

Diskless klaszter a meglévő gépeinkre



Edlinger Dávid
Szalai László

Nyugat-magyarországi Egyetem
Informatika i és Gazdasági Intézet

Rólunk: NYME INGA (inf.nyme.hu)



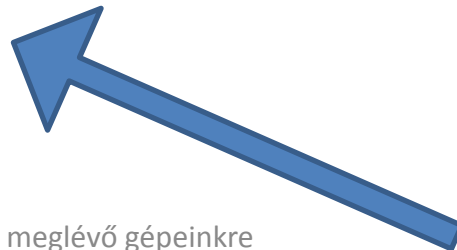
- Soproni gazdasági informatikus képzés (kb. 120 aktív hallgató)
- 10 éves múlt, BSc, MSc, műegyetemi segítség
- Saját hálózat az egyetemen belül, saját gondozásunkban
- Kb. 200 felhasználó, 60 db asztali gép, 20 db laptop
- 28db szerver (16db virtuális), 13 logikai alhálózat
- Főleg Linux szerverkörnyezet (de van Windows, Solaris is)
- Saját levelező és webes infrastruktúra, VPN, Wifi, laborok, stb.
- Virtualizáció: **Xen**, XCP, Vmware (,HyperV, KVM)
- Privát felhő (XCP alapú, failover, load balancing, snapshot backup, stb.)

Mi is az a „klaszter”?

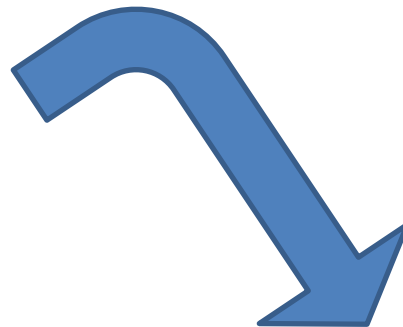
- Wikipedia:

„A számítógépfürt (angolul cluster) hasonló felépítésű és feladatokat ellátó számítógépek lazán összekapcsolt csoportja, ami sok szempontból egy rendszernek látszik.”

- HA, magas rendelkezésre állású klaszter
- „HPC”, számítási fürt



Számoló klaszterek fejlődése



NWS 2013, Diskless klaszter a meglévő gépeinkre

Node mi van ezekkel?



Kihasználatlan erőforrások...

NWS 2013, Diskless klaszter a meglévő gépeinkre



Meg velük?

Kismértékben kihasznált
számítógépek...



Gépidő kihasználtság

- Az intézetünk asztali gépeinek kihasználtsága
 - Laborgépek
 - 8-18 óráig: 25%
 - 18-8 óráig: 0 % (!)
 - Asztali oktatói gépek
 - 0-24 óráig: 5 %
 - Adminisztrációs gépek
 - 8-16 óráig: 40%
 - 16-8 óráig: 0 % (!)

Átlag: 17 %

Processzor kihasználtság

- Az intézetünk asztali gépeinek kihasználtsága
 - Laborgépek
 - 8-18 óráig: 13%
 - 18-8 óráig: 0 % (!)
 - Asztali oktatói gépek
 - 0-24 óráig: 2 %
 - Adminisztrációs gépek
 - 8-16 óráig: 40%
 - 16-8 óráig: 0 % (!)

Átlag: 12 %

Intézeti gépek darabszáma, ereje

- 3db számítógép labor
 - 3 x 13db = 39 db számítógép (2 CPU mag, 2 Gb RAM, 250 Gb HDD)
- Oktató + adminisztrációs gépek
 - 4 db számítógép (2 CPU mag, 2 Gb RAM, 250 Gb HDD)
- Összesen: 43 db gép – **86 db CPU mag**

