

INFORMÁCIÓS RENDSZEREK ÜZEMELTETÉSE

BME VIK TMIT

MÉRNÖK-INFORMATIKUS ALAPKÉPZÉS



BME VIK TMIT

7. ADATKÖZPONTOK (DATA CENTERS)

Adatközpontok definíciója

Adatközponti topológiák:

Multi-tier, Fat tree, Full mesh, Spine and leaf

Útválasztó protokollok:

Spanning tree, Equal Cost Multipath

High Performance Clusters

Demilitarizált zónák

HTML alapú alkalmazások

Adatközpontok felépítése

Adatközponti szolgáltatások

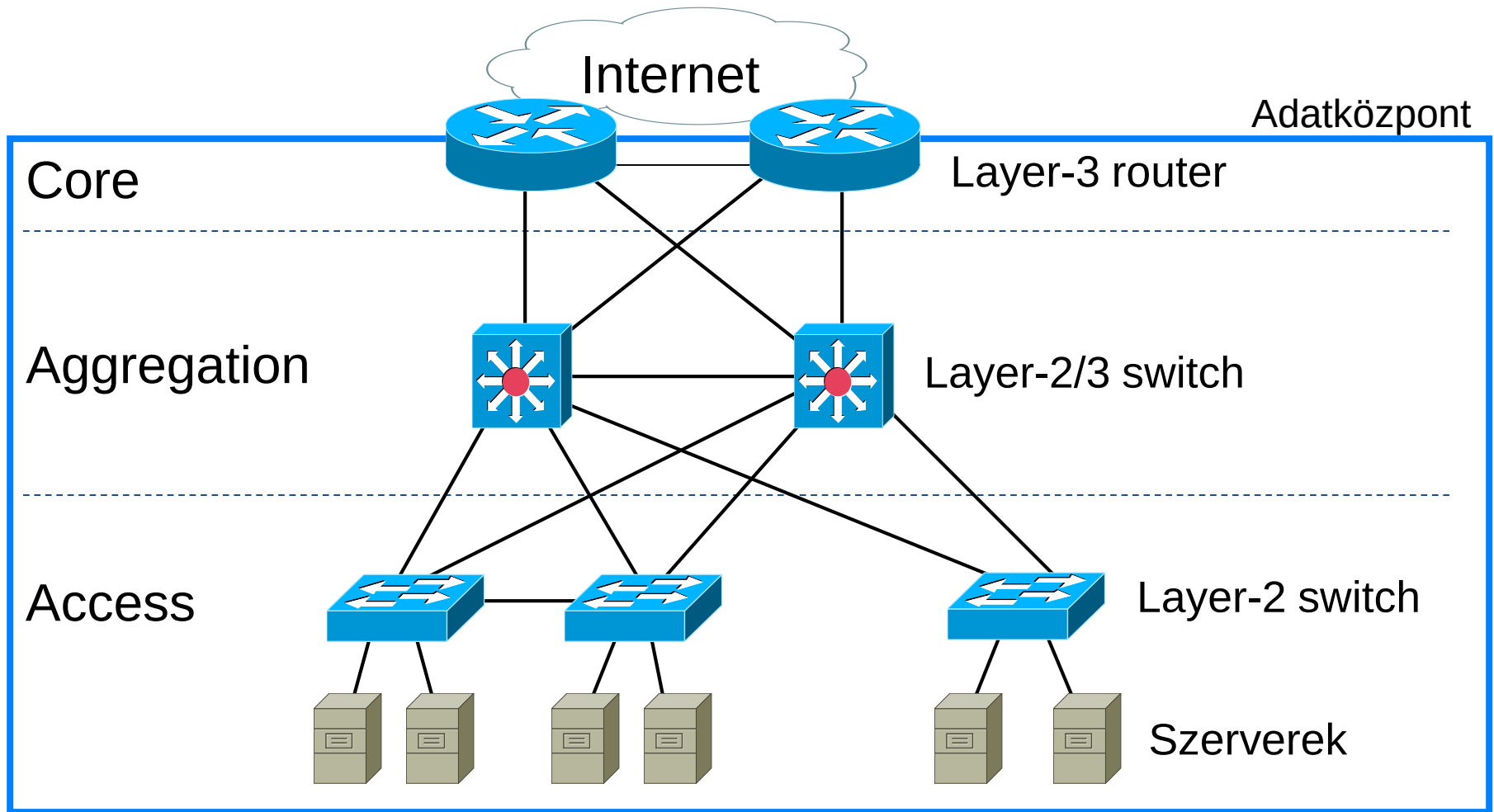


ADATKÖZPONTOK

- Az adatközpont számítógép rendszerek és komponenseik elhelyezésére szolgáló létesítmény
- Napjainkban a PC-k tízezreit tartalmazó egyetemi, kutatóközponti, vállalati adatközpontok megszokottak
- Egy adatközpontot tipikusan tudományos számításokra, üzleti analitikára, adatbányászatra, adattárolásra, és nagyteljesítményű hálózati szolgáltatások nyújtására használnak



TIPIKUS ADATKÖZPONTI TOPOLÓGIA

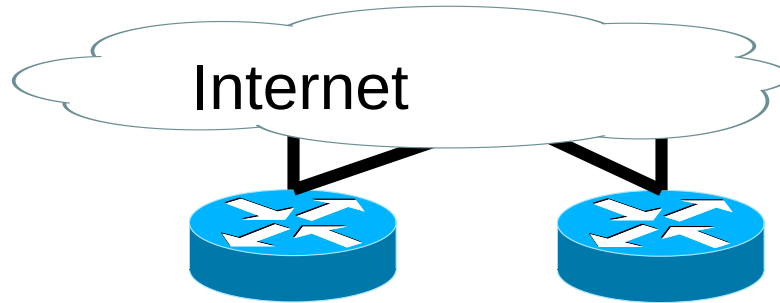


MULTI-TIER MODEL

- Multi-tier: többszintű
- Access – hozzáférési
 - Szerverek csatlakoztatása
- Aggregation – aggregációs
 - A hozzáférési réteg redundáns kapcsolatai
- Core – mag
 - Routing szolgáltatások nyújtása
 - “Összeköti az adatközpontot a külvilággal”
 - Az adatközpont más részeivel
 - Adatközponton kívüli szolgáltatásokkal (pl. Internet)
 - Földrajzilag elkülönülő adatközpontokkal
 - Más távoli helyszínekkel



