

MANAGEMENT OF INFORMATION SYSTEMS

BME VIK TMIT
SOFTWARE ENGINEERING, BSc



BME VIK TMIT

MANAGEMENT OF INFORMATION SYSTEMS

7. DATA CENTERS

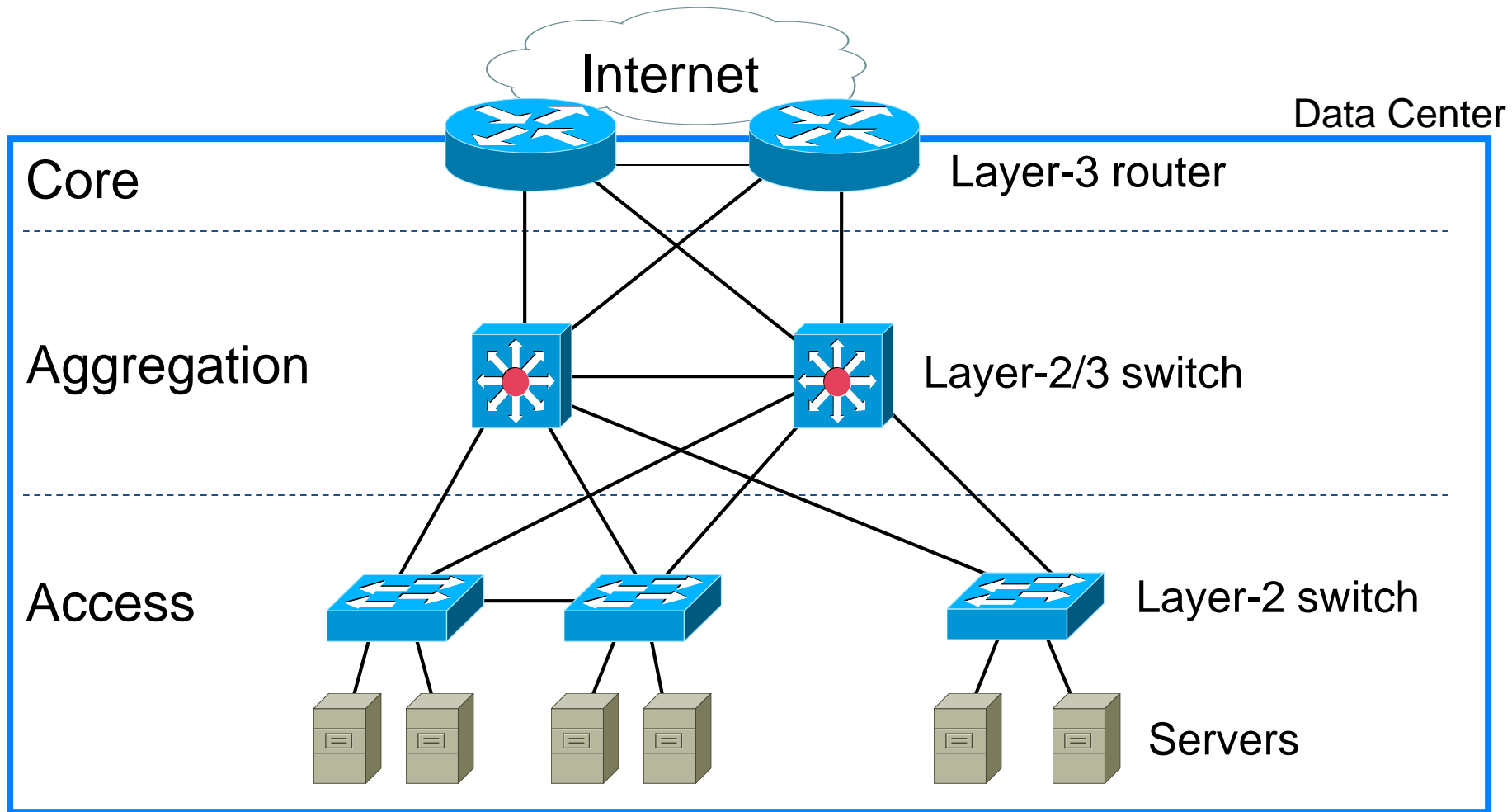


DATA CENTERS

- A **data center** is a facility used to house computer systems and associated components
- Nowadays, data center consisting of tens of thousands of PCs are increasingly common in universities, research labs, and companies
- Data center can be used to
 - do scientific computing,
 - financial analysis,
 - data analysis and warehousing,
 - and providing large-scale network services



COMMON DATA CENTER TOPOLOGY

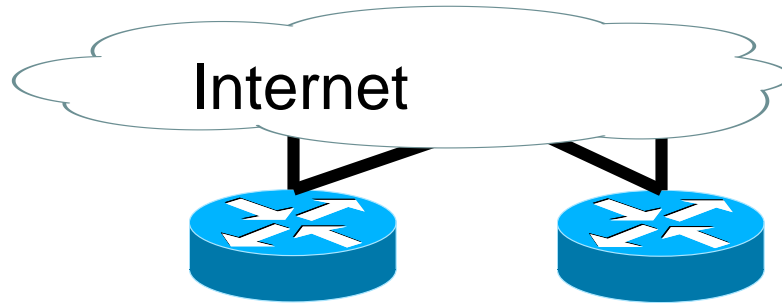


MULTI-TIER MODEL

- Access
 - Where hosts connect to the network
- Aggregation
 - To which the access layer is redundantly connected
- Core
 - Provides routing services
 - “Connects DC to external world”
 - to other parts of the data center
 - to services outside of the data center such as the Internet
 - geographically separated data centers
 - and other remote locations



