CERN

HPC User Meeting April 2018

Pablo Llopis, Nils Høimyr, Dan van der Ster, Maria Alandes

Agenda

- Introduction
- Scratch space improvements
  - Kernel cephfs mounts
  - CephFS backend tuning
- Windows HPC to Linux HPC migration
- Extended credentials for EOS access (AUKS)
- Profiling tools
- Near future
  - Scratch space area migration
  - Partitions re-organization
- User feedback

Introduction

- Update on HPC service since meeting last year
- Please refer to the presentation from May 2018 for technical details on SLURM HPC service and cluster
- We also refer to the regular HTC **batch service** on HTCondor in this meeting

High Performance Computing (HPC)

- Applications and use cases that do not fit the standard batch High Throughput Computing (HTC) model.
  - E.g. parallel MPI applications requiring 32-2000 cores for a single job
- Other applications all should go to our HTC batch environment (HTCondor)
- ~180 k cores
  - 1-8 cores for regular jobs
  - 16, 24, 32 and 48 core nodes for large jobs
  - Special “BigMem” facility for users with special requirements, e.g. CST and engineering applications
  - Ramping up 24 cores/120Gb batch nodes for BE/ABP use cases
  - Ref. KB0004192 for more information

HPC service for accelerator and technology sector

- Batch HPC facility using SLURM “HPC-batch”
  - 2 Infiniband clusters 2x72 20 cores nodes
  - Older “batch” cluster with 16 core batch nodes with low-latency 10Gb Ethernet interconnects
- The following slides will focus on the SLURM HPC batch facility

HPC Storage with CephFS

- Ceph is a software-defined storage used at CERN for 5 years
  - Network-attached block devices for OpenStack
  - Object storage with S3 (Amazon-compatible HTTP storage service)
  - NFS/Lustre-like filesystem with CephFS
- We invested significant efforts with the Ceph developers to validate CephFS for HPC use-cases
  - World first entry for CephFS in IO-500 list presented at SuperComputing 2018
  - Validating re-implementation of O_LAZY, a POSIX extension to optimize parallel IO
- Currently interested in benchmarking/tuning IO-bound HPC applications
  - Project with climate researchers in Trieste
  - Looking for more applications… some ideas?

HPC scratch space update

- Home scratch directories on /hpcscratch
- Faster scratch area on /bescratch with tuned CephFS in pilot use for a while
- CentOS 7.6 and later includes support for the cephfs kernel mount, improved performance compared to Ceph-FUSE

HPC scratch space migration

- We plan to migrate the default /hpcscratch user home environment to the new, faster, CephFS storage
- Migration of login user environment to new /hpcscratch planned for **May**.
- Only the first level of folders/directories to the new home
- Old /hpcscratch to be available as /hpcscratch_old after the update
- **Please delete** inactive project folders in your HPC scratch area or **archive** projects to EOS.

Extended Credentials

- **Today**: Existing jobs do not have Kerberos credentials
  - This is the reason you can’t copy back to EOS in your job submission file.

- **Next week**: Jobs **will have kerberos credentials** by default.
  - You may use “*eos cp*” to copy data back.
  - Your job’s credentials will be renewed by up to a week automatically. More than a week is not possible due to CERN’s authentication system setup (without compromising security).

Extended Credentials

- Known issue that could affect some users. Ref KB0006097

> "sbatch: error: spank-auks: cred forwarding failed : auks cred : input buffer is too large"

- This could happen if you're on many e-groups. Your **job should continue working as usual**, but without the ability to perform "eos cp".
- Let us know if that happens to you!
  We think it's possible for us to find a solution to this issue. Since it requires some effort, we'll prioritize this task according to your feedback.

Profiling Tools

- Still a Work In Progress

- Different levels of profiling
  - CPU performance profiling
  - **Computation vs MPI communication**
  - I/O Profiling

Profiling Tools

- Computation vs MPI Communication
- When using Intel MPI, Intel Parallel Studio tools may be used
  - Limited support for MVAPICH, but it still works.
- Intel Trace Analyzer and Collector is an easy low-hanging fruit

Instead of loading the normal way with:

```
module load mpi/openmpi/3.0.0
```

You load Intel MPI with:

```
source /cvmfs/projects.cern.ch/intelsw/psxe/linux/18-all-setup.sh
```

