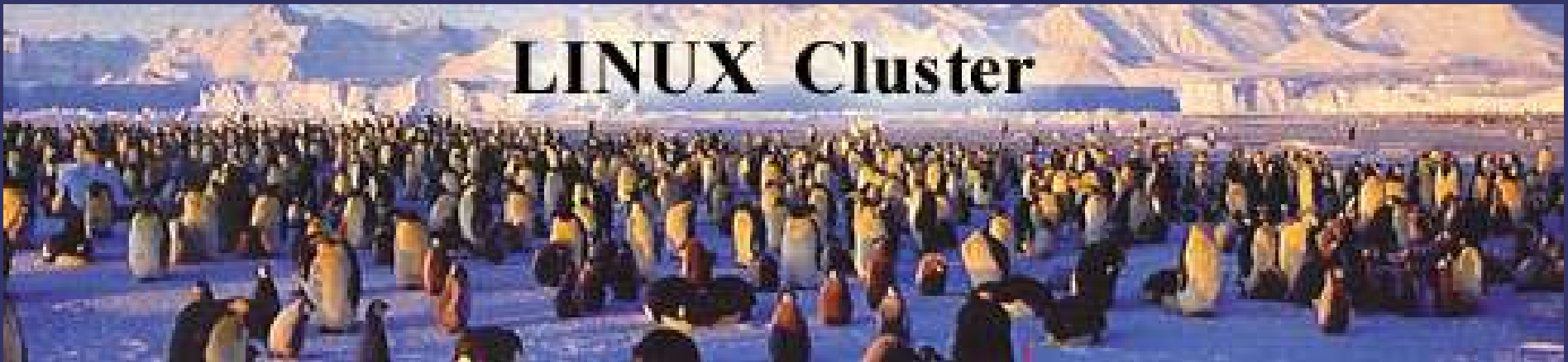


Live free() or die() Az openMosix cluster szoftver



Erdei Csaba
FSF.hu

Fürtök típusai

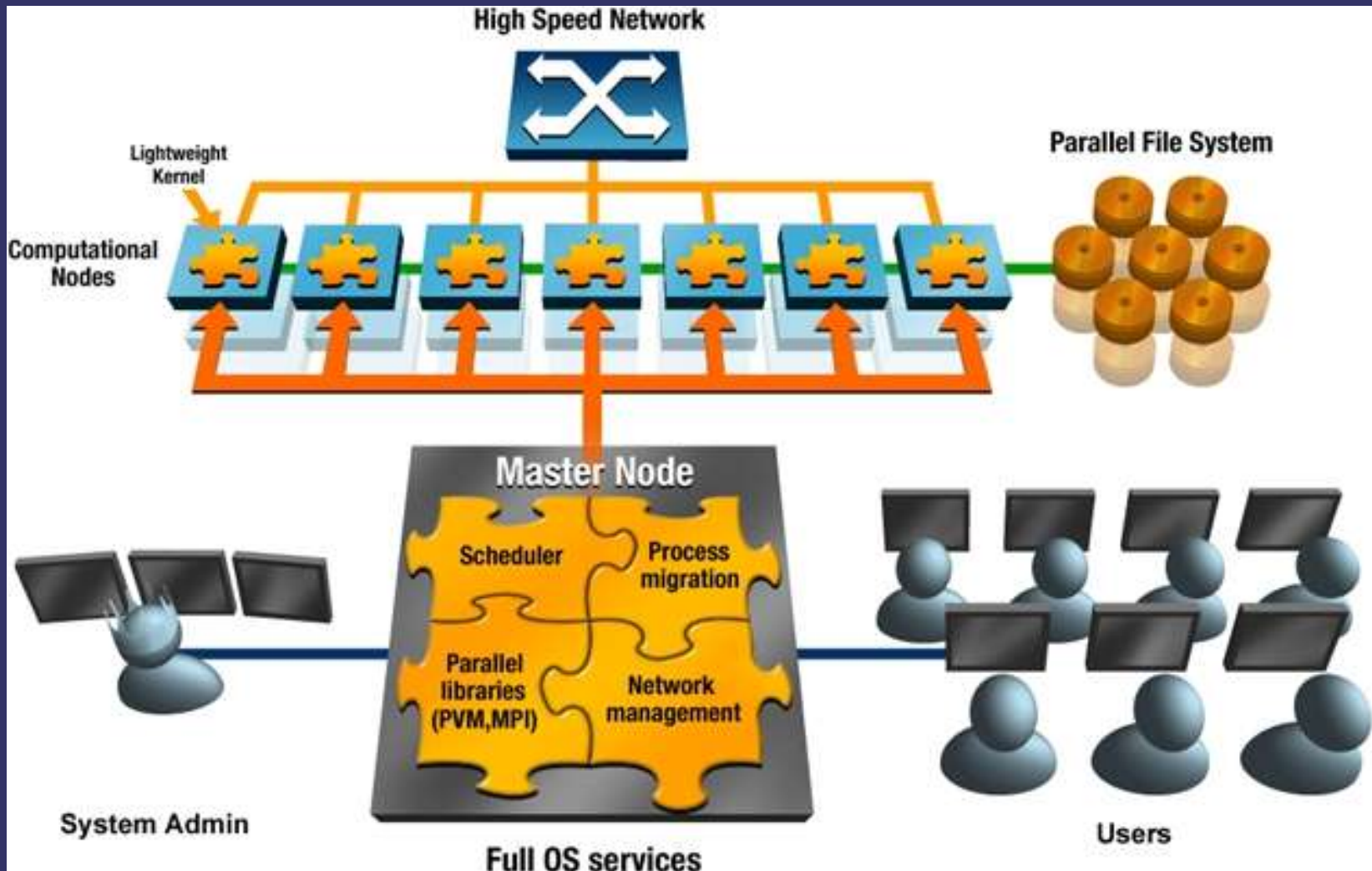
- Magas rendelkezésre állást biztosító
(HA - high availability)
- Terheléselosztó (Load balancing)
- Tudományos-technikai célú
(HPTC - High Performance Technical Computing)
- SSI – Single System Image
- GRID

