CSG2132 Module 2 Notes

What is a Datacentre?

- A datacenter is a specialised facility an organisation uses to consolidate their business-critical systems, systems infrastructure, enterprise applications and data necessary to continuity of operation.
- A datacentre contains a wide range of hardware assets such as servers, routers, switches, firewalls and storage systems, as well as environmental maintenance and incident management systems that ensure equipment is kept safe and can operate within specific tolerance guidelines.
- The top priorities of any datacentre are reliability, efficiency, security, standards compliance and interoperability with other relevant systems.
- Enterprise-level datacentres usually occupy entire buildings or warehouses devoted to storing the above-listed elements and the personnel that manage them.
- Large companies such as Google, Facebook, Amazon and Microsoft possess many massive datacentres that house their critical infrastructure, located in many parts of the world. The demand for growth and management of datacentres is constantly growing, and companies are always searching for innovative ways to expand their own datacentre assets.

What is a server?

- A server is a specific type of computer designed to receive data and service requests from other computers on a network such as LAN, WAN or Internet, and respond to these with the information or applications requested.
- The word server is understood by most to mean a web server where web pages can be accessed over the internet through a client like a web browser. However, there are several types of servers, including local ones like file servers that store data within an intranet network.
- There are many different types of servers, each with a specific task or purpose, with the main ones being as follows:

Server Type	Description
Web server	Comprised of software or hardware using HTTP and related protocols to receive and fulfill WWW requests from a wide range of digital devices and their resident browsers.
Mail server	Manage the sending and receiving of electronic mail using several different protocols including SMTP (Simple Mail Transfer Protocol), POP3 (Post Office Protocol v3) and IMAP (Internet Message Access Protocol).

Media server	Delivers video and audio content to client devices and apps. A prime example is VOD (Video On Demand), that retrieves video content from storage systems and delivers it to client endpoints across the Internet as offered by Netflix, Amazon, Stan and Hulu. Another is Live streaming, that delivers content as it is generated in real time such pay-per-view events.
File server	Provides access to files. May take the form of a dedicated system such as a NAS or SAN, or as a standard server that hosts and provides access to an organisation's files. Can use a variety of protocols to communicate with client workstations including Ethernet LAN, TCP/IP and FTP.
Game server	Runs and provides access to online multiplayer games using either TCP (Transmission Control Protocol) or UDP (User Datagram Protocol).
Proxy server	Acts as a gateway between client devices and applications, e.g. web browser, and the internet. Proxy servers use a range of protocols such as SOCKS (Secure Sockets), HTTP (Hypertext Transfer Protocol) / HTTPS (Secure Hypertext Transfer Protocol) and FTP (File Transfer Protocol).
VPN server	Establishes and manages a '*virtual encrypted tunnel*' between a client application, e.g. web browser, and a remote service. Traffic is then routed through this tunnel so that transmitted data and its source is not visible to interceptors.
Database server	Establishes and manages access to database services and supporting protocols to client applications across a network. Examples include SQL Server, MySQL, Oracle, PostgreSQL
Virtualisation server	More commonly referred to as a Virtual Server, is a physical server compartmentalised (partitioned) into several smaller virtual (software only) using virtualization software. Each virtual server runs independently of the others, running their own operating systems, applications and services. Examples include VMWare Workstation, Hyper-V, Citrix Hypervisor and Xen Project.

Physical Servers

- Physical servers are just computers.
- Just like any PC, a physical server will have CPU, RAM, Disks and NICs.
- However, they are generally designed to be on all the time and modular enough to be easily maintained

Multiple services

- A single physical server can provide access to multiple services, e.g. proxy and email management services usually run on the same server.
- What services a server provides can be defined by the software running on the server
- This is common in smaller work environments

Distributed services

- A single service can also be provided by multiple servers working together
- This can be done for scalability and load balancing or for redundancy and failover

Server Racks

- Physical servers are stored in racks.

RACK UNITS

- Each rack can fit a certain number of devices based on standard units
- Smaller devices can take up only one unit (1U) of space
- Larger devices can take up several units of space

Virtual Servers

- Servers can also be virtualized, allowing them to run entirely as software.
- Entire networks can be virtualised, allowing for the topology of the logical infrastructure to be configured separately from the physical hardware.
- This can drastically reduce ongoing costs as new servers can be provisioned through software without needing to purchase additional hardware.
- Virtual servers are a highly popular solution for a wide range of applications dues to the advantages they provide, which include:
    - They save time, space and money
    - They can be centrally managed by minimal admins
    - Act just the same as physical servers in terms of client and application services
    - Can run a wide range of OS and applications on one (1) physical server
    - Can be configured to multiple levels of service, some offering fewer resources and performance, others offering more resources and higher performance

Networking

- Physical networking in datacentres is an important consideration.
- The logical topology of the network must be designed to suit the business needs of the datacentre.
- The physical locations of switches and routing equipment must be carefully considered.

