Cluster bauen für Anfänger (Grundlagen)

Jens Weiße
jens.weisse@bsd-crew.de

29. Oktober 2005

Freie Software ist in Gefahr – Softwarepatente

Der Fortbestand Freier Software und auch vieler Klein - und Mittelständischer Unternehmen wird durch Software Patente akut bedroht. Selbst der Microsoft Gründer sieht diese Gefahr.

> *„Wenn manche Leute verstanden hätten, wie Patente erteilt werden würden, als die meisten der heutigen Ideen erfunden wurden, und wenn sie sich dann Patente geholt hätten, wäre unsere Branche heute im kompletten Stillstand.“Bill Gates (1991)*

- www.ffii.org
- www.nosoftwarepatents.com

Was kommt heute dran?

HPC Cluster

Konfiguration eines HPC Cluster

Software

Batchsysteme

fertige Lösungen / Quellen / weitere Informationen

Clusterarten

- High-Availability (HA) / Fail-Over Cluster – Hochverfügbarkeit
- Load-Balancing Cluster
- High-Performance Computing (HPC) – Hochleistungsrechnen

High Performance Computing (HPC) – Hochleistungsrechnen

Ziel: hohe/höchste Rechenleistung für:

- seti@home
- kompilieren
- Videos digitalisieren / CD oder DVD rippen und encodieren (ogg, mpeg, . . .)
- Wettervorhersage
- numerische Simulation (FEM, CFD, Visualisierung)
- . . .

alles was möglichste vielen Rechenoperationen je Sekunde erfordert

HPC – aber wie?

HPC Architekturen:

- Symmetric Multiprocessors (SMP)
  - ein Betriebssystem
  - gemeinsamer Speicher
  - teurer (ab 4 CPUs)
  - weniger skalierbar (max 256 CPUs mit SGI Altix)
- Vektorrechner
  (optimiert für Matrizen und Vektorenrechnung)
- Massively Parallel Processors (MPP) bzw. Cluster
  - jeder Knoten hat eigenen Speicher, Betriebssystem, I/O, Netzwerk (Kommunikation)
  - kein gemeinsamer Speicher
  - gut skalierbar (BlueGene/L mit 65536 CPUs mit 280,6 Teraflops / ASC Purple)

Einkaufszettel

möglichst:

- schnelle CPUs
- viele CPUs
- 2 oder 4 fach SMP
- schnelles Netzwerk mit geringen Latenzzeiten (Infiband, Myrinet, GigaBit Ethernet, FastEthernet, . . .)
- kleine Gehäuse
- (Energieeffizient oder eigenes Kraftwerk)
- . . .

Praxis: Alles was zwischen 0 und 1 unterscheiden kann.

Konfigurationsziel

- zentrale Nutzerverwaltung
- zentrale Arbeitsverzeichnisse und Softwareinstallation
- Anwendungssoftware
- Batchsystem
- (Monitoring: top, gkrellm, Ganglia, Nagios, MRTG, ...)
- ...

zentrale Nutzerverwaltung

- Passwörter synchronisiert werden
- UID müssen gleich sein (Dateirechte)

Wie:

- LDAP (mit Kerberos) (sicher, komplex)
- NIS (Network Information Service) (Sicherheitsprobleme, einfach)
- eigene Shell Skipte zur Synchronisierung der /etc/passwd, /etc/shadow, /etc/groups, /etc/hosts, ...

Beispiel: NIS

- Network Information Service
- von SUN als „Yellow Pages“ entwickelt
- passwd, hosts, group, ...
- NIS-Howto
- Managing NFS and NIS – O'Reilly Verlag

Hinweise:
NIS ist relativ unsicher und daher nicht für Netze mit „spielsüchtigen“ Nutzern geeignet. NIS liefert die Passwortdatenbank mit den verschlüsselten Passwörtern aus. Mit ausreichend viel Rechenleistung kann sich passende Passwörter berechnen (brute force). Sicherer ist LDAP mit Kerberos.

