# The Mosix HOWTO

## Kris Buytaert

buytaert@be.stone−it.com

**Revision History**

| | |
|---|---|
| Revision v0.15 | 13 March 2002 |
| Revision v0.13 | 18 Feb 2002 |
| Revision ALPHA 0.03 | 09 October 2001 |

# Table of Contents

# Table of Contents

# Table of Contents

# Chapter 1. Introduction

## 1.1. Introduction

This document gives a brief description to Mosix, a software package that turns a network of GNU/Linux computers into a computer cluster. Along the way, some background to parallel processing is given, as well as a brief introduction to programs that make special use of Mosix's capabilities. The HOWTO expands on the documentation as it provides more background information and discusses the quirks of various distributions.

Kris Buytaert got involved in this piece of work when Scot Stevenson was looking for somebody to take over the Job, this was during February 2002 The first new versions of this HOWTO are rewrites of the Mosix Howto draft and the Suse Mosix HOWTO

("FEHLT", in case you are wondering, is German for "missing"). You will notice that some of the headings are not as serious as they could be. Scot had planned to write the HOWTO in a slightly lighter style, as the world (and even the part of the world with a burping penguin as a mascot) is full of technical literature that is deadly. Therefore some parts still have these comments

Initially this was a draft version of a text intended to help Linux users with SuSE distributions install the Mosix cluster computer package − in other words, to turn networked computers running SuSE Linux into a Mosix cluster. This HOWTO is written on the basis of a monkey−see, monkey−do knowledge of Mosix, not with any deep insight into the workings of the system.

The original text did not cover Mosix installations based on the 2.4.* kernel. Note that SuSE 7.1 does not ship with the vanilla sources to that kernel series.

## 1.2. Disclaimer

Use the information in this document at your own risk. I disavow potential liability for the contents of this document. Use of th concepts, examples, and/or other content of this document is ent at your own risk.

All copyrights are owned by their owners, unless specifically no otherwise. Use of a term in this document should not be regarde affecting the validity of any trademark or service mark.

Naming of particular products or brands should not be seen as endorsements.

You are strongly recommended to take a backup of your system before major installation and backups at regular intervals.

## 1.3. Distribution policy

## 1.4. New versions of this document

New versions of this document can be found on the web pages of the Linux Documentation Project at http://www.linuxdoc.org in the appropriate subfolder. Changes to this document will usually be discussed on the Mosix Mailing List. See the Mosix Hompage http://www.mosix.org for details.

## 1.5. Feedback

Currently this HOWTO is being maintained by Kris Buytaert, please do send questions about Mosix to the mailing list.

Please send comments, questions, bugfixes, suggestions, and of course praise about this document to the author.

If you have a technical question about Mosix itself, please post them on the Mosix mailing list. Do not repeat not send them to the Scott who doesn't know squat about the internals, finds anything written in C++ terribly confusing, and learned Python mainly because the rat on the book cover was so cute.

# Chapter 2. So what is mosix Anyway ?

## 2.1. A very, very brief introduction to clustering

Most of the time, your computer is bored. Start a program like xload or top that monitors your system use, and you will probably find that your processor load is not even hitting the 1.0 mark. If you have two or more computers, chances are that at any given time, at least one of them is doing nothing. Unfortunately, when you really do need CPU power − during a C++ compile, or coding Ogg Vobis music files − you need a lot of it at once. The idea behind clustering is to spread these loads among all available computers, using the resources that are free on other machines.

The basic unit of a cluster is a single computer, also called a "node". Clusters can grow in size − they "scale" − by adding more machines. A cluster as a whole will be more powerful the faster the individual computers and the faster their connection speeds are. In addition, the operating system of the cluster must make the best use of the available hardware in response to changing conditions. This becomes more of a challenge if the cluster is composed of different hardware types (a "heterogenous" cluster), if the configuration of the cluster changes unpredictably (machines joining and leaving the cluster), and the loads cannot be predicted ahead of time.

### 2.1.1. A very, very brief introduction to clustering

#### 2.1.1.1. HPC vs Failover vs Loadbalancing

Basically there are 3 types of clusters, the most deployed ones are probably the Failover Cluster and the Loadbalancing Cluster, HIGH Performance Computing.

Failover 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 wether all the services are still in use, as soon as a service on one machine breaks down the other machine tries to take over.

With loadbalancing clusters the concept is that when a request for say a webserver comes in, the cluster checks wich machine is the lease busy and then sends the request to that machine. Actually most of the times a Loadbalancing cluster is also Failover cluster but with the extra load balancing functionality and often with more nodes.

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. Beowulfs have been developed especially to give research facilities the computing speed they need. These kind of clusters also have some loadbalancing features, they try to spread different processes to more machines in order to gain perfomance. But what it mainly comes down to in this situation is that a process is being parralellised 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.

#### 2.1.1.2. Mainframes and supercomputers vs. clusters

Traditionally Mainframes and Supercomputers have only been built by a selected number of vendors, a company or organisation that required the performance of such a machine had to have a hughe budget available for it`s Supercomputer. Lot`s of universities could not afford them the costs of a Supercomputer, therefore other alternatives were being researched by them. The concept of a cluster was born when people first tried to spread different jobs over more computers and then gather back the data those jobs produced. With cheaper and more common hardware available to everybody, results similar to real Supercomputers were only to be dreamt of during the first years, but as the pc platform developed further, the performance gap between a Supercomputer and a cluster of multiple personal computres became smaller.

#### 2.1.1.3. Cluster models [(N)UMA, DSM, PVM/MPI]

There are different ways of doing parallel processing, (N)UMA, DSM , PVM, MPI are all different kinds of Parallel processing schemes.

(N)uma , (Non−)Uniform Memory Access machines for example have shared access to the memory where they can execute their code. In the Linux kernel there is a NUMA implementation that varies the memory acces times for different regions of memory. It then is the kernel's task to use the memory that is the closest to the cpu it is using.

DSM

PVM / MPI are the tools that are most commonly being used when people talk about GNU/Linux based Beowulfs. 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.

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 userspace so no special kernel modifications are required, basically each user with enough rights can run PVM.

#### 2.1.1.4. Mosix's role

The Mosix software packages turns networked computers running GNU/Linux into a cluster. It automatically balances the load between different nodes of the cluster, and nodes can join or leave the running cluster without disruption. The load is spread out among nodes according to their connection and CPU speeds.

Since Mosix is part of the kernel and maintains full compatibility with normal Linux, a user's programs, files, and other resources will all work as before with no changes necessary. The casual user will not notice the difference between Linux and Mosix. To him, the whole cluster will function as one (fast) GNU/Linux system.

## 2.2. The story so far

### 2.2.1. Historical Development

The name "Mosix" comes from FEHLT. The 6th incarnation of Mosix was developed for BSD/OS. GNU/Linux was chosen as a development platform for the 7th incarnation in DATE_FEHLT because of

### 2.2.2. Current state

Like most active Open Source programs, Mosix's rate of change tends to outstrip the the follower's ability to keep the documentation up to date. See the Mosix Home Page for current news. The following relates to Mosix VERSION FEHLT for the Linux kernel FEHLT as of DATUM FEHLT:

### 2.2.3. openMosix

openMosix is in addition to whatever you find at mosix.org and in full appreciation and respect for Prof. Barak's leadership in the outstanding Mosix project .

Moshe Bar has been involved for a number of years with the Mosix project (www.mosix.com) and was co−project manager of the Mosix project and general manager of the commercial Mosix company.

After a difference of opinions on the commercial future of Mosix, he has started a new clustering company − Qlusters, Inc. − and Prof. Barak has decided not to participate for the moment in this venture (although he did seriously consider joining) and held long running negotiations with investors. It appears that Mosix is not any longer supported openly as a GPL project. Because there is a significant user base out there (about 1000 installations world−wide), Moshe Bar has decided to continue the development and support of the Mosix project under a new name, openMosix under the full GPL2 license. Whatever code in openMosix comes from the old Mosix project is Copyright 2002 by Amnon Bark. All the new code is copyright 2002 by Moshe Bar.

openMosix is a Linux−kernel patch which provides full compatibility with standard Linux for IA32−compatible platforms. The internal load−balancing algorithm transparently migrates processes to other cluster members. The advantage is a better load−sharing between the nodes. The cluster itself tries to optimize utilization at any time (of course the sysadmin can affect these automatic load−balacing by manuel configuration during runtime).

This transparent process−migration feature make the whole cluster look like a BIG SMP−system with as many processors as available cluster−nodes (of course multiplicated with 2 for dual−processor systems). openMosix also provides a powerful mized for HPC−applications, which unlike NFS provides cache consistency, time stamp consistency and link consistency.

There could (and will) be significant changes in the architecure of the future openMosix versions. New concepts about auto−configuration, node−discovery and new user−land tools are discussed in the openMosix−mailinglist.

To approach standardization and future compatibility the proc−interface changes from /proc/mosix to /proc/hpc and the /etc/mosix.map was exchanged to /etc/hpc.map. Adapted commandline user−space tools for openMosix are already available on the web−page of the project and from the current version (1.1) Mosixview supports openMosix as well.

The hpc.map will be replaced in the future with a node−autodiscovery system.

openMosix is supported by various competent people (see www.openMosix.org) working together around the world. The gain of the project is to create a standardize clustering−environent for all kinds of HPC−applications.

openMosix has also a project web−page at http://openMosix.sourceforge.net with a CVS tree and mailinglist for the developer and user.

