Network Fundamentals for Cloud

BITS Pilani
Pilani Campus

Nishit Narang
WILPD-CSIS

innovate achieve lead

BITS Pilani
Pilani Campus

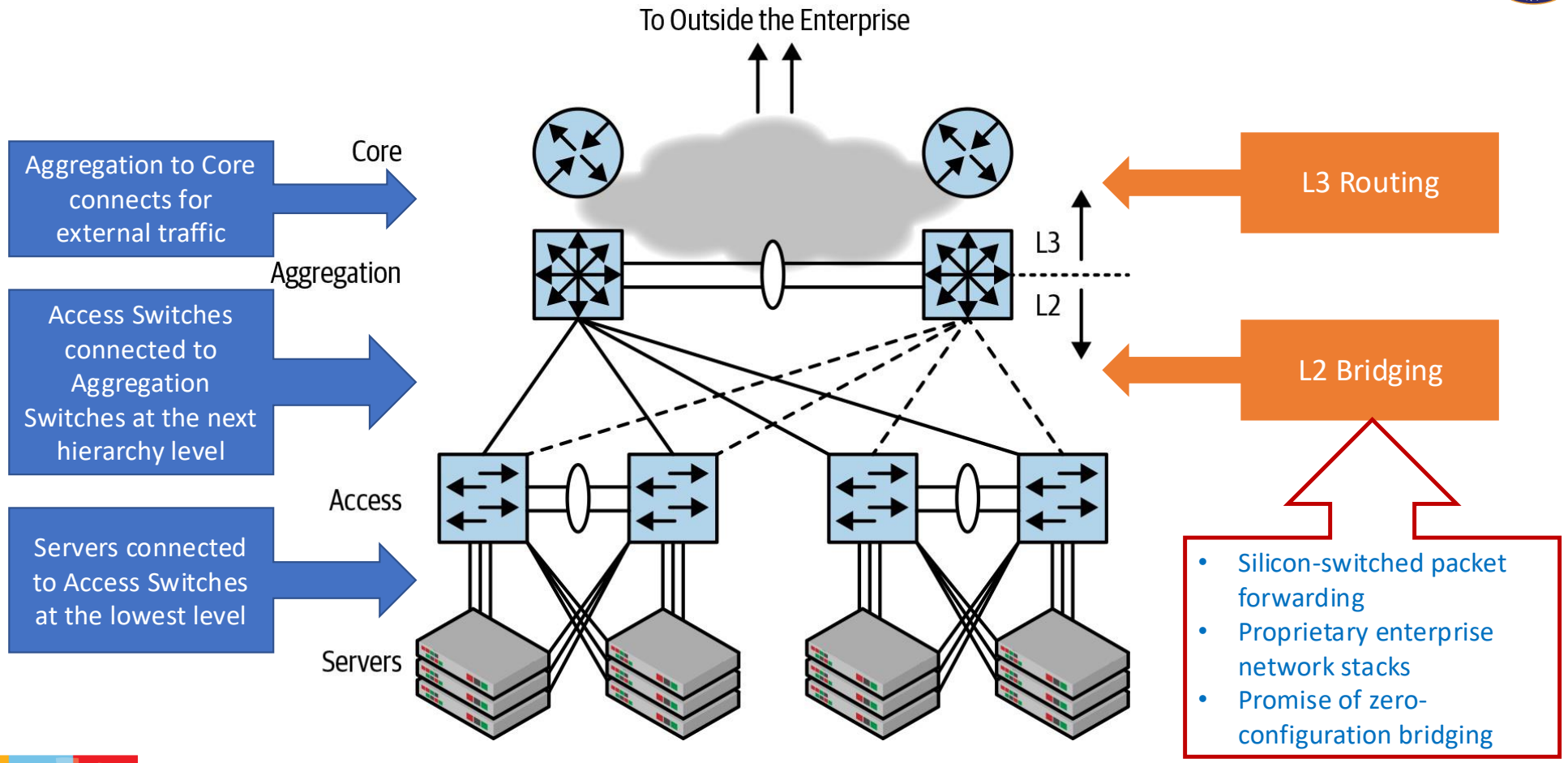
CC ZG503: Network Fundamentals for Cloud

Lecture No. 10: Data Center Networks (Contd.)