Tudományos-technikai célú fürtök

- Nagy számításigényű feladatok
- Általában hierarchikus
- Párhuzamosítható feladatok és programozás (MPI, PVM)
- Ütemezés
- Központi fájlserver
- Nem HA
- Tudomány és ipar
- Géntérkép, kémia, fizika ...



Tudományos-technikai célú fürtök sémája



SSI fürtök

- Single System Image
- Egy rendszernek látszó telep
- Minden egyben
 - hibatűrés
 - terheléselosztás
 - erőforrás-megosztás



SSI cluster szoftverek

- Mosix
- openMosix (GPL)
- clusterKNOPPIX(GPL)
- openSSI (GPL)



Mi az openMosix

Linux rendszermag kiterjesztés és felhasználói programok, amely segítségével egyedi gépeinkből SSI-fürtöt tudunk létrehozni.

A számítógépes rendszerünk „egy gépnek látszik”. Közös erőforrásokkal automatikus terheléselosztás valósul meg. A tagok bármikor beléphetnek a rendszerbe, illetve kiléphetnek onnan.



Az openMosix története

- A 80-as években született PDP-11/70-en. Egy teljes és egy lemez nélküli (diskless) PDP, innen jött a processz migrálás ötlete
- 1997: áttérés GNU/Linux-ra
- 1999: bináris verzió
- 1998-2001: Hebrew University - Prof. Barak és Dr. Moshe Bar



Az openMosix története

- 2001-ben Mosix / openMosix szétválás licenc problémák miatt
- 2002 júliusára a Mosix installációk 97%-a openMosix-ra váltott
- 2002 augusztusa óta a legaktívabb Linux-telep projekt



Az openMosix működése

- Memória elfogyásának megelőzése
- Erőforrás-elosztása
 - közzgazdasági kutatás
 - processzor, memória, ... alapján költségszámolás
 - mindig a legkisebb költséggel fut a feladat



Az openMosix működése

- Gép-párok közötti feladat elosztás (akár a hőterjedés)
- Nem központosított
- Folyamatos információgyűjtés, a teleptagok kapcsolatban vannak egymással (szívdobbanás-szerű)



