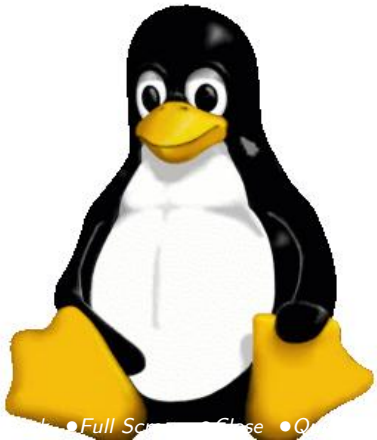
openMosix

Live free() or die()
A short intro to HPC

Kris Buytaert
buytaert@stone-it.be

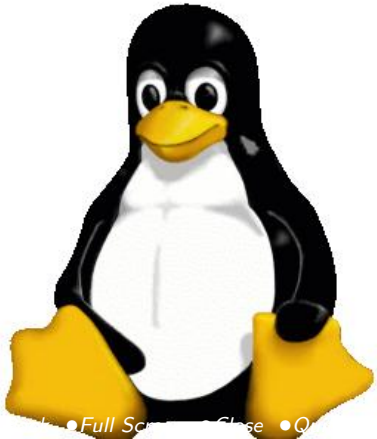
L2U , Leuven , March 2003



Welcome

agenda

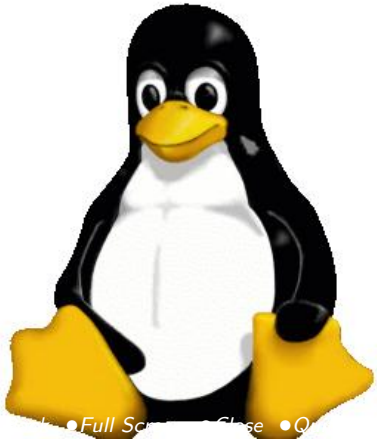
1. Clustering
2. High Performance Computing
3. openMosix
4. RealLife
5. Questions ?



1. Clustering

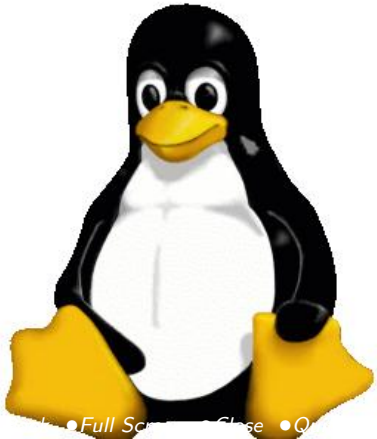
1.1. Different types of clustering

- Failover (Heartbeat, ...)
- Loadbalancing (LVS, ...)
- High Performance Computing



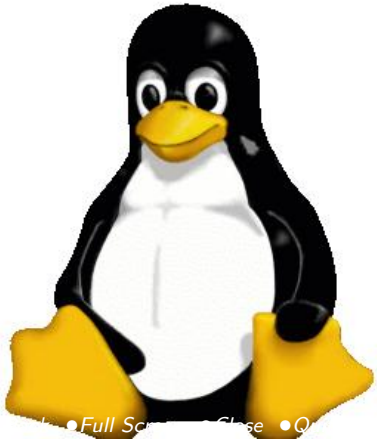
1.2. Failover

Fail-over Clusters consist of 2 or more network connected computers with a separate heartbeat connection between the 2 hosts. The Heartbeat connection between the 2 machines is being used to monitor whether all the services are still in use, as soon as a service on one machine breaks down the other machine tries to take over.



1.3. LoadBalancing

With load-balancing clusters the concept is that when a request for say a web-server comes in, the cluster checks which machine is the lease busy and then sends the request to that machine. Actually most of the times a Load-balancing cluster is also Fail-over cluster but with the extra load balancing functionality and often with more nodes.