NIS Server I

1. NIS Server installieren (ypserv oder nis und portmapper)
2. portmapper starten

```
root@server:> /etc/init.d/portmapper start
```

3. alle Rechner in die „/etc/hosts“ aufnehmen

```
root@server:> cat /etc/hosts
127.0.0.1       localhost.localdomain   localhost
192.168.23.42   server.lit              server
192.168.23.50   cluster1.lit            cluster1
[...]
```

4. NIS Domainname wird aus /etc/defaultdomain genommen

```
root@server:> cat /etc/defaultdomain
lit
```

NIS Server II

5. NIS-Server einrichten (mit Strg+D und y bestätigen).

```
root@server:> /usr/lib/yp/ypinit -m
```

6. NIS-Maps (Ersatz/Ergänzung für Linux Standarddateien)

```
root@server:> ls -l /var/yp/lit
group.bygid     group.byname     hosts.byaddr
hosts.byname        [...]        passwd.byname
passwd.byuid    shadow.byname    ypservers
```

7. /var/yp/Makefile steuert welche Maps NIS verwendet.

NIS Server III

8. Zugriff auf den NIS-Server einschränken

```
root@server:> cat /etc/ypserv.securenets
# This line gives access to everybody. PLEASE ADJUST!
# netmask           network
255.255.255.0       192.168.23.0
```

9. NIS-Server starten

```
root@server:> /etc/init.d/nis start
oder
root@server:>  ypserv
(-d ... Debug-Modus)
```

NIS Server IV

10. Daten für NIS-Maps stammen normalerweise aus den Dateien in „/etc“. Lässt sich in „/var/yp/Makefile“ anpassen.

```
root@server:> cat /var/yp/Makefile
# These are the source directories for the NIS
# files; normally that is /etc but you may want
# to move the source for the password and group
# files to (for example) /var/yp/ypfiles. The
# directory for passwd, group and shadow is
# defined by YPPWDDIR, the rest is taken from
# YPSRCDIR.
YPSRCDIR =  /var/yp/src                 #/etc
YPPWDDIR =  /var/yp/src                 #/etc
```

Vorher die Dateien aus „/etc“ nach „/var/yp/src“ kopieren. (siehe Makefile)

NIS Server V

11. Änderungen an den Daten für die Maps (z.B. hosts)

```
root@server:> cd /var/yp
root@server:> vi src/hosts
root@server:> make
```

NIS Client I

1. NIS Client installieren (ypbind)
2. NIS-Server bekannt machen

```
root@client:> cat /etc/ypconf
ypserver     192.168.23.42
```

3. NIS-Domainnamen konfigurieren

```
root@client:> cat /etc/defaultdomain
lit
```

4. NIS-Client starten

```
root@client:> /etc/init.d/nis start
oder
root@client:> /etc/init.d/ypbind start
oder
root@client:> ypbind
```

NIS Client II

5. Passwortänderung für User

```
jens@client:> yppasswd
Changing password for jens
Old Password:
[...]
```

Verzeichnisse

Ziel:

- ein globales /home
- zentrale Verzeichnisse für Software
- lokale Verzeichnisse für Zwischenergebnisse
- (evtl. Knoten über NFS booten)

Varianten:

- NFS (Network File System)
- OpenAFS
- Coda Distributed File System
- PVFS (Parallel Virtual File System)
- Lustre Cluster File System

NFS Server

NFS Server installieren (nfsd oder nfs-server oder nfs-kernel-server)

```
root@server:> cat /etc/exports
/export/home/    192.168.23.*(rw,sync)
/export/work/    cluster4(rw,no_root_squash) (rw)
/export/appl/    (ro)
root@server:> /etc/init.d/nfs-server start
```

Hinweise:
man exports
Der NFS-Server kann bei vielen gleichzeitgen Schreib-/Leseoperationen zum Flaschenhals werden. Deswegen lieber lokale „scratch“ Verzeichnisse und in denen nach Jobende aufräumen. Oder PVFS nutzen

NFS Client I

- NFS Client installieren (nfs-common)
- mounten erfolgt über „/etc/fstab“ oder „autofs“ (automounter)
- Mountpoints anlegen (fstab) / Symlinks anlegen (autofs)

NFS Client II – /etc/fstab

```
root@client:> cat /etc/fstab
[...]
server:/export/home /import/home   nfs   \
                                    rw,bg,hard,intr 0 0
192.168.23.42:/export/work /work   nfs   \
                                    rw,bg,soft,intr 0 0
```

Mounten erfolgt beim booten bzw. manuell falls die Option „noauto“ verwendet wird. Weitere Optionen siehe „man mount“.

