THE CASPER CLUSTER: PRESENT & FUTURE

Nicolò Nepote - Department of Control and Computer Engineering - March 12, 2014
AGENDA

1. CASPER IS...
2. ASK CASPER WHAT HE CAN DO FOR YOU
3. HOW ARE THINGS GOING ON?
4. CASPER IS GOING TO BE...

hpc.polito.it
hpc.dauin@polito.it
@hpc_polito
CASPER IS...
A friendly ghost with my child's face?

It would be fun...

...but it's not the correct answer
A buzzing and noisy kubrickian monolith?
Less fun...
...but far more useful to research
CASPER, BACK IN 2008

1. CASPER is the HPC system available at HPC@POLITO
2. Cluster Appliance for Parallel Execution and Rendering
3. First use case as a Blender 3D render farm
4. Built as a “Beowulf style” cluster
5. Based on Linux and free/open management software
6. Possibly use open standards for high speed interconnection
7. Must provide engineering support
8. Need users to publish their papers and students to learn better: this is our mission
9. Budget = 0
CASPER, BACK IN 2008

0.16 TFLOPS
Athlon XP single core
2 GB RAM/core
44 nodes
Gigabit Ethernet
Local storage
Huge useless power consumption

not exactly what is called “a supercomputer”
1. Slow network is useless network → InfiniBand
2. Lack of parallelism → Opteron
3. Master node as storage node → Dedicated NAS
4. We need more RAM, a lot more RAM, please! → 128 GB/node, 4 GB/core
5. Can you give priority to fellow research groups? → Priority queues
6. Help your users, they're your first resource (someone says) → Study, study, study!

Grant from the Board of Governors in 2012
CASPER TODAY

Architecture: Linux InfiniBand MIMD Distributed Shared-Memory Cluster
Node Interconnect: InfiniBand DDR 20 Gb/s on copper
Storage Interconnect: Ethernet 2x 10 Gb/s (bonding 802.3ad)
Service Network: Gigabit Ethernet 2x 1 Gb/s (bonding 802.3ad)
CPU Family: AMD Bulldozer
CPU model: Opteron 6276 2.3 GHz (turbo 3.0 GHz) 16 cores
Sustained performance: ~ 3 TFLOPS (recalculating)
Computing Cores: 448
Number of Nodes: 14 (dual socket)
Total RAM Memory: 1.8 TB DDR3 Registered ECC
Working Storage: 47 TB on RAID 6, throughput near 800 MB/s
OS/Scheduler: ROCKS Clusters + GridEngine
CASPER TODAY

- Compute nodes dedicated to fellow research groups
- Master node and Login node
- Ethernet switch
- InfiniBand DDR switching fabric
- Standard compute nodes
- 60 TB NFS-shared Storage via dual 10 Gbe
- 12 TB backup storage via Gbe
- Compute nodes dedicated to fellow research groups
**COMPUTE NODE HW CONFIGURATION**

- InfiniBand DDR/QDR HCA dual port with CX4/QSFP connectors
- Dual socket AMD Opteron 6276
- bootstrap HD (OS + temporary folders)
- 128 GB DDR3 Registered ECC
- Redundant PSU
- Cooling
- 1TB SATA 3 HD (dedicated to HDFS)
THE ADDED VALUE OF InfiniBand

1. Layer 1-2 open (and expensive) high speed network technology
2. Industry standard in HPC
3. Delivers large bandwidth and low latency
4. Carries MPI traffic as well as IP (IPoIB)
5. Can really boost your computation
1. CASPER (3 TFLOPS): 4 GB/core, 128 GB/node, 32 cores/node
2. CINECA's Fermi (2 PFLOPS): 1 GB/core, 16 GB/node, 16 cores/node
3. Tyane 2 (33 PFLOPS): 320 MB/core, 64 GB/node, 24 cores/node (plus 2 Xeon Phi accelerators)
4. Fat nodes provide the user with high parallelism and large memory available for non-MPI, multi-threaded or even sequential programs
CASPER LOGICAL ARCHITECTURE

- Working storage
  - NFS file sharing
    - Compute node
      - GbE
      - DDR IB
    - Compute node
      - GbE
    - Compute node
      - MPI + IPoIB
    - Master node
  - 10 GbE
  - HDFS storage
    - HDFS over Ethernet
    - HDFS over InfiniBand
ASK CASPER WHAT HE CAN DO FOR YOU
SOME USE CASES AND STRATEGIES

MPI parallel programming
In-house code (C, Fortran) using OpenMPI, MPICH, etc.
Truly parallel through InfiniBand

Domain splitting simulations
Parallelize the program by dividing the domain into (mostly) independent regions
Can exploit MPI and InfiniBand

MPI-capable 3rd party software suites
Star-CCM+, OpenFOAM, Quantum-Espresso, Gromacs...
Huston, we need a license!

