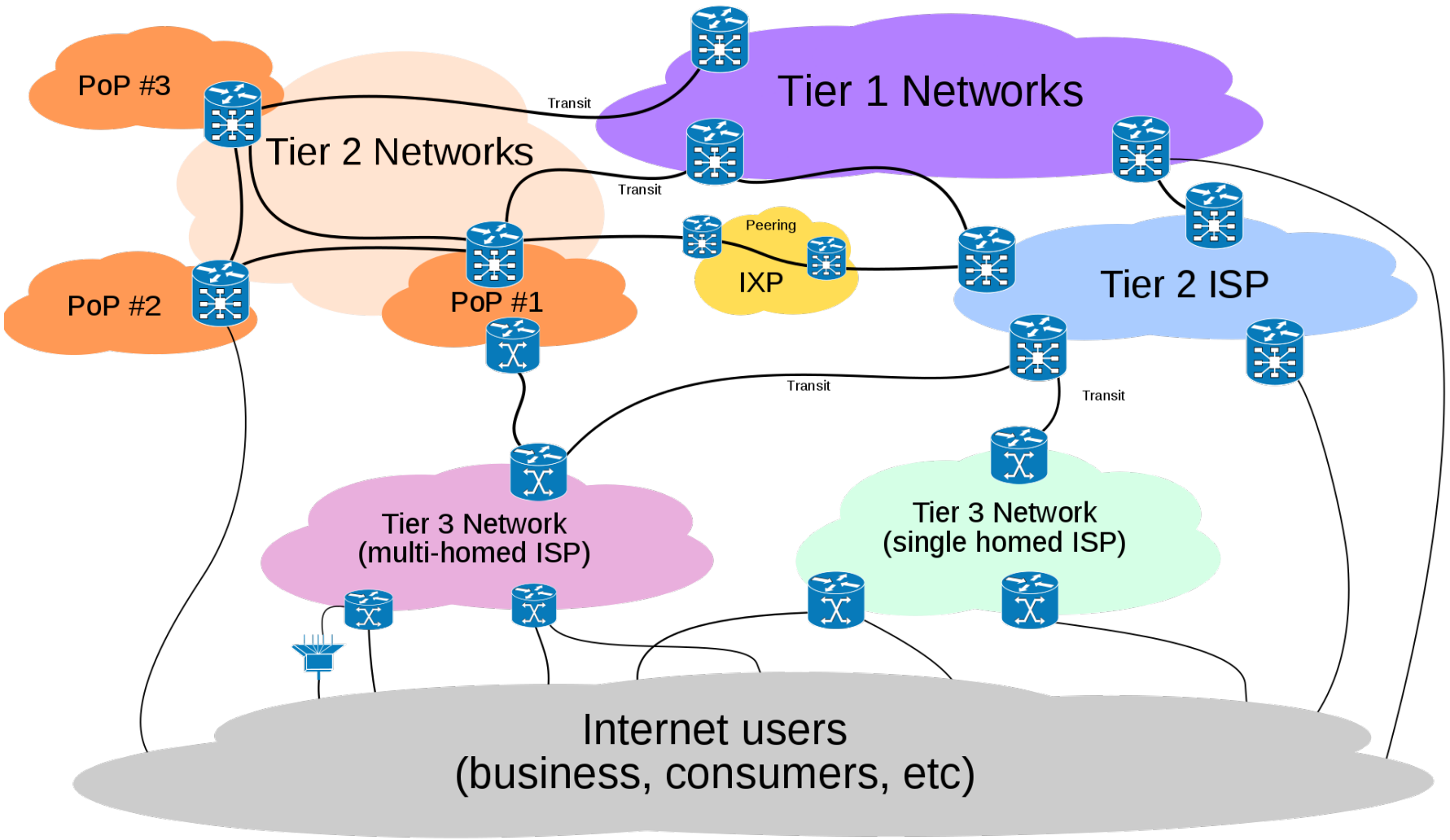
Datacenter Network

Peng Zhang
School of Computer Science and Technology
Xi'an Jiaotong University

Spring 2024

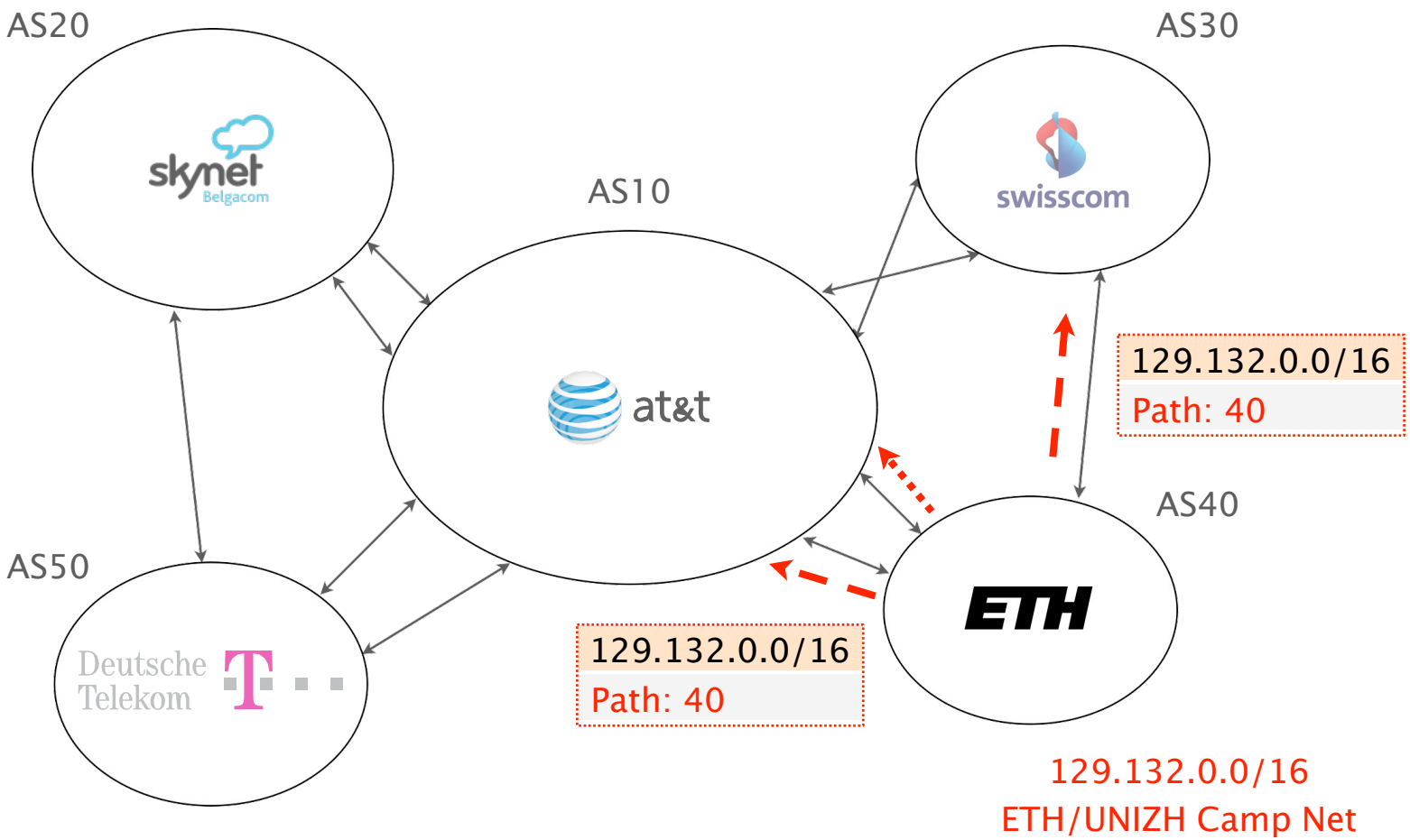
Recap: Internet



Internet connectivity: an overview **https://commons.wikimedia.org/wiki/User:Ludovic.ferre**

