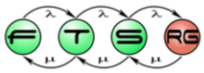
Intelligens rendszerfelügyelet

Fürtözés és replikáció

Micskei Zoltán

(részben Medgyesi Zoltán munkája alapján)



Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

Utolsó módosítás: 2015. 04. 27.

Bevezető

- Cél: hibatűrés
  - számítógép hibák tolerálása
- Mikor éri meg:
  - Egy géppel elérhető: ~99%-os rendelkezésre állás (évi max 3,5 nap kiesés)
  - Ha ennél jobbat akarunk
- Redundancia beépítése

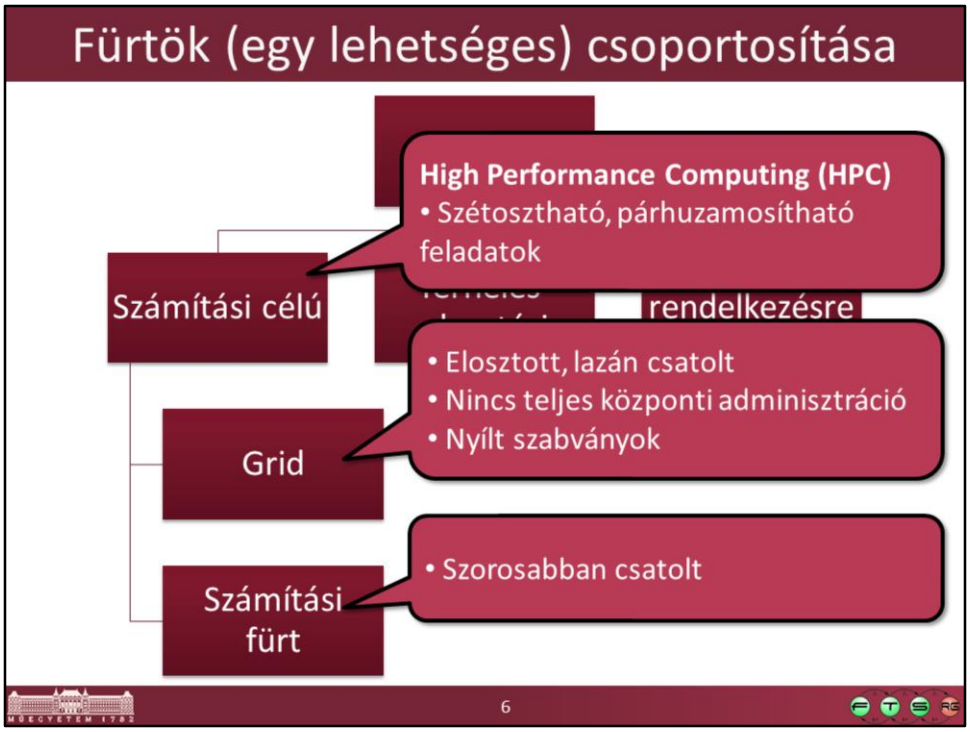
Tartalom

- Fürtök
  - **Fürtök csoportosítása**
  - Terheléselosztó fürtök
  - Feladatátvételi fürtök
- Replikáció
  - Elsődleges – másodlagos séma
  - Multimaster

A számítógépfürt

Fürt (cluster): különálló *számítógépek együttese*, amelyek egymással együttműködve és azonos szolgáltatásokat, alkalmazásokat futtatva egyetlen rendszerként, *virtuális kiszolgálóként* jelennek meg az ügyfelek számára.

Fürtök (egy lehetséges) csoportosítása

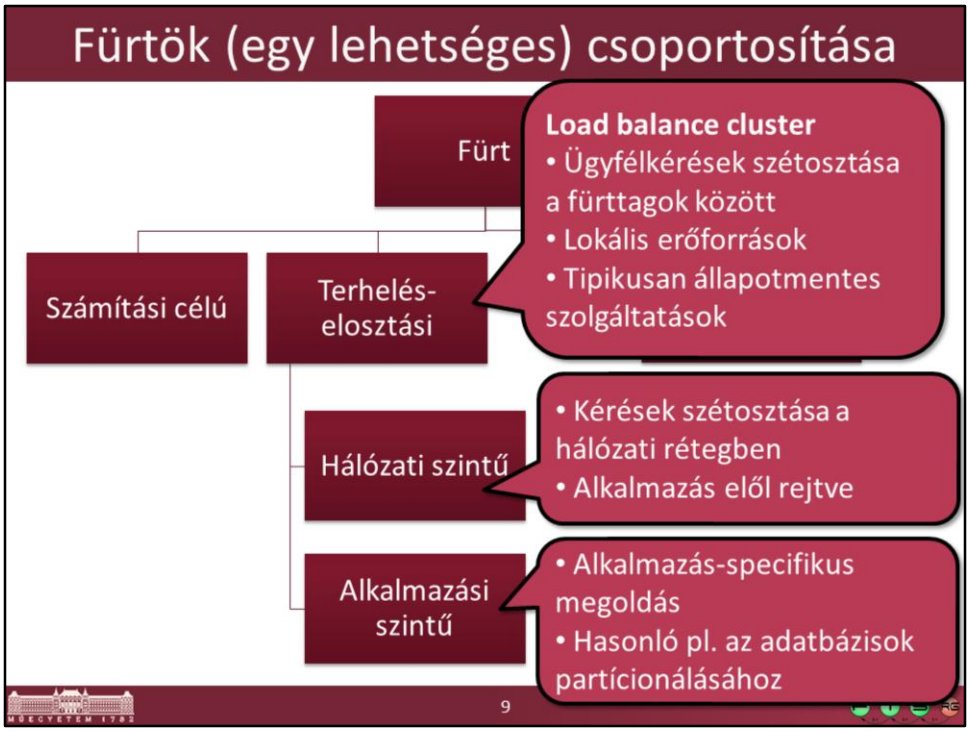


Grid: lásd pl. Ian Foster. „What is the Grid? A Three Point Checklist”, July 20, 2002.
URL: http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf

Számítási fürt példa: Cray TITAN

- CPU:
  - 18 688 darab AMD Opteron 6274 16-core CPUs
  - 18 688 darab Nvidia Tesla K20X GPUs
- Memória: 710 TB (598 TB CPU and 112 TB GPU)
- Sebesség: 17.59 petaFLOPS (LINPACK)
- 200 darab rack * 24 blade

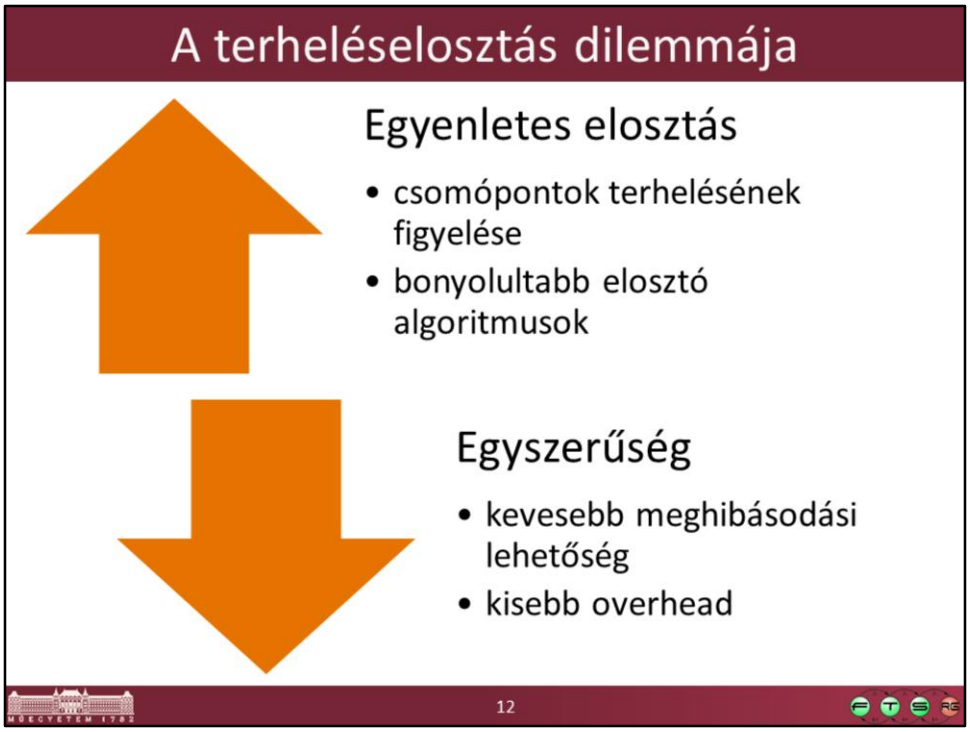
Forrás: http://en.wikipedia.org/wiki/Titan_(supercomputer)
2015-ben ez most épp már csak a másodk leggyorsabb szuperszámítógép


Fürtök (egy lehetséges) csoportosítása
Fürt
Számítási célú
Terhelés-elosztási
Hálózati szintű
Alkalmazási szintű
Load balance cluster
• Ügyfélkérések szétosztása a fürttagok között
• Lokális erőforrások
• Tipikusan állapotmentes szolgáltatások
• Kérések szétosztása a hálózati rétegben
• Alkalmazás elől rejtve
• Alkalmazás-specifikus megoldás
• Hasonló pl. az adatbázisok partícionálásához
9

Fürtök (egy lehetséges) csoportosítása

Tartalom

- Fürtök
  - Fürtök csoportosítása
  - **Terheléselosztó fürtök**
  - Feladatátvételi fürtök
- Replikáció
  - Elsődleges – másodlagos séma
  - Multimaster

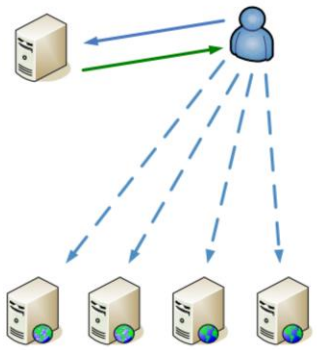


Nem biztos, hogy az a terheléselosztó algoritmus a leghatékonyabb, ami folyamatosan figyeli a csomópontok terhelését, megpróbálja nagyon pontosan megbecsülni az aktuális kérés munkaigényét, és ez alapján nagyon egyenletes elosztást produkálni, mert a sok plusz munka nagy terhelés esetén túl sok időt és erőforrást emészt fel, és emiatt nem lesz hatékony a kiszolgálás.

Hálózati terheléselosztó fürtök fajtái

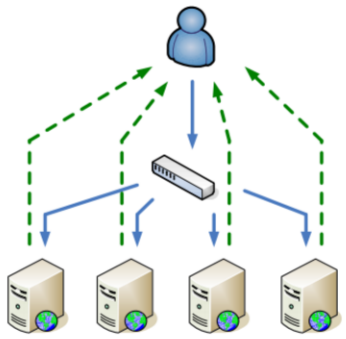
- Round-robin DNS
- Teljesen elosztott
- Központi elemre épülő

Round-robin DNS



- DNS szerver más-más címet ad vissza kérésenként
- Pl.: nslookup www.cnn.com
- Előny:
  - egyszerű
  - független fürttagok
- Hátrány:
  - statikus

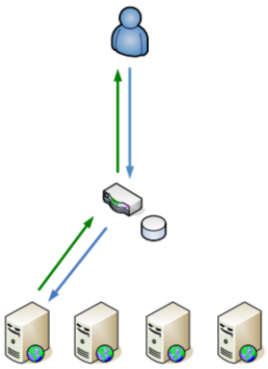
Teljesen elosztott NLB fürt



- Közös IP, MAC cím a fürtnek
- Kéréseket mindenki megkapja
- Egy csomópont válaszol
- Pl. Microsoft NLB

