Network Fundamentals for Cloud

Nishit Narang
WILPD-CSIS

BITS Pilani
Pilani Campus

BITS Pilani
Pilani Campus

CC ZG503: Network Fundamentals for Cloud

Lecture No. 9: Data Center Networks (Contd.)

RECAP: Role of DCN

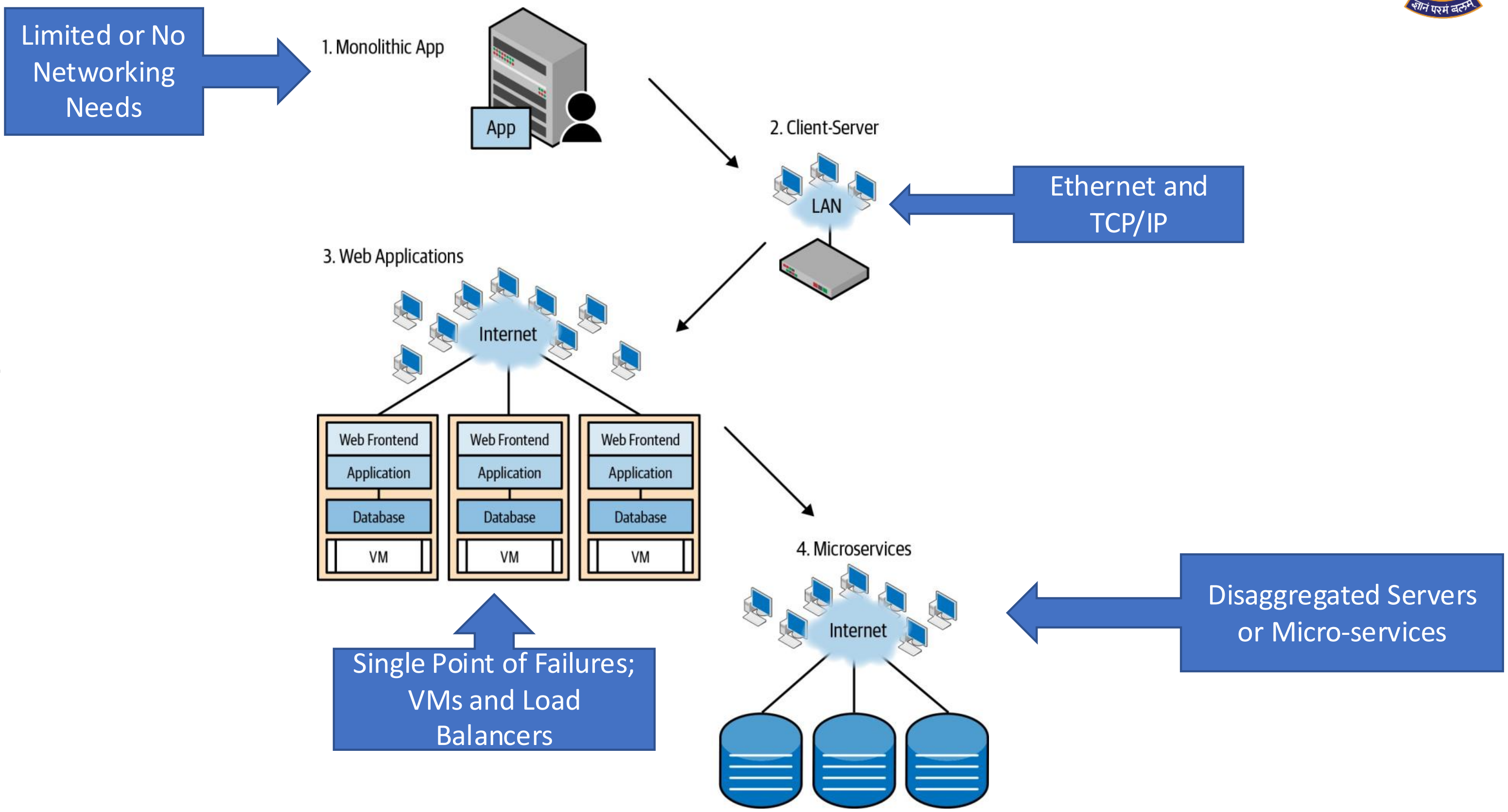
- Role of DC in realizing the “Cloud”
  - DC provides the ingredients for the Cloud
    - Compute (of different capacities & types)
    - Storage (of different capacities & types)
- Networking “Connects” the ingredients!!