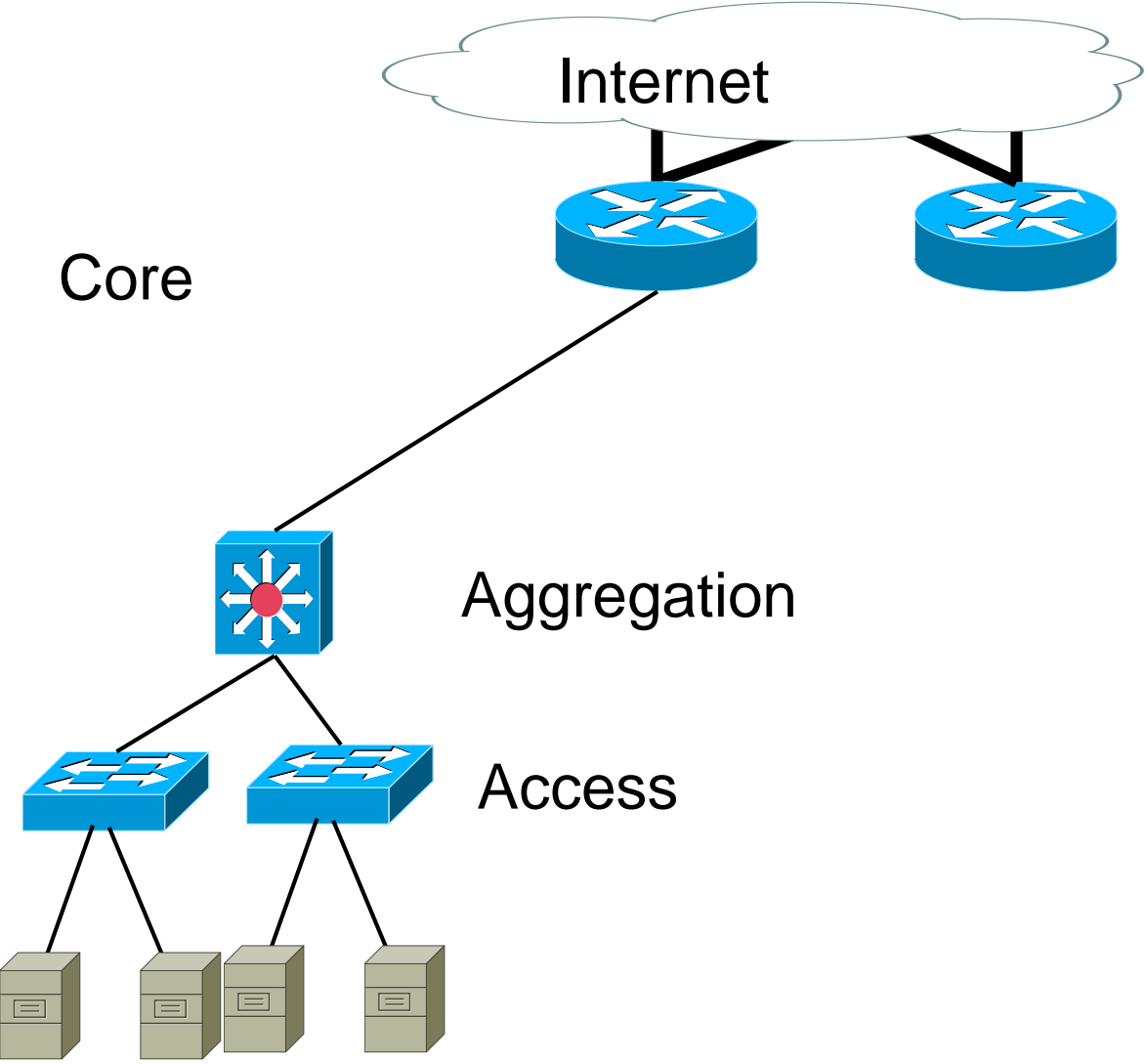
MULTI-TIER MODEL



Core



MULTI-TIER MODEL



Core

Internet

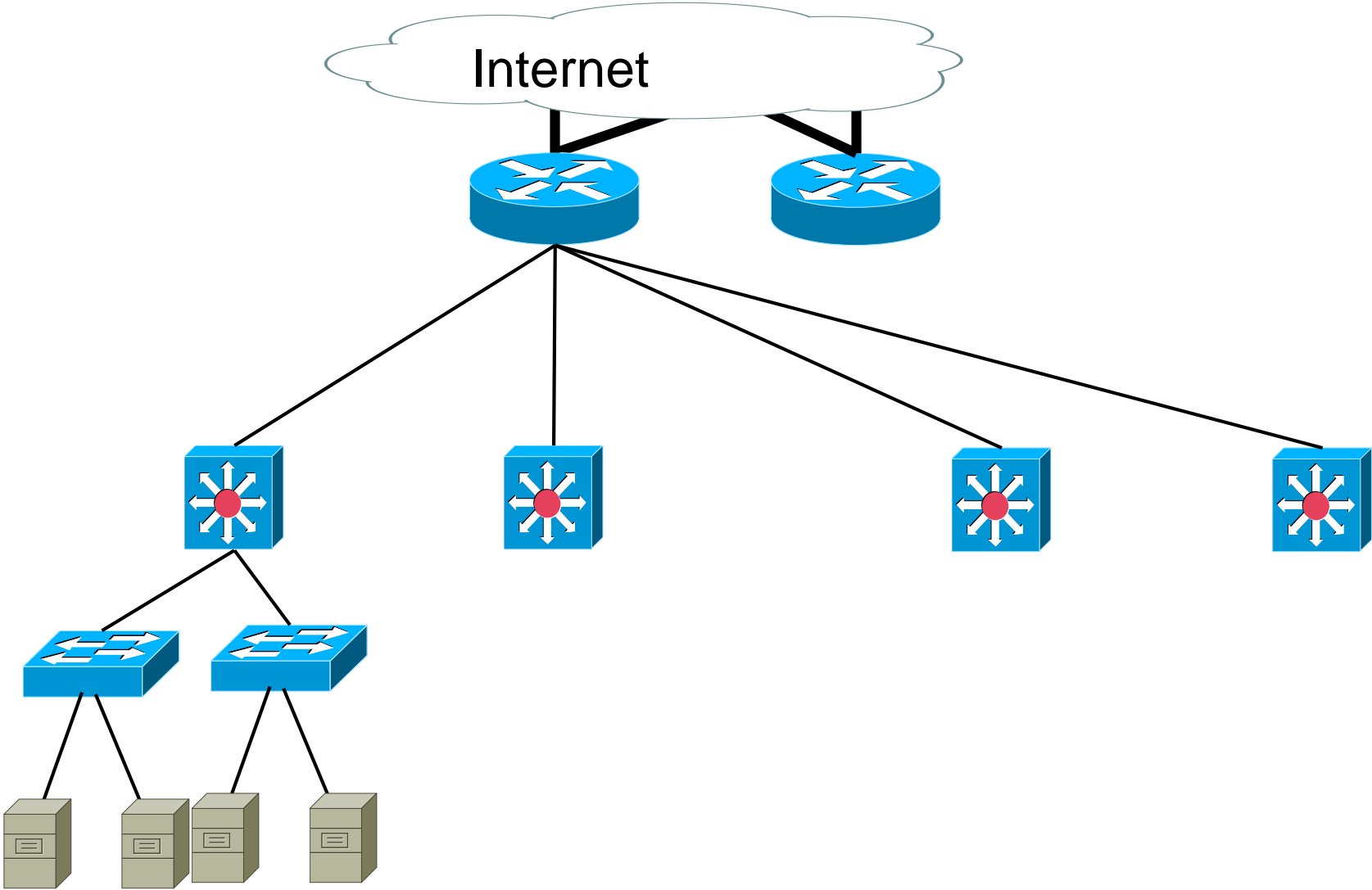
Aggregation

Access

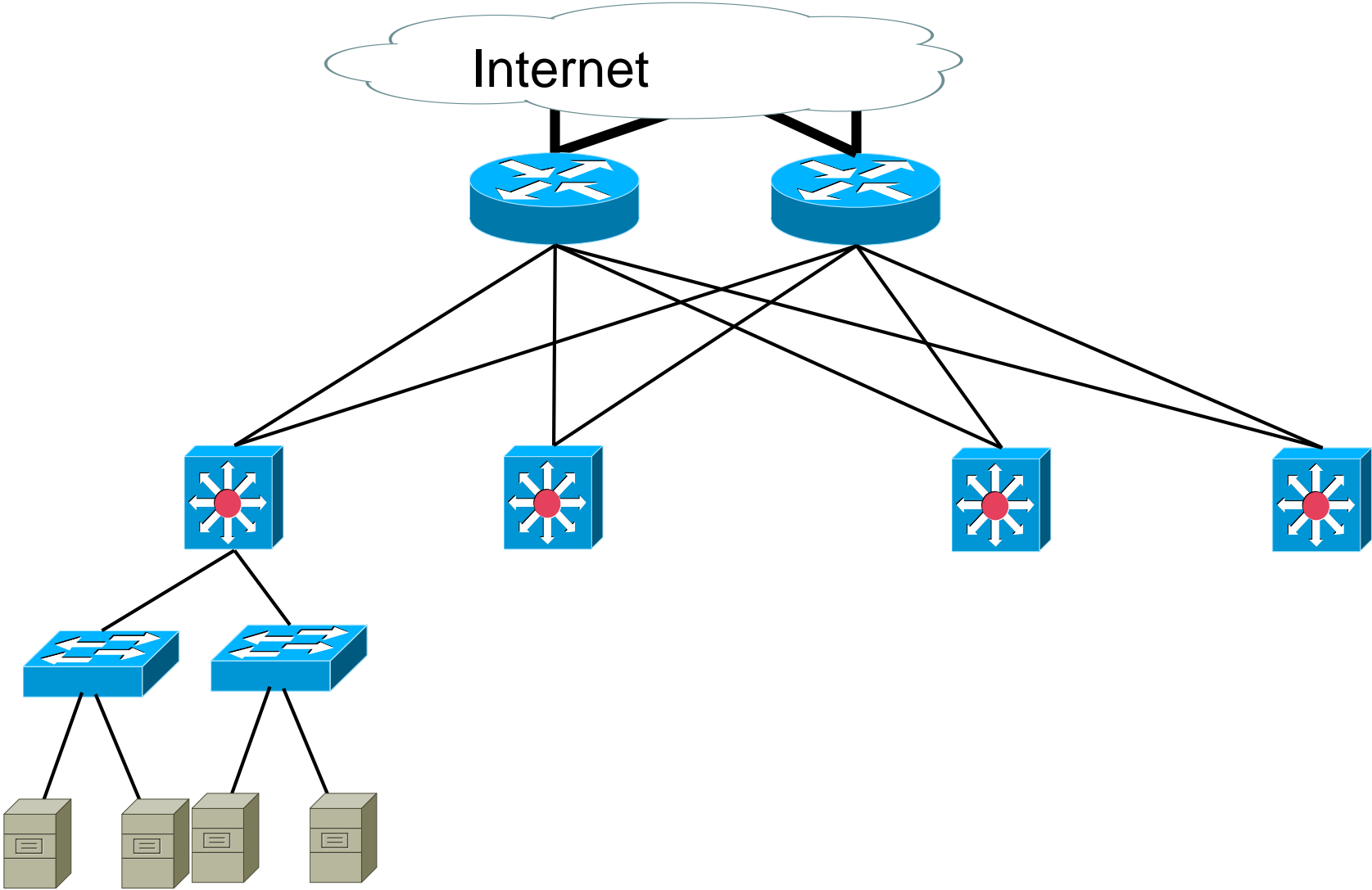
Servers



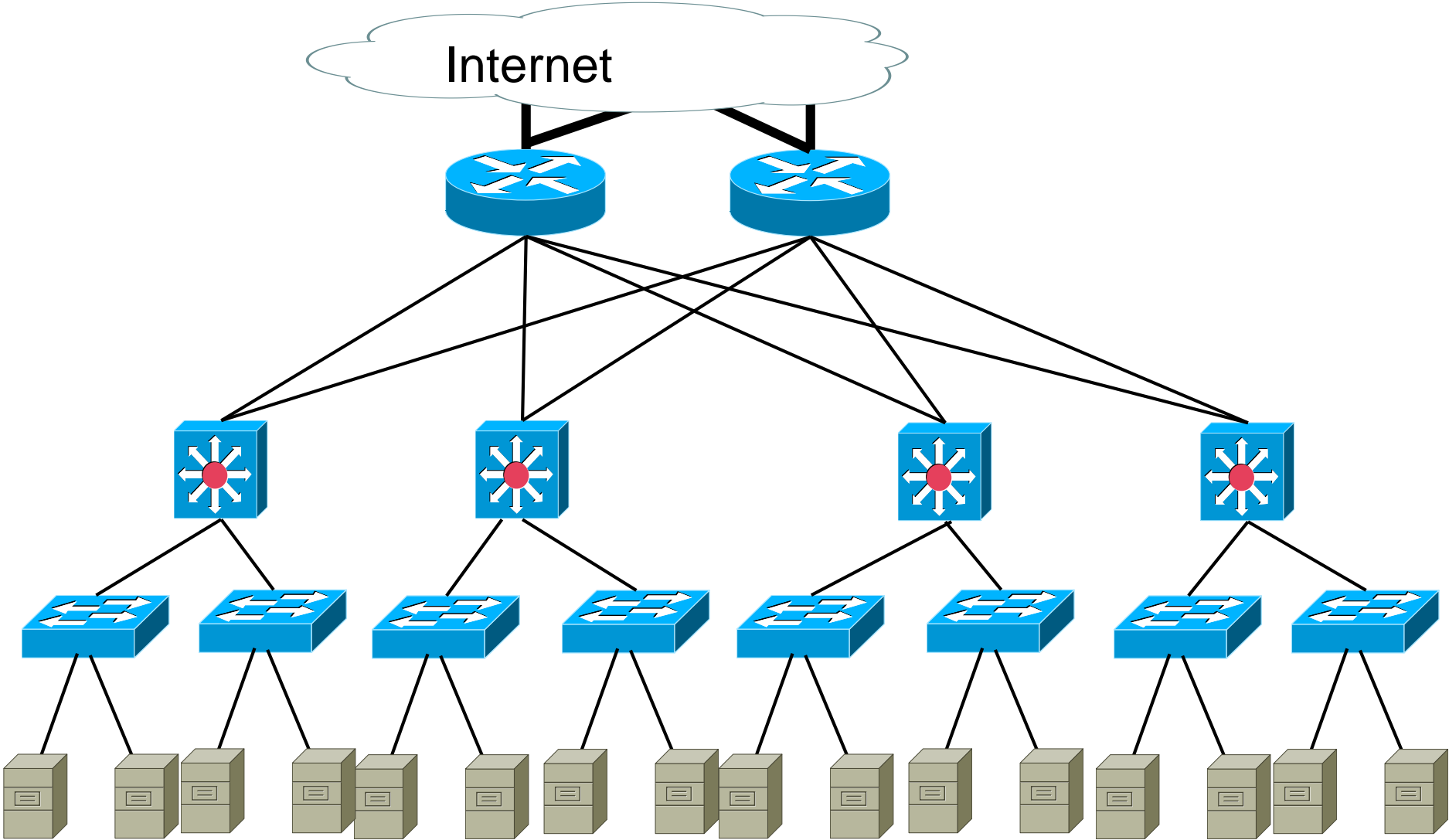
MULTI-TIER MODEL



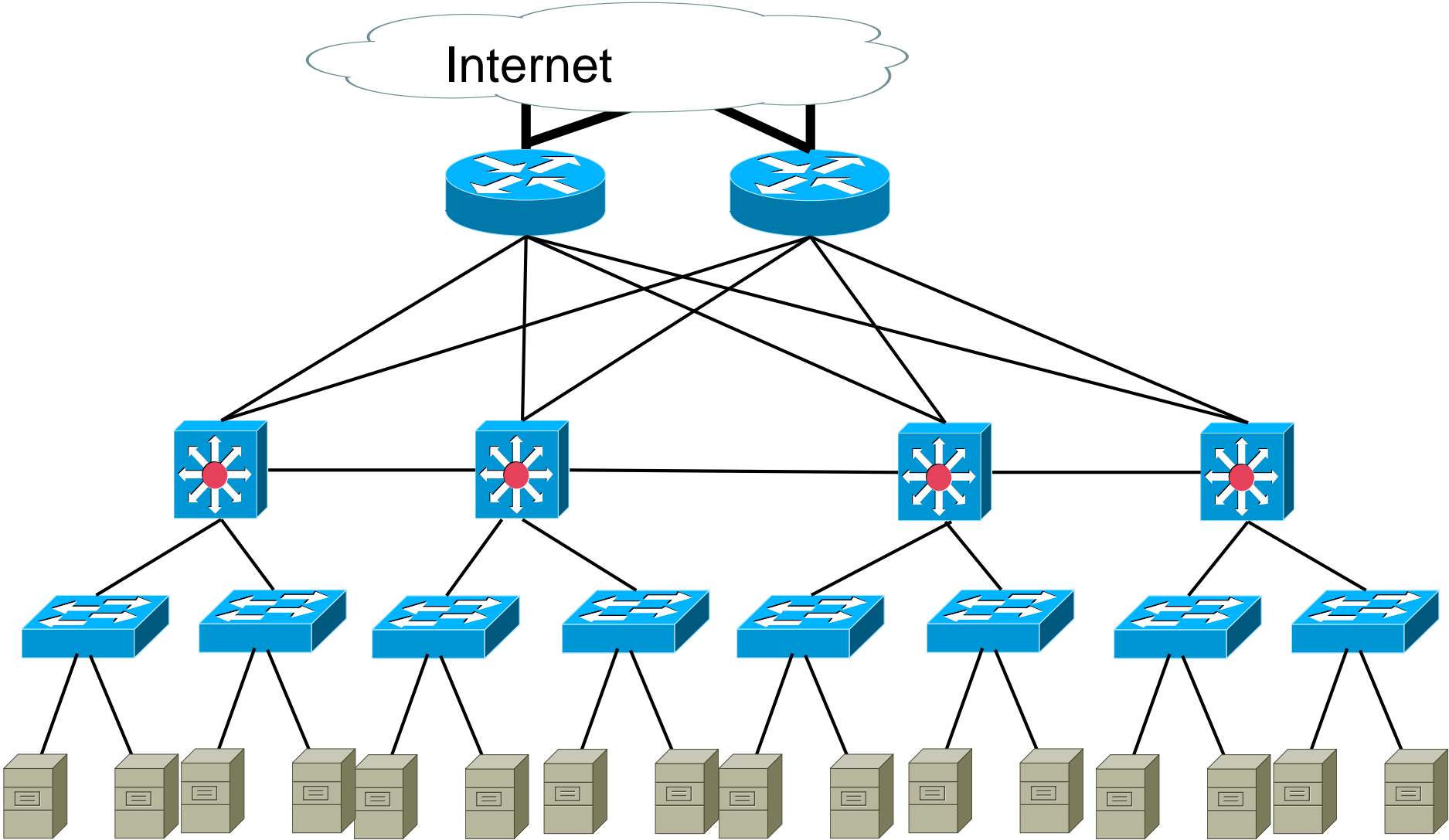
MULTI-TIER MODEL



MULTI-TIER MODEL

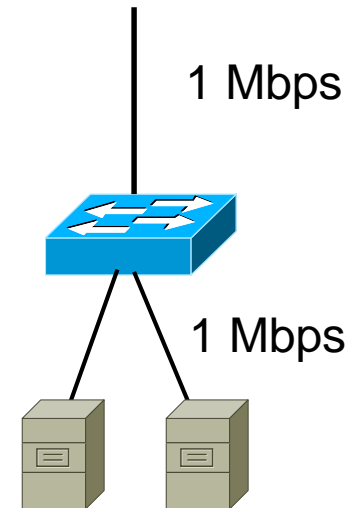


MULTI-TIER MODEL



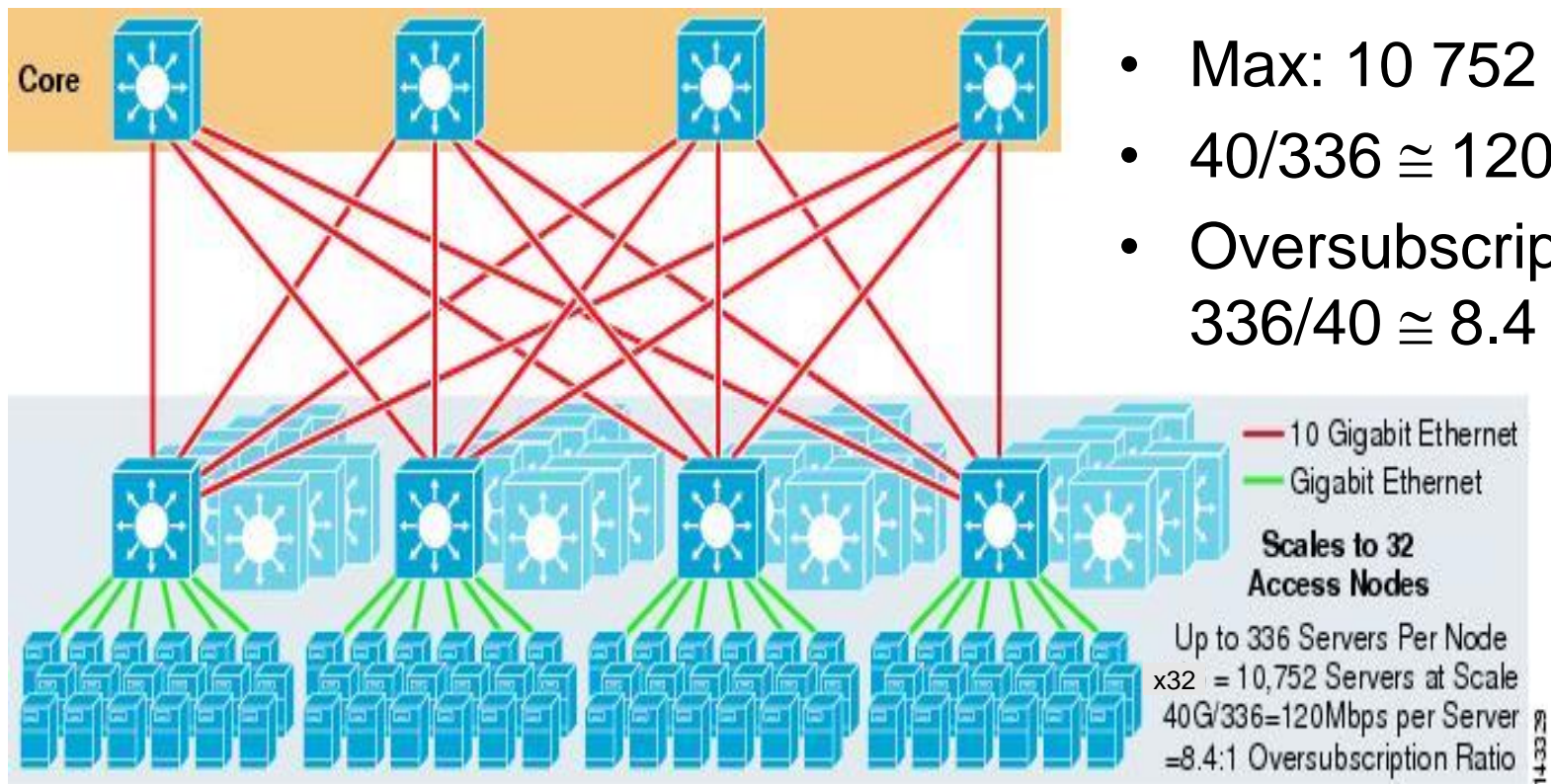
OVERSUBSCRIPTION

- Total output bandwidth / input bandwidth (down/up)
- $2 * 1 \text{ Mbps} / 1 \text{ Mbps} = 2$
- Not likely that all the bandwidth is needed at the same time