- Előny:
  - nincs SPOF szétosztó
- Hátrány:
  - Korlátos méret

Központi elemre épülő



- Központi elosztó (dispatcher)
- Dedikált HW-es megoldások is
- Kifinomult terhelésfigyelés és elosztás

- Előny
  - Elosztóban egyéb szolgáltatások (cache, SSL offload…)
- Hátrány
  - Elosztó SPOF lehet

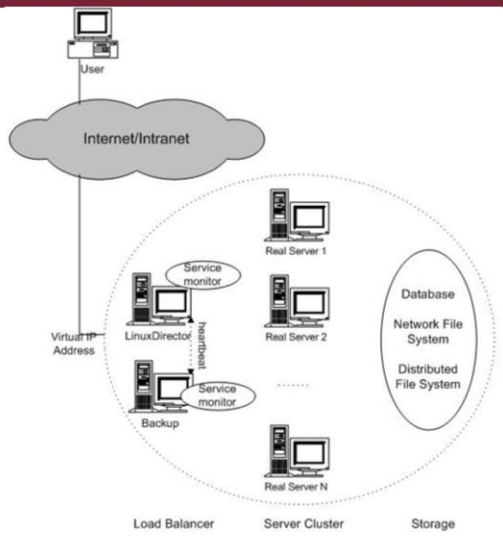
Probléma: munkamenet megőrzése

- Ügyfél munkamenete tipikusan a webszerver memóriájában tárolódik
  - De: ügyfél egymás utáni kéréseit különböző webszerver szolgálja ki
- Terheléselosztó szintű megoldás:
  - Affinitás: adott ügyfél kéréseit mindig ugyanaz a szerver szolgálja ki
- Alkalmazás szintű megoldás:
  - Munkamenet tárolása központi gépen / adatbázisban
  - Munkamenet tárolása a kliensen, elküldése minden kérésben

Példák: Hálózati terheléselosztók

- RRDNS:
  - majd minden DNS kiszolgáló (bind, MS DNS…)
- Elosztott megoldások:
  - Microsoft Network Load Balancing
- Központi elosztót használó:
  - HW (Cisco, BigIP, Juniper…)
  - Linux Virtual Server

Linux Virtual Server



High Availability of Linux Virtual Server

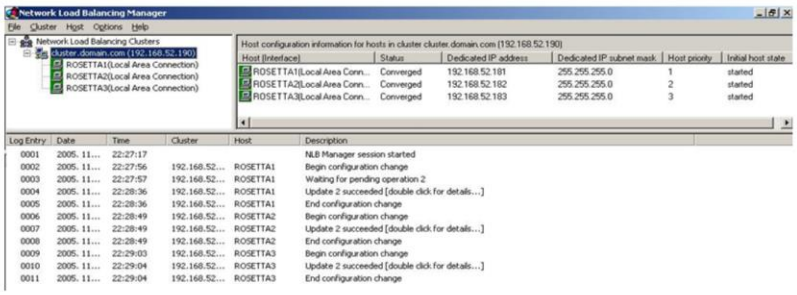
- Elterjedt (pl. sourceforge.net, linux.com…)
- Elosztó: aktív-passzív
- Layer 4 és 7 elosztás

19

Forrás: http://www.linuxvirtualserver.org/

Microsoft NLB

- maximum 32 csomópont
- kieső kiszolgálók detektálása 10 sec alatt
- Speciális szűrő hálózati meghajtó
- Portszabályok, affinitás



Érdeklődőknek bővebb leírás: Medgyesi Zoltán, Micskei Zoltán. Hálózati terheléselosztó fürtök. Mérési segédlet, http://mit.bme.hu/~micskeiz/meres/cluster_meres/files/01_load_balance_clusters_segedlet.pdf

Tartalom

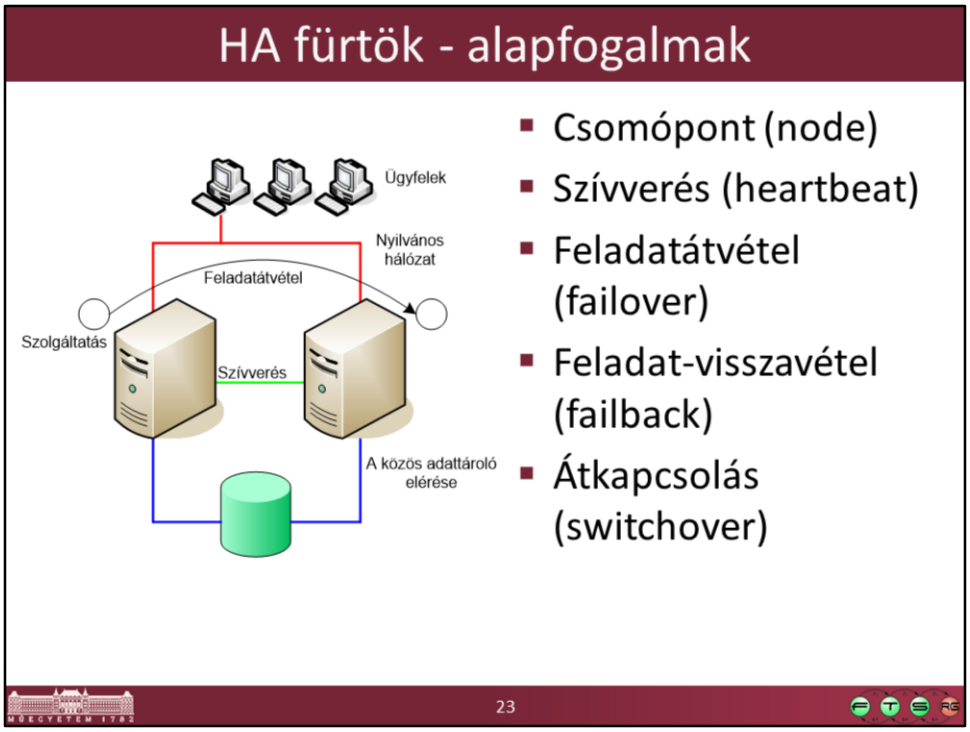
- Fürtök
  - Fürtök csoportosítása
  - Terheléselosztó fürtök
  - **Feladatátvételi fürtök**
- Replikáció
  - Elsődleges – másodlagos séma
  - Multimaster