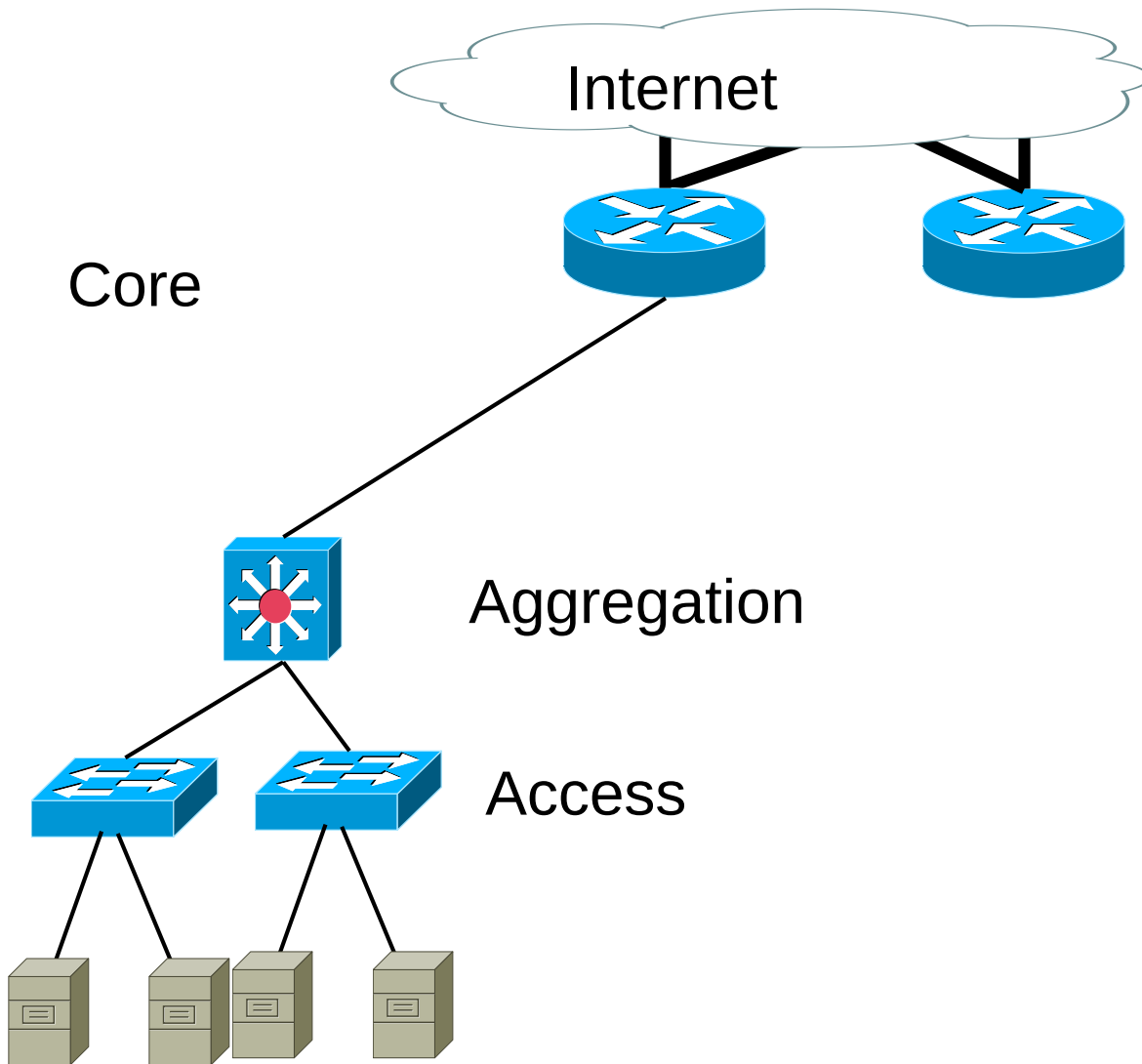
MULTI-TIER MODEL



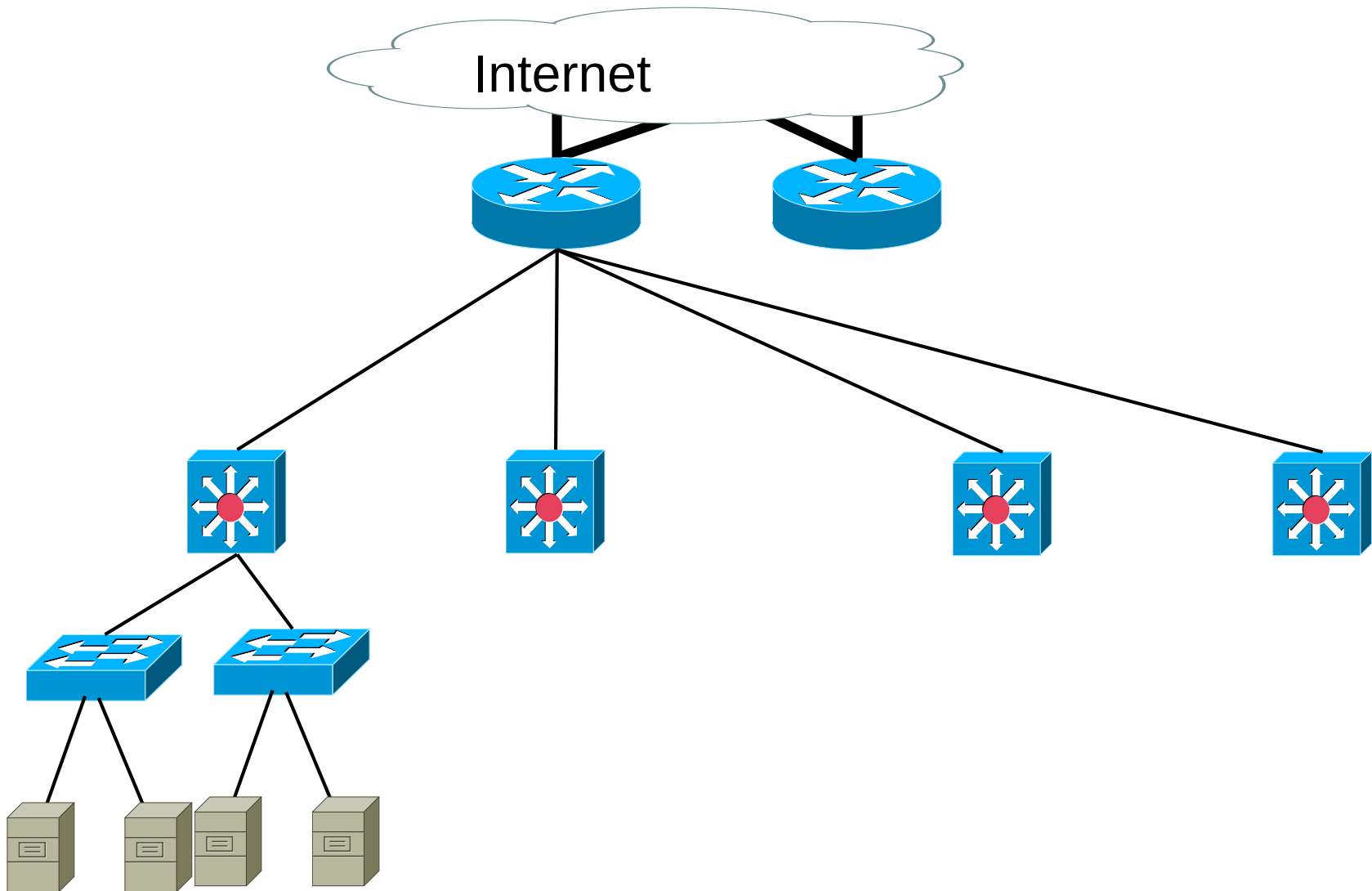
Core



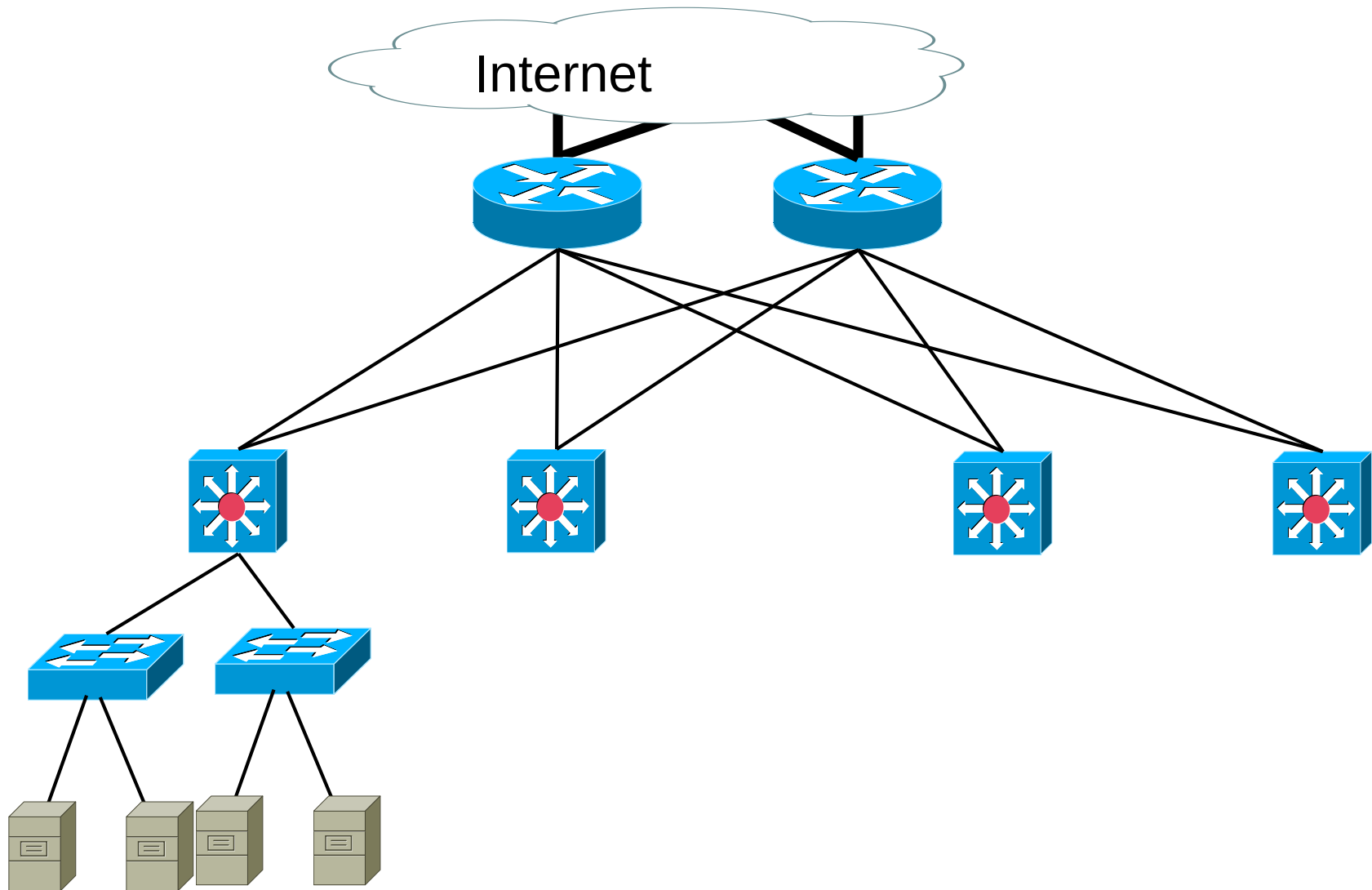
MULTI-TIER MODEL



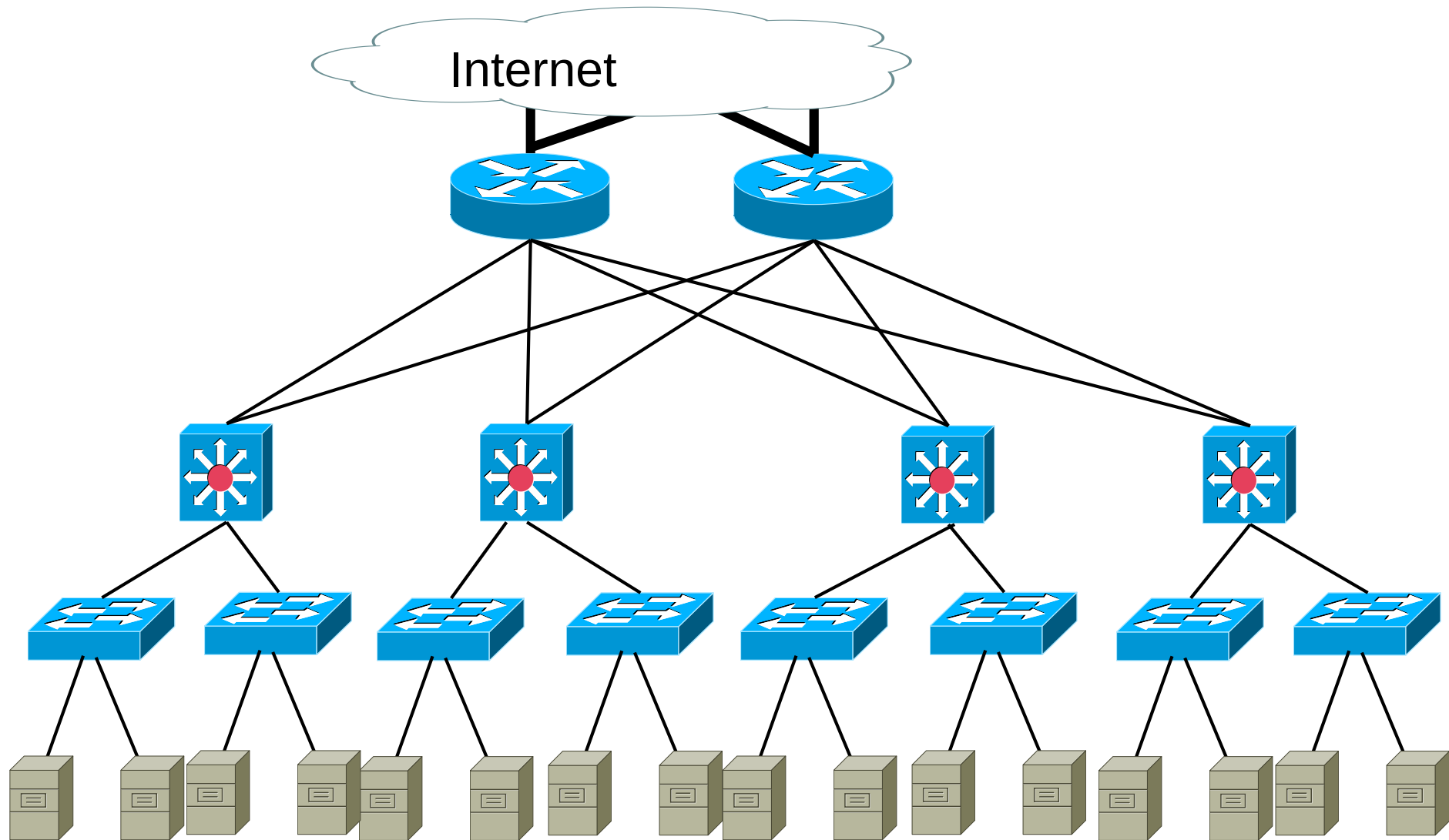
MULTI-TIER MODEL



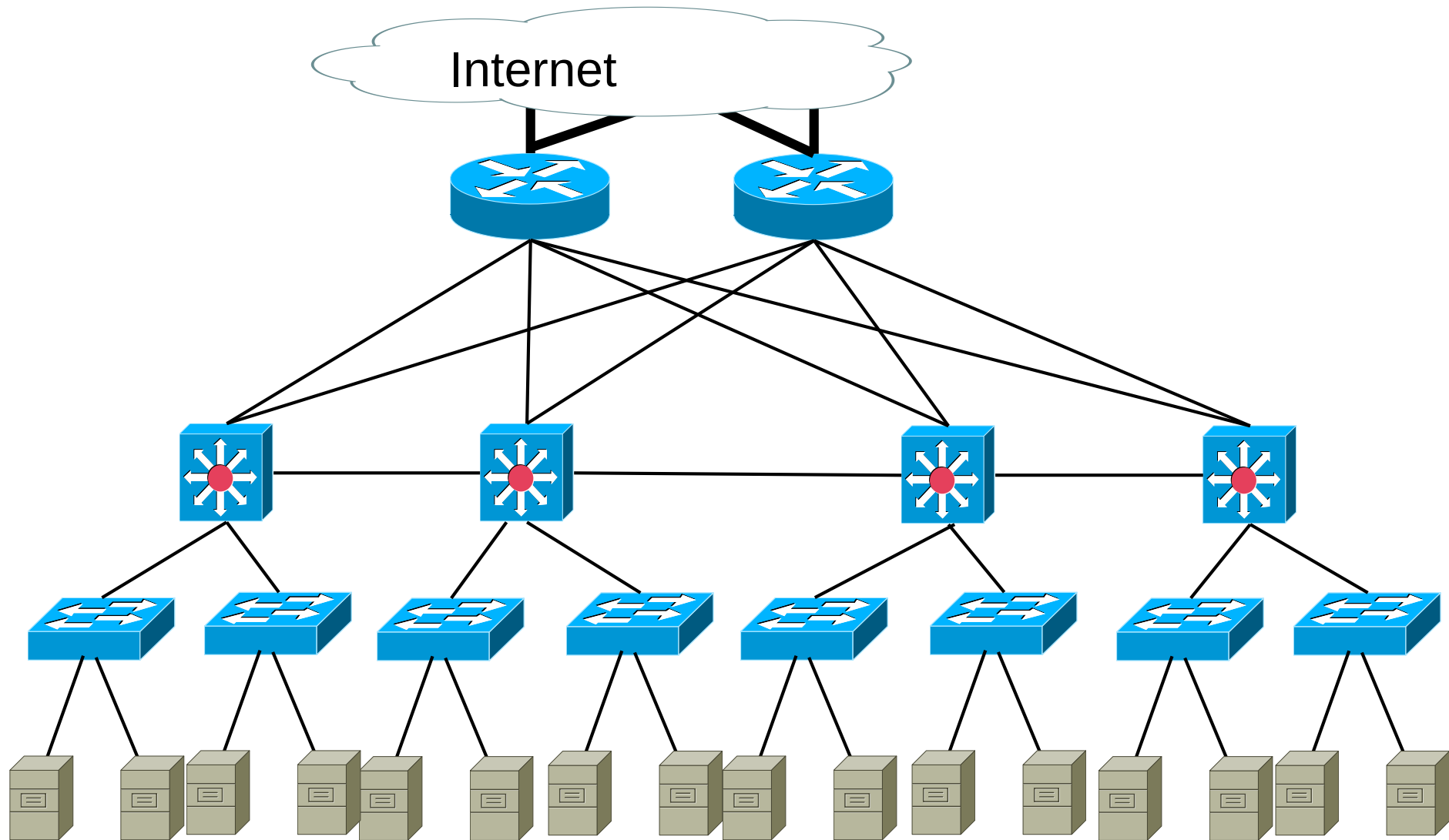
MULTI-TIER MODEL



MULTI-TIER MODEL

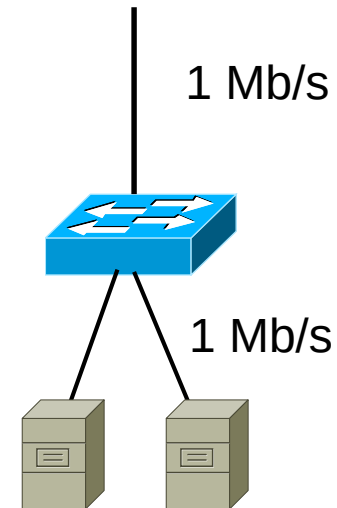


MULTI-TIER MODEL



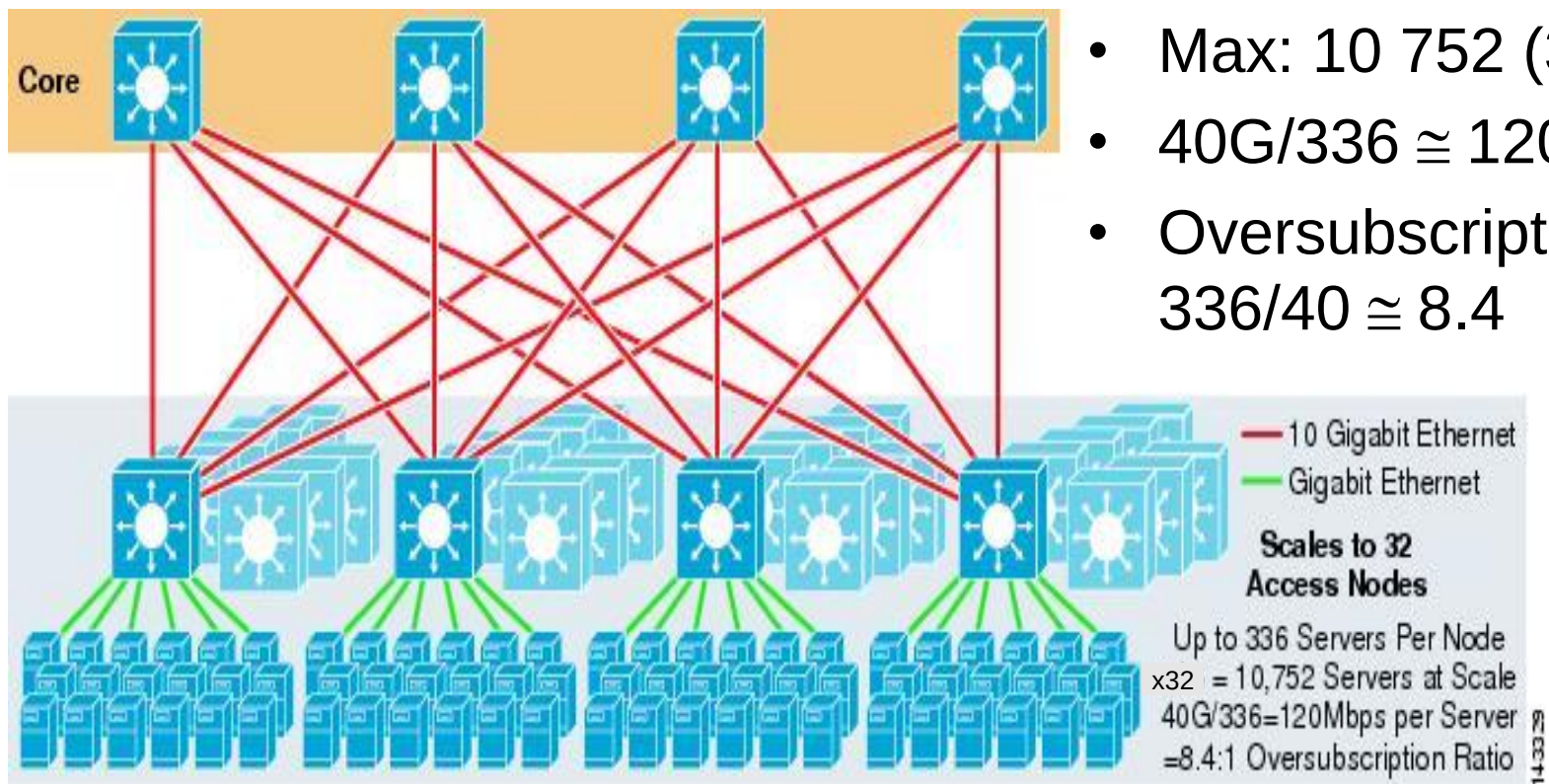
TÚLMÉRETEZÉS (OVERSUBSCRIPTION)

- Kimenő összes sávszélesség / bejövő sávszélesség (le/fel)
- $2 \times 1 \text{ Mb/s} / 1 \text{ Mb/s} = 2$
- Nem valószínű, hogy egyszerre kell az összes sávszélesség