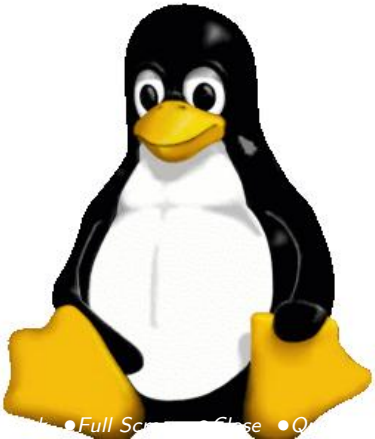
2. HPC

High Performance Computing

The last variation of clustering is the High Performance Computing Cluster, this machine is being configured specially to give data centers that require extreme performance the performance they need. Beowulf's have been developed especially to give research facilities the computing speed they need. These kind of clusters also have some load-balancing features, they try to spread different processes to more machines in order to gain performance. But what it mainly comes down to in this situation is that a process is being parallelized and that routines that can be ran separately will be spread on different machines in stead of having to wait till they get done one after another.

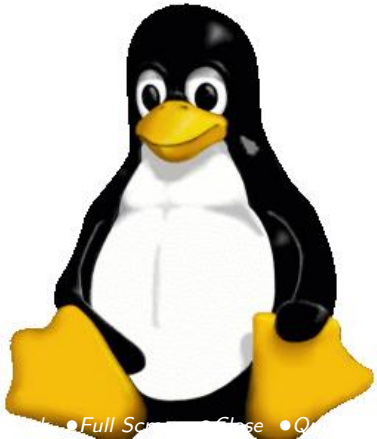
2.1. Beowulf

In late 1993 Donald Becker and Thomas Sterling began sketching the outline of a commodity-based cluster system designed as a cost-effective alternative to large supercomputers. In early 1994, working at CESDIS under the sponsorship of the ESS project, the Beowulf Project was start.
PVM / MPI are the tools that are most commonly being used when people talk about GNU/Linux based Beowulf's.



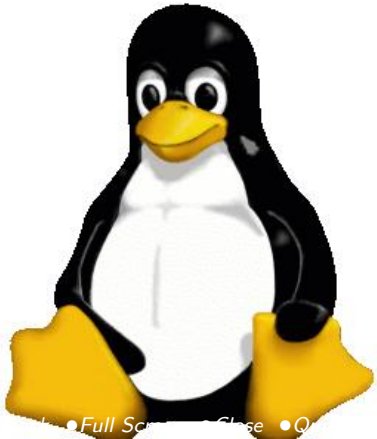
2.2. MPI

MPI stands for Message Passing Interface it is the open standard specification for message passing libraries. MPICH is one of the most used implementations of MPI, next to MPICH you also can use LAM , another implementation of MPI based on the free reference implementation of the libraries.



2.3. PVM

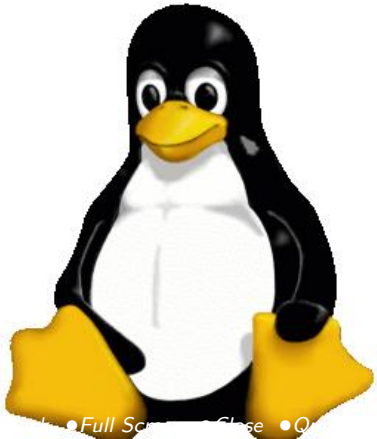
PVM or Parallel Virtual Machine is another cousin of MPI that is also quite often being used as a tool to create a Beowulf. PVM lives in user space so no special kernel modifications are required, basically each user with enough rights can run PVM.