- DCN → Data Center Network
  - The network that connects the assets within a Data Center

- Video
  - https://www.youtube.com/watch?v=avP5d16wEp0

RECAP: Application - Network Dance!

- A distributed application is in a dance with the network, with the application leading
- The story of the modern data center network begins when the network was caught flat-footed when the application began the dance to a different tune



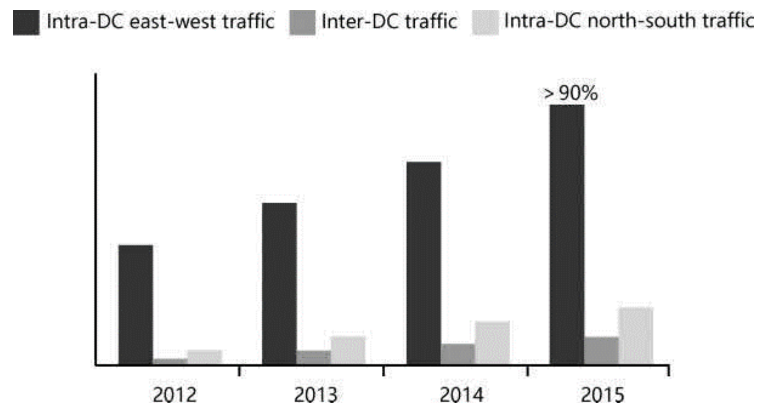
Cluster-based application architectures such as MapReduce have become prominent

Historic Shift from Client-Server Traffic to Server-Server Traffic

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

RECAP: DCN Challenges - Traffic

- Internal Traffic is BIG!!
  - In 2020, every minute there were
    - more than 1.6 million Google searches,
    - 260 million emails sent,
    - 47,000 apps downloaded,
    - 220,000 photos uploaded to Facebook, and
    - 660 million data packets transmitted…..



  - It is estimated that global DC IP traffic increases five-fold every year
- Big data requires wide pipes…..
  - East-West traffic accounts for more than 90% of total DC traffic

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

DCN Challenges – Network Faults and Capacity

- Need for Intelligent O&M and Network Fault Recovery
  - Driven by cloud-based DCs and Network Function Virtualization (NFV), the number of managed objects (MOs) in a cloud DCN is ten times greater than that of a legacy DCN
  - Network needs to detect dynamic VM migration and elastic scaling of applications, which results in frequent configuration changes and traffic surges
    - Example: LinkedIn data shows that the number of network faults saw an 18-fold increase from 2010 to 2015.
  - As network, computing, and storage boundaries are blurred, network faults become more difficult to locate and isolate
- Types of Network Faults:
  - Connection faults, such as a VM going offline unexpectedly or communication becoming intermittently interrupted.
  - Performance faults, such as network congestion during heavy loads.
  - Policy faults, such as unauthorized access and port scanning.
- Newer applications are demanding tight Network-I/O timings!
  - Also see next slide….

DCN Challenges – Network Faults and Capacity

- How to address Network Faults and Capacity Challenges?
  - Use of Self-healing networks
  - Use of Intelligent analysis engine (big data algorithms) to predict / detect / isolate network faults
  - Use of SDN and SDN Controller for simplified cloud network operations
  - Bandwidth Oversubscription (especially in core layer)

DCN Challenges – TCP Incast

- TCP incast is a recently identified network transport pathology that affects many-to-one communication patterns in datacenters
- Caused by a complex interplay between datacenter applications, the underlying switches, network topology, and TCP, which was originally designed for wide area networks

- The problem especially affects computing paradigms in which distributed processing cannot progress until all parallel threads in a stage complete
- Examples of such paradigms include distributed file systems, web search, advertisement selection, and other applications with partition or aggregation semantics

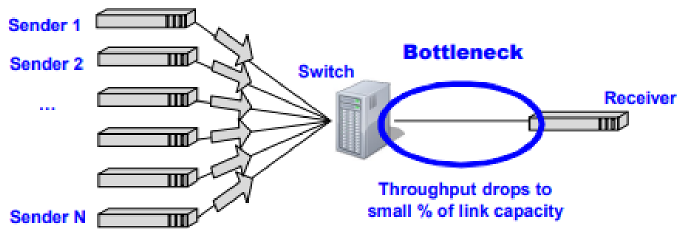


Figure 1. Simple setup to observe incast. The receiver requests k blocks of data from a set of N storage servers. Each block is striped across N storage servers. For each block request received, a server responds with a fixed amount of data. Clients do not request block $k+1$ until all the fragments of block k have been received.