HA fürtök csoportosítása

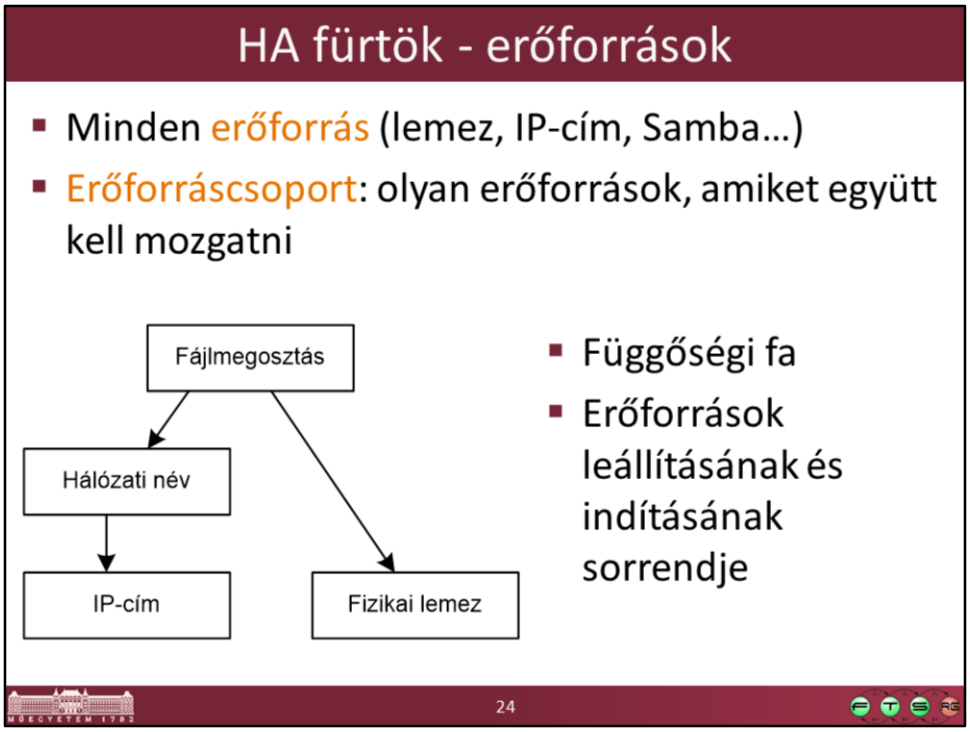
(Klasszikus csoportosítás*)

- Megosztott lemezes (shared disk)
  - Szolgáltatás több csomóponton fut(hat)
  - Közös erőforrást egyszerre írhatják-olvashatják
  - De: fizikai szinten sorosítás, zárolás használata
  - Pl.: Oracle RAC
- Megosztott elem nélküli (shared nothing)
  - Szolgáltatás egyszerre egy csomóponton fut
  - Egy erőforrást egyszerre egy csomópont birtokol
  - De: fizikai szinten lehet közös elérésű erőforrás

*M. Stonebraker, The Case for Shared Nothing, 1985, http://db.cs.berkeley.edu/papers/hpts85-nothing.pdf



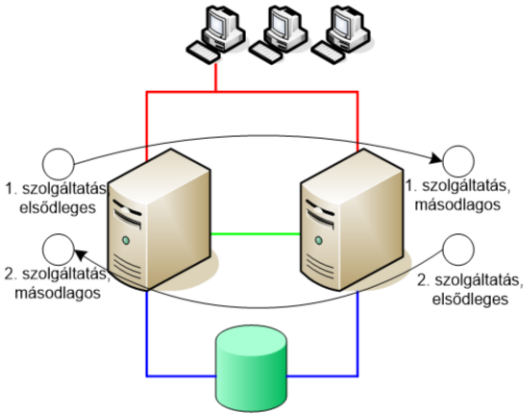
- A csomópontok egymásnak úgynevezett *szívverés* (heartbeat) üzeneteket küldenek, ezek segítségével lehet detektálni, hogy kiesett-e valaki.
- Ha a bal oldali számítógép meghibásodik, akkor a fürtszoftver érzékeli ezt, a jobb oldali számítógépen elindítja a szolgáltatást, a bal oldali gépen pedig leállítja – ezt nevezzük *feladatátvételnek* (failover). Ettől kezdve a tartalék gép használja a közös adattárolót és fogadja az ügyfelek kéréseit.
- Ha később a bal oldali gép ismét üzemképessé válik, akkor lehetőség van arra, hogy ismét ez futtassa a szolgáltatást. A feladatátvétellel ellentétes irányú műveletet *feladat-visszavételnek* (failback) nevezzük.
- A legtöbb gyártó megkülönbözteti azt az esetet, amikor a szolgáltatások áttétele hiba miatt történik, és azt, amikor a rendszergazda kezdeményezi a műveletet, például azért, hogy valamelyik fürttagot ideiglenesen, például karbantartási célból kivehesse a fürtből. Az ilyen áttételeket *átkapcsolásnak* (switchover), egyes esetekben *felügyeleti feladatátvételnek* (administrative failover), az ellenkező irányú műveletet pedig *visszakapcsolásnak* (switchback) nevezik.



A *függőségi fa* a fürtben definiált erőforrások közötti függőségeket tartja nyilván.