Kihasználtság: 12 %

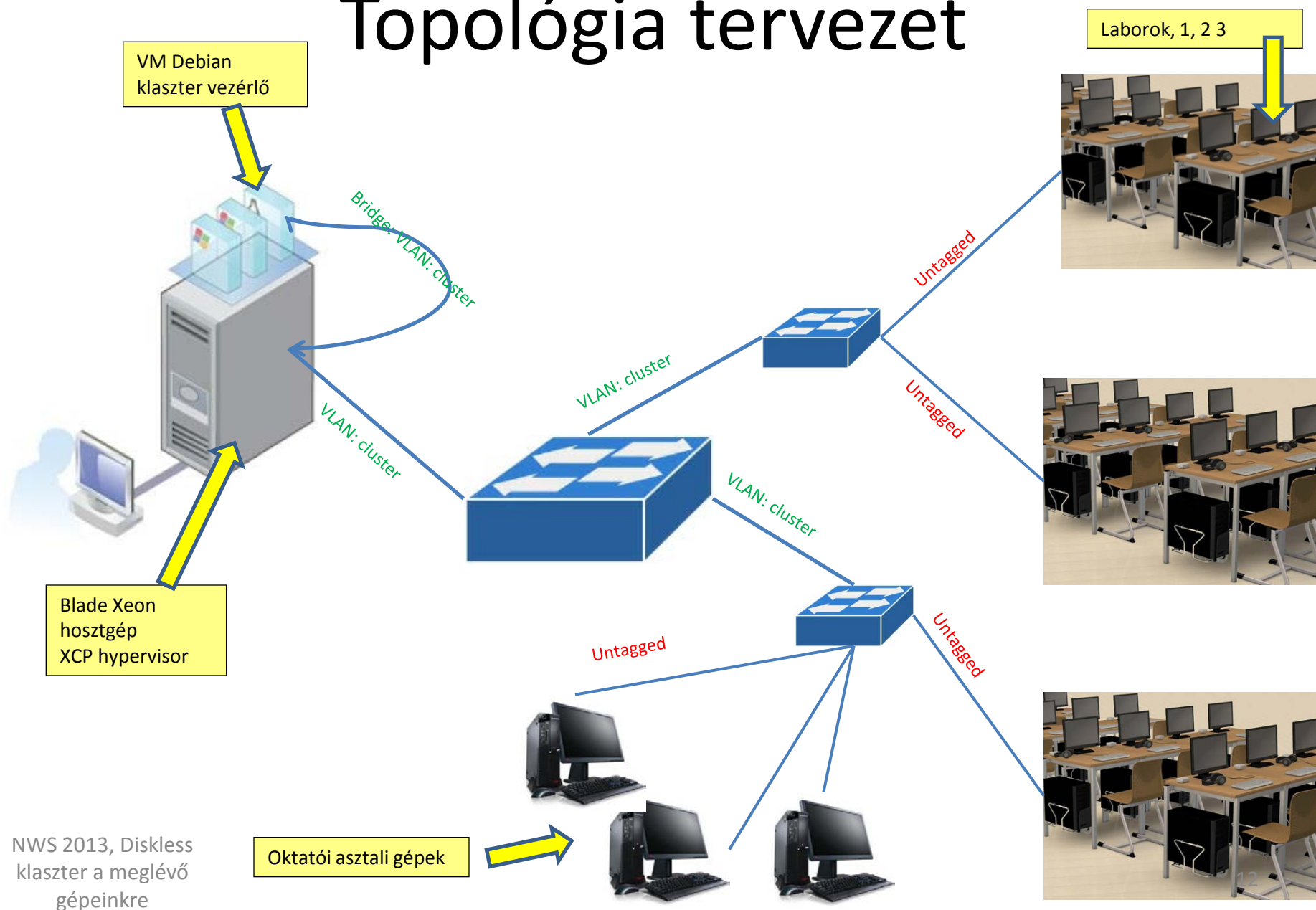
Ötlet

- Vonjuk be az erőforrásokat egy új, számoló klaszterbe
- Használjuk az eredeti rendszereket, minél kevesebb változtatással
- Telepítsünk rájuk klasztert kezelő szoftvereket
- Indítsuk be és állítsuk le őket automatikusan
- Számoltassunk velük processzorigényes (vagy IO igényes) feladatokat