NFS Client III – /etc/fstab

```
root@client:> mount
/dev/hda1 on / type reiserfs (rw)
server:/export/home on /import/home type    \
                                nfs (rw,bg,hard,intr,...)
[...]
```

```
root@client:> ln -s /import/home /home
root@client:> ls -l /home
lrwxrwxrwx  1 root root [...] /home -> /import/home
```

NFS Client IV – autofs

autofs installieren

```
root@client:> cat /etc/auto.master
# Mountpoint     Konfigurationsdatei
/import_auto     /etc/auto.import_auto
/import_usb      /etc/auto.import_usb
```

```
root@client:> cat /etc/auto.import_auto
# Mountpoint Optionen     NFS-Exports
home         -soft,intr    192.168.23.42:/export/home
work         -soft,intr    server:/export/work
```

Mounten erfolgt automatisch beim Zugriff auf eine Datei/Verzeichnis unterhalb von /import_auto/

NFS Client V – autofs

```
root@client:> ln -s /import_auto/home /home
root@client:> ls -l /home/
lrwxrwxrwx  1 root root [...] /home -> /import_auto/home
```

```
root@client:> mount
/dev/hda1 on / type reiserfs (rw)
automount(pid23074) on /import_auto type autofs
server:/export/home on /import_auto/home type    \
                                  nfs (rw,soft,intr,...)
```

Wie einloggen?

Ziel:

- automatisiertes einloggen ohne Eingabe von Passwort oder Passphrase
- keine passwortlosen Accounts
- sicheres/verschlüsseltes einloggen
- OpenSSH oder Kerberos

3 gute Artikel über OpenSSH Linux-Magazin

- Teil 1
- Teil 2
- Teil 3

OpenSSH I

1. Schlüsselgenerierung (mit Passphrase)

```
jens@client:> ssh-keygen -t rsa -b 2048
```

2. öffentlichen Teil in die „authorized_keys“ aufnehmen (Da unser /home über NFS auf jedem Rechner verfügbar ist, braucht dies nur einmal im eigenen /home geschehen.)

```
jens@client:> cd .ssh
jens@client:> cat id_rsa.pub >> authorized_keys
```

3. testen ob OpenSSH diesen Public-Key zur Authentifizierung nutzt (muss nach Passphrase fragen)

```
jens@client:> ssh jens@cluster1
Enter passphrase for key '/home/jens/.ssh/id_rsa':
```

OpenSSH II

4a. ssh-agent einrichten und Schlüssel hinzufügen (Debian)

```
root@client:> cat /etc/X11/Xsession.options
[...]
use-ssh-agent
```

Der ssh-agent wird beim einloggen in X automatisch gestartet. Nur die Schlüssel muss man hinzufügen.

```
jens@client:> ssh-add /home/jens/.ssh/id_rsa
```

OpenSSH III

4b. ssh-agent einrichten und Schlüssel hinzufügen (normal)

```
root@client:> cat  /home/jens/.xinitrc
[...]
if [ -x /opt/kde3/bin/startkde ]; then
  exec ssh-agent /bin/bash -c    \
  "ssh-add < /dev/null &&        \
  /opt/kde3/bin/startkde &>      \
  $HOME/.kde-errors"
fi
```

Network Time Protocol (NTP) I

alle Rechner im Netz sollten die selbe Uhrzeit haben

- nutzt Zeitserver über Internet und/oder Funk (GPS, DCF-77, . . .)
- stellt die Zeit automatisch
- berechnet die Drift der Rechneruhr gegenüber dem Zeitnormal
- beachtet sogar die Signallaufzeiten
- für ein großes Cluster ist die Installation eines eigenen Zeitservers sinnvoll
- Dokumentation: /usr/share/doc/ntp-doc/html/index.html
- http://www.eecis.udel.edu/ mills/ntp.html

