Cluster@cs.pub.ro

Razvan Dobre , Alexandru Herisanu, Emil Slusanschi

http://cluster.grid.pub.ro

Agenda

- Ce este un sistem de batch?
- A story about hardware and software
- Software: Librarii, compilatoare si dependinte
- Arhitecturi avansate – cei 10%
- Profiling & Debugging
- The data connection

2

Ce este un sistem de batch?

- MPI, OpenMP, thread-uri si socketi – singurele variante
- Batch system (server farm), Grid si Cloud – almost the same

```
$ qsub -q [queue]  -b y [executable] -> $ qsub -q queue_1 -b y /path/my_exec

$ qsub -pe [pe_name] [no_procs] -q [queue]  -b n [script]
e.g: $ qsub -pe pe_1 4 -q queue_1 -b n my_script.sh
```

3

Batch system How-To

- Cel mai simplu script hello world

```
$ cat script.sh
#!/bin/bash
‘pwd‘/script.sh
$ chmod +x script.sh
$ qsub -q queue_1 script.sh
```

- In cazul MPI discutia este interesanta: openmpi, openmpi*1, sun-hpc, sun-hpc*1, intel
- Ai un set de servicii pe care trebuie sa le menajezi:
  - Tight Integration – node/core bindings
  - Loose Integration – mpiboot…

4

Batch system How-To

- `qstat –f, qdel [job_id], qstat –g c, qstat –t, qstat –j [job_id]`

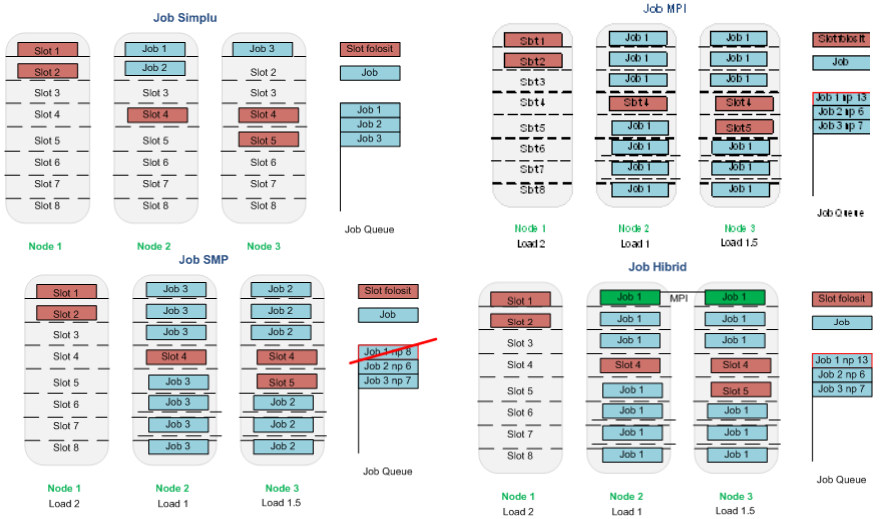
- Ce e un slot?
- Ce fel de aplicatie am?

```
[alexandru.herisanu@fep-53-1 ~]$ qstat -g c
CLUSTER QUEUE                   CQLOAD   USED    RES  AVAIL  TOTAL aoACDS  cdsuE
--------------------------------------------------------------------------------
all.q                             0.85      9      0    108    456    272     76
cnmsi-virtual.q                   -NA-      0      0      0     88      0     88
fs-dual.q                         0.00      0      0     16     16      0      0
fs-p4.q                           -NA-      0      0      0      0      0      0
ibm-cell-qs22.q                   -NA-      0      0      0     16      0     16
ibm-nehalem-8.q                   0.70      0      0      8     32      0     24
ibm-nehalem.q                     0.70      2      0     15     64      0     48
ibm-opteron.q                     0.94    109      0     46    168     12     12
ibm-quad-3.q                      0.83     11      0   2460   2760      0    300
ibm-quad.q                        0.79    198      0     26    224      0      8
[alexandru.herisanu@fep-53-1 ~]$
```

5

Batch system How-To

- Job simplu, SMP, MPI si hibrid



6

Batch system How-To

- Aaand … the environment variables OMP_NUM_THREADS
- Cate thread-uri am per procesor?
- Am impartit corect?
- Conteaza compilatorul?
- Cum functioneaza codul pe alte arhitecturi?