3. openMosix

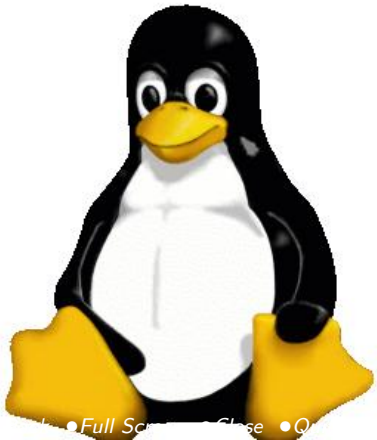
openMosix

- Kernel Patch + userland tools
- Automatic Migration of processes to other nodes
- Statistics define to which node a process can migrate
- not everything migrates, there are limitations



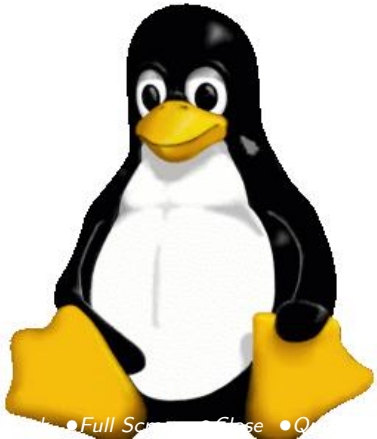
3.1. PRO openMosix

- no extra libraries required (vs PVM/MPI)
- no changes to code required
- no need to actually parrallize your application
- oMFS
- Autodiscovery
- Clustermask



3.2. CON openMosix

- DSM still in Beta
- Issues with Threads not gaining performance
- Kernel dependent
- no performance gain on 1 process

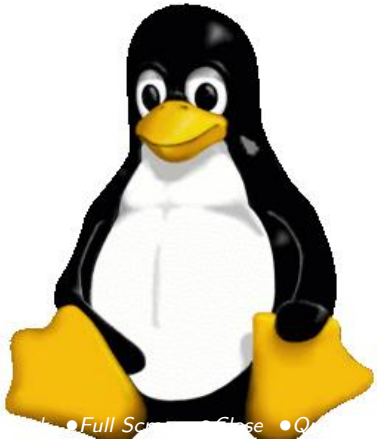


3.3. Limitations

- Threading vs threading issues with Linux

  At one point Neon asks what happens if somebody dies in the Matrix, if he would also die in the real world. Trinity answered to him that a body (a pthread) cannot live without a mind (ie memory, ie address space).

- Applications using shared memory (however a patch is available but I haven't tested it yet)



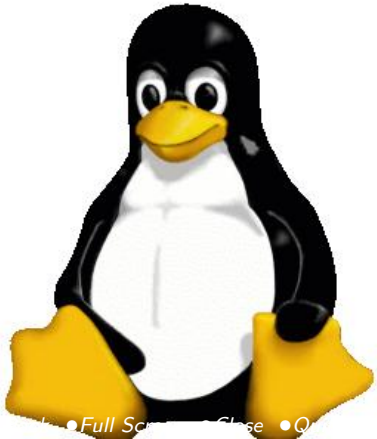
3.4. Planning

Planning your cluster

- Power
- Airco
- Rack

Types of Clusters

- Single Pool
- Server Pool
- Adaptive Pool

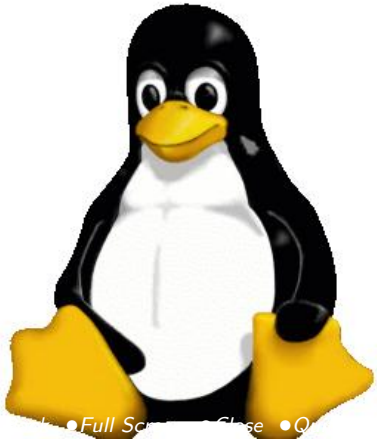


3.5. Compiling

openMosix Compilation
mv /root/openMosix-2.4.20-2.gz /usr/src/linux-2.4.20 cd /usr/src/linux-2.4.20 zcat openMosix-2.4.20-2.gz — patch -Np1

```
...
CONFIG_MOSIX=y
# CONFIG_MOSIX_TOPOLOGY is not set
CONFIG_MOSIX_UDB=y
# CONFIG_MOSIX_DEBUG is not set
# CONFIG_MOSIX_CHEAT_MIGSELF is not set
CONFIG_MOSIX_WEEEEEEEEE=y
CONFIG_MOSIX_DIAG=y
CONFIG_MOSIX_SECUREPORTS=y
CONFIG_MOSIX_DISCLOSURE=3
CONFIG_QKERNEL_EXT=y
CONFIG_MOSIX_DFSA=y
CONFIG_MOSIX_FS=y
CONFIG_MOSIX_PIPE_EXCEPTIONS=y
CONFIG_QOS_JID=y
...
```