Switch locations

- A network switch connects multiple devices on a network using packet switching protocols to receive and on-send data between them.
- Often, data centre server racks will contain an access switch to provide networking for the devices in the rack.
- These switches are often located either at the top or the bottom of the rack, depending on whether the cabling runs above the racks or below a raised floor.
- Alternately the switches will be placed in separate racks with cabling running through conduits overhead or beneath a raised floor

Datacentre Cabling

- With so many devices needing to be connected, cabling in datacentres is a huge issue. Poorly thought out datacenter cabling ends up in what is known in the industry as *cable spaghetti*. Poorly laid out cabling causes additional costs when it denies easy access to other equipment and when it must be reorganised due to such access issues. Poor cable design also results in excess heat when located to close to one another, which is another issue that needs to be avoided.
- There is an art to good cable management and datacentres are designed with structured cabling in mind.

Switch locations

- These switches may often be connected to a series of intermediary or aggregation switches
- These are used to consolidate the access switches in a way that cuts down on cabling to routers while still providing an equal path to each rack (no daisy chaining)
- This aggregation layer can also be used to add some network redundancy

Routing

- Layer 3 routing between networks often happens at two levels:
- Internal routing between different subnets within the datacentre
- Edge routing to the outside world or external networks
- Edge or Border routers are often separate from internal routers.

Datacentre Topology

- Datacentres can use one of several network topologies according to business requirements and constraints. One is the *tree-based* topology, of which a prime example is known as the *three-tier data centre* network that uses an access, aggregate and core layer. Another is the *hub* topology which many servers are cross-linked or cross-indexed for to achieve specific functionalities. There is also the *BCube* topology which is very

modular in design so that elements can be serviced or upgraded without affecting other parts of the network.

Oversubscription

- Oversubscription is the practice of assigning more users to a network access resource (switch, router) than it's designed to handle. It works on the principle that it's statistically unlikely that all users assigned to a network resource will attempt to use its bandwidth simultaneously, which of course, would cause performance degradation issues if that did indeed happen. In effect therefore, oversubscription is an exercise in calculating probability and risk. It's an important network concept because, when done properly, it allows a datacentre to service more users with less network assets, which it turn, cuts down long term expenditure.
- For example, if you *subscribed* two (2) 10Gb bit/sec PVCs (permanent virtual circuit) to a 20Gb bit/sec port, this would equate to *100% subscription*, and therefore, no oversubscription would exist. Both circuits in practice could run at maximum capacity and would be easily supported by the assigned PVC.
- However, four (4) 10Gb bit/sec PVCs were subscribed to a 20Gb port, this would result in *200% subscription*, aka a 2-to-1 oversubscription ration (OSR). Although this seems bad, it's not - if the network engineers have done their homework and neither PVC exceeds 50% capacity at any one time. If end users only make *intermittent* use of the resource, rather than constant, the oversubscription is a highly cost effective a logical network resource allocation solution.
- However, when calculating OSR, you must always keep one eye on the future, as network demands can easily grow and derail such a solution. The main issue with calculating sound OSR is that it's not always what kind of workload a network resource will have to handle in the future.

Datacentre Power

- Datacentres consume up to 1.5% of all electricity in the world.
- Some datacentre locations are chosen specifically to be close to cheap energy supplies such as hydroelectric dams
- High availability of power is important and large-scale UPS systems and emergency generators are often used for critical systems.

Datacentre Cooling

- Often, datacentres can use almost as much energy on cooling as on computation.
- Recently, there has been a large push towards lowering energy use with warmer data centres, but careful temperature control is still a huge part of datacentre design.

- Cooling systems are often a combination of air conditioning, cooling towers, chillers, pumps and humidifiers.