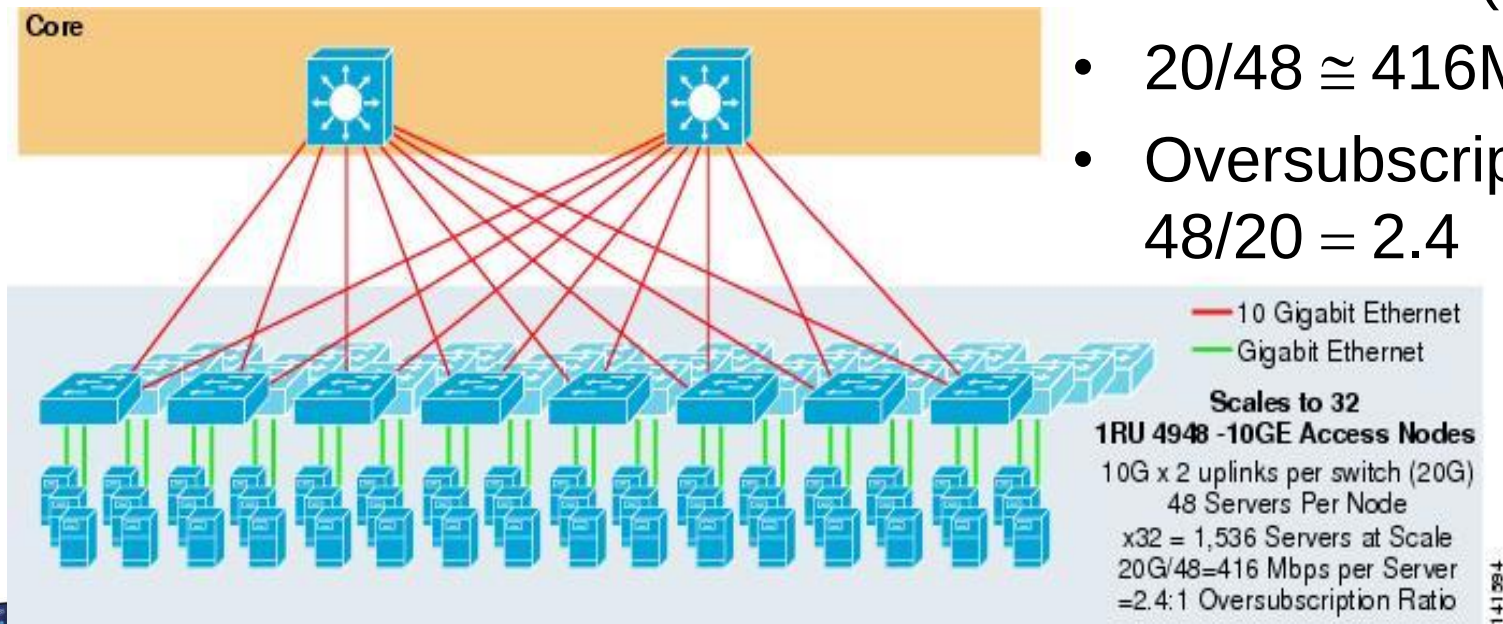
2-TIER MODEL

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



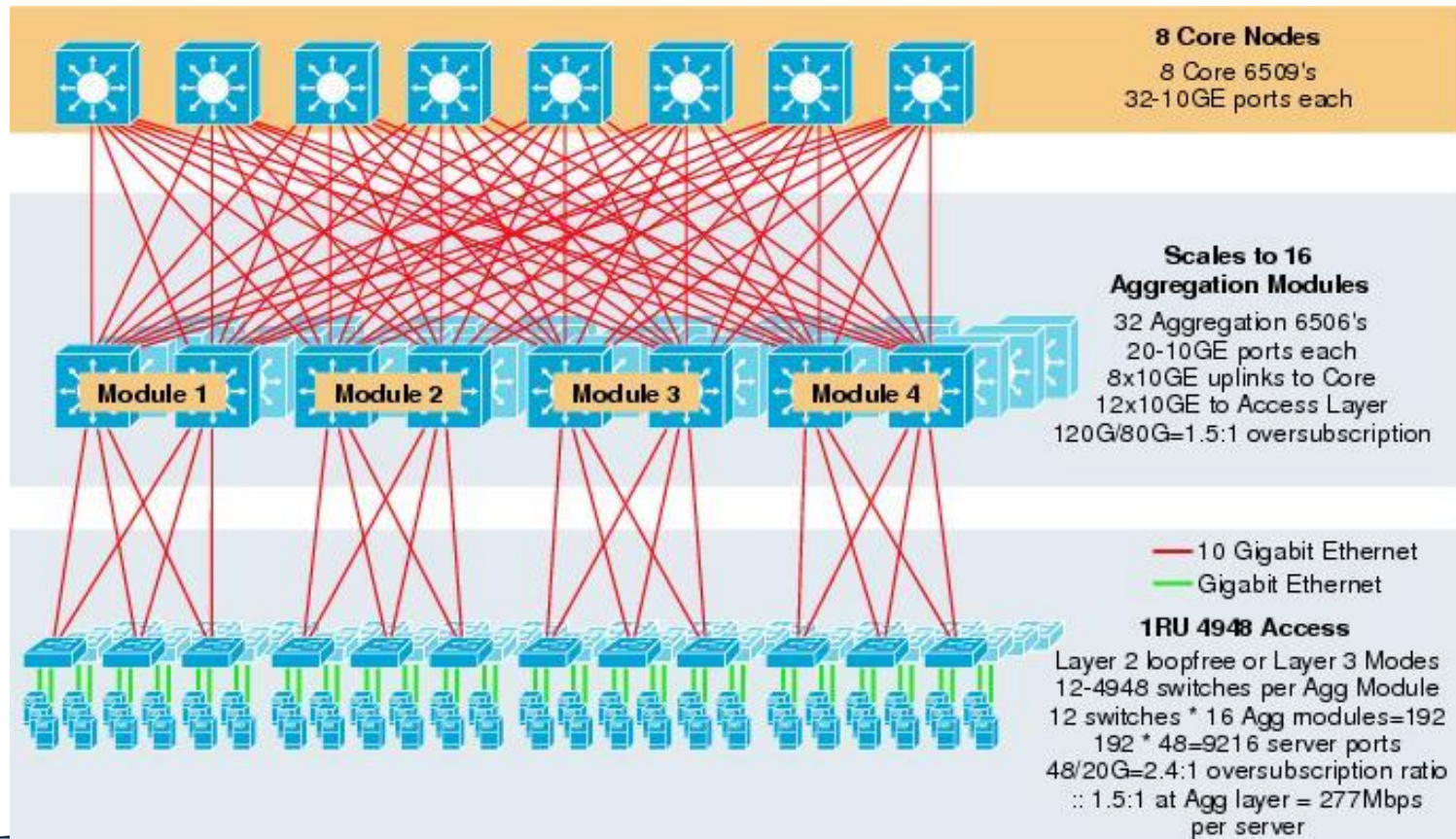
2-TIER ToR (1RU) SWITCH-EKKEL

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



3-TIER MODEL

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

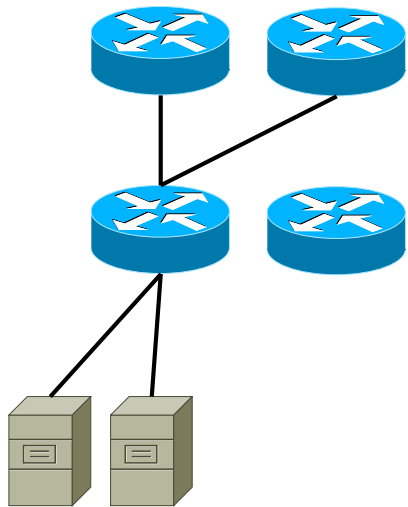


MULTI-TIER -> FAT TREE

- Nagy, speciális routerek / switchek
- -> Kisebb, köznapi
- -> Fat Tree Topology (vastagfa topológia)
- A következő példában switch $4 * 1\text{Gbit/s}$ porttal
- Fele a portoknak lefelé, a másik fele felfelé



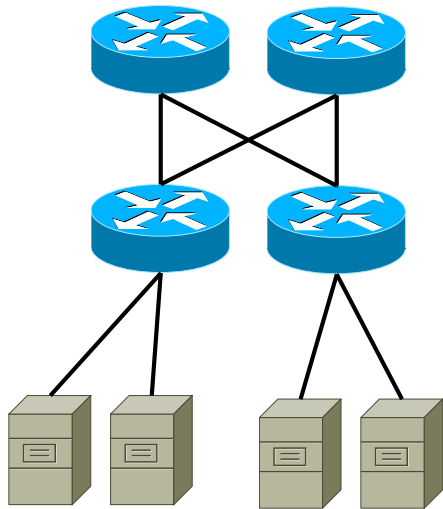
FAT TREE TOPOLOGY



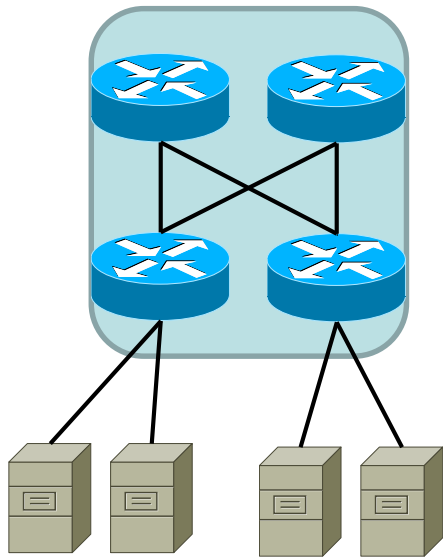
Fele a portoknak lefelé, a másik fele felfelé



FAT TREE TOPOLOGY



FAT TREE TOPOLOGY



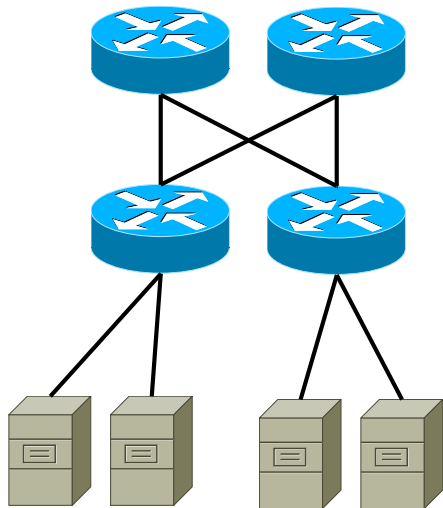
Pod



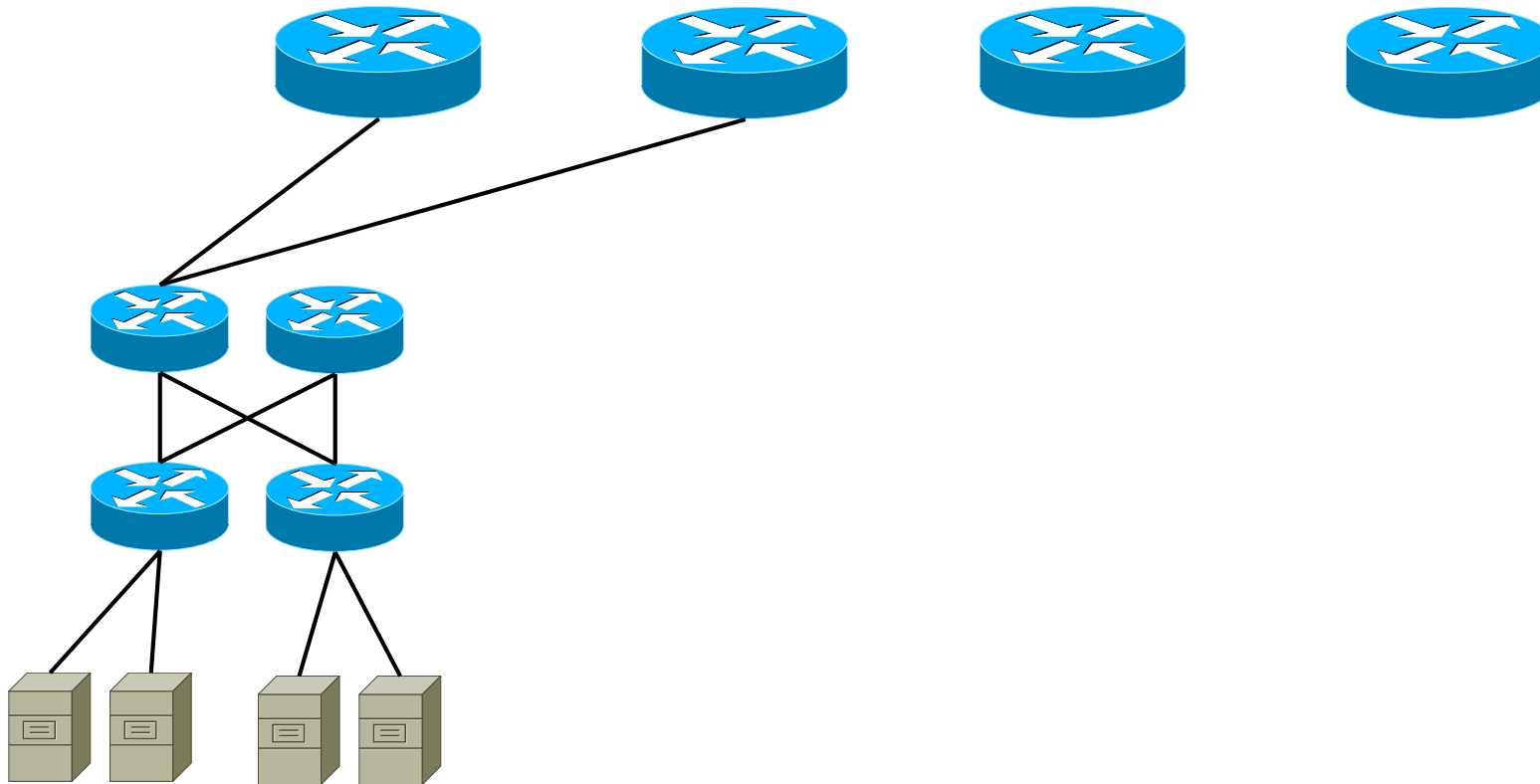
FAT TREE TOPOLOGY



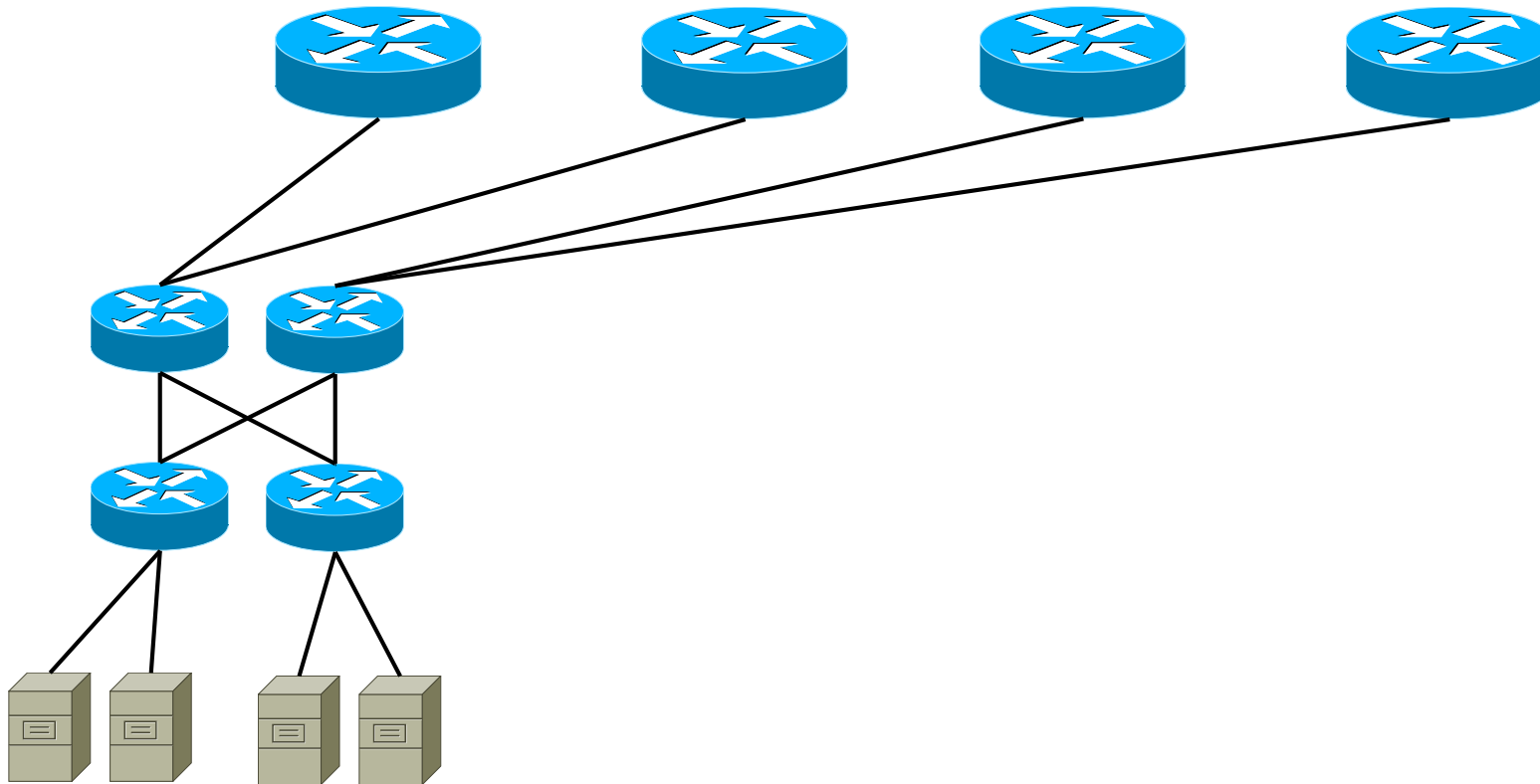
Fele a portoknak lefelé, a másik fele felfelé