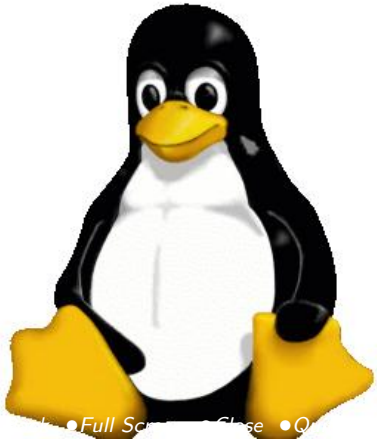
make dep bzImage modules modules_install



3.6. Configuration

openMosix Configuration

- using a manually created config file and setpe
- using autodiscovery



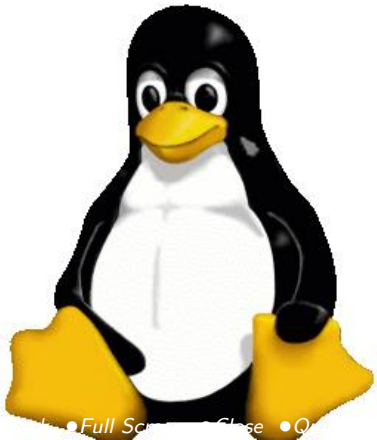
3.7. Config

A config file

```
1       192.168.1.1     1
2       192.168.1.2     1
3       192.168.1.3     1
4       192.168.1.4     1
```

idem

```
1   192.168.1.1  4
```



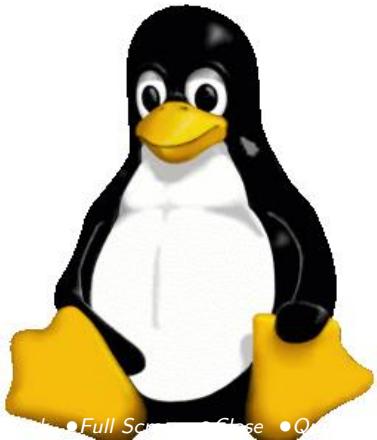
3.8. Starting

Starting openmosix

```
setpe -w -f /etc/openmosix.map
```

or

```
/etc/init.d/openmosix start
```



3.9. MFS

- openMFS is no replacement for NFS
- it is no clusterfilesystem.

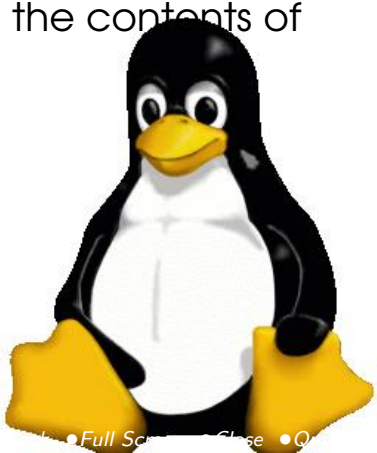
Modify your /etc/fstab

```
mfs_mnt  /mfs mfs   dfsa=0   0 0
```

on each node you will se in /mfs

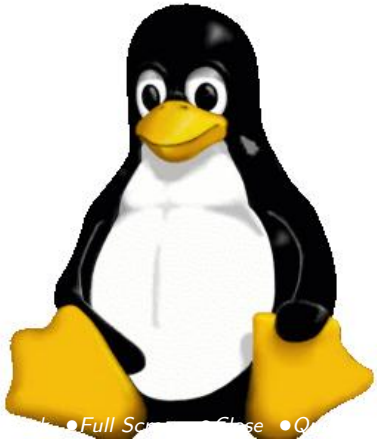
```
1
2
3
```

For each node that exists, these directories will contain the contents of the clusternodes filesystem.



