Data Center Infrastructure

Acknowledements

Images are borrowed from many publicly available Internet sources

What is a Data Center (DC)?

- "…a data center is a physical facility that organizations use to house their critical applications and data…" – Cisco
- "A data center is a facility that centralizes an organization's IT operations and equipment, as well as where it stores, manages, and disseminates its data." – Palo Alto Networks
- "A data center is a building, dedicated space within a building, or a group of buildings used to house computer systems and associated components, such as telecommunications and storage systems" – Wikipedia
- "A data center is a facility housing many networked computers that work together to process, store, and share data" - Cloudflare

- Important characteristics
  - Physical space somewhere
    - Size depends on amount of computing and associated resources to be placed based on needs of the organization
  - Contains computer systems and associated components
    - Servers, storage, network switches,….
  - For now, we will focus only on the physical aspects of the data center and not what services it can provide (will see later)
- At the backend of so many services you use every day
  - Google, Facebook, Twitter, Amazon, Amazon AWS service, Microsoft Azure cloud, Netflix, ….
- Many many other specialized services in most organizations
  - The term is used very broadly

- So what do you need in a data center?
  - The servers, storage, switches, racks etc. of course…
    - Number can range from a few servers to tens of thousands of servers
    - How do you organize them? How do you connect them?
  - They need continuous power to run, so UPS
    - How do you plan how much UPS power you need?
    - Input to the UPS?
  - They generate heat, so cooling components like AC
    - So just blow cold air from somewhere like in your rooms?
    - How do you plan how much cooling capacity you need?
    - May sound unrelated to CS, but cooling is one of the most important issues in data center design
      - In general, building energy-efficient systems/algorithms is a very important area of research

- Having so much power in one place increases chance of fire, so Fire monitoring/suppression systems
- Costly equipment, so physical security is important, so Access Control system
- Lots of space and lots of cables, easy target for rats/insects ☺ So Rodent-repellant system
- If it is a large space, how do you monitor everything is working, get alerts for faults etc. from one place? So BMS (Building Management System)
  - Can be from simple to very complex and powerful systems
- Some others like water leak detection system etc.
- What if the power fails? Need Generators
- If your service is critical, need to worry about failures
  - What if a UPS goes down?
  - Ask the same question for all other components
- A small data center may not have all of the above for cost reasons, but good to have

- And then there are the computers, storage, switches etc. to actually give the service
  - How do you manage them?
- Typically, the companies that build the data center (without the computing infrastructure) are different from the ones that provide the computing infrastructure
  - Some companies do both
- Two approaches
  - Build a data center, and then see what you can put in there
    - Some companies just build them and rent them out to you to keep your servers
  - Plan your system, and then build the data center around it
    - Some organizations build their own data centers

Aim for the next few lectures

- Give you a basic understanding of hardware devices of interest like servers, storage etc.
  - You may think you know them, but usually you don't from my experience ☺
  - There are devices that you have not seen which are very important from an end-to-end networking and service delivery point of view
- Give you an idea and visualization of what is a data center and what does it look like physically, from small to large
- Make you aware of some of the issues in building and operating a data center
- Will not make you experts in designing them!
  - But give you a holistic picture of where your network is supposed to work
  - Network design (physical interconnection and protocol design) will depend on the environment (will point out later)

Basic Hardware Devices

Typical Hardware Components

- Servers
- Storage
- Switches/Routers
- Access Points
- Wireless Controllers
- Firewalls
- Load balancers
- Racks
- Cables – fiber, copper
- KVM Switch
- …

How many do you know in some detail?

Servers

- Types of servers (form factor)
  - Tower
  - Rack
  - Blade
  - High-density servers
- Important considerations
  - Power efficiency
    - Affects both power needs and cooling needs
  - Space needed
  - Management issues
    - Cabling
    - Configuration complexity

Tower Servers

- One server in one box
- Not scalable in space
  - Typically, horizontal space is at a premium in a data center
  - Costly. Also, more space to cool
- Not always power-efficient
  - A typical server power supply – 700W or higher
  - A Xeon CPU takes around 150-165W peak
  - Even with 2 CPUs and others, power may go waste unless you have GPUs etc.
  - Cooling is dependent on actual dissipation
    - How will you design? Based on the power supply capacity or based on the actual configuration?