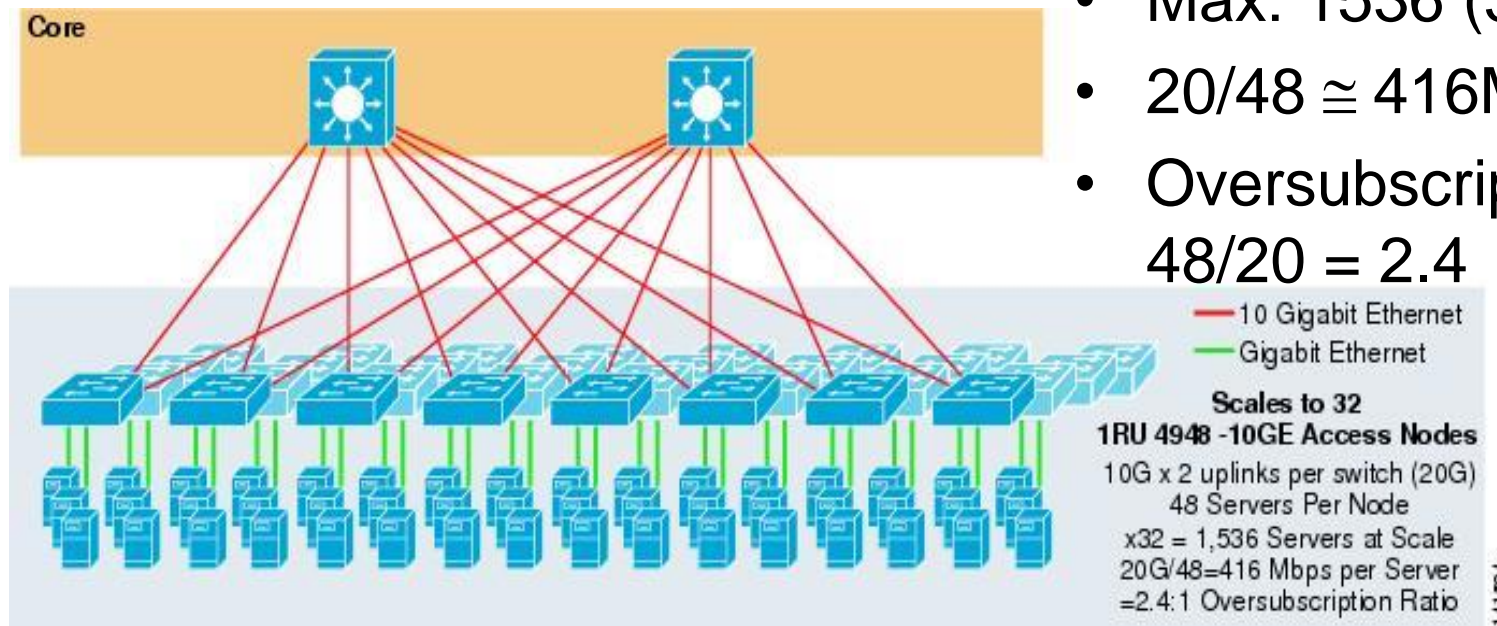
2-TIER MODEL

- Core: $32 \times 10\text{Gbit/s}$
- Access:
 $4 \times 10\text{G} + 336 \times 1\text{G}$
- Max: 10 752 (32×336)
- $40/336 \cong 120\text{M} / \text{server}$
- Oversubscription:
 $336/40 \cong 8.4$



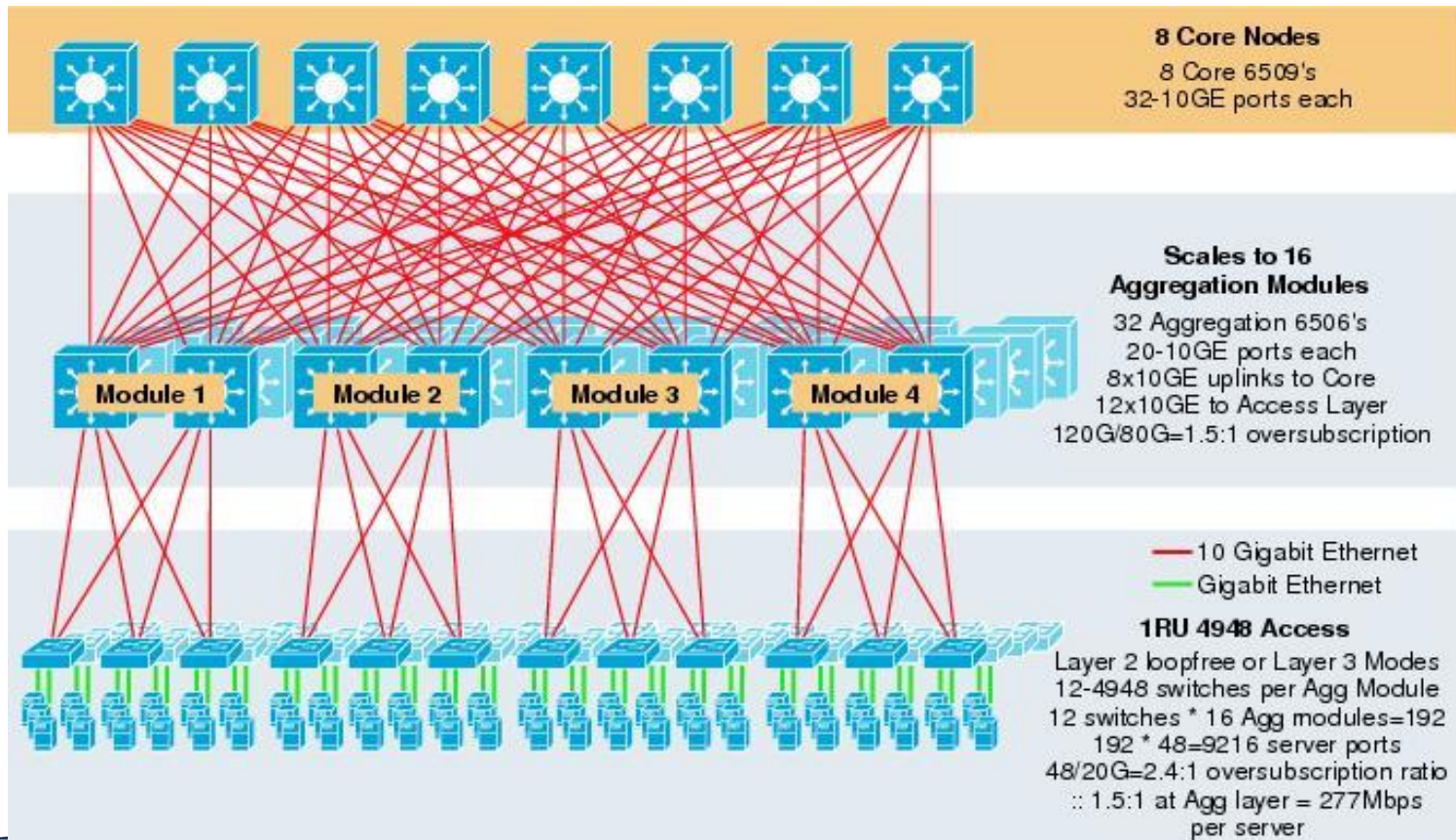
2-TIER WITH ToR (1RU) SWITCHES

- ToR – Top of Rack
- Core: 32*10G
- Access:
Top of Rack 2*10G +
48*1G
- Max: 1536 (32*48)
- $20/48 \cong 416\text{M} /\text{server}$
- Oversubscription:
 $48/20 = 2.4$



3-TIER MODEL

- Max. 8 core switch;
- 16 aggr. module, with 12 access switch each * 48 server = 9216 server
- $(2*80)/(12*48) \cong 277\text{M}/\text{server}$
- Overs.: $1.5*2.4=3.6$



141595



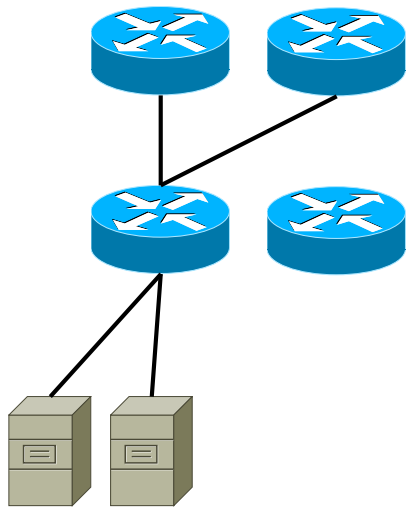
MULTI-TIER -> FAT TREE

- Large, special routers / switches
- -> Smaller, commercial
- -> Fat Tree Topology

- In the example, switches with $4 * 1\text{Gbit/s}$ ports are used
- Rule: half of the ports downward, other half upward



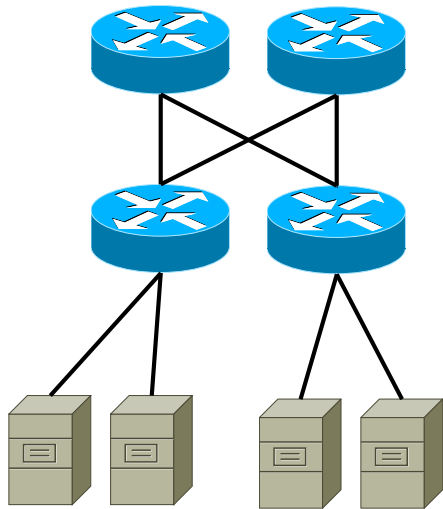
FAT TREE TOPOLOGY



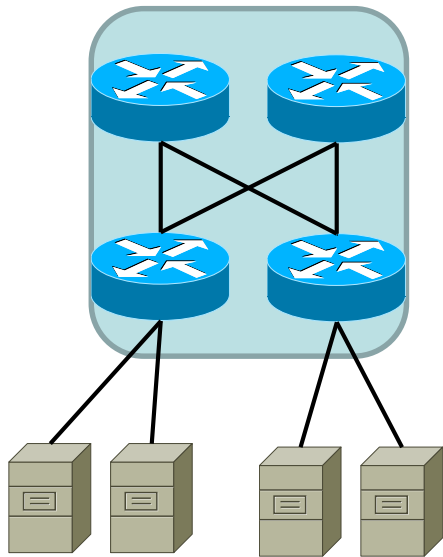
Rule: half of the ports downward, other half upward



FAT TREE TOPOLOGY



FAT TREE TOPOLOGY



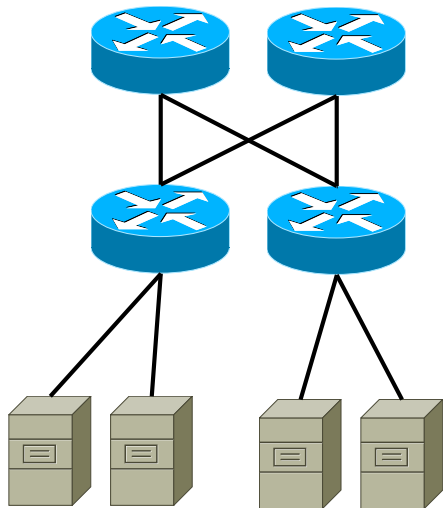
Pod



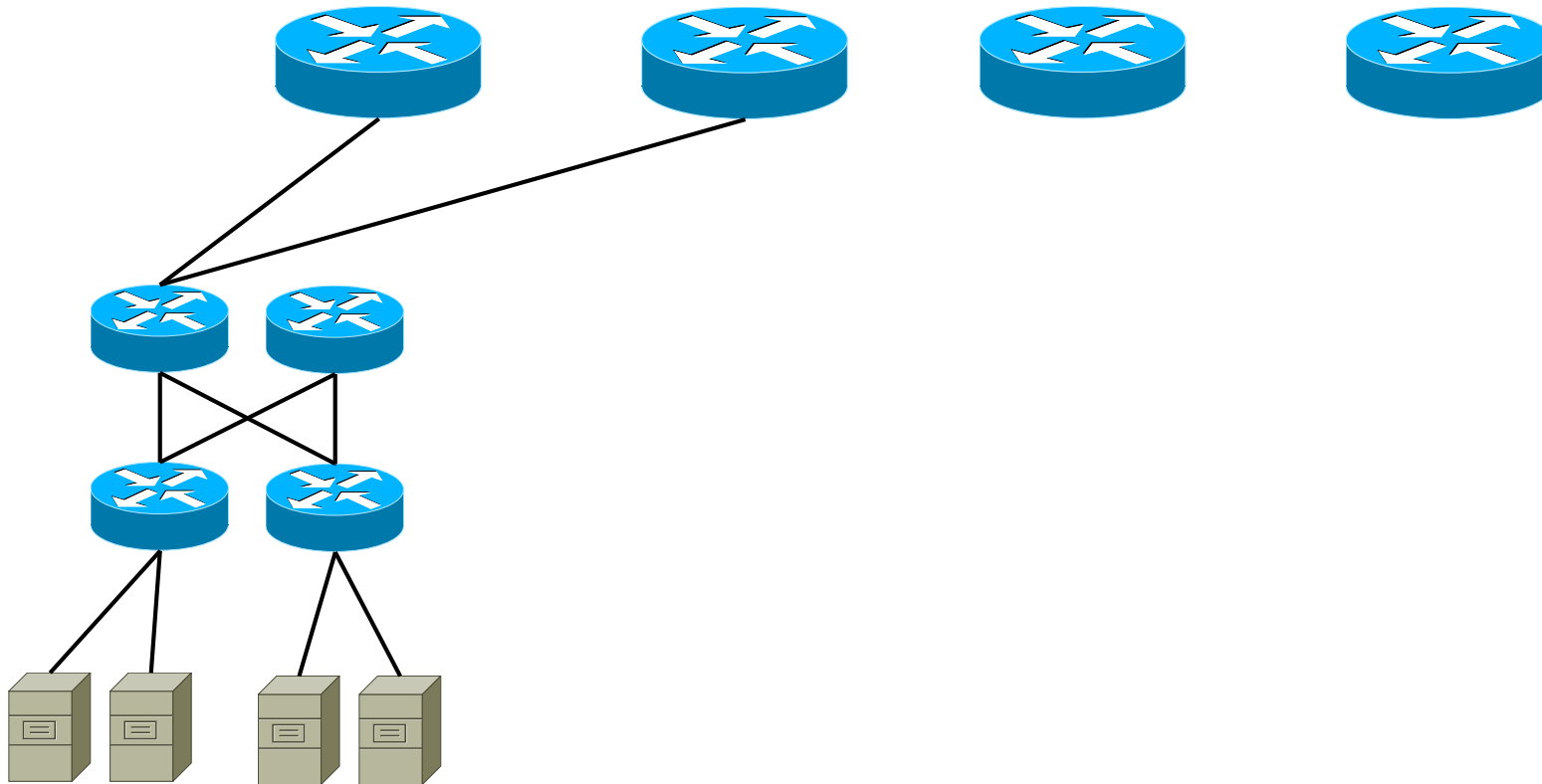
FAT TREE TOPOLOGY



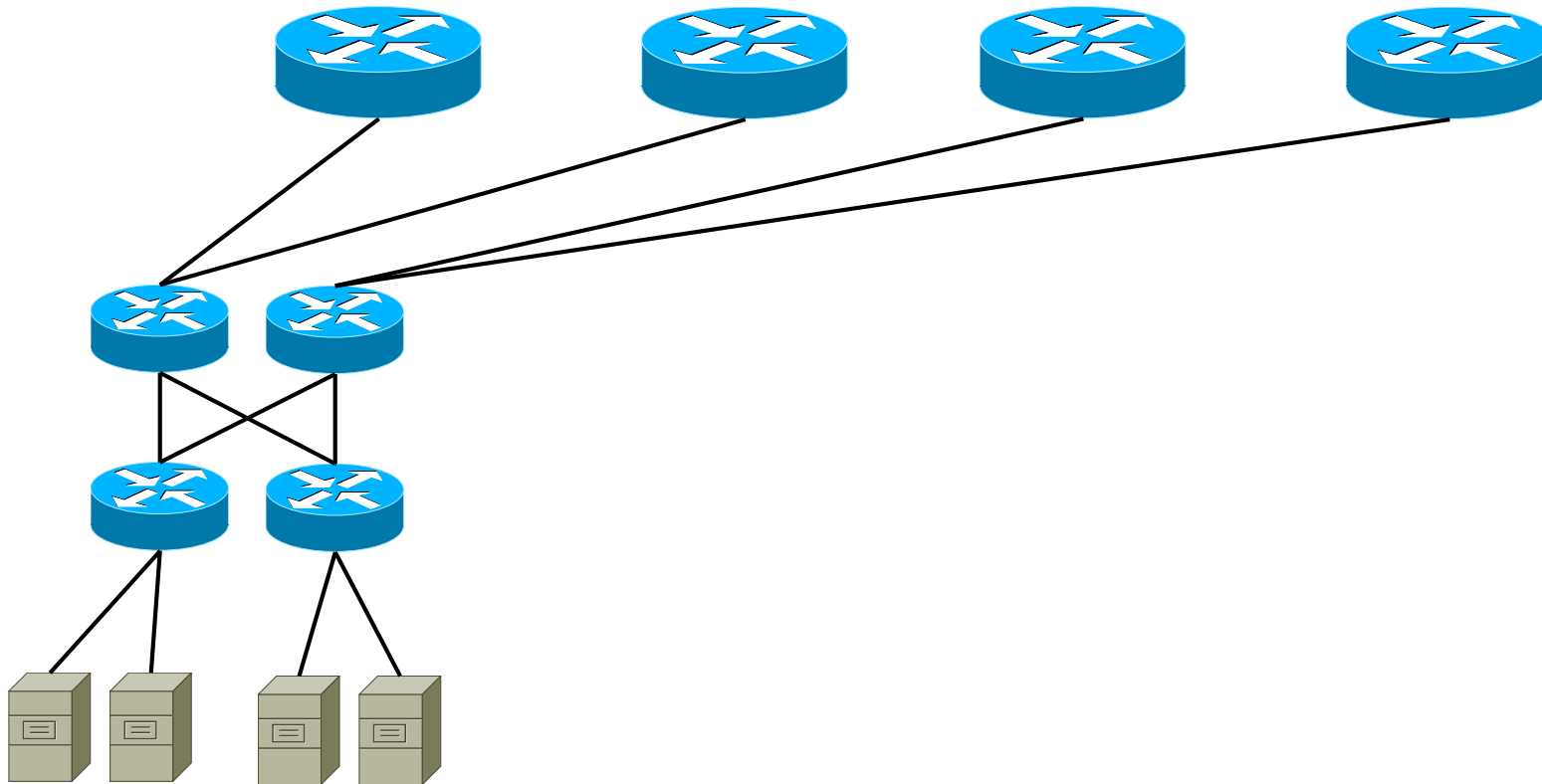
Rule: half of the ports downward, other half upward



