Sim@P1: Using Cloudscheduler for offline processing on the ATLAS HLT farm

F Berghaus for the Sim@P1 team

on behalf of the ATLAS Collaboration
Outline

- Definition: What is Sim@P1
- Status: Current operation
- Plan: Integration of Cloudscheduler
What is Sim@P1?

Sim@P1 = Simulation at point one

<table>
<thead>
<tr>
<th>Racks</th>
<th>Servers per rack</th>
<th>Cores per node</th>
<th>RAM per node</th>
<th>RAM per core</th>
<th>Total cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-4, 6-13, 94, 95</td>
<td>32</td>
<td>16</td>
<td>~24 Gbyte</td>
<td>~1.5 Gbyte</td>
<td>10K</td>
</tr>
<tr>
<td>64-69</td>
<td>40</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-26, 75-77</td>
<td>32</td>
<td>48</td>
<td>~64 Gbyte</td>
<td>~1.3 Gbyte</td>
<td>64K</td>
</tr>
<tr>
<td>70-74, 79-90</td>
<td>40</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44-54</td>
<td>40</td>
<td>56</td>
<td>64 Gbyte</td>
<td>~1.1 GByte</td>
<td>74K</td>
</tr>
<tr>
<td>Total: 58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ATLAS Detector → Level 1 Trigger → High Level Trigger → CERN Data Centre
Sim@P1: Current Operation

- Dedicated VLAN for *offline* access to list of hosts in CERN General Purpose Network
- Compute resources isolated by virtualization
Sim@P1: Current Operation

- **Boot:**
  - Puppet launches nova on worker nodes
  - Puppet executes scripts to launch instances
  - Instances connect to condor
    - CM + 2 Sched in GPN
- **APF submits to each Sched**
- **Shutdown:**
  - Puppet kills nova on worker nodes
  - Puppet calls cleanup scripts
- 20MB CernVM3 micro-kernel distributed from glance
- CernVM3 caches in ATLAS software and operating system
- Two SQUID servers in P1 are sufficient to provide software
Issues with current operation

- Hard to maintain:
  - Many undocumented scripts
  - Scripts spread over many servers in P1 and in GPN
- No error handling for running instances
- Hard to update or modify
Proposal for Sim@P1
Cloudscheduler

- Batch system on distributed cloud infrastructure
- In production for offline processing for
  - ATLAS (2012 - present)
  - Belle-II (2014 - present)

Cloud Scheduler

Cloud Interface
- (Nova, Boto, OCCI)

Cloud Interface
- (Amazon, Google, Microsoft)

Job Scheduler
- (HTCondor, Torque, etc.)

User
Cloudscheduler at Point 1

- Proposal for long shutdown two [LS2]:
  - Cloudscheduler & OpenStack run in P1 Network
  - Polling thread and HTCondor run in CERN GPN
  - Cloudscheduler and polling thread interact with database
Cloudscheduler at Point 1

- Communication flow for Cloudscheduler
- Requires channel to database between P1 & GPN
Harvester Job Submission

- Harvester pull mechanism allows job-specific resource request
- Condor reports resources availability to Harvester to improve PanDA job brokering
Summary

- Sim@P1 is successfully operating
- Cloudscheduler setup to ease operation under evaluation
- PanDA Harvester setup for job more flexible job submission
Thanks to many contributors

Cloudscheduler Team
K Casteels, C Driemel, M Ebert, C Leavett-Brown, M Paterson, R Seuster, R Sobie, R P Taylor, T Weiss-Gibbons

Sim@P1 Team
A Di Girolamo, C Lee, P Love, J Schovancová, R Walker

TDAQ Team
F Brasolin, D A Scannicchìo, M E Pozo Astigarraga
Tentative future

- GPN Uplink
- Non-modular Switch
- Point 1 Gateways
- Control Network Core Routers
- OpenStack Controller
- Glance Service
- DHCP + Ganglia
- CVMFS + Frontier
- Squid Proxy
- HLT Server
- Castor Router

1 Gbps
2x 10GbE
2x 10GbE
2x 10GbE
2x 10GbE
2x 10GbE
2x GbE
2x GbE
10GbE
10GbE
10GbE

Hiding some heterogeneity