Source: Understanding TCP Incast and Its Implications for Big Data Workloads by Yanpei Chen, Rean Griffith*, David Zats, Anthony D. Joseph, Randy Katz (University of California, Berkeley, *Vmware)

DCN Challenges – TCP Incast

- There have been many proposed solutions for TCP incast. Approaches include:
  - modifying TCP parameters or its congestion control algorithm,
  - optimizing application level data transfer patterns,
  - switch level modifications such as larger buffers or explicit congestion notification (ECN) capabilities, and
  - link layer mechanisms such as Ethernet congestion control.
- Application level solutions are the least intrusive to deploy, but require modifying each and every datacenter application
- Switch and link level solutions require modifying the underlying datacenter infrastructure, and are likely to be logistically feasible only during hardware upgrades
- TCP incast is fundamentally a transport layer problem, thus a solution at this level may be best
  - e.g. An existing solution is reducing the minimum length of TCP retransmission time out (RTO) from 200ms to 1ms

Source: Understanding TCP Incast and Its Implications for Big Data Workloads by Yanpei Chen, Rean Griffith*, David Zats, Anthony D. Joseph, Randy Katz (University of California, Berkeley, *Vmware)

DCN Challenges - Infrastructure

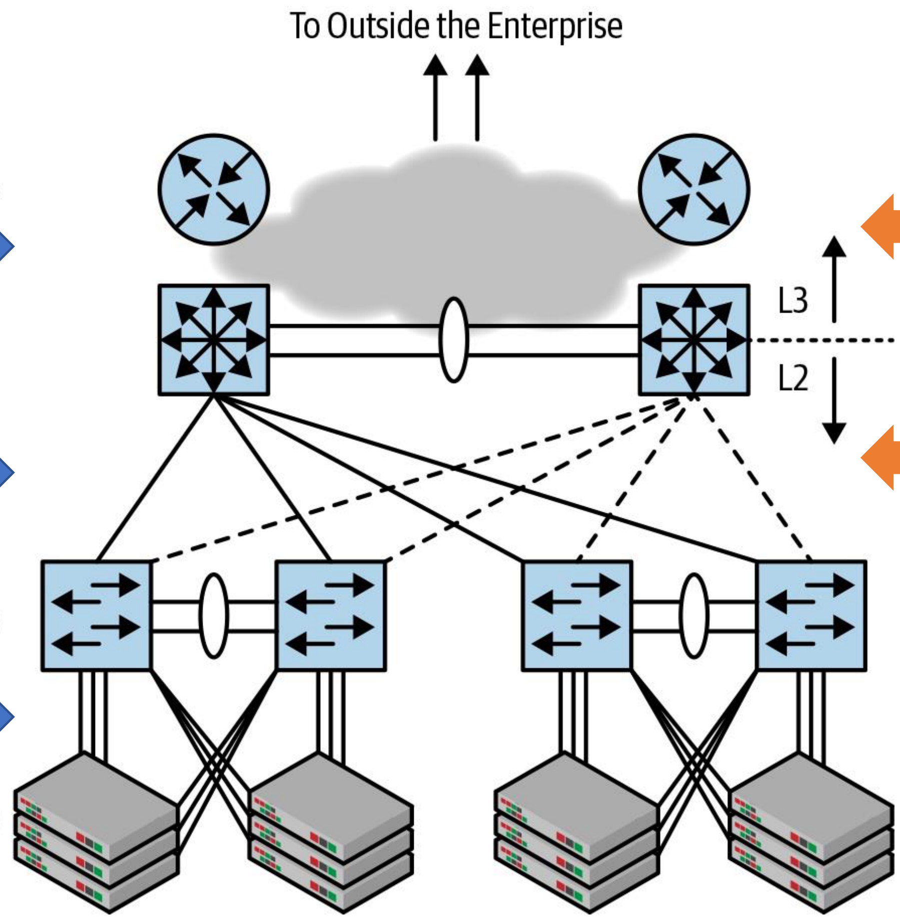
- DC Network Cost
- DC Cooling
- DC Cabling