Előzmények

- 6 évvel ezelőtt elkészült egy HPC klaszter
 - 22db AMD egymagos node gép + vezérlő
 - 250 Gb HDD + Gigabites háló +KVM switch a gépekhez
 - Amit feltelepítettünk:**
 - Debian Cluster Component megoldás (Linux, 2.4 -> 2.6)
 - Diszkek egy része AOE + RAID6 + LVM-be szervezve, NFS
 - MATÁV számára számolásigényes feladatmegoldás
(bash szkriptek, perl feldolgozó program)
 - Windows-os gép is (feldolgozó progi, NFS bevonása)

Topológia tervezet



Cluster controller, node

XCP VM

Debian 6.0.6, kernel 2.6.32-5-686-bigmem

Xe-daemon (Xe tools)

TORQUE Resource Manager

Apache + Munin

NFSd

NTPd

SSHd

DNSmasq

TFTPd

WOL szkriptek

XCP VM node indítási szkriptek

LABORGÉP

BIOS: PXE first



Windows7 + EXT3
partíció

Debian LIVE Netboot 6.0.6, kernel 2.6.32-5-686

TORQUE client

Munin-node

NFSd

NTPd

SSHd

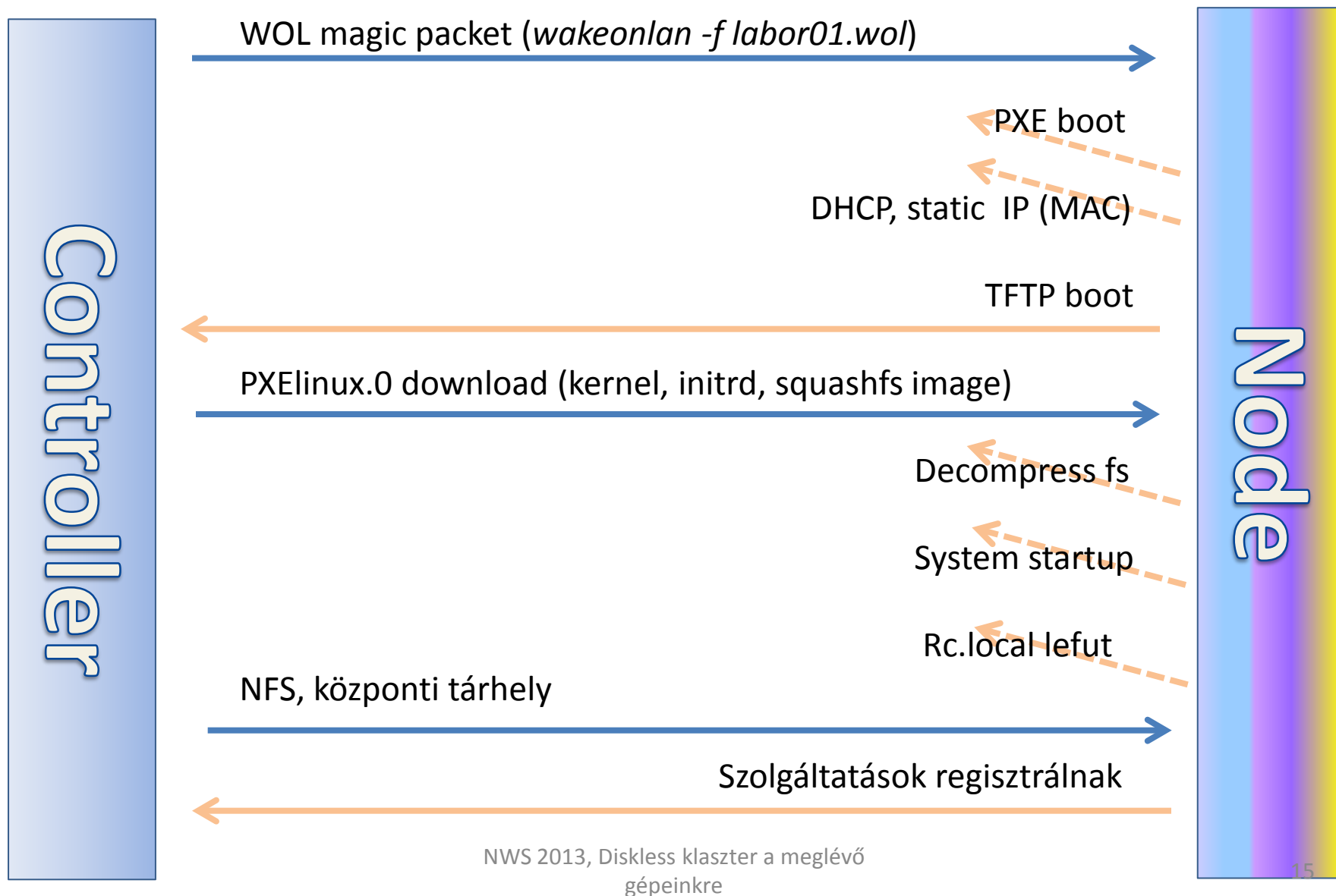
Rc.local szkriptek

- ***Indítási korrekciók***
- ***NFS terület csatolása***
- ***Lokális diszk csatolása***

Cluster node indítás

- Automatizálható legyen
- Laborban tanórák, ne zavarjon be
- Esti időszakban terheléstől függően több node
- Kezdetben GRUB2-es időzíteni ötlet, **nem muxik.**