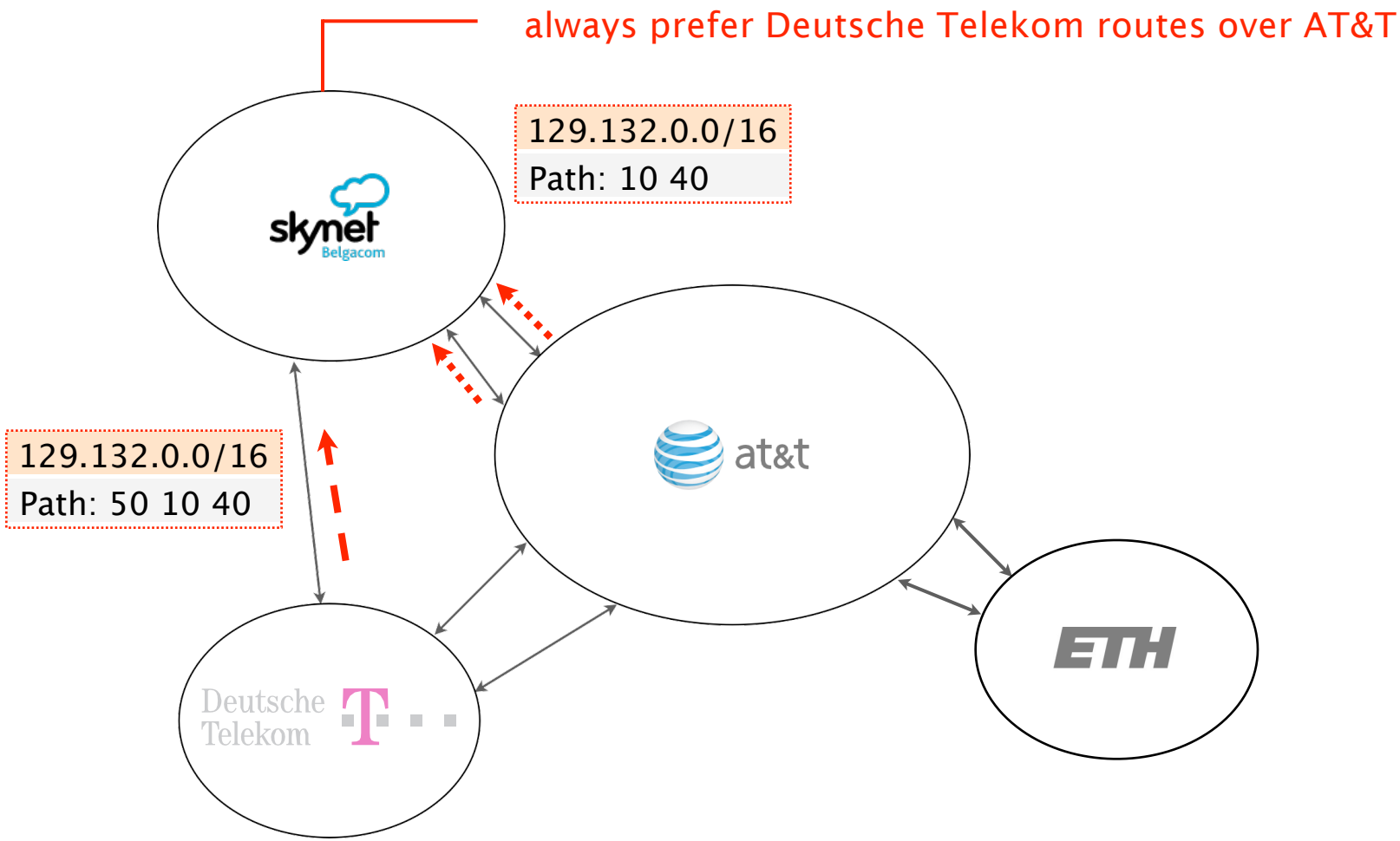
Recap: BGP is Path Vector Protocol

BGP announcements carry complete path information instead of distances



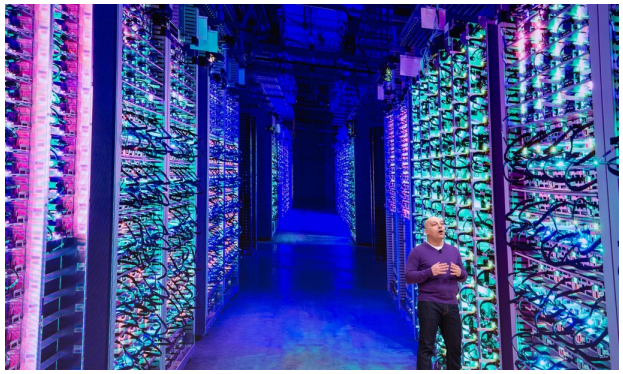
Recap: BGP is Policy Based

Each AS is free to select and use any path preferably, the cheapest one

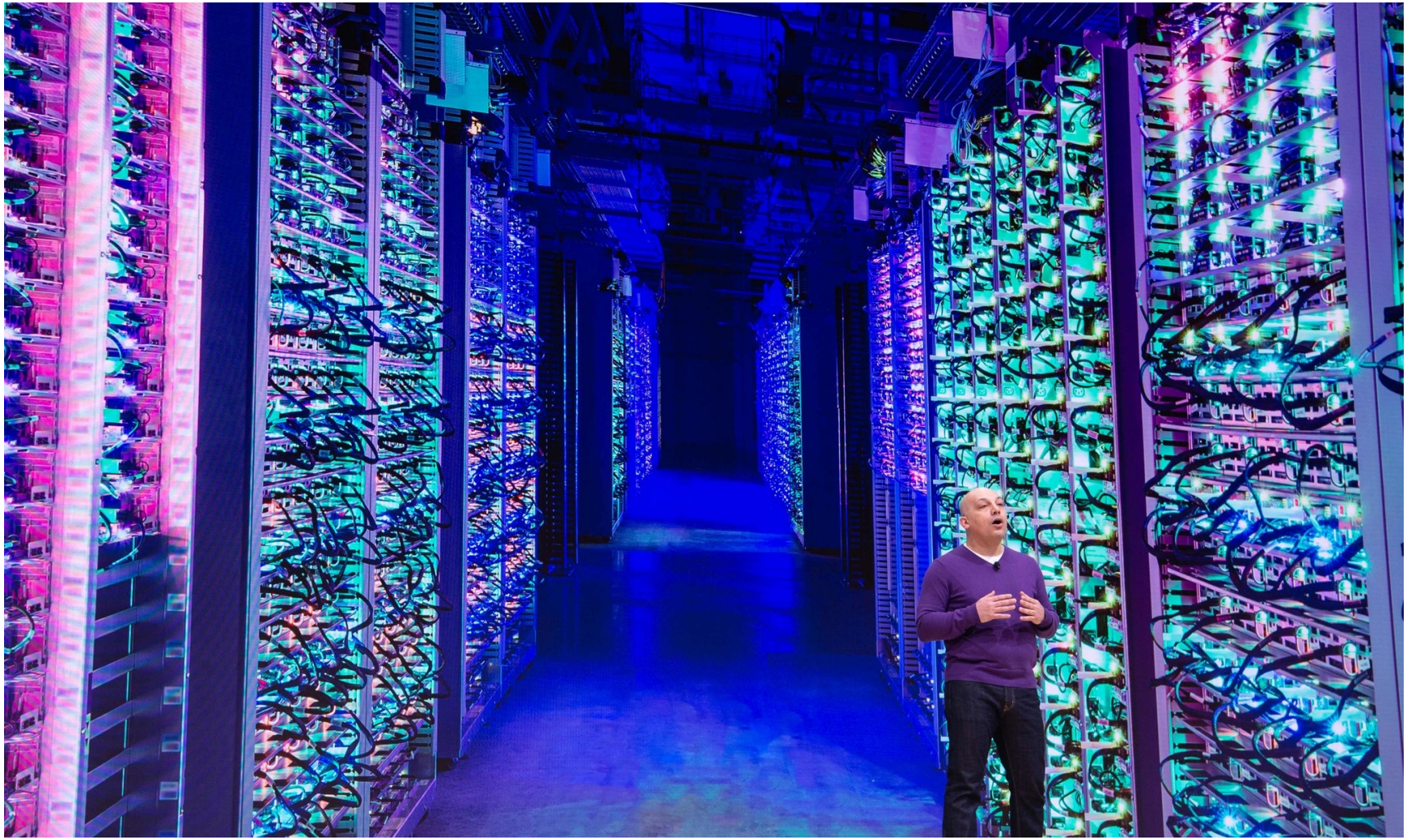


Finally Arriving at Datacenter

dstmac	??:??:??:??:??:??
dstip	8.8.8.8
payload	What is the IP address of www.baidu.com?



Datacenters



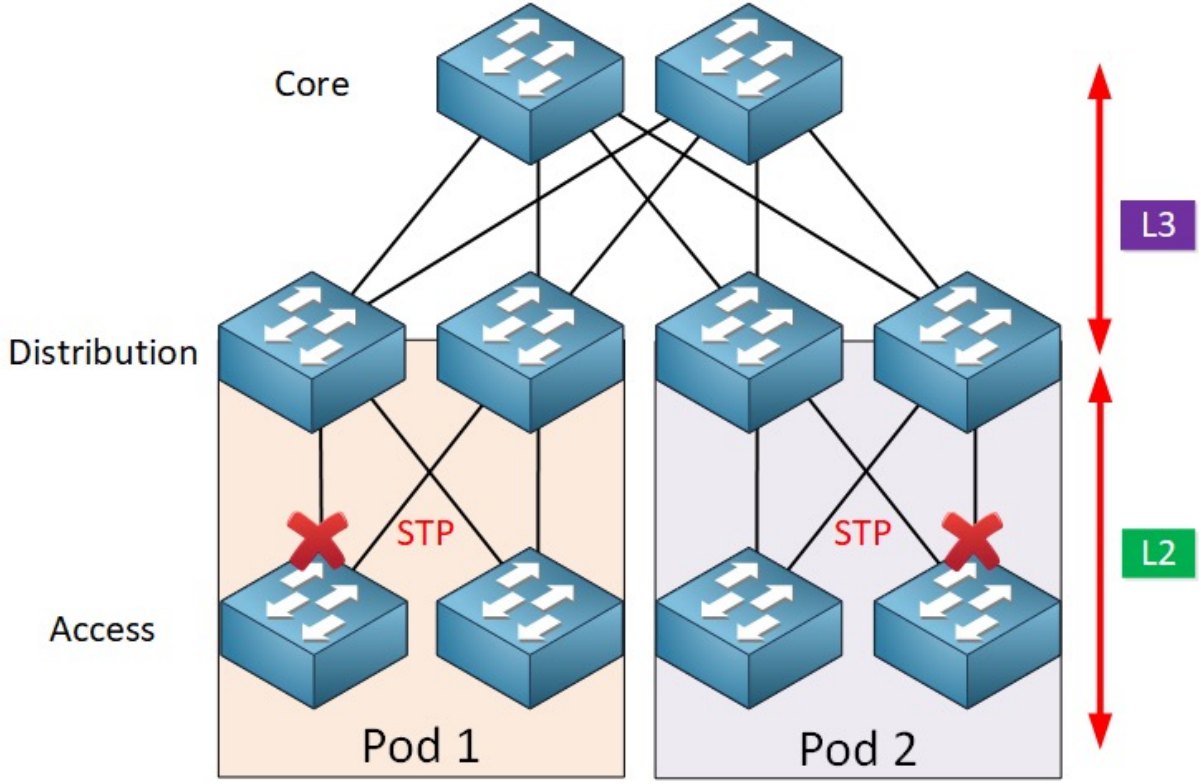
https://www.nytimes.com/2020/02/27/technology/cloud-computing-energy-usage.html

Datacenter Networks

P5.NR-A.10
P5.NR-A.12
P5.NR-A.11

Traditional Architecture

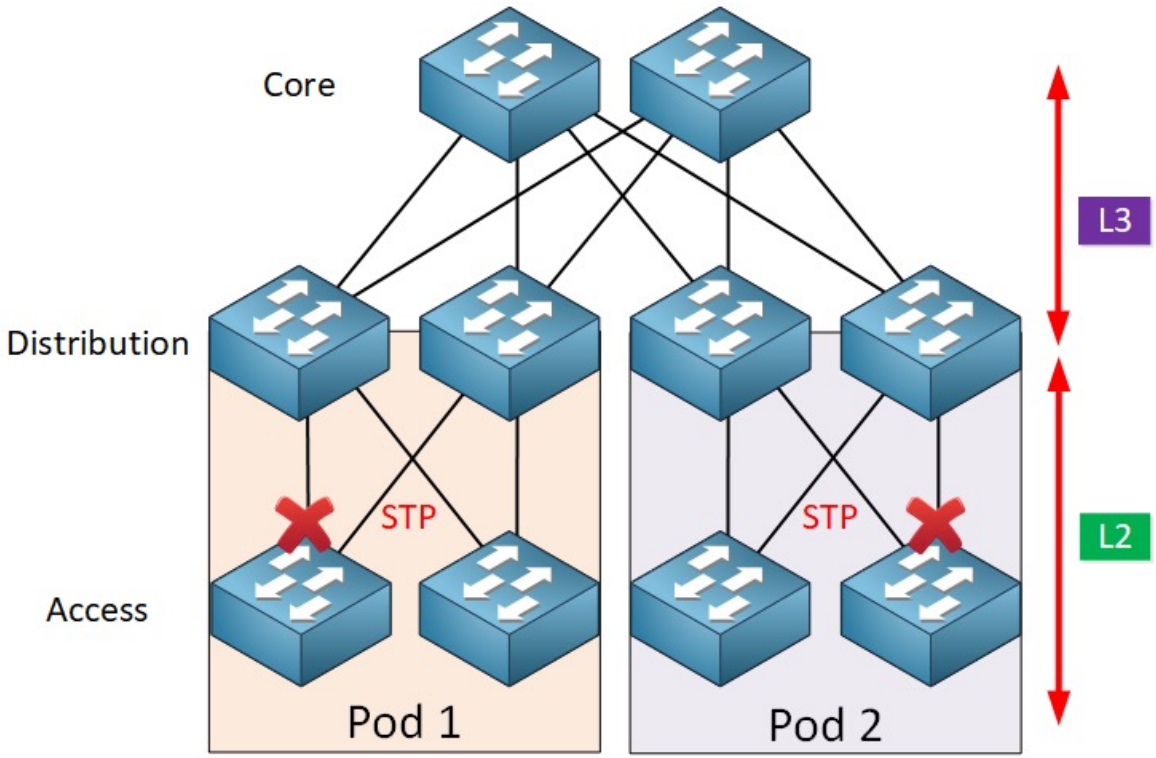
- Three layers: core, distribution (aggregation), access
- Following campus network design