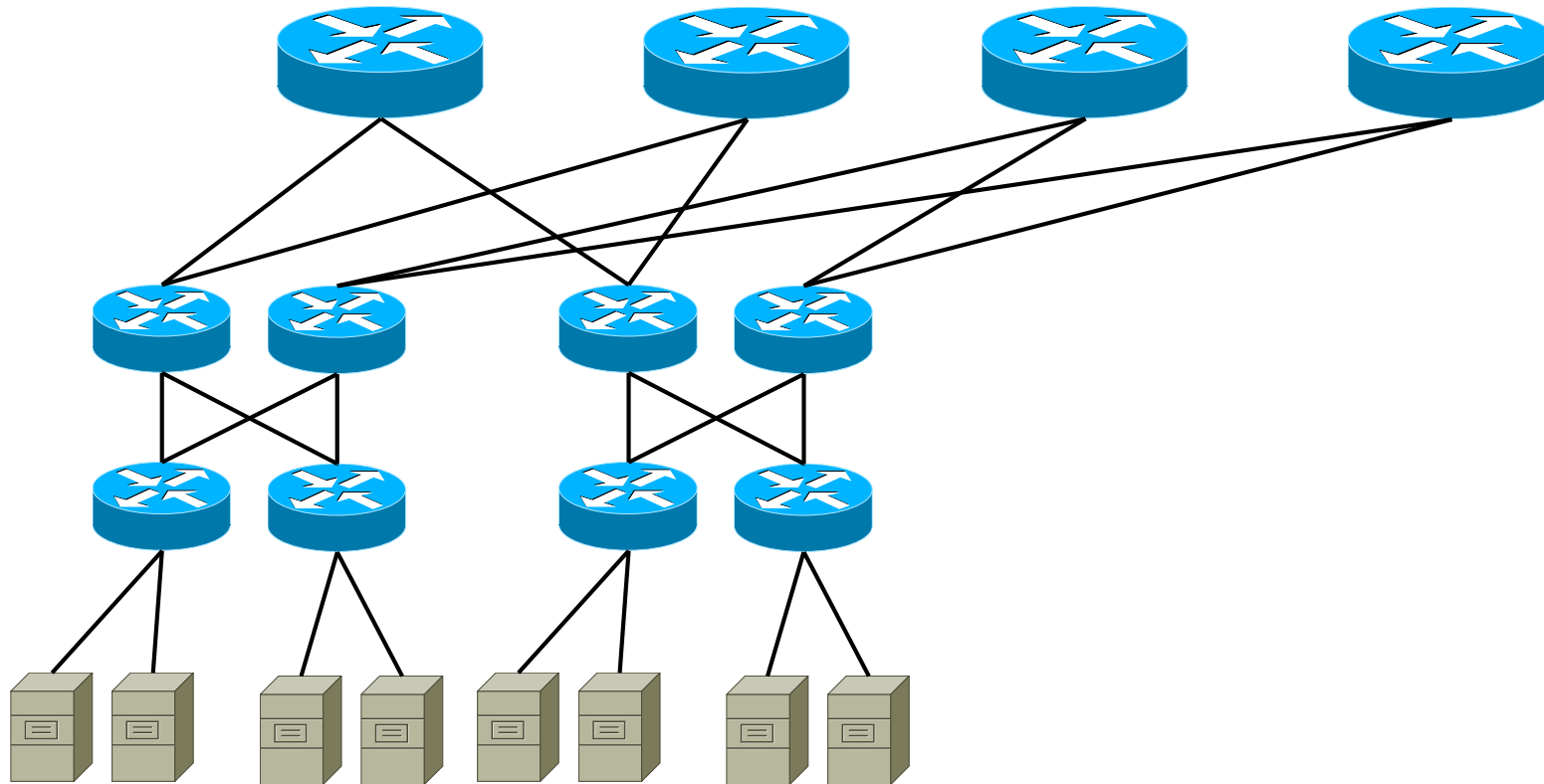
FAT TREE TOPOLOGY



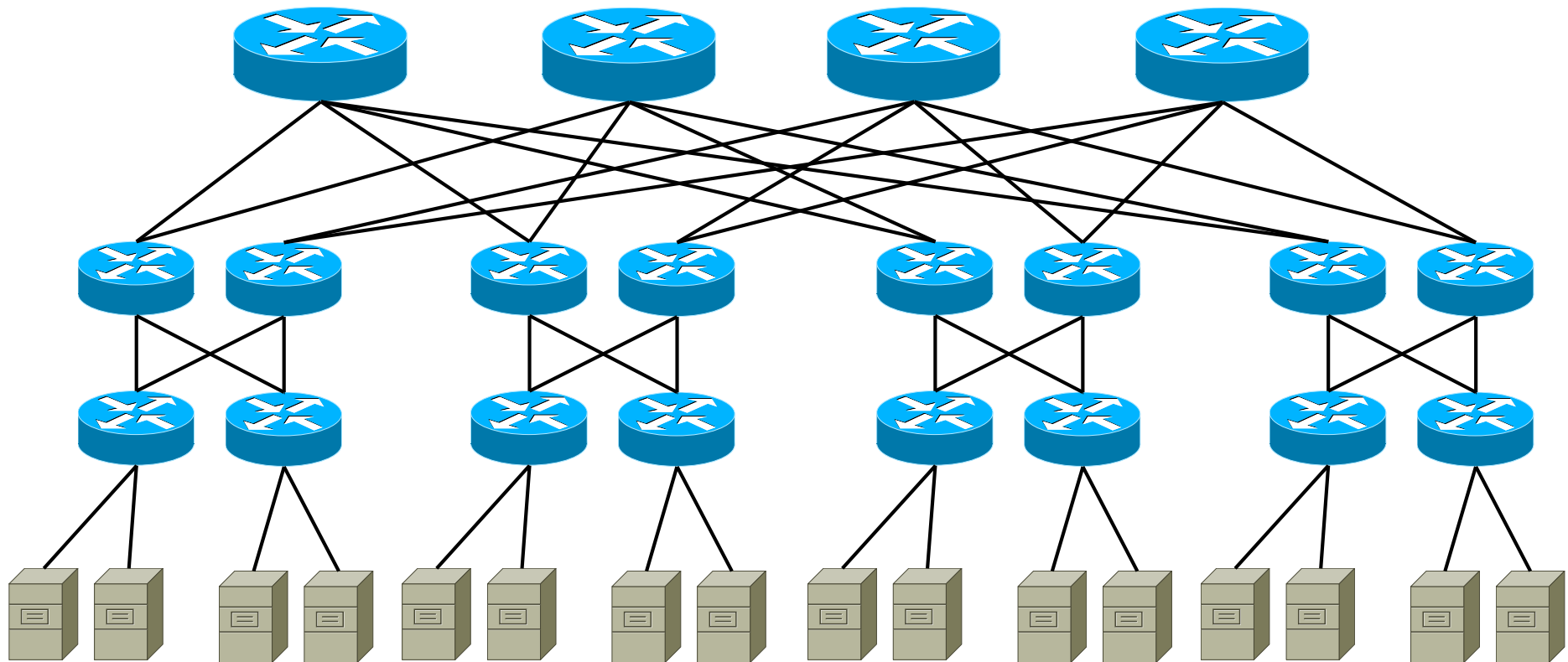
FAT TREE TOPOLOGY



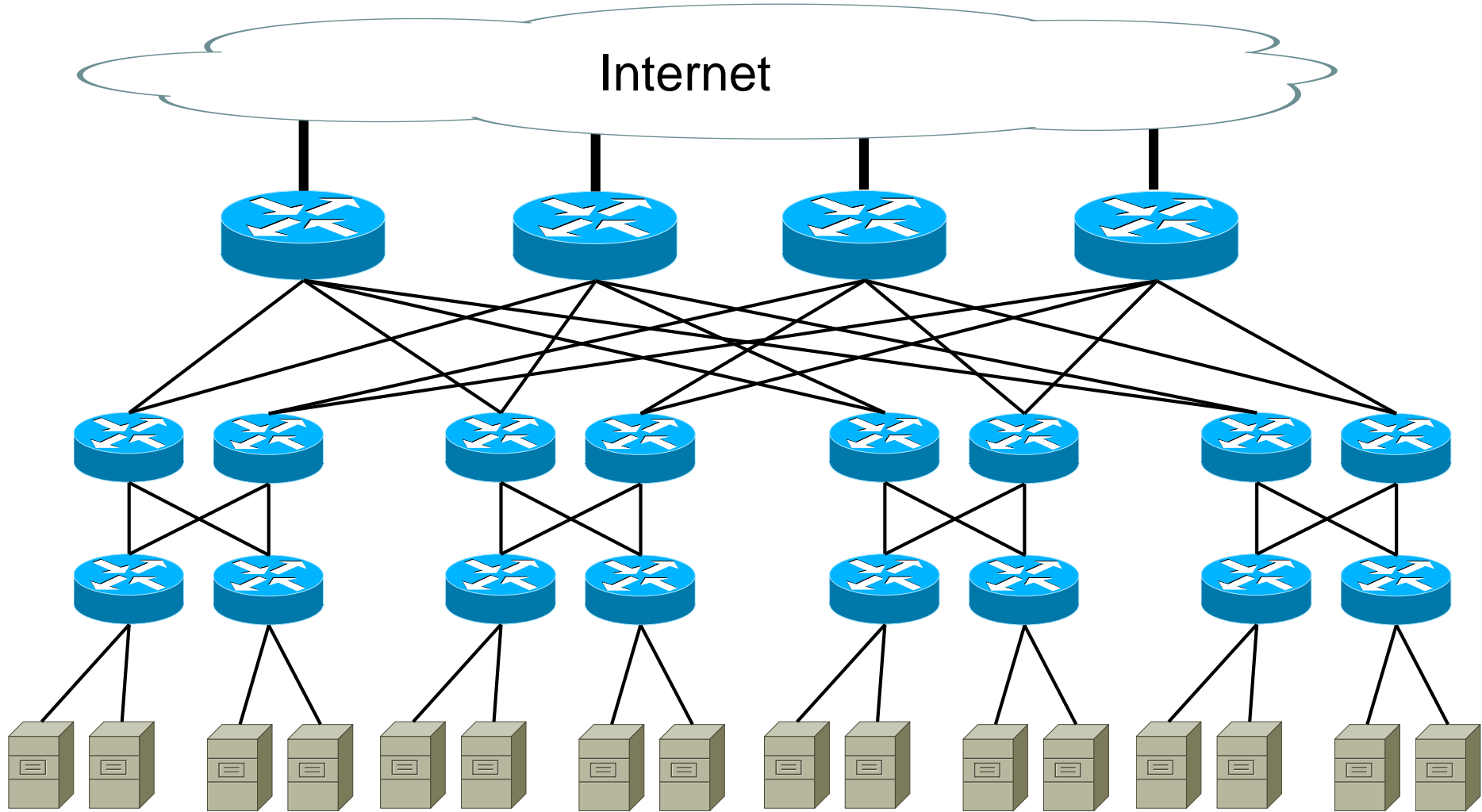
FAT TREE TOPOLOGY



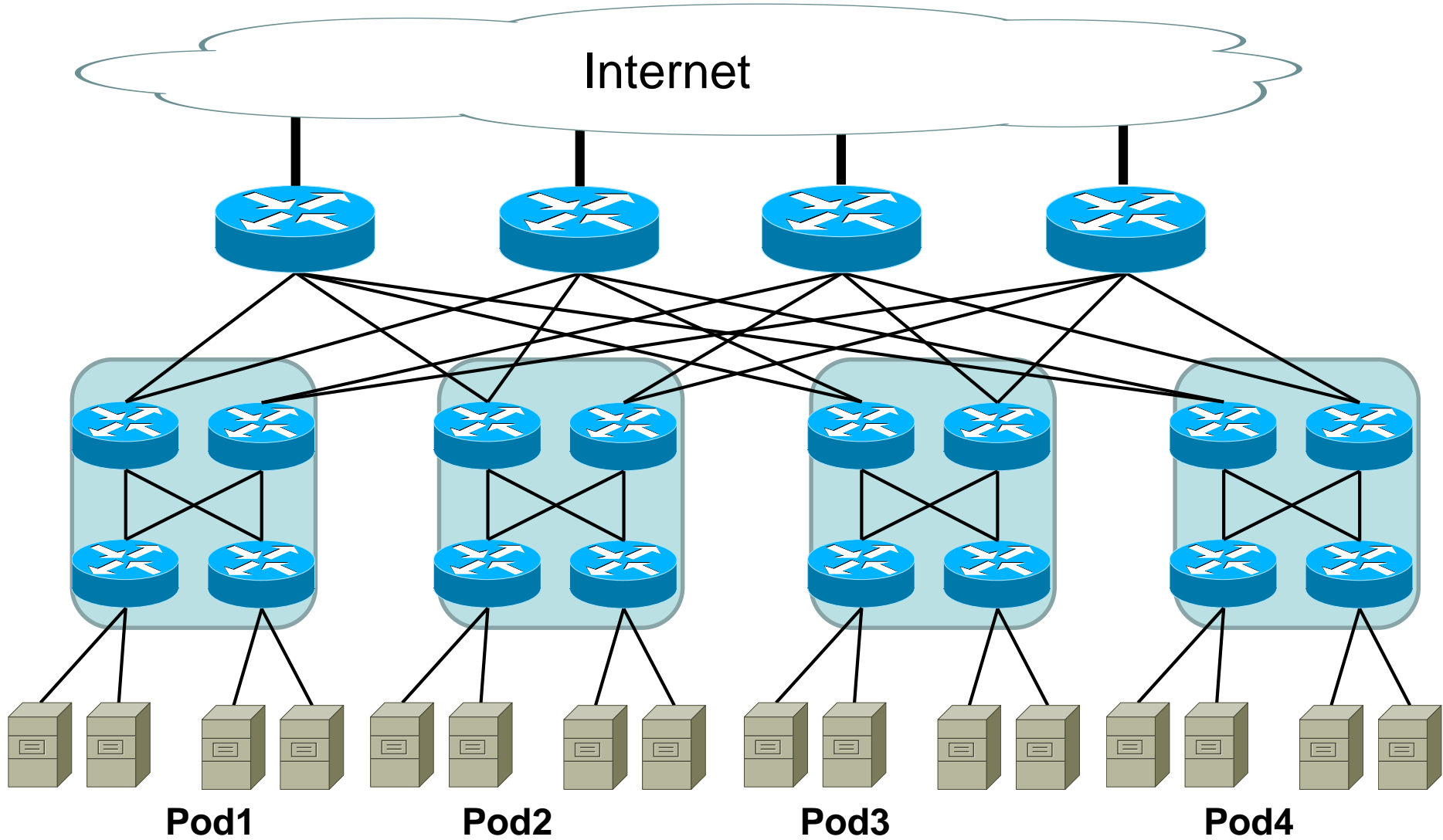
FAT TREE TOPOLOGY



FAT TREE TOPOLOGY

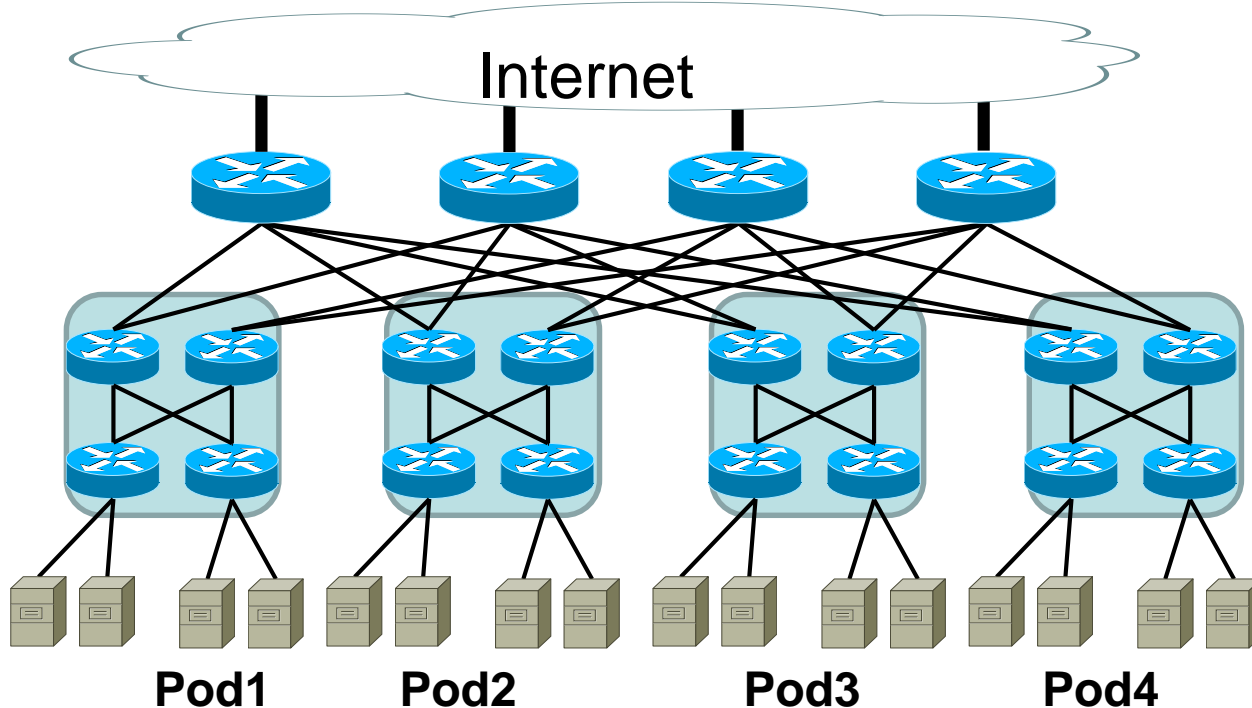


FAT TREE TOPOLOGY



FAT TREE TOPOLOGY

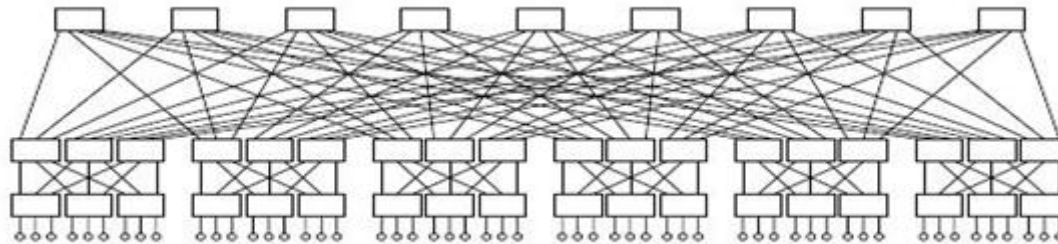
- 1 switch: k ports
 - k pods
- 1 pod contains:
 - $k/2$ aggregate switches
 - $k/2$ access switches
 - $(k/2) * (k/2) = (k/2)^2$ servers
- Servers = $k * (k/2)^2$
- $(k/2) * (k/2) = (k/2)^2$ core switches



FAT TREE TOPOLOGY

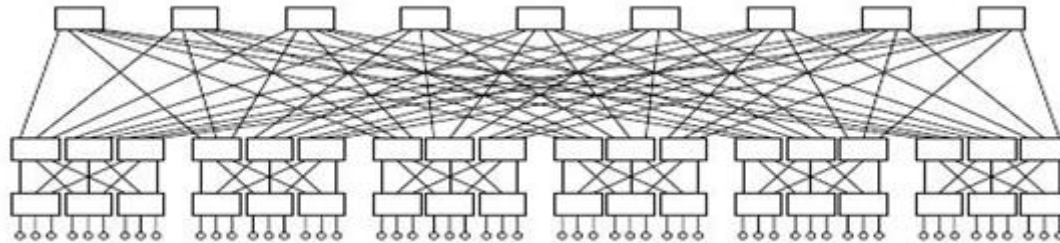
k-ary fat tree:

- three-layer topology (access, aggregation and core)
- each switch has k ports (-> k pods)
 - half of them up, other half down
- each access switch connects to $k/2$ servers & $k/2$ aggregation switches
- each aggregation switch connects to $k/2$ access & $k/2$ core switches
- $(k/2)^2$ core switches: each connects to k pods
- each pod consists of $(k/2)^2$ servers & 2 layers of $k/2$ k-port switches
- altogether $k * (k/2)^2 = k^3/4$ servers



6-ARY FAT TREE TOPOLOGY

- each switch has 6 ports -> 6 pods
- each access switch connects to $6/2 = 3$ servers & $6/2 = 3$ aggregation switches
- each aggregation switch connects to $6/2 = 3$ access & $6/2 = 3$ core switches
- $(6/2)^2 = 9$ core switches: each connects to 6 pods
- each pod consists of $(6/2)^2 = 9$ servers & 2 layers of $6/2 = 3$ 6-port switches
- altogether: $6 * 2 * (6/2) = 36$ access+aggregation switches
- altogether $6 * (6/2)^2 = 6^3/4 = 54$ servers



64-ARY FAT TREE TOPOLOGY

- Number of servers?