Feladatátvételi topológiák (1)

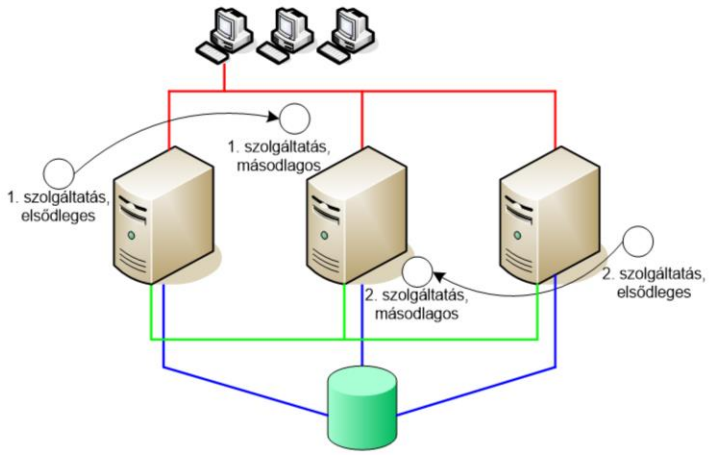
- Feladatátvételi pár (aktív-aktív)



Bár aktív-aktívnak is szokás ezt a sémát hívni, de azt fontos észrevenni, hogy egy szolgáltatás egyszerre csak egy gépen fut a másik gépen ugyanakkor egy másik szolgáltatás aktív.

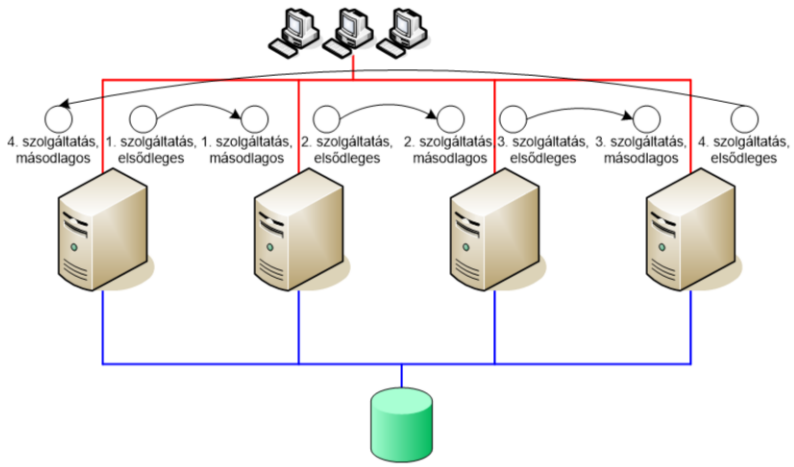
Feladatátvételi topológiák (2)

- Forró tartalék (N+1)
- Több tartalék (N+M)



Feladatátvételi topológiák (3)