https://networklessons.com/cisco/ccna-200-301/spine-and-leaf-architecture

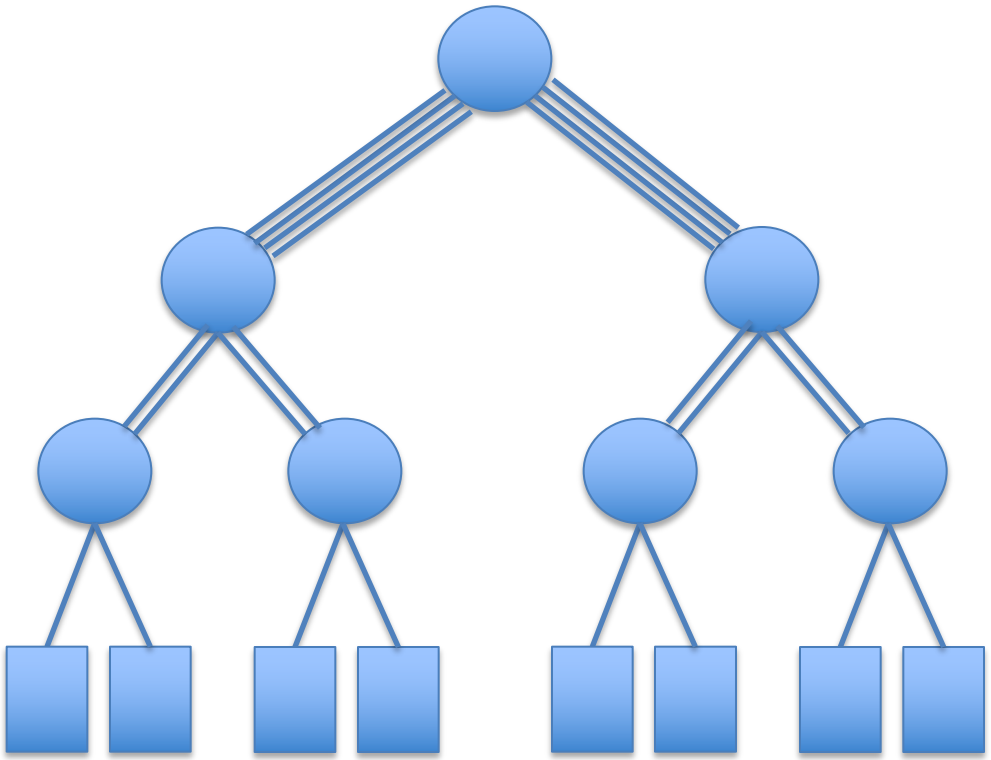
Limitations

- Waste of bandwidth
  - STP block ports except one
- Latency
  - Server to server traffic: 5 hops
- No for the cloud
  - rise of east-west traffic
  - VM migration



https://networklessons.com/cisco/ccna-200-301/spine-and-leaf-architecture

Solution: “Fat” Tree



More ports = Much more cost

Cisco Nexus C36180YC-R Switch

48 ports 1, 10, or 25 Gigabit Ethernet SFP

Product	List Price (USD)
N3K-**C36180YC-R**	$39,285.00

https://itprice.com/cisco-gpl/c36180yc-r

Telephone Switching

- Problem: too many crosspoints

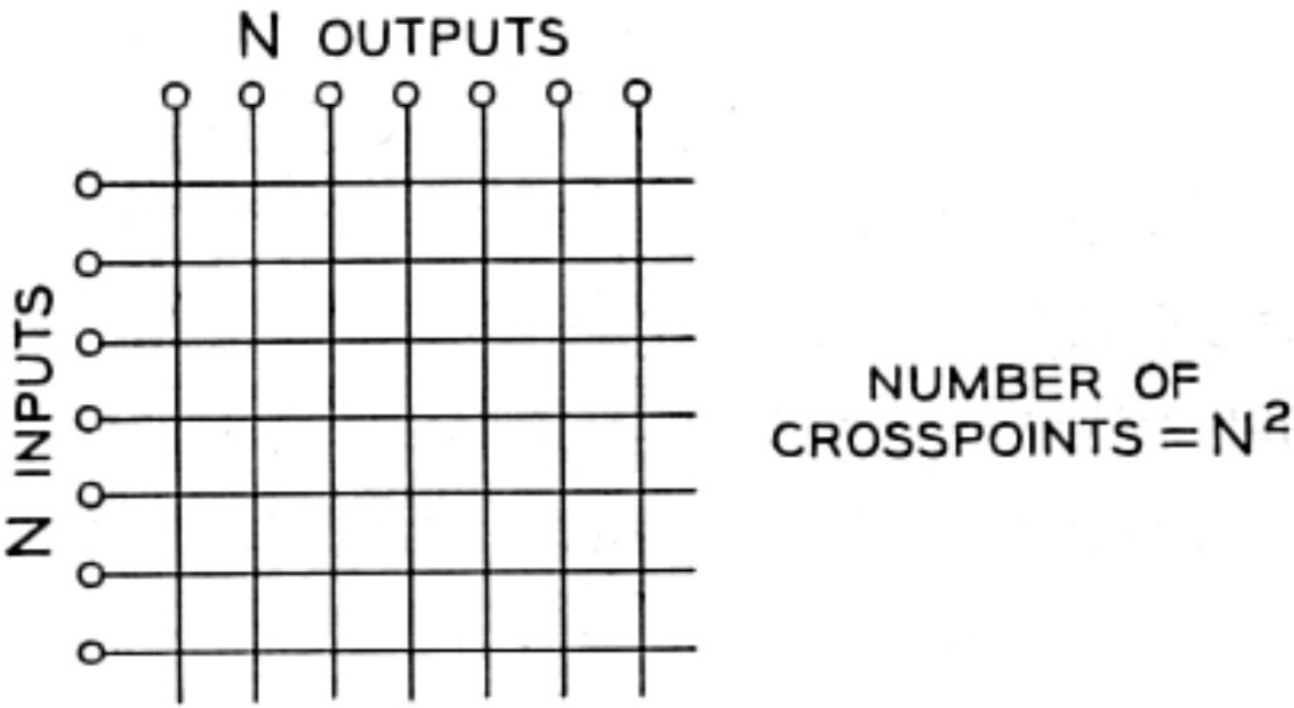
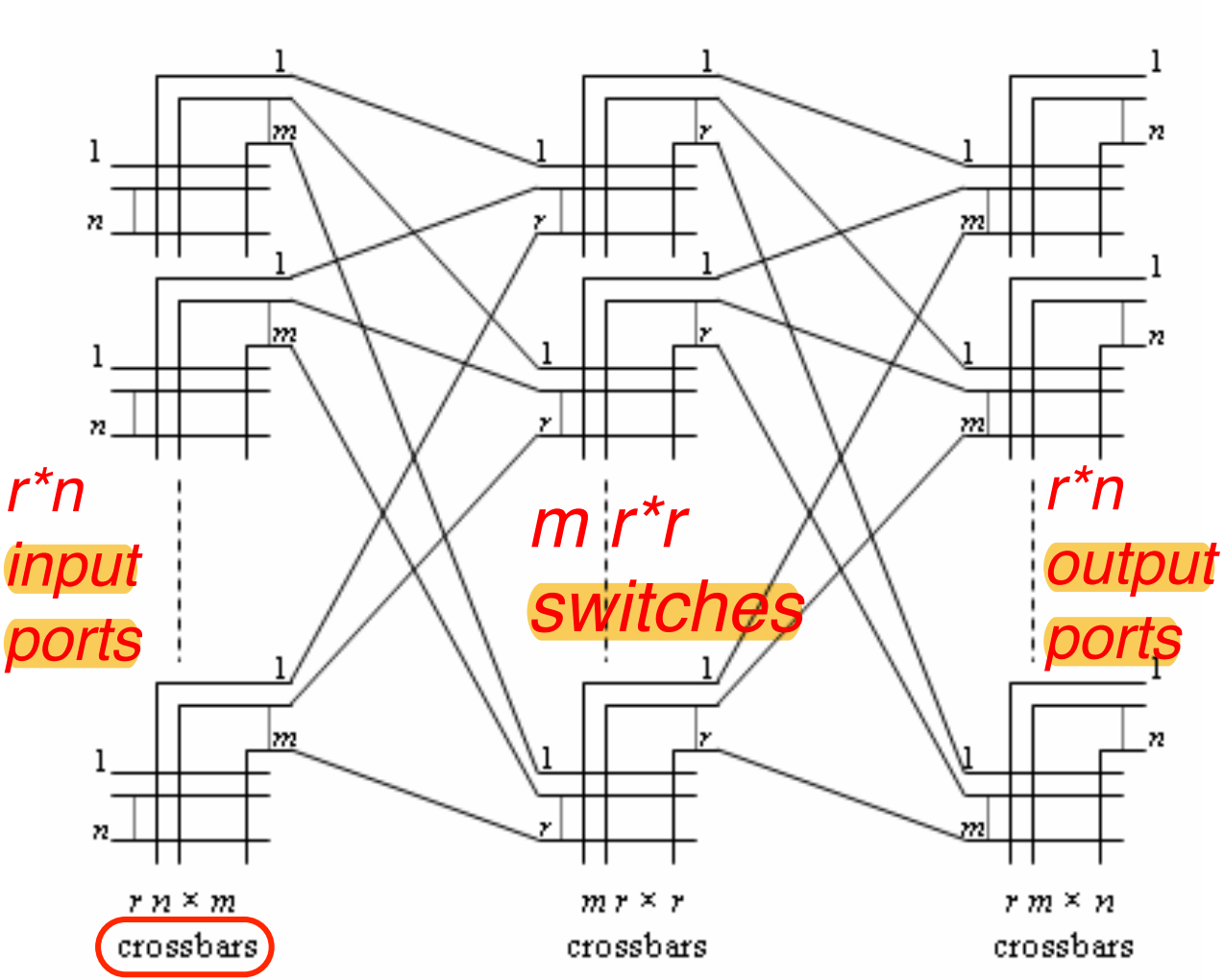


Fig. 1 — Square Array.