RECAP: DCN Evolution

- Traditional network topology
  - *access-aggregation-core*
  - Became prominent around year 2000
  - Considered fast, cheap and easy to administer
  - well suited to the north-south traffic pattern of client-server application architecture
  - Not suited, however, to the server-server traffic pattern of DCNs
- Modern DCN topologies:
  - The structure of the new world is the Clos topology (*named after one of its inventors, Charles Clos*)
  - Basic Clos topology is also called the ***leaf-spine*** topology
  - ***Fat Tree*** topology, a special instance of the Clos topology is extremely popular

RECAP:Traditional Network Topology



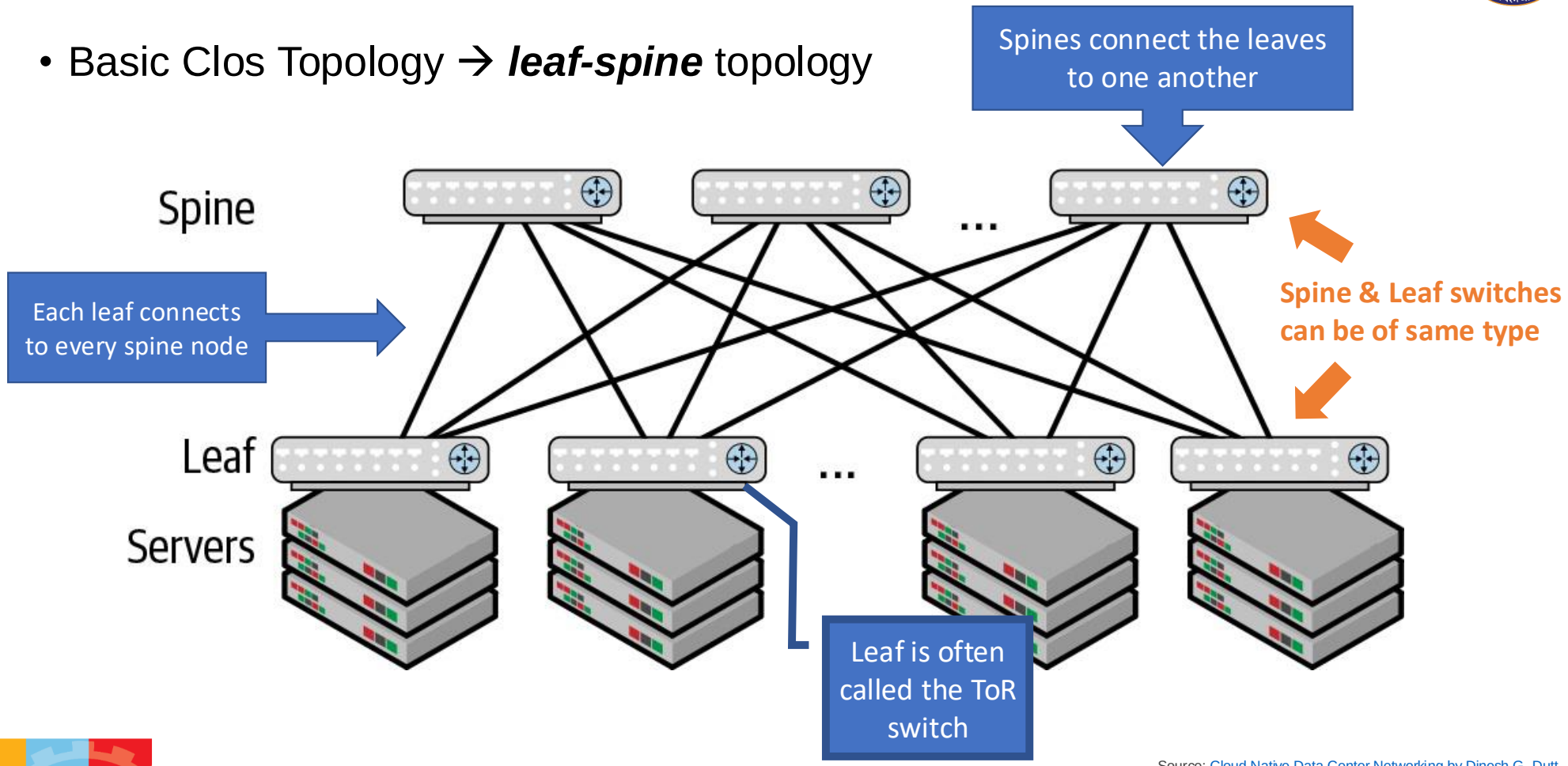
Source: Cloud Native Data Center Networking by Dinesh G. Dutt