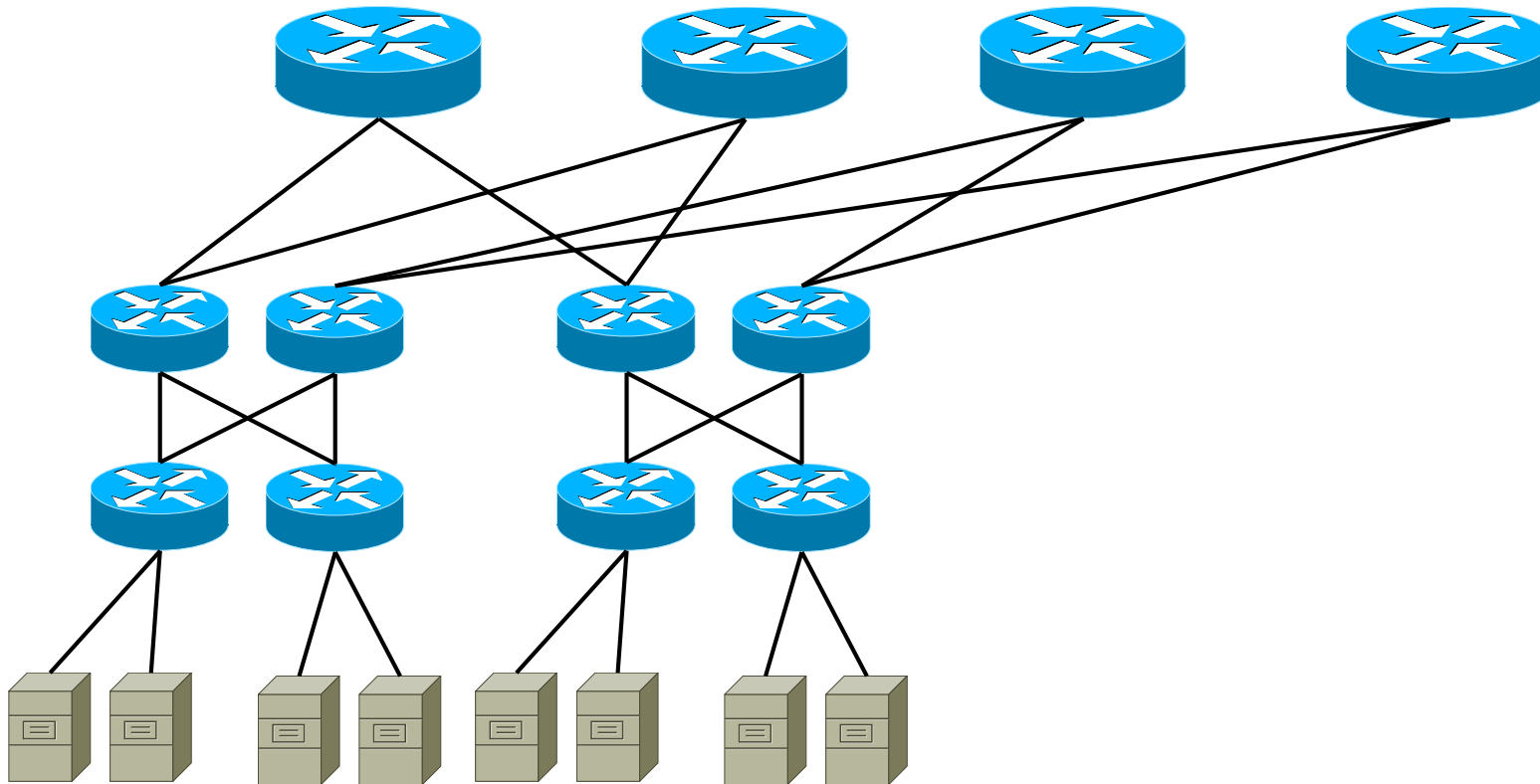
FAT TREE TOPOLOGY



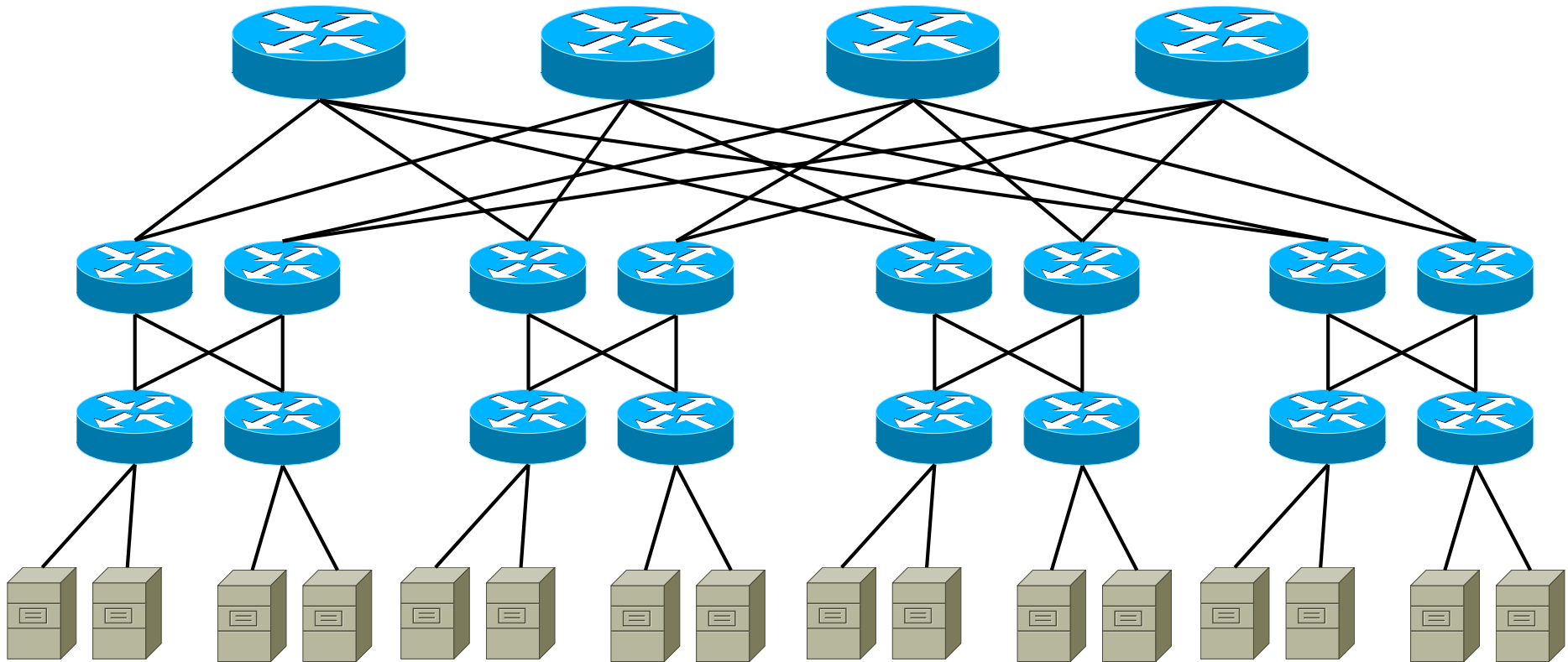
FAT TREE TOPOLOGY



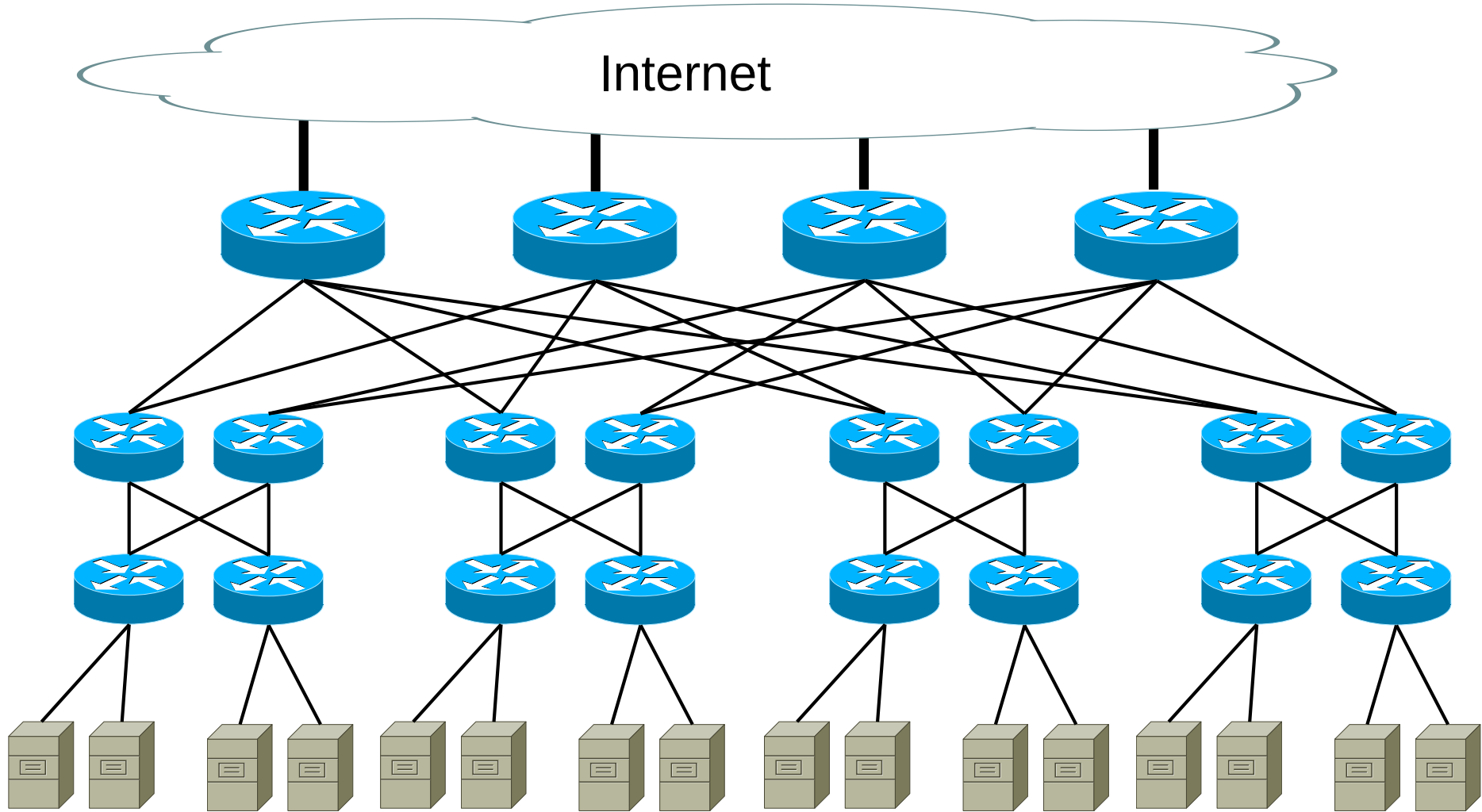
FAT TREE TOPOLOGY



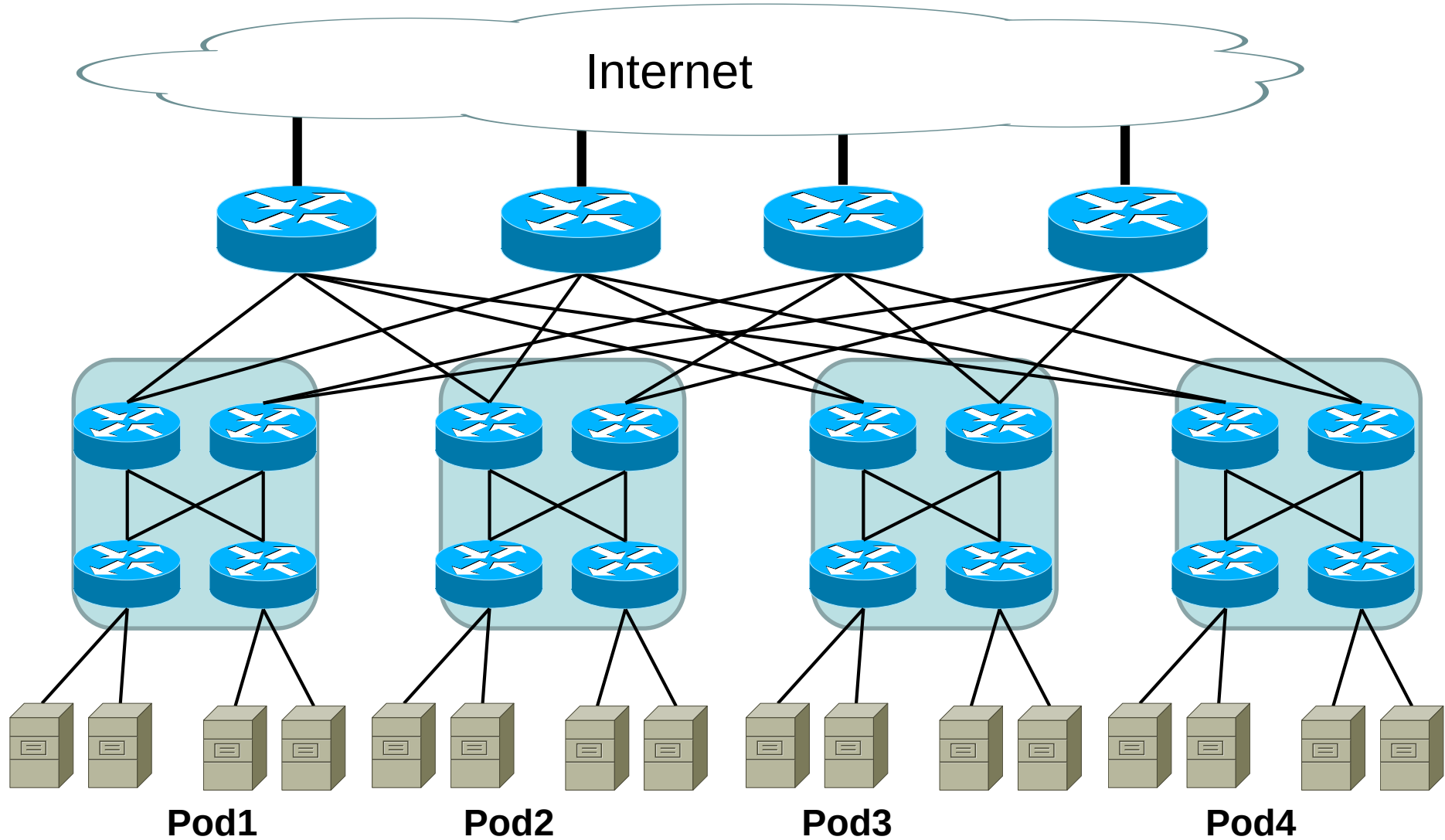
FAT TREE TOPOLOGY



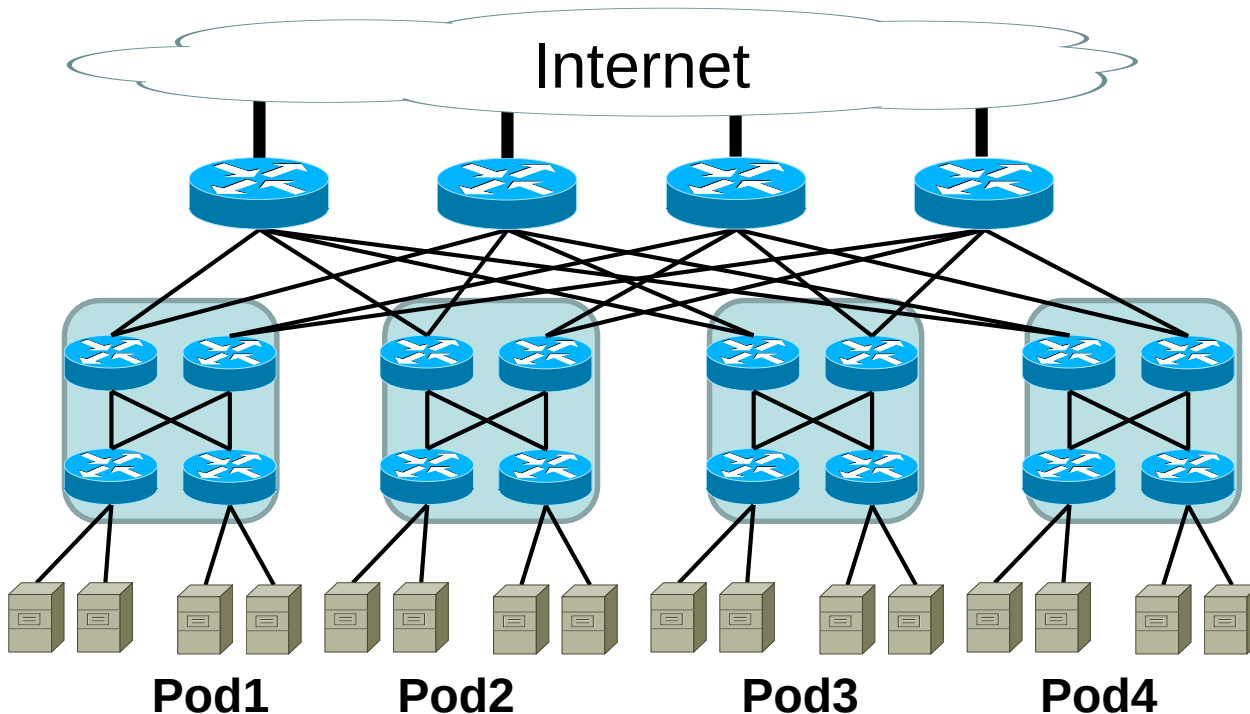
FAT TREE TOPOLOGY



FAT TREE TOPOLOGY



FAT TREE TOPOLOGY



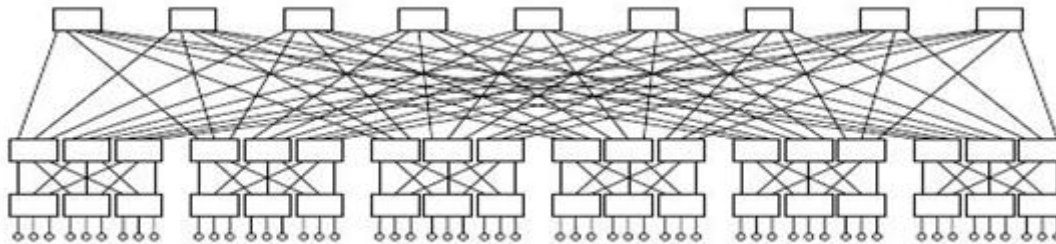
- 1 switch: k port
 - k pod
- 1 pod tartalma:
 - $k/2$ aggregate switch
 - $k/2$ access switch
 - $(k/2) * (k/2) = (k/2)^2$ szerver
- Össz.: $k * (k/2)^2 = k^3/4$ szerver
- $(k/2) * (k/2) = (k/2)^2$ core switch



FAT TREE TOPOLOGY

k-adrendű (k-ary) fat tree:

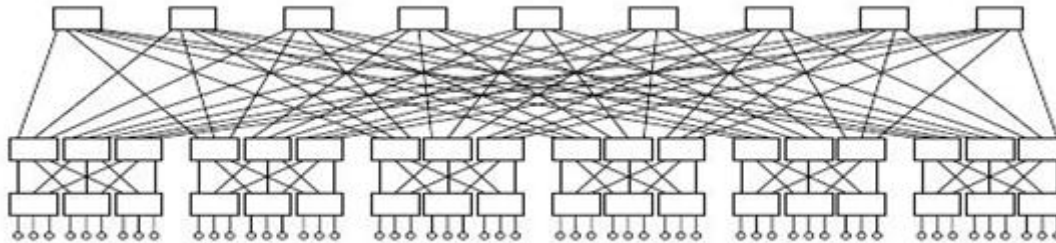
- 3-rétegű topológia (access, aggregation, core)
- minden switch-nek k portja van (-> k pod)
 - fele felfelé, fele lefelé
- minden access switch $k/2$ serverhez és $k/2$ aggregation switch-hez kapcsolódik
- minden aggregation switch $k/2$ access és $k/2$ core switch-hez kapcsolódik
- $(k/2)^2$ core switch: mindegyik k pod-hoz kapcsolódik
- minden pod $(k/2)^2$ szervert és 2 réteg $k/2$ db k-portú switch-et tartalmaz
- összesen $k * (k/2)^2 = k^3/4$ szerver



6-ODRENDŰ FAT TREE TOPOLOGY

k-adrendű (k-ary) fat tree:

- minden switch-nek 6 portja van -> 6 pod
 - fele felfelé, fele lefelé
- minden access switch 6/2 szerverhez és 6/2 aggregation switch-hez kapcsolódik
- minden aggregation switch 6/2 access és 6/2 core switch-hez kapcsolódik
- $(6/2)^2 = 9$ core switch: mindegyik 6 pod-hoz kapcsolódik
- minden pod $(6/2)^2 = 9$ szerveret és 2 réteg 6/2 db 6-portú switch-et tartalmaz
- összesen $6 * (6/2)^2 = 6^3/4 = 54$ szerver



64-EDRENDŰ FAT TREE TOPOLOGY

- Szerverek száma?
- $k * (k/2)^2 = k^3/4$ szerver = $64 * 32^2 = 65536$ szerver