7

A story about hardware and software

- Exista un ghid al programatorului
- ClusterGuide v3.0 & v.3.1 – http://cluster.grid.pub.ro

Model	Processor Type	Sockets/Cores	Memory	Hostname
ICF IBM HS21 65 nodes	Intel Xeon E5405, 2 GHz	2/8	16 GByte	quad-wnXX.hpc-icf.ro
CNMSI VCluster 120 nodes	Intel Xeon X5570, 2.93 G	1/1*	1 GByte	cnmsi-wn

Model	Processor Type	Sockets/Cores	Memory	Hostname
IBM HS21 28 nodes	Intel Xeon E5405, 2 GHz	2/8	16 GByte	quad-wn
IBM HS22 4 nodes	Intel Xeon E5630, 2.53 GHz	2/16	32 GByte	nehalem-wn
IBM LS22 14 nodes	AMD Opteron	2/12	16 GByte	opteron-wn
IBM QS22 4 nodes	BECell Broadband	2/4	8 GByte	cell-qs
Fujitsu Esprimo 61 nodes	Intel P4 3 GHz	1/1	2 GByte	p4-wn
Fujitsu Celsius 2 nodes	Intel Xeon 3 GHz	2/2	2 GByte	dual-wn

The NCIT Cluster Resources User's Guide
Version 3.1

Release date: 7 April 2010

8

Infrastructura curenta

- Cuvantul magic: diversitate

  32 dual quad-core Xeon + **20 dual hex-core Opteron**

  + 4 dual PowerXCell 8i + ~~50 P4 HT + 32 dual Xeon~~

  = 642 cores

  (total: 918 cores / HPC 642 – Virtualization 232)
- GbE/Infiniband Interconnect
- Total storage of 36TB

HP-SEE

+ other "friendly clusters"

Software

- Sun Grid Engine 6.2u5 / cfEngine software provisioning
- Compilers
  - Ibm XL, Intel C/Fortran, PGI, SunStudio, gcc
- Debuggers
  - TotalView
- Profilers
  - VTune, Sun Studio Analyser
- Libraries
  - Intel MKL, NAG, Blas
- Tools
  - Code Saturne, Charm++, Gaussian09, OpenFoam, Paraview etc.

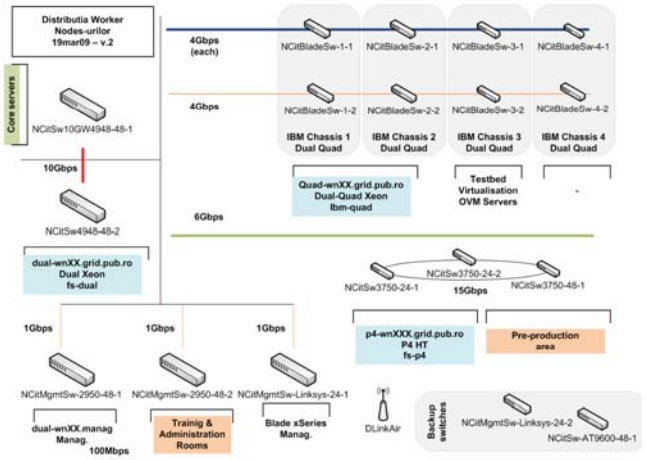
hadoop

lustre®

http://cluster.grid.pub.ro

Our Network

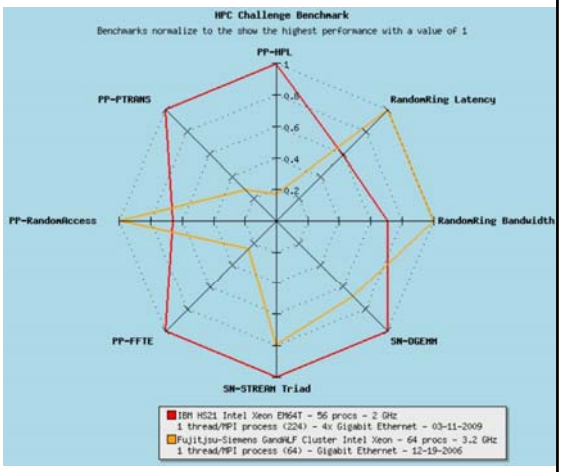
- full IPv6 stack (routed, not tunneled)
- Infiniband
- Dual Gigabit