- Feladatátvételi gyűrű

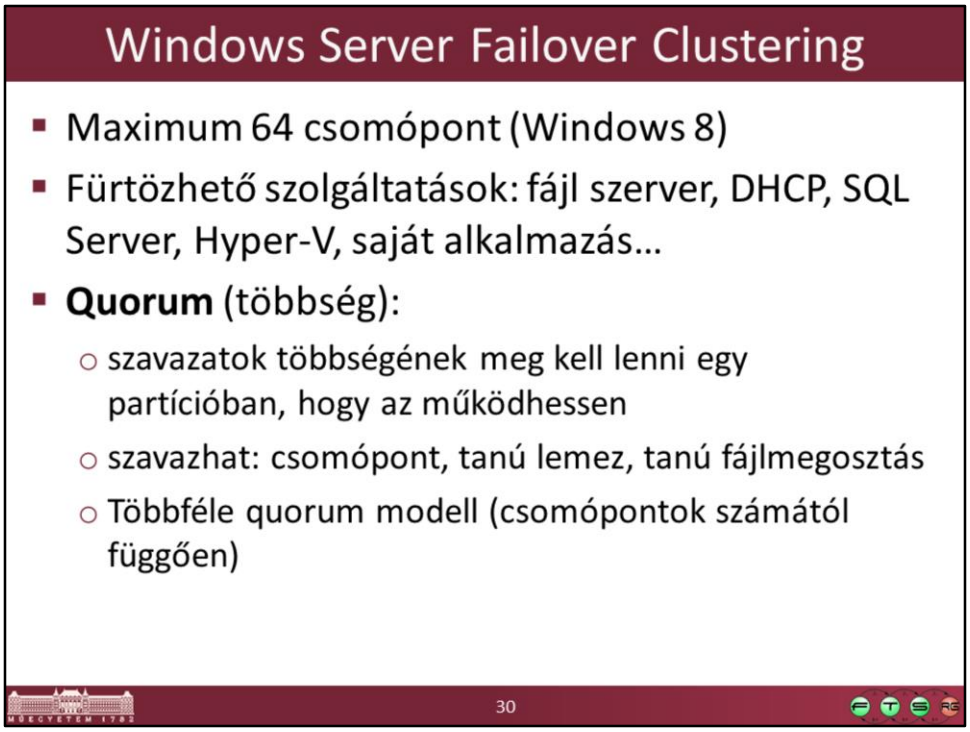


Feladatok, problémák a fürtökben

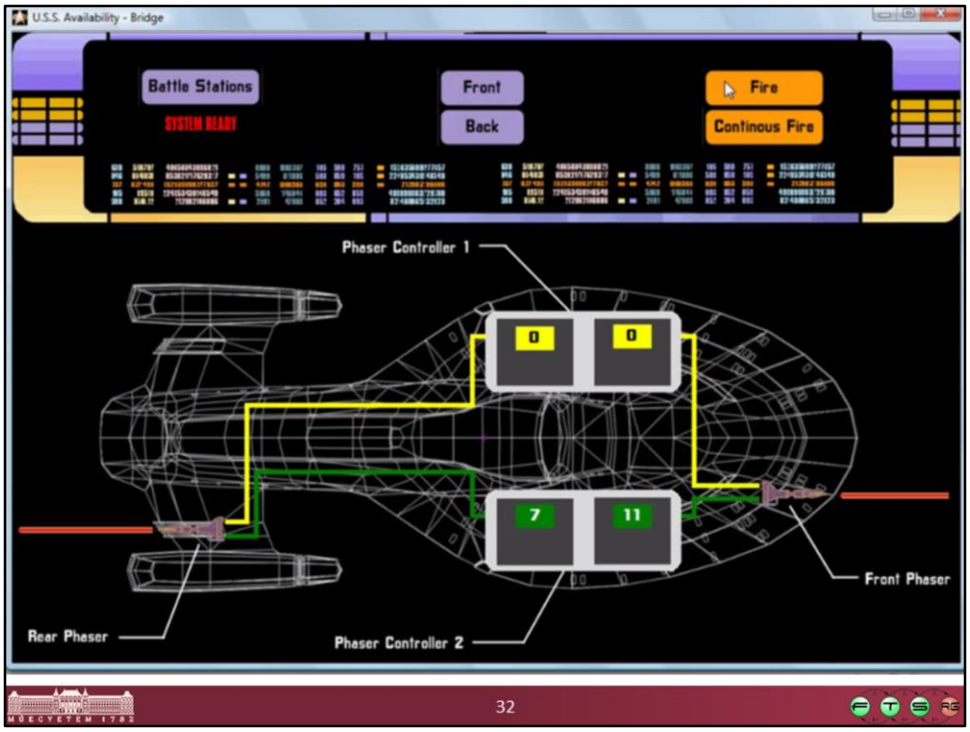
- Tagsági kép fenntartása (group membership): ki működik a csomópontok közül
- Csoportkommunikáció (group communication): üzenetek eljuttatása a többieknek hibák esetén is
- Tudathasadás (split brain): fürt több, független részre szakad
- Amnézia: kiesés után újrainduló csomópontot értesíteni a közben történt változásokról
- Gördülő frissítés (rolling upgrade): csomópontok frissítése egyesével, többi működik közben

Megoldások

- IBM High Availability Cluster Multiprocessing
- Oracle Clusterware
- Linux-HA
- SA Forum AIS
- Windows Server Failover Clustering
- VMware vSphere HA
- …



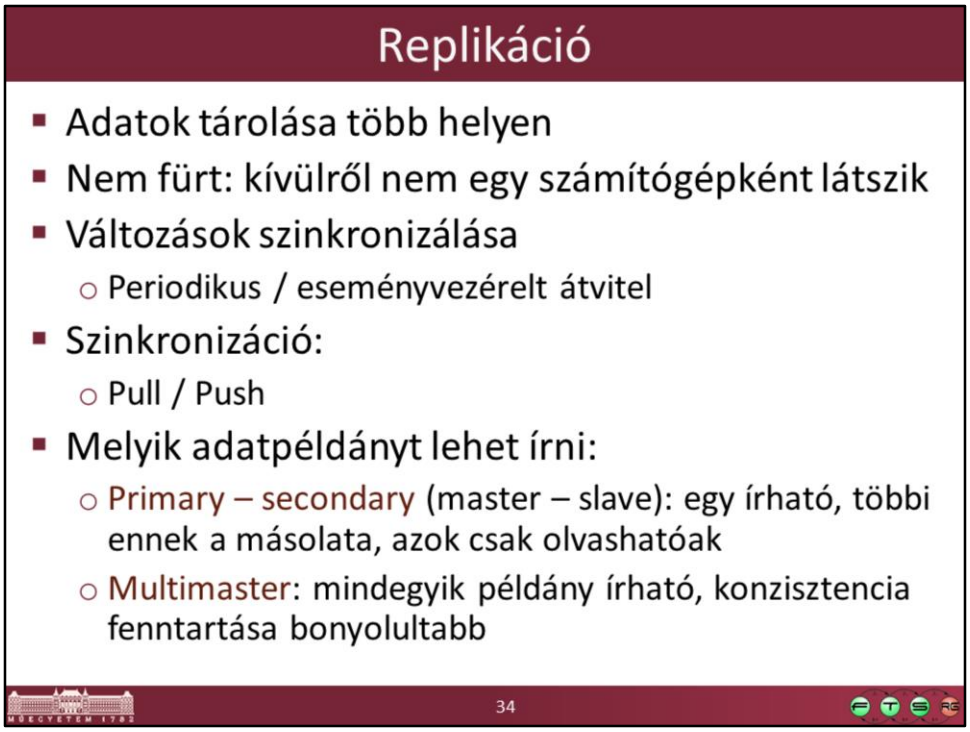
A Windows Server Failover Clusteringet például majd a *Szolgáltatásbiztos rendszertervezés szakirány* laborjában lehet kipróbálni.



Warm-up videó: (http://www.inf.mit.bme.hu/edu/specialization/presentations oldalról elérhető)
Közvetlen link:
http://www.inf.mit.bme.hu/sites/default/files/szakirany/demo/msc/uss-availability-demo-hun.avi
Nagy rendelkezésre állású fürtök kialakítás az SA Forum AIS szabványának segítségével

Tartalom

- Fürtök
  - Fürtök csoportosítása
  - Terheléselosztó fürtök
  - Feladatátvételi fürtök
- **Replikáció**
  - Elsődleges – másodlagos séma
  - Multimaster

