Distributed filesystem experiments at the High Performance Computing Center of Strasbourg

Romaric DAVID, Michel RINGENBACH
david@unistra.fr, mir@unistra.fr
Direction Informatique
05/11/2015

UNIVERSITÉ DE STRASBOURG

Plan

- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

This talk focuses on some extra experiments of the HPC Center of the University of Strasbourg (Unistra)

- Unistra is one of the major universities in France:
  - 48 000 students
  - 4800 employees
  - 3 Nobel prizes since 1987
  - major research institute in many scientific domains
  - ...some of them need HPC
- HPC Center (http://hpc.unistra.fr) serves the whole Alsace Region

High Performance Computing Center : Fundings

The HPC Center of the Unistra is funded by:

- Unistra: hosts the engineers responsible for the HPC Center
- The research labs fundings: until 2013, 100% of compute servers had been bought by the labs
  Labs are located not only in Strasbourg, but in all the Alsace region (too many logos to show)
- The French national initiative *Investissements d'Avenir*, *via* a national project: Equip@Meso
- French government, Alsace Region and Strasbourg Eurométropole

HPC in Strasbourg: Flops, Bytes....

- Around 350 servers, 5500 cores
- 500 TB of GPFS Storage
- 60 GPUs, from Tesla M2050 to K80
- 223 Tflops
- More than 250 active users
- More than 150 softwaremodules

HPC in Strasbourg

A team composed of 5 people:

- Operating all the HPC facilities (datacenter, clusters)
- Supporting more than 50 HPC scientific software by:
  - Defining a standard set of tools we strongly support: Intel compilers + in-house built OpenMPI, Cuda
  - In most cases, building/linking the scientific apps against these standard tools
  - **Writing and optimizing code**
- Doing all the training
- Promoting HPC for SMEs

Plan

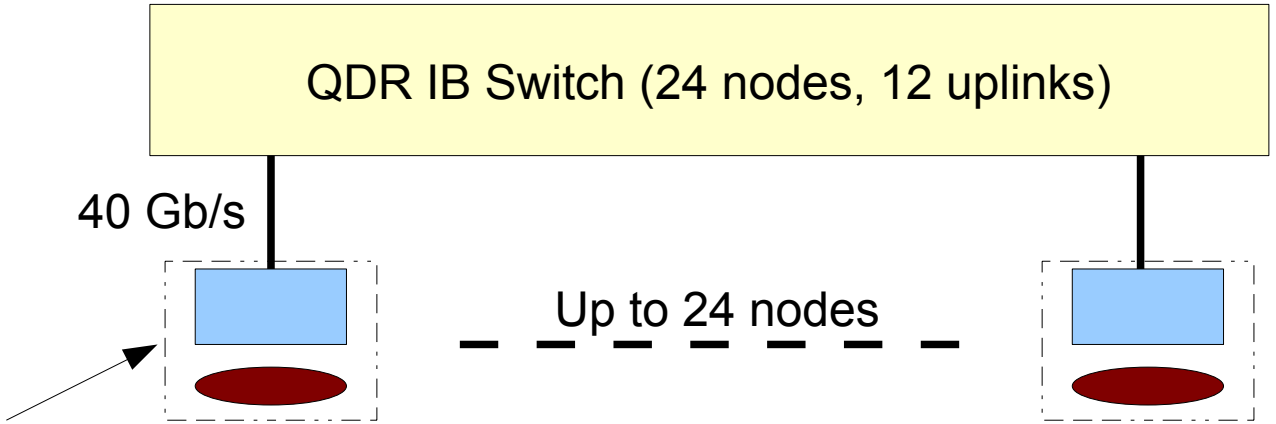
- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

« My simulation is slow »

- Once upon a time...
  - We were challenged by users of the Relion application http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page, used for reconstruction of 2D or 3D classes in cryo-electron microscopy
  - The execution times on the computing centers were much more longer than in the lab computers
  - Strategically not acceptable for the HPC Center
- At first we went wrong:
  - Compared speed-up against standard experiments of the author of the code (Sjors Scheres)
  - Profiled the application on users data-sets: I/O problems
  - Several hundreds of small files in real Datasets

« My simulation is slow »

- In the meantime, we were trying BeeGFS to answer the question « what can we do with those nearly unused local disks on the compute nodes ? »
- Very simple BeeGFS Setup:



Compute Nodes
1 x 1TB disk
No Raid, no nothing

« My simulation is slow »

- Very simplified BeeGFS setup:
  - Ext3
  - Uses system disk (or even /dev/shm on some test-cases)
  - Data lays on a specific directory, is visible outside BeeGFS
  - 1 Meta-data server per (max) 24 nodes
  - Volumes named after the IB switch they belong
- Which usage for this data ?
  - Temporary (scratch) data of jobs
  - No backup
  - Warning !!!!