Profiling Tools

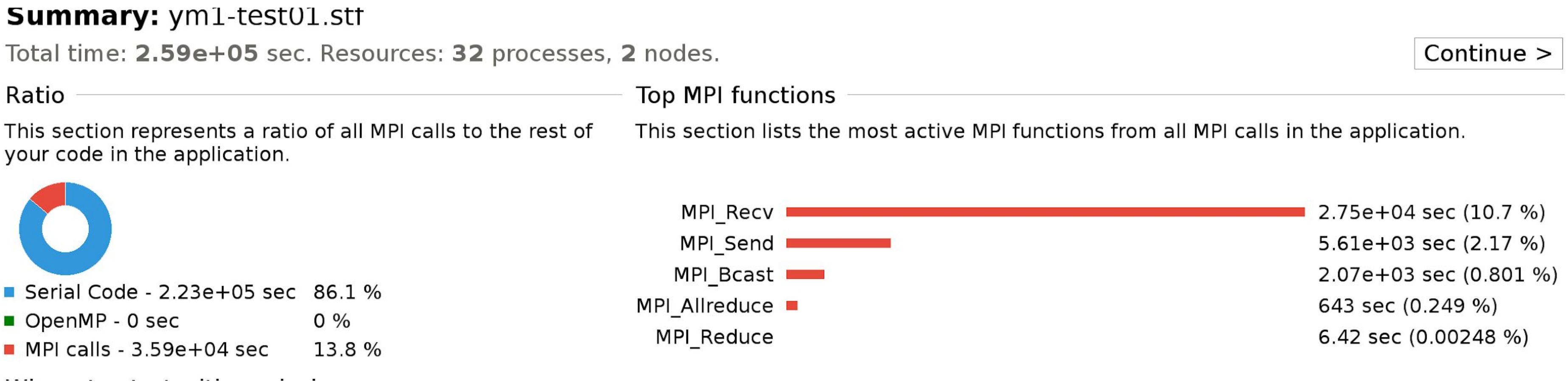
```
#!/bin/bash

#SBATCH --partition batch-long
#SBATCH --time 1-23:00:00
#SBATCH -N 6
#SBATCH --cpus-per-task 16
#SBATCH --exclusive

source /cvmfs/projects.cern.ch/intelsw/psxe/linux/18-all-setup.sh
export OMP_NUM_THREADS=8

mpirun -trace -np 12 ./fds_impi_intel_linux_64_i2018 test.fds 2> out.out
```

Profiling Tools: Trace Analysis

- Results in an **.stf** file being written (and many other files).
- Make sure you are using X11 forwarding! Ref KB0005052

→ **traceanalyzer simulation.stf**

Profiling Tools: Trace Analysis



Summary: ym1-test01.stf

Total time: **2.59e+05** sec. Resources: **32** processes, **2** nodes.

Continue >

Ratio

This section represents a ratio of all MPI calls to the rest of your code in the application.

