Automated Testing of Installed Software

or so far, How to validate MPI stacks of an HPC cluster?

Xavier Besseron

HPC and Computational Science @ FOSDEM 2014
February 1, 2014

UNIVERSITÉ DU
LUXEMBOURG

Outline

1. Context & Motivations
2. Basic tests & Automation
3. ATIS
4. Main issues with MPI stacks
5. Quick overview / demo
6. Future work

Context: HPC clusters and Software

Large variety of software on HPC clusters

- Example: HPCBIOS
- Huge work to install, maintain, update, etc.

Tools to manage software

- EasyBuild: build, (re-)install
- Module: switch from one flavor to another

I counted 2211 EasyConfig files in EasyBuild

Example: HPC platform of University of Luxembourg

General statistics

- 2 clusters: Chaos and Gaia
- providing 1115 modules
- 376 different software/libraries
- 25 different flavors of zlib
- 15 different flavors of GCC
- 10 different flavors of GROMACS, OpenBLAS, ScaLAPACK
- 9 different flavors of WRF
- ...

⇒ explosion of the number of available software

Let's focus on MPI stacks

On Gaia cluster at University of Luxembourg

- 4 MPI families: OpenMPI, MVAPICH2, MPICH, IntelMPI
- 5 versions of OpenMPI: `1.4.5 1.6.3 1.6.4 1.6.5 1.7.3`
- 3 versions of MVAPICH2: `1.7 1.8.1 1.9`
- 3 versions of MPICH: `2-1.1 3.0.4 3.0.3`
- 8 versions of IntelMPI: `3.2.2.006 4.0.0.028 4.0.2.003 4.1.0.027 4.1.0.030 4.1.1.036 4.1.2.040 4.1.3.045`
- over 14 toolchains

⇒ 31 different modules provide MPI

And so what?	Some are not working out-of-the-box	
Why?	Let's try to find out	
What can we do?	~~Spam/complain to the sysadmins~~	Fix it!

How to test an MPI stack?

- Check for binaries

```
which mpicc mpirun
```

- Compile and run a small example

```
mpicc hello.c -o hello
mpirun -np 2 -machinefile <hostfile> hello
```

- Compile and run micro-benchmarks

```
tar -xzf osu-micro-benchmarks-3.9.tar.gz
cd osu-micro-benchmarks-3.9
./configure && make
cd mpi/pt2pt
mpirun -np 2 -machinefile <hostfile> osu_bw
mpirun -np 2 -machinefile <hostfile> osu_latency
```

- Check the performance is correct
- Run HPL?
- ...

How to test many MPI stacks?

Repeat the previous slides multiple times!

How to test many MPI stacks?

Repeat the previous slides multiple times!

- Make a script that test one MPI stack
- List the MPI stacks you want to test
- Run the script for all of them
- Collect data from all the tests
- Present the results in a synthetic way
- Repeat all this periodically

⇒ ATIS framework (Automated Testing of Installed Software)

Not reinventing the wheel!

Based on existing testing framework: CMake Cross-platform Make CDash

CTest

- Testing tool distributed as a part of CMake
- Automates updating, configuring, building, testing, performing memory checking, performing coverage
- Submits results to a CDash or Dart dashboard system

CDash

- Open source, web-based software testing server
- Aggregates, analyzes and displays the results of software testing
- Nice feature: can spam the sysadmins when tests fail

But also Shell script, R, numdiff, cron, ...

ATIS Current status

Current focus

- Only on MPI testing
- Only on general behavior of MPI
- Only testing a couple of nodes, i.e. not the whole cluster

User-oriented testing

- Run in the same environment as a user
- Try to mimic what a normal user would do

Source code `https://github.com/besserox/ATIS`

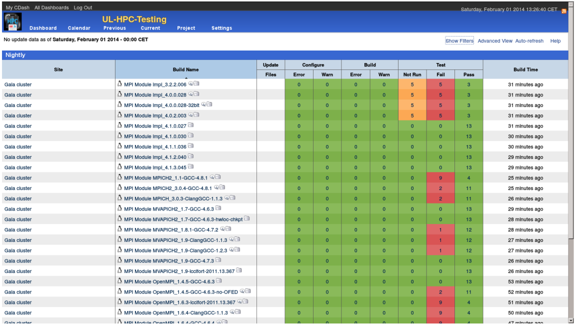
- About 15 files
- 247 lines of CMake/CTest
- 212 lines of Bash
- 98 lines of R

Main issues with MPI stacks

- Configuration issues
  - specific connector (i.e. `oarsh` instead of `ssh`)
  - InfiniBand interface
  - ...
- Dynamic libraries issues,
  i.e. `LD_LIBRARY_PATH` not set properly
  - for MPI libraries itself
  - for other dependencies (hwloc, cuda, ...)
- Bug in the MPI stacks
  - bashism in IntelMPI 3.X
  - ...
- Performance issues
  - need better tuning?