- Jumbo frames
- Network stack configuration, bonding

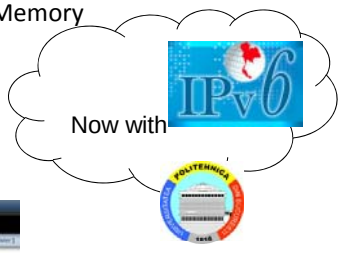
HPL Performance

- 1TFlops sustained for 230.000 dense linear systems
- Current work:
  - Use IBM XL C/C++ Compiler Suite
  - Use the Intel Cluster Toolkit Compiler Suite
  - Use optimized Intel MKL, MASS libraries
  - Optimize for LS22 Infiniband
  - Use OpenCL version of LINPACK for GP-GPU computing



Infrastructure Outlook

- GPGPU Programming in curricula probably in 1-2 years
- At least 4 NVidia Fermi Engines in IBM Blade Servers
- At least 22 Blades with x86/RISC computing architecutres & Infiniband
- RDMA over Infiniband GPU Memory to GPU Memory
- OpenCL & CUDA Programming
- Everything is a ~~Object~~ VM



Software: Librarii, compilatoare si dependinte

- Nu totul se rezuma la gcc si nu totul este open-source
- Ce e acela un linker?

```
[alexandru.herisanu@fep-53-1 openmpi-1.5]$ ./configure --help | more
`configure' configures Open MPI 1.5 to adapt to many kinds of systems.

Usage: ./configure [OPTION]... [VAR=VALUE]...

To assign environment variables (e.g., CC, CFLAGS...), specify them as
VAR=VALUE.  See below for descripti

Defaults for the options are specif

Configuration:
  -h, --help              display t
      --help=short        display c
      --help=recursive    display t
  -V, --version           display v
  -q, --quiet, --silent   do not pr
      --cache-file=FILE   cache tes
  -C, --config-cache      alias for
  -n, --no-create         do not cr
      --srcdir=DIR        find the

Installation directories:
  --prefix=PREFIX         install a
                          [/usr/loc
  --exec-prefix=EPREFIX   install a
                          [PREFIX]
```

```
Some influential environment variables:
CC          C compiler command
CFLAGS      C compiler flags
LDFLAGS     linker flags, e.g. -L<lib dir> if you have libraries in a
            nonstandard directory <lib dir>
LIBS        libraries to pass to the linker, e.g. -l<library>
CPPFLAGS    (Objective) C/C++ preprocessor flags, e.g. -I<include dir> if
            you have headers in a nonstandard directory <include dir>
CPP         C preprocessor
CXX         C++ compiler command
CXXFLAGS    C++ compiler flags
CXXCPP      C++ preprocessor
CCAS        assembler compiler command (defaults to CC)
CCASFLAGS   assembler compiler flags (defaults to CFLAGS)
F77         Fortran 77 compiler command
FFLAGS      Fortran 77 compiler flags
FC          Fortran compiler command
FCFLAGS     Fortran compiler flags
YACC        The `Yet Another C Compiler' implementation to use. Defaults to
            the first program found out of: `bison -y', `byacc', `yacc'.
YFLAGS      The list of arguments that will be passed by default to $YACC.
            This script will default YFLAGS to the empty string to avoid a
            default value of `-d' given by some make applications.
```

14

Software: Librarii, compilatoare si dependinte

- Write your own configure.sh
- Folositi diff pentru a gasi fisierele instalate

```
#!/bin/bash

CC="pgcc" \
CXX="pgCC" \
F77="pgf77" \
F90="pgf90" \
FC="pgf95" \
./configure \
   --prefix=/opt/libs/openmpi/openmpi-1.3.2_pgi-7.0.7/ \
--with-sge
```

```
find /opt > fis1
make install
find /opt > fis2
diff fis2 fis1 > MyProgram-Installed
```

```
%files
%defattr(-,root,root,-)
> /opt/Accelrys
> /opt/Accelrys/bin
> /opt/Accelrys/bin/exec
```

15

Software: Librarii, compilatoare si dependinte

- When it all goes wrong … it's only C (does not hurt much, does it?)

```
LD_LIBRARY_PATH, LDCONFIG,
CFLAGS="-I /home/heri/include",
LDFLAGS="-L /home/heri/lib -l mylib"
```