FAT TREE TOPOLOGY ÉRTÉKELÉSE

- Sáv szélesség (Bandwidth)
 - Mindegyik rétegnek ugyanakkora az aggregált (összesített) sáv szélessége
 - Oversubscription = 1
- Olcsó, egyforma kapacitású eszközökből építhető
 - Minden port ugyanazt a sebességet támogatja (end host sebesség azonos)
 - Minden eszköz a teljes vonali sebességgel adhat, ha a csomagokat egyenletesen szétosztjuk az összes elérhető útvonalon
- Jól skálázható: k -portú switch $k^3/4$ szerveret támogat
- Kisebb áramfogyasztás / hő / hűtés



A HAGYOMÁNYOS ARCHITEKTÚRÁK PROBLÉMÁI

- Változás a forgalmi modellben (traffic pattern)
 - Tradícionálisan: észak-dél
 - Most: megjelenik a kelet-nyugat is
 - Miért?
 - Virtualizáció / Cloud
 - Bármely (virtuális) szerver bármelyik fizikain futhat
 - Terhelés újra allokálása
 - Kommunikáció kell közöttük: UGYANAZ a sebesség TETSZŐLEGES clusterek között
 - Nagy megbízhatóság

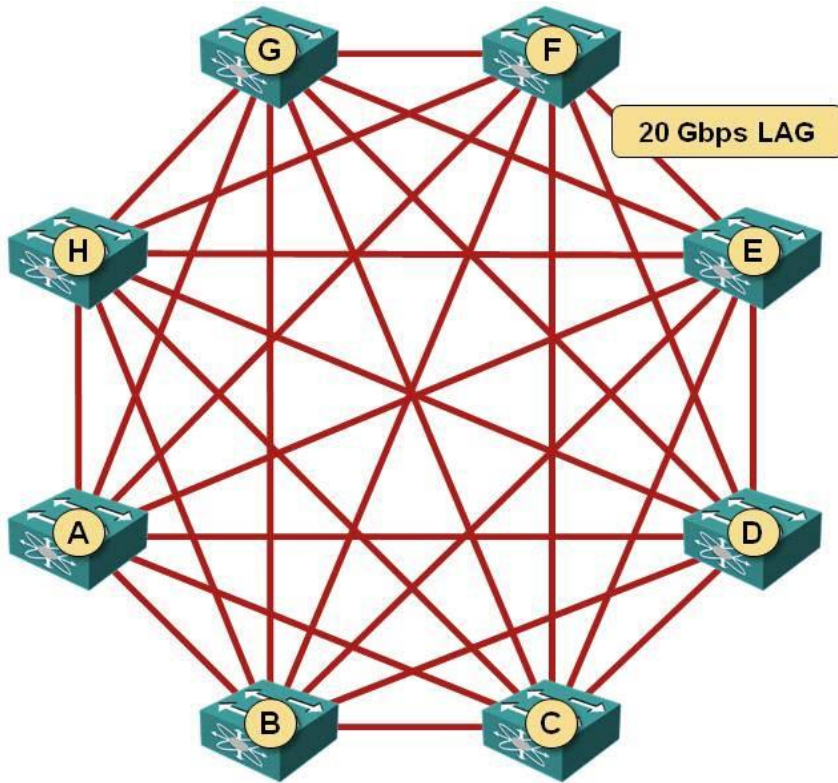


FULL MESH

- Tipikus adatközponti switch:
 - 48 x 10GE port és
 - 4 x 40GE port, ami 16 x 10GE port-ként is használható (160 GE)
 - Összesen 64*10GE port (640 GE)
- Használjunk:
 - 48 port-ot szerverek felé
 - 16 (14) port-ot intra-fabric kapcsolatra
- Oversubscription = ?
 - 3:1



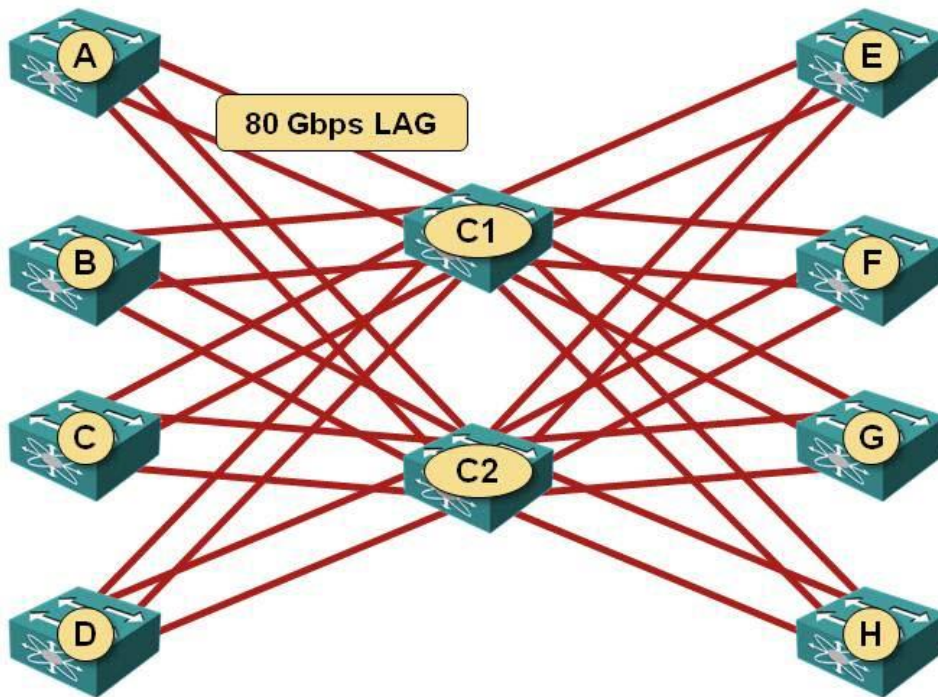
FULL MESH



- Bár 140 Gb/s uplink kapacitásunk van, mégis csak 20 Gb/s érhető el bármely két node között
- Nincs alternatív irány
 - Hibatűrés nincs
- Sok $(n*(n-1)/2)$ link
- LAG: Link aggregation



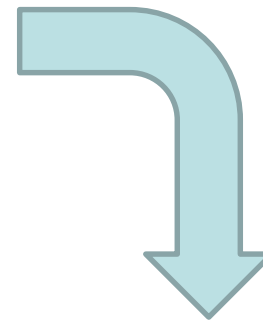
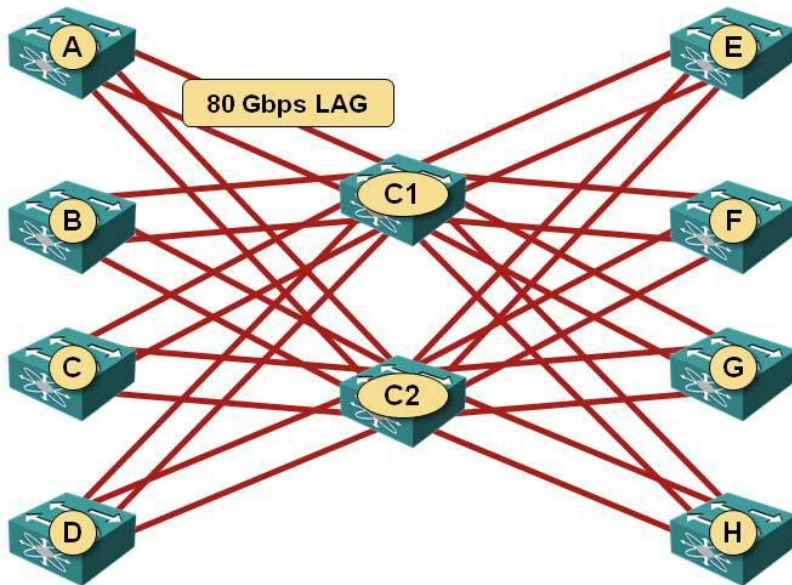
HATÉKONYABB ARCHITEKTÚRA



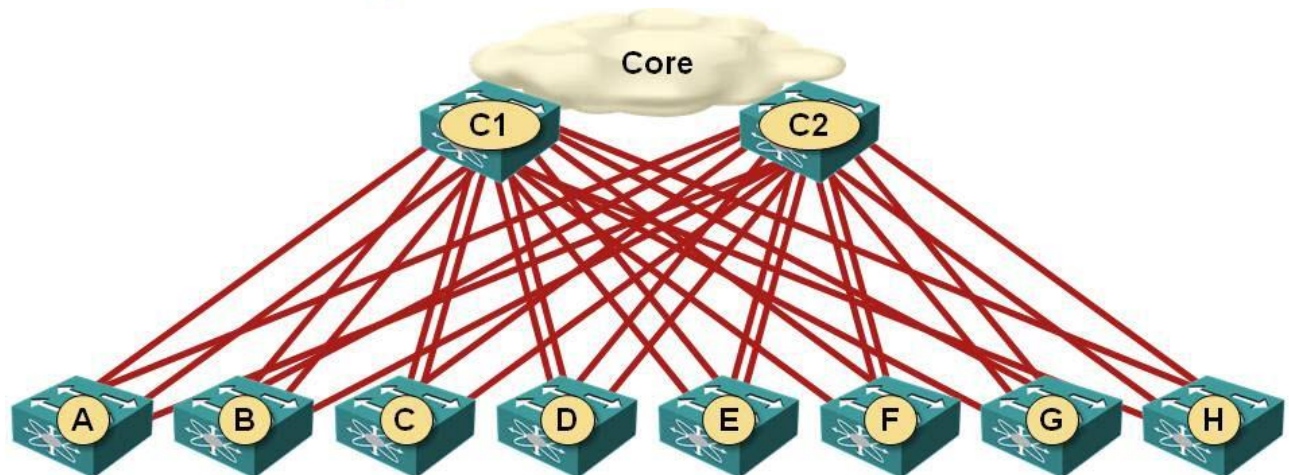
- 160 Gb/s bármely két node között
- Redundáns
- DE:
 - több switch
 - lassabb



RAJZOLJUK CSAK EGY KICSIT MÁSKÉNT...