Rack servers

- One server in the basic form
- Scalable in space
  - Can be stacked in a rack to take vertical space
  - Amount of space taken depends on server configuration (mostly no. of disks, add-on cards like GPU etc., CPU/Memory/Network do not add much space)
    - 1U, 2U, 4U (1 U = 1-3/4 inches) typically
- Same issues with power-efficiency

Servers in a Rack

High Density Rack Servers

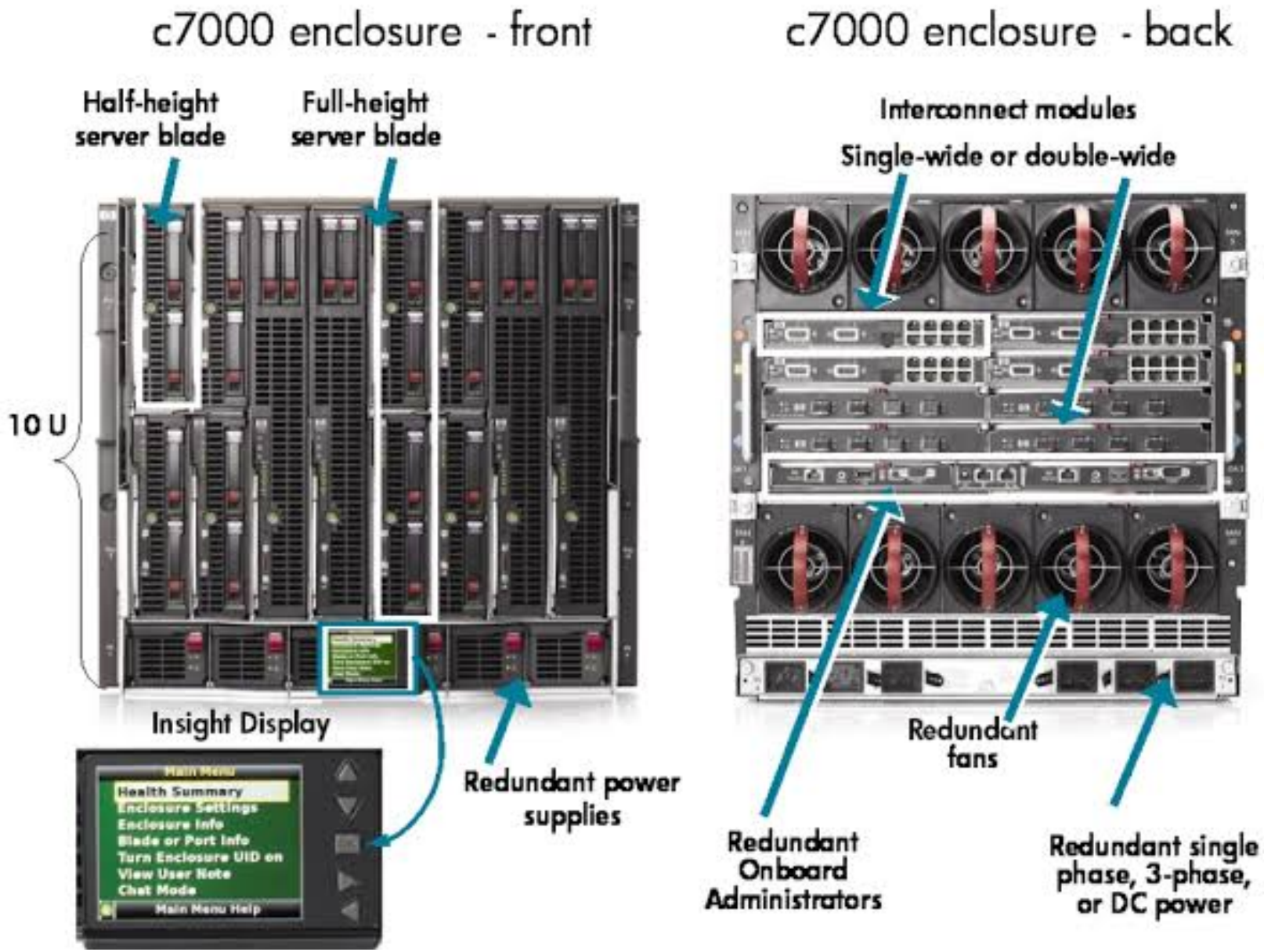
- A variation where you pack more than one server (typically up to 4) in a single 2U frame
- All 4 servers share the same power supply
- Fans may not be shared
- More space efficient and power efficient
- Fixed size, so limit on what you can pack in a server
  - Not too much disk
  - Not many additional cards for additional features