DCN Evolution

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

Traditional Network Topology



Aggregation to Core connects for external traffic

Access Switches connected to Aggregation Switches at the next hierarchy level

Servers connected to Access Switches at the lowest level

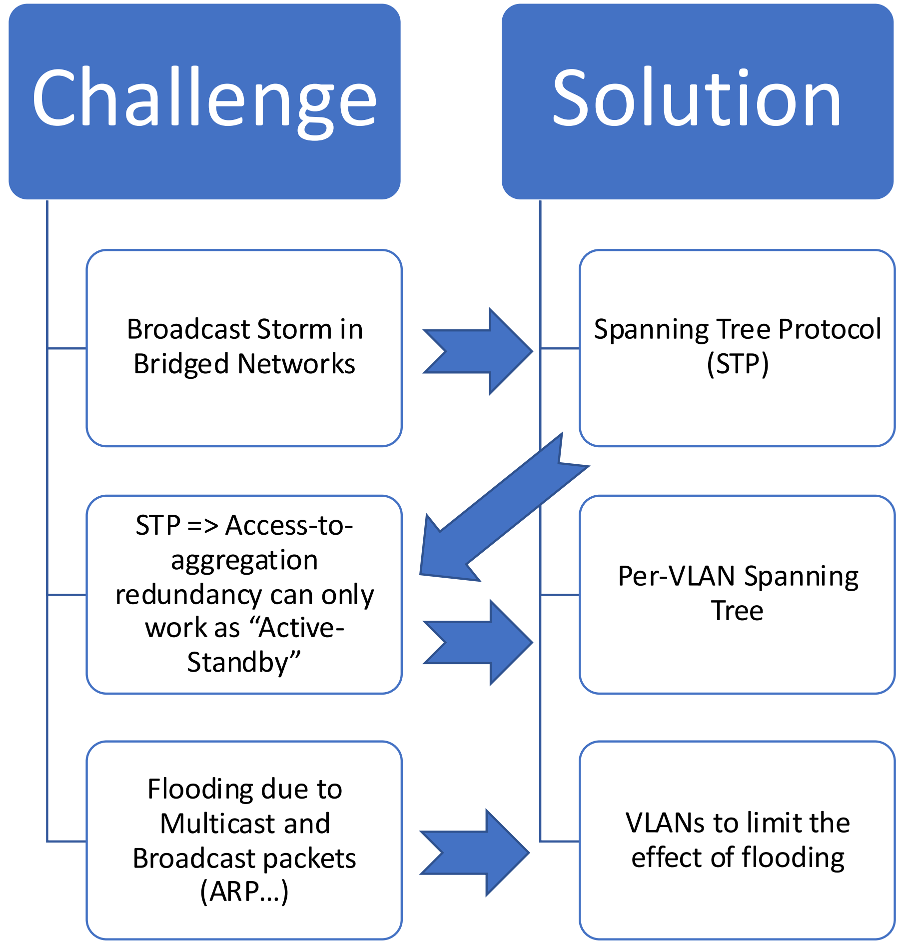
L3 Routing

L2 Bridging

- Silicon-switched packet forwarding
- Proprietary enterprise network stacks
- Promise of zero-configuration bridging