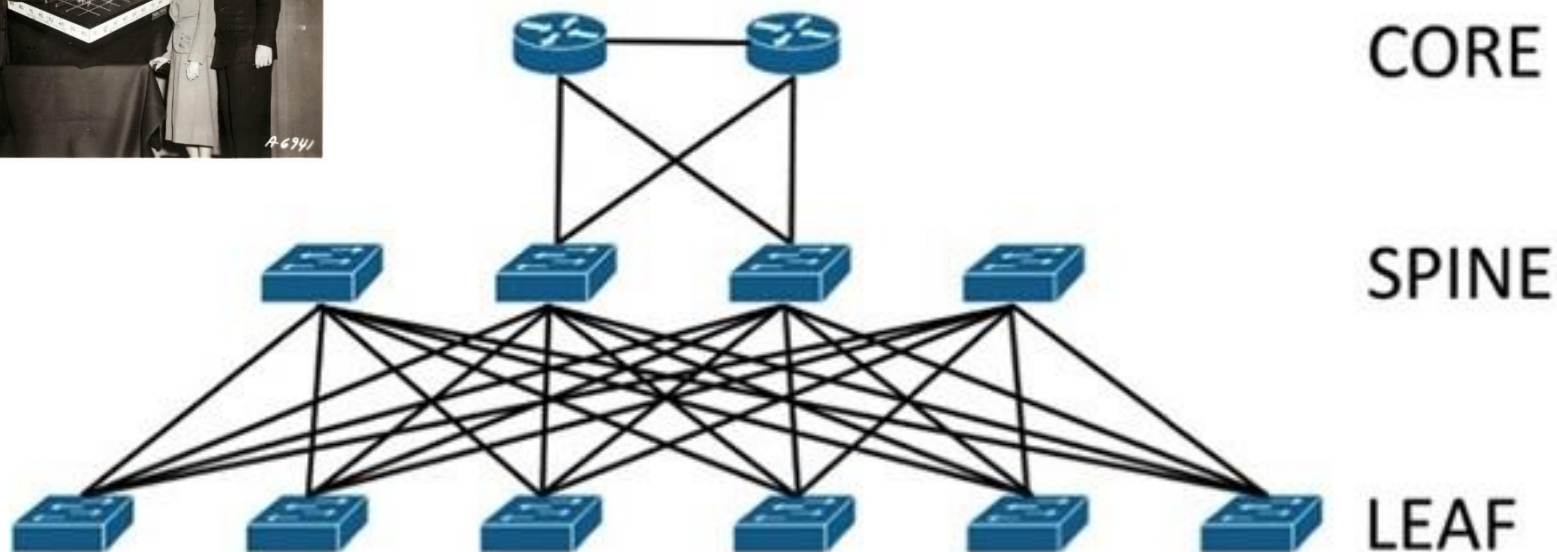


Folded array



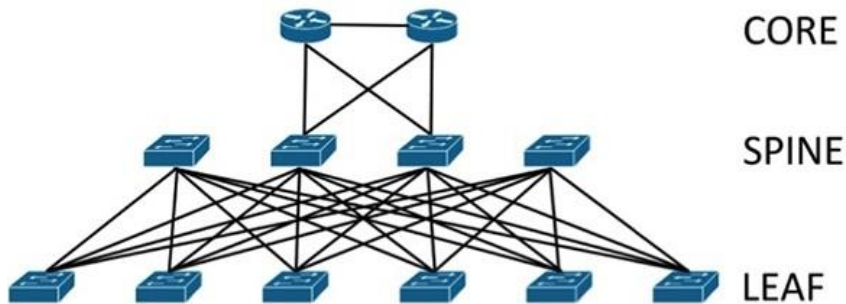
ÚJ TREND

- Clos Network (linkkapcsolás) / Folded Array / Spine and Leaf architecture



SPINE AND LEAF SZABÁLYOK

- Összeköttetések száma = $\#leaf * \#spine$

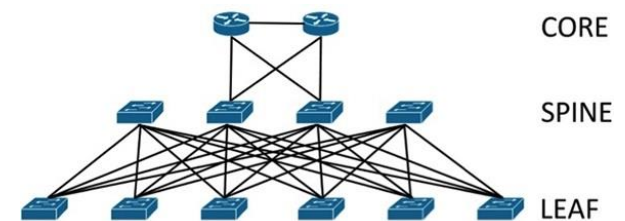


- $\#spine = \#leaf_port$
- $\#spine_port = \#leaf$



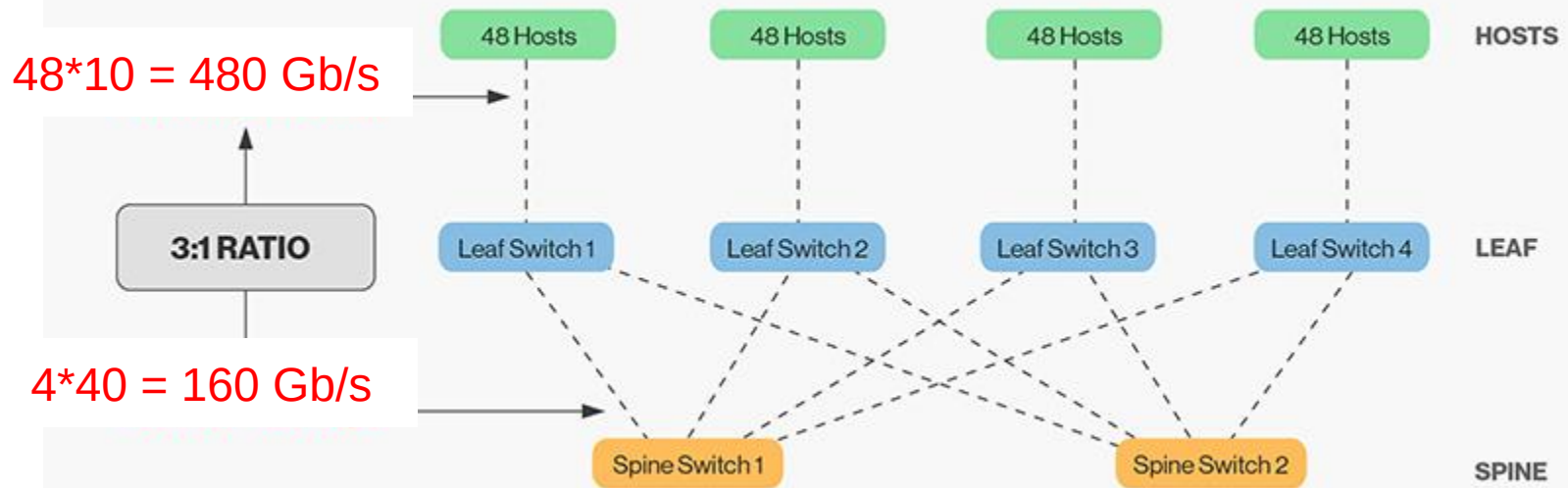
SPINE AND LEAF ARCHITEKTÚRA

- Leaf ~ access
- Spine ~ aggregation
- DE: **minden leaf minden spine-hoz kapcsolódik**
 - Ethernet fabric
 - Maximum távolság: 3 hop (4 link)
 - High performance clusters (HPC)



TIPIKUS ARCHITEKTÚRA

- Leaf:
 - Szerverek felé: $48 * 10 \text{ GbE port}$
 - Spine felé: $4 * 40 \text{ GbE port}$
- Oversubscription: 3:1



TERHELÉSMEGOSZTÁS, PÁRHUZAMOS UTAK

- Tradicionális routing algoritmus – Spanning Tree Protocol (STP – feszítőfa)
 - Hogy elkerüljük a hurkokat – minden node-hoz a legjobb útvonal
 - Mi a fő hátránya?
 - Csak EGY útvonal két node között
- Equal-Cost Multipath (ECMP) protokoll
 - Több, azonos hosszúságú (költségű) útvonalat talál
 - Gyakorlatban tipikusan 8 vagy 16
 - Terhelésmegosztás
 - Jobb sávszélesség kihasználás
 - Csomag újra sorrendezés kell



SPANNING TREE PROTOCOL (STP)

- Feszítőfa protokoll
- Layer 2 protokoll
 - Bridge-eken (switch-eken)
 - IEEE 802.1D (eredetileg)
 - IEEE 802.1Q-2014 (új verziók is)
- Logikai, hurokmentes topológiát épít ki Ethernet hálózatokban



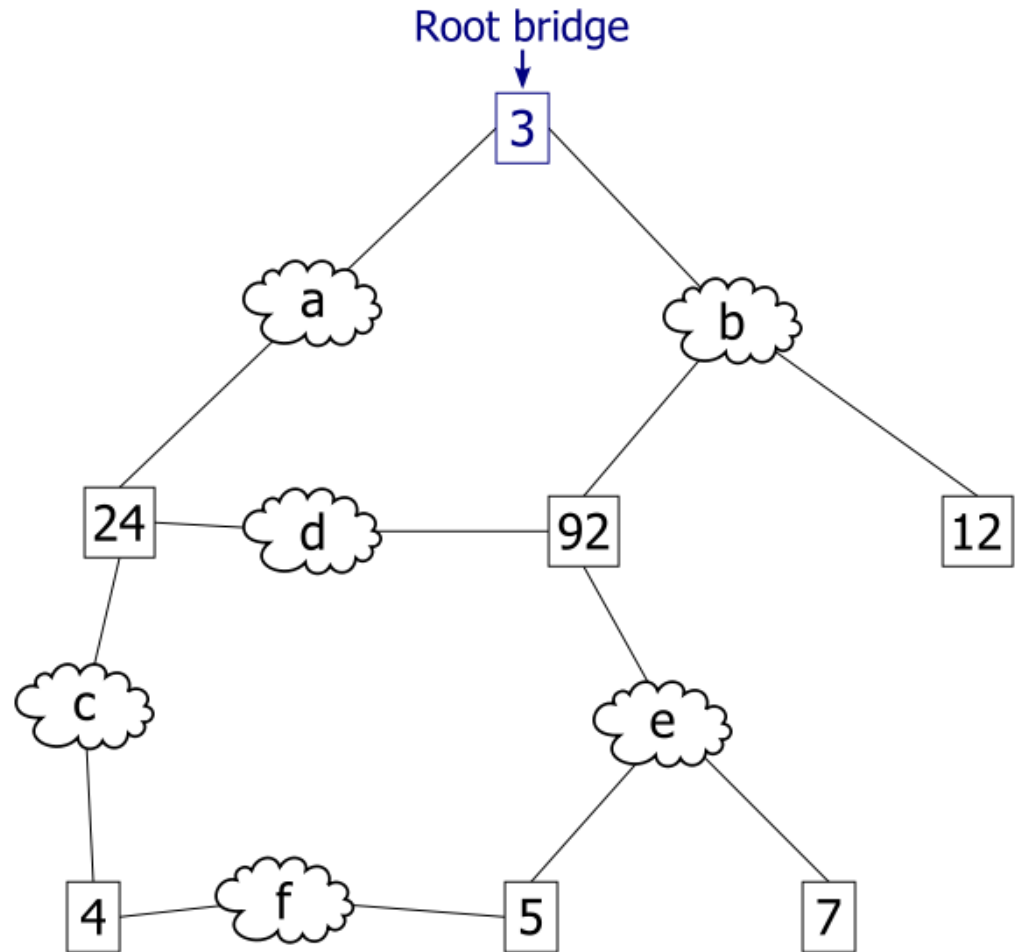
SPANNING TREE PROTOCOL

- Minden bridge-nak meg kell tudnia, milyen „messze” van a root-tól
 - A linkek „költsége” (cost)
 - Root port – az a root felé menő port, ami a „legkisebb költségű”
- Minden hálózati szegmensnek meg kell tudnia, melyik bridge-n keresztül van a „legközelebb” a root
 - Designated (kijelölt) port – a „legközelebbi” bridge azon portja, amihez az adott hálózati szegmensnek kapcsolódik



SPANNING TREE PROTOCOL (STP)

- Kiinduló (root) bridge/ switch választás:
 - Legkisebb ID
 - ID: prioritás (default 32768) + MAC cím
 - Jó, ha központi (backbone switch)



DEFAULT COSTS

Cost of a link – a sebességgel fordítottan arányos

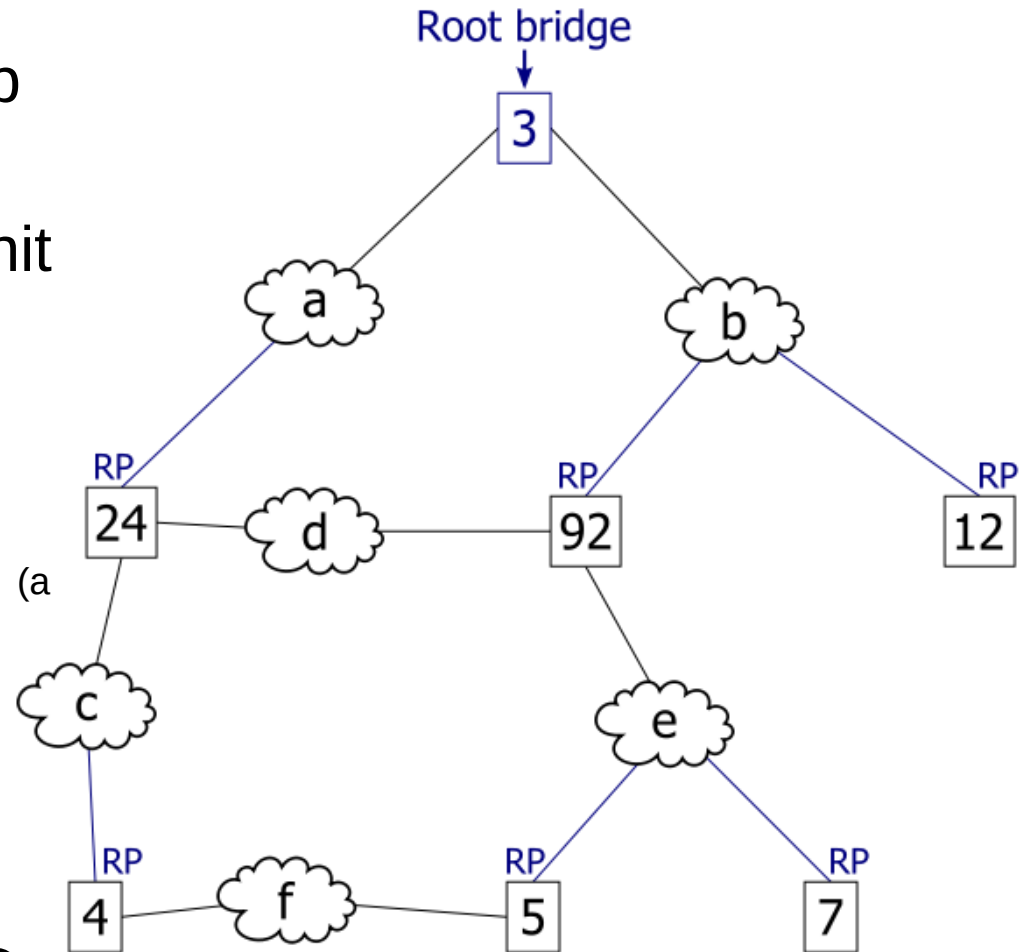
Data rate	STP cost (802.1D-1998)	RSTP (Rapid STP) cost (802.1W-2004, default value)
4 Mbit/s	250	5,000,000
10 Mbit/s	100	2,000,000
16 Mbit/s	62	1,250,000
100 Mbit/s	19	200,000
1 Gbit/s	4	20,000
2 Gbit/s	3	10,000
10 Gbit/s	2	2,000

