



HÁLÓZATI RENDSZEREK
ÉS SZOLGÁLTATÁSOK
TANSZÉK



Budapest,
2020.05.31.

SZÁMÍTÓGÉP ARCHITEKTÚRÁK

Skálázható és Hibatűrő rendszerek

Horváth Gábor, Belső Zoltán

BME Hálózati Rendszerek és Szolgáltatások Tanszék
ghorvath@hit.bme.hu, belso@hit.bme.hu

Ez a videó a BME Mérnökinformatikus BSc képzésének a tananyaga.
Személyes használatra bárki szabadon megnézheti és birtokolhatja.
A BME Villamosmérnöki és Informatikai Karán kívüli képzésekben csak
a Kar engedélyével lehet felhasználni!

A complex network diagram with numerous nodes of varying sizes and colors (gray, white, and light blue) connected by thin gray lines. Some nodes are highlighted with dashed circles. The background is white.

Skálázhatóság

Nagyvállalati informatikai környezet kihívásai:

- Nagy adatmennyiség feldolgozása lehet szükséges
→ diszk kapacitás
- Nagy lehet a számítási kapacitás igény, sok program futhat egyszerre
→ processzor teljesítmény, memória
- Az adatok képviselik az értéket
→ mentési rendszer
- Sok felhasználó egyidejű kiszolgálása lehet szükséges
→ hálózati kapacitás
- Az adatok titkot képezhetnek
→ adatbiztonság
- Az üzletmenet múlhat az informatikai környezet elérhetőségén
→ rendelkezésre állás

- Hogyan lehet „nagy” rendszert összerakni?
 - Elegendő-e mindenből „sokat” venni?
 - Vagy egy ponton túl a mennyiség már átcsap minőségbe?
- Ezek a skálázhatóság kérdései

Mit jelent a nagy rendszer?

- Sok párhuzamos folyamat → sok processzor
- Nagy adatmennyiség → sok diszk
- Hatékony adatfeldolgozás → sok memória

De vajon elegendő-e egy csomó asztali PC-t venni,
és bezsúfolni egy nagy gépterembe egy csomó polcra?



Ez a feladattól függ!

A feladat a CERN-ben:

- Hihetetlenül nagy adatmennyiség,
- de felbontható nagyon sok viszonylag kicsi,
- egymástól függetlenül feldolgozható részre.
 - jól szétosztható egymáshoz lazán (hálózaton) kapcsolódó, független gépekre
- A feldolgozás sok számítás elvégzését jelenti
 - a fő erőforrás a processzor számítási kapacitás
- Egy gép meghibásodása nem hátráltatja a többi munkáját,
„csak” a ráeső számítást nem végzi el. Adat sem vesz el ezáltal.
 - a gépek egymástól függetlenül cserélhetőek

Egy tipikus nagyvállalati környezet nem ilyen!

Tipikus nagyvállalati feladat: Online Transaction Processing (OLTP)

Például: credit/debit tranzakció

- Pénzt kell átvezetni egyik számláról a másikra
- Egy adatbázis két rekordja érintett
- Az egyik csökken a tranzakció összegével, a másik nő
- Mindezt „egyszerre”, oszthatatlanul: külső szemlélő számára
 - vagy már mind a két művelet sikeres
 - vagy még egyik sem teljesült.
 - lehetnek megkötések
(pl: a számlaegyenleg nem lehet negatív)
- Ilyen tranzakcióból kell adott idő alatt minél többet megcsinálni,
- amelyek nem feltétlenül függetlenek!
(pl: ugyan azt a számlát érintik)

Az adatbázis tranzakciók jellemzői:

- Egymással párhuzamosan is elvégezhetőek
- Egymástól megfelelő lock mechanizmussal izolálni kell
 - Ehhez Inter Process Communication (IPC) eszköz kell
 - Ilyen lock-ból sok lehet
 - Hatékony implementációra van szükség
- Nagy adatmennyiséget mozgat a diszk és a memória között
- Több tranzakció is hivatkozhat ugyanarra az adat tartalomra
 - Előnyös, ha ezeket csak egyszer kell a memóriában tartani
 - Processzek közötti osztott memória használata előnyös

Kifejezetten igénylik a nagy mértékben párhuzamosított rendszereket!

Párhuzamos rendszerek

- Szorosan csatolt, például:
 - A processzorok közös memóriaterületet látnak:
Symmetric MultiProcessing (SMP)
- Lazán csatolt, például :
 - A rendszer tagjai önállóak,
saját memória, független operációs rendszer:
Massively Parallel Processing (MPP)
 - A csatolást valamilyen hálózati kapcsolat biztosítja
- A kettő közötti, például :
 - Van közös memória az egyes processzorok között, de
 - van saját memóriája is a processzoroknak
 - és/vagy nem egyformán hatékonyan érhető el:
Non Uniform Memory Access (NUMA)

Mekkora nagy lehet egy “nagy” rendszer?