Challenges with Acc-Agg-Core Topologies

- Lack of scalability for DCN traffic patterns / applications
  - Flooding → ***flood-and-learn*** model of self-learning bridges doesn't scale!
  - VLAN limitations → 12-bit VLAN ID => 4096 VLANs, a paltry value at the scale of the cloud
  - Burden on Aggregation switches (2) to respond to all ARP messages
  - STP limitations → more east-west traffic => more aggregation switches. Unpredictable / unusable topologies emerged due to link/node failures.
- Complexity
  - Unless the access-agg-core network is carefully designed, congestion can quite easily occur in such networks → over-subscription of network bandwidth
- Failure Impact
  - access-agg-core model is prone to very coarse-grained failures; In other words, failures with large blast radiuses.
    - For example, the failure of a single link halves the available bandwidth
- Inflexibility: It is not possible to have the same VLAN be present across two different pairs of aggregate switches

Clos Network Topology

- Basic Clos Topology → ***leaf-spine*** topology



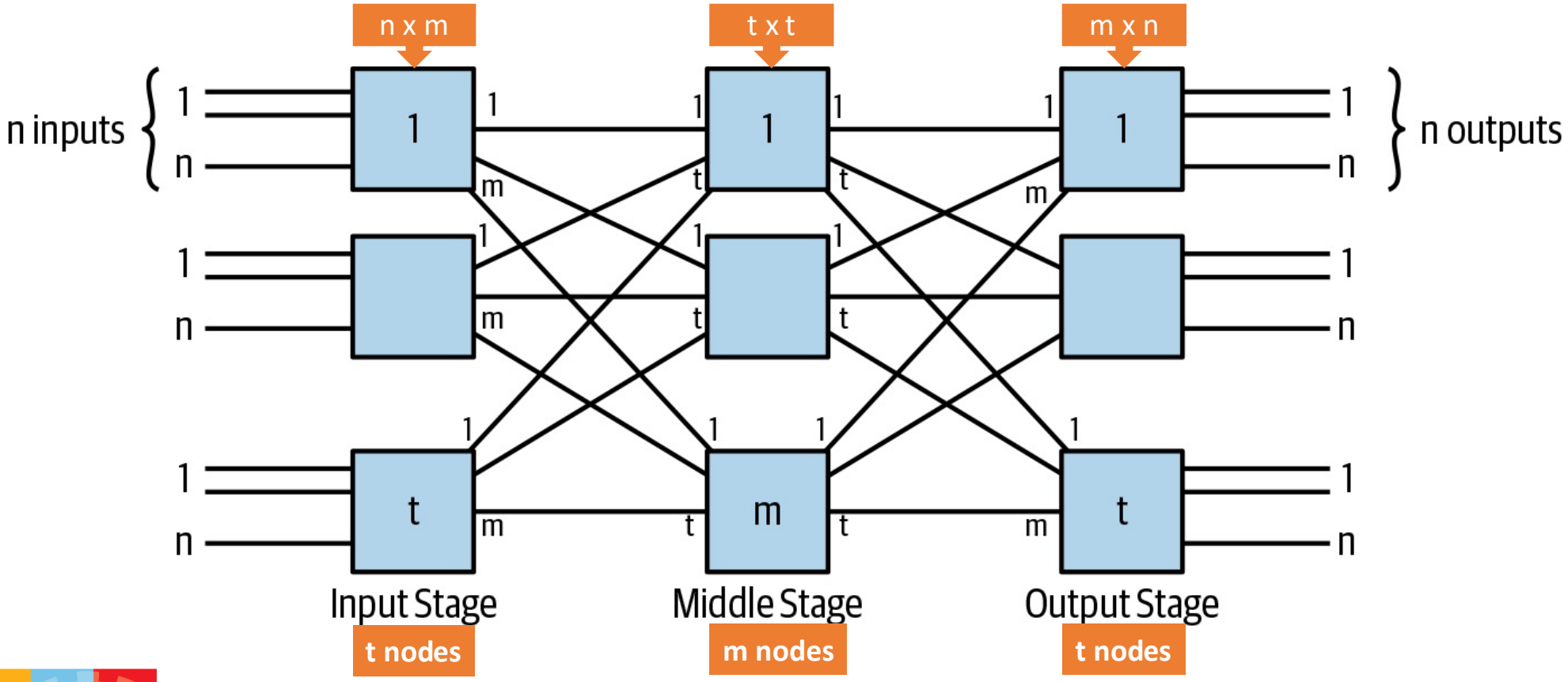
Source: Cloud Native Data Center Networking by Dinesh G. Dutt