Az openMosix működése

- Feladat migrálás
 - UHN (Unique Home Node)
 - Deputy: a feladat helyi része
 - Remote: a feladat távoli része
 - Rendszerhívások kezelése!!!
 - Fájlműveletek
 - Socketek
 - Egyéb műveletek (shared memory, thread-ek)



Az openMosix működése

- oMFS/DFSA: a tagok elérnek más tagok fájlrendszerét. Lehet, hogy a program megy az adathoz.
- DSM (Distributed Shared Memory) – Migshm patch – MAASK csoport (5 kernelhacker lány :-)
- CHPOX – a feladatok felfüggesztéséhez és újraindításához (checkpointing)

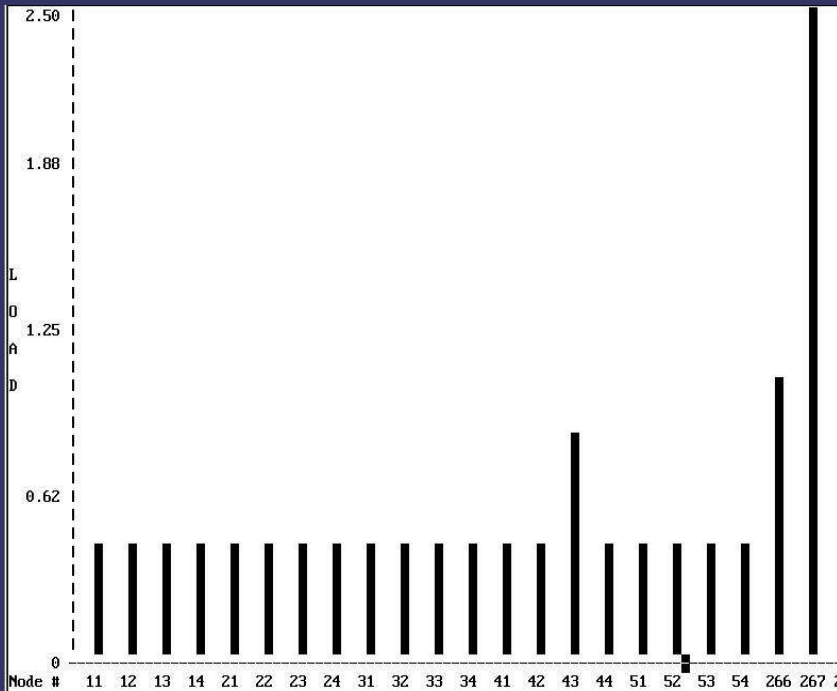


Az openMosix előnyök - hátrányok

- Nem kell a programokat módosítani (PVM, MPI nem kell)
- Könnyű telepítés és konfigurálás, autodiscovery
- Dinamikus
- Gyorsan fejlődik, az igényeknek megfelelően
- Kernelfüggetlen
- Egy processz esetében nem gyorsul a munka (természetesen)



Userspace-tools



mosmon

mtop

```

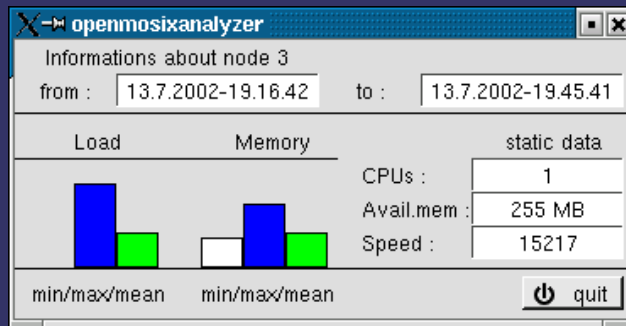
11:33am up 110 days, 5:52, 4 users, load average: 2.94, 2.95, 2.91
103 processes: 98 sleeping, 5 running, 0 zombie, 0 stopped
CPU states: 1248.0% user, 35.3% system, 4.6% nice, 0.0% idle
Mem: 4004308K av, 2507540K used, 1496768K free, OK shrd, 81312K buff
Swap: 4191488K av, 5948K used, 4185540K free 2052344K cached
  