IBM z13



IBM Power 795



Lenovo x3960 X6



Dell PowerEdge R930

	IBM Power E980 (1 NODE)	IBM Power E980 (2 NODE)	IBM Power E980 (3 NODE)	IBM Power E980 (4 NODE)
Product Line				
Machine type	9080-M9S	9080-M9S	9080-M9S	9080-M9S
System Packaging	19" rack drawer (7U) One 5U system node & one 2U system control unit	19" rack drawer (12U) Two 5U system nodes & one 2U system control unit	19" rack drawer (17U) Three 5U system nodes & one 2U system control unit	19" rack drawer (22U) Four 5U system nodes & one 2U system control unit
Microprocessor type	64-bit POWER9	64-bit POWER9	64-bit POWER9	64-bit POWER9
# of processor sockets per server	4	8 (4 per system node)	12 (4 per system node)	16 (4 per system node)
Processor Options: GHz (cores/socket) # of cores	3.9 to 4.0 GHz (8) 32 3.7 to 3.9 GHz (10) 40 3.58 to 3.9 GHz (11) 44 3.55 to 3.9 GHz (12) 48	3.9 to 4.0 GHz (8) 64 3.7 to 3.9 GHz (10) 80 3.58 to 3.9 GHz (11) 88 3.55 to 3.9 GHz (12) 96	3.9 to 4.0 GHz (8) 96 3.7 to 3.9 GHz (10) 120 3.58 to 3.9 GHz (11) 132 3.55 to 3.9 GHz (12) 144	3.9 to 4.0 GHz (8) 128 3.7 to 3.9 GHz (10) 160 3.58 to 3.9 GHz (11) 176 3.55 to 3.9 GHz (12) 192
Minimum number cores active	8	8	8	8
Energy Scale	Y	Y	Y	Y
Level 2 (L2) cache per core	512 KB	512 KB	512 KB	512 KB
Level 3 (L3) cache per core	10 MB	10 MB	10 MB	10 MB
Level 4 (L4) cache per socket	Up to 128 MB	Up to 128 MB	Up to 128 MB	Up to 128 MB
System memory: min / max / (min % active) 1600 MHz DDR3 or DDR4	256 GB / 16 TB / (50%)	512 GB / 32 TB / (50%)	768 GB / 48 TB / (50%)	1024 GB / 64 TB / (50%)

Xeon E7-88xx v4 (octa-processor) (Wikipedia)

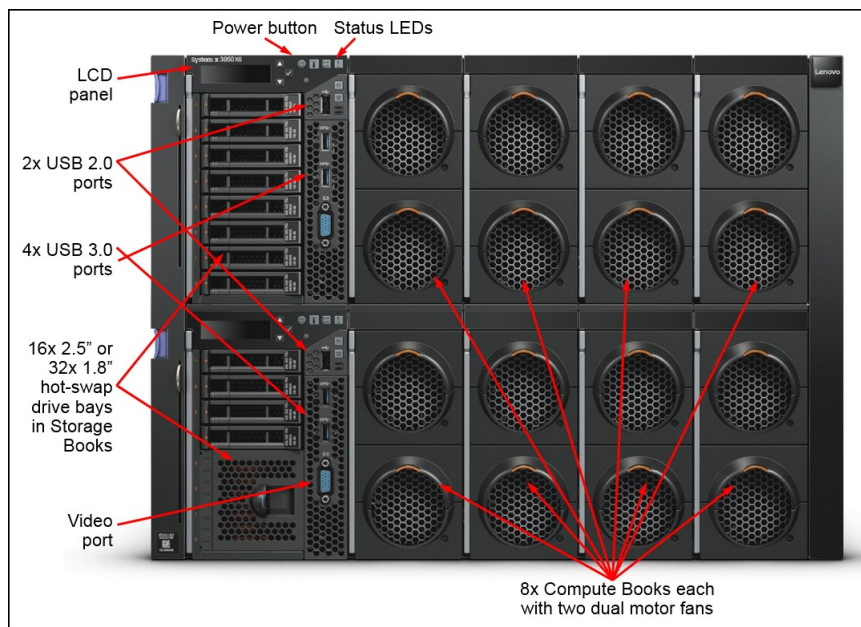
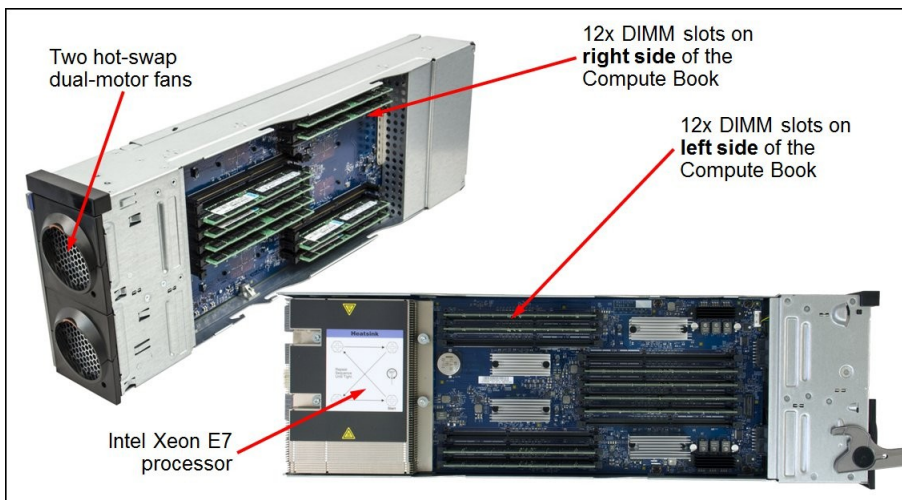
Model number	Cores	Frequency	Turbo	L2cache	L3 cache	TDP	I/O bus	Memory	Release date	Release price (USD)
Xeon E7-880 v4	22	2.2 GHz	4/4/4/4/4/4/4/4/4/4/4/4/4/5/6/7/8/9/11/11	22 × 256 KB	55 MB	150 W	3 × 9.6 GT/s QPI	4 × DDR4-1866 4 × DDR3-1600	June 6, 2016	\$5895
Xeon E7-8890 v4	24	2.2 GHz	4/4/4/4/4/4/4/4/4/4/4/4/4/5/6/7/8/9/10/12/12	24 × 256 KB	60 MB	165 W	3 × 9.6 GT/s QPI	4 × DDR4-1866 4 × DDR3-1600	June 6, 2016	\$7174
Xeon E7-8894 v4	24	2.4 GHz	2.8 GHz(all-core) 3.4 GHz(one-core)	24 × 256 KB	60 MB	165 W	3 × 9.6 GT/s QPI	4 × DDR4-1866 4 × DDR3-1600	February 9, 2017	\$8898