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

Challenges with Bridged Network Topologies

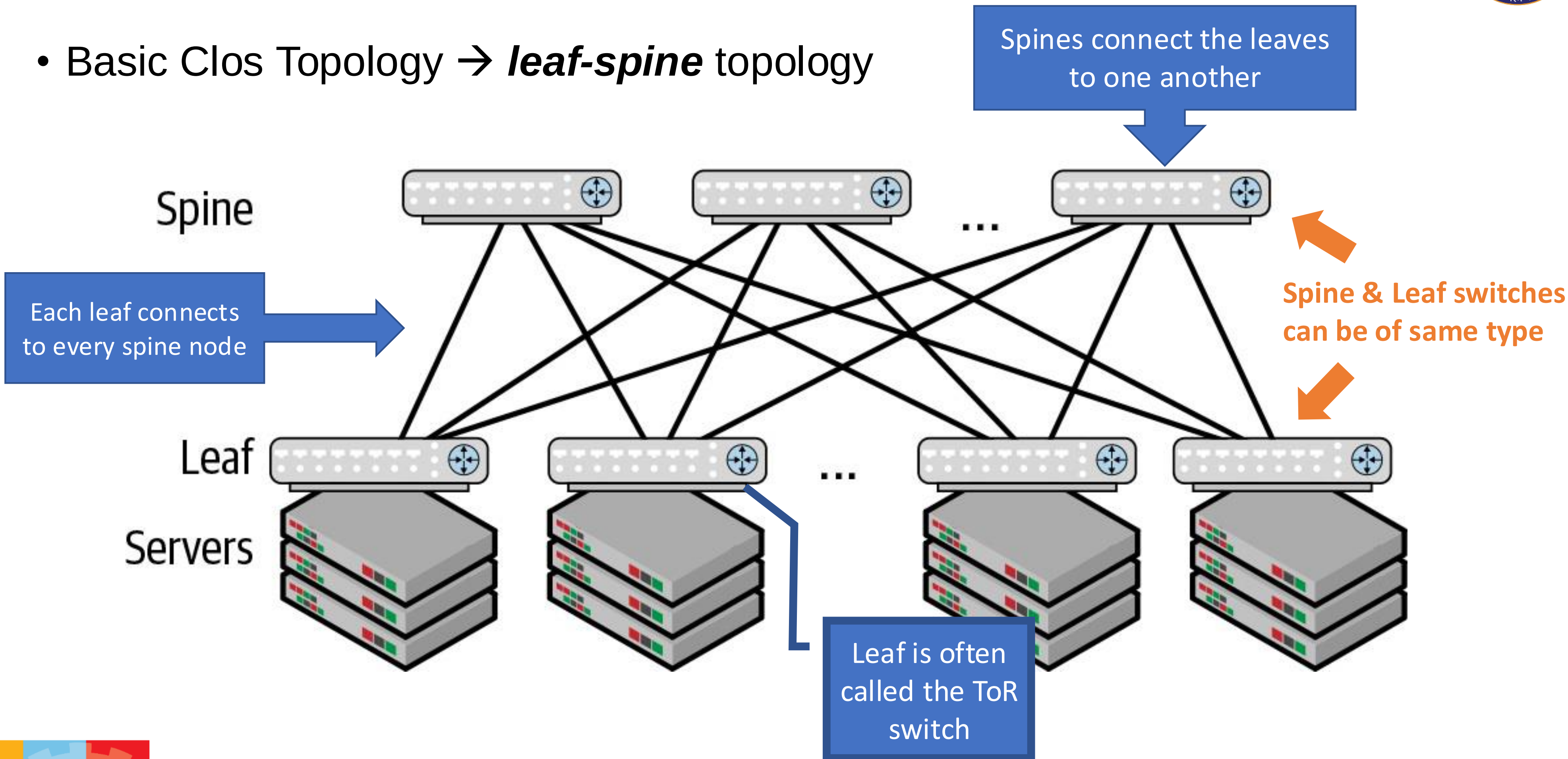


Challenges with Acc-Agg-Core Topologies

- 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



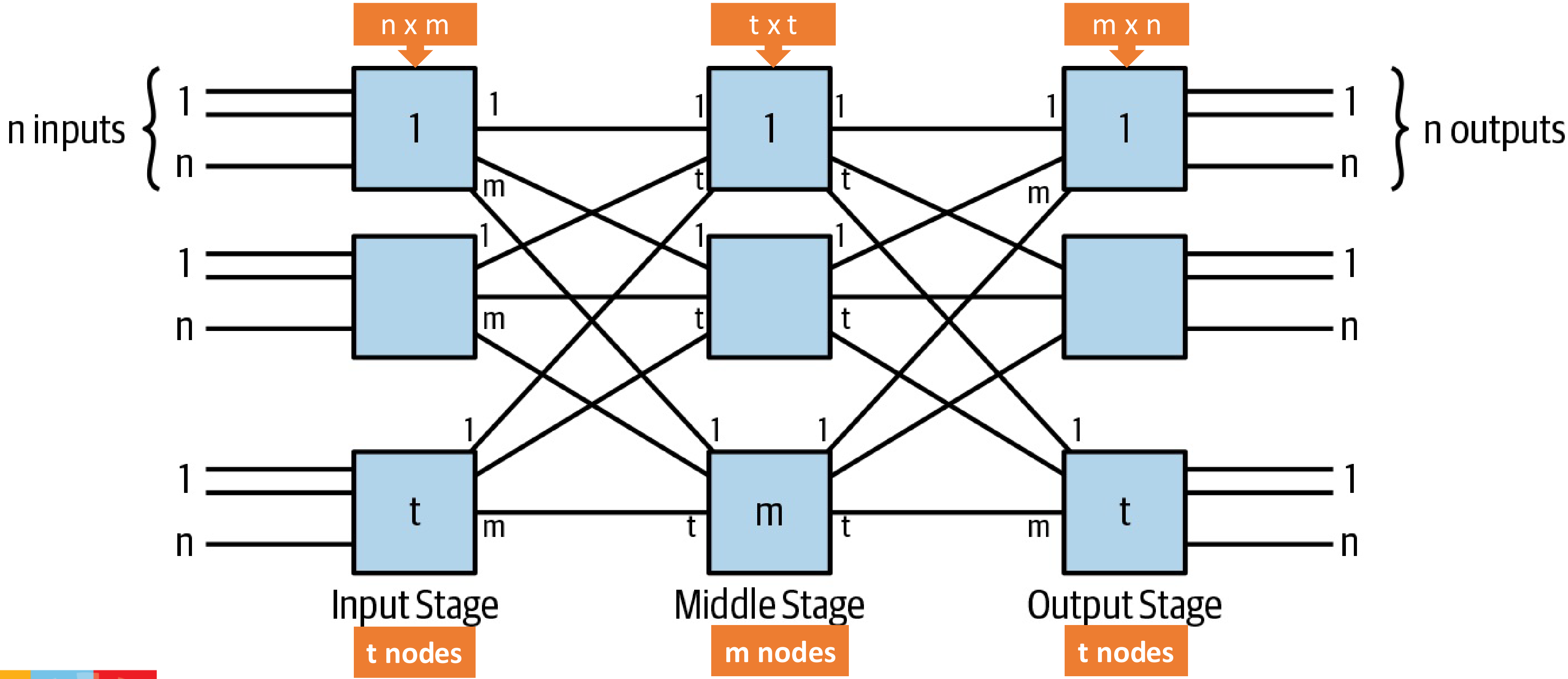
Source: [Cloud Native Data Center Networking by Dinesh G. Dutt]()