- $k * (k/2)^2 = k^3/4$ servers = $64 * 32^2 = 65536$ servers



EVALUATION OF FAT TREE TOPOLOGY

- Bandwidth
 - Each layer has the same aggregated bandwidth
 - Oversubscription = 1
- Can be built using cheap devices with uniform capacity
 - Each port supports same speed as end host
 - All devices can transmit at line speed if packets are distributed uniform along available paths
- Great scalability: k -port switch supports $k^3/4$ servers
- Smaller power consumption
 - Lower heat / air condition



PROBLEMS OF THE TRADITIONAL ARCHITECTURES

- Change in the traffic pattern
 - Traditionally: north-south
 - Now: east-west as well
 - Why?
 - Virtualisation / Cloud
 - Any (virtual) server can be placed at any physical
 - Reallocation of workload
 - Communication is required between them: SAME speed between ANY cluster
 - Higher reliability

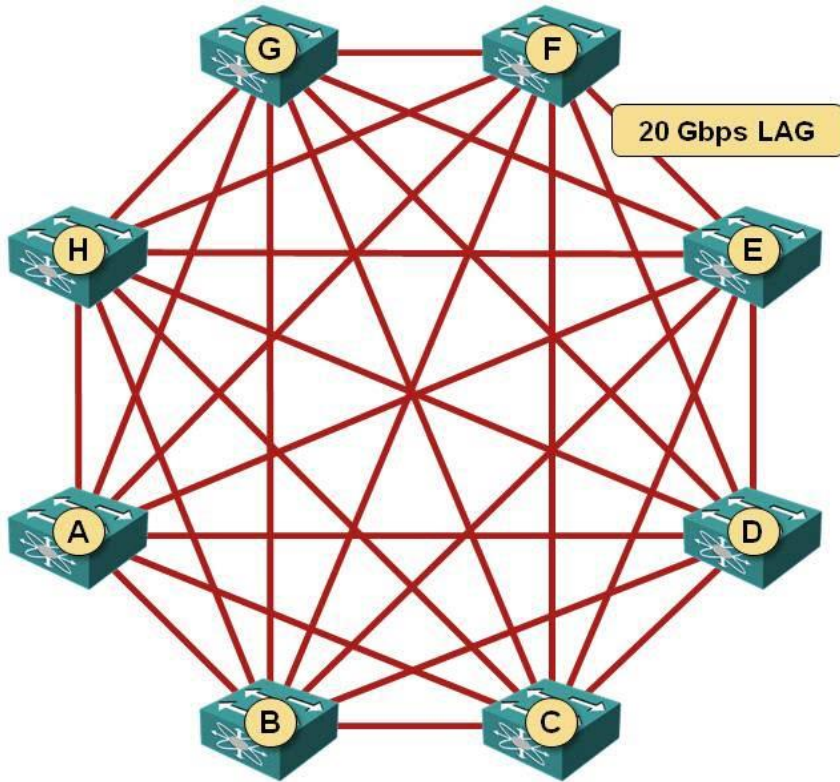


FULL MESH

- Typical data center switches:
 - 48 x 10GE ports and
 - 4 x 40GE ports that can be used as 16 x 10GE ports (160 GE)
 - Altogether 64*10GE ports (640 GE)
- Use:
 - 48 ports for servers
 - 16 (14) ports for intra fabric connections
- Oversubscription = ?
 - 3:1



FULL MESH

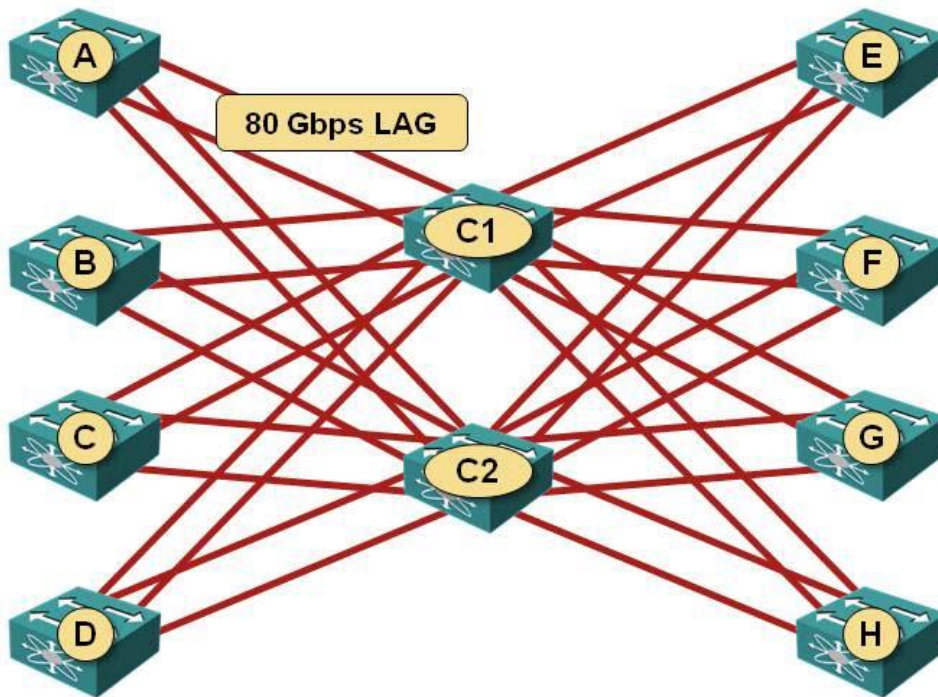


- Though we have 140 Gbps uplink capacity, we can achieve only 20 Gbps between any two nodes
- No alternative routes
 - Error prone
- Lot of $(n*(n-1)/2)$ links

LAG: Link aggregation



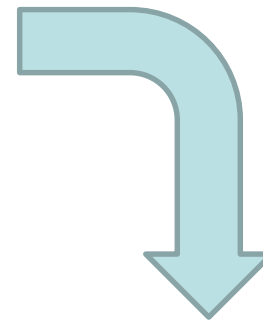
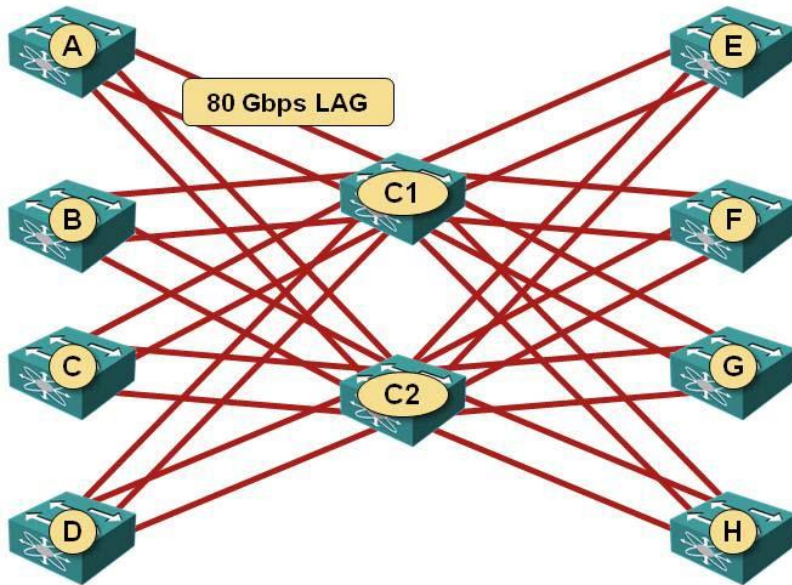
MORE EFFICIENT ARCHITECTURE



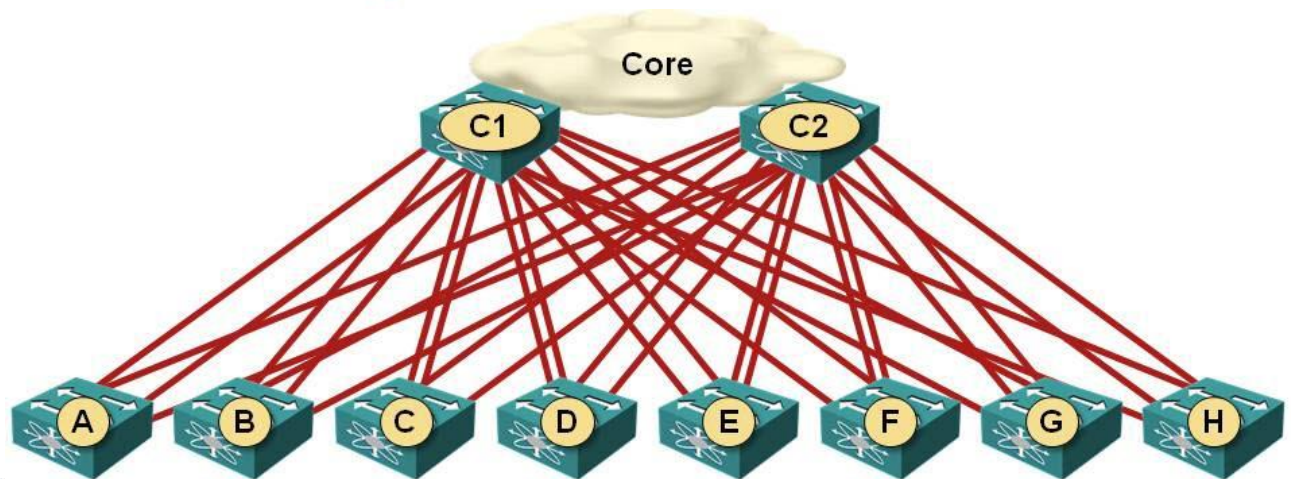
- 160 Gbps between any two nodes
- Redundant
- BUT:
 - more switches
 - slower



DRAW IT IN A MODIFIED WAY...