"Cascade Lake-SP" (14 nm) Scalable Performance Xeon Platinum (octa processor) (Wikipedia)

Model number	Cores (threads)	Frequency	Turbo Boost all-core/ 2.0	L2 cache	L3 cache	Memory	Release date	Release price (USD)
Xeon Platinum 8280M	28 (56)	2.7 GHz	3.3/4.0 GHz	28 × 1 MiB	38.5 MiB	6× DDR4-2933	2 April 2019	\$13,012
Xeon Platinum 8284	28 (56)	3 GHz	?/4.0 GHz	28 × 1 MiB	38.5 MiB	6× DDR4-2933	Q2 2019	\$15,460

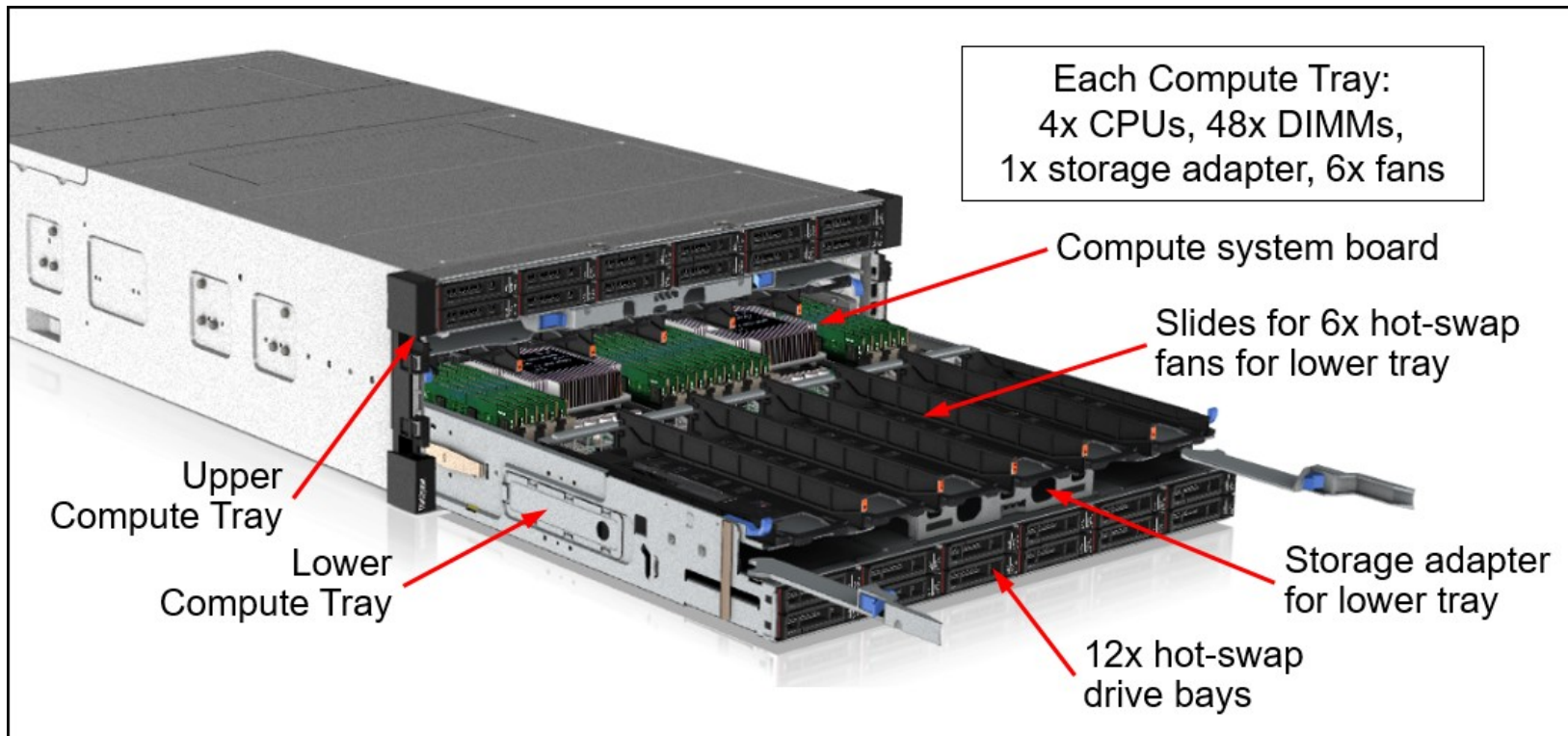
Lenovo System x3950 X6 szerver:

- 8x Intel Xeon E7-8800 v4 processzor, 24 core / socket
összesen $8 \times 24 = 192$ core (384 thread SMT)
- 24 DIMM socket per processzor modul
összesen $192 \times 64 \text{ GB} = 12 \text{ TB}$ memória