 - `LUA, os.date ([format [, time]])`
- Aztán PXE boot, DHCP szabályokkal, TFTPd időzítéssel, **BEVÁLT!**

Cluster node indítás folyamata



Computing node

```
login as: root
root@10.16.4.14's password:
Linux node 2.6.32-5-686 #1 SMP Sun Sep 23 09:49:36 UTC 2012 i686

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Wed Mar 20 10:08:04 2013 from 10.16.4.254
root@node:~# mount
aufs on / type aufs (rw)
tmpfs on /lib/init/rw type tmpfs (rw,nosuid,mode=0755)
proc on /proc type proc (rw,noexec,nosuid,nodev)
sysfs on /sys type sysfs (rw,noexec,nosuid,nodev)
udev on /dev type tmpfs (rw,mode=0755)
tmpfs on /dev/shm type tmpfs (rw,nosuid,nodev)
devpts on /dev/pts type devpts (rw,noexec,nosuid,gid=5,mode=620)
10.16.4.50:/srv/tftp/node on /live/image type nfs (ro,relatime,vers=3,rsize=
eo=7,retrans=3,sec=sys,mountport=65535,addr=10.16.4.50)
tmpfs on /live/cow type tmpfs (rw,noatime,mode=755)
tmpfs on /live type tmpfs (rw,relatime)
tmpfs on /tmp type tmpfs (rw,nosuid,nodev)
controller.cluster:/srv/share on /tmp/share type nfs (rw,addr=10.16.4.50)
/dev/sda6 on /mnt/local type ext3 (rw)
root@node:~#
```

```
# pbstop -J
```

