**Data center architecture (characterization, comparison, disaggregation)**

Plan of the **presentation**:

Introduction (Duc)

I) Overview of Data Center Architecture (Duc)

II) Characterization of Data Center Architectures (Duc)

III) Comparison of Data Center Architectures (Jan)

IV) Disaggregation in Data Centers (Jan)

V) Case Studies and Real-World Examples (Jan)

VI) Future Trends and Challenges (Duc)

Conclusion (Duc)

I) Overview of DCA

A data center refers to the structural design and organization of a data center facility. It encompasses various components and subsystems, each serving a specific purpose in facilitating data storage, processing, and transmission.

Data centers are very important in many different fields. There are various type of data centers:

* Enterprise data centers: these data centers are constructed and owned by companies for their good functioning.
* Managed services data centers: they are deployed, managed and monitored by third-party service providers. Companies choose this kind of data center to access the features using a managed service platform, so that they don’t need to purchase equipment and infrastructures.
* Colocation data centers: They allow companies to rent spaces to host their data center structures. They provide power supplies, cooling and security.
* Cloud data centers: these data centers are also monitored by a third-party provider, allowing customers to access resources via the internet.

A data center has various components:

* The facility: A data center must obviously have space to handle IT equipment, with the ability to generate a vast amount of energy. Moreover, the facility must have a specific temperature and humidity to allow a good functioning of the data center.
* Servers: These are the computational engines responsible for executing tasks and processing data within the data center. They come in different forms, such as rack-mounted servers, blade servers, or modular servers, and play a crucial role in fulfilling the computing needs of various applications and services.
* Storage: Data centers house vast amounts of information, requiring robust storage systems to store and retrieve data efficiently. Storage technologies such as hard disk drives (HDDs), solid-state drives (SSDs), and network-attached storage (NAS) devices are used to accommodate the data storage requirements of diverse workloads.
* Networking: Seamless communication and connectivity are vital within a data center. Networking infrastructure ensures smooth data flow between servers, storage systems, and external networks. Switches, routers, and cables form the backbone of data center networking, facilitating high-speed data transmission and facilitating efficient intercommunication between various components.
* Power Infrastructure: Data centers demand a significant power supply to sustain their operations. Power infrastructure includes generators, Uninterruptible Power Supply (UPS) systems, and power distribution units. These components ensure a stable power supply, backup power during outages, and efficient power distribution to support the computing and cooling needs of the data center.

We also have to consider the aspect of security. It is important to have trained personnel to ensure the security of the site. Moreover, with the vast amount of servers in a data center, it is crucial to have an advanced system of cybersecurity to prevent potential cyberattacks. Data centers also have to feature redundant connections to different service providers, to eliminate single points of failure.

A crucial aspect in data centers is also energy efficiency. With the rising concerns about environmental sustainability, energy efficiency has become a critical aspect of data center architecture. Data centers consume significant amounts of electricity, and optimizing energy usage is essential to reduce operating costs and minimize their carbon footprint. Efficient cooling mechanisms, such as hot and cold aisle containment, precision cooling, and free cooling, help maintain suitable temperatures within the data center while minimizing energy consumption. Power management strategies, such as server virtualization, power capping, and dynamic frequency scaling, also contribute to energy efficiency. Data centers consume about 1000 kWh per square meter. This is about ten times the power consumption of a typical American home. This important energy consumption comes from the fact that a lot of energy is necessary in order to maintain and cool down the components.

Data centers in the world consume about 3 percent of the global electric supply and account for about 2 percent of total GHG emissions.

II) Characterization of DCA

When it comes to DCA, there are many things to consider to characterize it.

Scalability: one of the fundamental characteristics of data center architecture is scalability. As data demands continue to grow exponentially, it becomes crucial for data centers to scale their resources effectively. Scalability ensures that the infrastructure can accommodate increasing workloads, adapt to evolving technology, and meet the ever-expanding needs of users and businesses. Scaling can occur vertically, by adding more resources to existing servers, or horizontally, by adding more servers to distribute the workload.

Reliability: A data center must be reliable, which first means that they should be available at all times. A small downtime can lead to a significant amount of financial loss. It is estimated that an hour of downtime can lead to an average of $140,000 and $540,000 per hour, depending on the organization. It can also affect the reputation of the organization and even worse, it can result in some data loss.