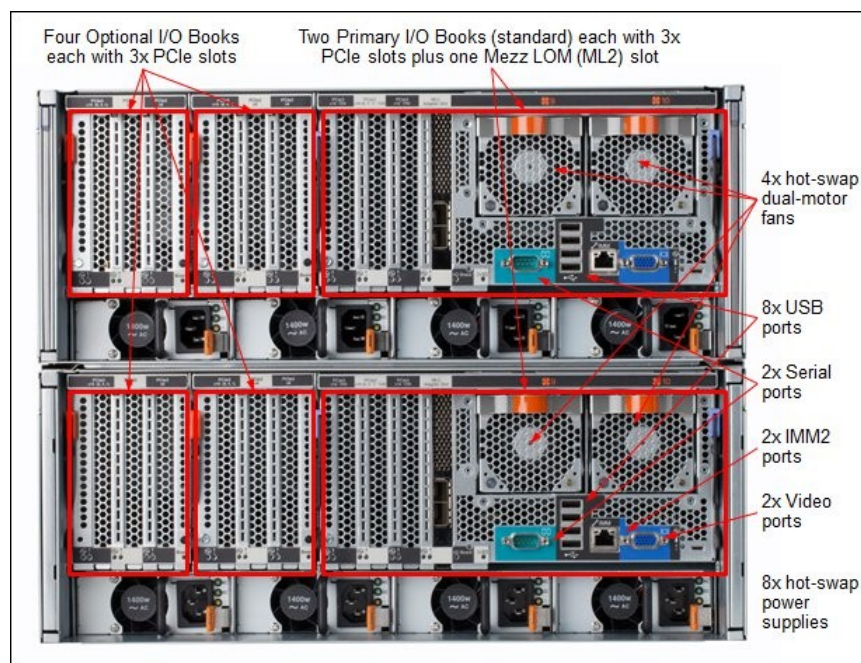
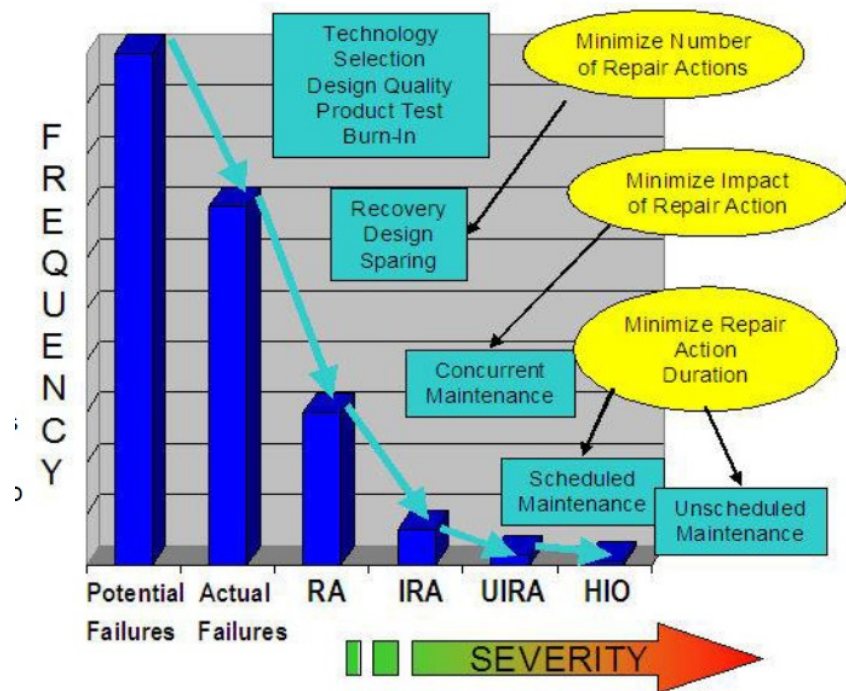
Lenovo ThinkSystem SR950 Server (Xeon SP Gen 2) szerver:

- 8x CPU, 28 core/CPU
- 96 DIMM slot, max: 24 TB (96x 256 GB, 2933 MHz)



Nem csak a processzorok száma vagy a memória mérete számít!

- **Balanced System Design:**
a rendszer komponensei (memória, IO sávszélesség, belső buszok) együtt skálázódnak. Nincs szűk keresztmetszet
- **RAS: Reliability, Availability and Serviceability**





Magas rendelkezésre állás

A rendelkezésre állás mértékegysége: “Hány kilences?”

- Az idő 90%-ában rendelkezésre áll (egy kilences rendelkezésre állás): az év 365 napjából ~328.5 napig üzemel, ~36.5 napig áll.
- Két kilences (99%-os) rendelkezésre állás: évi ~361.35 napig üzemel, ~3.65 napig áll.
- Három kilences (99.9%): ~364.6 nap üzem, 0.365 nap (8 óra 45 perc) állás
- Négy kilences (99.99%): ~355.96 nap üzem, ~52 perc állás
- Öt kilences (99.999%): évi 5.3 perc állás
 - Ez még egy reboot-ra sem elég!

De a rendszerünk nem egyetlen komponensből áll!

- „Egy rendszer olyan erős, mint a leggyengébb láncszeme.”
 - Sajnos nem, sokkal gyengébb.
- Példa: Tekintsünk egy rendszert, amely két komponensből áll:
 - Mindegyik egyenként 90% rendelkezésre állású.
 - A komponensek egymástól függetlenül képesek hibázni.
 - A rendszer akkor áll rendelkezésre, ha mindkét komponense rendelkezésre áll. Bármelyik hibázik, a rendszer hibázik.
 - A teljes rendszer rendelkezésre állása:

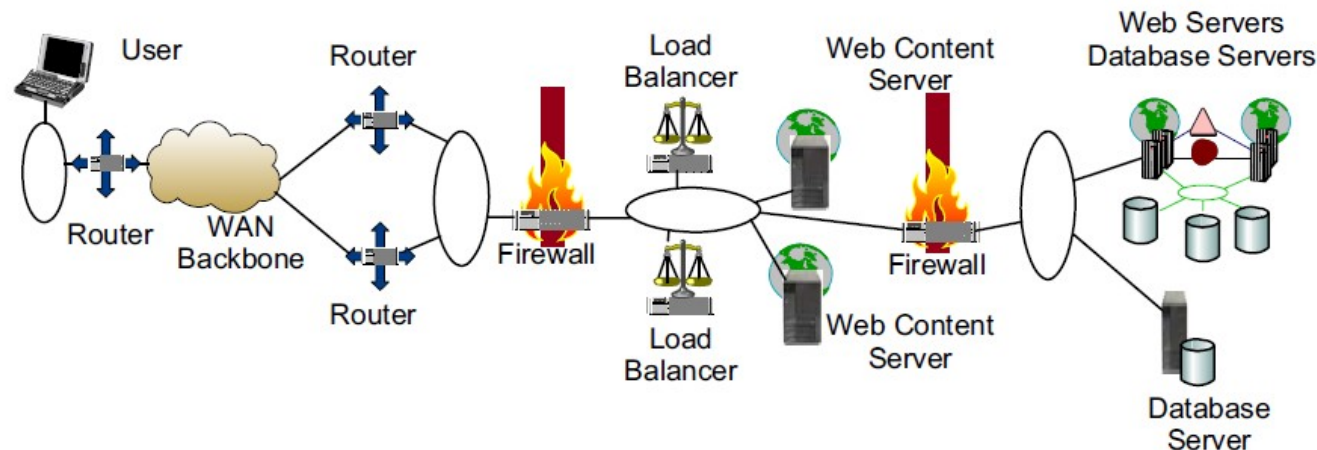
$$0.9 \cdot 0.9 = 0.81 = 81\%$$

De egy rendszer nem csupán két független komponensből áll!

Példa:

- 30 komponens, egyenként 99% rendelkezésre állás
- A teljes rendszer rendelkezésre állása:

$$0.99^{30} = 0.7397 \approx 74\%$$



A teljes rendelkezésreállást az alacsonyabb rendelkezésre állású komponensek határozzák meg.

Példa:

- 30 komponens
- 27 közülük 100% rendelkezésre állású,
- 3 pedig 90% rendelkezésre állású
- A teljes rendszer rendelkezésre állása:

$$1^{27} \cdot 0.90^3 = 0.729 \approx 73\%$$

Hogyan lehet javítani a rendelkezésre álláson?

- Kritikus komponensek duplikálása
- Például: diszk tükrözése
- Ha a duplikált komponens pár bármelyike rendelkezésre áll
→ a rendszer rendelkezésre áll
- Más szóval: a rendszer az egyik komponens meghibásodását még elviseli
- A második meghibásodása már „végzetes”

Példa:

- 99% rendelkezésre állású komponens duplikálva
- A teljes rendszer rendelkezésre állása:

$$1 - (1 - 0.99)^2 = 0.9999 = 99.99\%$$

Kis túlzással a duplikált komponenst (a rendszer többi részéhez képest) tekinthetjük hibamentesnek!

Duplikáljunk hát meg minden komponenst!

Ez persze általában nem ilyen egyszerű, nem elég a komponenseket megduplikálni, az egész rendszert úgy kell felépíteni, hogy bármely komponens meghibásodását képes legyen eliminálni az alternatív komponens használatával.

Olyan rendszer kell, amelynek nincs olyan komponense, amelynek a meghibásodása (egymagában) a teljes rendszer leállítását okozza.

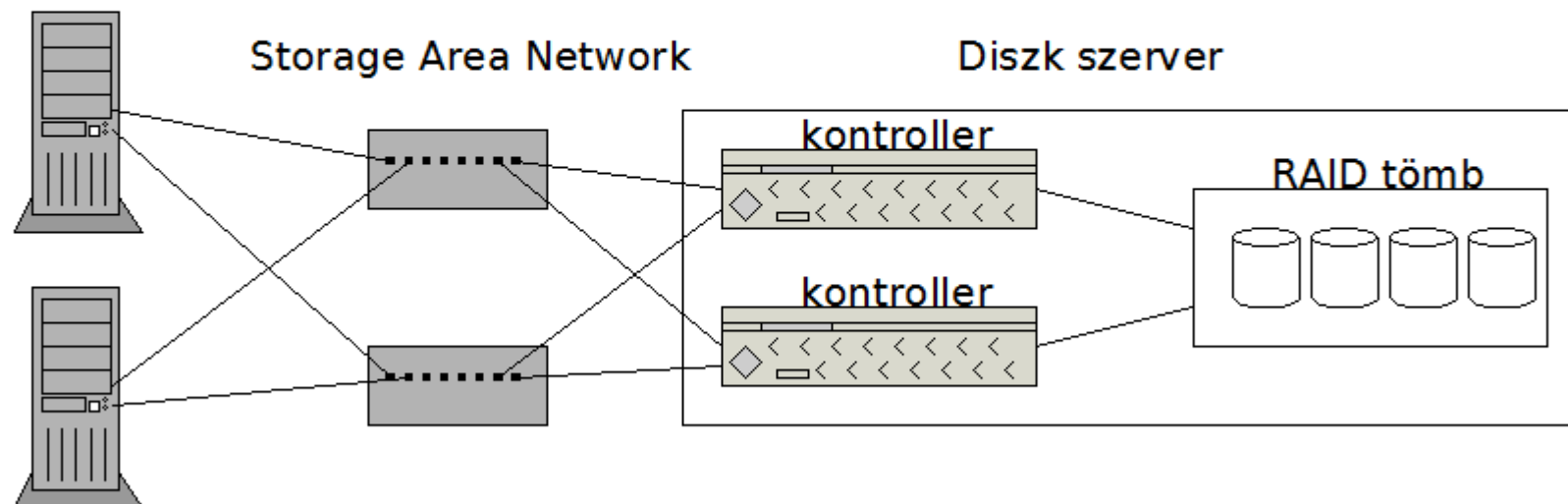
„No Single Point of Failure”
azaz: nincs egy pontú hibahely