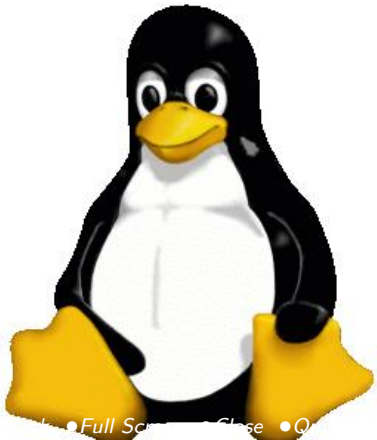
3.10. Running apps

- Nothing fancy , just start the application, it will migrate to the least busy node
- mosrun , run-on , as the name says, you define on which node your applicatoin runs.



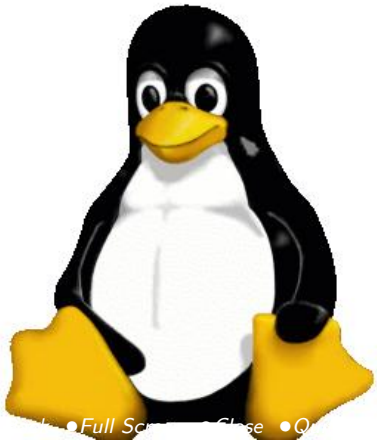
3.11. openMosixView

- View : monitoring
- procs : managing processes
- collector : logs
- analyser : analyse the data collected
- history : a process history
- migmon : a migration monitor / drag and drop



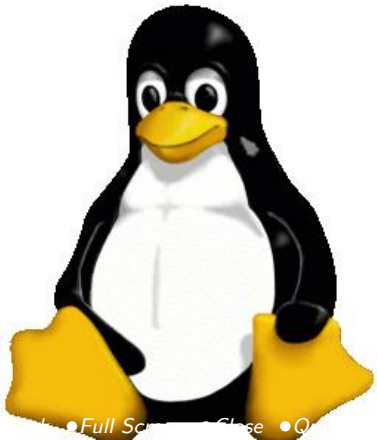
3.12. Testing

- a for lus with
  `awk 'BEGIN {for(i=0;i<10000;i++)for(j=0;j<10000;j++);}' &`
- openMosix stress test



3.13. Examples that Work

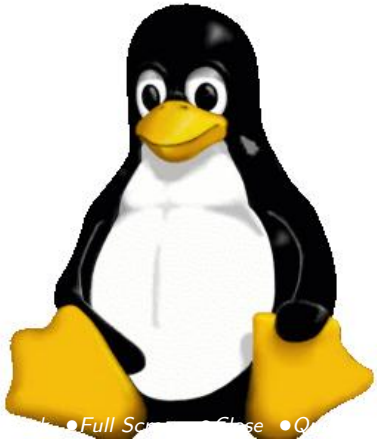
- Mathlab 5
- Blast (Bio Informatics), (patched version)
- MJPEG tools
- bladeenc
- povray
- mpi
- flac



3.14. Not Migrating

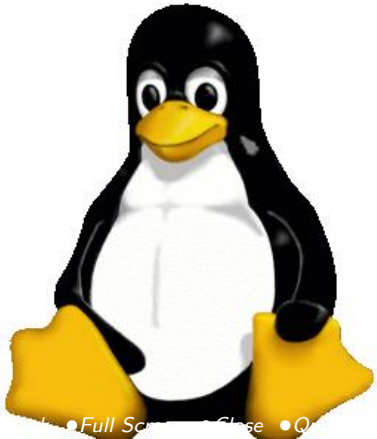
However .. with the current shared memory patch these might be wrong already.

- Apache
- mySQL
- pthreads apps
- Oracle
- Mathlab 6
- VMWare



4. RealLife

An Example



5. Questions

Questions ?