Cost: The conception of the architecture of a data center should consider the costs associated with its construction. Nowadays, a data center costs between 10 and 12 millions dollars per megawatt to construct.

Different types of architecture exist for data centers.

* Monolithic architecture: this architecture has been the foundation of data centers for many years. All components, including servers, storage, and networking, are tightly integrated into a single unit.
* Three-tier or multi-tier model: this data center architecture consists of switches and routers organized into three layers:
  + The access layer: It is responsible for connecting end devices such as servers, storage devices, and end-user devices to the network. It acts as an interface between the devices and the rest of the data center infrastructure. It provides connectivity and access control for end devices and implements security policies / enforces access controls. It also handles local VLANs (Virtual Local Area Networks) and subnets and often includes top-of-rack (ToR) switches, which connect individual servers and devices directly to the network.
  + The aggregation layer: The aggregation layer sits between the access layer and the core layer and serves as a bridge connecting the two. Its primary purpose is to aggregate traffic from multiple access layer switches and route it efficiently to the core layer. This layer aggregates and consolidates network traffic from access layer switches and implements policies for Quality of Service (QoS) and traffic prioritization.It also provides redundancy and high availability through link aggregation and load balancing. It often includes distribution switches that connect to multiple access layer switches and provide connectivity to the core layer.
  + The core layer: The core layer forms the backbone of the data center architecture, responsible for high-speed and reliable transport of data between different parts of the network. It provides fast and efficient routing of data across the data center network, and ensures high availability and fault tolerance through redundant connections and protocols. It also facilitates high-speed data transfers and low-latency communications and typically consists of high-performance routers or switches capable of handling large volumes of traffic
* Mesh point of delivery architecture: It is a network architecture that aims to improve the distribution of content and services by deploying content delivery infrastructure at multiple strategic locations. In this architecture, content is stored and delivered from various distributed points, known as mesh nodes or mesh PoDs, rather than relying on a central point of delivery.

It contains multiple leaf switches interconnected within the PoDs. It is a repeatable design pattern and its components maximize the modularity, scalability, and manageability of data centers. Besides, this architecture can realize the efficient connection between multiple PoDs and a super-spine tier. Therefore, data center managers can add new data center architecture to their existing three-tier topology easily for the low-latency data flow of new cloud applications.

Mesh PoD architecture utilizes dynamic routing mechanisms to determine the optimal path for content delivery. These mechanisms consider factors such as network congestion, latency, and proximity to select the most efficient path to serve content requests.

* The Super Spine Mesh architecture is a network architecture design that combines the benefits of a spine-leaf topology with the scalability and redundancy of a mesh topology. It is specifically designed to meet the demands of large-scale data centers and cloud environments. This architecture provides high performance, low latency, and efficient communication between network devices.

The Super Spine Mesh architecture is based on a spine-leaf topology. It consists of two main layers: the spine layer and the leaf layer. The spine layer comprises a set of high-performance switches, often referred to as spine switches, which are interconnected. The leaf layer consists of multiple leaf switches that are connected to the spine switches. Within the spine layer, the Super Spine Mesh architecture utilizes a mesh connectivity pattern. This means that each spine switch is directly connected to every other spine switch in a full-mesh fashion. This mesh connectivity provides multiple paths for traffic to flow between spine switches, enhancing redundancy and fault tolerance.

[A spine-leaf architecture is data center network topology that consists of two switching layers—a spine and leaf. The leaf layer consists of access switches that aggregate traffic from servers and connect directly into the spine or network core. Spine switches interconnect all leaf switches in a full-mesh topology.]

* The Clos (or Spine-Leaf) topology and the Super Spine Mesh topology are similar in that they both aim to provide high capacity, scalability, and redundancy. However, the key difference lies in the level of hierarchy and how the traffic flows within these architectures.
* Clos (Spine-Leaf) Topology:
* The Clos topology is based on a three-stage design: Leaf (edge), Spine, and again Leaf. Every Leaf switch is connected to every Spine switch, providing multiple equal-cost paths for traffic to traverse. This structure is inherently scalable and resilient as additional Spine or Leaf switches can be added to scale the network.
* Super Spine Mesh Topology:
* Super Spine Mesh topology builds upon the traditional Spine-Leaf model by introducing an additional layer - the Super Spine layer. This new layer sits above the Spine layer and connects multiple Spine-Leaf 'pods'. It offers more flexibility and scalability to support larger, more complex data center environments. In this model, the Leaf switches connect to Spine switches, which in turn connect to Super Spine switches, forming a mesh-like architecture.
* In summary, the main structural difference is that Super Spine Mesh introduces an additional layer to the Spine-Leaf design to further scale the network and provide additional paths for traffic, which can be particularly beneficial in very large or geographically distributed data centers.