- autoconf si automake

```
It was created by configure, which was
generated by GNU Autoconf 2.61.  Invocatio

  $ ./configure --prefix=/opt/utils/octave

## --------- ##
## Platform. ##
## --------- ##

hostname = quad-wn15.grid.pub.ro
uname -m = x86_64
uname -r = 2.6.18-128.1.10.el5
uname -s = Linux
uname -v = #1 SMP Thu May 7 12:48:13 EDT 2009
```

```
configure:3104: gcc -E  conftest.c
conftest.c:10:28: error: ac_nonexistent.h: No such file or directory
configure:3110: $? = 1
configure: failed program was:
| /* confdefs.h.  */
| #define PACKAGE_NAME ""
| #define PACKAGE_TARNAME ""
| #define PACKAGE_VERSION ""
| #define PACKAGE_STRING ""
| #define PACKAGE_BUGREPORT ""
| #define OCTAVE_SOURCE 1
| #define _GNU_SOURCE 1
| /* end confdefs.h.  */
| #include <ac_nonexistent.h>
configure:3143: result: gcc -E
configure:3172: gcc -E  conftest.c
configure:3178: $? = 0
configure:3209: gcc -E  conftest.c
conftest.c:10:28: error: ac_nonexistent.h: No such file or directory
configure:3215: $? = 1
configure: failed program was:
```

16

We got your back like no other

- Daca e de la noi, in general functioneaza

```
$ module help

   To get the list of available modules type:

$ module avail
--------------------------- /opt/modules/modulefiles ---------------------------
apps/bullet-2.77                java/jdk1.6.0_23-32bit
apps/codesaturn-2.0.0RC1        java/jdk1.6.0_23-64bit
apps/gaussian03                 mpi/Sun-HPC8.2.1c-gnu
apps/gulp-3.4                   mpi/Sun-HPC8.2.1c-intel
```

- Environment-ul este incarcat automat din script

An available module can be loaded with
`$ module load [module name] -> $ module load compilers/gcc-4.1.2`

17

Mprun Framework

- The user perspective: vreau sa testez programul meu pe mai multe arhitecturi, folosind compilatoare diferite

```
$ mprun.sh -h
Usage: mprun.sh --job-name [job-name] --queue [queue-name] \
                --pe [Paralell Environment Name] [Nr. of Slots]   \
                --modules [modules to load] --script [My script]  \
                --out-dir [log dir] --show-qsub --show-script     \
                --batch-job

Example:
   mprun.sh --job-name MpiTest --queue ibm-opteron.q \
        --pe openmpi*1 3 \
        --modules "compilers/gcc-4.1.2:mpi/openmpi-1.5.1_gcc-4.1.2" \
        --script exec_script.sh \
        --show-qsub --show-script
```

18

Mprun Framework (2)

- Pasii de profiling si scalare sunt in general:
  - masor un baseline
  - fac **o singura modificare**, masor din nou
  - compar datele, explic
- Variabile de lucru: arhitectura, compilatorul, nr. de CPU-uri

```
mprun.sh --job-name MpiTest --queue ibm-opteron.q --pe openmpi 1 \
        --modules "compilers/gcc-4.1.2:mpi/openmpi-1.5.1_gcc-4.1.2" \
        --script exec_script.sh --show-qsub --show-script
mprun.sh --job-name MpiTest --queue ibm-nehalem.q --pe openmpi 2 \
        --modules "compilers/gcc-4.1.2:mpi/openmpi-1.5.1_gcc-4.1.2" \
        --script exec_script.sh --show-qsub --show-script
mprun.sh --job-name MpiTest --queue ibm-quad.q --pe openmpi*1 4 \
        --modules "compilers/gcc-4.1.2:mpi/openmpi-1.5.1_gcc-4.1.2" \
        --script exec_script.sh --show-qsub --show-script
```

19

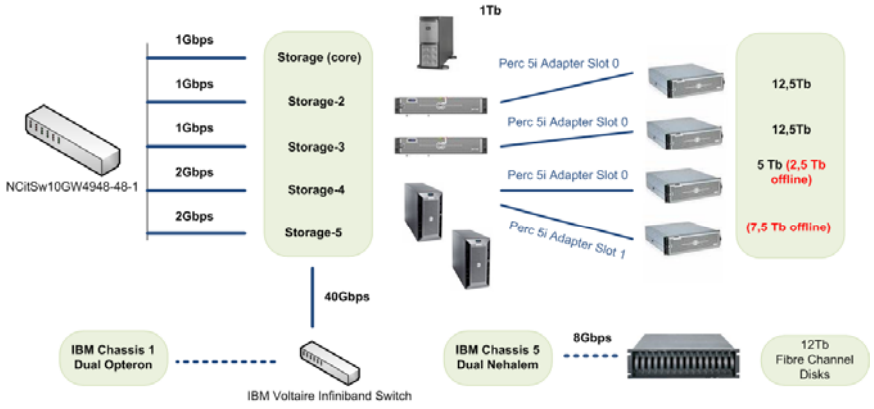
Arhitecturi avansate – cei 10%

- Arhitecturi diferite de procesoare (inclusiv CELL)
- Afinitate de procesor:
  - C – pthreads
  - MPI – MPI rank-map, SunGridEngine Core Binding
- Afinitate de memorie:
  - NUMA – libnuma
- Storage & I/O knowledge
  - Diverse subsisteme
  - NFS, LustreFS (sist. de fisiere distribuit)
  - Discuri locale, discuri pe fibra, discuri peste retea