Blade Server

- Two parts (supplied separately)
  - A chassis containing power supply, fans, and a backplane
  - Servers that slide into the chassis and connect to power supply etc. through the backplane
    - You can put any number of servers up to a maximum number (typically, around 8, 10, 14, 16, 18 depending on OEM and model)
  - Can also take other devices like network switch that can slide into the chassis backplane and provide interconnection between the servers etc.

A chassis that can take 16 servers

Figure 1: Front and back views of the BladeSystem c7000 Enclosure (single phase version)

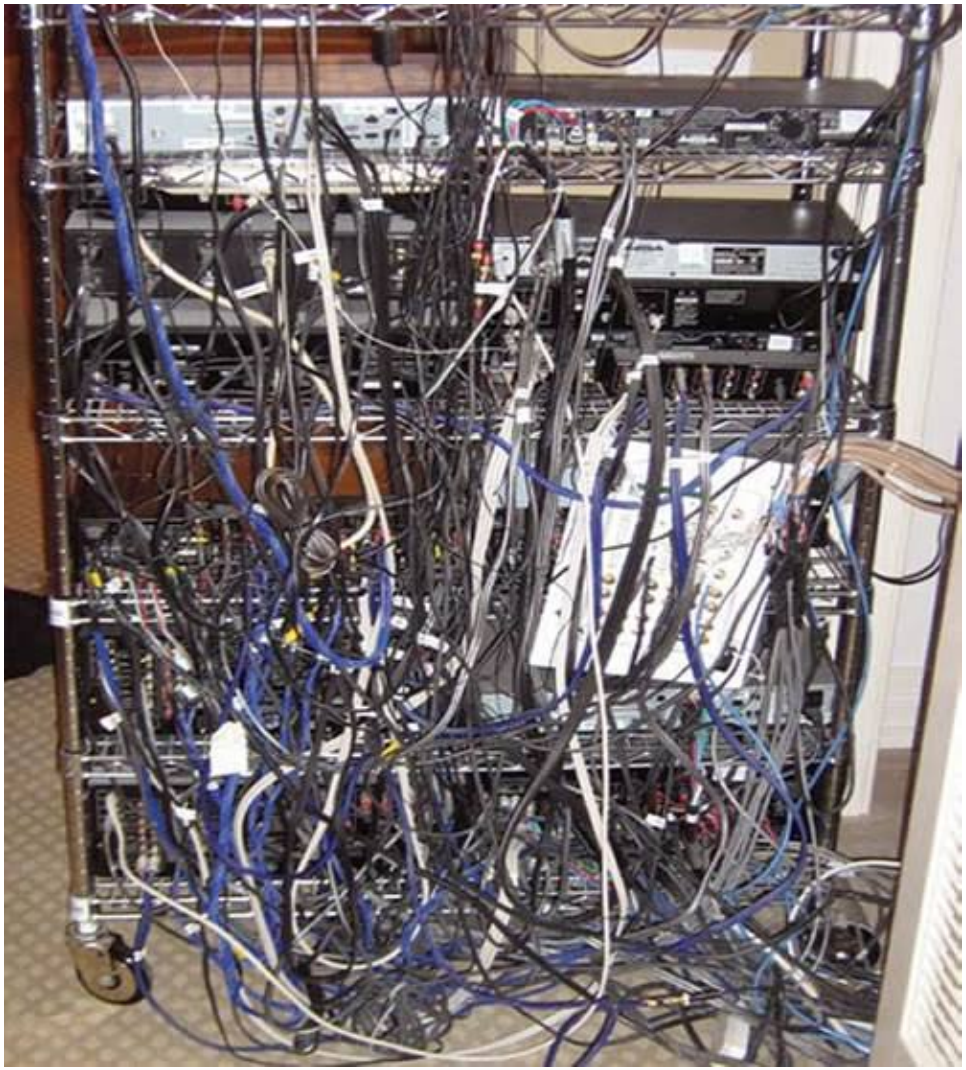


- Common power supply and fans for the servers, switches etc.
- Extremely space efficient
  - The chassis is typically between 5U to 10U depending on number of servers that can be put in
  - The chassis slides into the rack
- Power-efficient
- All servers can be managed from a single console
- Fixed size, so limits what can be put on each server
  - Storage per server is limited
  - Additional cards beyond standard configurations may be a problem

Cabling Aspects

- Consider 10 rack servers in a rack
- Each server will at least have 1 power cord, 1 network cord
  - Typically more than one for fault-tolerance
  - There may be other cables for management, storage connectivity etc.
- Bad cabling can soon get out of hands

Bad!



Good

- Blade servers can reduce cabling complexities even further
- Consider 16 blade servers in a single chassis
  - Put the network switch inside the chassis (rack-integrated switch)
  - No. of network cords coming out of the chassis is 1-2 (only the uplink ports of the switch)
  - No of power cables = 1 or 2 (dual supply, typical)
  - Compare that with no. of cables that will come out of 16 rack servers

- Data centers typically will use high-density servers or blade servers
  - Not always, there are other issues we will discuss later
- How many servers can you pack in a rack?
  - Depends on rack size
  - Max rack size (height) is 42U usually
    - Depth and width varies, but within a set so that servers from any vendor can go in
  - Cooling needs will limit how much you can pack in a rack

Typical Cabling

- Each rack has one or two PDU (power strip) to connect the server power supplies
- For network connection
  - Put a Top-of-Rack (TOR) switch in each rack
  - Connect the server network cords to it
  - Take the uplink port connections out to next level switch
- May need other connections to storage etc.
- We will see later and build a hypothetical system exactly

Storage

- Direct Storage
  - Directly attached to the server (inside the same box or in a box just outside the server (DAS or Direct Attached Storage))
- Storage Area Networks (SAN)
  - Shared storage over specialized protocols primarily
- Network Attached Storage (NAS)
  - Shared storage over IP network

Direct Attached Storage

- Extend the server's storage by putting additional disks in a box next to the server
- Directly connected to the server with a Host Bus Adapter (HBA)
- Any standard disk access protocol for accessing
  - SAS, SCSI, SATA,…
- Accessible only to the server, no direct sharing of the space with other servers (if other servers want to access, must be done through this server)
- Good for local extension of disk space beyond what can be fit into your server's box

Storage Area Network (SAN)

- Shared storage space that can be shared between servers
- Block level access similar to normal disks
- Connection over Fibre Channel (FC) protocol
  - Allows connection over IP network also with iSCSI protocol
    - In that case, acts similar to any IP device
- Typically one or more switches between servers and SAN storage
  - FC switch for FC connection (Fiber cables) or Ethernet switch for iSCSI connection
- Two parts
  - SAN controller (front end for connection and protocol)
  - Disk subsystem (the backend actual disks)
- Looks similar to rack servers and can be put on racks

Network Attached Storage (NAS)

- Shared storage that can be shared between servers
- File-level access (read/write files, not disk blocks)
- Two parts
  - NAS controller
    - Basically a file server with added functionality
  - Disk subsystem
- Accessed over the network, connected to a network switch the same way as any other IP device
- Looks similar to rack servers and can be put on racks for higher capacity NAS boxes
- Smaller capacity NAS boxes can be standalone units

NAS or SAN?