**III) Comparison of Data Center Architectures (Jan)**

* VLANs (L2), Subnets (L3 - IP), Broadcast (ARP) => Flooding Because of redundancy which introduces loops => STP
* Here we are going to describe a transition from three layer architecture to Spine - Leaf DC architecture.
* STP + VLAN Trunking + VRRP + OSPF
* STP - spanning tree protocol - prevents L2 loops in frame forwarding
  + The problem with STP and vLANs:
    - It spans a big area as we scale our data center.
    - The packets have to go to the router and back down when we use VLANs because **subnets use different IP ranges**-> we should be using l3 switches
* **VLAN Trunking -** splits the switch topology into **multiple broadcast domain:**
  + *if we connect two default setting switches, as shown in figure 1, any broadcast frame received by either switch is forwarded to the other one and then out all its ports. Therefore a broadcast domain is not limited to one switch only, it includes all devices that get a copy of any broadcast frame, even if they are connected to other switches. If we scale this logic to a LAN with tens of interconnected switches, we could have a broadcast domain consisting of hundred of end devices. This at some point can contest the network*
  + **VLAN on multiple switches -** using virtual LANs we can split the switch topology into **multiple broadcast domains:**

1. configuring ports 1 through 4 of both switches to VLAN 10 and ports 5 through 9 to VLAN20. Although it is a valid design and it works, it simply does not scale very well. It requires a physical link between the switches per VLAN. => **DOES NOT SCALE**
2. In order to overcome this scaling limitation, we can use another Ethernet technology called **VLAN trunking**. It creates only one link between the switches that support as many VLAN as needed. At the same time, it also keeps the VLAN traffic separate, so frames from VLAN 20 won't go to devices in VLAN 10 and vice-versa.

* **VLAN Tagging** - VLAN trunking allows switches to forwards frames from different VLANs over a single link called trunk. This is done by adding an additional header information called tag to the Ethernet frame. The process of adding this small header is called VLAN tagging. When the ethernet frames are sent between the switches over the trunk link, they are tagged with VLAN header. When the receiving switch gets them, removes the VLAN tag and sends them to the clients in the VLAN, the frames are untagged.
* The **Virtual Router Redundancy Protocol** (**VRRP**) is a computer [networking protocol](https://en.wikipedia.org/wiki/Networking_protocol) that provides for **automatic assignment of available** [**Internet Protocol**](https://en.wikipedia.org/wiki/Internet_Protocol) **(IP) routers to participating** [**hosts**](https://en.wikipedia.org/wiki/Host_(network)). This increases the **availability and reliabilit**y of [routing](https://en.wikipedia.org/wiki/Routing) paths via a**utomatic** [**default gateway**](https://en.wikipedia.org/wiki/Default_gateway) **selection**s on an IP [subnetwork](https://en.wikipedia.org/wiki/Subnetwork).  
  The protocol achieves this by the c**reation of virtual router**s, which are an **abstract representation of multiple routers, i.e. primary/active and secondary/Standby** [routers](https://en.wikipedia.org/wiki/Router_(computing)), **acting as a group**. The virtual router is assigned to act as a default gateway of participating hosts, instead of a physical router. If the physical router that is routing packets on behalf of the virtual router **fails**, another physical router is s**elected to automatically replace it.**
* **So our L3 switches are running STP, VRRP, and OSPF.** This is hard to monitor, configure and troubleshoot, very complex and **does not scale well**.
* Transition to OSPF -> only one protocol
  + STP -> VXLAN
  + VRRP -> Anycast - **enables redundancy and multiple paths**
* **Virtual Extensible LAN** (**VXLAN**) is a [network virtualization](https://en.wikipedia.org/wiki/Network_virtualization) technology that attempts to address the [scalability](https://en.wikipedia.org/wiki/Scalability) problems associated with large [cloud computing](https://en.wikipedia.org/wiki/Cloud_computing) deployments. It uses a [VLAN](https://en.wikipedia.org/wiki/VLAN)-like encapsulation technique to encapsulate [OSI](https://en.wikipedia.org/wiki/OSI_model) [layer 2](https://en.wikipedia.org/wiki/Layer_2) [Ethernet frames](https://en.wikipedia.org/wiki/Ethernet_frame) within [layer 4](https://en.wikipedia.org/wiki/Layer_4) [UDP](https://en.wikipedia.org/wiki/User_Datagram_Protocol) datagrams
  + **Basically encapsulates our L2 frame into L3 packet to achieve VLAN trunking - r**equires VTEP - virtual tunneling endpoints
* **There is no more one uplink, there are multiple uplinks all the time because of anycast.**
* **To keep costs down, they built their switches from standard commodity switch chips. They found that the features of traditional data center switches that were used in part to justify their high costs—such as decentralized network routing and protocols to manage support of arbitrary deployment scenarios—were unnecessary in a WSC because the network topology could be planned in advance. Google instead used centralized control that relied on a common configuration that was copied to all data center switches. The modular hardware design and robust software control allowed these switches to be used both for inside the WSC and for wide area networks between WSCs.**
* When comparing the architectures, the choice depends on the specific requirements of the network and data center. Factors to consider include **scalability needs, flexibility, fault tolerance, latency requirements, management complexity, and cost consideration**s. Each architecture has its strengths and weaknesses, and organizations should evaluate which one best aligns with their specific needs and priorities.