Clos Network

- Solution: multi-stage switching

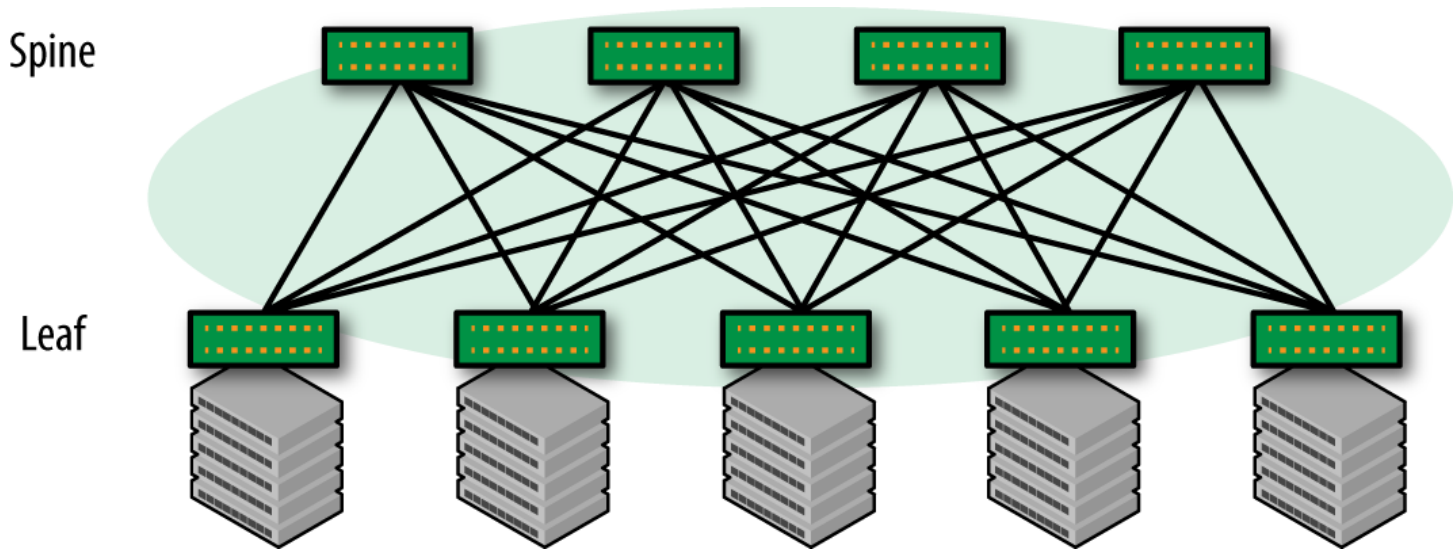


Charles Clos first proposed the design in 1952

Theorem:

If $m \geq 2n-1$, the Clos network is ***strict-sense nonblocking***

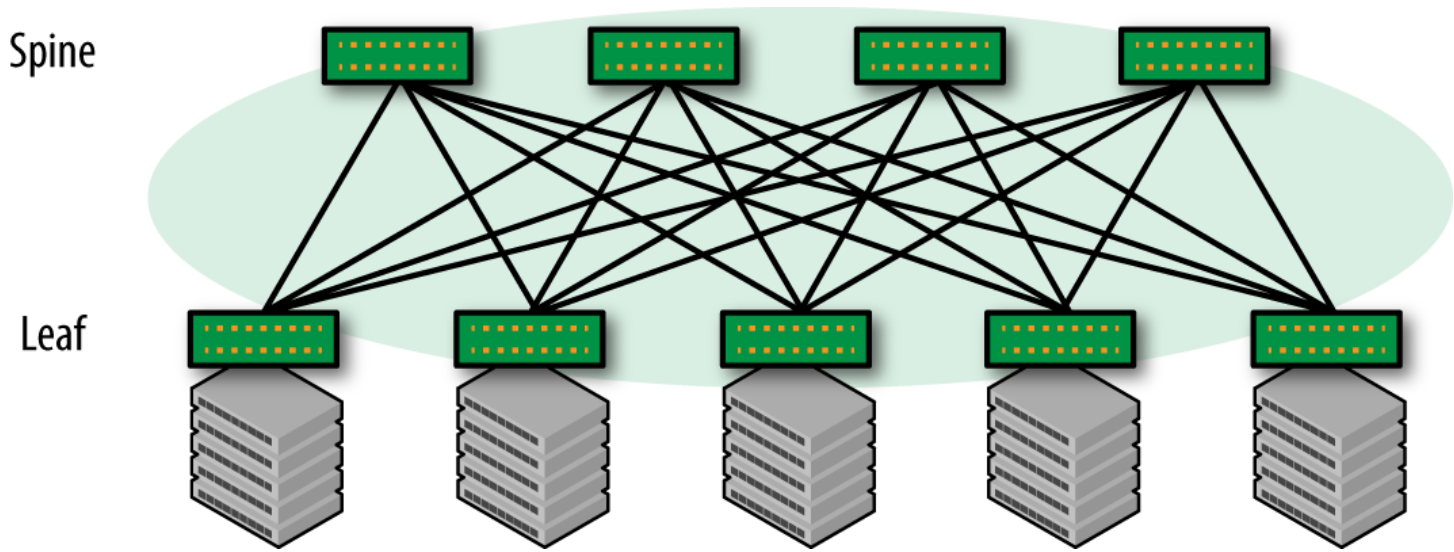
Folded 3-Stage Clos Network



- Equal-cost multi-path (ECMP): IP and Routing
- Server to server traffic: 3 hops
- Large bandwidth for east-west traffic
- VM migration

https://www.oreilly.com/library/view/bgp-in-the/9781491983416/ch01.html

Folded 3-Stage Clos Network

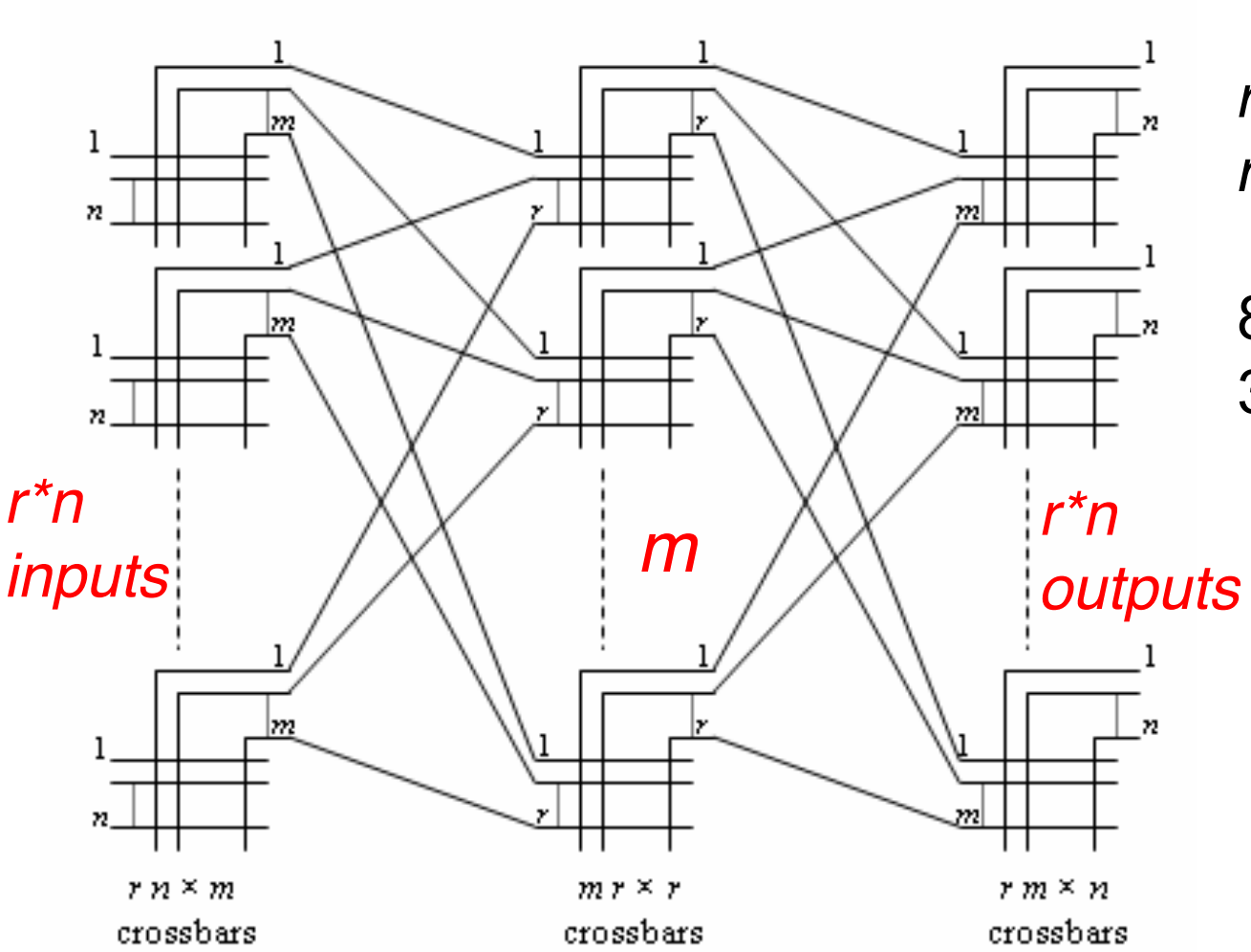


Problem: each switch still needs a lot of ports

https://www.oreilly.com/library/view/bgp-in-the/9781491983416/ch01.html

Example

- Connect 16 servers with 3-stage Clos network

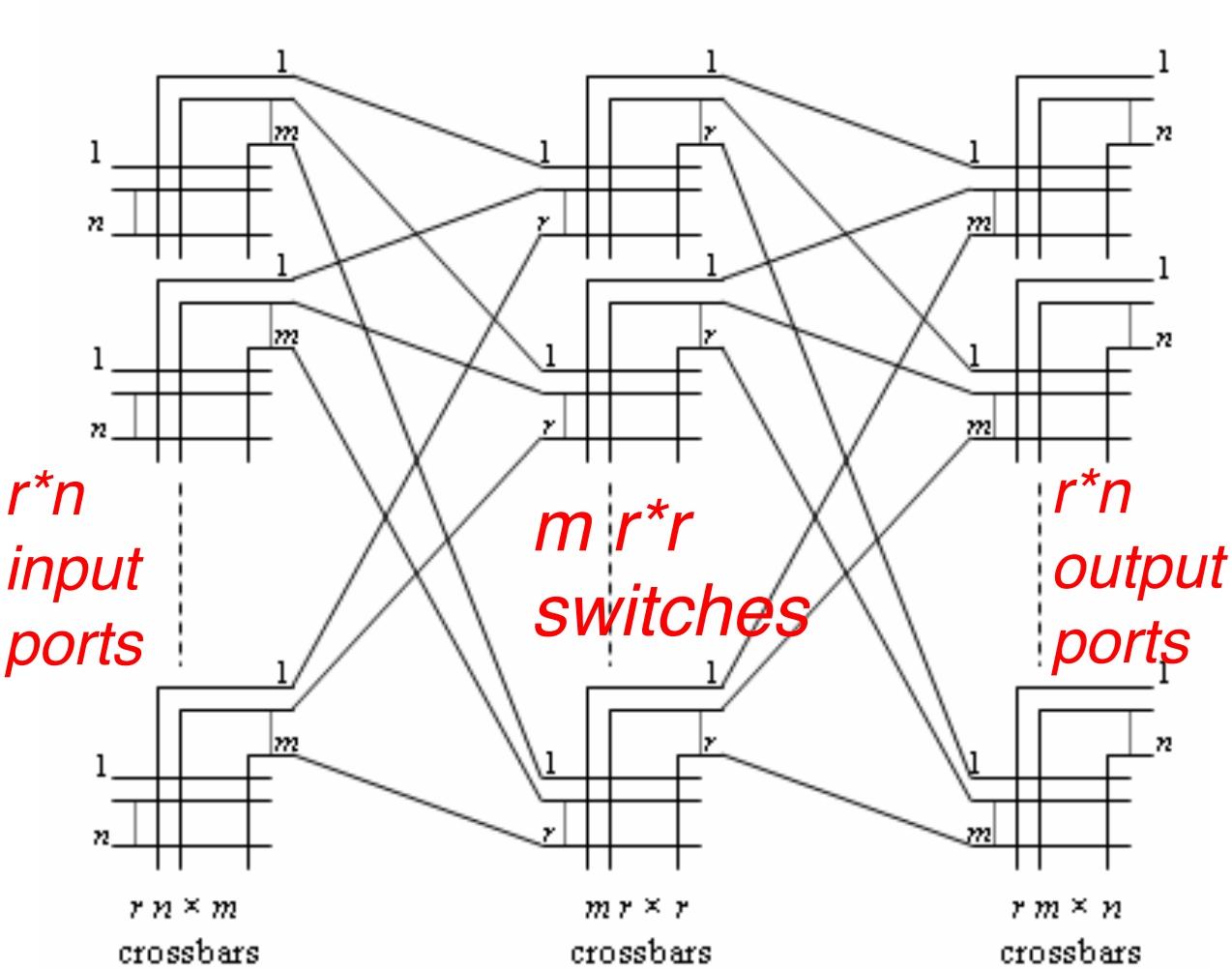


*r*n*2=16*
n=2, r=4, m≥3

8 5-port switches,
3 8-port switches

A Generalization of Clos Network

- Rearrangeably Nonblocking



Theorem:

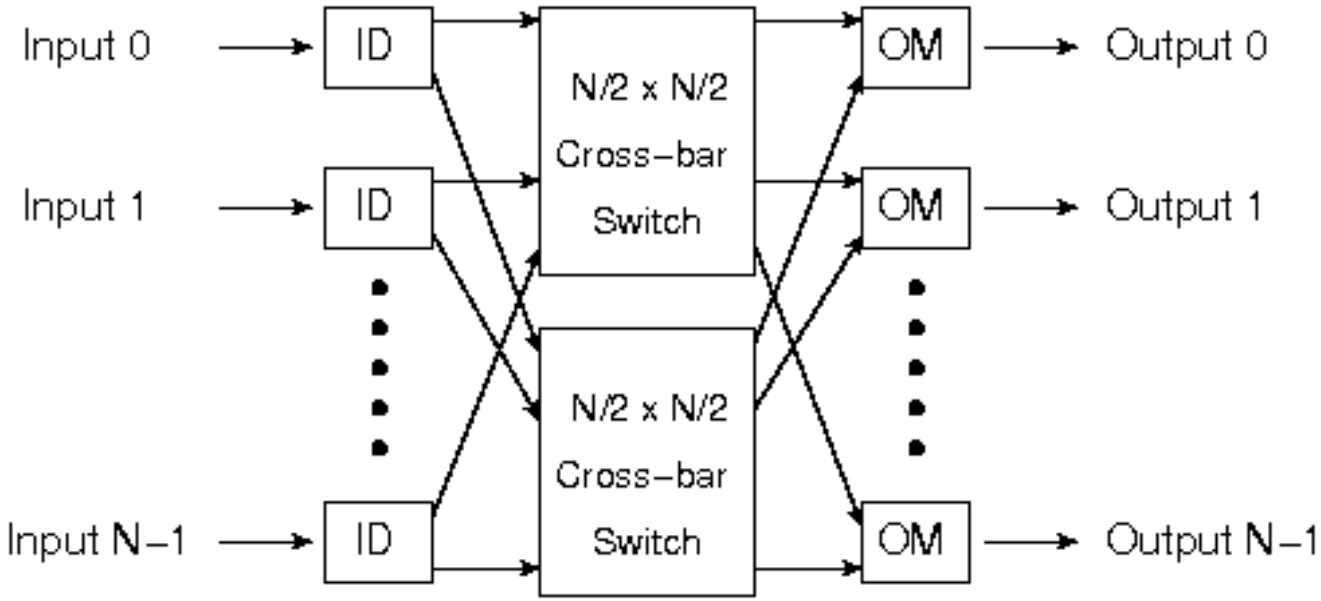
If ***m* ≥ *n***,
the Clos network
is ***Rearrangeably nonblocking***

*r*n*2=16*
n=2, r=4, m=2

8 4-port switches,
2 8-port switches

Beneš network

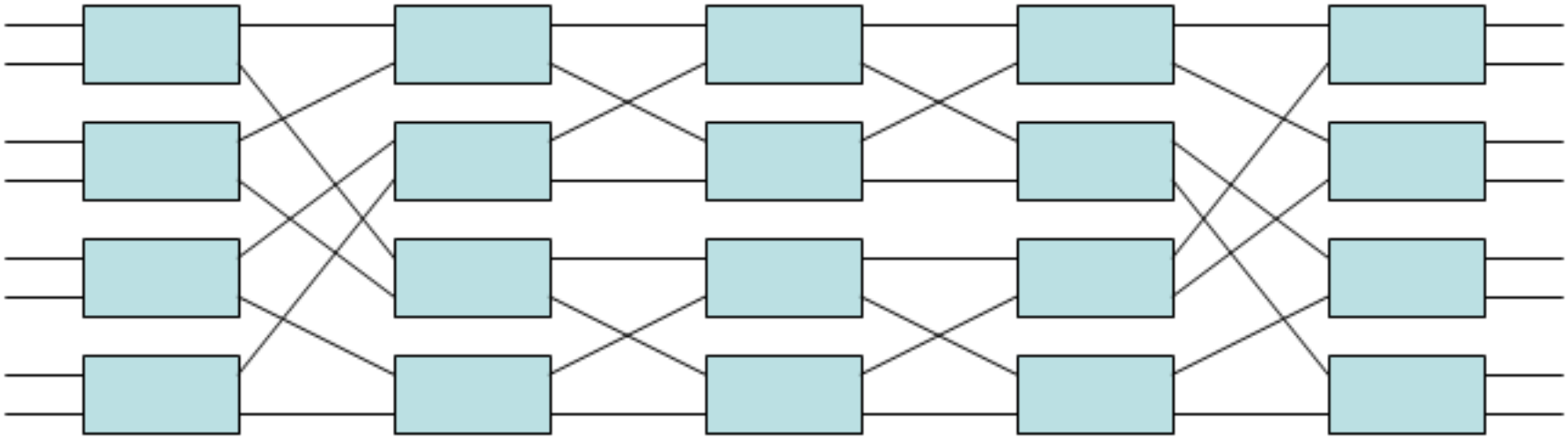
- Solution: adding more stages
  - reducing an *N* x *N* cross-bar switch to two *N/2* x *N/2* cross-bar switches and two *N*-input exchange switches



N×N Beneš network

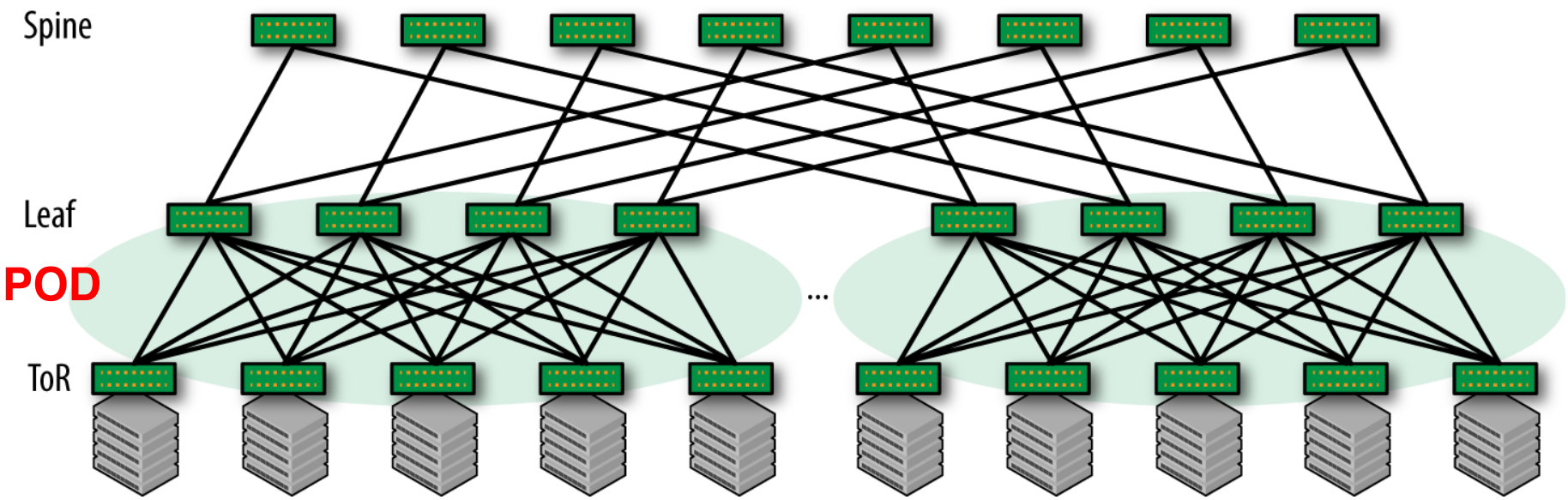
- *2 log2N − 1* stages, each containing *N/2* 2×2 crossbar switches, and use a total of *N log2N − N/2* 2×2 crossbar switches.

N=8: 5 stages, 20 4-port switches



Folded 5-Stage Clos Network

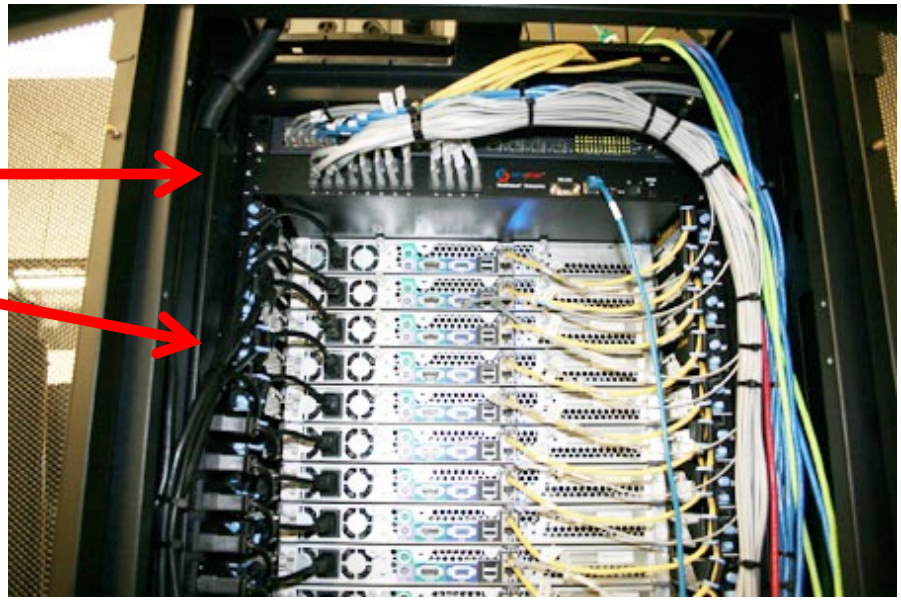
- Solution: adding more stages
  - Connect many spine-leaf networks by another layer of spine switches
  - Each spine-leaf network is called a *pod* or *cluster*,