- SAN offers high speed access over FC, NAS is slower in general
  - 8/16 Gbps for FC SAN (typical), shared only between the servers accessing the storage
  - NAS speed depends on switch speed (1/10Gbps) and how many devices (may not be accessing the storage) are using the network
- SAN needs separate file server for file-level access
- SAN needs more cabling, and separate switch cost
  - FC switches are more costly compared to Ethernet switches (even 10G switches)

Disk Types

- Depending on access protocol
  - SATA
    - Typically enterprise SATA for servers
  - NL-SAS
  - SAS
  - SSD
- SATA to SSD – higher access speed, higher reliability, higher cost for same amount of storage, lower capacity per disk
- Disks are populated typically in RAID 5 or 6 for protection from disk failure

Switches

- Provides interconnectivity between devices
- Devices connect to ports on the switch
- Uplink ports to connect to other devices connected to other switches
- Switch type depends on protocol
  - Ethernet switch to connect devices using Ethernet protocol
    - Switches based on Ethernet address
  - L3 Switch
    - Switches based on IP address
  - FC switch to connect devices using FC protocol
    - Switches based on FC address
  - Infiniband switch
    - High speed network protocol for interconnecting large clusters
  - Others..

Network Switch

- L2/L3
  - L2 switches based on Ethernet address, L3 based on IP address
- Can be small capacity, low cost to very high capacity and high cost
  - Switching speed from 1 Gbps to 100's of Tbps
- Managed/Unmanaged switches
  - Managed switches can be monitored/configured from a central place
  - For unmanaged switch, someone has to physically go to the switch and do that
- Cost varies with managed/unmanaged, number of ports and switching capacity among others