Quick Demonstration / Overview

HPC @ Uni.lu CDashboard



Saturday, February 01 2014 13:26:40 CET

UL-HPC-Testing

No update data as of **Saturday, February 01 2014 - 00:00 CET**

Nightly

Site	Build Name	Update	Configure		Build		Test			Build Time
		Files	Error	Warn	Error	Warn	Not Run	Fail	Pass	
Gala cluster	MPI Module Impl_3.2.2.006		0	0	0	0	5	5	3	31 minutes ago
Gala cluster	MPI Module Impl_4.0.0.028		0	0	0	0	5	5	3	31 minutes ago
Gala cluster	MPI Module Impl_4.0.0.028-32bit		0	0	0	0	5	5	3	31 minutes ago
Gala cluster	MPI Module Impl_4.0.2.003		0	0	0	0	5	5	3	31 minutes ago
Gala cluster	MPI Module Impl_4.1.0.027		0	0	0	0	0	0	13	31 minutes ago
Gala cluster	MPI Module Impl_4.1.0.030		0	0	0	0	0	0	13	30 minutes ago
Gala cluster	MPI Module Impl_4.1.1.036		0	0	0	0	0	0	13	30 minutes ago
Gala cluster	MPI Module Impl_4.1.2.040		0	0	0	0	0	0	13	29 minutes ago
Gala cluster	MPI Module Impl_4.1.3.045		0	0	0	0	0	0	13	29 minutes ago
Gala cluster	MPI Module MPICH2_1.1-GCC-4.8.1		0	0	0	0	0	9	4	25 minutes ago
Gala cluster	MPI Module MPICH2_3.0.4-GCC-4.8.1		0	0	0	0	0	2	11	25 minutes ago
Gala cluster	MPI Module MPICH_3.0.3-ClangGCC-1.1.3		0	0	0	0	0	2	11	26 minutes ago
Gala cluster	MPI Module MVAPICH2_1.7-GCC-4.6.3		0	0	0	0	0	0	13	29 minutes ago
Gala cluster	MPI Module MVAPICH2_1.7-GCC-4.6.3-hwloc-chkpt		0	0	0	0	0	0	13	28 minutes ago
Gala cluster	MPI Module MVAPICH2_1.8.1-GCC-4.7.2		0	0	0	0	0	1	12	28 minutes ago
Gala cluster	MPI Module MVAPICH2_1.9-ClangGCC-1.1.3		0	0	0	0	0	1	12	27 minutes ago
Gala cluster	MPI Module MVAPICH2_1.9-ClangGCC-1.2.3		0	0	0	0	0	1	12	27 minutes ago
Gala cluster	MPI Module MVAPICH2_1.9-GCC-4.7.3		0	0	0	0	0	0	13	26 minutes ago
Gala cluster	MPI Module MVAPICH2_1.9-iccifort-2011.13.367		0	0	0	0	0	0	13	26 minutes ago
Gala cluster	MPI Module OpenMPI_1.4.5-GCC-4.6.3		0	0	0	0	0	0	13	53 minutes ago
Gala cluster	MPI Module OpenMPI_1.4.5-GCC-4.6.3-no-OFED		0	0	0	0	0	2	11	52 minutes ago
Gala cluster	MPI Module OpenMPI_1.6.3-iccifort-2011.13.367		0	0	0	0	0	9	4	51 minutes ago
Gala cluster	MPI Module OpenMPI_1.6.4-ClangGCC-1.1.3		0	0	0	0	0	9	4	50 minutes ago
Gala cluster	MPI Module OpenMPI_1.6.4-GCC-4.6.4		0	0	0	0	0	9	4	47 minutes ago

Future directions

- Test other software/features
  - Checkpoint/Restart of a process using BLCR
  - ...

- Test features specific to a given MPI stack
  - alternative launcher (e.g. `mpirun_rsh` for MVAPICH2)
  - disable InfiniBand
  - distributed Checkpoint/Restart of an MPI job

- More reliable detection of performance issues
  - how to tolerate temporary variation of the performance?

Any feedback?

Thank you for your attention!

- Any feedback, comments, questions?
- New ideas or features?