Extension of the INFN Tier-1 on a HPC system

Tommaso Boccali (INFN Pisa) on behalf of many people, see last slide
The 3 parties in the game

- LHC Experiments
  - The Italian groups in ALICE, ATLAS, CMS, LHCb, deliver 10-15% of the experiments' computing pledges.

- CNAF
  - Italian Tier-1 in Bologna. Main INFN data center providing computing and storage to more than 30 experiments

- CINECA
  - HPC center in Bologna. EU/Prace Tier-0.
  - Current top machine is Marconi, partially KNL and partially Skylake based. At 21st position in top500.org
  - Selected site for 1 of the 3 EU pre-exascale machines

**CNAF**
- Standard "GRID-like"HTC farm (30k cores, 400 kHS06)
- 38 PB of disk
- 90 PB of tapes on 2 libraries

**CINECA**

- **Marconi cluster**
  - Based on Omnipath
  - ~19 Pflop/s
  - 17 PB of local storage

- **Marconi A2 Partition**
  - 3600 nodes with 1 Xeon Phi 2750 (KNL) at 1.4 GHz and 96 GB of RAM
  - 68 cores/node, 244800 cores
  - Peak Performance: ~11 Pflop/s

- **Marconi A3 Partition**
  - 3216 nodes with Skylake at 2.1 GHz
  - Peak Performance: ~8 Pflop/s
An integration effort!

Matching LHC workloads with HPCs is known not be an easy task! You need to match (at least)

- **“Usually strict” site policies**
  - Ad-hoc operating system, limited/absent external connectivity, user policies, node hardware setup, ...

- **The typical requirements of our workloads**
  - Base architecture, access to external conditions / data, largish local work areas, access to sw via CVMFS, RAM requirements, virtualization...

- **The expectations** from a WLCG site
  - accounting, traceability, security, accountability, monitoring, job late binding requiring external access, ...
  - Use generic grid framework and tools to incorporate HPC resources

- **NOT a solved a problem in general**
  - Ad-hoc experimentations with specific sites
  - Experiments trying to sketch a more general path, but we are not there yet

- **An additional requirement is to allow transparent access to CINECA resources as if they were part of the CNAF site**
  - CINECA Marconi A2 as an “elastic extension” of CNAF - first step for being part of the pledge
<table>
<thead>
<tr>
<th><strong>A typical Marconi A2 node configured with</strong></th>
<th><strong>A typical WLCG node has</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A KNL CPU: 68 or 272(HT) cores, x86_64, rated at ~¼ the HS06 of a typical Xeon</td>
<td>1/2 Xeon-level x86_64 CPUs: typically 32-64 cores, O(10 HS06/thread) with HT on</td>
</tr>
<tr>
<td>96 GB RAM, with ~10 to be reserved for the OS: 1.3-0.3 GB/thread</td>
<td>2GB/thread, even if setups with 3 or 4 are more and more typical</td>
</tr>
<tr>
<td>No external connectivity</td>
<td>Full outgoing external connectivity, with sw accessed via CVMFS mounts</td>
</tr>
<tr>
<td>No local disk (large scratch areas via GPFS/Omnipath)</td>
<td>O(20 GB/thread) local scratch space</td>
</tr>
<tr>
<td>Access to batch nodes via SLURM; Only Whole nodes can be provisioned, with 24 h lease time</td>
<td>Access via a CE. Single thread and 8 thread slots are the most typical; 48+ hours lease time</td>
</tr>
<tr>
<td>Access granted to individuals</td>
<td>Access via pilots and late binding; VOMS AAI for end-user access</td>
</tr>
</tbody>
</table>
Does it make sense to try an integration?

● **Today**
  ○ we are not relying on HPC resources. **Not on the critical path**

● **Tomorrow (2027+)**
  ○ There are strong hints that HL-LHC processing will need to access HPC resources depending on **specific Funding Agency's policies**
    ○ The current modelling of HL-LHC computing does not fit a reasonable budget with only in-house resources
    ○ If we do not get experience on today's systems, it is difficult to contribute to the specs of next gen systems
    ○ HPC systems give massive access to accelerators and in general to heterogeneous systems