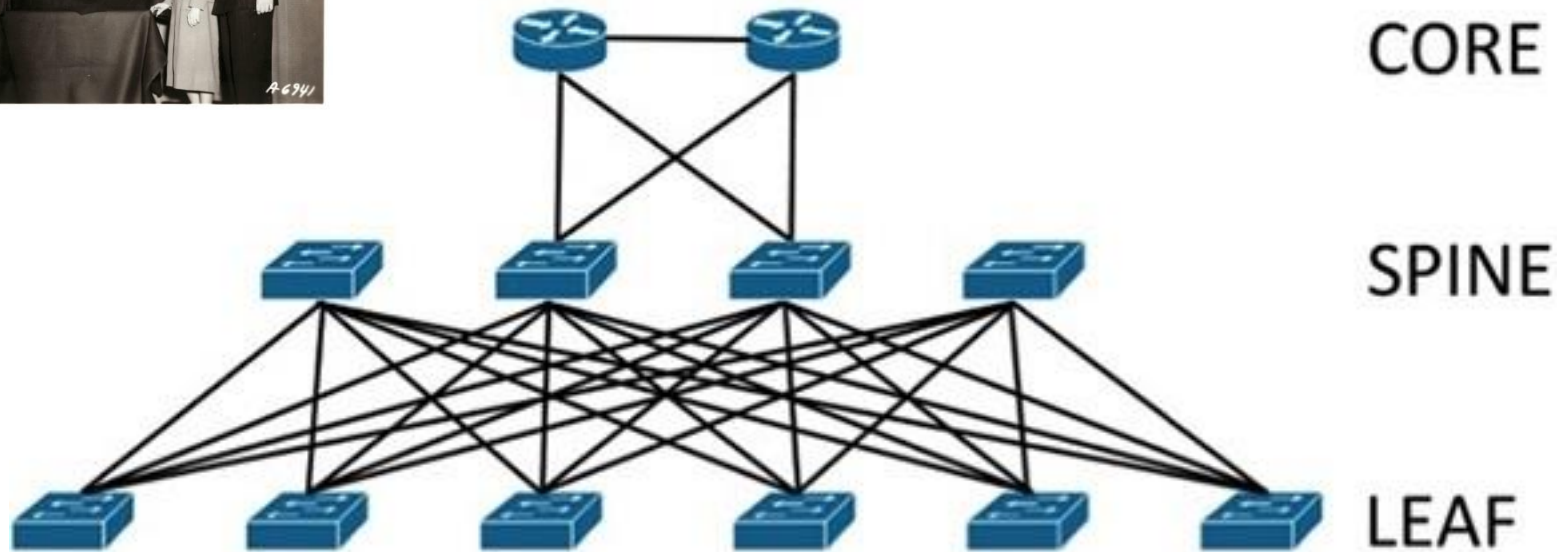


Folded array



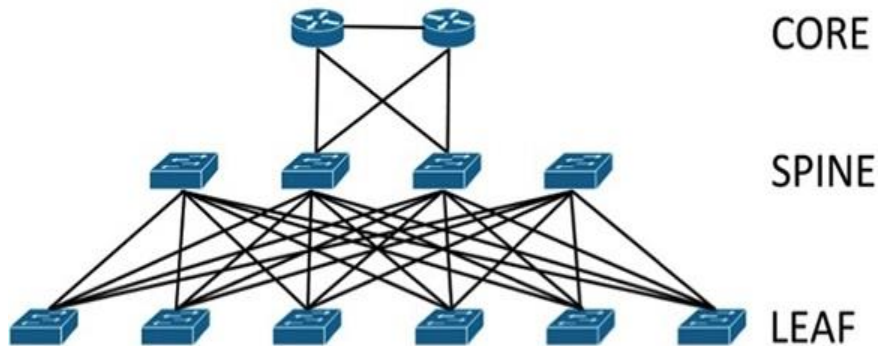
NEW TREND

- Clos Network / Folded Array / Spine and Leaf architecture



SPINE AND LEAF RULES

- Total # of interconnections = $\#leaf * \#spine$

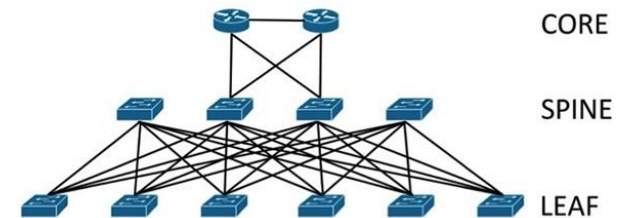


- $\#spine = \#port_of_a_leaf$
- $\#port_of_a_spine = \#leaf$



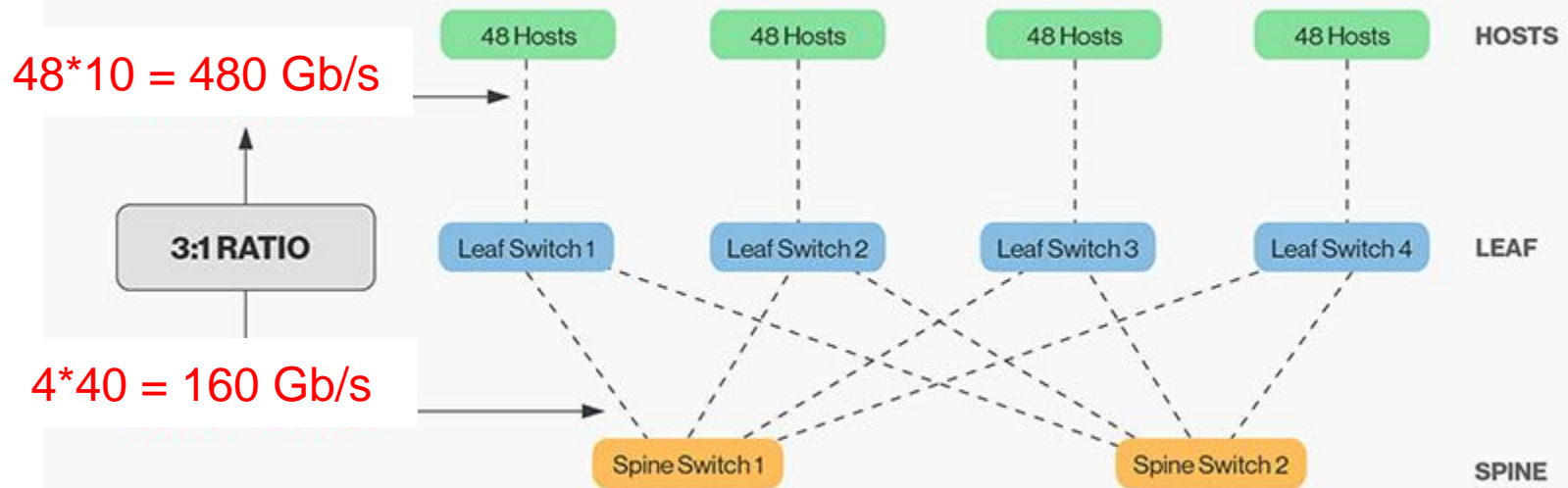
SPINE AND LEAF ARCHITECTURE

- Leaf ~ access
- Spine ~ aggregation
- But: **each leaf is connected to each spine**
 - Ethernet fabric
 - Maximum distance: 3 hops (4 links)
 - High performance clusters (HPC)



TYPICAL ARCHITECTURE

- Leaf:
 - To Servers: 48 * 10 GE ports
 - To Spine: 4 * 40 GE ports
- Oversubscription: 3:1



LOAD BALANCING, PARALLEL PATHS

- Traditional routing algorithm – Spanning Tree Protocol (STP)
 - To avoid loops finds the best path to every node
 - What is the main disadvantage?
 - Only ONE path between any two nodes
- Equal-Cost Multipath (ECMP) protocol
 - Finds more paths with the same length (cost)
 - In practice typically 8 or 16
 - Load distribution
 - Higher bandwidth utilization
 - BUT: Packet reordering needed



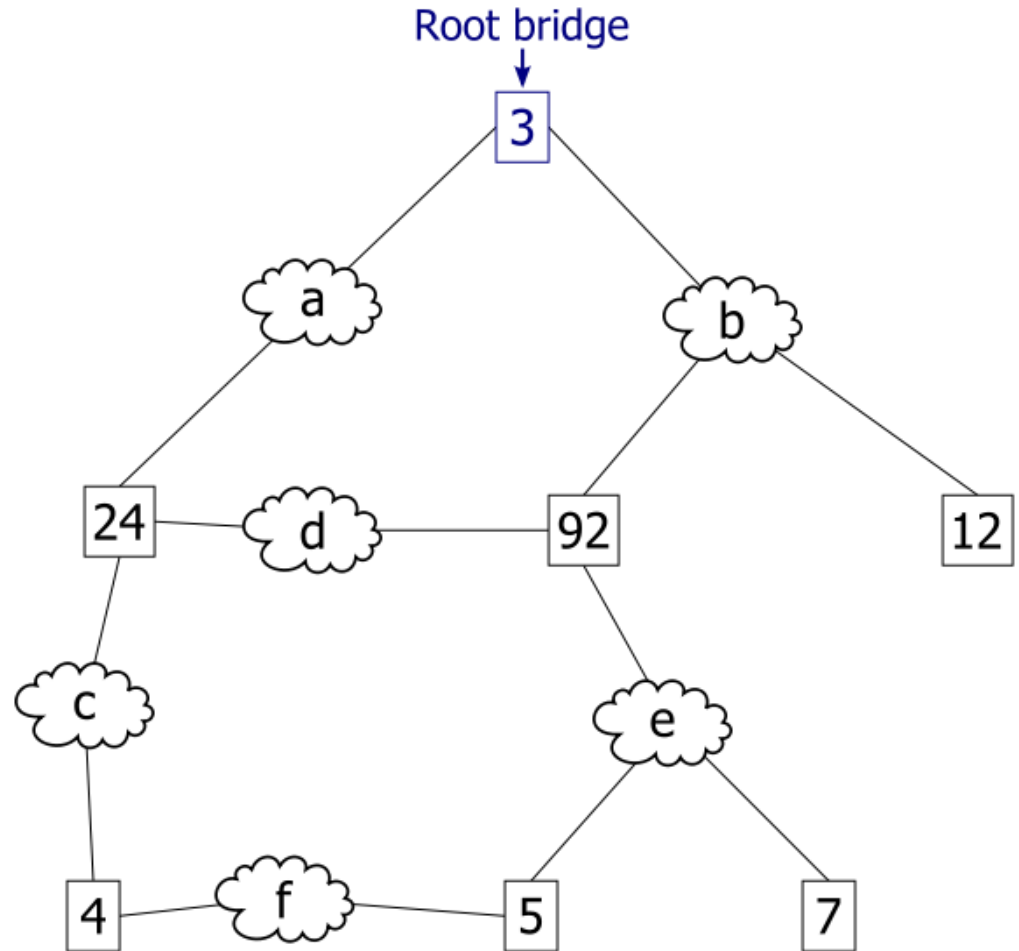
SPANNING TREE PROTOCOL (STP)

- Layer 2 protocol
 - Runs on bridges and switches
 - IEEE 802.1D (original)
 - IEEE 802.1Q-2014 (incl. different new versions)
- Builds a logical loop-free topology for Ethernet networks



SPANNING TREE PROTOCOL (STP)

- Select a root bridge/switch
 - With smallest ID
 - ID: priority (default 32768) + MAC address
 - Good if central, no end station connected (backbone switch)



SPANNING TREE PROTOCOL

- Each bridge has to learn how 'far' it is from the root
 - Cost of the links
 - Root port – the port to the root which is the cheapest
- Each network segment has to learn which bridge is the 'closest' to the root
 - Designated port – the port of the 'closest' bridge to which the network segment is connected



DEFAULT COSTS

Cost of a link – inversely proportional to speed

Data rate	STP cost (802.1D-1998)	RSTP (Rapid STP) cost (802.1W-2004, default value)
4 Mbit/s	250	5,000,000
10 Mbit/s	100	2,000,000
16 Mbit/s	62	1,250,000
100 Mbit/s	19	200,000
1 Gbit/s	4	20,000
2 Gbit/s	3	10,000
10 Gbit/s	2	2,000