## 2.3. Mosix in action: An example

Mosix clusters can take various forms. To demonstrate, let's assume you are a student and share a dorm room with a rich computer science guy, with whom you have linked computers to form a Mosix cluster. Let's also assume you are currently converting music files from your CDs to Ogg Vobis for your private use, which is legal in your country. Your roommate is working on a project in C++ that he says will bring World Peace. However, at just this moment he is in the bathroom doing unspeakable things, and his computer is idle.

So when you start a program called FEHLT to convert Bach's .... from .wav to .ogg format, the Mosix routines on your machine compare the load on both nodes and decide that things will go faster if that process is sent from your Pentium−233 to his Athlon XP. This happens automatically − you just type or click your commands as you would if you were on a standalone machine. All you notice is that when you start two more coding runs, things go a lot faster, and the response time doesn't go down.

Now while you're still typing ...., your roommate comes back, mumbling something about red chile peppers in cafeteria food. He resumes his tests, using a program called 'pmake', a version of 'make' optimized for parallel execution. Whatever he's doing, it uses up so much CPU time that Mosix even starts to send subprocesses to your machine to balance the load.

This setup is called *single−pool*: All computers are used as a single cluster. The advantage/disadvantage of this is that you computer is part of the pool: Your stuff will run on other computers, but their stuff will run on your's, too.

## 2.4. Components

### 2.4.1. Process migration

### 2.4.2. The Mosix File System (MFS)

### 2.4.3. Direct File System Access (DFSA)

Both Mosix and openMosix provide a cluster−wide filesystem (MFS) with the DFSA−option (direct filesystem access). It provides access to all local and remote filesystems of the nodes in an Mosix or openMosix cluster.

## 2.5. Work in Progress

### 2.5.1. Network RAM

### 2.5.2. Migratable sockets

### 2.5.3. High availablility

# Chapter 3. Features of Mosix

## 3.1. Pros of Mosix

No extra packages required  No Code changes required

## 3.2. Cons of Mosix

Kernel Dependent  Not Everything works this way  Shared memory issues

Issues with Multiple Threads not gaining performance.

You won't gain performance when running 1 single process such as your  Browser on a Mosix Cluster , the process won't spread itselve over the cluster. Except of course your process wil migrate to a more performant machine.

## 3.3. Extra Features in OpenMosix

# Chapter 4. Requirements and Planning

## 4.1. Hardware requirements

Installing a basic clusters requires at least 2 machines with network connected. Either using a crosscable between the two network cards or a switch or hub. Off course the faster your networkcards the easier you will get better performance for your cluster. These days Fast Ethernet is standard, putting multiple ports in a machine isn`t that difficult, but make sure to connect them through other physical networks in order to gain the speed you want. Gigabit ethernet is getting cheaper any day now but I suggest that you don`t rush to the shop spending your money before you have actually tested your setup with multiple 100Mbit cards and noticed that you really do need the extra network capacity.

## 4.2. Hardware Setup Guidelines

Setting up a big cluster requires some thinking to be done, where are you going to put the machines, not under a table somehwere or in the middle of your office. It`s ok if you just want to do some small tests , but if you are planning to deploy a N node cluster you will have to make sure that the environment that will hold this machine is capable of doing so. I`m talking about preparing one or more 19" racks to host the machines, configure the appropirate network topology, either straight, single connected or even a 1 to 1 cross connected network between al your nodes. You will also need to make sure that there is enough power to support such a range of machines. That your airconditioning system supports the load and that in case of powerfailure your UPS can cleanly shut down al the required systems. You might want to invest in a KVM Switch in order to fasciliate access to the machines consoles. But even if you don`t have the number of nodes that justify these investments, make sure that you can always easily access the different nodes, you never know when you have to replace the fan or a harddisk of a machine in trouble. If that means that you have to unload a stack of machines to reach the bottom one hence shutting down your cluster you are in trouble.

## 4.3. Software requirements

The systems we plan to use will need a basic Linux installation of your choice, RedHat , Suse , Debian or another distribution, it doesn`t really matter which one. What does matter is that the kernel is at least on 2.4 level, and that your networkcards are configured correctly, next to that you`ll need a healthy space of swap.

## 4.4. Planning your cluster

How to configure MOSIX clusters with a pool of servers and a set of (personal) workstations:

- Single−pool = all the servers and workstations are used as a single cluster: install the same "mosix.map" in all the computers, with the IP addresses of all the computers. Advantage/disadvantage: your workstation is part of the pool.
- Server−pool = servers are shared while workstations are not part of the cluster: install the same "mosix.map" in all the servers, with the IP addresses of only the servers. Advantage/disadvantage: remote processes will not move to your workstation. You need to login to one of the servers to use the cluster.

- Adaptive−pool = servers are shared while workstations join or leave the cluster,  e.g. from 5PM to 8AM: install the same "mosix.map" in all the computers, with the IP addresses of all the servers and workstations,  then use a simple script, to decide whether MOSIX should be activated or deactivated. Advantage/disadvantage: remote processes can use your workstation when you are  not using it.

# Chapter 5. Distribution specific installations

## 5.1. Installing Mosix

This chapter deals with installing Mosix and OpenMosix on different distributions. It won't be an exhaustive list of all the possible combinations. However thoughout the chapter you should find enough information on installing Mosix in your environment.

Techniques for installing multiple machines with Mosix will be discussed in the next chapter.

## 5.2. Getting Mosix

## 5.3. Getting OpenMosix

## 5.4. openMosix General Instructions

### 5.4.1. Kernel Compilation

Always use pure vanilla kernel−sources from e.g. www.kernel.org to compile an openMosix kernel! Be sure to use the right openMosix version dependend on the kernel−version. Do not use the kernel that comes whith any linux−distribution; it won't work.

Download the actual version of openMosix and untar it in your kernel−source directory (e.g. /usr/src/linux−2.4.16). If your kernel−source directory is other than "/usr/src/linux−[version_number]" at least the creation of a symbolic link to "/usr/src/linux−[version_number]" is required. Now apply the patch using the patch utility:

```
patch −Np1 < openMosix1.5.2moshe
```

This command displays now a list of patched files from the kernel−sources. Enable the openMosix−options in the kernel−configuration e.g.

```
...
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
...
```

and compile it with:

```
make dep bzImage modules modules_install
```

After compilation install the new kernel with the openMosix options within you boot−loader e.g. insert an entry for the new kernel in /etc/lilo.conf and run lilo after that.

Reboot and your OpenMosx−cluster(node) is up!

## 5.4.2. hpc.map

Syntax of the /etc/hpc.map file Before starting openMosix there has to be a /etc/hpc.map configuration file (on each node) which must be equal on each node. The hpc.map contains three space seperated fields:

```
openMosix−Node_ID               IP−Adress(or hostname)          Range−size
```

An example hpc.map could look like this:

```
1       node1   1
2       node2   1
3       node3   1
4       node4   1
```

or

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

or with the help of the range−size these both exampels are equal with:

```
1       192.168.1.1     4
```

openMosix "counts−up" the last byte of the ip−adress of the node according to its openMosix−ID. (if you use a range−size greater than 1 you have to use ip−adresses instead of hostnames)

If a node has more than one network−interfaces it can be configured with the ALIAS option in the range−size field (which is equal to set the range−size to 0) e.g.

```
1       192.168.1.1     1
2       192.168.1.2     1
3       192.168.1.3     1
4       192.168.1.4     1
4       192.168.10.10   ALIAS
```

Here the node with the openMosix−ID 4 has two network−interfaces (192.168.1.4 + 192.168.10.10) which are both visible to openMosix.

Always be sure to run the same openMosix version AND configuration on each of your Cluster nodes!

Start openMosix with the "setpe" utility on each node : setpe −w −f /etc/hpc.map Execute this command (which will be described later on in this howto) on every node in your openMosix cluster. Installation finished now, the cluster is up and running :)

## 5.4.3. MFS

At first the CONFIG_MOSIX_FS option in the kernel configuration has to be enabled. If the current kernel was compiled without this option recompilation with this option enabled is required. Also the UIDs and GUIDs in the cluster must be equivalent. The CONFIG_MOSIX_DFSA option in the kernel is optional but of course required if DFSA should be used. To mount MFS on the cluster there has to be an additional fstab−entry on each nodes /etc/fstab.

for DFSA enabled:

```
mfs_mnt         /mfs            mfs     dfsa=1          0 0
```

for DFSA disabled:

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

the syntax of this fstab−entry is:

```
[device_name]           [mount_point]   mfs     defaults        0 0
```

After mounting the /mfs mount−point on each node, each nodes filesystem is accessable through the /mfs/[openMosix_ID]/ directories.

With the help of some symbolic links all cluster−nodes can access the same data e.g. /work on node1

```
on node2 :      ln −s /mfs/1/work /work
on node3 :      ln −s /mfs/1/work /work
on node3 :      ln −s /mfs/1/work /work
...
```

Now every node can read+write from and to /work !

The following special files are excluded from the MFS:

the /proc directory

special files which are not regular−files, directories or symbolic links e.g. /dev/hda1

Creating links like:

```
ln −s /mfs/1/mfs/1/usr
```

or

```
ln −s /mfs/1/mfs/3/usr
```

is invalid.