Controller node

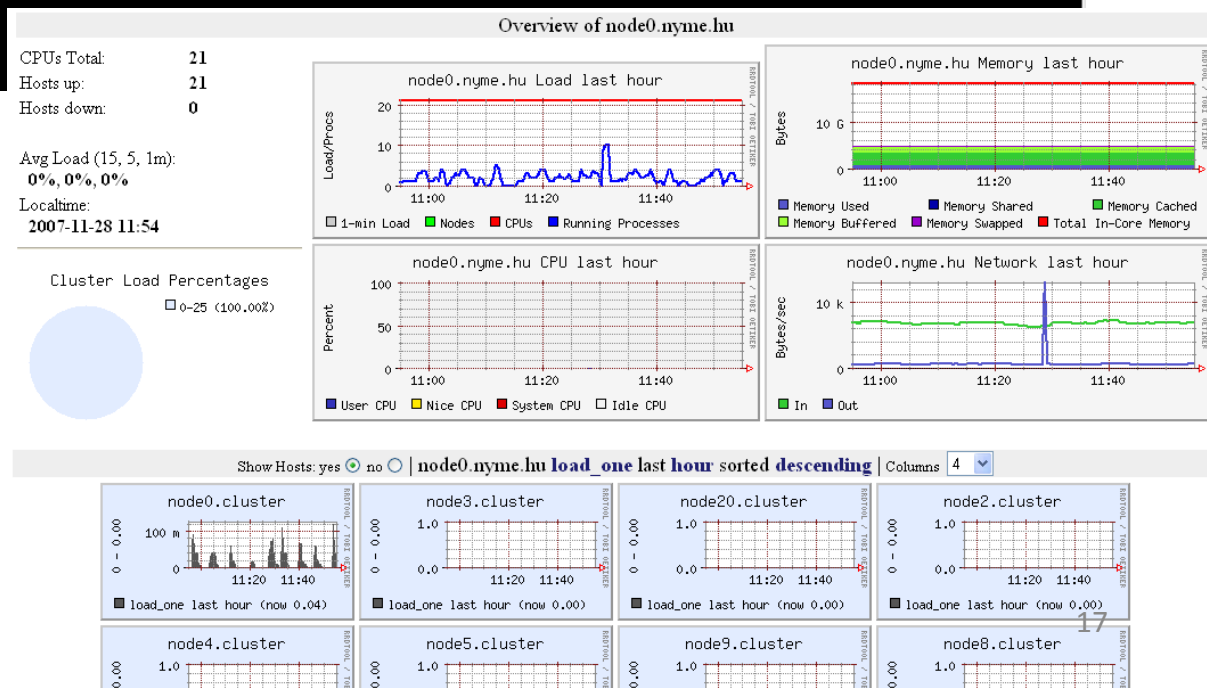
```
Usage Totals: 10/39 Procs, 10/25 Nodes, 1/1 Jobs Running 13:47:54
Node States:      22 down      3 free

CPU 0
 1 2 3 4 5 6 7 8 9 0  1 2 3 4 5 6 7 8 9 0  1 2 3 4 5 6 7 8 9 0
-----
node01 u u u . u u u u u u  * * * . * * * * * * *  * * * * .
-----

Job#   Username   Queue   Jobname   Nodes   S   Elapsed/Requested
u = 4   user         batch   torque-nod   10    R   --/ 01:00

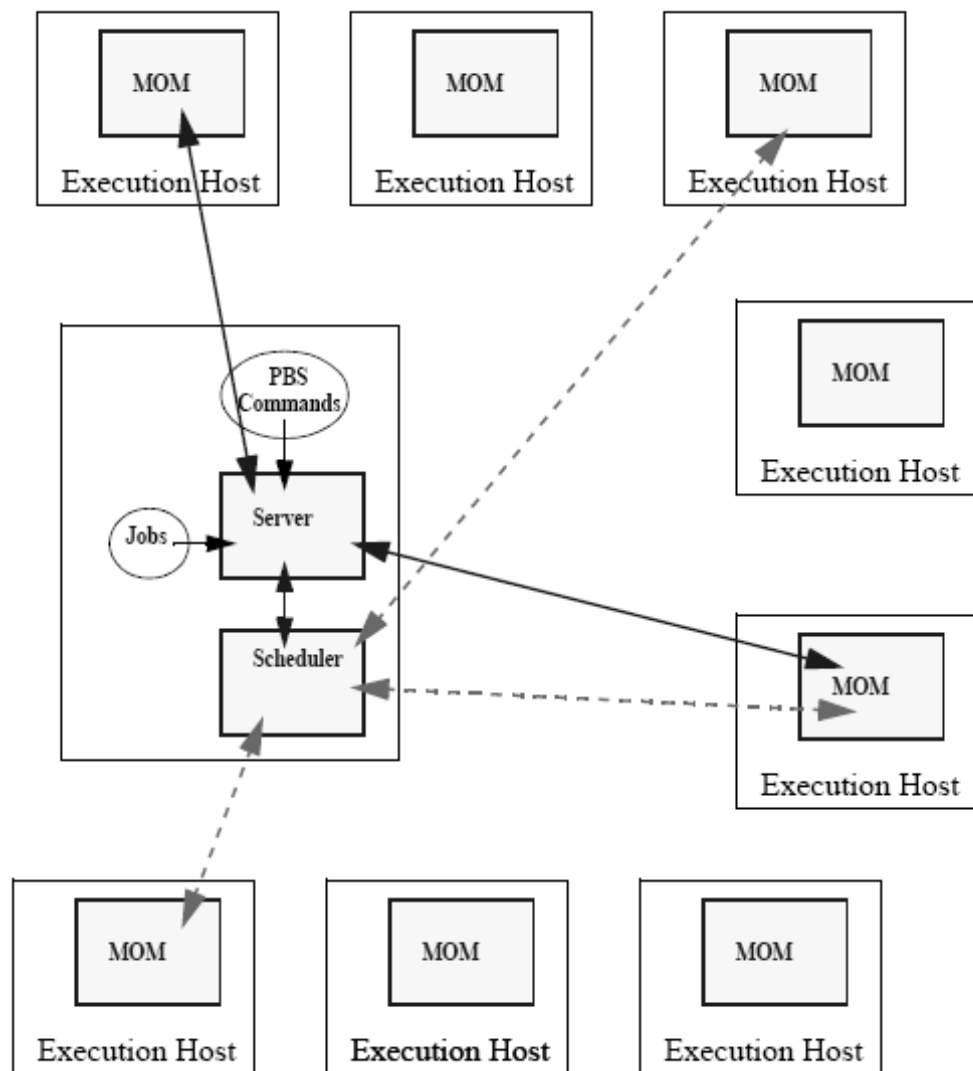
[?] unknown  [@] busy   [*] down  [.] idle  [%] offline  [!] other
```