A 48-port switch

A chassis-based higher capacity switch

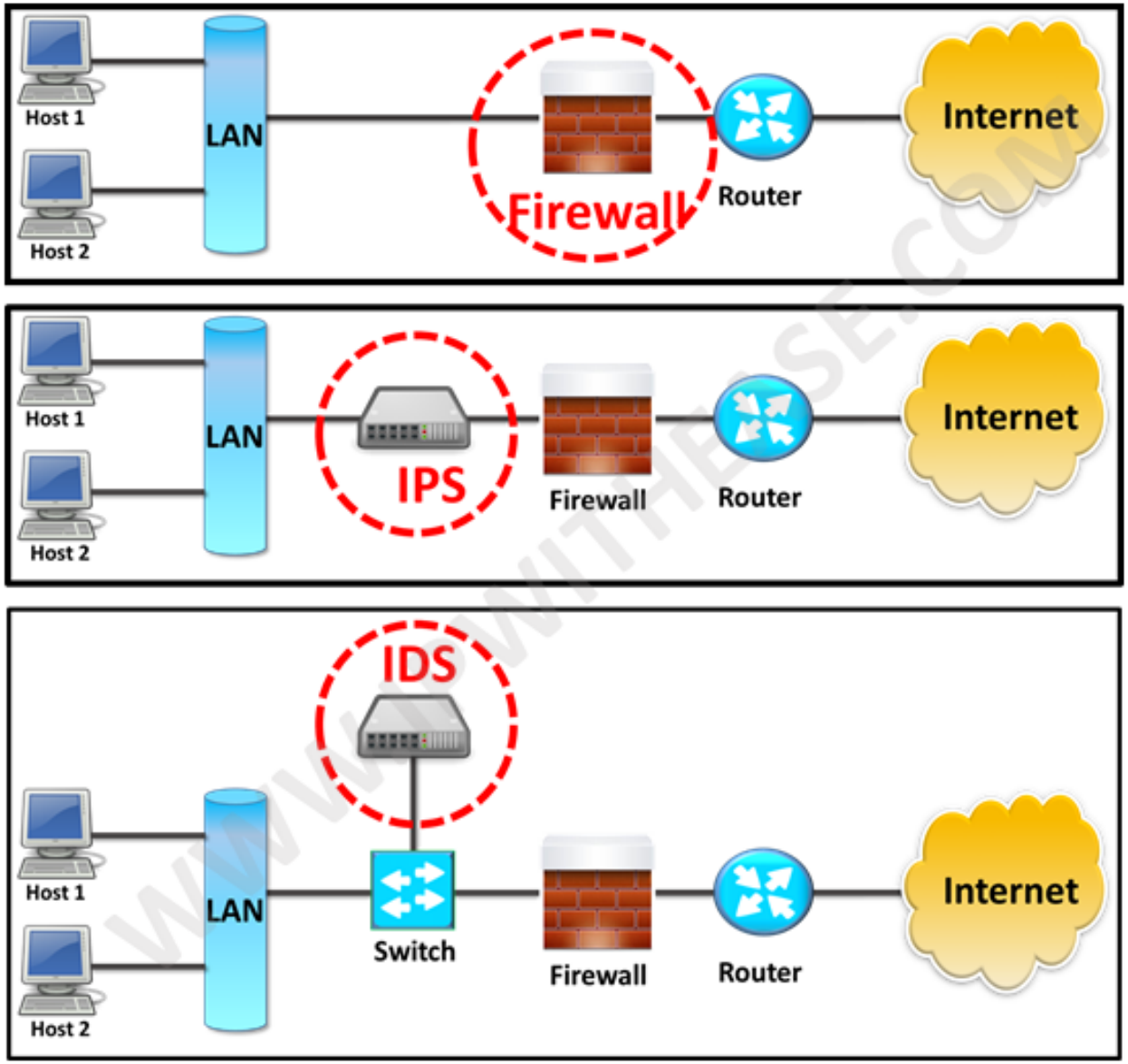
Other Important Active Devices

- Routers
  - Runs routing algorithms and forwards IP packets
  - Terms Routers and L3-switch are used interchangeably, as both switch based on L3 (IP) information
    - In practice the term routers is nowadays used for high end devices that give lot more features and configuration options
    - Still confusing, no clear accepted distinction
- Firewalls
  - Looks at packets and applies rules (block/allow) based on packet header information
    - Example: "drop packet if source IP = 10.20.3.4", "allow packet to server with IP 10.30.20.4 and port 80 only if coming from the subnet 10.30.20.0/24",....
  - All incoming and outgoing packets pass through the firewall

- Server Load Balancers (or called simply Load Balancers mostly)
  - Sits between an incoming connection on one side and multiple servers on the other side (servers can be connected directly to LB or through switch)
  - Monitors the status (up/down) and load of servers
  - Sends incoming connections to different servers to balance load on the servers
  - Needed only if load balancing is needed (essential in DCs)
- Link Load Balancers
  - Connects more than one ISP connection to a network, and balances the outgoing traffic across the multiple connections as per policy set

- Intrusion Prevention System (IPS)
  - Monitors potential security threats and attacks over the network, including virus/malware detection, url filtering etc.
  - All incoming and outgoing packets go through the IPS if it is there
  - Packets may be blocked if threats detected as per policy set to stop atttacks
- Intrusion Detection System (IDS)
  - Monitors packets to detect potential attacks and raises alert, but does not block them

A Typical Connection



Indicative connection shown. Other complex connections possible with multiple firewalls in large networks based on security needed