The following system calls are supported without sending the migrated process (which executes this call on its home (remote) node) going back to its home node:

read, readv, write, writev, readahead, lseek, llseek, open, creat, close, dup, dup2, fcntl/fcntl64, getdents, getdents64, old_readdir, fsync, fdatasync, chdir, fchdir, getcwd, stat, stat64, newstat, lstat, lstat64, newlstat, fstat, fstat64, newfstat, access, truncate, truncate64, ftruncate, ftruncate64, chmod, chown, chown16, lchown, lchown16, fchmod, fchown, fchown16, utime, utimes, symlink, readlink, mkdir, rmdir, link, unlink, rename

Here are situations when system calls on DFSA mounted filesystems may not work:

diffrent mfs/dfsa configuration on the clusternodes

dup2 if the second file−pointer is non−DFSA

chdir/fchdir if the parent dir is non−DFSA

pathnames that leave the DFSA−filesystem

when the process which executes the system−call is being traced

if there are pending requests for the process which executes the system−call

## 5.5. RedHat

## 5.6. Suse 7.1 and Mosix

### 5.6.1. Versions Required

The following is based on using SuSE 7.1 (German Version), Linux  Kernel 2.2.19, and Mosix 0.98.0.

The Linux Kernel 2.2.18 sources are part of the SuSE distribution. Do  not use the default SuSE 2.2.18 kernel, as it is heavily patched with  SuSE stuff. Get the patch for 2.2.19 from your favorite mirror such as  . If there are further patches for the 2.2.* kernel  RROR URL HERE by the time you read this text, get those, too.

If one of your machines is a laptop with a network connection via  pcmcia, you will need the pcmcia sources, too. They are included in  the SuSE distribution as MISSING: RPM HERE.

Mosix 0.98.0 for the 2.2.19 kernel can be found on  http://www.mosix.org as MOSIX−0.98.0.tar.gz . While you are there, you  might want to get some of the contributed software like qps or mtop.  Again, if there is a version more current than 0.98.0 by the time you  read this, get it instead.

SuSE 7.1 ships with a Mosix−package as a rpm MISSING: RPM HERE  Ignore this package. It is based on Kernel 2.2.18 and seems to have  been modified by SuSE (see /usr/share/doc/packages/mosix/README.SUSE).  You are better off installing the Mosix sources and installing from  scratch.

### 5.6.2. Installation

We're assuming your hardware and basic Linux system are all set up correctly and that you can at least telnet (or ssh) between the different machines. The procedure is described for one machine.  Log in as root. Install the sources for the 2.2.18 Kernel in /usr/src. SuSE will place them there automatically as /usr/src/linux−2.2.18 if you install the RPM RPM NAME. Rename the directory to /usr/src/linux−2.2.19. Remove the existing link /usr/src/linux and create a new one to this directory with

```
        ln −s /usr/src/linux−2.2.19 linux
```

(assuming you are in /usr/src).  Patch the kernel to 2.2.19 (or whatever the current version is). If you do not know to do this, check the Linux Kernel HOWTO.  Make a directory /usr/src/linux−2.2.19−mosix and copy the contents of the vanilla kernel /usr/src/linux−2.2.19 there with the command

```
        cp −rp linux−2.2.19/* linux−2.2.19−mosix/
```

This gives you a clean backup kernel to fall back on if something goes wrong. Remove the /usr/src/linux link (again). Create a link /usr/src/linux to /usr/src/linux−2.2.19−mosix with

```
        ln −s /usr/src/linux−2.2.19−mosix linux
```

to make life easier.  Change to /tmp, copy the Mosix sources there and unpack them with the command

```
        tar xfz MOSIX−0.98.0.tar.gz
```

Do not unpack the resulting tar archives such as /tmp/user.tar that appear.

### 5.6.3. Setup

- Run the install script /tmp/mosix.install and follow instructions.

  Mosix should be enabled for run level 3 (full multiuser with network, no xdm) and 5 (full multiuser with network and xdm). There is no run level 4 in SuSE 7.1.

  The Mosix install script does not give you the option of creating a boot floppy instead of an image. If you want a boot floppy, you will have to run "make bzdisk" after the install script is through.

  Do not repeat /not/ reboot.

- The install script in Mosix 0.98.0 is made for RedHat distributions and therefore fails to set up some SuSE files correctly. It tries to put stuff in /sbin/init.d/, which in fact is /etc/init.d/ (or /etc/rc.d/) with SuSE. Also, there is no /etc/rc.d/init.d/ in SuSE. So:

  - Copy /tmp/mosix.init to /etc/init.d/mosix and make it executable with the command

    ```
                    chmod 754 /etc/init.d/mosix
    ```
  - MISSING − MODIFY ATD stuff "/etc/rc.d/init.d/ATD" BY HAND
  - MISSING − MODIFY THE "/etc/cron.daily/slocate.cron" FILE
  - The other files − /etc/inittab, /etc/inetd.conf, /etc/lilo.conf − are modified correctly.
- Edit the file /etc/inittab to prevent some processes from migrating to other nodes by inserting the command "/bin/mosrun −h" in the following lines:

  Run levels:

  ```
          l0:0:wait:/bin/mosrun −h /etc/init.d/rc 0
          l1:1:wait:/bin/mosrun −h /etc/init.d/rc 1
          l2:2:wait:/bin/mosrun −h /etc/init.d/rc 2
          l3:3:wait:/bin/mosrun −h /etc/init.d/rc 3
          l5:5:wait:/bin/mosrun −h /etc/init.d/rc 5
          l6:6:wait:/bin/mosrun −h /etc/init.d/rc 6
  ```
  (Remember, there is no run level 4 in SuSE 7.1)

  Shutdown and sulogin:

  ```
      ~~:S:respawn:/bin/mosrun −h /sbin/sulogin
          ca::ctrlaltdel:/bin/mosrun −h /sbin/shutdown −r −t 4 now
          sh:12345:powerfail:/bin/mosrun −h /sbin/shutdown −h now THE \
             POWER IS FAILING
  ```

  It is not necessary to prevent the /sbin/mingetty processes from migrating − in fact, if you do, all of the child processes started from your login shell will be locked, too [Note to German readers: This is mistake in the article "Zwischen Multiprocessing und Cluster−Computing" on Mosix in "Linux−Magazin" 6/2000].

- To enable the processes started by your window manager to migrate, edit the files ~/.xinitrc and ~/.xsession by going to the end of the file and changing the line "exec $WINDOWMANAGER" to

  ```
         exec /bin/mosrun −l $WINDOWMANAGER
  ```
  You should be able to enable migration for all users' window mangers by modifying the equivalent line in /etc/X11/xdm/Xsession MISSING: NOT TESTED YET. However, see section 8 "Notes" for reasons why you might not want to do this by default.

- The command to start and stop Mosix (do not repeat /not/ do this now) is

  ```
         /etc/init.d/mosix {start|stop|status|restart|reload}
  ```

  To have Mosix start automatically at boot time, go to /etc/init.d/ . In the subdirectories ./rc3.d and ./rc5.d, create the following links:

  ```
          ln −s ../mosix S30mosix
          ln −s ../mosix K01mosix
  ```

  The first line causes Mosix to be called as the last part of the install procedure for the given run level, the second line closes it down as one of the first services.
- Create a file /etc/mosix.map following the instructions in the Mosix documentation. In the most simple case, you will have n computers which have their IP−addresses in sequence so that the map file will simply look like

  ```
          1   IP−address of first node  n
  ```

  This is where a lot of errors occur, let me clarify this with an example. Suppose you have 5 machines, 10.0.0.1, 10.0.0.2 , 10.0.0.100 , 10.0.0.101 and 10.0.0.150 your mosix.map would look like

  ```
          1  10.0.0.1     2
          3  10.0.0.100   2
          5  10.0.0.150   1
  ```

  PLEASE VERIFY THIS !!!!!!!
- Run "/etc/versionate", which will most probably tell you that the Mosix module already has a version. Do it anyway.
- Now, finally, reboot. The computer should come up running Mosix.

## 5.7. Debian and Mosix

Installing mosix on a Debian based machine can be done as described below First step is downloading the packages from the net. Since we are using a debian setup we needed

```
http://packages.debian.org/unstable/net/mosix.html
http://packages.debian.org/unstable/net/kernel−patch−mosix.html
http://packages.debian.org/unstable/net/mps.html
```

You can also apt−get install them ;) Next part is making the kernel mosix capable. Copy the patch.$kernel version in to your /usr/src/linux−$version directory run

```
patch −p0 < patches.2.4.10
```

Check your kernel config and run

```
make dep ; make clean ; make bzImage ; make modules ; make modules_install
```

You now wil need to edit your /etc/mosix/mosix.map This file has a bit a strange layout. We have 2 machines 192.168.10.65 and 192.168.10.94 This gives us a mosix.map that looks like

```
1 192.168.10.65 1
2 192.168.10.94 1
```

After rebooting with this kernel (lilo etc you know the drill), you then should have a cluster of mosix machines that talk to eachoter and that do migration of processes. You can test that by running the following small script ..

```
awk 'BEGIN {for(i=0;i<10000;i++)for(j=0;j<10000;j++);}'
```

a couple of times, and monitor it`s behaviour with mon where you will see that it spreads the load between 2 different nodes. If you have enabled Process−arrival messages in your kernel you will notice that each time a remote (guest) process arrives on your node a Weeeeeee will be printed and each time a local proces returns you will see a Woooooo on your console. So basically If you don`t see any of those messages during the running of a program and if you have this option enabled in your kernel you might conclude that no processes migrate. We also setup Mosixview (0.8) on the debian machine

```
apt−get install mosixview
```

In order to be able to actually use Mosixview you will need to run it from a user who can log in to the different nodes as root. We suggest you set this up using ssh. Please note that there is a difference between the ssh and ssh2 implemtations .. if you have a identity.pub ssh wil check authorized_keys, if you have id_rsa.pub you will need authorized_keys2 !! Mosixview gives you a nice interface that shows the load of different machines and gives you the possibility to migrate processes manually. A detailed discussion of Mosixview can be found elsewhere in this document.