Benefits of the *Leaf-Spine* Topology

- Ability to use homogeneous switching equipment
- Redundancy → more than two paths between any two servers
- High-capacity → Adding spines increases the capacity between leaf nodes
- Simplicity
  - Spines only connect leaves; no other functionality (e.g. ARP etc) [[unlike Aggregation switches]]
  - All network functions are supported by edge devices
  - Routing as the interconnect model using ECMP (bridging only within rack. Or, using VXLANs across racks)
- Scalability → the Clos topology is a scale-out architecture!
  - Adding leaves and servers increases the amount of work performed by the network
  - Adding spines increases the bandwidth between the edges

Classic (three-stage) Clos Topology

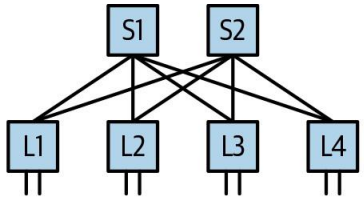
If: m=n=t AND Flip the Output Stage onto the Input Stage → *Leaf-Spine* Topology



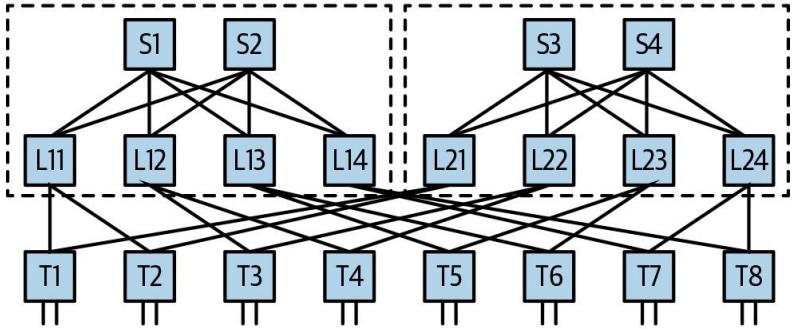
Source: Cloud Native Data Center Networking by Dinesh G. Dutt