- Terminologies galore, confusing unless you deal with them regularly
  - ADC (Application Delivery Controllers)
    - Usually server load balancers with many other features for effective load balancing across services running on the servers
  - UTM (Unified Threat Management)
    - Threat prevention, including virus/malware detection, attack prevention, url filtering etc., sort of firewall +IPS + other features
    - In the connection shown in the last slide, you can replace the firewall + IPS with a UTM box in the same place
  - NGFW (Next generation Firewall)
    - Similar functionalities to UTM, though usually higher end
    - Can be connected in the same place as the UTM as mentioned above
  - Features vary widely depending on OEM and product
    - Major OEMs – Cisco, Checkpoint, Palo Alto Networks, Fortinet, Sophos, Radware,..
- A DC will have load balancing and threat protection implemented in some manner (we will look at load balancing later)

Racks

- Typically 42U racks in data centers for servers, storage, network switches
- For smaller data centers, smaller racks can be used for network switches etc.
- Typical sizes – 42U, 32U, 22U, 18U, 15U, 12U, 9U, 6U, 3U etc.

Cable Types

- CAT5/CAT5A/CAT6/CAT6A
  - Copper cables for standard network connection
  - Limited distance (around 100 meters max)
  - Good for lower bandwidth use (upto 1 Gbps usually, CAT6 can support 10Gbps)
  - Cheaper, more durable
  - Prone to electrical interference, puts constraints on cabling
- Optical Fiber
  - For higher speed use, no electrical interference
  - MMF (Multi-mode Fibers) for limited distance (upto 300-400 meters), SMF (Single-mode Fibers) for large distance (hundreds of km)
  - Native medium for some protocols like FC (though FC over copper is also supported)
  - Costlier, prone to damage
- So which one should you use?