## 5.8. Other distributions

Based on the explanations above you should be able to install Mosix on most other Linux platforms.

# Chapter 6. Cluster Installation

## 6.1. Cluster Installations

This chapter does not deal with installing Mosix as such, it does however deal with installing multiple machines with mosix. Automated or semi automated mass installs.

## 6.2. Installation scripts [LUI, ALINKA]

## 6.3. The easy way: Automatic installation

## 6.4. The hard way: When scripts don't work

## 6.5. Kick Start Installations

## 6.6. DSH,  Distributed Shell

# Chapter 7. ClumpOS

## 7.1. What is Clump/OS

clump/os is a CD−based Linux /MOSIX mini−distribution designed to allow users to quickly, or temporarily, add nodes to a MOSIX cluster; As I write this in march 2002 the version (release 5.x) is a 5.3M ISO download.

This chapter has been contributed by Jean−David Marrow who is the main author of Clump/OS.

## 7.2. How does it work

At boot−time, clump/os will autoprobe for network cards, and, if any are detected, try to configure them via DHCP. If successful, it will create a mosix.map file based on the assumption that all nodes are on local CLASS C networks, and configure MOSIX using this information. clump/os Release 4 best supports machines with a single connected network adapter. The MOSIX map created in such cases will consist of a single entry for the CLASS−C network detected, with the node number assigned reflecting the IP address received from DHCP. (On the 192.168.1 network, node #1 will be 192.168.1.1, etc.) If you use multiple network adapters Expert mode is recommended as the assignment of node numbers is sensitive to the order in which network adapters are detected. (Future releases will support complex topologies and feature more intelligent MOSIX map creation.) clump/os will then display a simple SVGA monitor (clumpview) indicating whether the node is configured, and, if it is, showing the load on all active nodes on the network. When you've finished using this node, simply press [ESC] to exit the interface and shutdown.

Alternatively, or if autoconfiguration doesn't work for you, then you can use clump/os in Expert mode. Please note that clump/os is not a complete distribution or a rescue disk; the functionality present is the bare minimum required for a working MOSIX server node.

It works for us, but may not work for you; if you experience difficulties, please email us with as much information about your system as possible −− after you have investigated the problem. (See Problems? and Expert mode. You might also consider subscribing to the clump/os mailing list.)

## 7.3. Requirements

As the purpose of clump/os is to add nodes to a cluster, it is assumed that you already have a running MOSIX cluster −− or perhaps only a single MOSIX node −− from which you will be initiating jobs. All machines in the cluster must conform to the following requirements:

clump/os Machine(s) 586+ CPU,

bootable CDROM

NIC

64M+ RAM (the system is loaded entirely into a ramdisk; this means that you should have at least 64M of RAM (and likely more) to accomodate the approx. 16M ramdisk, space needed for Linux itself, and space for your work. This approach was chosen so that the same CDROM can be used to configure multiple systems.)

Master Machine(s) Linux 2.4.17, MOSIX 1.5.7 (manually configured)

Network Environment Running DHCP server (f you don't, or won't, run DHCP, you can still manually configure your system; see Problems? and Expert Mode. Using DHCP is highly recommended, however, and will greatly simplify your life in the long run. )

The following network modules are present, although not all support autoprobing; if you don't see support for your card in this list, then clump/os will not work for you even in Expert Mode.

*3c501.o 3c503.o 3c505.o 3c507.o 3c509.o 3c515.o 3c59x.o 8139cp.o 8139too.o 82596.o 8390.o ac3200.o acenic.o at1700.o cs89x0.o de4x5.o depca.o dgrs.o dl2k.o dmfe.o dummy.o e2100.o eepro.o eepro100.o eexpress.o epic100.o eth16i.o ewrk3.o fealnx.o hamachi.o hp−plus.o hp.o hp100.o lance.o lp486e.o natsemi.o ne.o ne2k−pci.o ni5010.o ni52.o ni65.o ns83820.o pcnet32.o sis900.o sk98lin.o smc−ultra.o smc9194.o starfire.o sundance.o sungem.o sunhme.o tlan.o tulip.o via−rhine.o wd.o winbond−840.o yellowfin.o*

Please also note that clump/os may not work on a laptop, definately doesn't support PCMCIA cards, and will probably not configure MOSIX properly if your machine contains multiple connected ethernet adapters; see Note 1. This is a temporary limitation of the configuration scripts, and the Release 3/4 kernels which are compiled without CONFIG_MOSIX_TOPOLOGY

## 7.4. Getting Started

You can download the latest clump/os ISO under the terms of the GPL, without warranty of any kind, from the clump os website. Afterwards you have to burn the image to CDROM, insert the CD into your drive, and reboot. (More detailed instructions are in the works, but all the information you need is somewhere on this page −− please read the notes in the margin!)

## 7.5. Problems ?

- *The CDROM doesn't boot*

  Check your BIOS settings to make sure that your machine is configured to boot from the CDROM drive; also make sure that the CDROM is the first boot device.
- *The SVGA interface doesn't work, or the display is incorrect*

  Boot into Expert mode, and send us mail describing your video hardware so that we can correct this in future versions. (You won't be able to use clumpview for now.) If at all possible, please send us a working libsvga configuration file for the machine in question.
- *The network adapter isn't detected/autoconfigured (or no DHCP)*

  If you see a message (in clumpview) stating that no ethernet devices were configured, or that this node isn't configured yet, then either your ethernet card was not detected or the system was not able to configure the card via DHCP.

  If you don't have a DHCP server configured and running on your local network, clump/os will never autoconfigure; if you have multiple connected network adapters, then clump/os may not configure

MOSIX properly. If autoprobing for your network adapter doesn't work, or if you aren't using DHCP, then you'll have to configure your card manually in Expert mode −− using insmod, ifconfig, and route −− and then configure MOSIX via setpe.

If you do need to manually configure your network adapter, please advise us. We'd like to solve this problem, if possible, or at least document which network cards autoprobe correctly.

- *Migrating processes generate errors ("Network Unreachable")*

  This rare problem can be caused by conflicts resulting from differently configured kernels −− even if you are using the correct MOSIX and Linux kernel versions. If clump/os correctly detects all your nodes, but migrating processes generate errors, then please compare your master node's kernel configuration file with the R4.x kernel .config.
- *Migrating processes generate errors ("Process migration failed: incompatible topology")*

  You are likely using master nodes with CONFIG_MOSIX_TOPOLOGY defined, which is not supported by clump/os at this time. See Requirements, and compare your kernel configuration as per the previous FAQ; you will need to recompile your master node kernel(s).

If you don't find your issue here, please consider posting to the clump/os mailing list. (Please note that only subscribers are permitted to post; click on the link for instructions.) You should also make certain that you are using the latest versions of MOSIX and clump/os, and that the versions −− clump/os R4.x and MOSIX 1.5.2 at the time of this writing −− are in sync.

## 7.6. Expert Mode

If you hold down shift during the boot process, you have the option of booting into Expert mode; this will cause clump/os to boot to a shell rather than to the graphical interface. From this shell you can attempt to insert the appropriate module for your network adapter (if autoprobing failed), and/or configure your network and MOSIX manually. Type "halt" to shut down the system. (Note that since the system resides in RAM you can't hurt yourself too badly by rebooting the hard way if you have to −− unless you have manually mounted any partitions rw, that is, and we don't recommend doing so at this point.)

If you want to run clumpview, execute:

```
     open −s −w −− clumpview −−drone −−svgalib
```

This will force the node into 'drone' mode (local processes will not migrate), and will force clumpview to use SVGALIB; the open command will ensure that a separate vt is used.

Please be advised that the environment provided was initially intentionally minimalistic; if you require additional files, or wish to copy files from the system to another machine, your only option is nc (netcat −− a great little utility, btw), or mfs if MOSIX is configured. From version R5.4 on size is no longer a primary consideration.

Expert mode (and clump/os for that matter) is 'single−user'; this is one of the reasons that utilities such as ssh are not included. These and other similar decisions were made in order to keep clump/os relatively small, and do not affect cluster operation.

From version R5.4, if you experience problems in Expert Mode, you can boot into Safe Mode; in Safe Mode no attempt is made at autoconfiguration.

# Chapter 8. Administrating openMosix

## 8.1. Basic Administration

openMosix provides the advantage of process migration to HPC−applications.  The administrator can configure and tune the openMosix−cluster by using  the openMosix−userspace−tools or the /proc/hpc interface which will be now described in detail.