```

PID	USER	PRI	NI	SIZE	RSS	SHARE	STAT	N#	%CPU	%MEM	TIME	COMMAND
982	zzzzzzzz	15	0	149M	149M	52	R	0	99.9	3.8	2077h	magma.exe.mall
9149	zzzzzzzz	18	0	9604	1624	516	R	0	99.9	0.0	15531m	zzzzzzzz
18106	zzzzzzzz	16	0	42748	41M	24	S	4	99.8	1.0	224:27	magma.exe
26534	zzzzzzzz	16	0	4120	4120	8	S	5	98.9	0.1	9463m	magma.exe
30568	zzzzzzzz	17	0	559M	559M	452	S	4	98.9	14.3	6191m	magma.exe
26477	zzzzzzzz	17	0	7336	7336	1060	S	2	98.5	0.1	9509m	magma.exe
30567	zzzzzzzz	15	0	523M	523M	2432	S	4	98.5	13.3	6164m	magma.exe
13684	zzzzzzzz	16	0	122M	122M	56	S	5	98.3	3.1	2365h	magma.exe.mall
17989	zzzzzzzz	16	0	4624	4624	1296	S	5	98.3	0.1	275:02	magma.exe
22775	zzzzzzzz	18	0	4324	4324	140	S	3	98.1	0.1	20874m	magma.exe
22764	zzzzzzzz	14	0	3760	3760	140	S	3	98.0	0.0	20874m	magma.exe
23671	zzzzzzzz	11	0	712M	712M	2416	S	2	96.3	18.2	10592m	magma.exe
23672	zzzzzzzz	15	0	709M	709M	2416	S	2	80.7	18.1	10598m	magma.exe
18214	zzzzzzzz	9	0	408M	408M	1276	S	4	5.0	10.4	31:34	magma.exe
27215	zzzzzzzz	19	18	33168	32M	1280	S N	4	4.3	0.8	1920m	magma.exe
27329	zzzzzzzz	9	0	39580	38M	1192	S	3	3.7	0.9	1732m	magma.exe
18007	zzzzzzzz	9	0	29116	28M	1264	S	5	3.6	0.7	150:34	magma.exe



Userspace-tools

openmosixanalyzer



openmosixview

openMosixview Advanced Execution

/usr/bin/mybigjob
 (you can now specify additional command-line arguments)

- no migration
- run home
- run on
- cpu job
- io job
- no decay
- slow decay
- fast decay
- parallel

host-chooser: run job on cluster-node [2]

execute close

openmosixproc

openMosixprocs-Migrator

- 192.168.88.2
- 192.168.88.3
- 192.168.88.4
- 192.168.88.5
- 192.168.88.6

Name: x-pympov
 State: S (sleeping)
 Tgid: 1806
 Pid: 1806
 PPid: 1799
 TracerPid: 0
 Uid: 0
 0
 0
 0

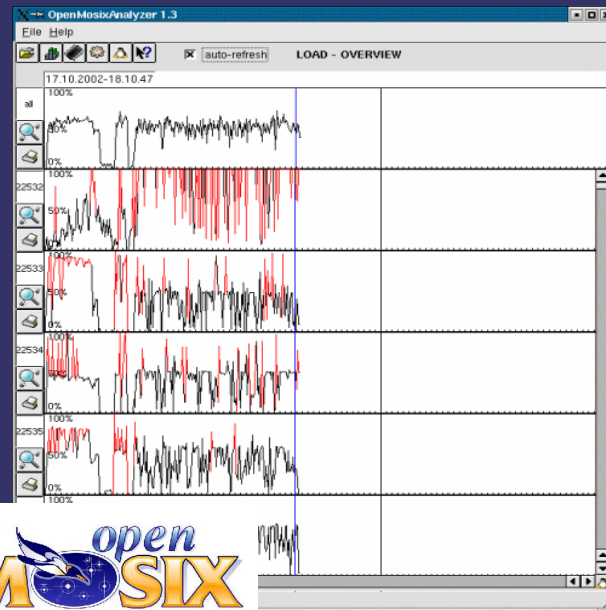
doubleclick a node for migrating PID:1806 running on node 3

send to home! node
 send to best node
 or kill proc.