Network Time Protocol (NTP) II

```
root@server:> cat /etc/ntp.conf
server pool.ntp.org
server 127.127.1.0
fudge 127.127.1.0 stratum 13
restrict default kod notrap nomodify nopeer noquery
restrict 127.0.0.1 nomodify

logfile /var/log/ntpd
driftfile /var/lib/ntp/ntp.drift
statsdir /var/log/ntpstats/
statistics loopstats peerstats clockstats
filegen loopstats file loopstats type day enable
filegen peerstats file peerstats type day enable
filegen clockstats file clockstats type day enable
```

Network Time Protocol (NTP) III

Differenz zwischen Atomzeit und Rechnerzeit anzeigen lassen:

```
root@server:> ntpq -p
     remote           refid      st t when poll reach   delay   offset  jitter
==============================================================================
 LOCAL(0)        LOCAL(0)         13 l   61   64  377    0.000    0.000   0.008
*ntp0-rz.rrze.un .GPS.            1 u  955 1024  377    9.108    1.613   0.690
```

Message Passing Interface (MPI)

- Protokoll / Standard
- Datenaustausch zwischen den Knoten (kein SMP)
- Ziel: mehrere Knoten berechnen Teillösungen und tauschen die Daten an den „Schnittkanten“ aus
- Implementationen: LAM/MPI, MPICH, ...
- http://www.mpi-forum.org/

Anwendungsbeispiele für MPI

- MPI-Povray – Patch für den OpenSource Raytracer Povray
- BlenderMods – Erweiterung für Blender um MPI basiertes paralleles rendern
- MPI Toolbox for Octave (Octave ist eine Sprache für numerische Berechnungen)
- viele kommerzielle Simulations- und Berechnungswerkzeuge: Matlab/Simulink, Ansys, LS-Dyna, ...
- Trefferliste für die Suche nach MPI auf sourceforge.net

LAM/MPI I

- LAM (Local Area Multicomputer)
- Indiana University (USA)
- http://www.lam-mpi.org/

LAM/MPI II

1. Auf welchen Knoten soll LAM/MPI laufen?

```
jens@client:> cat  lamhosts_file
cluster1 cpu=2
cluster4 cpu=2
```

2. LAM/MPI Umgebung „booten“

```
jens@client:> lamboot lamhosts_file
LAM 6.5.9/MPI 2 C++/ROMIO - Indiana University
```

LAM/MPI III

3. Kontrolle ob LAM/MPI „gebootet“ hat

```
jens@client:> lamnodes
n0      cluster1:2
n1      cluster4:2
```

4. Programm starten

```
jens@client:> mpirun -np $Prozessanzahl $Programm
```

5. LAM/MPI anhalten

```
jens@client:> lamhalt
LAM 6.5.9/MPI 2 C++/ROMIO - Indiana University
```

MPICH I

- sehr portabel
- Basis für viele angepasste Versionen
- http://www-unix.mcs.anl.gov/mpi/mpich/

MPICH II

1. Machinefile

```
jens@client:> cat  nodes_1_4
# where n is the number of processors in an
# SMP.  The hostname should be the same as
# the result from the command "hostname"
cluster1:2
cluster4:2
```

2. Programm starten

```
jens@client:> mpirun -np $Prozessanzahl     \
                -machinefile nodes_1_4 $Programm
```

OpenMosix

- Linux Kernel Erweiterung
- Programme können zwischen den Knoten wandern – Loadbalancing (auf einem „Master“ starten und der Prozess migriert bei Bedarf auf freie oder schnellere Knoten)
- Cluster erscheint als SMP-Maschine
- lässt sich mit MPI kombinieren
- stabil für 2.4.x Kernel; in Entwicklung für 2.6.x Kernel
- nur IA-32 und Itanium (IA 64) CPUs unterstützt (Opteron in Planung)
- http://openmosix.sourceforge.net/
- The openMosix HOWTO

Übersicht

- distcc – verteilter C/C++ Compiler
- dvd::rip
- seti@home und viele weitere Projekte zum verteilten Rechnen
- ...