## 8.2. Configuration

The values in the flat files in the /proc/hpc/admin directory presenting  the current configuration of the cluster. Also the administrator can write its own values into these  files to change the configuration  during runtime, e.g.

```
echo 1 > /proc/hpc/admin/block          −blocks the arrival of remote processes
echo 1 > /proc/hpc/admin/bring          −bring all migrated processes home
...
/proc/hpc/admin/ (binary files)          config
                                        −the main configuration file (written by the setpe util)

(flat files)            block           −allow/forbid arrival of remote processes
                        bring           −bring home all migrated processes
                        dfsalinks       −list of current symbolic dfsa−links
                        expel           −sending guest processes home
                        gateways        −maximum number of gateways
                        lstay           −local processes shoud stay

                        mospe           −contains the openMosix node id
                        nomfs           −disables/enables MFS
                        overheads       −for tuning
                        quiet           −stop collecting load−balacing informations
                        decayinterval   −interval for collecting informations about load−balancin
                        slowdecay       −default 975
                        fastdecay       −default 926
                        speed           −speed relative to PIII/1GHz)
                        stay            −enables/disables automatic process migration
```

Writing a 1 to the following files

```
/proc/hpc/decay/

                        clear           −clears the decay statistics
                        cpujob          −tells openMosix that the process is cpu−bound
                        iojob           −tells openMosix that the process is io−bound
                        slow            −tells openMosix to decay its statistics slow
                        fast            −tells openMosix to decay its statistics fast
```

## 8.3. Informations about the other nodes

```
/proc/hpc/nodes/[openMosix_ID]/cpus             −how many cpu's the node has
/proc/hpc/nodes/[openMosix_ID]/load             −the openMosix load of this node
/proc/hpc/nodes/[openMosix_ID]/mem              −available memory as openMosix believes
/proc/hpc/nodes/[openMosix_ID]/rmem             −available memory as Linux believes
/proc/hpc/nodes/[openMosix_ID]/speed            −speed of the node relative to PIII/1GHz
```

```
/proc/hpc/nodes/[openMosix_ID]/status           −status of the node
/proc/hpc/nodes/[openMosix_ID]/tmem             −available memory
/proc/hpc/nodes/[openMosix_ID]/util             −utilization of the node
```

## 8.4. Additional Informations about processes

local processes

```
/proc/[PID]/cantmove                    −reason why a process cannot be migrated
/proc/[PID]/goto                        −to which node the process should migrate
/proc/[PID]/lock                        −if a process is locked to its home node
/proc/[PID]/nmigs                       −how many times the process migrated
/proc/[PID]/where                       −where the process is currently being computed
/proc/[PID]/migrate                     −same as goto remote processes
/proc/hpc/remote/from                   −the home node of the process
/proc/hpc/remote/identity               −additional informations about the process
/proc/hpc/remote/statm                  −memory statistic of the process
/proc/hpc/remote/stats                  −cpu statistics of the process
```

## 8.5. the userspace−tools

These following tools are providing easy adminitration to openMosix  clusters.

```
migrate −send a migrate request to a process
                syntax:
                        migrate [PID] [openMosix_ID]


mon             −is a ncurses−based terminal monitor
                 several informations about the current status are displayed in bar−charts

mosctl          −is the openMosix main configuration utility
                syntax:
                        mosctl  [stay|nostay]
                                [stay|nolstay]
                                [block|noblock]
                                [quiet|noquiet]
                                [nomfs|mfs]
                                [expel|bring]
                                [gettune|getyard|getdecay]

                        mosctl  whois   [openMosix_ID|IP−address|hostname]

                        mosctl  [getload|getspeed|status|isup|getmem|getfree|getutil]   [openMosi

                        mosctl  setyard [Processor−Type|openMosix_ID||this]

                        mosctl  setspeed        interger−value

                        mosctl  setdecay interval       [slow fast]

more detailed:

stay            −no automatic process migration
nostay          −automatic process migration (default)
```

```
lstay           −local processes should stay
nolstay         −local processes could migrate
block           −block arriving of guest processes
noblock         −allow arriving of guest processes
quiet           −disable gathering of load−balancing informations
noquiet         −enable gathering of load−balancing informations
nomfs           −disables MFS
mfs             −enables MFS
expel           −send away guest processes
bring           −bring all migrated processes home
gettune         −shows the current overhead parameter
getyard         −shows the current used Yardstick
getdecay        −shows the current decay parameter
whois           −resolves openMosix−ID, ip−addresses and hostnames of the cluster
getload         −display the (openMosix−) load
getspeed        −shows the (openMosix−) speed
status          −displays the current status and configuration
isup            −is a node up or down (openMosix kind of ping)
getmem          −shows logical free memory
getfree         −shows physical free mem
getutil         −display utilization
setyard         −sets a new Yardstick−value
setspeed        −sets a new (openMosix−) speed value
setdecay        −sets a new decay−interval





mosrun          −run a special configured command on a chooosen node
                syntax:
                        mosrun  [−h|openMosix_ID| list_of_openMosix_IDs] command [arguments]
```

The mosrun−command can be executed with several more comandline options. To ease this up there are several preconfigured run−scripts for executing jobs with a special (openMosix) configuration.

```
nomig           −runs a command which process(es) won't migrate
runhome         −executes a command locked to its home node
runon           −runs a command which will be directly migrated and locked to a node
cpujob          −tells the openMosix cluster that this is a cpu−bound process
iojob           −tells the openMosix cluster that this is a io−bound
process
nodecay         −executes a command and tells the cluster not to refresh the load−balancing stati
slowdecay       −executes a command with a slow decay interval for collecting load−balancing stat
fastdecay       −executes a command with a fast decay interval for collecting load−balancing stat



setpe           −manuell node configuration utility
                syntax:
                        setpe   −w −f   [hpc_map]
                        setpe   −r [−f  [hpc_map]]
                        setpe   −off

−w reads the openMosix configuration from a file (typically /etc/hpc.map)
−r writes the current openMosix configuration to a file (typically /etc/hpc.map)
−off turns the current openMosix configuration off


tune            openMosix calibration and optimizations utility.
                (for further informations review the tune−man page)
```

Additional to the /proc interface and the commandline−openMosix utilities (which are using the /proc interface) there is a pachted "ps" and "top" available (they are called "mps" and "mtop") which displays also the openMosix−node ID on a column. This is usefull for finding out where a specific process is currently being computed.

The administrator can have a overview about the current status of the cluster and its nodes with the "Mosix Cluster Information Tool PHP" which can be found at http://wijnkist.warande.uu.nl/mosix/ . (the path to the NODESDIR has to be adjusted to $NODESDIR="/proc/hpc/nodes/")

For smaller cluster it might also be usefull to use Mosixview which is a GUI for the most common administration tasks.

# Chapter 9. Tuning Mosix

## 9.1. Optimising Mosix

Login a normal terminal as root. Type

```
       setpe −r
```

which, if everything went right, will give you a listing of your /etc/mosix.map. If things did not go right, try

```
        setpe −w −f /etc/mosix.map
```

to set up your node. Then, type

```
       cat /proc/$$/lock
```

to see if your child processes are locked in your mode (1) or can migrate (0). If for some reason you find your processes are locked, you can change this with

```
        echo 0 > /proc/$$/lock
```

until you fix the problem. Repeat the whole configuration scheme for a second computer. The programs tune_kernel and prep_tune that Mosix uses to calibrate the individual nodes do not work with the SuSE distribution. However, you can fake it. First, bring the computer you want to tune and another computer with Mosix installed down to single user mode by typing

```
        init 1
```

as root. All other computers on the network should be shutdown if possible. On both machines, run the following commands:

```
        /etc/init.d/network start
        /etc/init.d/mosix start
        echo 1 > /proc/mosix/admin/quiet
```

This fakes prep_tune and the first parts of tune_kernel. Note that if you have a laptop with a pcmcia network card, you will have to run

```
        /etc/init.d/pcmcia start
```

instead of "/etc/init.d/network start". On the computer you want to tune, run tune_kernel and follow instructions. Depending on your machines, this can take a while − if you have a dog, this might be the time to go on that long, long walk you've always promised him. tune_kernel will create a program called "pg" in /root for testing reasons. Ignore it. After tuning is over, copy the contents of /tmp/overheads to the file /etc/overheads (and/or recompile the kernel). Repeat the tuning procedure for each computer. Reboot, enjoy Mosix, and don't forget to brag to your friends about your new cluster.

## 9.2. Where to place your files

# Chapter 10. Special Cases

## 10.1. Laptops and PCMCIA Cards

If you are installing Mosix on a Laptop, you will have to recompile the pcmcia sources, because they are distributed as a separate package and not as kernel modules. On a Suse 7.1 machine, in theory, this should work by installing the packages and then running

```
        rpm −ba /usr/src/packages/SPECS/pcmcia.spec
```

as described in the SuSE manual [on page 358 of the German edition]. However, the script tends to get confused by the location of the libraries of the vanilla version and the Mosix version, so after running the above line, you will have to go to the sources in /usr/src/kernel−modules/pcmcia and run

```
        make config
```

When prompted for the "Module install directory", change the default setting of "/lib/modules/2.2.19" to

```
        /lib/modules/2.2.19−mosix
```

Then run "make" and "make install", which should put the pcmcia modules in /lib/modules/2.2.19−mosix/pcmcia . Note that you must be running the Mosix kernel before you recompile the pcmcia sources.