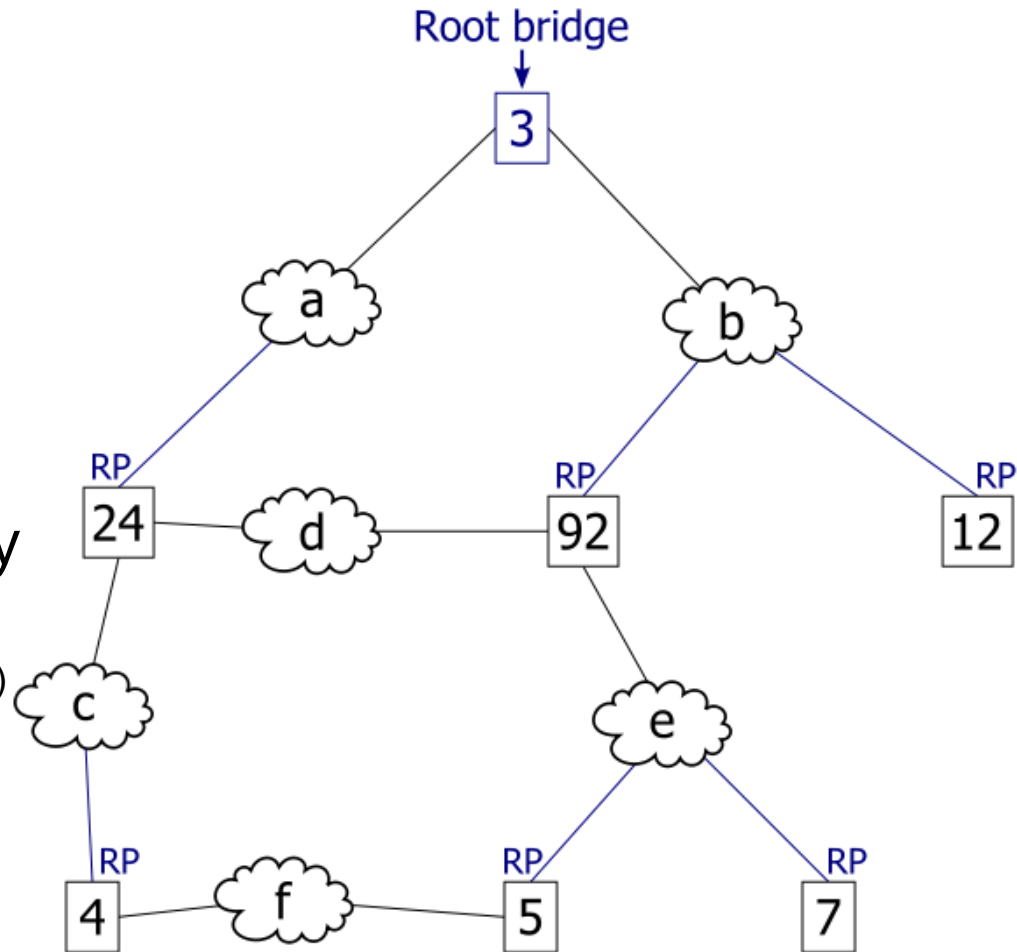
1 Gigabit/second/bandwidth

20 Terabit/second/bandwidth



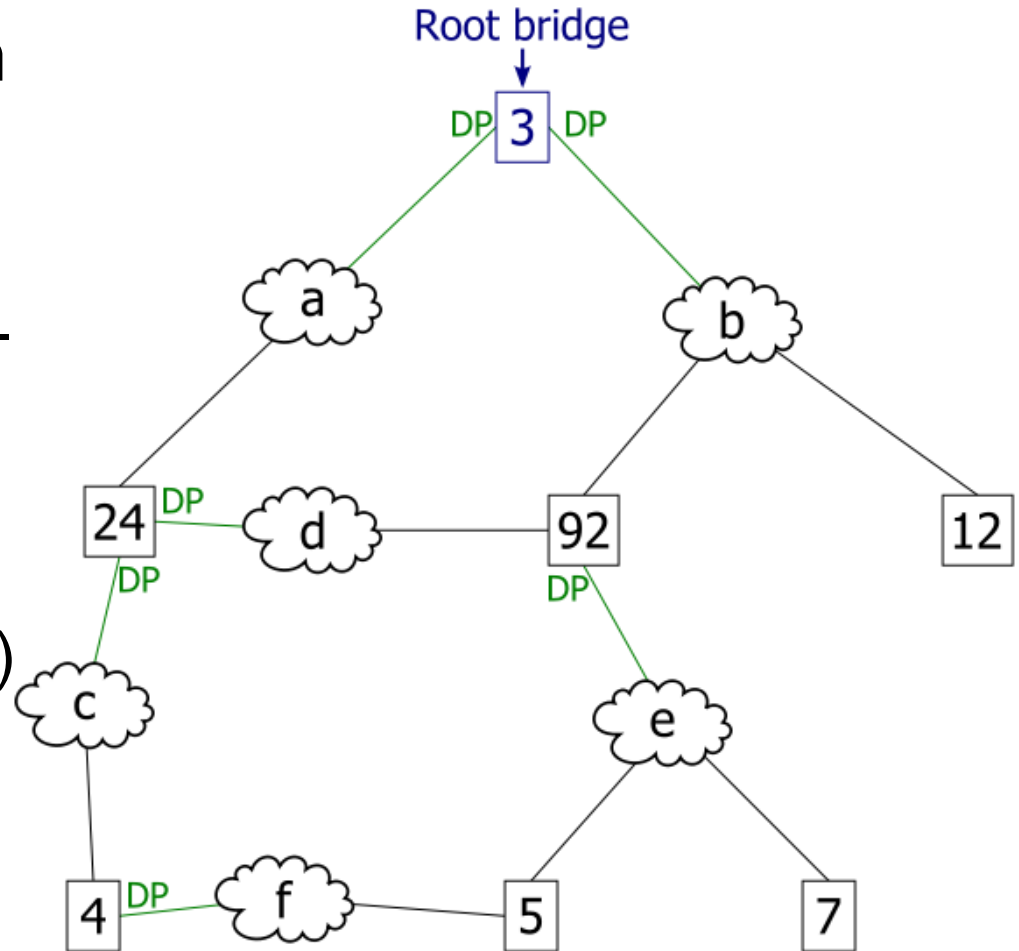
SPANNING TREE PROTOCOL (STP)

- Least cost path from the root for each bridge
 - Bridge protocol data units (BPDUs) transmitted by root with cost=0
 - Others increment cost by the cost of incoming link (all links have the same cost on the figure)
 - Find the smallest
 - Root Port (RP)
 - When tie: from bridge with smaller ID



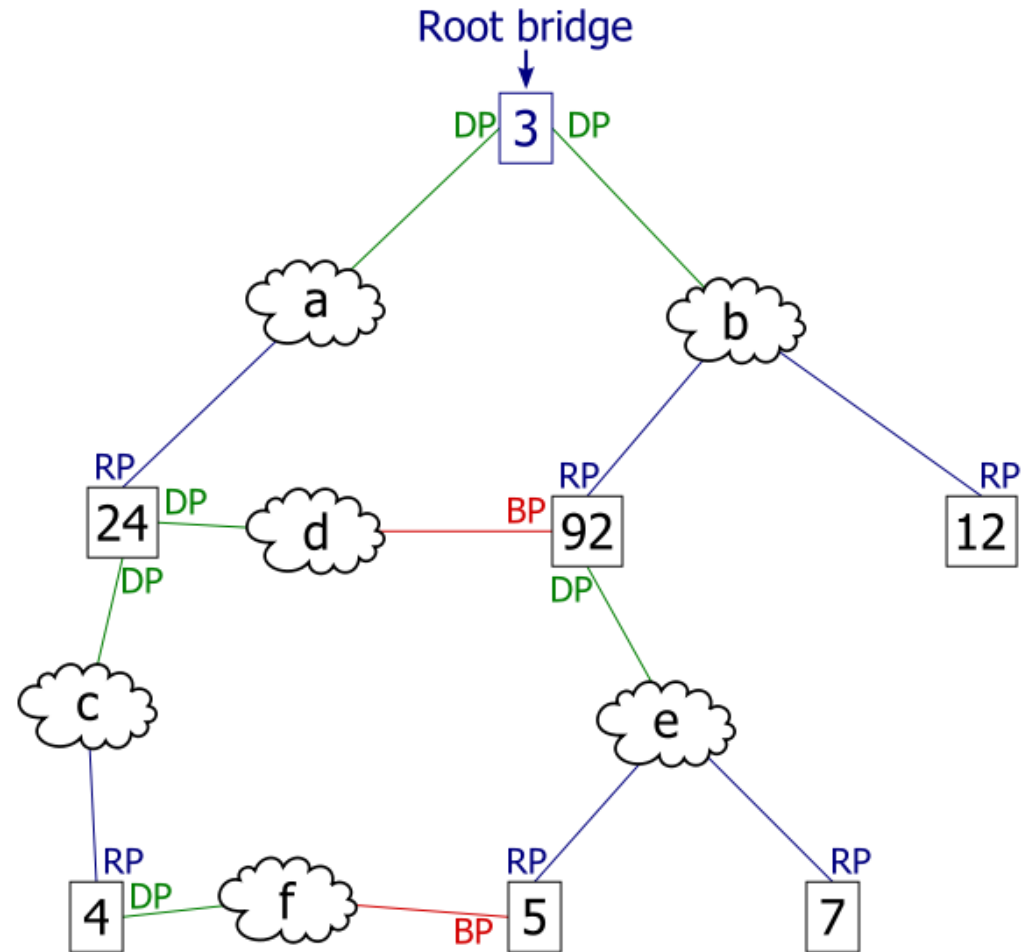
SPANNING TREE PROTOCOL (STP)

- Least cost path from each network segment
- Bridge with the least-cost path from the network segment to the root
- Designated *port* (DP) for the segment
- When tie: from bridge with smaller ID (see d)



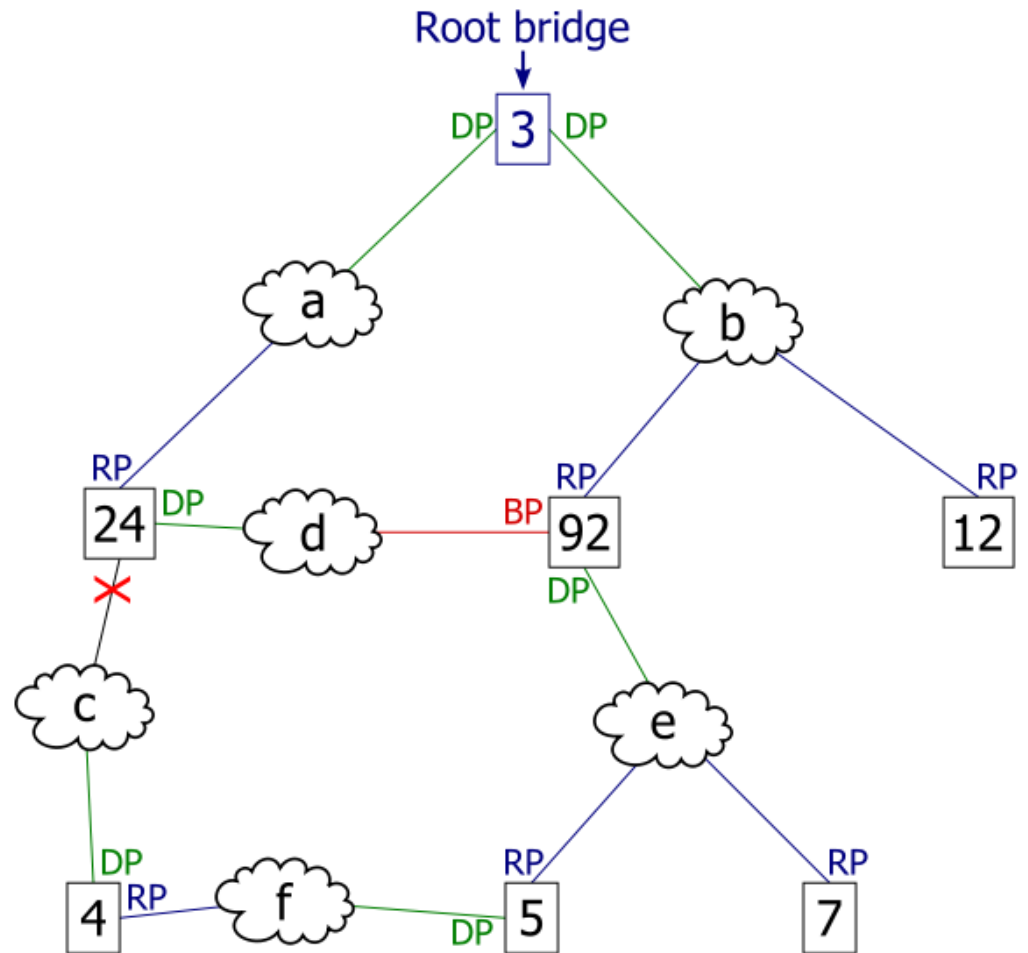
SPANNING TREE PROTOCOL (STP)

- The DPs will be set to forwarding mode
- The rest: blocked (BP)



SPANNING TREE PROTOCOL (STP)

- At link failure:
 - Reconfiguration
 - Takes time



EQUAL-COST MULTI-PATH (ECMP)

- For Layer2 and Layer3
- If n path exist between two nodes
 - Modulo n hash algorithm used to select
 - \sim remainder at division
 - Different hash algorithms
 - From header
 - Source/destination MAC/port
 - Source/destination IP+port and Protocol number
 - To direct same TCP stream onto the same link

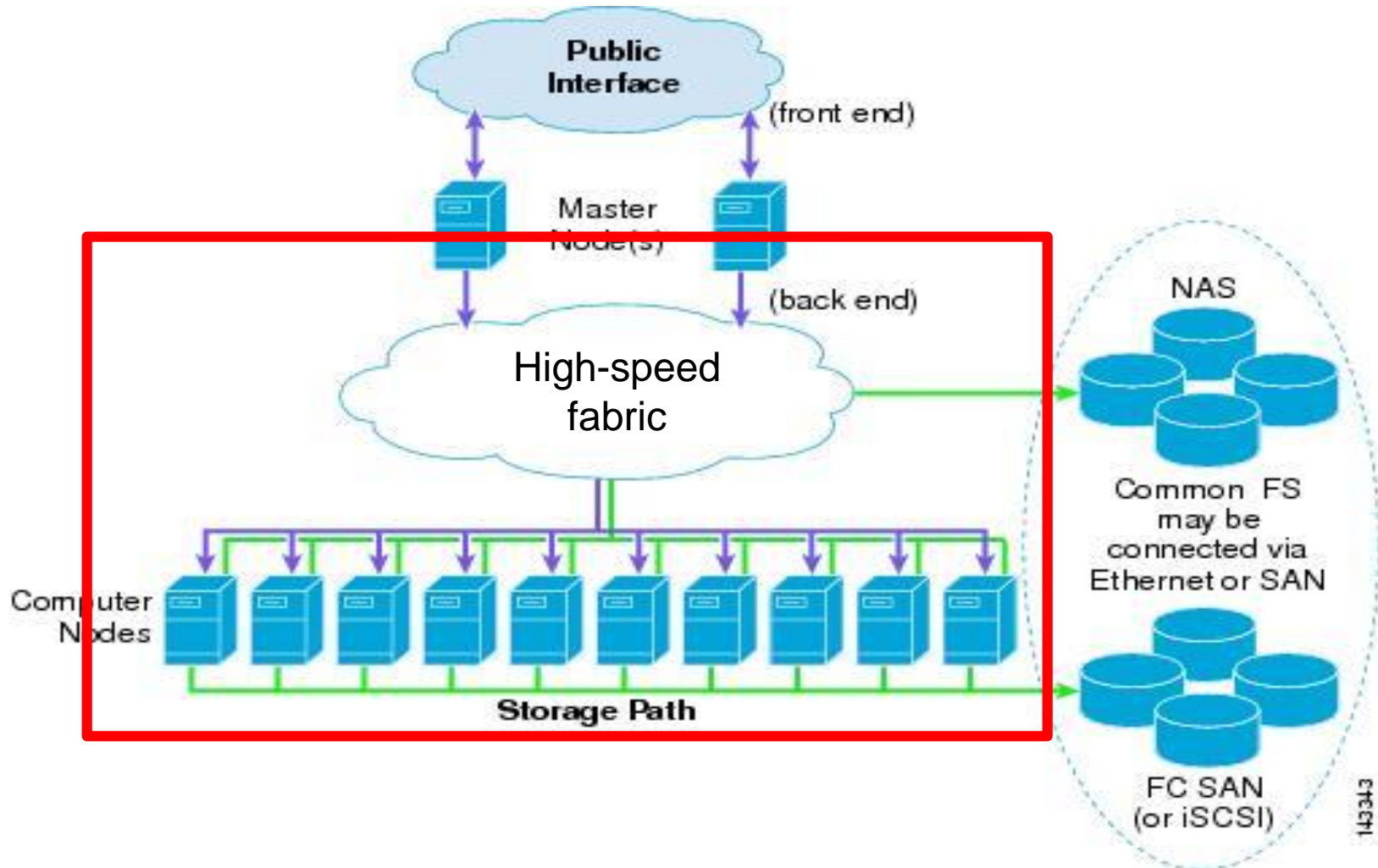


HIGH PERFORMANCE CLUSTERS (HPC)