https://www.oreilly.com/library/view/bgp-in-the/9781491983416/ch01.html

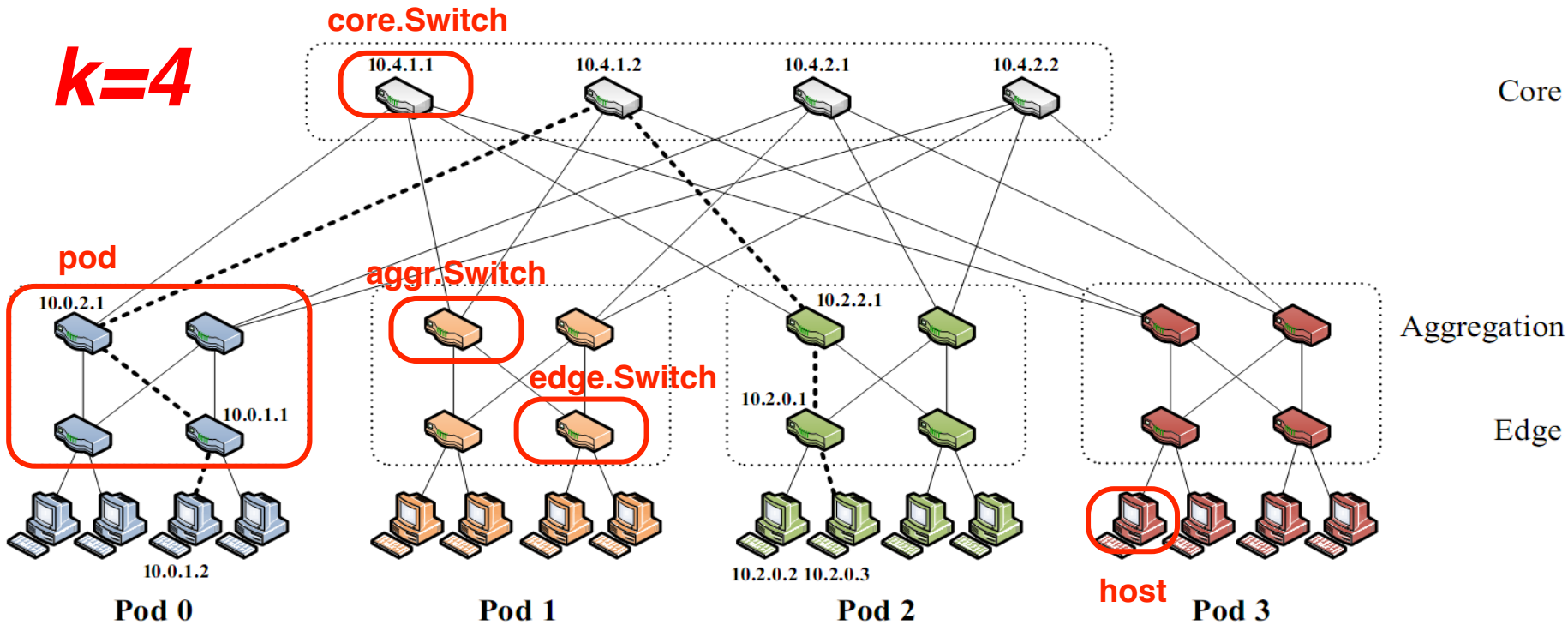
Top-of-Rack Architecture

- Rack of servers
  - top-of-rack switch
  - commodity servers



- Modular design
  - Preconfigured racks
  - Power, network, and storage cabling

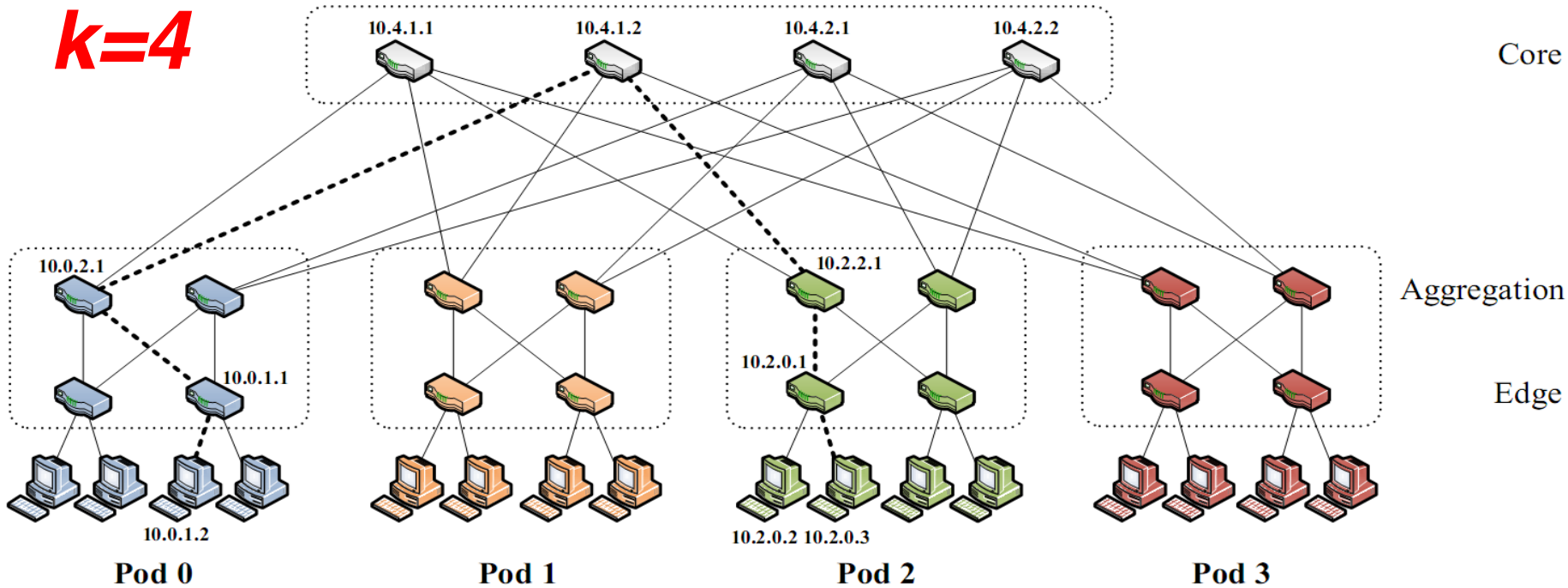
Fat Tree



- Each switch has k ports
- There are k pods
- Each pod has k/2 aggr. switches and k/2 edge switches
- Each edge switch is connected to k/2 hosts and k/2 aggr. switches
- Each core switch is connected to an aggr. switch for each pod

A Scalable, Commodity Data Center Network Architecture, SIGCOMM 2008

Fat Tree



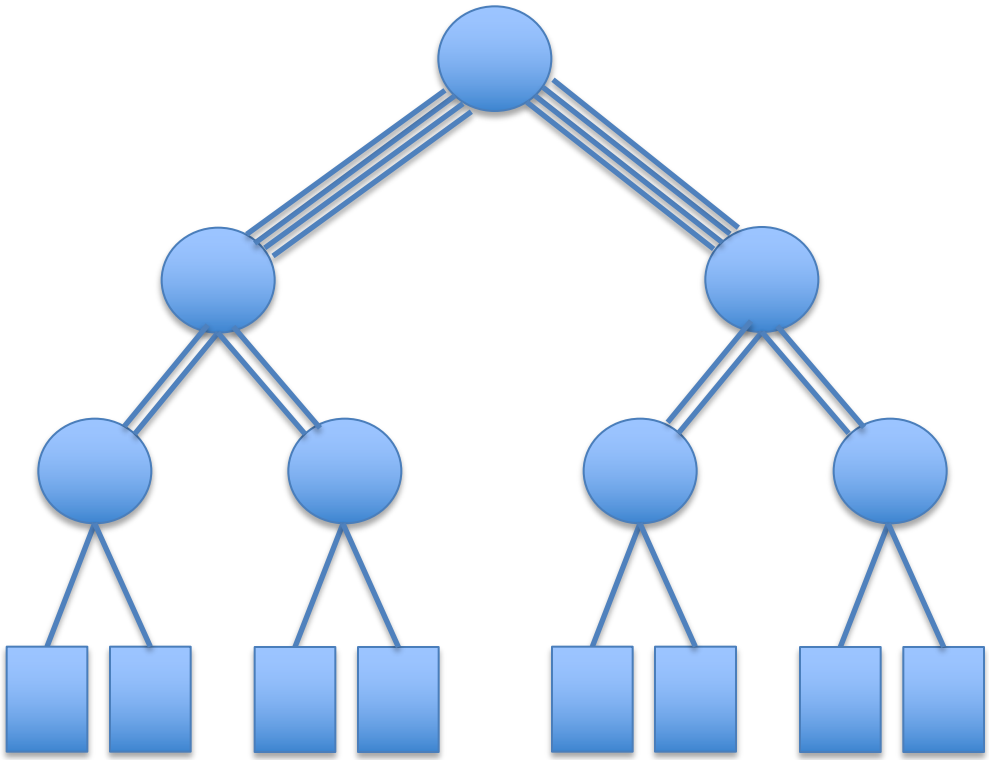
- How may switches are needed? $k*(k/2+k/2)+(k/2)*(k/2)=(5/4)k^2$
- How many servers that it can hold? $k*(k/2)*(k/2)=(1/4)k^3$
- How many hops between each pair of hosts at different pods? **>=6**
- How many 6-hop paths between hosts at different pods? $(k/2)^2$

65536 servers for **k=64**

A Scalable, Commodity Data Center Network Architecture, SIGCOMM 2008

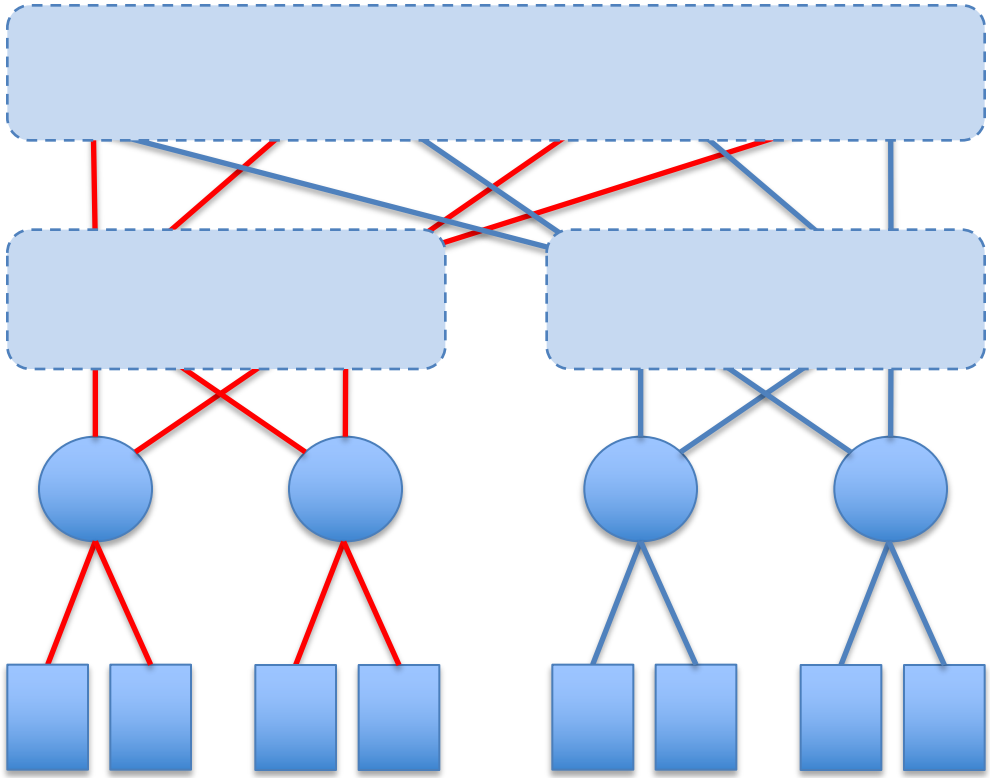
Why “Fat” tree?