SIGSTOP SIGCONT

renice process
 -20 0 20
 fast slow

close



openMosix-configuration

node : 192.168.88.3

on off auto-migration on/off
 yes no talk to others nodes
 yes no local procs stay
 yes no send away guest procs
 start stop start/stop

apply cancel

console remote proc-box

-display : 0 : 0

clear clear history close

openmosixprocs

from openMosix-node 22532 with IP-adress 192.168.88.4

remote identity statistics

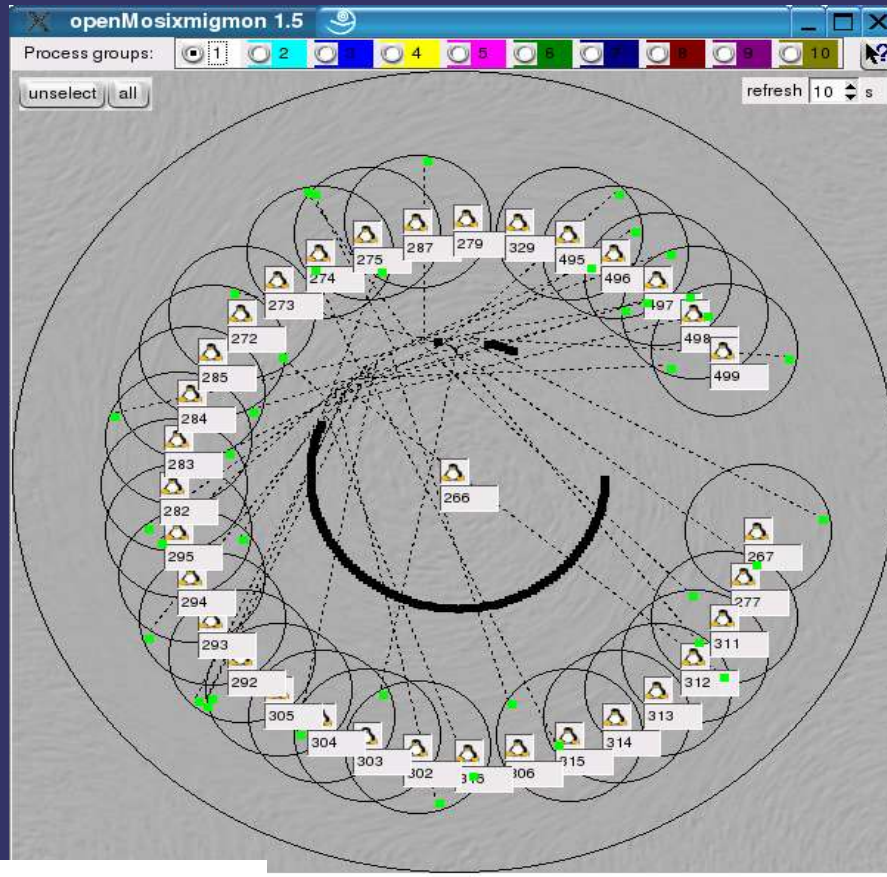
goto home node
 goto best node

pid=12795 utime=254
 tgid=520 cutime=0
 uid=0 nice=0
 gid=0 state=R
 pgrp=1 vsz=2613248
 session=1 rss=114
 nmigs=1 nswap=0
 cnszap=0