- Clusters have common goal
 - Combining multiple CPUs
 - To appear as a unified, high performance system
 - High Availability
 - Load Balancing
 - Increased Computing Power
- E.g.:
 - Meteorology (weather simulation)
 - Military research
 - Financial trend analysis
 - Film animation
 - Manufacturing (automotive/aircraft design, aerodynamic simulation)
 - Search engines



LOGICAL VIEW OF A SERVER CLUSTER



HPC TYPES

- Type 1 (tightly coupled)
 - Parallel message passing (Parallel processing)
 - Applications runs on each node in parallel
 - A Master Node determines the input for each node
 - Communication between nodes
 - Weather forecast, simulation (collision, el.magn. field, radio propagation, fluid/air streams)
- Type 2
 - Distributed I/O processing
 - E.g.: search engines
 - Master Node sprays the request
- Type 3 (loosely coupled)
 - Parallel file processing
 - Source file/image divided up and distributed for manipulation in parallel



EVALUATION OF SPINE AND LEAF

- Advantages:
 - Supports east-west traffic
 - Supports virtualization (cloud)
 - High and uniform east-west (inter-cluster) speed
 - Uses all interconnection links
 - Spanning Tree Protocol (STP) -> Equal-Cost Multipath Protocol (ECMP)



EVALUATION OF SPINE AND LEAF

- Advantages:
 - Uniform switches
 - Fix configuration switches
 - Smaller power consumption
 - Inexpensive devices -> high performance, reliable construct
 - If a switch/link fails, only slightly degrades the performance



EVALUATION OF SPINE AND LEAF

- Disadvantages:
 - More switches
 - Number of ports of a switch constrains the max. size
 - More cables
 - When adding a new spine new cables from all leafs
 - Long cables
 - Fiber optic, optical modules (expensive), instead of coaxial cables (cheap but only for short distance)



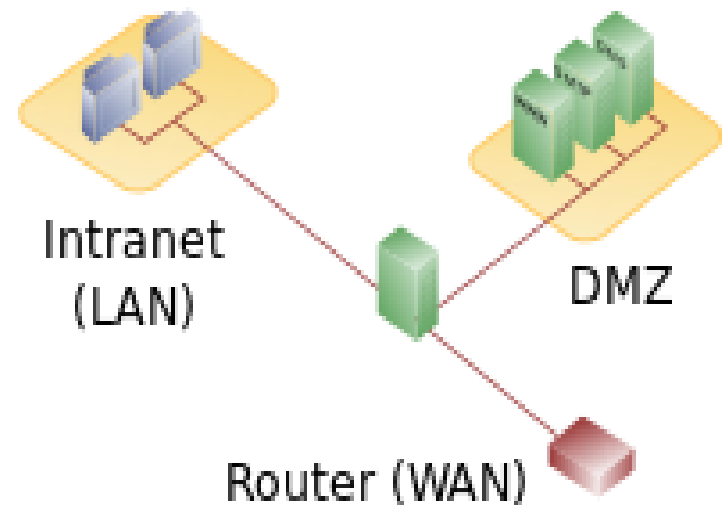
DEMILITARIZED ZONES

- Hosts most vulnerable to attack: provide services to users outside LAN
 - E-mail, web, DNS, FTP, VoIP, ...
 - Place them in a separate subnetwork - Demilitarized Zone (DMZ) or Perimeter Network
 - More secure than Internet, less secure than Intranet
 - Hosts in DMZ have limited connectivity to (some) internal servers and reduced/controlled communication to Internet
 - Firewall(s)
 - Protects against external attacks, but does not deal with internal attacks



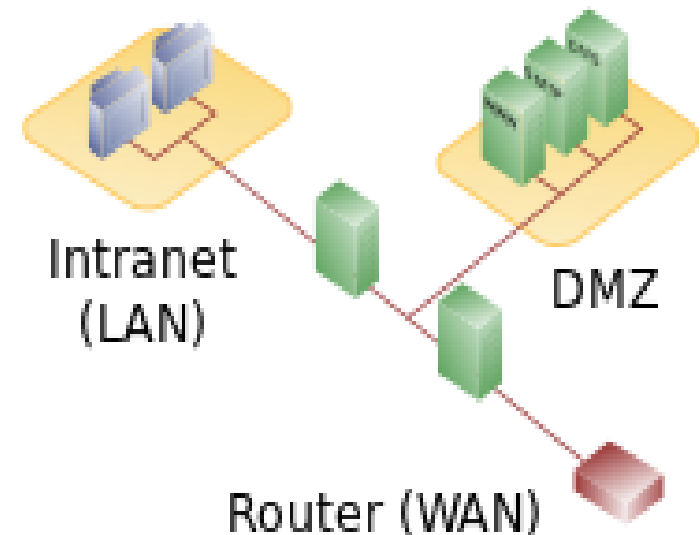
DMZ ARCHITECTURE I.

- Single firewall – three legged model
 - Handles traffic between DMZ-Intranet and DMZ-Internet
 - Single point of failure



DMZ ARCHITECTURE II.

- Dual firewall
 - Front-end/Perimeter
 - Back-end/Internal
- Typically from different vendors
 - Not the same security holes
 - Not the same configuration methods
 - More expensive



PROXY SERVERS

- In DMZ
 - For outgoing traffic
 - Obliges (internal) users to use the proxy to Internet
 - Security
 - Monitoring
 - Centralized web content filtering
 - To reduce Internet traffic – cache in proxy



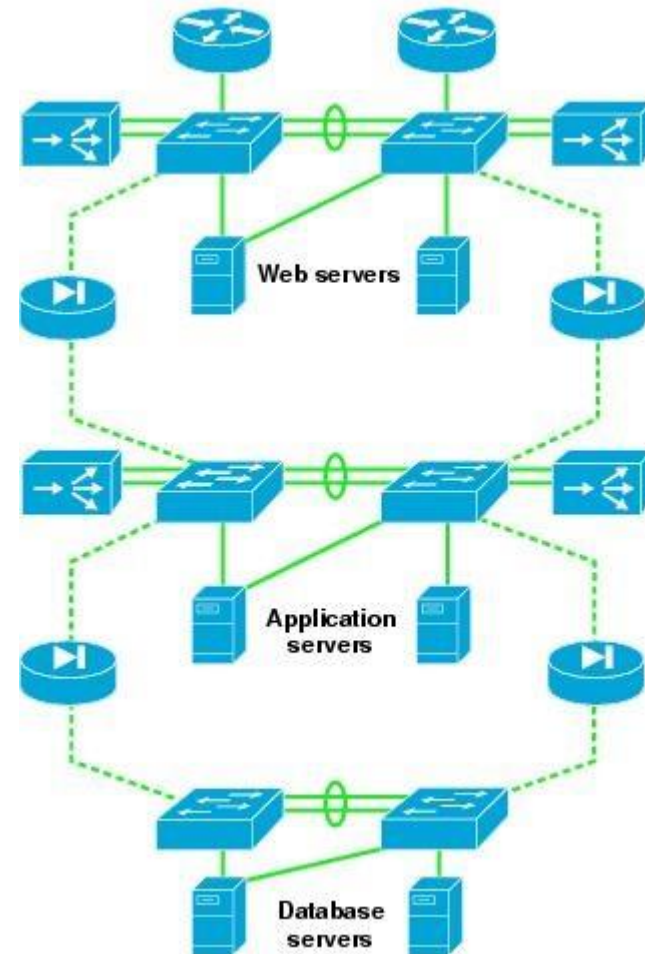
REVERSE PROXY SERVERS

- For incoming traffic
- Provides indirect access for an external network to internal resources
 - To read e-mails from outside the company
- Only the reverse proxy server can have an access to mail server
- Extra layer of security
- Typically by an application layer firewall



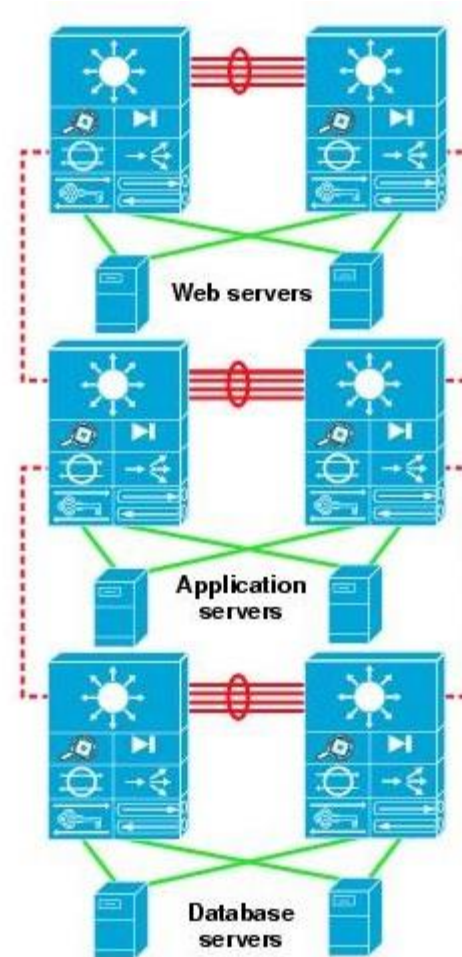
HTTP-BASED APPLICATIONS I.

- 3 tiers
 - Web servers
 - Application servers
 - Database servers
- Separated by firewalls



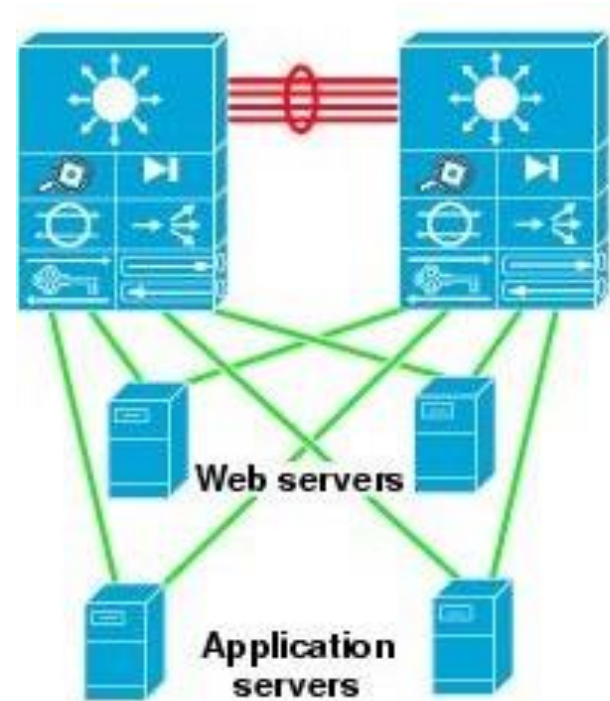
HTTP-BASED APPLICATIONS II.

- Integrated Service Modules
 - Router/switch
 - Firewall
 - Load balancing
 - Security
 - IDS (Intrusion Detection System)

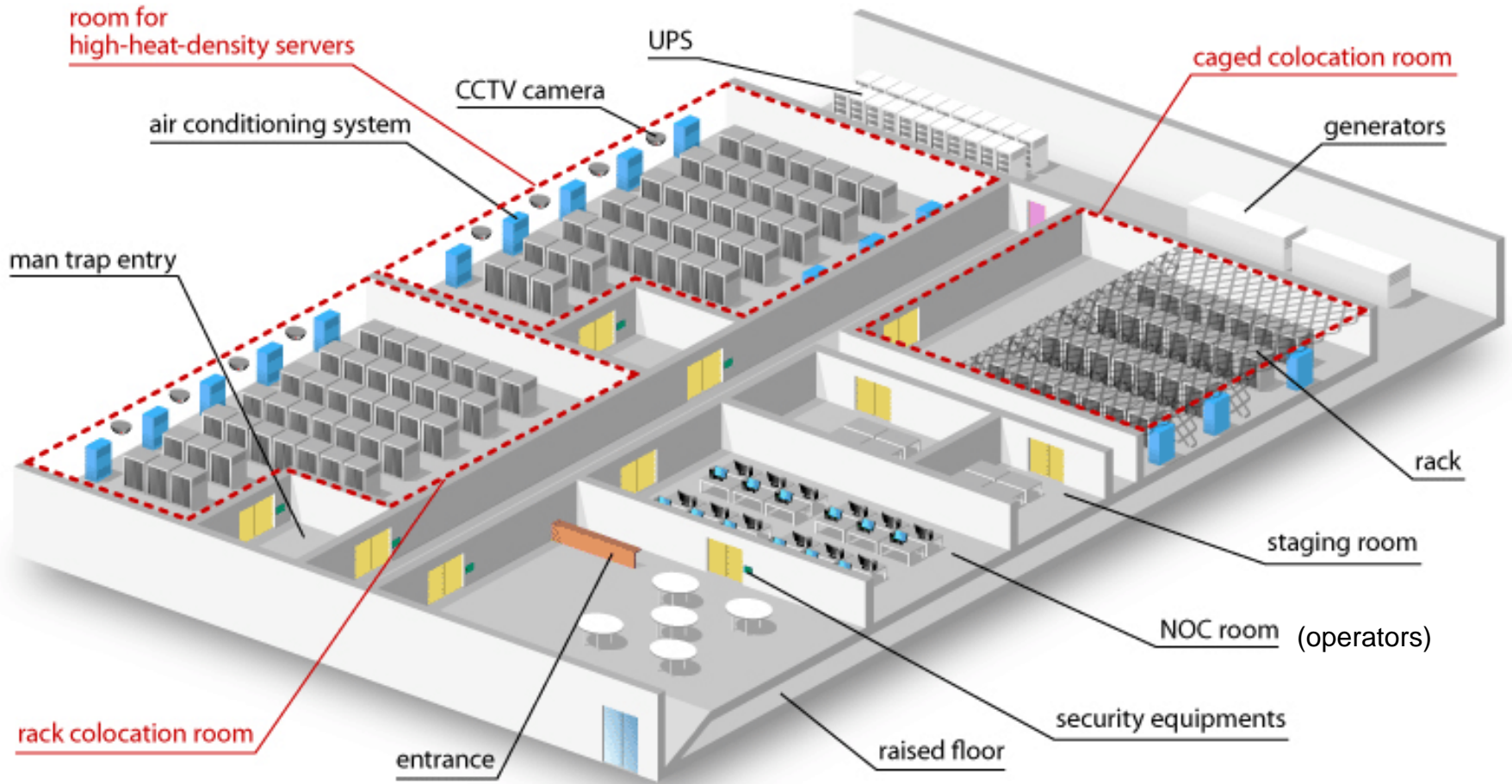


HTTP-BASED APPLICATIONS III.

- Logical segregation
 - With VLANs
 - Reduced performance
 - But fewer devices
- + Database servers
 - typically physically separated



DATA CENTER OVERLOOK



COLOCATION ROOM (COLO)

- Where the customers can place their equipment
 - “carrier hotel”
 - Closed rooms
 - Cages
 - Racks (lockable)
 - Device owned by the customer
 - Data Center provides the operation & maintenance
 - Space, power, cooling, physical security
 - Connections to telecommunications and network service providers



CAGED COLOCATION ROOM



RACK COLOCATION ROOM



STAGING ROOM

- Equipment
 - Packing / unpacking
 - Verification tests
 - Configuration
- Outside the DC



DATA CENTER SERVICES

- Colocation / Server rental
- 24/7 supervision
- High reliable environment
 - georedundancy
- Operation
- Design and implementation of movement
- Data back-up and restoration
- Activation



DATA CENTER SERVICES

- Redundant power supply from two different systems
- Redundant transformers
- Redundant diesel generator
 - Automatic switching
- DC, n+1 redundant
- HVAC, n+1 redundant
- Raised floor, cable tables
- Redundant fibre optic ring