● On top of that ...
The peculiarities of CNAF-CINECA situation

- They are **close** (less than 8 km)
- They are already collaborating: a large part of the CNAF pledge (180 kHS06) comes from “standard GRID-like” nodes hosted @ CINECA (former A1 partition of Marconi)
- Connected via a dark fiber and a pair of Infinera CloudXpress systems
  - Capping at 1.2 Tbit/s; currently at half capacity
  - High bandwidth + low latency → no strict need for caches
- CINECA and CNAF are planning to move to a common location (the “Technopole”) by 2021, and integration experience is welcome in that perspective
The Grant

- The LHC Italy community successfully applied for a “PRACE Project Access” on the CINECA KNL partition
- **30McoreH allocated** after a demonstration the project was “feasible” (via a 20kcoreH test)
- “Feasible” meant many handshaking / changes to initial setup on both sides - thanks to the mutual understanding and the flexibility from CINECA’s side on what is seen as a use case of mutual interest

### Project scope and plan

<table>
<thead>
<tr>
<th>Project name</th>
<th>LHC@BPHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research field</td>
<td>High Energy Physics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Principal Investigator (PI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title (Dr., Prof., etc.)</td>
</tr>
<tr>
<td>Last name</td>
</tr>
<tr>
<td>First name</td>
</tr>
<tr>
<td>Organisation name</td>
</tr>
<tr>
<td>Department*</td>
</tr>
<tr>
<td>Group*</td>
</tr>
<tr>
<td>Country</td>
</tr>
</tbody>
</table>

- **CVMFS (and its squids) ok**
  - Interest also from non HEP community
- **External networking enabled to CNAF and CERN**
  - Enough to guarantee access to CNAF storage and conditions for experiments' workflows
  - We can use it also for accessing external data
- Partial routing to CNAF storage / squids over the dark fiber @ 40 Gbit/s (technical limitation to be removed with next machine)
  - The rest over GPN
- **Singularity** audited by CINECA’s sysadmins and green lighted
- A **HTCondorCE/Slurm** allowed at the CINECA edge nodes

Some personal comments:
- Being able to speak to the center makes things so much easier
- Often the site limitations are there since “nobody needed the feature before”
- In general, getting an agreement on the changes was easier than feared!
Technical setup #1: jobs

- Storage Site #1
  - XrootD
- Storage Site #2
  - XrootD
- XrootD Caching Proxy
- SRM
- GPFS
- Xrd

- CVMFS
- Experiment conditions data
- Experiment late binding WMS

- HTC CE
- Slurm
- Singularity
- Squid1..n
- Routing1..4

- GPN 40 Gbit/s direct link

- KNL Node
- payload
- pilot
Technical setup #2: sw and conditions

- Storage Site #1
  - Xrd
- Storage Site #2
  - Xrd
- XrootD Caching Proxy
- CVMFS
  - Stratum 0
- Experiment conditions data
- Experiment late binding WMS
- GPN
- Xrd Storage Site #1
- Xrd Storage Site #2
- 40 Gbit/s direct link
- 40 Gbit/s direct link
- HTC CE
- KNL Node
- Singularity
- Squid1..n
- Routing1..4
- SRM
- GPFS
- Xrd
Technical setup#3: data access

Storage Site#1: XrootD
Storage Site #2: XrootD

XrootD Caching Proxy

SRM
GPFS

40 Gbit/s direct link

40 Gbit/s direct link

CVMFS Stratum 0
Experiment conditions data
Experiment late binding WMS

HTC CE
Singularity

Squid1..n
Routing1..4

Xrd

Xrd

Xrd

Xrd

Xrd
... but: what to run?

- Nodes with 68(272) threads each
  - Insufficient RAM (86 GB) to run 68 (272) single threaded payloads
- Clear advantage in using MT/MP codes to reduce the footprint
  - But inevitably throughput goes down (Amdahl and friends ...)
  - Need to find a the best configuration, different from the one used on machines with standard RAM (2 GB/thread)
- Some examples of analyses done for/after the grant submission from CMS (MT) and LHCb (MP)

LHCb: GaussMP simulation
(see Integrating LHCb workflows on HPC resources: status and strategies @CHEP20-19)

- Easily able to fit the memory @ 68 threads
- Even able to fit @ 68x2 HT threads using MP@17, but with minimal gain in ev/sec

(See F.Stagni @CHEP2019)
Able to use up to 256 cores in multiple configurations

#P=32, #T=8 leaves not enough memory margin, going for #P=8, #T=32 (20% loss in performance)