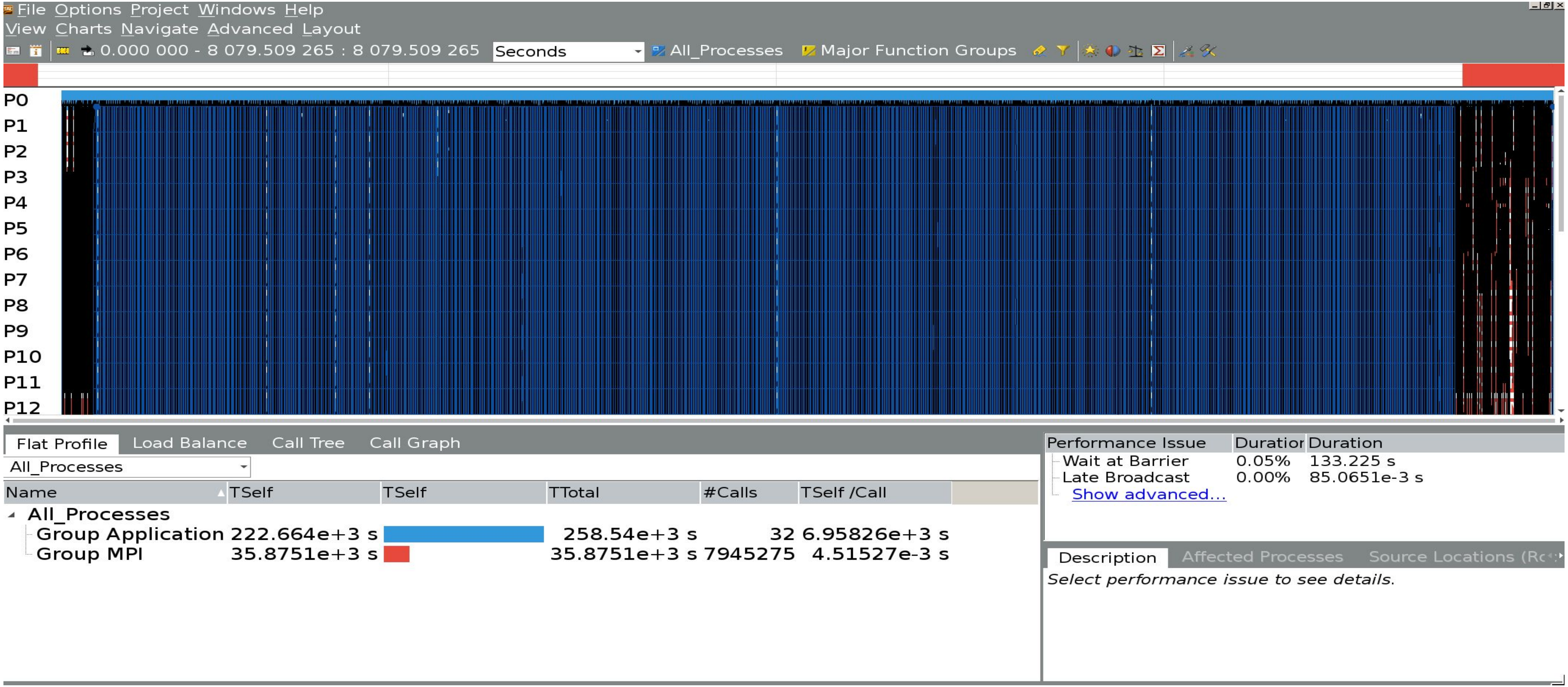
🟦 Serial Code - 2.23e+05 sec	86.1 %
🟩 OpenMP - 0 sec	0 %
🟥 MPI calls - 3.59e+04 sec	13.8 %

Top MPI functions

This section lists the most active MPI functions from all MPI calls in the application.

MPI_Recv	2.75e+04 sec (10.7 %)
MPI_Send	5.61e+03 sec (2.17 %)
MPI_Bcast	2.07e+03 sec (0.801 %)
MPI_Allreduce	643 sec (0.249 %)
MPI_Reduce	6.42 sec (0.00248 %)

