - TCP provides end-to-end communication.
   - TCP provides flow control and quality of service.

b. 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.

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. Thus IPsec is used as an security protocol.

(References: Textbook: Cloud Computing: Business Trends and Technologies, https://www.tutorialspoint.com/data_communication_computer_network/transmission_control_protocol.htm)

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

ANS.

One problem with using TCP/IP as transport is under-utilization of the underlying physical media. As a remedy, the notion of an iSCSI session is introduced.

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.

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.

(References: Textbook: 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.

ANS.

Granular access control is fundamental in Cloud Computing. 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) of a capability with a shared key. More specifically, a credential is a structure:

<capability, 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.

The standardized scheme derives a proof based on the capability key. The proof is a quantity computed with the capability key over selective request components according to the negotiated security method.

(References: Textbook: 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.

ANS.

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

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. A major drawback of the approach is that per-host control is not favorable to optimal storage utilization in a multi-host environment, not to mention that the operational overhead of the volume manager is multiplied.

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. Nevertheless, it has the drawback of being vendor-dependent and difficult (if not impossible) to work across heterogeneous storage systems.

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).

In-band approach, where the virtualization function for mapping and I/O redirection is always in the path of both the control and application traffic. Naturally the virtualization function could become a bottleneck and a single point of failure. Caching and clustering are common techniques to mitigate these problems. On the positive side, 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. Nevertheless, there is a trade-off as in this case the performance of other virtual machines on the same host may suffer when the snapshot of a virtual machine is being taken.

Out-of-band approach, where the virtualization function is in the path of the control traffic but not the application traffic. The virtualization function directs the application traffic. 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. Per-host caching, however, faces the challenging problem of keeping the distributed cache consistent.

(References: Textbook: 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.

ANS.

NOR flash because its basic construct has properties resembling 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 memory removes this limitation (while also reducing the cost). It is called NAND flash because its basic construct has properties similar to those of a NAND gate. NAND flash, however, allows random access only in units that are larger than a byte.

Feature	NOR Flash		NAND Flash	
	General	S70GL02GT	General	S34ML04G2
Capacity	8MB – 256MB	256MB	256MB – 2GB	256MB
Cost per bit	Higher	6.57×10^{-9} USD/bit for 1ku	Lower	2.533×10^{-9} USD/bit for 1ku
Random Read speed	Faster	120ns	Slower	30µS
Write speed	Slower		Faster	
Erase speed	Slower	520ms	Faster	3.5ms
Power on current	Higher	160mA (max)	Lower	50mA (max)
Standby current	Lower	200µA (max)	Higher	1mA (max)
Bit-flipping	Less common		More common	
Bad blocks while shipping	0%		Up to 2%	
Bad block development	Less frequent		More frequent	
Bad block handling	Not mandatory		Mandatory	
Data Retention	Very high	20 years for 1K program-erase cycles	Lower	10 years (typ)
Program-erase cycles	Lower	100,000	Higher	100,000
Preferred Application	Code storage & execution		Data storage	

(References: Textbook: Cloud Computing: Business Trends and Technologies, Flash 101: NAND Flash vs NOR Flash - Embedded.com)

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

ANS.

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

1. A write operation over the existing content requires that this content be erased first. (This makes write operations much slower than read operations.)

2. Erase operations are done on a block basis, while write operations on a page basis;

3. 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).

(References: Textbook: Cloud Computing: Business Trends and Technologies)

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

ANS.

Memcached supports a simple key-value store for small chunks of arbitrary data in DRAM on commodity computers. It is specific to caching to allow applications to bypass heavy operations such as database queries. Data durability is never part of the equation; each cached item is valid only for a certain period. Memcached is client-server-based, employing a request/response protocol (which may run over TCP or UDP).

A server stores data in a hash table. Keys are unique strings used to index into the table. For example, a result from a database query can be cached in a memcached server with the query string as the key. Although each data item has a limited lifetime, memcached does not implement garbage collection to actively reclaim memory. Instead, memory is reclaimed only when an expired item is being retrieved or when the space is needed for caching a new item.In the latter case, one of the least-recently-used items is subject to eviction. If an expired item exists, it is selected for reclaiming first. Otherwise, a still-valid item is selected.

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 naive scheme to distribute the hash table so that the same server is selected for the same key might be as follows:

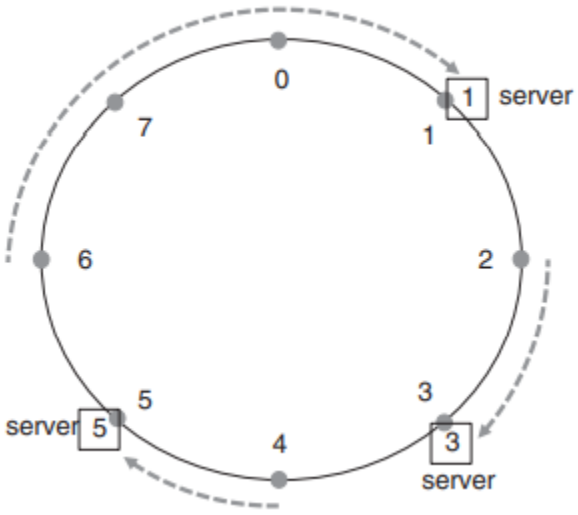
$s = H(k) \bmod n$,

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.

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 [36] 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 identifier; and
- To cache a data item of key k, select the server whose identifier is equal to or larger than H(k).

The server selected for key k is called k's successor, which is responsible for the arc between k and the identifier of the previous server. Below figure shows a circle of three servers, where server 1 is responsible for caching the associated data items for keys hashed to 6, 7, 0, and 1; server 3 for keys hashed to 2 and 3; and server 5 for keys hashed to 4 and 5.



An immediate result of consistent hashing is that a departure or an arrival of a server only affects its immediate neighbors.

Overall, memcached proves to be an effective, scalable mechanism to improve application performance. It is widely used by high-traffic websites such as Facebook, Twitter, and YouTube. In particular, Facebook has deployed thousands of memcached servers to support its social networking services, creating the largest key-value store in the world—where over a billion requests per second are processed and trillions of items are stored

(References: Textbook: Cloud Computing: Business Trends and Technologies)