Scaling Clos Topology

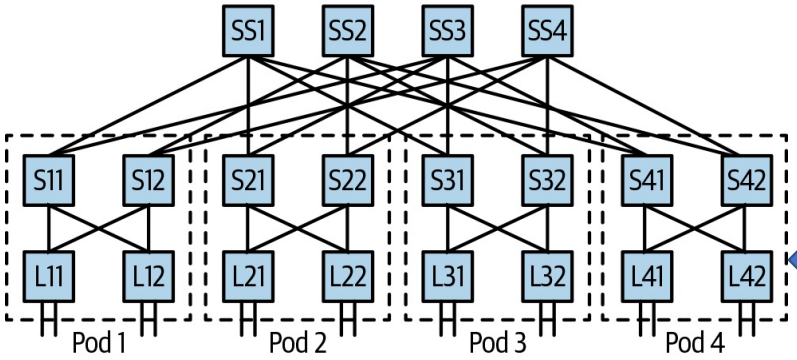
Examples with four-port switches



(a) Two-tier Clos using four-port switches



(b) Virtual chassis-based three-tier Clos using four-port switches



(c) Pod-based three-tier Clos using four-port switches

Model popularized by **Facebook**

Model used by Microsoft and Amazon

Source: Cloud Native Data Center Networking by Dinesh G. Dutt

DCN Design Aspects

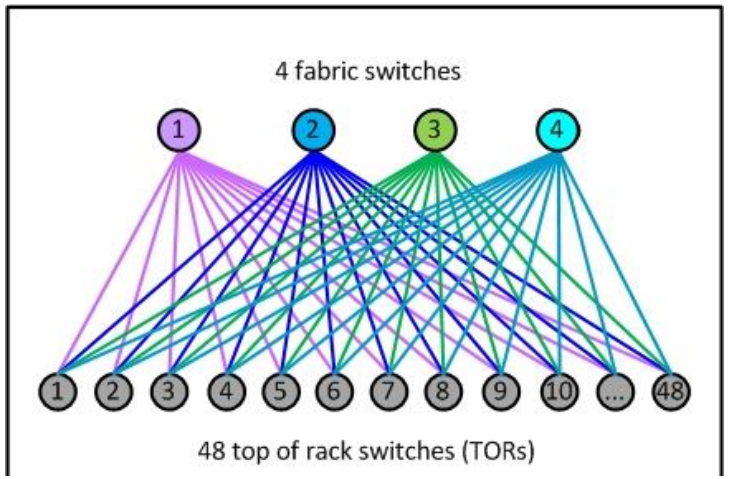
- Choice of Topology
- Oversubscription of Bandwidth
- Multipath Routing
- Overall Cost

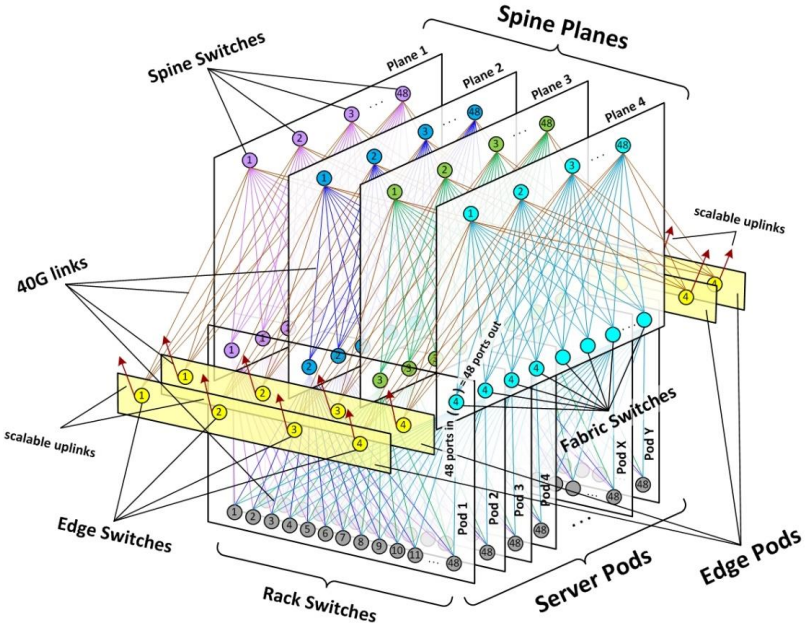
Case Studies

- A Scalable, Commodity Data Center Network Architecture (Research Paper)
  - Introduction (Sec 1: DC Applications and their traffic patterns
  - Section 2.1: DC Network Topologies and Bandwidth Oversubscription
  - Section 2.2: Clos Networks and Fat-Tree Topology

Case Studies

- [Introducing data center fabric, the next-generation Facebook data center network - Engineering at Meta]
  - Facebook is a good example application to explain cloud application traffic patterns and networking needs
  - The example shows performance limitations encountered by FB in M2M traffic when using cluster-design. And the migration to the next-generation fabric design to overcome this challenge. This introduces modularity, leading to gradual scalability.
  - Role of BGP4 (distributed routing) alongside a centralized BGP controller (for centralized override). FB calls this hybrid approach as "***distributed control, centralized override***".
  - High-capacity 40G links connecting fabric switches, TOR switches and Spine switches. Rack servers connected to TOR via 10G links.