## 10.2. Diskless nodes

At first you have to setup a DHCP−server which answers the DHCP−request for an ip−adress when a diskless client boots. This DHCP−Server (i call it master in this howto) acts additional as an NFS−server which exports the whole client−filesystems so the diskless− cluster−nodes (i call them slaves in this howto) can grab this FS (filesystem) for booting as soon as it has its ip. Just run a "normal"−MOSIX setup on the master−node. Be sure you included NFS−server−support in your kernel−configuration. There are two kinds (or maybe a lot more) types of NFS:

```
kernel−nfs
or
nfs−daemon
```

It does not matter which one you use but my experiences shows to use kernel−nfs in "older" kernels (like 2.2.18) and daemon−nfs in "newer" ones. The NFS in newer kernels sometimes does not work properly. If your master−node is running with the new MOSIX−kernel start with one filesystem as slave−node. Here the steps to create it: Calculate at least 300−500 MB for each slave. Create an extra directory for the whole cluster−filesystem and make a symbolic−link to /tftpboot. (The /tftpboot−directory or link is required because the slaves searches for a directory named /tftpboot/ip−adress−of−slave for booting. You can change this only by editing the kernel−sources) Then create a directory named like the ip of the first slave you want to configure, e.g. mkdir /tftpboot/192.168.45.45 Depending on the space you have on the cluster−filesystem now copy the whole filesystem from the master−node to the directory of the first slave. If you have less space just copy:

```
/bin
/usr/bin
/usr/sbin
/etc
/var
```

You can configure that the slave gets the rest per NFS later. Be sure to create empty directories for the mount−points. The filesystem−structure in /tftpboot/192.168.45.45/ has to be similar to / on the master.

```
/tftpboot/192.168.45.45/etc/HOSTNAME                    //insert the hostname of the slave
```

```
/tftpboot/192.168.45.45/etc/hosts                       //insert the hostname+ip of the slave
```

Depending on your distribution you have to change the ip−configuration of  the slave :

```
/tftpboot/192.168.45.45/etc/rc.config
/tftpboot/192.168.45.45/etc/sysconfig/network
/tftpboot/192.168.45.45/etc/sysconfig/network−scripts/ifcfg−eth0
```

Change the ip−configuration for the slave as you like.  Edit the file

```
/tftpboot/192.168.45.45/etc/fstab               //the FS the slave will get per NFScoresponding t
/etc/exports                                    //the FS the master will export to the slaves
```

e.g. for a slave fstab:

```
master:/tftpboot/192.168.88.222  /       nfs     hard,intr,rw    0 1
none    /proc   nfs     defaults        0 0
master:/root     /root   nfs     soft,intr,rw    0 2
master:/opt      /opt    nfs     soft,intr,ro    0 2
master:/usr/local        /usr/local      nfs     soft,intr,ro    0 2
master:/data/      /data nfs     soft,intr,rw    0 2
master:/usr/X11R6        /usr/X11R6      nfs     soft,intr,ro    0 2
master:/usr/share        /usr/share      nfs     soft,intr,ro    0 2
master:/usr/lib        /usr/lib      nfs     soft,intr,ro    0 2
master:/usr/include        /usr/include      nfs     soft,intr,ro    0 2
master:/cdrom        /cdrom      nfs     soft,intr,ro    0 2
master:/var/log  /var/log        nfs     soft,intr,rw    0 2
```

e.g. for a master exports:

```
/tftpboot/192.168.45.45         *(rw,no_all_squash,no_root_squash)
/usr/local                      *(rw,no_all_squash,no_root_squash)
/root                           *(rw,no_all_squash,no_root_squash)
/opt                            *(ro)
/data                           *(rw,no_all_squash,no_root_squash)
/usr/X11R6                      *(ro)
/usr/share                      *(ro)
/usr/lib                        *(ro)
/usr/include                    *(ro)
/var/log                        *(rw,no_all_squash,no_root_squash)
/usr/src                        *(rw,no_all_squash,no_root_squash)
```

If you mount /var/log (rw) from the NFS−server you have on central  log−file! (it worked very well for me. just "tail −f /var/log/messages" on  the master and you always know what is going on)

The cluster−filesystem for your first slave will be ready now. Configure the slave−kernel now. If you have the same hardware on your cluster you can reuse the configuration of the master−node. Change the configuration for the slave like the following:

```
CONFIG_IP_PNP_DHCP=y
and
CONFIG_ROOT_NFS=y
```

Use as less modules as possible (maybe no modules at all) because the  configuration is a bit tricky. Now (it is well described in the beowulf−howtos) you have to create a  nfsroot−device. It is only used for patching the slave−kernel to boot from NFS.

```
mknod /dev/nfsroot b 0 255
rdev bzImage /dev/nfsroot
```

Here "bzImage" has to be your diskless−slave−kernel you find it in  /usr/src/linux−version/arch/i386/boot after succesfull compilation.  Then you have to change the root−device for that kernel

```
rdev −o 498 −R bzImage 0
```

and copy the kernel to a floppy−disk

```
dd if=bzImage of=/dev/fd0
```

Now you are nearly ready! You just have to configre DHCP on the master.  You need the MAC−adress (hardware adress) of the network card of your  first slave.  The easiest way to get this adress is to boot the client with the already  created boot−floppy (it will fail but it will tell you its MAC−adress). If  the kernel was configured alright for the slave the system should come up from the floppy, booting the diskless−kernel, detecting its network−card and sending an DHCP− and ARP request. It will tell you its hardware adress during that moment! It looks like : 68:00:10:37:09:83. Edit the file /etc/dhcp.conf like the following sample:

```
option subnet−mask 255.255.255.0;
default−lease−time 6000;
max−lease−time 72000;
subnet 192.168.45.0 netmask 255.255.255.0 {
     range 192.168.45.253 192.168.45.254;
     option broadcast−address 192.168.45.255;
     option routers 192.168.45.1;
}
host firstslave
{
     hardware ethernet 68:00:10:37:09:83;
     fixed−address firstslave;
     server−name "master";
}
```

Now you can start DHCP and NFS with their init scripts:

```
/etc/init.d/nfsserver start
/etc/init.d/dhcp start
```

You got it!! It is (nearly) ready now!

Boot your first−slave with the boot−floppy (again). It should work now. Shortly after recognizing its network−card the slave gets its ip−adress from the DHCP−server and its root−filesystem (and the rest) per NFS.

You should notice that modules included in the slave−kernel−config must exist on the master too, because the slaves are mounting the /lib−directory from the master. So they use the same modules (if any).

It will be easier to update or install additional libraries or applications if you mount as much as possible from the master. On the other hand if all slaves have their own complete filesystem in /tftpboot your cluster may be a bit faster because of not so many read/write hits on the NFS−server.

You have to add a .rhost file in /root (for user root) on each cluster−member which should look like this:

```
node1   root
node2   root
node3   root
....
```

You also have to enable remote−login per rsh in the /etc/inetd.conf. You should have these two lines in it

if your linux−distribution uses inetd:

```
shell   stream  tcp     nowait  root    /bin/mosrun mosrun −l −z /usr/sbin/tcpd in.rshd −L
login   stream  tcp     nowait  root    /bin/mosrun mosrun −l −z /usr/sbin/tcpd in.rlogind
```

And for xinetd:

```
service shell
{
socket_type     = stream
protocol        = tcp
wait            = no
user            = root
server          = /usr/sbin/in.rshd
server_args     = −L
}
service login
```

```
{
socket_type     = stream
protocol        = tcp
wait            = no
user            = root
server          = /usr/sbin/in.rlogind
server_args     = −n
}
```

You have to restart inetd afterwards so that it reads the new configuration.

```
/etc/init.d/inetd restart
```

or There may be another switch in your distribution−configuration−utility where you can configure the security of the system. Change it to "enable remote root login". Do not use this in insecure environments!!! Use SSH instead of RSH! You can use MOSIXVIEW with RSH or SSH. Configuring SSH for remote login without password is a bit tricky. Take a look at the "HOWTO use MOSIX/MOSIXVIEW with SSH?" at this website. If you want to copy files to a node in this diskless−cluster you have now two possibilities. You can use rcp or scp for copying remote or you can use just cp and copy files on the master to the cluster−filesystem of one node. The following two commands are equal:

```
rcp /etc/hosts 192.168.45.45./etc
cp /etc/hosts /tftpboot/192.168.45.45/etc/
```

## 10.3. Very large clusters

# Chapter 11. Common Problems

## 11.1. My processes won't migrate

Where to find the reason ... checking on /proc/

## 11.2. setpe reports

## 11.3. I don`t see all my nodes

# Chapter 12. Other Programs

## 12.1. mexec

## 12.2. mosixview

### 12.2.1. Requirements for Mosixview

QT > 2.3.0 root rights ! rlogin and rsh (or ssh) to all cluster−nodes without password  the MOSIX−tools mosctl, migrate, runon, iojob, cpujob ...  (included in every MOSIX distribution)

### 12.2.2. Documentation on MOSIXVIEW

There is a full HTML−documentation on MOSIXVIEW included in every package (>=0.9). You find the startpage of the docu in your MOSIXVIEW installation directory in the following path: mosixview/mosixview/docs/en/index.html The RPM−packages have their installation directories in /usr/local/mosixview

### 12.2.3. Installation of the RPM−distribution

Download the latest version of MOSIXVIEW rpm−package for your linux−distribution Then just execute e.g.:

```
rpm −i mosixview−1.0.suse72.rpm
```

This will install the all binaries in /usr/bin To uninstall:

```
rpm −e mosixview
```

Installation of the source−distribution  Download the latest version of MOSIXVIEW and unzip+untar the sources and copy the tarball to e.g. /usr/local/.

```
gunzip mosixview−1.0.tar.gz
tar −xvf mosixview−1.0.tar
```

Automatic setup−script  Just cd to the mosixview−directory and execute

```
./setup [your_qt_2.3.x_installation_directory]
```

Manual compiling  Set the QTDIR−Variable to your actual QT−Distribution, e.g.

```
export QTDIR=/usr/lib/qt−2.3.0  (for bash)
or
setenv QTDIR /usr/lib/qt−2.3.0          (for csh)
```