Fat Tree

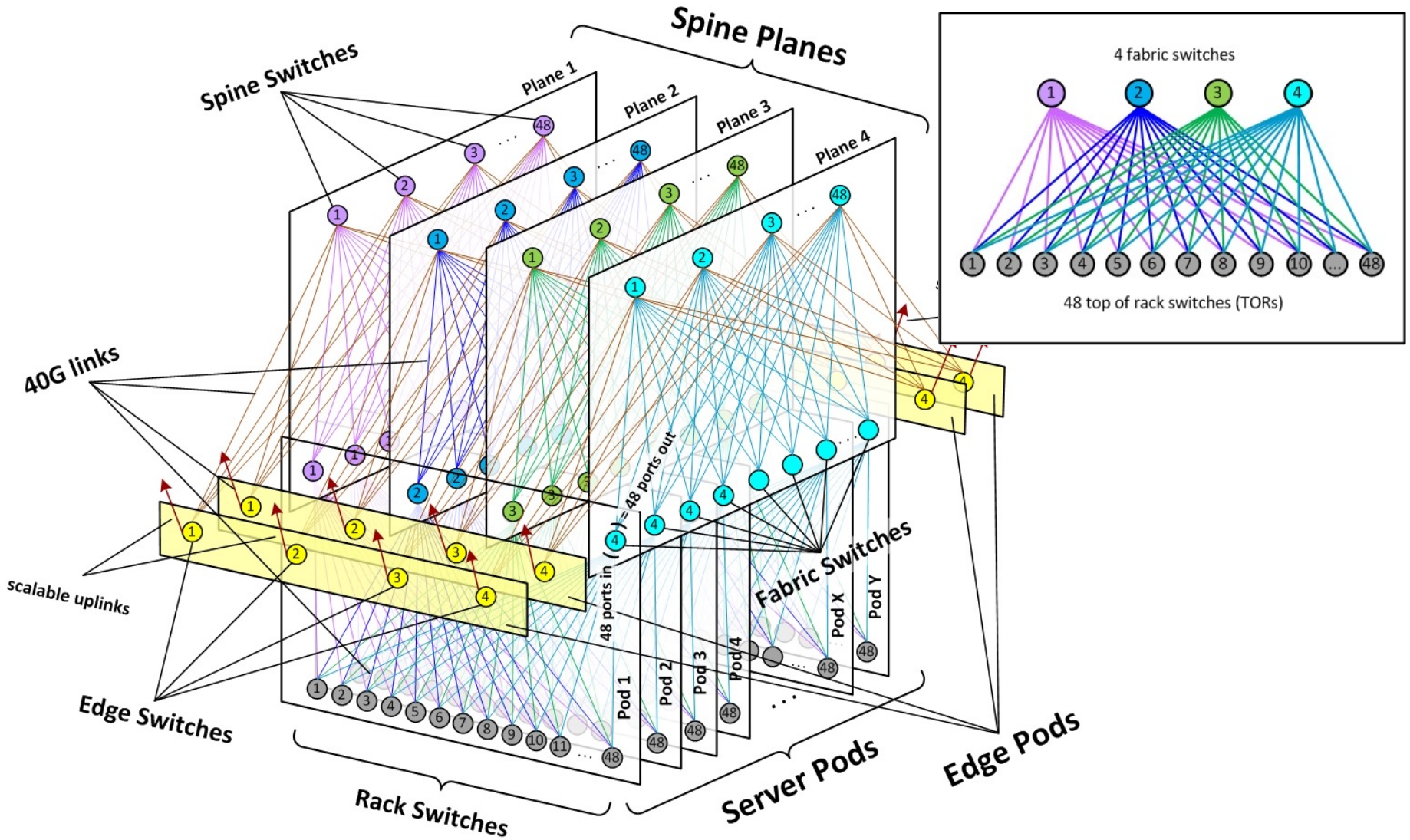


How to Connect 10k Servers?

- Add more cheap switches!

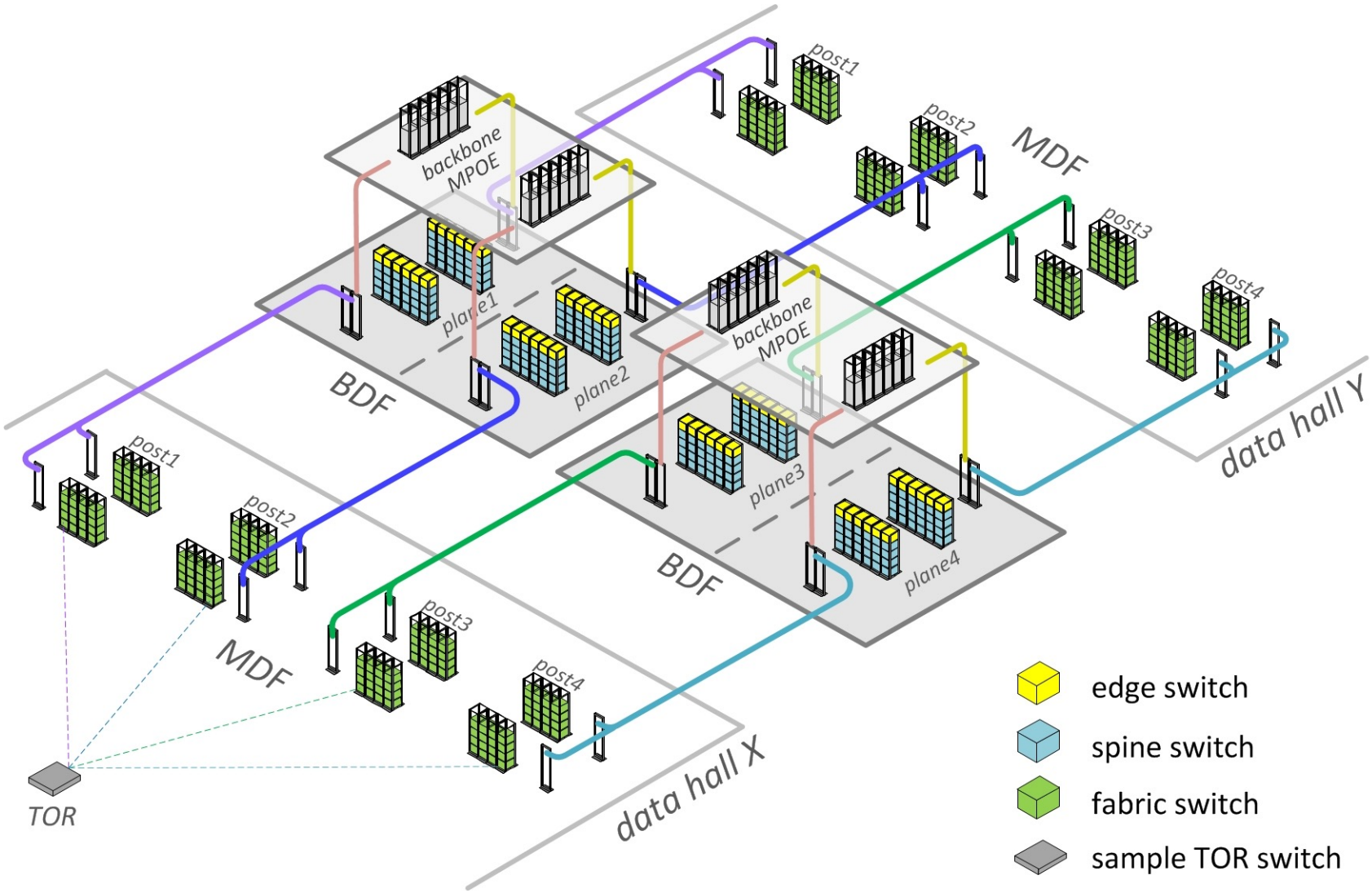


Facebook DCN Topology



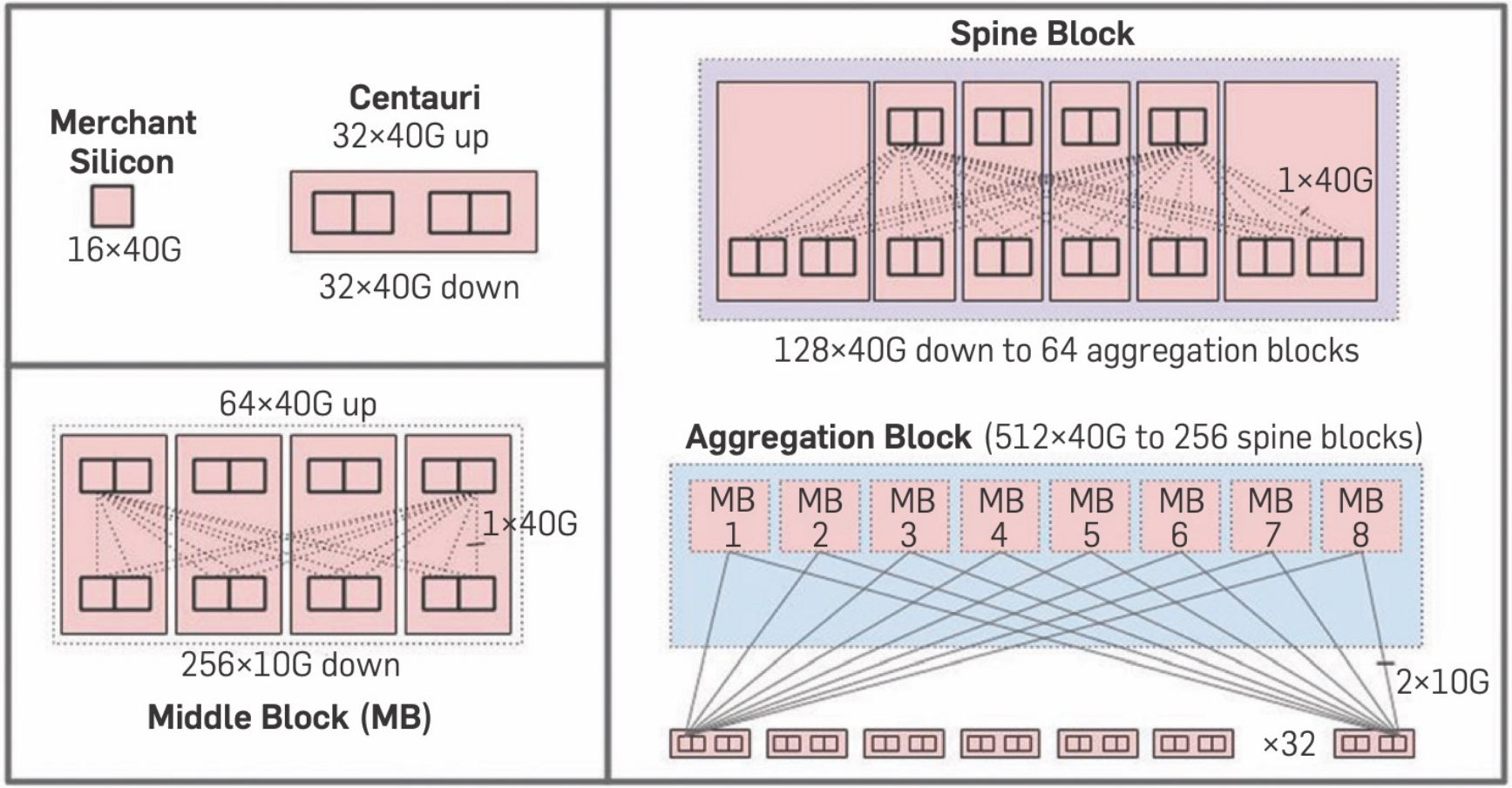
https://engineering.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

Facebook DCN Topology (Physical)