Példa: diszk alrendszer

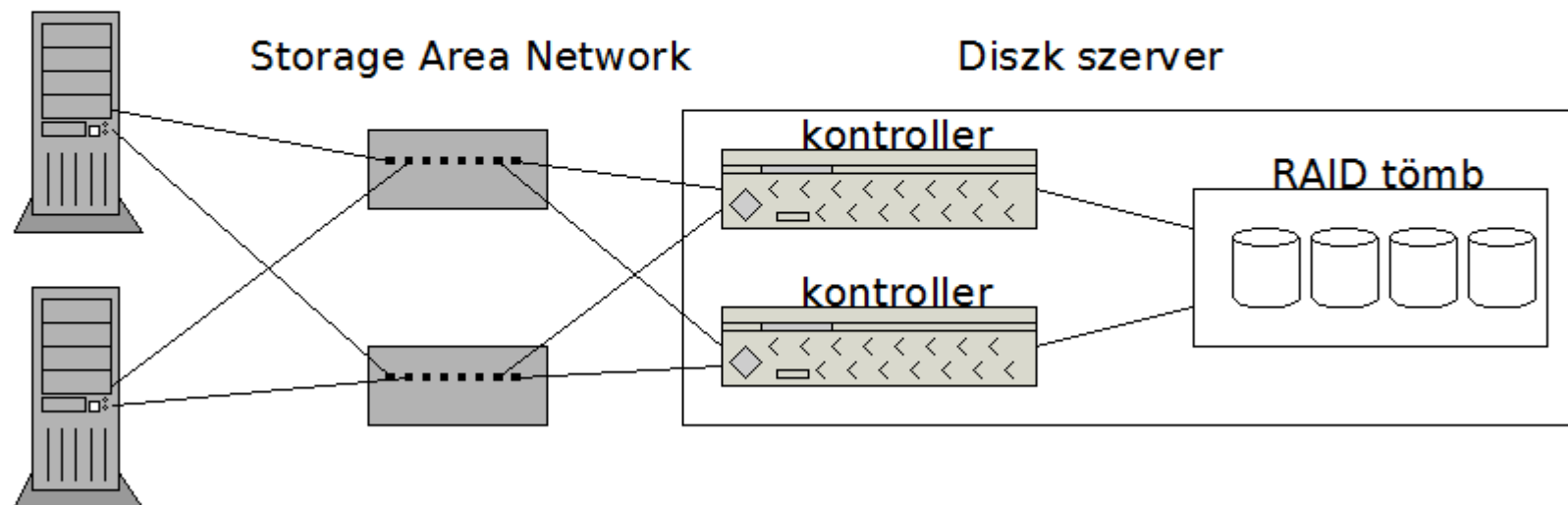
- Az egyes lemezek meghibásodása ellen véd a RAID rendszer kialakítása.
- De mi véd a RAID vezérlő kártya meghibásodása ellen?
- És a RAID vezérlő kártya és a processzor (memória) közötti busz meghibásodása ellen?
- Tápegység meghibásodása?