CPU efficiency (from “top”) declines as expected with Amdahl

Previous study (C.Jones): also RECO follows the same pattern of usability
ATLAS: Geant4 simulation runs @ KNL in MP

---

ALICE:
- submission via slurm / htcondor from our alice-htc vobox (XrootD used to get input files needed for simulation)
- **1st set of tests (slurm and condor CE):**
  - job type: simulation in O2/Run-3 framework (Pythia events + Geant3) with different settings ($n_{jobs}$ per node, $n_{thread}$ per jobs)
- **2nd set of tests (condor CE):**
  - the job and type queue ($n_{jobs}$ per node, $n_{thread}$ per jobs) will be set at the central services level
  - the jobs will be pre-equipped with monitoring tools

(ALICE was not present at the time of project submission, recent addition)

---

It makes sense to use up to 2x HT, then suffering from memory
The system in action

- Grant ends by April 2020; plan is to start large scale productions @ Xmas
- Complete dress rehearsal already done with CMS:
  - Submit via CRAB to the CMS HTCondor Pool, directing to the CE @CINECA (logically belonging to CNAF T1)
  - Slurm starts singularity as a pre-exec (optional, some experiments can run the pilot also in the host system)
  - Payload starts from CVMFS and reads data from the CNAF XrootD proxy cache provided via the XDC EU Project
  - Results are stored on CNAF's storage
  - All the communications including MONIT callbacks in place

```bash
WARN TMPDIR(/marconi scratch/usuario07/cms09/slrjm 5420552/glise U7J2ge execute/dir 028) is not a writable directory setting $TMPDIR = $PWD
******** gMS-CMSRunAnalysis.sh STAFFING at Mon Oct 7 18:47:08 GMT 2019 on r105c06s83
---
Local time: Mon Oct 7 18:47:08 UTC 2019
Current system: Linux r105c06s83 3.10.0-327.36.3.e17.xppsl /51.1.41511.x86 64 #1 SMP Tue Febl 7 22:40:59 UTC 2017 x86_64 x86_64 x86 64 GNU/Linux
```
The proxy cache in action...

- Based on **XDC**(EU project) Xcache implementation
- **Yellow**: starting from cold cache, files are loaded into the cache
- They are served in successive runs (**Green**)

- Given the design where files @ CNAF do not pass through the cache, **1 host @ 10Gbit/s should suffice**
- Cache is needed here mostly in order to **overcome routing limitations**, and to allow accessing the whole federation (and not only CNAF and CERN files)
Conclusions and future plans

- Scalability tests are needed before we get into production ~ Xmas
  - Idea is to use 250 nodes as baseline (0.5 M coreH/day)
- Use Marconi as a concrete example to study HPC pledge evaluation (HS06 and HEP-workload benchmarks)
  - More than the real “value” of the granted CPU time, *we especially value the establishing of close interactions with the HPC site*
  - The successful handshaking in a working configuration will be used to *testify that a collaboration is possible* when meeting other HPC centers
- More intriguing “local” scenario in 2022
  - By the end of 2021, new CINECA pre-exascale machine (a.k.a. *Leonardo*) will be installed (at the new Bologna Technopole); by 2022, also INFN Tier-1 will be moved at the Technopole. *The two centers will be literally in the same room*
  - A tested collaboration protocol, with mutual understanding of needs, will open the path to very interesting collaborations, with *INFN having a privileged link to a 250 PetaFlops machine*
The team

● LHC-Italy
  ○ Tommaso Boccali
  ○ Alessandro de Salvo
  ○ Concezio Bozzi
  ○ Stefano Perazzini
  ○ Francesco Noferini
  ○ Anna Lupato
  ○ Alessio Gianelle
  ○ Daniele Spiga
  ○ Diego Ciangottini
  ○ Daniele Bonacorsi
  ○ (...)

● CNAF
  ○ Stefano Dal Pra
  ○ Stefano Zani
  ○ Gaetano Maron
  ○ Luca dell’Agnello
  ○ (...)

● CINECA
  ○ Marcello Morgotti
  ○ Daniela Galetti
  ○ Carlo Cavazzoni
  ○ (...)

● CERN
  ○ Andrea Valassi
  ○ Federico Stagni
  ○ (...)