Hints :  (from the testers of mosixview who compiled it on diffrent linux−distributions, thanks again)  Create the link /usr/lib/qt pointing to your QT−2.3.x installation  e.g. if QT−2.3.x is installed in /usr/local/qt−2.3.0

```
ln −s /usr/local/qt−2.3.0 /usr/lib/qt
```

Then you have to  set the QTDIR environment variable to

```
export QTDIR=/usr/lib/qt        (for bash)
or
setenv QTDIR /usr/lib/qt                (for csh)
```

There is no need to "make clean" and delete config.cache and Makefile because all versions >= 0.6 are already contains "cleaned" source−code. That means there are no precompiled binaries any more and (maybe) less problems to compile by yourself! // (If compiling fails because of not finding qwidget.h, qobject.h or any other header files you have to delete the files config.cache and Makefile and then configure+make. (happens on my RedHat−Cluster)) // After that the rest should work fine:

```
./configure
make
```

then do the same in the subdirectory mosixcollector, mosixload and mosixview_client.

```
cd mosixcollector
./configure
make
cd ..
cd mosixload
./configure
make
cd ..
cd mosixmem
./configure
make
cd ..
cd mosixhistory
./configure
make
cd ..
cd mosixview_client
./configure
make
cd ..
```

Copy all binaries to /usr/bin

```
cp mosixview/mosixview /usr/bin
cp mosixview_client/mosixview_client/mosixview_client /usr/bin
cp mosixcollector/mosixcollector_daily_restart /usr/bin
cp mosixcollector/mosixcollector/mosixcollector /usr/bin
cp mosixload/mosixload/mosixload /usr/bin
cp mosixload/mosixload/mosixmem /usr/bin
cp mosixload/mosixload/mosixhistory /usr/bin
```

And the mosixcollector init−script to your init−directory e.g.

```
cp mosixcollector/mosixcollector.init /etc/init.d/mosixcollector
or
cp mosixcollector/mosixcollector.init /etc/rc.d/init.d/mosixcollector
```

Now copy the mosixview_client binary on each of your cluster−nodes to /usr/bin/mosixview_client

```
rcp mosixview_client/mosixview_client your_node:/usr/bin/mosixview_client
```

You can now execute mosixview (cd .. to quit the subdirectory mosixview_client)

```
./mosixview/mosixview
```

(do not use the & to force mosixview in the background!) If the "make install" fails just copy the mosixview binary wherever you want or create a symbolic link from /usr/bin/install (or wherever install is) to /usr/bin/ginstall and "make install" again.

## 12.2.4. the main window

This picture shows the main application−window of MOSIXVIEW. The function will be explained in the following HowTo. (Click to enlarge)

MOSIXVIEW reads the /etc/mosix.map at startup and builds a raw with a lamp, a button, a slider, a lcd−number two progressbars and a some labels for each cluster−member. The green lights at the left are displaying the MOSIX−Id and the status of the cluster−node. Red if down, green for avaiable. The status can set to autorefresh with the checkbox like the other dynamic objects. If you click on a button displaying an host−name (or ip) a configuration−dialog will pop up. It default shows the MOSIX−Name and some buttons to execute the most common used "mosctl"−commands. (described later in this HowTo) Use the "nslookup−checkbox" to get even hostname+ip in the config−dialog. Do not enable this option if your cluster−nodes only have ip−adresses and no hostnames in DNS! With the "speed−sliders" you can set the MOSIX−speed for each host. The current speed is displayed by the lcd−number. The load−balancing of the whole cluster can be influenced by this values. Processes in a MOSIX−Cluster are migrating easier to a node with more MOSIX−speed than to nodes with less speed. Sure it is not the physically speed you can set but it the speed MOSIX "thinks" a node has. e.g. a cpu−intensive job on a cluster−node which speed is set to the lowest value of the whole cluster processes will search for a better processor for running on and migrate away easily. The progressbars in the middle gives an overview of the load on each cluster−member. It displays in percent so it does not represent exactly the load written to the file /proc/mosix/nodes/x/load (by MOSIX), but it should give an overview. The next progressbar is for the used memory the nodes. I shows the currently used memory in percent from the avaiable memory on the hosts (the label to the right displays the avaiable mem). How many CPUs your cluster have is written in the box to the right. The last line of the main windows contains a configuration button for "all−nodes". You can configure all nodes in your cluster similar by this option. How good the load−balancing works is displayed by the progressbar in the last line. 100% is very good and means that all nodes nearly have the same load.

## 12.2.5. the configuration−window

This dialog will popup if an "cluster−node"−button is clicked. If your all cluster−members have DNS−hostnames the "nslookup"−option in the main−window can set to "enabled". The hostname and the ip−adress will be shown, otherwise only the MOSIX−name will be displayed. The MOSIX−configuration of each host can be changed easily now. All commands will be executed per "rsh" or "ssh" on the remote hosts (even on the local node) so "root" has to "rsh" (or "ssh") to each host in the cluster without prompting for a password (it is well described in a beowulf documentation or on the HowTo's on this page how to configure it). The commands are:

```
automigration on/off
quiet yes/no
bring/lstay yes/no
exspel  yes/no
mosix start/stop
```

If the MOSIXVIEW−client is properly installed on the remote cluster−nodes click the "remote proc−box"−button to open the MOSIXVIEW−client (proc−box) from remote. xhost +hostname will be set and the display will point to your localhost. The client is executed on the remote also per "rsh" or "ssh". (the binary mosixview_client must be copied to e.g. /usr/bin on each host of the cluster) The MOSIXVIEW−client is a process−box for managing your programs. It is usefull to manage programs started and running local on the remote nodes. The client is also described later in this HowTo. If you are logged on your cluster from a remote workstation insert your local hostname in the edit−box below the "remote proc−box". Then the MOSIXVIEW−Client will be displayed on your workstation and not on the cluster−member you are logged on (maybe you have to set "xhost +clusternode" on your workstation). There is a history in the combo−box so you have to write the hostname only once.

### 12.2.6. the migrator−window

This dialog will popup if process from the processbox is clicked.

The MOSIXVIEW−migrator window displays all nodes in your MOSIX−cluster. This window is for managing one process (with additional status−information since version 0.7). By doubleclicking on an host from the list the process will migrate to this host. After a short moment the process−icon for the managed process will be green, which means it is running remote. The "home"−button sends the process to its home node. In this example the process already running local. With the "best"−button the process is send to the best avaiable node in your cluster. This migration is influenced by the load, speed, cpu's and what MOSIX "thinks" of each node. It maybe will migrate to the host with the most cpu's and/or the best speed. With the "kill"−button you can kill the process immediatly. To pause a program just click the "SIGSTOP"−button and to continue the "SIGCONT"−button. With the renice−slider below you can renice the current managed process (−20 means very fast, 0 normal and 20 very slow)

### 12.2.7. managing processes from remote

This dialog will popup if the "manage procs from remote"−button beneath the process−box is clicked This TabView displays processes that are migrated to the local host. The procs are coming from other nodes in your cluster and currently computed on the host mosixview is started on. Similar to the two buttons in the migrator−window the process is send home by the "goto home node"−button and send to the best avaiable node by the "goto best node"−button.

### 12.2.8. advanced−execution

If you want to start jobs on your cluster the "advanced execution"−dialog may help you.

Choose a program to start with the "run−prog" button (fileopen−icon) and you can specify how and where the job is started by this execution−dialog. There are several options to explain. the command−line You can specify additional commandline−arguments in the lineedit−widget on top of the window. how to start

```
−no migrationstart a local job which won't migrate
−run homestart a local job
−run onstart a job on the node you can choose with the "host−chooser"
−cpu jobstart a computation intensive job on a node (host−chooser)
−io jobstart a io intensive job on a node (host−chooser)
−no decaystart a job with no decay (host−chooser)
−slow decaystart a job with slow decay (host−chooser)
−fast decaystart a job with fast decay (host−chooser)
−parallelstart a job parallel on some or all node (special host−chooser)
```

### 12.2.9. the host−chooser

For all jobs you start non−local simple choose a host with the dial−widget. The MOSIX−id of the node is also displayed by a lcd−number. Then click execute to start the job. the parallel host−chooser You can set the first and last node with 2 spinboxes. Then the command will be executed an all nodes from the first node to the last node. You can also inverse this option.

### 12.2.10. the MOSIXVIEW−client

This process−box is really usefull for managing the processes running on your cluster. (MOSIXVIEW−client and the "local proc−box" are the same; you should install it on every cluster−node)

This processlist gives an overview what is running where. The second column displays the MOSIX−node ID of each process. 0 means local, all other values are remote nodes. Migrated processes are marked with a green icon and nonmoveable processes have a lock. By doubleclicking a process from the list the migrator−window will pop−up for managing e.g. migrating the process. If you click on the "manage procs from remote" button a new window will come up (the remote−procs windows) displaying the process currently migrated to this host. There are also options to send the remote processes away.