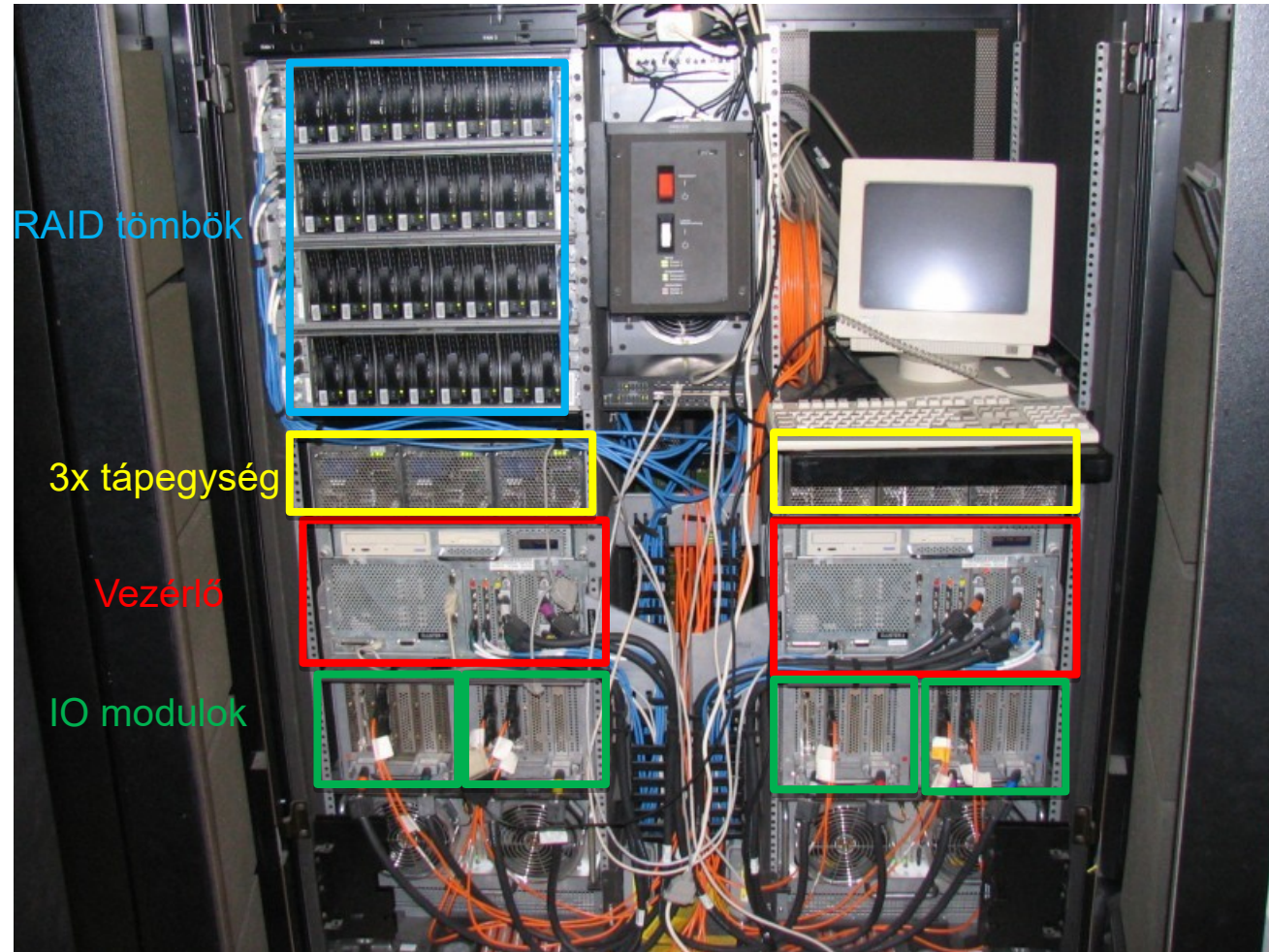
Megoldás: Diszk szerver

- A lemezeket emeljük ki a számítógépből, egy külön szerverbe
- Minden lemez két portos (két kontrollerhez képes csatlakozni)
- Redundáns kontroller (ezek önálló szerverek)
- A számítógép(ek) mindkét kontrollerrel kapcsolatban állnak
→ alternatív elérési utak léteznek a RAID tömbhöz (MultiPath IO).



- Kapcsolat a számítógép(ek) és a diszk server között:
 - Közvetlen: SCSI vagy FC (Fiber Channel)
 - Hálózat: SAN (Storage Area Network) vagy LAN
- Az egyik kontroller vagy bármelyik összeköttetés meghibásodása esetén I/O error keletkezik, és az operációs rendszer megismétli az I/O műveletet egy alternatív útvonalon.





IBM Enterprise Storage Server „Shark”

További bonyolítás: a kontrollerekben gyorsítótár (cache) van, amelyet az írási műveletek gyorsítására is szeretnénk használni.

- Az operációs rendszer és a programok arra számítanak, hogy amikor egy (írási) I/O művelet (sikeresen) véget ért, akkor az adat biztonságban van a diszken.
- De az I/O gyorsítása érdekében a diszk szerver már akkor késznek „jelenti” az írási műveletet, amikor még csak a cache-be íródott be.
 - Egy áramszünet ekkor adatvesztéshez vezet!
 - Hasonlóképpen a kontroller meghibásodása!
- Ez probléma, mert a programok nem erre számítanak!!!

Védekezés:

- Nemfelejtő cache használata:
Például akkumulátorral megtáplált SRAM alapú memóriával
 - Drága és korlátozott kapacitású
 - Az olvasási műveletekhez sürgősszerűen a nemfelejtő cache
 - Továbbra sem véd a controller meghibásodása ellen
- Legyen két cache:
 - egy nagy méretű normál (DRAM alapú), felejtő
 - és egy kisebb, nemfelejtő
- Írási műveletnél a controller az adatot beírja:
 - a saját normál cache-ébe
 - és a másik controller nemfelejtő cache-ébe

- Normális üzemmenet esetén:
 - később a normál cache-ből kiírja a lemezre,
 - illetőleg visszaolvasás esetén ugyaninnen szolgálja ki
- Ha a kontroller hibázik:
 - a másodlagos kontroller ezt észleli és
 - először is kiírja, amit a nemfelejtő cache-ében talál,
 - majd átveszi a további I/O kérések kiszolgálását
- Áramszünet vagy más nemtervezett leállás esetén:
 - Újraindítás után mindkét kontroller:
 - először ellenőrzi, hogy mi van a nemfelejtő cache-ben
 - ha van dirty adat, azt kiírja,
 - majd az egyikük megkezdí az I/O kérések kiszolgálását



Klaszter

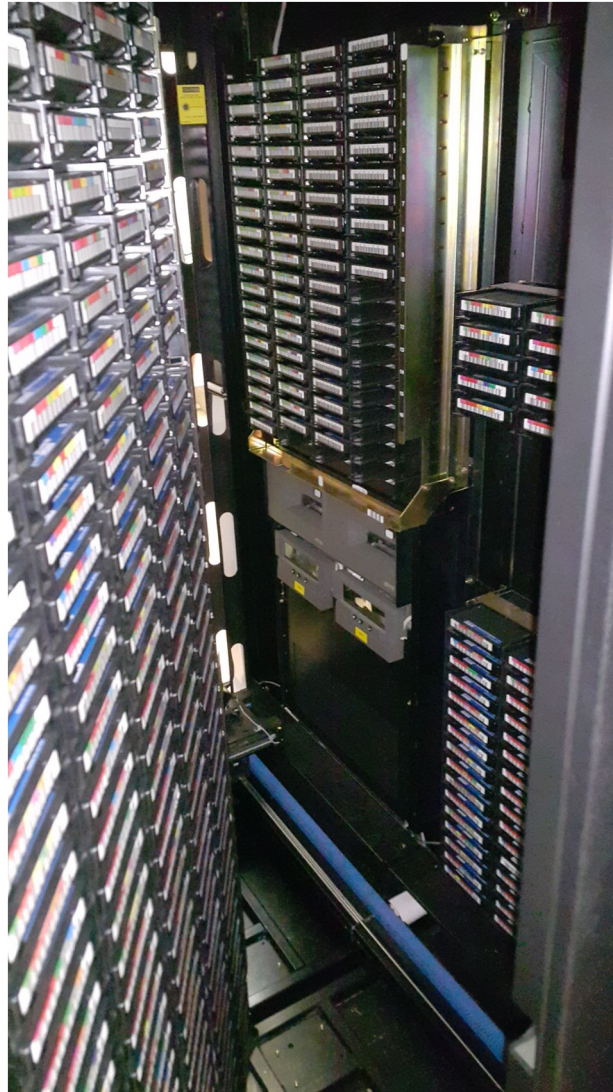
- Mi a helyzet a host processzorral és a memóriával?
 - Ezeket nem könnyű duplikálni
 - Ezért inkább az egész gépet duplikáljuk
→ Massive Parallel Processing (MPP)
- Két gép két független operációs rendszerrel fut
 - Amelyek képesek ugyanazt az adatot (diszket) elérni
 - Folyamatosan figyelik egymást (heartbeat), észlelve ha a másik hibázik (elérhetetlenné válik)
 - A két gép kívülről egy rendszernek látszik → Klaszter
 - Tipikusan:
 - az egyik gép az éles (szolgáltató),
 - a másik a tartalék
- High Availability Cluster MultiProcessing (HACMP)
Parallel Sysplex