Munin képernyő



NWS 2013, Diskless klaszter a meglévő gépeinkre

Cluster elvi működése



- Mindennek az alapja a Torque manager
- User-ek SSH belépés után parancsokat adnak ki a *Server* alkalmazásnak
- A *Scheduler* alkalmazás besorolja a parancsokat *JOB* sorba, majd kiosztja a szabad számoló node-ok felé
- A lefutott *JOB*-ok kikerülnek a QUEUE-ból

Diszk IO megfontolások

- Node-onként
 - Közös NFS tárhely (Gigabit ethernet)
 - Ide kerülnek a futtatandó szkriptek
 - Kicsi IO igény esetén a „nyersanyag” adatok helye, olvasható minden node által
 - Lokális tárhely
 - Temporális nagysebességű diszkművelethez
 - Nagy IO esetén a „nyersanyag” adatok egyszeri átmozgatása

Parancs (Job) futtatási példák

- Egyszerű job indítás user által:

@ qsub /mnt/nfskozos/testjob.sh

- A qsub program gazdagon paraméterezhető, egy újabb példa bemutatja azt, hogy a job kétprocesszoros gépeknek (2 db) lesz kiosztva:

@ qsub -l nodes=2:ppn=2 /mnt/nfskozos/testjob.sh

- Lehetőség van interaktív job-futtatásra is:

@ qsub -I

Párhuzamosítás a cluster-en

- A feladat megoldása **TERVEZENDŐ!**
- Az egyes node-ok között nincs kommunikáció
- Definiálhatók (ha nehezen is) függőségek a node-ok közötti számolásnál
- Nagy adattömeg feldolgozása esetén előzetes előkészítés, akár lokális diszkre másolás
- Tipikusan „darabolhatóan megoldható” jellegű feladatoknál használható