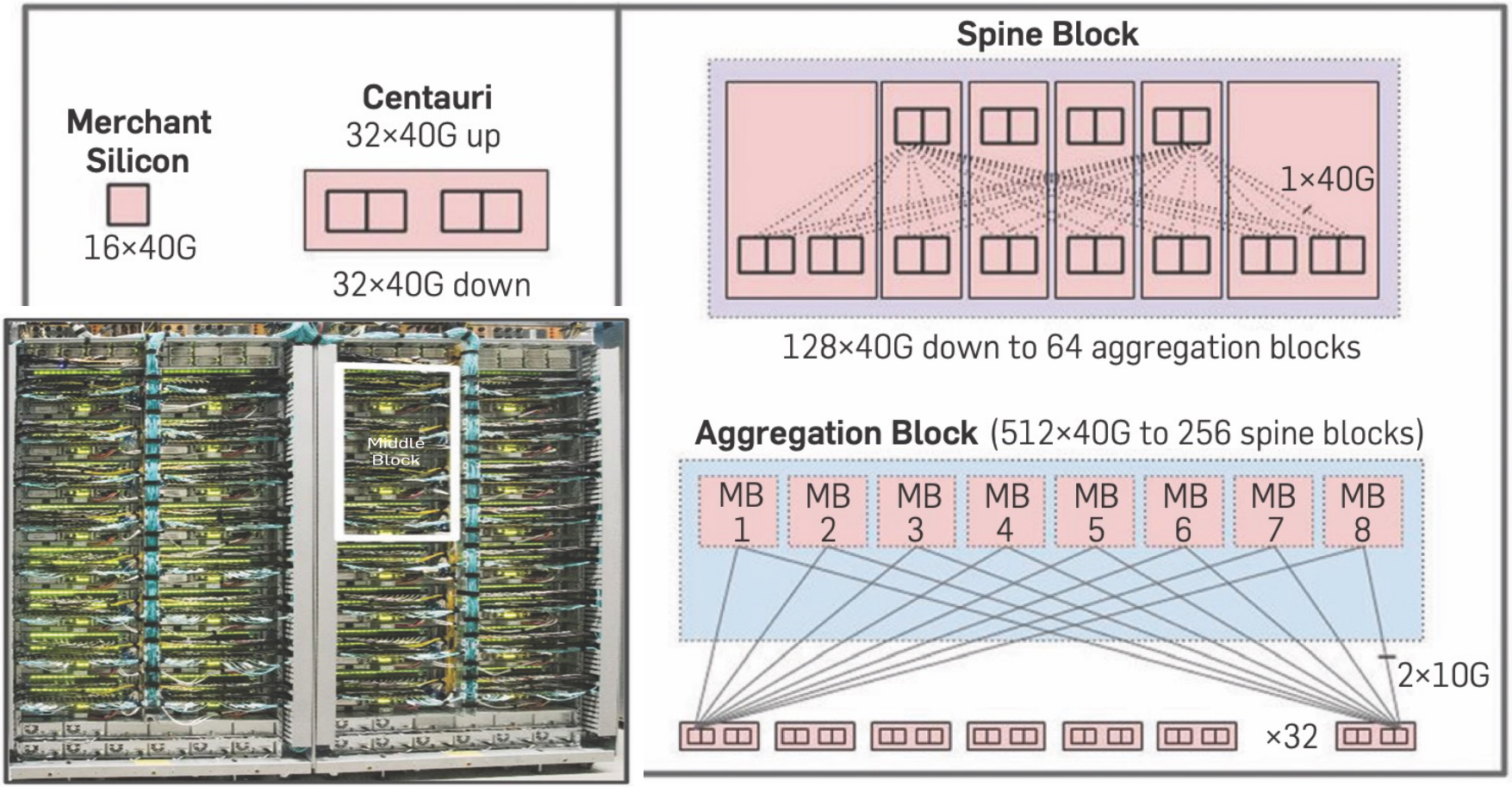
https://engineering.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

Google Jupiter Topology



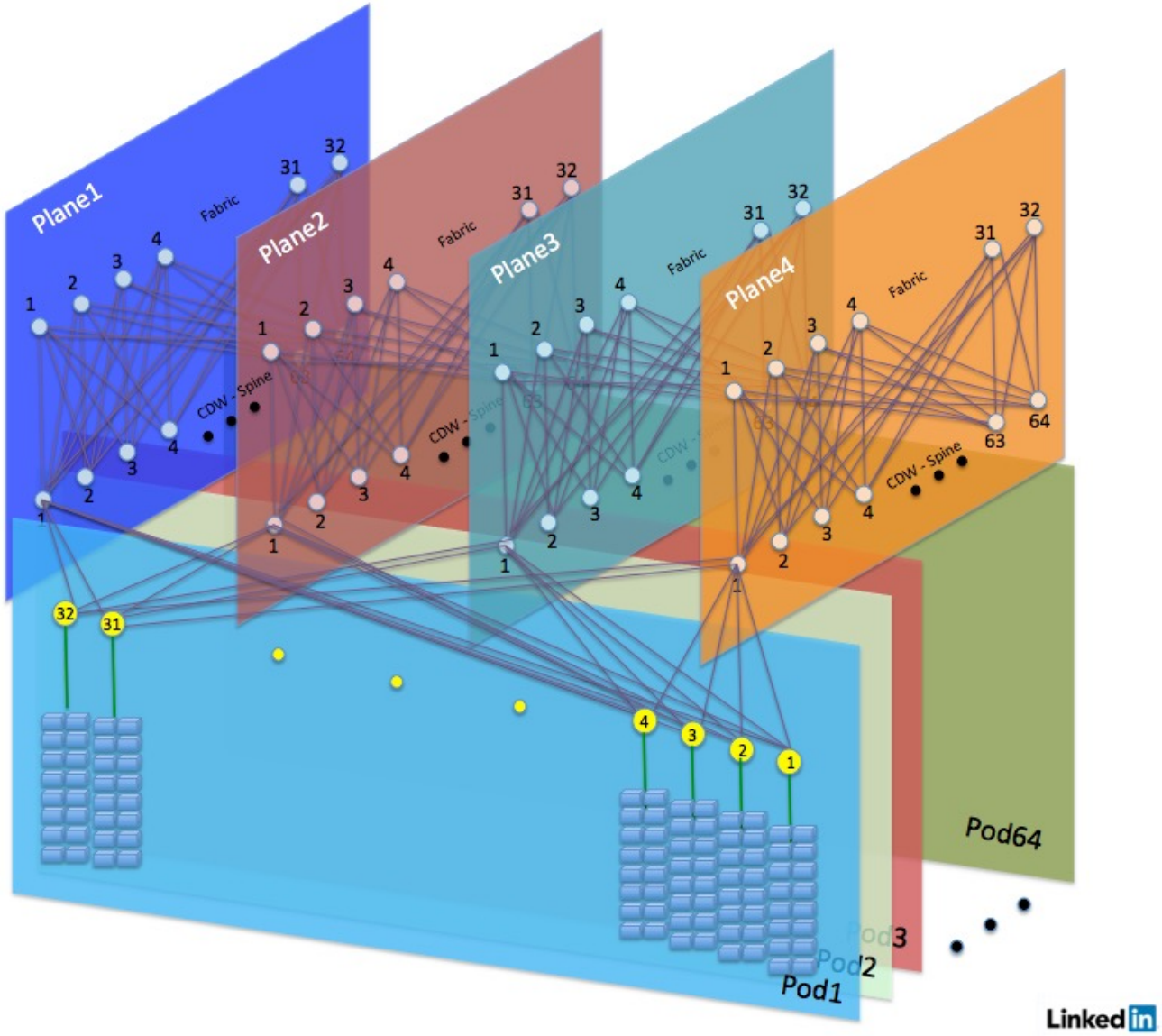
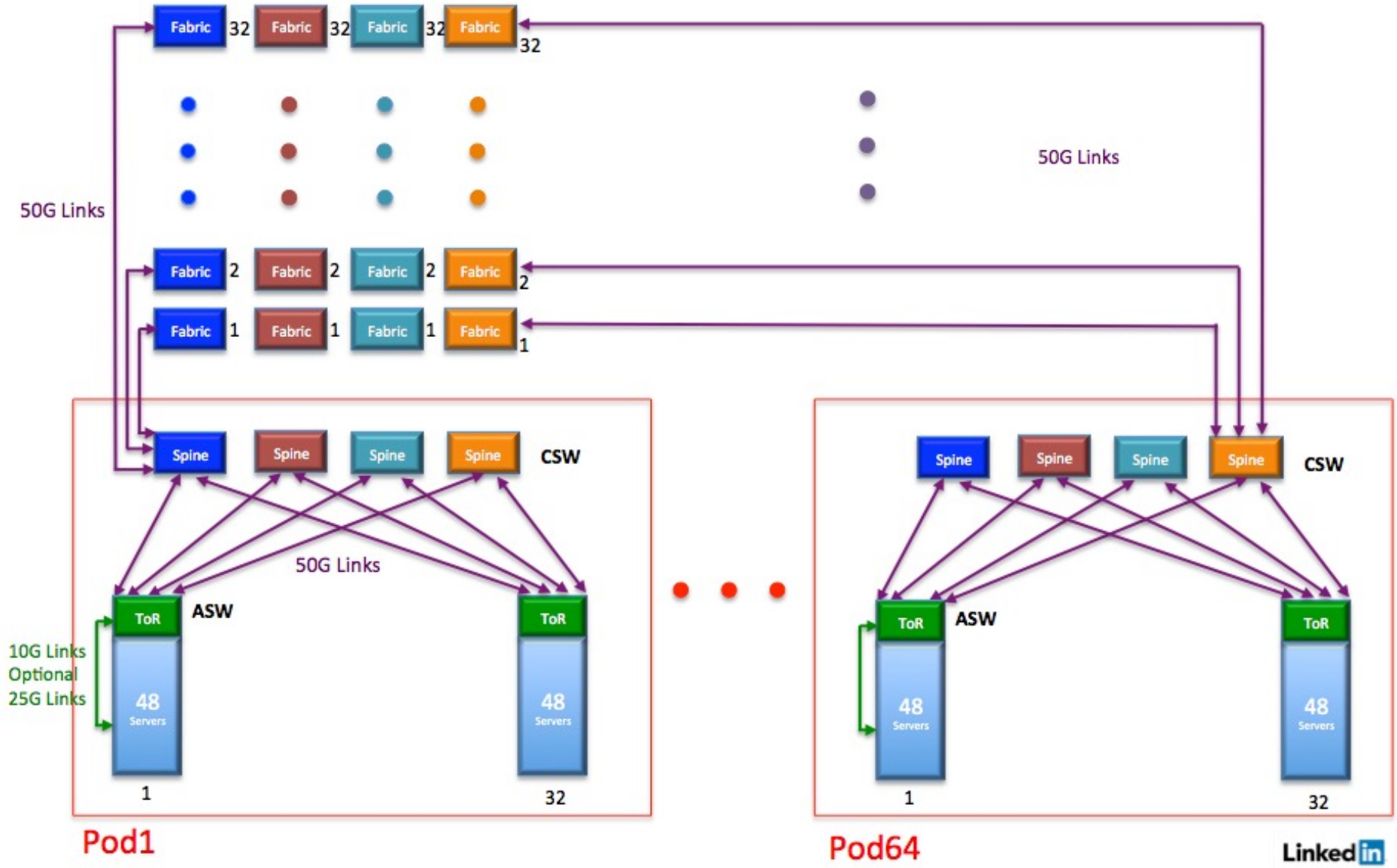
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network, SIGCOMM 2015

Google Jupiter Topology



# of servers: 64 (aggregation) * 32 (tor) * 32 (servers) = 65536
# of switches: 6 (spine) + 64 (aggregation) * [8*4 (middle) +32 (tor)] = 4102
Each server has 40Gbps, bisection bandwidth among servers: 1.3Pbps

LinkedIn DCN Topology



LinkedIn DCN Topology