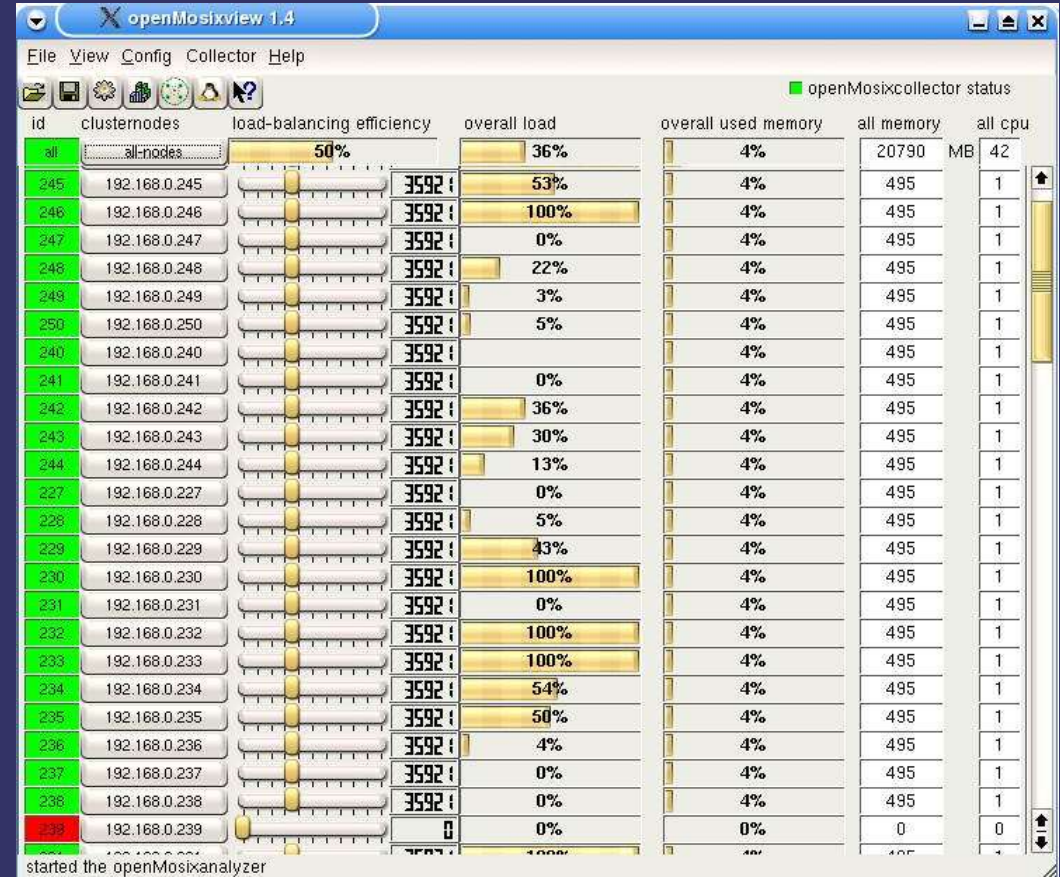
3 remote processes running on this node OK

Userspace-tools

openMosixmigmon



openMosixview



The screenshot shows the openMosixview 1.4 interface. It displays a table of system metrics for various nodes. The table has the following columns: id, clustermodes, load-balancing efficiency, overall load, overall used memory, all memory, and all cpu. The data is as follows:

id	clustermodes	load-balancing efficiency	overall load	overall used memory	all memory	all cpu
all	all-nodes	50%	36%	4%	20790 MB	42
245	192.168.0.245	3592	53%	4%	495	1
246	192.168.0.246	3592	100%	4%	495	1
247	192.168.0.247	3592	0%	4%	495	1
248	192.168.0.248	3592	22%	4%	495	1
249	192.168.0.249	3592	3%	4%	495	1
250	192.168.0.250	3592	5%	4%	495	1
240	192.168.0.240	3592	0%	4%	495	1
241	192.168.0.241	3592	0%	4%	495	1
242	192.168.0.242	3592	36%	4%	495	1
243	192.168.0.243	3592	30%	4%	495	1
244	192.168.0.244	3592	13%	4%	495	1
227	192.168.0.227	3592	0%	4%	495	1
228	192.168.0.228	3592	5%	4%	495	1
229	192.168.0.229	3592	43%	4%	495	1
230	192.168.0.230	3592	100%	4%	495	1
231	192.168.0.231	3592	0%	4%	495	1
232	192.168.0.232	3592	100%	4%	495	1
233	192.168.0.233	3592	100%	4%	495	1
234	192.168.0.234	3592	54%	4%	495	1
235	192.168.0.235	3592	50%	4%	495	1
236	192.168.0.236	3592	4%	4%	495	1
237	192.168.0.237	3592	0%	4%	495	1
238	192.168.0.238	3592	0%	4%	495	1
239	192.168.0.239	3592	0%	0%	0	0

started the openMosixanalyzer

Felhasznált irodalom, képek, lapok

- openmosix.sourceforge.net
- www.mclx.hu - Bodnár Csaba előadásai
- www.pingvintelep.hu
- <http://www.redhat.com/software/rha/cluster/manager/>
- www.freshmeat.net
- www.sourceforge.net
- www.top500.org



Köszönöm a figyelmet!

?

?

?

?

?

?

?

?

KÉRDÉSEK ?

?

?

?

?

?

?