20

Arhitecturi avansate – cei 10%

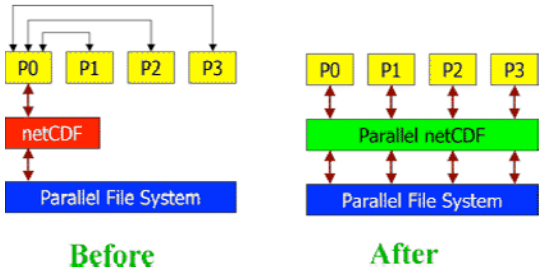
- MPI I/O, NFS, iSCSI
- How much? How fast? Where?



21

Arhitecturi avansate – cei 10%

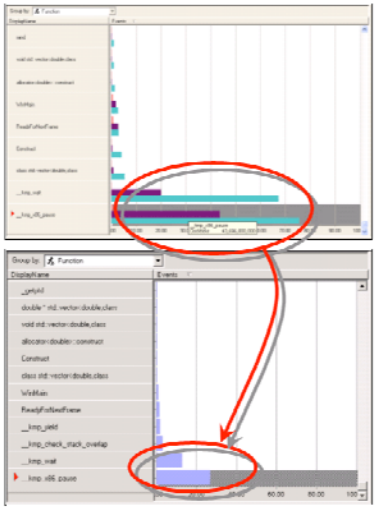
- Librariile te pot ajuta, dar de baza este utilizatorul
- NetCDF
- BLAS, Intel MKL

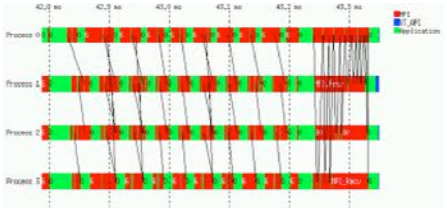


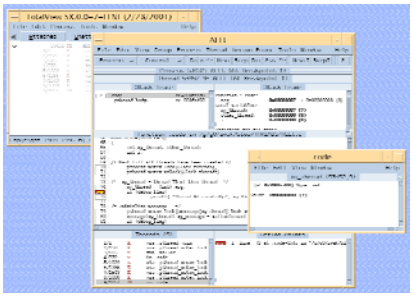
22

Profiling & Debugging

- Vtune, Sun Studio Analyser, MPI Tracing, HW Counters



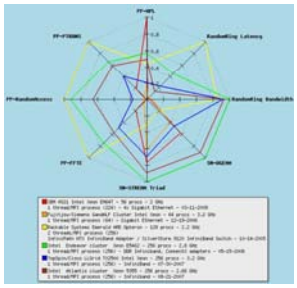




23

The data connection

- Sisteme de monitorizare:
  - MonAlisa http://monalisa.cern.ch
  - Cacti https://monitor.grid.pub.ro/cacti
    (autentificarea este cea de pe curs.cs)
- Modul de interconectare, trunk-uri, VLAN-uri, versiuni software se afla in cluster guide



24

HPC Related Lectures & Training @ CS

- Grid/HPC Initiative Summer school:
  - First GridInit was in **2004**
  - Usually debated grid middleware tasks
  - Fom 2008 the main focus is on developing **HPC Applications** using architectures with **multicore** processors
- Undergraduate Lectures:
  - Parallel Computing Algorithms and Data Structures, (Parallel) Computer Systems Architecture, Distributed Programming Languages
- Graduate Lectures:
  - Distributed Systems, Cluster & Grid Computing, High Performance Computing – Numerical Methods and Programming Techniques, Distributed Algorithms
- HPC Industry Training @cs.pub.ro:
  - Intel Multi-core Programming for Academia – 2007
  - IBM Basic and Advanced Cell Programming – 2008
  - IBM BlueGene Programming – 2009
  - Intel Parallelism Faculty – 2009
  - NVidia Cuda Programming – 2012

The End

?

cluster.grid.pub.ro
cs.pub.ro

26