How to Size UPS Power Supplies

- Compute the peak load of the system
  - For servers, compute primarily from the number and type of CPUs/GPUs and number and type of disks
    - Memory takes only around tens of milliwatts per GB
    - Disks can take around 5-10 W per disk (a ballpark figure, depends on type, less for SSDs)
    - Typically < 500W max for a standard 2-CPU server with current technology
  - For storage, compute from storage controller power specification and number and type of disks
  - For switches, compute from switch specification

- UPS output should support the peak load at around 80% capacity
  - Rule of thumb followed by many, there is no sacrosanct rule
  - Will depend a lot on cost
  - Duration of battery backup (15/30/45/60 minutes) at full load will depend on need and cost
    - Battery is the costliest item in a UPS (around 50-60% of the cost of a UPS by some estimates)
    - Need to be replaced every 2-3 years for maintenance-free sealed batteries
    - Larger duration means larger number/capacity of batteries, and higher cost

- If you already have a running data center, can make exact measurements for more precise planning
- Not everything in a data center needs to be on UPS
  - Many lights, some comfort ACs etc…..
- Input power needed
  - Power for IT load (computers, storage, network etc.)
  - Power for non-IT load (AC, lights, access control, BMS, UPS losses etc.
  - Very important operational cost
- Power efficiency of a data center is measured by PUE (Power Utilization Efficiency)
  - (Total power consumed)/(Power consumed by IT systems)
  - Should be low (between 1.5 or lower to 2 typically; depends a lot on environmental conditions)
- Leave the exact electrical design to experts!

Cooling Issues

- Single most important issue in data centers
- How to estimate cooling need?
  - Look at the peak power dissipation at full load
  - Calculate AC tonnage at 1 Tonne = 3.5KW (approx.)
  - Can take 70-80% of that if system is not expected to run at full load always
  - But capacity is not enough, type of cooling is important

Types of Cooling

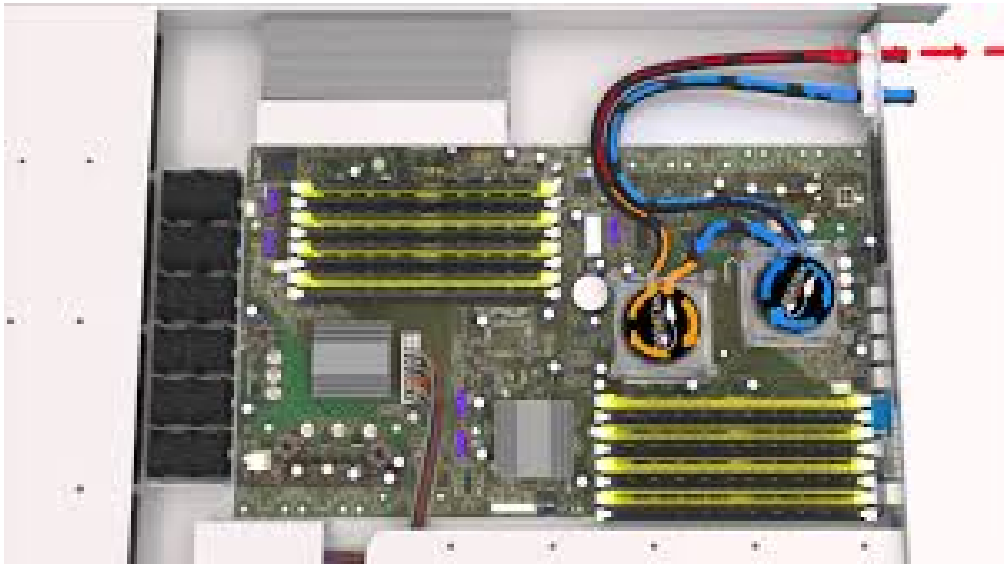
- Split/Window/Cassette /Tower AC
  - Ok only for very small data centers
- Precision AC (PAC)
- Inline cooling
- Rear Door Heat Exchangers (RDHX)
- Direct Liquid Cooling or Direct Cooling
- Basic philosophy
  - Processors generate the maximum heat (dissipated by heat sink)
  - Put a cooled gas/liquid near the heat which will absorb it and take the heat away
  - Re-cool the gas/liquid and re-circulate
  - Closer to the heat you can take the cooling agent, the better
  - Lesser the cooling agent travels after being cooled, the better