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

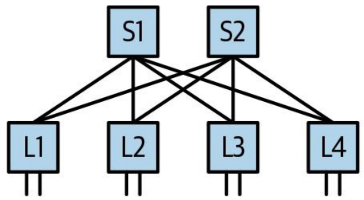
If: m=n=t AND Flip the Output Stage onto the Input Stage → *Leaf-Spine* Topology



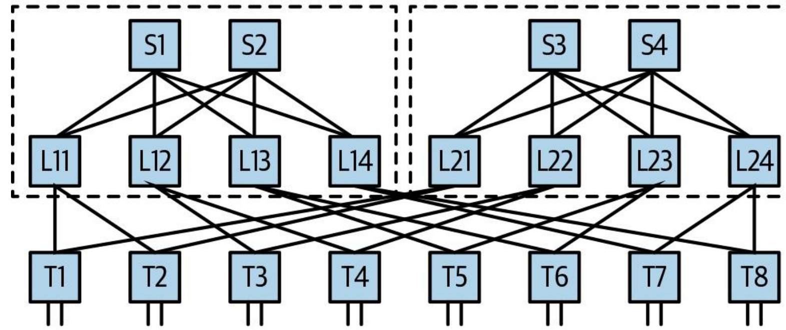
Source: Cloud Native Data Center Networking by Dinesh G. Dutt