Cooling Solution	Description
Calibrated Vectored Cooling (CVC)	For high-density servers; maximises airflow through equipment to increase ratio of circuit boards per server chassis and using less fans
Chilled Water System	For mid-to-large-sized datacentres; uses chilled water to cool air circulated through equipment racks and chassis.
Cold Aisle/Hot Aisle Design	A server rack configuration that employs alternating rows of *cold aisles* and *hot aisles*. Cold aisles use cold air intakes on the front of the racks. Hot aisles use hot air exhausts on the back of the racks. In essence therefore, the hot aisles are expelling hot air into chilling units insert into the cold aisles.
Computer Room Air Conditioner	Known as CRAC units, use compressor powered air con units to draw air across refrigerant-filled cooling arrays.
Computer Room Air Handler	Known as CRAH units, use chilled water flowing through cooling coil and modulating fans to draw air from outside facility. Best for colder climates.
Direct-to-Chip Cooling	Liquid cooling method that delivers coolant to a cold plate incorporated into a motherboard's processors to disperse heat.
Evaporative Cooling	Controls temperature by exposing hot air to water, with the resulting evaporation drawing the heat out of the air. Requires a lot of water to work.
Free Cooling	Uses the outside atmosphere to bring cool air into the servers without powered chilling. Requires cool climates to be effective.
Immersion System	Liquid cooling solution that submerges hardware into non-conductive, non-flammable dielectric fluid.
Raised Floor	Uses a frame that lifts the datacentre floor above the concrete foundation. Intermediate space is used for water-cooling pipes or increased airflow.

Location

- The primary factors to be considered when selecting a datacentre location are:

Server Type	Description
Power cost	This will be impacted by local climatic conditions, electricity provider competitions, state and local taxes and charges, energy efficiency rebates and tax benefits, quality of energy supply, use of energy modulation and regulation equipment.
Network connection bandwidth	Local access to high bandwidth, high reliability, high QoS backbones connections is superior to having to go through lower capacity channels first to reach the backbone.
Latency to clients	The distance between a datacentre and its end users/application has a major effect on latency, i.e. the time I t takes a user request to reach the datacentre and a response to arrive back. As a general rule, the closer datacentres are located to their users, the better.
Cost of land	Generally speaking, the closer land is located to optimal resources for datacentres, the more expensive the land will be to purchase and pay ongoing duties and rates for.

Datacentre management

- DCIM (Datacentre Infrastructure Management) combines information technology (IT) with facility management technology to centrally monitor and manage a datacentre's critical systems. DCIM usually employs specialised software, hardware and sensors that facilitate common, real-time monitoring of network assets and supporting infrastructure. DCIM aims to identify and ameliorate sources of risk and improve the availability of performance of servers, storage, environment control hardware, data communications systems and staff.

Modularity

- As datacentres are often required to scale as the business requirements grow, modular systems are very popular.
- Shipping containers filled with data storage systems that can be easily attached to existing datacentres are becoming more common.
- Modular datacentre components are usually standardised and come with built in cooling and power options.
- They also have the added benefit of being portable. This allows for entire datacentres to be easily shipped between locations.

Cloud

- One of the largest uses for datacentre technology today is providing cloud services.
- Cloud services involve providing access to shared resources over the internet.

- Rather than developing their own infrastructure and configuring their own software systems, companies are instead outsourcing them to cloud providers such as Microsoft or Amazon.

Infrastructure as a Service (IaaS)

- Cloud providers give access to virtual servers and virtual infrastructure.
- Distributed pools of hypervisor servers support large numbers of virtual machines that can be easily provisioned
- Examples of IaaS providers include DigitalOcean, Azure Virtual Machines, Rackspace, Alibaba E-HPC and IBM Cloud Virtual Servers

Platform as a Service (PaaS)

- Application development and deployment services
- Compilation and software builds
- Web hosting and testing
- Examples of PaaS providers include Amazon Web Services (AWS), Oracle Cloud Platform (OCP), Google App Engine, Microsoft Azure, Salesforce and Red Hat OpenShift

Software as a Service (SaaS)

- Email services
- CRM
- Office applications
- Examples of SaaS providers include HubSpot, MailChimp, Shopify, SurveyMonkey, Zoom, Salesforce, Dropbox and Wix

SDN

- Software Defined Networking can be used to allow more flexibility in the network configuration
- The foundational principle of SDN is to separate the Control plane from the Data plane
- The logical flow of the network (routing paths and administration) can be physically separated from the actual flow of traffic.
- One of the goals of SDN is to get away from the traditional style of jumping between individual CLI terminals to administer each device.
- Instead the entire network can be controlled together at a higher level.
- For example, Google arranges their networks around groups of small switches that are linked together into a much larger logical switch.
- They use a custom software control stack named “Jupiter” that can manage the thousands of switches within a datacentre as a single fabric.
- Many SDN systems are proprietary and not publicly available.
- There are some open source solutions (such as Opendaylight) based on the Openflow standard.