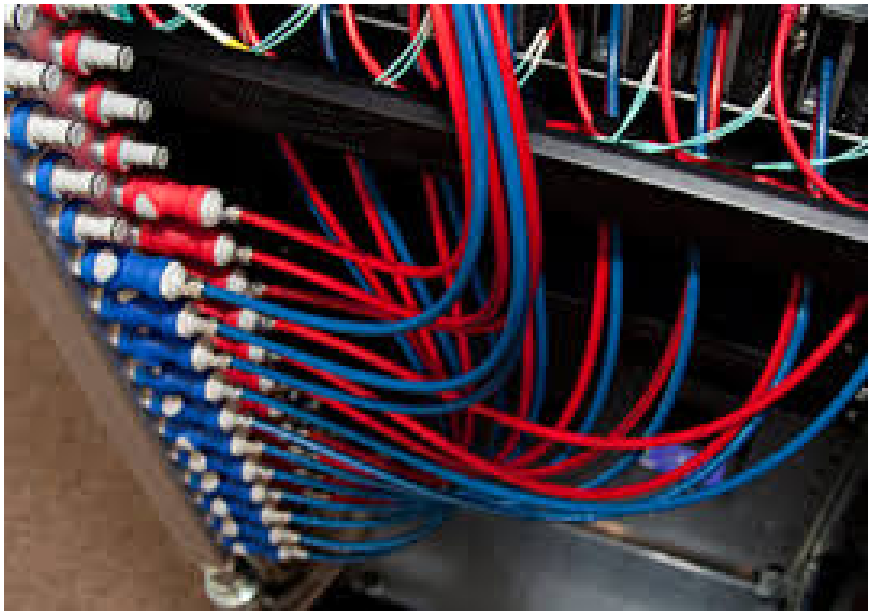
PAC Cooling

Inline Cooling

RDHX

Direct Liquid Cooling





- All of these will need the liquid to be re-cooled
- Done by chiller plants outside the data center room
- Capacity of chiller plants to be decided based on amount of heat to be cooled, the technology, temperature needed etc.
- Cooling system design is a very complex task
  - CS people are not expected to do it
  - But the guidelines help you in
    - Rough planning and budgeting
    - Sometimes preventing the vendor from feeding you wrong data! Extremely important for a systems manager

Rack Packing Thumb Rule

- PAC can support around 7-8 KW power dissipation per rack
- With inline cooling or RDHX, can go upto 15-20 KW typically
- Beyond that, requires direct liquid cooling typically
- Rough rule-of thumb, will vary widely with actual technology and make used
  - We also do a mixture, like some PAC and some RDHX to pack more in a rack

Putting it all together

- Lets make some rough design for an actual case
- Suppose you want to install
  - 100 servers, each with Dual Xeon Gold 6252 CPUs, 256GB RAM, 2x1.2 TB SAS disks, dual 1Gbps network ports
  - 200 TB of shared storage with 1.8 SAS disks in a NAS (What if SAN?)
  - How do we decide this? Another story, more later maybe…
- Design should be such that failure of any one device (including UPS, AC) should not stop the system
  - A standard requirement in most data centers
- We will design the infrastructure, interconnection, and a small DC layout in your classroom

A typical DC will have

- A server room with servers, storage, network etc.
- An Electrical Room with electrical panels, UPS, power distribution unit etc.
  - This is where the main power from transformers and DG sets will come in
- A battery room to keep the UPS batteries
  - Usually kept in a separate room for large data centers
- BMS/NOC room(s) – control room(s) where the operators/maintenance people will sit to monitor the data center
- Chillers to create the chilled waters needed for the AC
  - Depending on size and space availability, can be on roof, basement, or space outside the main data center rooms
- DG sets
  - Depending on size and space availability, can be on roof, basement, or space outside the main data center rooms

Electrical Room

Electrical Room

Battery Room

Chiller inside a building

Chiller on the roof (pipings inside)

A Fire Suppression Systems Schematic