Scaling Clos Topology

Examples with four-port switches

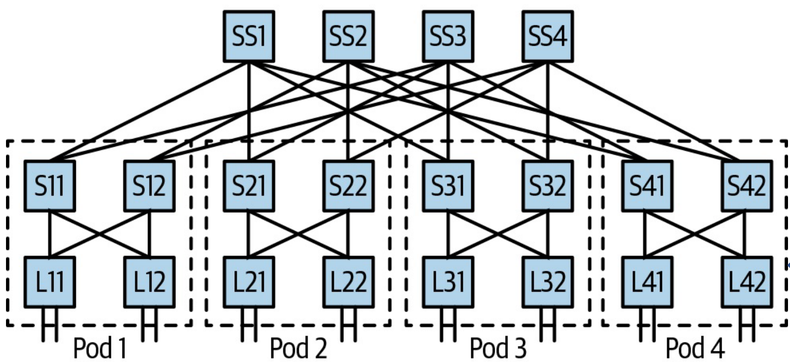


(a) Two-tier Clos using four-port switches



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

Model popularized by **Facebook**



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

Model used by Microsoft and Amazon

Source: [Cloud Native Data Center Networking by Dinesh G. Dutt]()

DCN Design Aspects

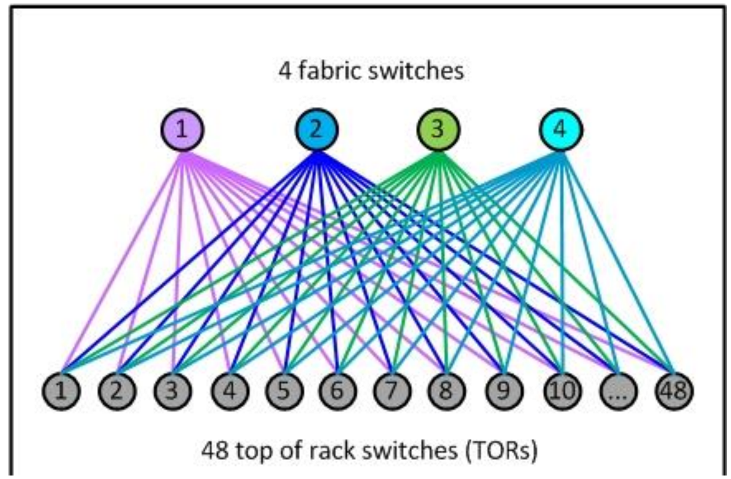
- Choice of Topology
- Oversubscription of Bandwidth
- Multipath Routing
- Overall Cost

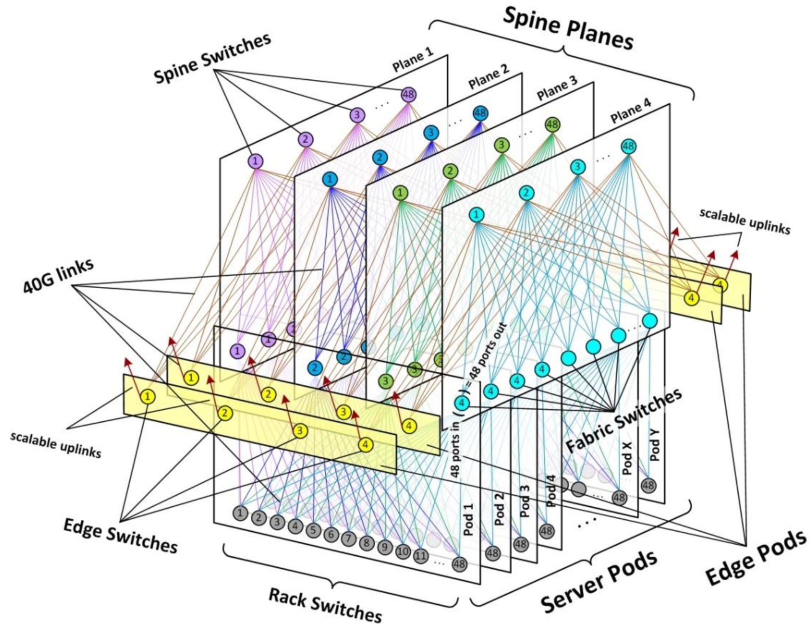
Case Studies

- 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.

Thank You!