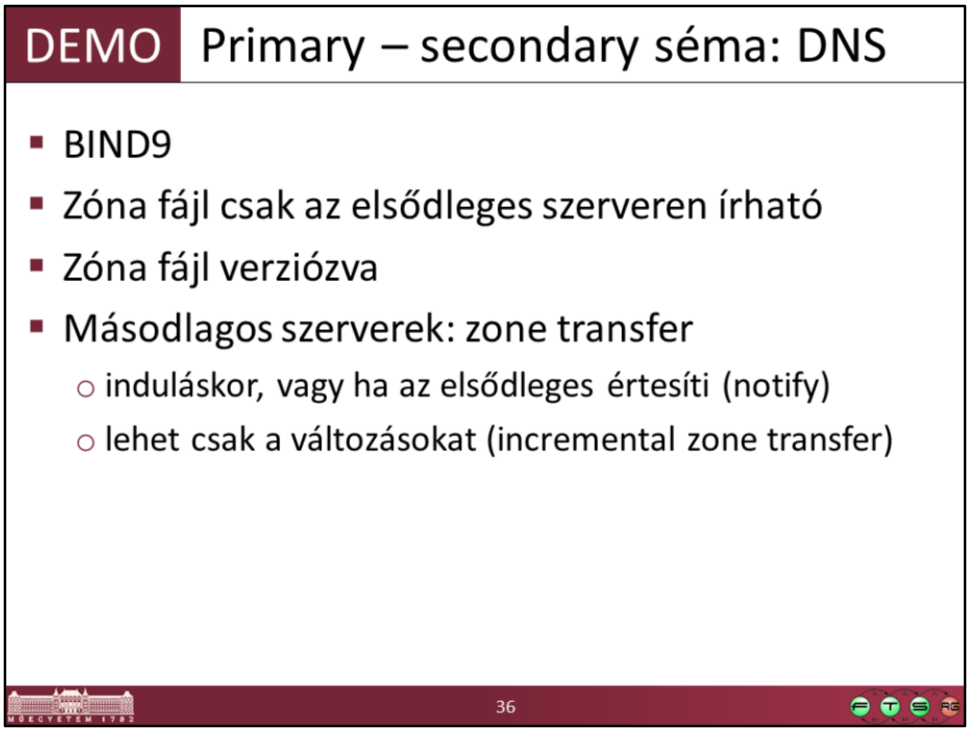


A csoportosítás természetesen megint nem fekete-fehér, a fürtök is használhatnak belül különböző replikációs technikákat.

Szinkronizáció:
-Push: akinél volt a változás, az „nyomja” át a többieknek
- Pull: a replikáció kliensei „húzzák” le a változásokat

Tartalom

- Fürtök
  - Fürtök csoportosítása
  - Terheléselosztó fürtök
  - Feladatátvételi fürtök
- Replikáció
  - **Elsődleges – másodlagos séma**
  - Multimaster



Installing A Bind9 Master/Slave DNS System,
http://www.howtoforge.com/debian_bind9_master_slave_system

BIND 9 Administrator Reference Manual, http://www.bind9.net/manuals

Konfig fájl: /etc/bind/named.conf, /etc/bind/named.con.local,
/etc/bind/named.con.options
Zóna fájlok: /var/cache/bind könyvtárban

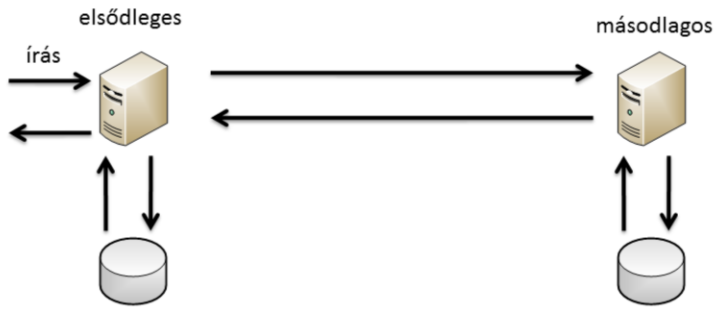
Log üzenetek: /var/log/syslog

Parancsok:
ellenőrzés: named-checkconf, named-checkzone
adminisztrálás: rndc

Zone transfer megnézése:
- masteren:
  - a zónában módosítani: serial változtatása, új A rekord felvétele
  - sudo rndc reload
  - tail /var/log/syslog
- slave:
  - cat /var/cache/bind/zona_fajl.db
  - nslookup

Primary – secondary séma: adatbázisok

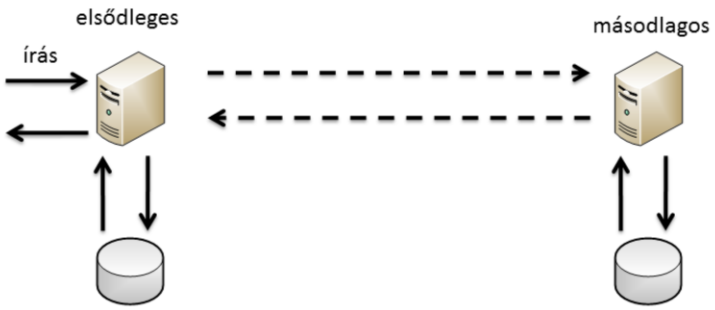
- Szinkron:



- „Zero data loss”, atomi írás
- Teljesítményveszteség az ára

Primary – secondary séma: adatbázisok

- Aszinkron:



- Helyi írás befejezése után egyből visszatér
- Mi legyen, ha a másodlagos írása közben hiba lesz?

Tartalom

- Fürtök
  - Fürtök csoportosítása
  - Terheléselosztó fürtök
  - Feladatátvételi fürtök
- Replikáció
  - Elsődleges – másodlagos séma
  - **Multimaster**

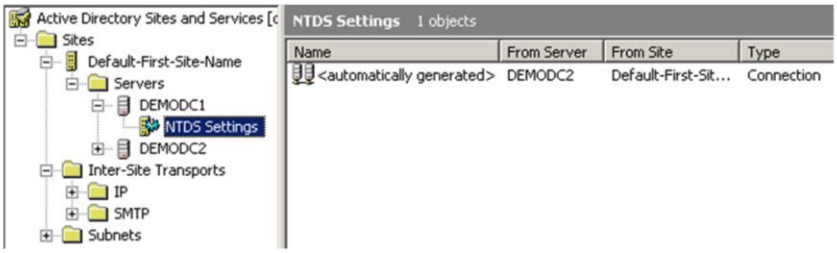
Multimaster replikáció: Active Directory

- Multimaster replikáció
  - bármelyik DC-n módosíthatunk
- Flexible Single Operations Master (FSMO)
  - 5 szerep, amiből egyszerre csak egy lehet
  - RID master, Schema master…
- Optimalizációk
  - csak a változott attribútum megy át
  - store and forward elv: változások továbbterjesztése

Active Directory Replication Technologies
http://technet2.microsoft.com/windowsserver/en/library/53998db6-a972-495e-a4e7-e3ca3f60b5841033.mspx