Finomságok, Live image creation

```
root@controller:/srv/tftp/build# cat create.sh
lb clean --binary
lb config --apt apt \
--apt-recommends true \
--architecture i386 \
--binary-image net \
--bootstrap debootstrap \
--bootstrap-flavour standard \
--bootappend-live noautologin \
--cache false \
--cache-packages false \
--chroot-filesystem squashfs \
--clean \
--distribution squeeze \
--hostname node \
--interactive shell \
--linux-flavours 686 \
--linux-packages "linux-image-2.6" \
--memtest none \
--mirror-binary "http://localhost:3142/ftp.debian.org/debian/" \
--mode debian \
--net-root-filesystem nfs \
--net-root-path /srv/tftp/node \
--net-root-server 10.16.4.50 \
--packages "bzip2 gcc gpp locales make munin-node munin-plugins-extra nfs-common ntp openssh-server perl tar torque-mom" \
--virtual-root-size 1024 \
--username user \
--win32-loader false

lb build
```


Finomságok, DHCPd

```
class "PXE" {  
    match if substring(option vendor-class-identifier, 0, 3) = "PXE";  
    next-server 10.16.4.50;  
    filename "pxelinux.0";  
}
```

```
host cluster1 {  
    hardware ethernet 00:19:D1:93:FA:3B;  
    fixed-address 10.16.4.1;  
}
```

Finomságok, DNSmasq

```
root@controller:/# cat /etc/resolv.conf
```

```
domain cluster
```

```
search cluster
```

```
nameserver 127.0.0.1
```

```
nameserver 193.2XX.9X.XX
```

```
root@controller:/# cat /etc/hosts
```

```
127.0.0.1 localhost controller.cluster controller
```

```
10.16.4.50 controller.cluster controller
```

```
10.16.4.1 node01.cluster node01
```

```
10.16.4.2 node02.cluster node02
```

```
...
```

Finomságok, Node startup config

```
root@controller:/srv/share# cat startup.sh
#!/bin/bash

cat > /etc/resolv.conf << EOF
domain cluster
search cluster
nameserver 10.16.4.50
EOF

mkdir -p /mnt/local
mount /dev/sda6 /mnt/local

echo "server 10.16.4.50" >> /etc/ntp.conf
sleep 3
/etc/init.d/ntp restart

echo "allow ^10\.16\.4\.50$" >> /etc/munin/munin-node.conf
sleep 1
/etc/init.d/munin-node restart

touch /var/spool/torque/mom_priv/config
cat > /var/spool/torque/mom_priv/config << EOF
\${logevent} 255
\${log_file_max_size} = 100
\${pbsserver} 10.16.4.50 #controller.cluster
#\${remote_reconfig} true
EOF

/etc/init.d/networking restart
sleep 3

echo "10.16.4.50" > /etc/torque/server_name
sleep 5
/etc/init.d/torque-mom restart
```

Gyakorlatban mi hogyan csináljuk?

- A laborok holtidejében egy szkript bekapcsolja a TFTPd progit, innentől a boot erről (nem Windows)
- Újabb feladatnál irányítottan WOL felébresztődnek a gépek a megfelelő számban
- Ha lefutott az összes JOB, SSH távkikapcs (C3 Suite hiánya, saját szkriptek)
- **Szükség esetén újabb VM node-ok bevonása XCP hoszt alól, controller-ről indítható indító szkripttel!**

Extra használat, extra lehetőségek

- Eddig számoló klaszterről volt szó, holtidőben!

Node! Jó a fantáziánk!

- A futó Windows-ok alatt IS lehet node gép!!
 - Virtualizációs szoftverek használata (Vmware, Virtualbox, stb.)
- Reverse proxy mögötti webkiszolgáló node-ok!
 - Apache + PHP becsomagolva, NFS alatt a weboldalak, stb.

(FIGYELEM: BIZTONSÁGI AGGÁLYOK !!)

Összefoglalás

- Asztali gépes környezetek is használhatók intenzív számoló feladatok elvégzésére, ha ilyenre van igényünk!
- Linux segítségével ingyenesen kihasználhatjuk ennek az előnyét, az eredeti infrastruktúrán, nagymértékű hardveres beavatkozás nélkül...

Köszönöm a figyelmet!

Email: szalai@inf.nyme.hu