1 Gigabit/second/bandwidth 20 Terabit/second/bandwidth



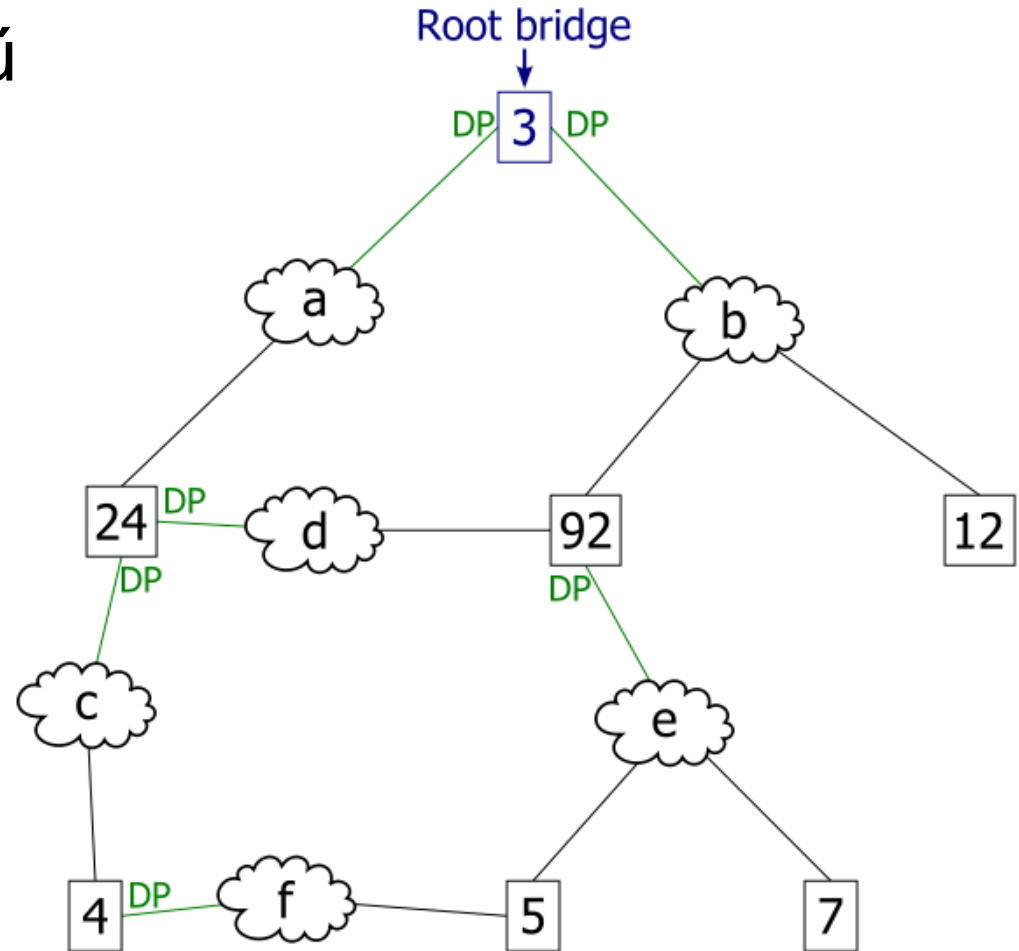
SPANNING TREE PROTOCOL (STP)

- Minden bridge meghatározza a legkisebb költségű utat a root felé
 - Bridge protocol data unit (BPDU) – gyökér cost=0-val küldi
 - A többiek növelik a bejövő link költségével (a példában egyformák)
 - Megkeresni a legkisebbet
 - Root Port (RP)
 - Döntetlennél: kisebb ID-jű bridge felől



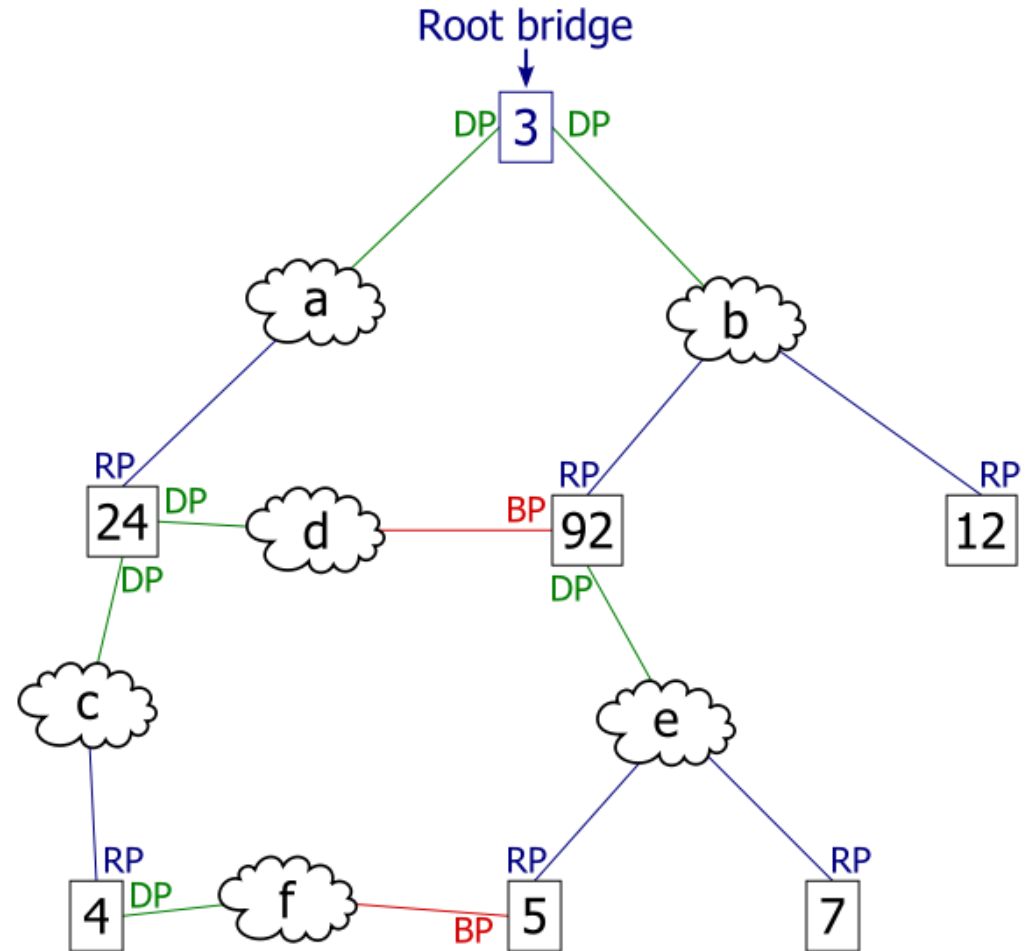
SPANNING TREE PROTOCOL (STP)

- A legkisebb költségű út minden hálózati szegmens felől:
 - A gyökér felé vezető „legkisebb költségű” bridge
 - A szegmens kijelölt (Designated) *port*-ja (DP)
 - Döntetlennél: kisebb ID-jű bridge felől (lásd d)



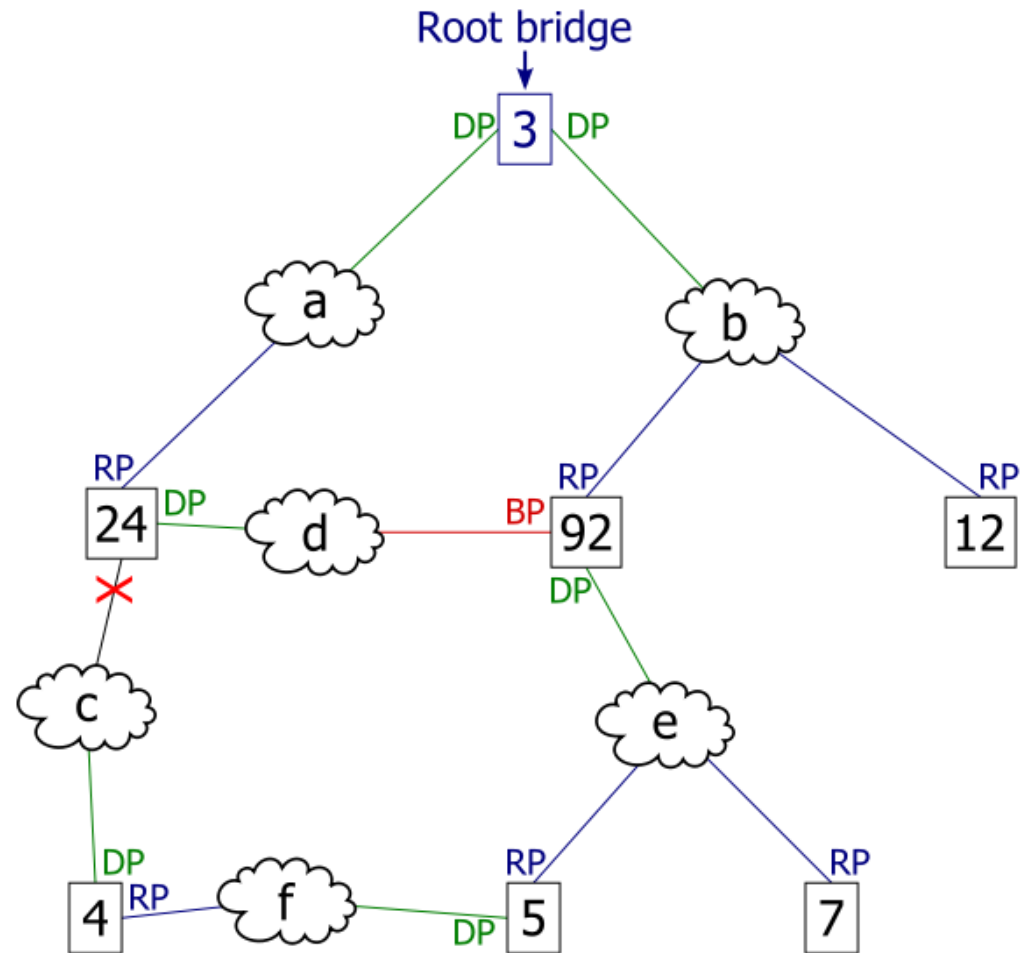
SPANNING TREE PROTOCOL (STP)

- A DP-ket használjuk továbbításra (set to forwarding mode)
- A többi port blokkolt (Blocked Port – BP)



SPANNING TREE PROTOCOL (STP)

- Link hibánál:
 - átkonfigurálás



EQUAL-COST MULTI-PATH (ECMP)

- Layer2 ill. Layer3
- Ha n útvonal van két node között:
 - Modulo n hash algoritmus a választáshoz
 - $\sim$ maradék az osztásnál
 - Különböző hash algoritmusok
 - Header alapján
 - Source/destination MAC/port
 - Source/destination IP+port and Protocol number
 - Egy TCP stream egy linken tartására



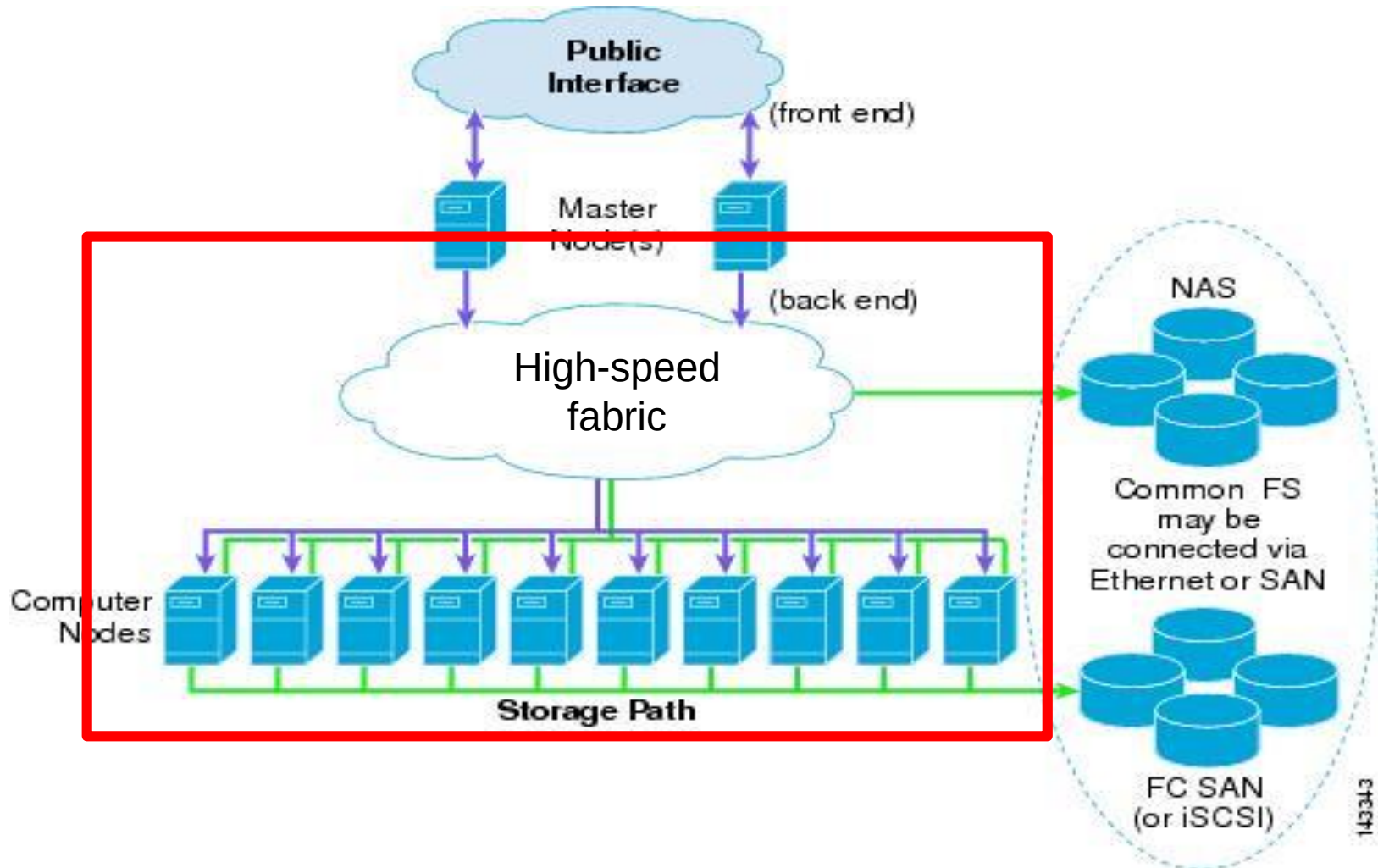


HIGH PERFORMANCE CLUSTERS (HPC)

- Nagyteljesítményű fürtök
- Cluster: közös cél
 - Több CPU kombinálásával
 - Hogy egy egységes, nagyteljesítményű rendszernek tűnjön
 - Nagy rendelkezésre állás
 - Terhelésmegosztás
 - Megnövelt számítási teljesítmény
- Pl.:
 - Meteorológia (időjárás szimuláció)
 - Katonai kutatás
 - Trendanalízis
 - Film animáció
 - Gyártás (autó/repülő tervezés, aerodinamikai szimuláció)
 - Keresőmotorok



SERVER CLUSTER BLOKCDIAGRAM



HPC TÍPUSOK

- 1. típus (szorosan csatolt – tightly coupled)
 - Párhuzamos üzenettovábbítás (párhuzamos műveletvégzés)
 - Az alkalmazások minden node-on párhuzamosan futnak
 - A Master Node határozza meg a node-ok bemenetét
 - Kommunikáció a node-ok között
 - Szimuláció (időjárás, el.mágn. mező, rádióhullám terjedés, folyadék áramlás, jet stream)
- 2. típus
 - Elosztott I/O processzálás
 - Pl.: keresőmotor
 - Master Node szétteríti a kérést
- 3. típus (lazán csatolt – loosely coupled)
 - Párhuzamos fájlfeldolgozás
 - A forrásfájl (kép) feldarabolása és párhuzamos feldolgozásra való szétosztása



SPINE AND LEAF ÉRTÉKELÉS

- Előnyök:
 - Kelet-nyugati forgalom támogatása
 - Virtualizáció (cloud) támogatása
 - Nagy és egyforma kelet-nyugati (inter-cluster) sebesség
 - Minden összekötő linket használ
 - Spanning Tree Protocol (STP) -> Equal-Cost Multipath Protocol (ECMP)



