Enabling data analysis at the PROOF on the Italian ATLAS Tier2s using PoD

PROOF and PoD

- PROOF (Parallel ROOT Facility) allows interactive analysis on a set of distributed resources using a multi-tier master-worker model to achieve dynamic workload-balancing.

- PoD (PROOF on Demand) → tool-kit to setup a PROOF cluster on any Resource Management System (RMS).
  - currently supported backends: LSIF, PBS, OGE, Condor, LoadLeveler and gLite-WMS

- PoD and ROOT setup from the CVMFS ATLAS distribution.
  - PoD version 3.10
  - ROOT 5.32/02

Distributed Analysis

- Four back-ends: PanDA (Production and Distributed Analysis), WMS (Workload Management System), CREAM and ARC.
- Job submission with: pox and Patella for PanDA back-ends
- Job submission with: George for PanDa, WMS, Cream and Arc back-ends
- ATLAS central accounting systems only report job submitted via PanDA
- User access Grid resources connecting to a User Interfaces

Readout Performances

- Testbed
  - 5rm type: Disk Pool Manager
  - Protocol access: write (configured for local access only)

- Input rate in Mb/s second using the PROOF statistics tools as a function of the number of workers.
  - Test performed into Romal and Frascati Tier2
  - Conservative estimation

- Slope for small number of workers measures the rate of reading and decompressing the event per worker.
  - 8 Mb/s per process, depend on the type of analysis and on the structure of the event read and built in memory.

- The scalability for increasing number of workers indicates how the system reacts to increasing worker loads (e.g. increasing number of users).

References


INFN – Laboratori Nazionali di Frascati
Cern – Geneva

ATLAS Italian Cloud

<table>
<thead>
<tr>
<th>Tier 2</th>
<th>2012 total resources</th>
<th>Computing</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Core</td>
<td>Hi/Perc</td>
<td>Batch system</td>
</tr>
<tr>
<td>Frascati</td>
<td>870</td>
<td>850</td>
<td>PBS</td>
</tr>
<tr>
<td>Milano (T2D)</td>
<td>1050</td>
<td>1090</td>
<td>PBS/Condor</td>
</tr>
<tr>
<td>Napoli (T2D)</td>
<td>1200</td>
<td>1240</td>
<td>PBS</td>
</tr>
<tr>
<td>Romal (T2D)</td>
<td>1300</td>
<td>13100</td>
<td>LSIF</td>
</tr>
</tbody>
</table>

Startup latency

- Startup latency: time necessary to allocate a certain number of nodes with PoD before to run PROOF analysis.
- First tested in Frascati’s Tier2. Startup latency tested in Frascati, Romal and Napoli. Latency depends on the share allocated in the scheduler and many other parameters (cluster load, ...)
- A job with 50 cores requested every 30 minutes for a total of 21 submissions.
- Color scale proportional to job submission time (from darker to lighter)
- In ganglia plot the cluster load reports different situations:
  - Deep colors running jobs, lighter colors queued jobs
  - Green Monte Carlo production jobs, yellow Panda analysis, olive green ATLAS Italian group (used for PoD submission)
- First submissions suffer from resource competition with Panda analysis jobs

- Two users with the same VOMS credential in competition for the same resources
  - Low Tier2 load: no competition for resources
  - Medium Tier2 load: larger startup latency
  - In the worst case the second user is not even able to allocate requested nodes in the monitoring time window (20 min.)

- Tiers 2 results
  - Good linear scalability over the range tested
  - Dataset used for the test distributed over 20 data servers connected via a 10 Gbit/s network switch
  - Saturation at 3200 processes requiring each 8 Mb/s.

- Frascati
  - Deviation from linear scalability over the range tested
  - Dataset used distributed over 5 data servers, 3 of which temporarily connected via a 1 Gbit/s switch
  - Saturation expected at ~1000 Mb/s.