Multiple instances like level-0 parallelism
Same non-MPI program runs on different inputs under different conditions inside thousands of jobs
Often memory consuming
Easy to deploy and very effective

Rendering of movies and hi-res static images
Blender 3D + Cycles
CASPER is a render farm again

Map-reduce applications
Hadoop + HDFS
This is HPC for Big Data
HDFS runs fast on IPoIB
USER EXPERIENCE

1. Modeling & development on your workstation
2. SFTP file transfer
3. Secure Shell connection
4. Job definition, in-queue submission and monitoring on CASPER
5. email notification for jobs
6. On-line statistics (ganglia)
1. All ROCKS 6.1 included software - Programs and libraries for technical and high performance computing in bundle with the ROCKS cluster distribution
2. AMD Open64 - AMD compilers optimized for Opteron Bulldozer architecture
3. AMD Core Math Library - Opteron optimized BLAS, LAPACK, FFTs and random number generators
4. Blender 3D 2.59 and 2.69 - 3D creation for everyone, free to use for any purpose (cit.)
5. IBM ILOG CPLEX 12.6.0.0 - High-performance mathematical programming solver
6. GotoBLAS2 & ATLAS - very fast multi-threaded BLAS implementations
7. Gromacs 4.6.5 and 5.0-beta2 - A versatile tool for molecular dynamics
8. OpenFOAM 2.1.1 and 2.2.x - Free, open source Computational Fluid Dynamics software package, plus 3rd party SW
9. Quantum Espresso 5.0.3 - Integrated suite of Open-Source computer codes for electronic-structure calculations and materials modeling at the nano-scale
10. Star CCM+ 8.06.005 - Simulation of Turbulent flow in Arbitrary Regions and Computational Continuum Mechanics
11. Octave 3.6.4 and 3.8.0 - High-level interpreted language, primarily intended for numerical computations
12. Matlab 8.1 R2013a - The language of technical computing (cit.)
CASPER SOFTWARE ARCHITECTURE

User Space (qsub - qstat)

ROCKS Cluster 6.1

CentOS 6.3
Linux x86_64
2.6.32
compute node

CentOS 6.3
Linux x86_64
2.6.32
compute node

CentOS 6.3
Linux x86_64
2.6.32
compute node

CentOS 6.3
Linux x86_64
2.6.32
compute node

CentOS 6.3
Linux x86_64
2.6.32
compute node

High speed network

ROCKS shell

MPI

ILOG CPLEX

Quantum Espresso

Star CCM+

Matlab Octave

Open FOAM

Blender 3D

Hadoop

OGS GridEngine 2011.11p1 (SGE 6.2u5)
GridEngine IS YOUR FRIEND

1. Remember that clusters are batch systems with different queues mastered by some scheduler

2. Know your software needs and behavior

3. Know the cluster and watch the situation before submitting

4. Choose the queue that better fits your expectations and ask what you need to the scheduler
QUEUES AVAILABLE ON CASPER

Job submission via qsub command

all.q

public.q    lisa.q    bioeda.q    nqs.q    t.b.d.

fast.q

node 0 → node 1 → node 4 → node 5 → node 6 → node 8 → node 9 → node 10 → node 11 → node 12

Fine tuning in GE: load thresholds, queue subordination, core over subscription, job priorities
HOW ARE THINGS GOING ON?
SOME FACTS ABOUT HPC@POLITO

1. 47 research projects have been hosted and more are coming
2. 17 projects are still active on CASPER
3. 26 research groups from 7 departments were involved
4. 39 papers were published
5. CASPER received 3 hardware upgrades in 5 years
6. HPC@POLITO was accepted as a member of the “HPC Advisory Council” in 2013
7. HPC@POLITO has no budget for human resources
PAPERS

Papers published per year

Papers published per area

Data source: hpc.polito.it
HARDWARE EVOLUTION

Total Cores and Nodes

- Cores
- Nodes

2011 2012 2013 2014

Total Memory and Storage

- RAM
- HDFS Storage
- Shared Storage

2011 2012 2013 2014

CASPER IS GOING TO BE...
HARDWARE EVOLUTION

1. Goal: more rough power
   - 12-core Intel Xeon E5 2695 v2 dual socket

2. Goal: faster storage
   - Lustre or pNFS over InfiniBand or Fiber Channel

3. Goal: reach 10 TFLOPS
   - InfiniBand FDR 56 Gb/s on fiber, ~ 20 nodes

4. Goal: use less power
   - multi-node highly-packed chassis (like HP s6500)

Why not general purpose GPUs?

Why not HPC coprocessors?

Announcing: Intel Xeon Phi™ Coprocessor Breakthrough Performance!

1 TFLOPs
Linpack (HPL)
in a node

118 TFLOPs
Entry into the Top500
1. Deploy already available NICE EnginFrame portal:

2. Develop a tile rendering system for Blender, but tightly integrated with GridEngine...
TO WHOM IT MAY CONCERN


A bibliography about HPC@POLITO
THANKS!

hpc.polito.it
hpc.dauin@polito.it
@hpc_polito