distcc I

- verteilter Kompiler für C und C++
- Client/Server
- wesentlich schneller als nur lokales kompilieren
- benötigt gleiche Kompiler auf allen Rechnern
- benötigt keine gemeinsames Dateisystem, synchronisierte Uhren, identische Bibliotheken oder Header Files
- keine Verschlüsselung oder Authentifizierung
- http://distcc.samba.org/

distcc II

Anleitung von Webseite distcc.samba.org

```
30-second instructions:
  For each machine, download distcc, unpack, and do
    ./configure && make && sudo make install
  On each of the servers, run distccd --daemon,
  with --allow options to restrict access.
  Put the names of the servers in your environment:
  export DISTCC_HOSTS='localhost red green blue'
  Build!
  cd ~/work/linux-2.4.19; make -j8 CC=distcc
```

Anforderung

- Starre Verteilung von Computern auf bestimmte Nutzer ist nicht optimal.
  - Urlaub
  - wechselnder Bedarf
  - 100 CPUs / 5 Nutzer entspricht 20 CPUs für jeden :-)
  - 20 CPUs / 100 Nutzer entspricht 0,2 CPUs für jeden :-(
- Ziel: maximale Auslastung der Resourcen (CPUs, Lizenzen)
- beachten der Prioritäten / Nutzergruppen
- faire Verteilung der Resourcen
- automatischer Start der Jobs falls Resource(n) verfügbar
- Information der Nutzer über Status der Jobs
- ...

Welche gibt es?

- Sun Grid Engine
- Torque+Maui
- mit Schikane: OpenPBS (unfrei: PBS Pro)
- unfrei: LFS (Load Sharing Facility – Platform Computing)
- ...

OpenPBS (/ Torque+Maui)

- PBS – Portable Batch System
- ursprünglich von der NASA entwickelt
- OpenPBS ist die ältere, ursprüngliche Version des PBS. PBS Pro ist eine verbesserte, kommerzielle Version.
- Torque ist die freie Weiterentwicklung des OpenPBS
- Maui ist ein erweiterter und auf Torque abgestimmter Scheduler (Bestimmt welcher Job in welcher Reihenfolge gestartet wird)
- Torque/Maui sind in der Konfiguration sehr ähnlich zu OpenBPS

OpenPBS

- OpenBPS besteht aus drei Komponenten:
  - Job Server pbs_serv (empfängt Jobs, Kommunikation, Überwachung der Jobs)
  - Job Scheduler pbs_sched (entscheidet welcher Job in welcher Reihenfolge startet)
  - Job Executors pbs_mom (startet Jobs auf Hosts/Knoten, meldet die Systemlast an den Server, oftmals nur MOM genannt)
- sehr gute (Kurz)anleitungen:
  Quick Guide to Setting Up OpenPBS und
  Building HPC Cluster with Linux . . . – IBM Redbook
- OpenPBS Admin Manual

http://www. . . .

- OSCAR (Open Source Cluster Application Resources)
- Ten Tips for Building Your First High-Performance Cluster
- Grundlagen über Cluster
- Buildung HPC Cluster with Linux . . . (IBM Redbook)
- The Aggregate
- RocksClusters.org
- Cluster Monkey
- High Performance Linux Clusters with OSCAR, Rocks, OpenMosix and MPI (O'Reilly-Verlag, ca 40 Euro)
- Beispielkapitel zu Managment Software – lesenswert
- http://sourceforge.net
- . . .

Freie Software ist in Gefahr – Softwarepatente

Der Fortbestand Freier Software und auch vieler Klein - und Mittelständischer Unternehmen wird durch Software Patente akut bedroht. Selbst der Microsoft Gründer sieht diese Gefahr.

> *„Wenn manche Leute verstanden hätten, wie Patente erteilt werden würden, als die meisten der heutigen Ideen erfunden wurden, und wenn sie sich dann Patente geholt hätten, wäre unsere Branche heute im kompletten Stillstand.“Bill Gates (1991)*

- www.ffii.org
- www.nosoftwarepatents.com

Fragen?

- Fragen/Kritik zum Vortrag, zu den Folien, . . . :
  jens.weisse@bsd-crew.de
- Fragen zu Linux: Linux User Group Dresden
- Fragen zu *BSD: BSD-Crew