« My simulation is slow »

- Performances : GPFS / BeeGFS
  - BeeGFS: Maximum bandwidth (dd, large files): 1GB/s
  - GPFS: 1GB/s or more but totally flooded when small files
- How to use this scratch space?
  - Users have to deal with 2 namespaces: `/home`, several `/scratch-XYZ` ← Named after the IB switch
  - Data staging mandatory (cp, **parallel cp**, ...)
  - Need to know where data is

« My simulation is slow »

- Users point of view
  - BeeGFS is great !
  - On the Relion code, speed up x 4
- What can users do with a 4x speedup ?
  - Run more simulations
  - Get more results
  - In this case, this lead to a publication in Nature

« My simulation is slow »

- Administrator point of view
  - Easy to deploy
  - *Sort of* fault-tolrant: generally when loosing a compute node part of a BeeGFS array, no data loss
  - Free Storage : 0 more U needed and lots of TB !
  - BeeGFS is the scratch solution we promote and deploy
  - Filesystem space = Job Node space
  - Makes node maintenance more difficult
  - Disposable high performance storage !

Plan

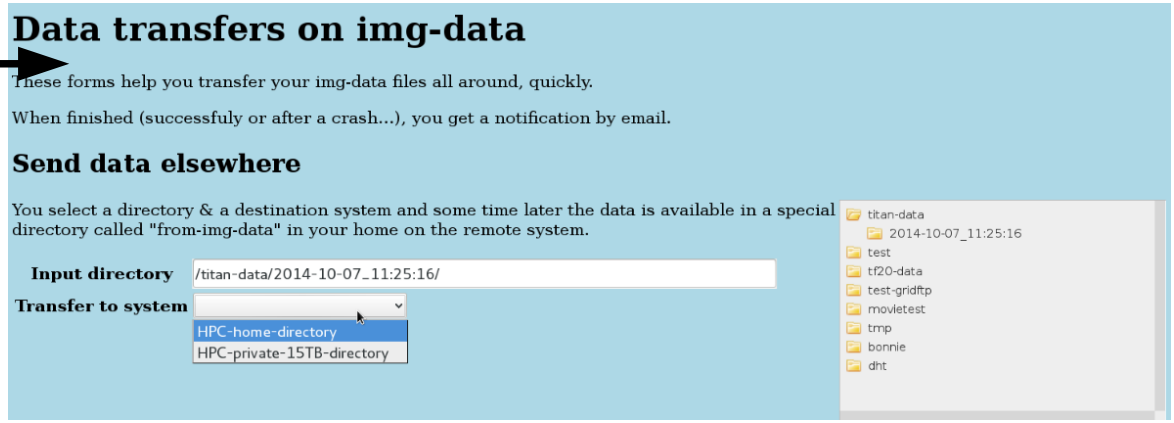
- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

Cascade effect

Now the users are able to deal with big data, how to transfer this data from the laboratory to the HPC ?

- We added a 10Gb metropolitan Vlan'ned network link: 5 kms between sites
- Transfer protocol: Grid-FTP with ssh authentication
- 700 MB/s point-to-point

Web interface developed by Jonathan Michalon (Igbmc)



Titan Microscope Hi-Res images

Cascade effect

Funny little things we never thought about before

- HPC Cluster = Nodes + IB
- Hold-on, what did I need Infiniband for, anyway?
  - IB is used for MPI
  - IB is used for GPFS
- For MPI, IB stay on the switch : since 2013, no job is allowed to spread on more than 1 switch
  - IB islands
- The overall IB network blocking factor 1:2
  - The « inter-switch » IB network is only used for GPFS

Cascade effect

- Should we really keep all these useless IB links ?
  - We can probably lower the blocking factor (1:3,...) ✓
  - Would lead to bigger IB islands → bigger MPI jobs ✓
- Given that GPFS is not that performant for **home directories** (lots of small files sometimes), we have to replace it by something else
- By the way, do we really need a parallel filesystem from the home directories ?
- Why not use Gb Ethernet for file access ?

Cascade effect

- **Statement: we want to build upon capacitive drives (7200 RPM, >= 4 TB)**
- Since June 2015, we've been trying on-site several filesystems (Thanks to Dell and Rozo Systems)
- Benchmarks : FIO (http://linux.die.net/man/1/fio) and *dd*, in parallel on up to 128 nodes.
- We tried :
  - Dell Compellent : pseudo-parallel NAS (up to 4 NAS heads) delivering CIFS, NFS, …
  - RozoFS: SDS, based on standard hardware, NFS and *native* mode via fuse

Cascade effect

- Dell Compellent: average good results, but scalability probably limited (size and performance). Not SDS....
- RozoFS: very good performance in NFS mode. Native mode works very well afer set-up
- We choosed RozoFS:
  - Standard hardware
  - 9 I/O servers (Dell R730xd)
  - 10 GbE network babkcone
  - 576 TO at the moment, up to 1.7 PO with 6TB disks

Plan

- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

Conclusion

- A single application infulenced the whole system
  - Regional computing centers are adaptive !
- *scratch* filesystems are the perfect sandbox
- Data needs to be close to the compute... during the compute !
- SDS, SSD and 7200 RPM disks are the keys to scale-up (capacity) and scale-out (performance) storage
- 10GbE and 40GbE is a game changer... makes parallel I/O possible on Ethernet
- We now use 2 SDS systems : BeeGFS and Rozo