SPINE AND LEAF ÉRTÉKELÉS

- Előnyök:
 - Egyforma switch-ek
 - Fix konfigurációjú switch-ek
 - Kisebb áramfelvétel
 - Olcsó eszközök -> nagyteljesítményű, megbízható konstrukció
 - Ha egy switch / link meghibásodik, csak kismértékű teljesítménycsökkenés



SPINE AND LEAF ÉRTÉKELÉS

- Hátrányok:
 - Több switch
 - A switch-ek portszáma limitálja a maximális méretet
 - Több kábel
 - Egy új spine hozzáadásakor kábelek minden leaf felől
 - Hosszú kábelek
 - Üvegszálak, optikai modulok kellenek (drága) a koax kábelek helyett (olcsók, de csak rövid távra jók)



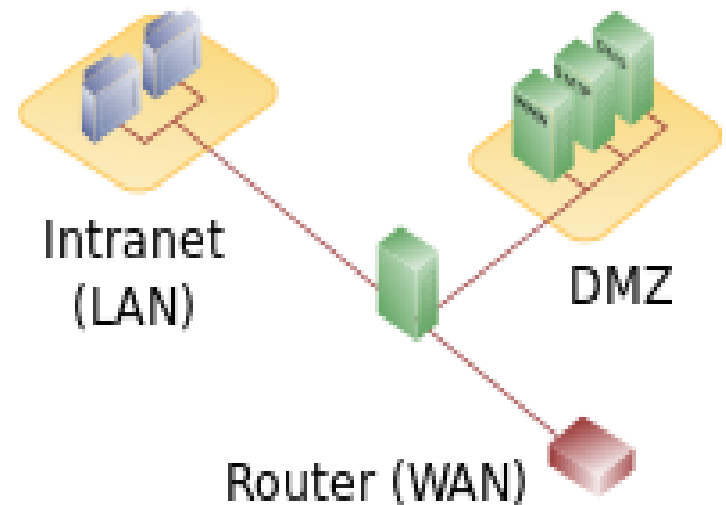
DEMILITARIZÁLT ZÓNÁK

- A legsérülékenyebb host-ok: amik LAN-on kívülre nyújtanak szolgáltatásokat
 - E-mail, web, DNS, FTP, VoIP, ...
 - Helyezzük őket egy szeparált alhálózatba – Demilitarizált zóna (Demilitarized zone – DMZ) vagy biztonsági hálózat (Perimeter Network)
 - Biztonságosabb, mint az Internet, de kevésbé biztonságos, mint az Intranet
 - A DMZ host-jainak korlátozott összeköttetése van (bizonyos) belső szerverekkel és korlátozott/ellenőrzött kommunikáció az Internet-tel
 - Tűzfal(ak)
 - Véd a külső támadások ellen, de nem foglalkozik a belsőkkel



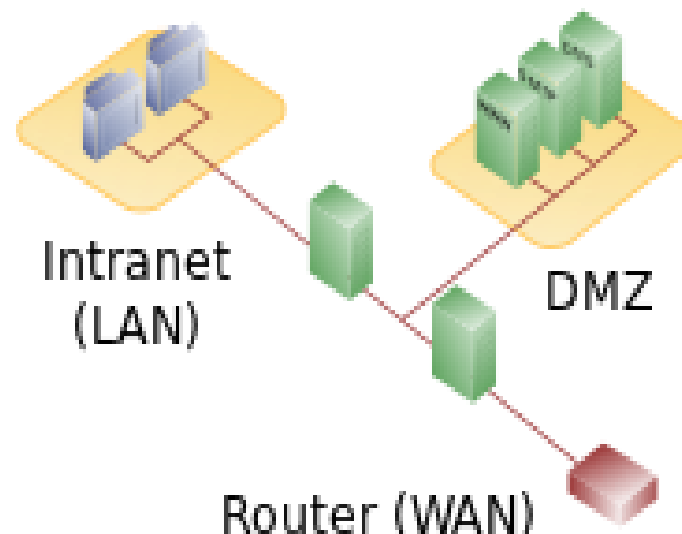
DMZ ARCHITEKTÚRA I.

- Egy tűzfal – háromágú (three legged) modell
 - DMZ-Intranet és DMZ-Internet közötti forgalmat kezeli
 - Single point of failure



DMZ ARCHITEKTÚRA II.

- Kettős tűzfal
 - Front-end/Perimeter
 - Back-end/Internal
- Tipikusan különböző gyártóktól
 - Nem ugyanazok a biztonsági rések
 - Nem ugyanaz a konfiguráció módszere
 - Drágább



PROXY SZERVEREK

- DMZ-ben
 - Kötelezik a (belső) használókat a proxy igénybevételére az Internet felé
 - Biztonság
 - Monitorozhatóság
 - Központi web tartalomszűrés
 - Csökkenti az Internet forgalmat – caching a proxy-ban



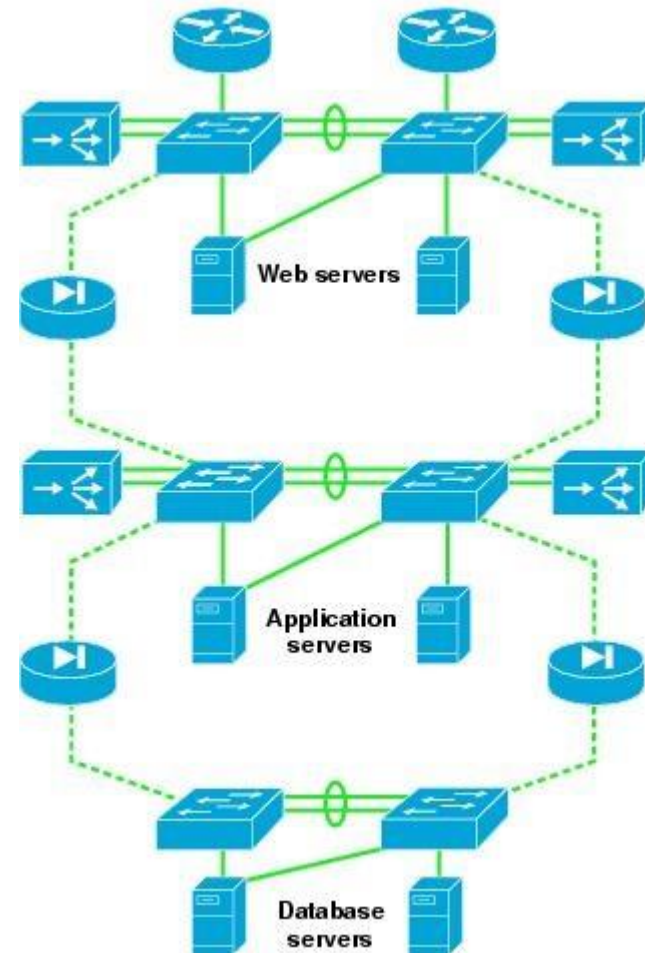
REVERSE PROXY SZERVEREK

- Indirekt hozzáférést biztosít külső hálózatok felől a belső erőforrásokhoz
 - Pl. e-mail hozzáférés cégen kívülről
- Csak a reverse proxy szervernek van hozzáférése a levelezőszerverhez
- Extra biztonsági layer
- Tipikusan applikációs rétegbeli tűzfallal



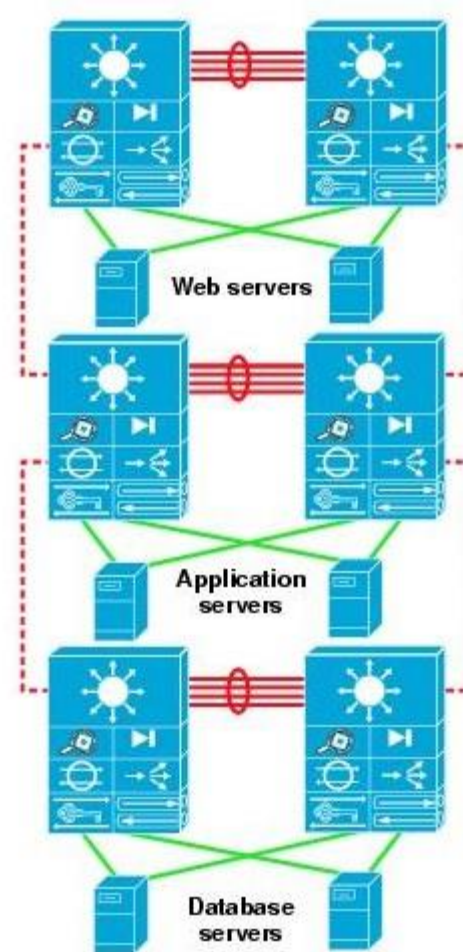
HTTP-ALAPÚ ALKALMAZÁSOK I.

- 3 tiers
 - Web szerverek
 - Alkalmazás szerverek
 - Adatbázis szerverek
- Tűzfalakkal elválasztva



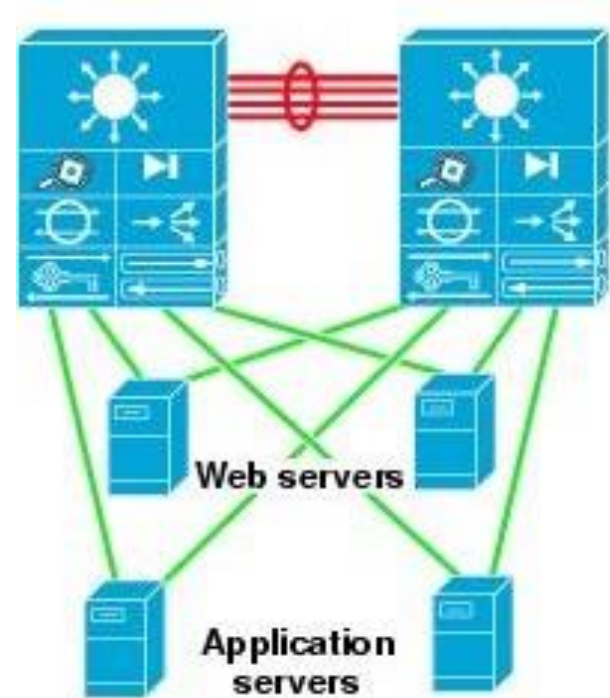
HTTP-ALAPÚ ALKALMAZÁSOK II.

- Integrált szolgáltatás modulok (Integrated Service Modules)
 - Router/switch
 - Tűzfal
 - Terhelésmegosztás
 - Biztonság
 - IDS (Intrusion Detection System – behatolás detektálás)

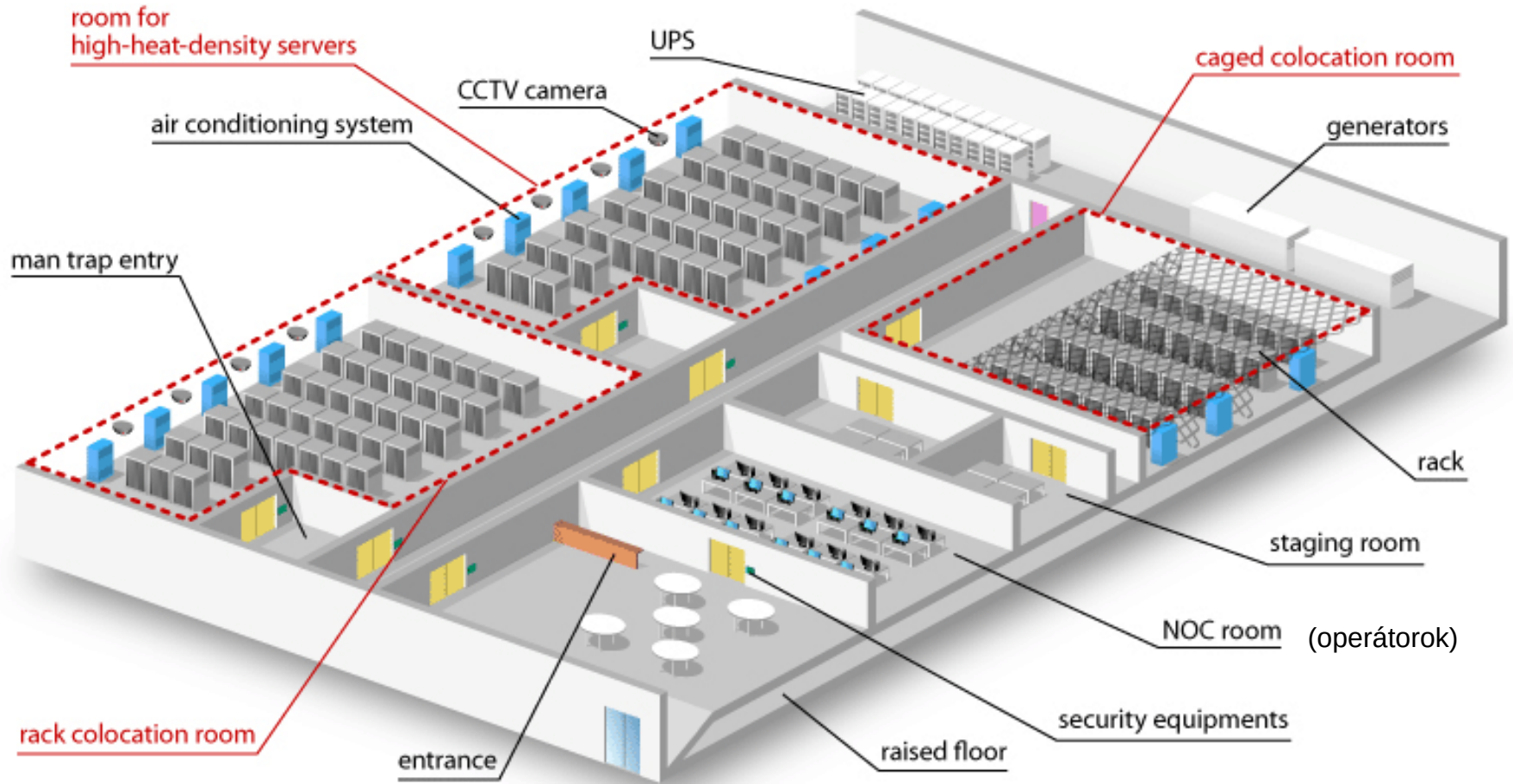


HTTP-ALAPÚ ALKALMAZÁSOK III.

- Logikai elválasztás
 - VLAN-okkal
 - Kisebb teljesítmény
 - De kevesebb eszköz
- + Adatbázis szerverek tipikusan fizikailag is elkülönülnek



ADATKÖZPONT ÁTTEKINTÉS



COLOCATION ROOM (COLO)

- „Közös használatú terem” – az ügyfelek ide helyezhetik berendezéseiket – “carrier hotel”
 - Zárt terek
 - Ketrecek (cage)
 - Rack-ek (zárhatóak)
 - Az eszköz az ügyfél tulajdona
 - Az adatközpont a működtetést és karbantartást biztosítja
 - Hely, áram, hűtés, fizikai biztonság
 - Kapcsolatok távközlési és hálózati szolgáltatók felé



CAGED COLOCATION ROOM



RACK COLOCATION ROOM



STAGING ROOM

- Érkeztető terület
- Eszközök
 - Kicsomagolása
 - Ellenőrző tesztelése
 - Konfigurálása
- A szerver-termeken kívül



ADATKÖZPONTI SZOLGÁLTATÁSOK

- Saját eszköz tárolása – bérlet
- 24 órás felügyelet
- Nagy megbízhatóságú környezet
– georedundancia
- Teljes körű üzemeltetés
- Költözés tervezése és kivitelezése
- Adatmentés és adatvisszaállítás
- Biztonsági mentés vészhelyzet esetére



ADATKÖZPONTI SZOLGÁLTATÁSOK

- Kettős tápellátás két független alállomásról
- Redundáns transzformátor
- Redundáns diesel generátor
 - Automatikus átkapcsolás és indítás
- DC $n+1$ redundanciájú egyenirányító rendszer
- Klimatizálás: $n+1$ redundancia
- Emelt padló, kábeltálcarendszer
- Redundáns optikai gyűrű