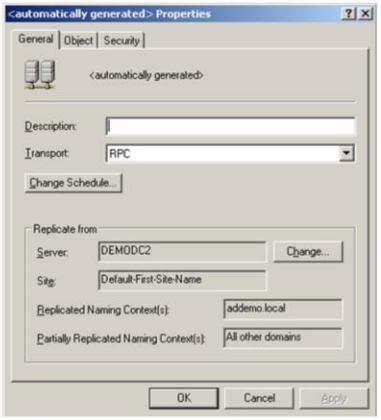
Replikációs topológia

- Telephely: gyors kapcsolattal összekötött DC-k
  - Intra-site: gyakori replikáció, RPC
  - Inter-site: ritkábban, IP/SMTP
- Knowledge Consistency Checker
  - Topológia automatikus létrehozása és frissítése



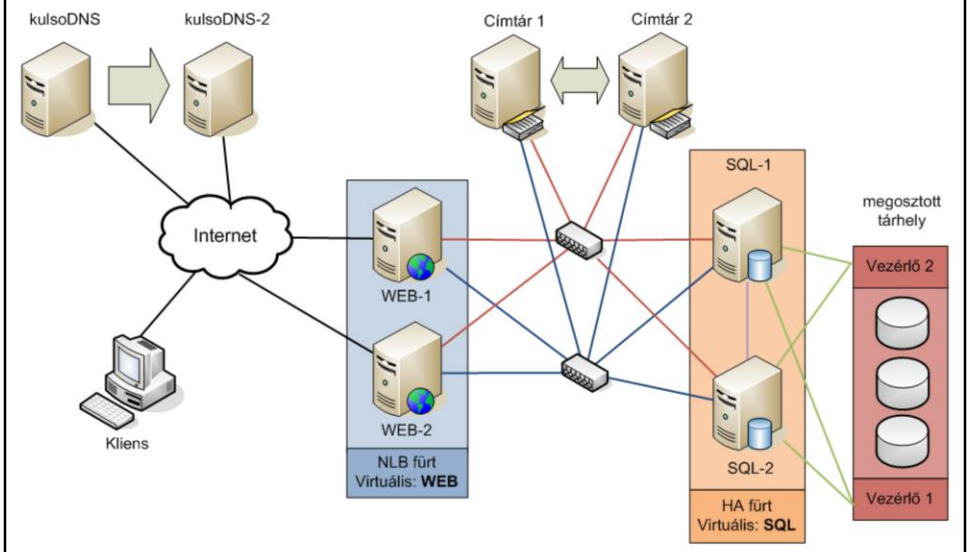
DEMO Active Directory replikáció

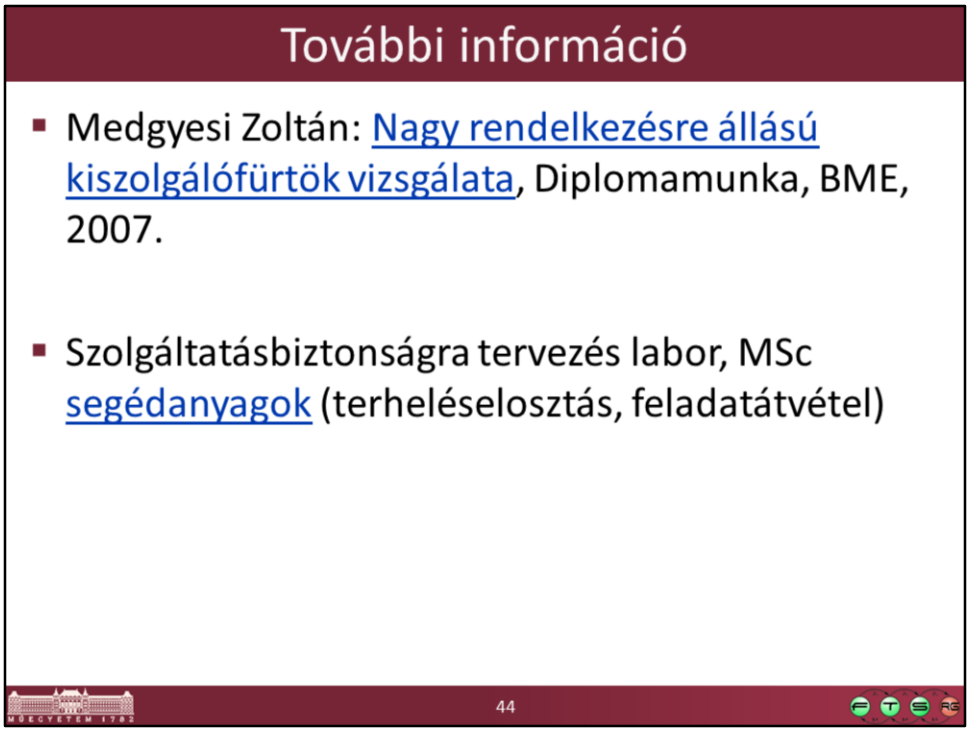
- Változás nyomon követése
- AD Sites and Services
  - Replikáció kikényszerítése
  - Telephelyek beállítása
- Ütközés feloldása
- Eseménynapló



Ütközés:
-legyen mindkét DC Global Catalog
-szakítsuk meg a kapcsolatot közöttük
-hozzuk létre ugyanolyan cn-ű objektumot
-kapcsolat vissza, replikáljunk
-korábbi objektumot átnevezi a replikációs komponens
-Eseménynapló / Directory Service

Technikák alkalmazása





- http://mit.bme.hu/~micskeiz/education/onlab/medgyesi_zoltan/medgyesi-zoltan-diploma.pdf
- http://www.inf.mit.bme.hu/edu/courses/szbtlab

Összefoglalás

- Fürtök, replikációs módszerek
- Többféle technika a számítógép és hálózati utak kiesésének kivédésére
  - Különböző előnyök és hátrányok
  - Különböző bonyolultság és költség
- DE: fürt se véd minden ellen
  - katasztrófa, adminisztrátor hibája, rongálás…
  - Kombinálni kell más módszerekkel