Profiling Tools: Trace Analysis



0.000 000 - 8 079.509 265 : 8 079.509 265

Seconds

All_Processes

Major Function Groups

All_Processes

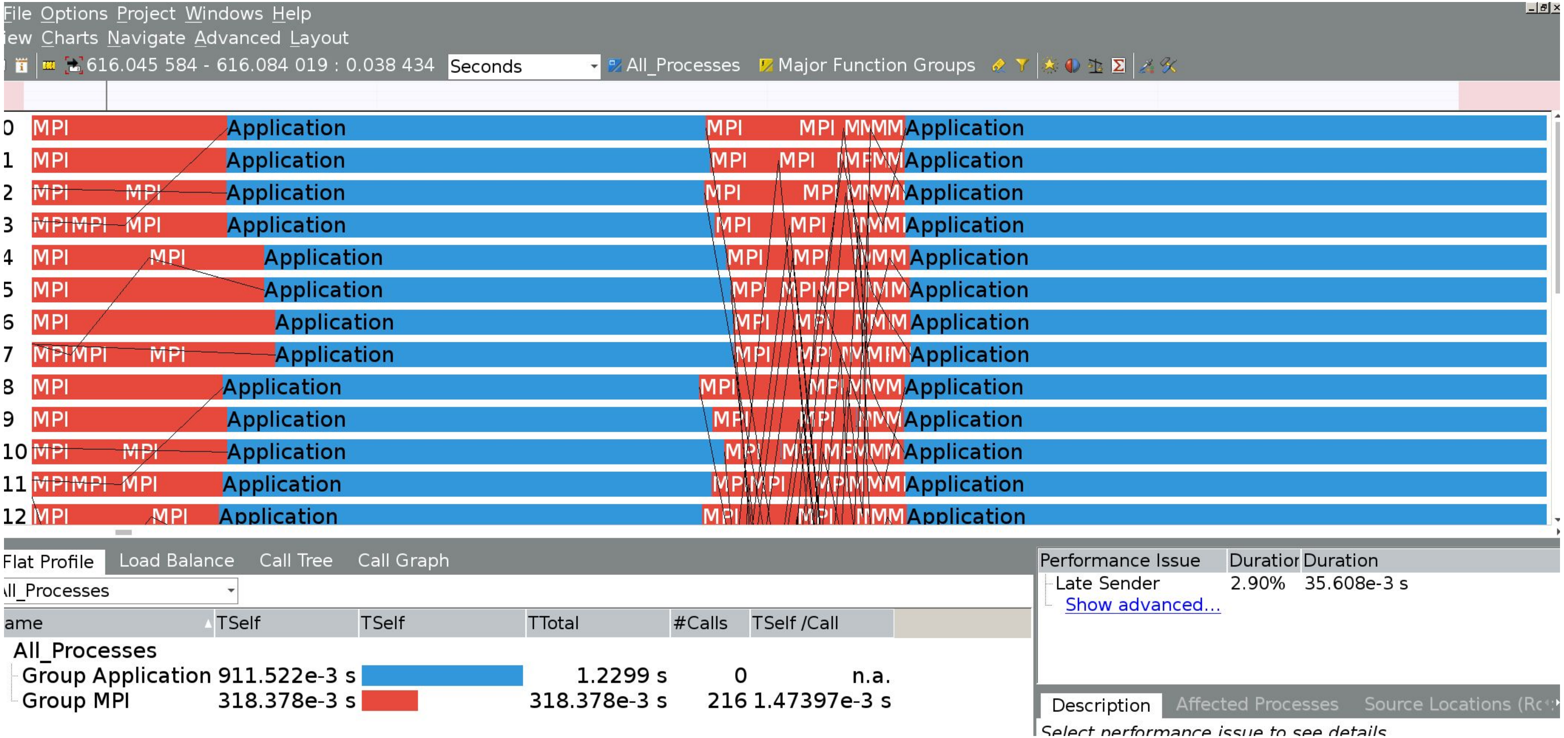
Name	TSelf	TSelf	TTotal	#Calls	TSelf /Call
All_Processes					
Group Application	222.664e+3 s	🟦	258.54e+3 s	32	6.95826e+3 s
Group MPI	35.8751e+3 s	🟥	35.8751e+3 s	7945275	4.51527e-3 s

Performance Issue	Duratior	Duration
Wait at Barrier	0.05%	133.225 s
Late Broadcast	0.00%	85.0651e-3 s

Show advanced...

Select performance issue to see details.

Profiling Tools: Trace Analysis


File Options Project Windows Help
iew Charts Navigate Advanced Layout
616.045 584 - 616.084 019 : 0.038 434
Seconds
All_Processes
Major Function Groups
0 MPI Application MPI MPI MMMM Application
1 MPI Application MPI MPI MFMM Application
2 MPI MPI Application MPI MPI MMM Application
3 MPI MPI MPI Application MPI MPI MMM Application
4 MPI MPI Application MPI MPI MMM Application
5 MPI Application MPI MPI MPI MMM Application
6 MPI Application MPI MPI MMM Application
7 MPI MPI MPI Application MPI MPI MMIM Application
8 MPI Application MPI MPI MMM Application
9 MPI Application MPI MPI MMM Application
10 MPI MPI Application MPI MPI MPMMM Application
11 MPI MPI MPI Application MPI MPI MPI MMM Application
12 MPI MPI Application MPI MPI MMM Application
Flat Profile
Load Balance
Call Tree
Call Graph
ll_Processes