### 12.2.11. the MOSIXCOLLECTOR

The MOSIXCOLLECTOR is a daemon which should/could be started on one cluster−member. It logs the MOSIX−load of each node to the directory /tmp/mosixview/* These history log−files analyzed by MOSIXLOAD, MOSIXMEM and MOSIXHISTORY (as described later) gives an nonstop overview of the load, memory and processes in your cluster. There is one main log−file called /tmp/mosixview/mosix.load. Additional to this there are additional files in this directory to which the data is written. At startup MOSIXCOLLECTOR writes its PID (process id) to /tmp/mosixcollector.pid. It won't start if this file exist! The MOSIXCOLLECTOR−daemon restarts once a day (depending on when started) and saves the current history to /tmp/mosixview[date]/* These backups are done automatically but you can also trigger this manual. There is an option to write a checkpoint to the history. These checkpoints are graphically marked as a blue vertical line if you analyze the history log−files with MOSIXLOAD or MOSIXMEM. For example you can set a checkpoint when you start a job on your cluster and another one at the end.. Here is the explanation of the possible commandline−arguments:

```
mosixcollector −d//starts the collector as a daemon
mosixcollector −k//stops the collector
mosixcollector −c//stops the collector and deletes the history−files
mosixcollector −n//writes a checkpoint to the history
mosixcollector −r//saves the current history and starts a new one
mosixcollector −help//print out a short help
mosixcollector −h//print out a short help
```

You can start this daemon whith its init−script in /etc/init.d or /etc/rc.d/init.d. You just have to create a symbolic link to one of the runlevels for automatic startup. How to analyze the created logfiles is described in the following MOSIXLOAD−section.

### 12.2.12. MOSIXLOAD

This picture shows the graphical Log−Analyzer MOSIXLOAD

With MOSIXLOAD you can have a non−stop MOSIX−load history. The history log−files created by MOSIXCOLLECTOR are displayed in a graphically way so that you have a long−time overview what happened and happens on your cluster. MOSIXLOAD can analyze the current "online" logfiles but you can also open older backups of your MOSIXCOLLECTOR history logs by the filemenu. The logfiles are placed in /tmp/mosixview/* (the backups in /tmp/mosixview[date]/*) and you have to open only the main history file "mosix.load" to take a look at older load−informations. (the [date] in the backup directories for the log−files is the date the history is saved) The start time is displayed on the top/left and you have a full−day view in MOSIXLOAD (24 h). If you are using MOSIXLOAD for looking at "online"−logfiles (current history) you can enable the "refresh"−checkbox and the view will auto−refresh (or use the manual refresh−button). The load−lines are normally black if the load of one node is smaller 50. If the load increases to >50 the lines are drawn yellow and red if load is higher 80. These values are MOSIX−informations. MOSIXLOAD gets these informations from the files /proc/mosix/nodes/[mosix ID]/load. The X−button of each nodes calculates the nodes avarage MOSIX−load. Clicking it will open a small new window in which you get the avarage load−value and a graphic which displays it coloured (black ok, yellow critique, red alert). If there are checkpoints written to the load−history by the MOSIXCOLLECTOR they are displayed as a vertical blue line. You now can compare the load values at a certain moment much easier.

## 12.2.13. MOSIXMEM

This picture shows the graphical Log−Analyzer MOSIXMEM

With MOSIXMEM you can have a non−stop memory history similar to MOSIXLOAD. The history log−files created by MOSIXCOLLECTOR are displayed in a graphically way so that you have a long−time overview what happened and happens on your cluster. MOSIXMEM can analyze the current "online" logfiles but you can also open older backups of your MOSIXCOLLECTOR history logs by the filemenu. The logfiles are placed in /tmp/mosixview/* (the backups in /tmp/mosixview[date]/*) and you have to open only the main history file "mosix.load" to take a look at older load−informations. (the [date] in the backup directories for the log−files is the date the history is saved) The start time is displayed on the top/left and you have a full−day view in MOSIXMEM (24 h). If you are using MOSIXMEM for looking at "online"−logfiles (current history) you can enable the "refresh"−checkbox and the view will auto−refresh (or use the manual refresh−button). The displayed values are MOSIX−informations. MOSIXMEM gets these informations from the files

```
/proc/mosix/nodes/[mosix ID]/mem.
/proc/mosix/nodes/[mosix ID]/rmem.
/proc/mosix/nodes/[mosix ID]/tmem.
```

The X−button of each nodes calculates the nodes avarage MOSIX−mem. Clicking it will open a small new window in which you get the avarage mem−value. If there are checkpoints written to the load−history by the MOSIXCOLLECTOR they are displayed as a vertical blue line. You now can compare the load values at a certain moment much easier.

## 12.2.14. MosixHistory

MOSIXHISTORY displays the processlist from the past MOSIXHISTORY gives a detailed overview which process was running on which node. The MOSIXCOLLECTOR saves the processlist from the host the collector was started on every minute and you can browse this log−data with MOSIXHISTORY. You can easy change the browsing time in MOSIXHISTORY by the time−slider. The rest is nearly similar to

MOSIXLOAD and MOSIXMEM. MOSIXHISTORY can analyze the current "online" logfiles but you can also open older backups of your MOSIXCOLLECTOR history logs by the filemenu. The logfiles are placed in /tmp/mosixview/* (the backups in /tmp/mosixview[date]/*) and you have to open only the main history file "mosix.load" to take a look at older load−informations. (the [date] in the backup directories for the log−files is the date the history is saved) The start time is displayed on the top/left and you have a full−day view in MOSIXHISTORY (24 h).

## 12.3. mpi

## 12.4. mps

## 12.5. pmake

## 12.6. pvm

## 12.7. qps

# Chapter 13. Hints and Tips

## 13.1. Locked Processes

If for some reason you find your processes are always locked in your home node and you can't find the reason, you can put the following lines into your ~/.profile as a stop−gap measure to automatically enable migration:

```
        if [ −x /proc/$$/lock ]; then
           echo 0 > /proc/$$/lock
        fi
```

However, you should make an effort to find out what the problems is − see the Mosix FAQ at http://www.mosix.org for details.

## 13.2. Choosing your processes

You will probably want to test your setup before deciding which programs you want to enable migration for. For example, if you are running KDE2 on a slow machine and have a significantly faster machine has part of your Mosix cluster, you might find resource−hungry programs like kmail are migrated out. This is not a bad thing as such, however, it can lead to a brief moment when your writing is not displayed on the screen immediately.

# Appendix A. More Info

## A.1. Further Reading

## A.2. Links

*Mosix  Debian Howto Mosix Mandrake Linux  Terminal Server Project*

## A.3. Supporting Mosix

# Appendix B. Credits

Scot W. Stevenson

I have to thank Scot W. Stevenson for al the work he did on this HOWTO before I took over.  He made a great start for this document.

Assaf Spanier

worked together with Scott in drafting the layout and the  chapters of this howto.  and now promised to help me out with this document.

Matthias Rechenburg

Matthias Rechenburg should be thanked for the work he did on Mosixview and the accompaning documentation , which we included in this howto.

Jean−David Marrow

is the author of Clump/OS,  he contributed the documentation  on  his distribution to the Howto.

# Appendix C. GNU Free Documentation License

Version 1.1, March 2000

> Copyright (C) 2000 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111−1307 USA Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

## 0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other written document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

## 1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you".

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front−matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (For example, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License.

The "Cover Texts" are certain short passages of text that are listed, as Front−Cover Texts or Back−Cover

Texts, in the notice that says that the Document is released under this License.

A "Transparent" copy of the Document means a machine−readable copy, represented in a format whose specification is available to the general public, whose contents can be viewed and edited directly and straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup has been designed to thwart or discourage subsequent modification by readers is not Transparent. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard−conforming simple HTML designed for human modification. Opaque formats include PostScript, PDF, proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine−generated HTML produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

## 2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

## 3. COPYING IN QUANTITY

If you publish printed copies of the Document numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front−Cover Texts on the front cover, and Back−Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine−readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a publicly−accessible computer−network location containing a complete Transparent copy of the Document, free of added material, which the general network−using public has access to download anonymously at no charge using public−standard network protocols. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

## 4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has less than five).
C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
D. Preserve all the copyright notices of the Document.
E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
H. Include an unaltered copy of this License.
I. Preserve the section entitled "History", and its title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
K. In any section entitled "Acknowledgements" or "Dedications", preserve the section's title, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section entitled "Endorsements". Such a section may not be included in the Modified Version.
N. Do not retitle any existing section as "Endorsements" or to conflict in title with any Invariant Section.

If the Modified Version includes new front−matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties−−for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front−Cover Text, and a passage of up to 25 words as a Back−Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front−Cover Text and one of Back−Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

# 5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections entitled "History" in the various original documents, forming one section entitled "History"; likewise combine any sections entitled "Acknowledgements", and any sections entitled "Dedications". You must delete all sections entitled "Endorsements."

# 6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

## 7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, does not as a whole count as a Modified Version of the Document, provided no compilation copyright is claimed for the compilation. Such a compilation is called an "aggregate", and this License does not apply to the other self−contained works thus compiled with the Document, on account of their being thus compiled, if they are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one quarter of the entire aggregate, the Document's Cover Texts may be placed on covers that surround only the Document within the aggregate. Otherwise they must appear on covers around the whole aggregate.

## 8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License provided that you also include the original English version of this License. In case of a disagreement between the translation and the original English version of this License, the original English version will prevail.

## 9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

## 10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

## How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright (c) YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.1 or any later version published by the Free Software Foundation; with the Invariant Sections being LIST THEIR TITLES, with the Front−Cover Texts being LIST, and with the Back−Cover Texts being LIST. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have no Invariant Sections, write "with no Invariant Sections" instead of saying which ones are invariant. If you have no Front−Cover Texts, write "no Front−Cover Texts" instead of "Front−Cover Texts being LIST"; likewise for Back−Cover Texts.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.