Source: [Introducing data center fabric, the next-generation Facebook data center network - Engineering at Meta]

Case Studies

Google DC Topologies

- Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network
  - CLOS Topologies (Leaf'n'spine) in data centers
  - Google datacenter networks run at dozens of sites across the planet, scaling in capacity by 100x over ten years to more than 1Pbps of bisection bandwidth.
  - Much of the general, but complex, decentralized network routing and management protocols supporting arbitrary deployment scenarios are overkill for single-operator, pre-planned datacenter networks. Here, a centralized control and management mechanism is discussed, based on a global configuration that is pushed to all datacenter switches.
    - Granular control over ECMP tables with proprietary, scalable in-house IGP
    - Use standard BGP between Cluster Border Routers and external vendor gear
    - Proprietary Neighbor Discovery (ND) protocol for online liveness and peer correctness checking - used for correcting cabling errors, one of the key challenges in a large data center
  - Data Center Challenges (viz. Fabric Congestion and Outages) discussed in section 6.

DCN Networking Technology Evolution

- xSTP technologies to eliminate loops in the L2 Network broadcast domain
  - STP had many issues (as summarized earlier).
  - Led to the evolution of Virtual Chassis Technologies
- Virtual Chassis Technology
  - Virtual chassis technologies implement N:1 virtualization
  - Integrates the control planes of multiple devices to form a unified logical device
    - I.e. Different physical devices share the same control plane, which is equivalent to creating a cluster for physical network devices
  - This logical device has a unified management IP address and also works as one node in various Layer 2 and Layer 3 protocols
    - Link aggregation allows this logical device to connect to each physical or logical node at the edge through only one logical link, solving the problem of dual-homing of terminals that inevitably causes loops on the network.
    - Therefore, the network topology after the integration is loop-free for xSTP, which indirectly avoids problems of xSTP
    - Master election and a master/standby switchover can also be performed in the cluster

Several virtual chassis technologies exist in the industry, such as Cisco's VSS, Huawei's CSS, etc

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

DCN Networking Technology Evolution (contd.)

- Disadvantages of Virtual Chassis Technology
  - Limited scalability: the scalability of any virtual chassis technology is limited by the performance of the master switch, as it provides the control plane for the entire virtual chassis system
  - Reliability: Because the control plane is on the master switch, packet loss may last for a long time or the entire system may stop running if the master switch fails
  - Upgrade Challenges: Control plane integration also makes it difficult to upgrade a virtual chassis system. Common restart and upgrade operations cause interruption of the control plane, resulting in packet loss for a long time.
  - Bandwidth waste: Dedicated links in a virtual chassis system are used for status exchange and data forwarding between devices

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

DCN Networking Technology Evolution (contd.)

- Layer 2 Multi-Pathing (L2MP) Technologies
  - L2MP technologies attempt to address the below twin challenges of xSTP and Virtual Chassis technologies:
    - They cannot support large DCNs with massive amounts of data.
    - Their link utilization is low.
  - It is recommended that link- state routing protocols widely used on Layer 3 networks be employed
  - These protocols not only support a large number of devices, but they are also loop-free and have high link utilization
    - Example: OSPF and IS-IS → they support ECMP load balancing and use the Shortest Path First (SPF) algorithm
  - The basic principle of L2MP technologies is to introduce mechanisms of routing technologies used on Layer 3 networks to Layer 2 networks

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

DCN Networking Technology Evolution (contd.)

- Layer 2 Multi-Pathing (L2MP) Technologies (contd.)
  - A link-state routing protocol usually requires that each node on a network be addressable
  - Each node uses the link-state routing protocol to calculate the network topology and then calculates the forwarding database based on the network topology
  - Therefore, L2MP technologies need to add an addressable identifier, which is similar to an IP address on an IP network, to each device on a network
  - TRILL is a standard L2MP protocol defined by the Internet Engineering Task Force (IETF) that became popular.

Source: Lei Zhang, Le Chen. Cloud Data Center Network Architectures and Technologies, CRC Press 2021

Thank You!