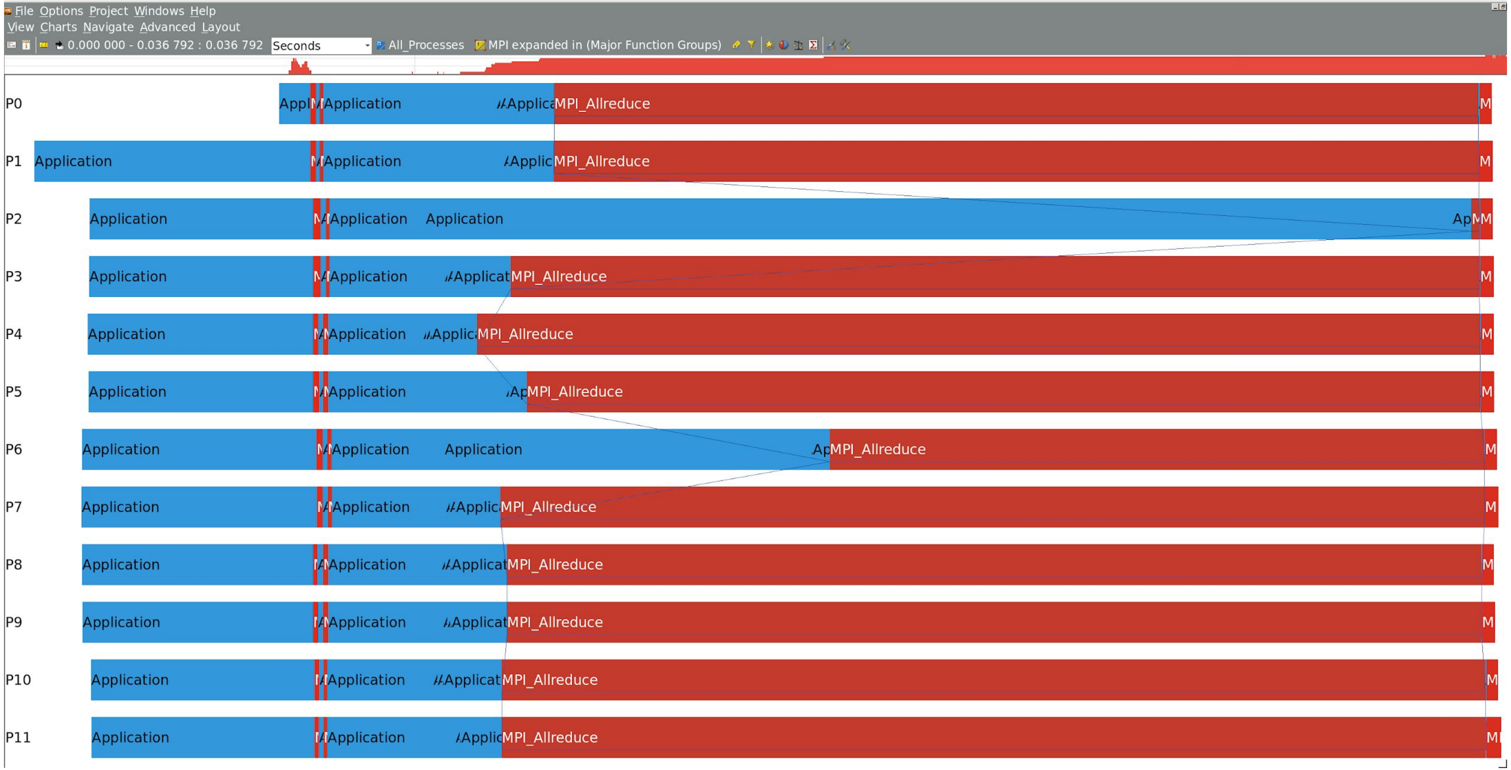
ame	TSelf	TSelf	TTotal	#Calls	TSelf /Call
All_Processes					
Group Application	911.522e-3 s	🟦	1.2299 s	0	n.a.
Group MPI	318.378e-3 s	🟥	318.378e-3 s	216	1.47397e-3 s

Performance Issue	Duratior	Duration
Late Sender	2.90%	35.608e-3 s
Show advanced...		

Description
Affected Processes
Source Locations (Rc

Select performance issue to see details

Profiling Tools: Trace Analysis



SLURM partition review

- Current partitions: “*batch-short*” and “*batch-long*”, as well as “*be-long*” and “*be-short*”
- Plan to re-organize these partitions as follows.
  - *batch-short*
  - *batch-long*
  - *inf-long* (combination of old be-short + be-long)
- Time limitation of 1 weeks for *long* partitions, which would simplify transparent upgrades.

SLURM partition review

- Advantages:
  - Less “thinking” about where to submit to.
- Disadvantages:
  - Resources are actually separate and independent clusters. “*inf-long*” would encompass different clusters, hiding underlying details.
  - Still possible to select “infiniband-only”, but less obvious to see how many infiniband-only resources are available or in use.

SLURM cluster backfill

- During periods with idle HPC cluster capacity, there will be **backfill** with HTCondor batch or grid jobs

- HPC MPI jobs will have priority and no backfill will take place when the SLURM clusters are congested

Questions and discussion