- Amennyiben az éles rendszer elérhetetlen, a tartalék:
 - Megfogja (mount-olja) a kieső rendszer lemezeit
 - Elindítja a kieső szolgáltatás folyamatait
 - Szükség esetén átcímzi a hálózati interface-eit
- A tartalék rendszer normál üzemben lehet:
 - Tétlen várakozó (cold standby)
 - Dolgozhat más szolgáltatáson (hot standby)
 - Dolgozhat ugyanazon a szolgáltatáson (load balance)
- Tervezett leállások, karbantartások is kivédhetőek:
 - Átterheljük a forgalmat a tartalék rendszerre,
 - majd a felszabaduló oldalon elvégezzük a szükséges karbantartást, pl:
 - hardver javítás
 - Szoftver frissítés

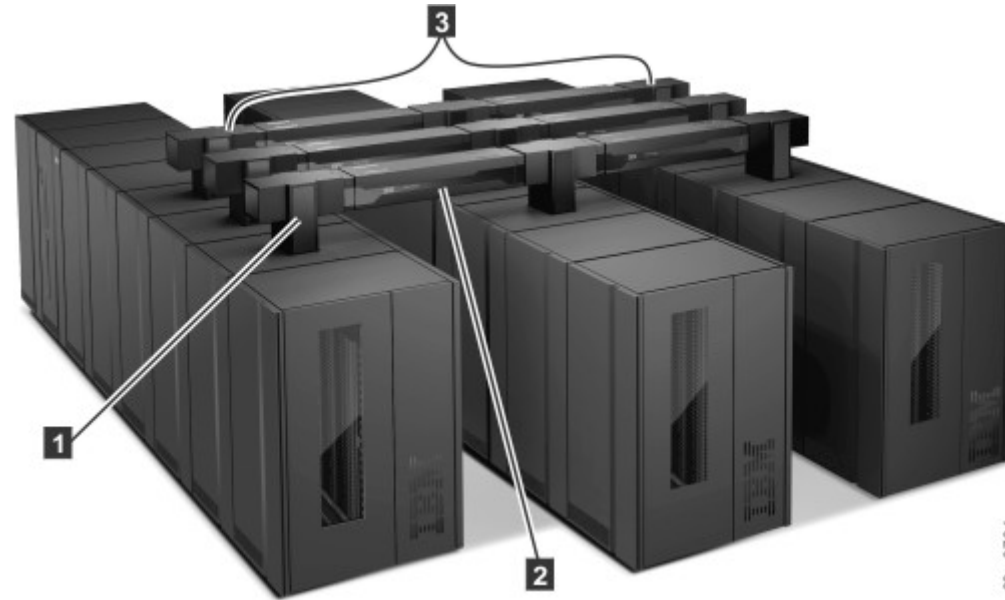
- Mi van, ha a redundáns elemeket egyszerre, közös okból éri kiesés?
 - Árvíz, földrengés
 - Terrortámadás
 - De akár „csak” az áramszolgáltatás szünetelése
- A tartalék rendszert oly mértékben kell elkülöníteni, hogy ne lehessen ilyen „közös ok”
 - Földrajzilag elkülönített redundáns rendszer
Geographically Dispersed High Available Cluster
 - Két gépterem, két diszk szerver, kétszeres szerverek
 - Köztük dedikált optikai kábelén LAN és SAN kapcsolat
 - A lemez írárok mindkét diszk szerverre megtörténnek:
Metro Mirror, Remote Copy

- Egy nagyvállalati rendszerben a legértékesebb elem az ADAT!
- Adatvesztés nem csak hardver meghibásodásból adódhat:
 - Szoftver hiba
 - Emberi tévedés
- Ezek ellen az adat duplikálása (tükrözés, RAID) nem véd, mert a másolat is hibás lesz!
- Offline mentési rendszerre van szükség
- Nagy adatmennyiség tárolására a legpraktikusabb a mágnes szalag
- Kivehető, eltehető a páncélszekrénybe
- Időben több kópiát lehet tárolni
- Véd a véletlen törlések, esetleg vírus támadások ellen is.

- Sok adat
→ sok kazetta
→ nyilvántartási
rendszer is kell
hozzá!
- Tape Library:
kazetta tároló
és mozgó
robot rendszer
- Mai szalagok:
4-10 TB per
kazetta
- A SAN hálózat
része



- A mentési rendszer nem egyenlő szalagok teleírásával, visszatölthetőnek kell lennie!
- Konzisztens állapotot kell kimenteni
- Több szintje van:
 - File szintű mentés: egyedi file-ok tölthetők vissza
 - Rendszer szintű mentés: teljes lemezek mentése
 - Adatbázis mentés: táblák mentése
- Katasztrófa mentés:
Bare Metal Restore





HÁLÓZATI RENDSZEREK
ÉS SZOLGÁLTATÁSOK
TANSZÉK