**IV) Disaggregation in Data Centers (Jan)**

Disaggregation in data centers is the s**eparation of integrated components like compute, storage, and networking into independent units** to enhance **flexibility, scalability, and resource utilizatio**n. Each standalone unit can be independently managed and scaled, allowing for more efficient resource allocation, cost reduction, and adaptation to changing workload demands.

Key disaggregation components include c**ompute disaggregation, which decouples compute resources like CPUs and memory from specific servers for dynamic allocation**; s**torage disaggregation**, which separates storage resources from servers and consolidates them into a **shared pool for more efficient utilization and centralized management**; and n**etwork disaggregation**,

Benefits of disaggregation include improved **scalability and flexibility**, allowing data centers to **adjust resources according to workload requirements**. It facilitates better resource utilization by **pooling compute, storage, and networking resources, leading to cost savings**. It also enhances **agility,** facilitating faster response to workload changes and faster application deployment. Furthermore, disaggregation promotes vendor independence by separating hardware and software components, avoiding vendor lock-in.

However, disaggregation introduces challenges, including **complexity in resource management, potential performance overhead, and the need for robust network connectivity**. Therefore, careful planning, implementation, and management are vital for successful disaggregation in data centers.

**V) Case Studies and Real-World Examples (Jan) - Google WSC**

1. Power Distribution:
   1. To prevent the whole WSC from going offline if power is lost, WSCs have their version of an uninterruptible power supply (UPS), just as most servers do in conventional data centers. Also, diesel generators are connected to the power distribution system at this level to provide power in the event of an issue with the utility power.
   2. The power is distributed in a hierarchy in a WSC with each level of the hierarchy (utility tower, on-site substations, building, racks) corresponding to a distinct failure and maintenance unit.
   3. There is little opportunity left to optimize power efficiency here.
2. Cooling:
   1. In a Google WSC, the server fans work synergistically with dozens of giant fans in the room to ensure airflow for the whole room (Figure 6.28). This division of labor means the small server fans use as little power as possible while delivering maximum performance at the worst-case power and ambient conditions.
   2. Careful airflow planning and the use of hot and cold aisles help maintain temperature separation.
   3. Water-side economization and evaporative cooling towers are used to leverage the local environment for cooling.
3. Racks:
   1. A WSC consists of multiple arrays (which Google calls clusters). Although arrays vary in size, some have one to two dozen rows with each row holding two to three dozen racks. The 20 slots shown the middle of the rack in Figure 6.30 hold the servers. Depending on their width, up to four servers can be placed in a single tray
   2. Racks in a Google data center contain multiple servers, with power converters to provide the appropriate voltage for server boards.
   3. Small batteries are placed at the bottom of each rack to provide backup power during utility power disruptions.
4. Networking in a Google WSC:
   1. Google utilizes a Clos network topology, named after its inventor, to build its data center network. This multistage network architecture incorporates low port-count switches, providing fault tolerance, scalability, and increased network scale.
   2. The network design includes multiple stages to accommodate the growing scale, and the inherent redundancy ensures that failures of individual links have minimal impact on overall network capacity.
   3. Google's network switches are built using commodity switch chips, and centralized control is employed for network routing and management. Each switch receives a consistent copy of the current network topology, simplifying routing in a complex Clos network.
   4. The latest Google switch, Jupiter, utilizes 16x16 crossbars with 40 Gbps links. The Top of Rack switch incorporates four of these chips, enabling connections of up to 48 40-Gbps links to servers and 16 40-Gbps links to the network fabric. The oversubscription ratio is kept at an impressive 3:1.
   5. Middle blocks housing the switch chips are interconnected with 256 10-Gbps links for Top of Rack connectivity and 64 40-Gbps links to connect with the rest of the network fabric through the spine. Each aggregation block is connected to the spine block via 512 40-Gbps links. At the largest scale, Google employs 64 aggregation blocks for dual redundant links, achieving a remarkable bisection bandwidth of 1.3 Pbit per second.
5. Servers in a Google WSC:

The example server in Figure 6.34 has two sockets, each containing an 18-core Intel Haswell processor running at 2.3 GHz (see Section 5.8). The photo shows 16 DIMMs, and these servers are typically deployed with 256 GB total of

DDR3-1600 DRAM. The Haswell memory hierarchy has two 32 KiB L1 caches,

a 256 KiB L2 cache, and 2.5 MiB of L3 cache per core, resulting in a 45 MiB L3

cache. The peak power of

the baseline is about 150 watts

The text highlights Google's efforts to improve energy efficiency and reduce environmental impact in its data centers through techniques such as higher operating temperatures, efficient cooling infrastructure, utilization of local environmental conditions, and monitoring for operational efficiency.

Overall, the information provides insights into the power, cooling, networking, racks, and servers within a Google data center, showcasing the company's commitment to energy efficiency and sustainability.

VI) Future trends and challenges

The field of data centers is subject to a constant evolution, and there are a lot of challenges for the future. The architecture of data centers has to evolve with the evolution of our modern society and the new challenges that come with this evolution

* Supply and demand meets constraints and delays: the increase of the demand is exponential, and this could make the cost of data centers grow extremely fast. Companies may face some budget restrictions in the future while the demand will not stop growing. It will be crucial for engineers to find new architecture to respond to the new challenges especially for hyperscale architecture.  
  The emergence of AI and deep learning also requires a new and massive use of data centers, especially for specialized HPC platforms.
* Environmental challenges: Data centers account for around 2% of all global carbon emissions. That is about the same as the airline industry. To put this into perspective, [the global carbon emissions for the chemical and petrochemical industry are slightly higher at 3.6%.](https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions) It should rise up to 14% by 2040. This is worrying, and we need to find architectures that can handle a better energy consumption

Conclusion:

Data centers are very present in our modern world, and we saw that there exists multiple types of architecture, each with their pros and cons. Each architecture has its own cost, scalability, efficiency… and each one has to be precisely monitored in order for the enterprises to work. But the hardest part is yet to come: new challenges will arrive and future engineers will have to come up with new ideas.  
  
  
Jan:

* A higher target for cold air temperature helps put the facility more often within

the range that can be sustained by cooling towers, which are more energy-

efficient than traditional chillers.

* Adding large fans for entire rooms to work in concert with the small fans of the servers to reduce energy while satisfying worst-case scenarios.
* Averaging the cooling per server to whole racks of servers by deploying the cooling coils per row to accommodate warmer and cooler racks.
* Deploying extensive monitoring hardware and software to measure actual PUE versus designed PUE improves operational efficiency.
* Operating more servers than the worst-case scenario for the power distribution

system would suggest. It is safe since it’s statistically improbable that thou-

sands of servers would all be highly busy simultaneously as long as there is

a monitoring system to off-load work in the unlikely case that they did (Fan

et al., 2007; Ranganathan et al., 2006). PUE improves because the facility is

operating closer to its fully designed capacity, where it is at its most efficient

because the servers and cooling systems are not energy-proportional. Such

increased utilization reduces demand for new servers and new WSCs