OSG Technology prep for LHC Run 2

Brian Bockelman
OSG Council meeting, January 2015

Preparations for Run II

- The goal of this presentation is to summarize the LHC Run 2 preparatory work we have done in the past three months..
- The basic question was whether the OSG Software stack was prepared for the challenges of Run II. In particular,
  - Are there was anything additional scale tests OSG needed to perform that were not covered by the LHC data challenges in Fall 2014? **Answer: Yes!**
  - What's the priority for Run 2 (and, correspondingly, risk) across the components of the software stack? **Answer: See extra slides.**

Scale tests in general

- LHC Run 2 increases data rates and required processing. Rough guidance was each increased by a factor of two.
  - It does not, for example, increase the number of sites.
- It was determined that data and interaction rates to filesystems are already thoroughly tested for storage systems we support.
  - For HDFS in particular, CMS covered this with their 20Gbps PhEDEx tests last Fall.
- We determined the major piece of work needed was probing the scale limits of the CE and of glideinWMS/HTCondor.

glideinWMS and HTCondor Scale Tests

- CMS had been meeting with HTCondor team for about 6 months; OSG joined in 3 months ago with a sole focus on scale testing.
- OSG utilized previous experience and tools for scaling HTCondor, especially using glideinWMS.
  - UW and UCSD provided hardware.
- Existing CMS scale: 100k parallel running jobs, 20k jobs per schedd.

glideinWMS and HTCondor Scale Tests

- Highlights of results as of December 31, 2014:
    - Idle jobs in a sched: goal 200k; achieved 500k.
    - Running jobs in pool: goal 200k; achieved 200k.
    - Running jobs in a single schedd: goal 40k; achieved 100k.
- Additionally, we tested several failure modes -

HTCondor-CE Scale Tests

- Goal for HTCondor-CE was 20k running jobs on a single CE from a single HTCondor-G factory.
  - UNL provided the hardware and test batch system.
  - Achieved: 16k running jobs. Confirmed the limit was the test batch system, not the CE.
  - UNL will be upgrading test pool in the future, allowing us to verify.
- **However**, we believe this is acceptable for now; largest single GRAM CE is currently 10k. CEs also scale horizontally - most sites will want to run multiple for redundancy regardless.

OSG Software Priorities for Run 2

- Unlike the LHC itself, OSG has been in continuous production throughout the long shutdown.
  - Hence, the startup has a slightly smaller impact.
- OSG ships about 200 packages; too many for a human to draw a conclusion about.
  - Hence, in the extra slides, I group the packages into “major components” and rate the *impact* on LHC and the *risk*.
    - Apologies if your favorite package was dropped - I’ll be happy to answer questions afterward!
- Overall messages:
  - All major components are ready for Run 2.
  - There are several pieces of software that have been orphaned which LHC depends on (whose outage may result in degraded or no physics). Responsibility for these have been taken over by OSG; this reduces overall risks, but is a significant cost!
  - OSG’s monitoring, information, and accounting software, while not critical, has many components with series concerns.

Future Activities

- **SCALING**: All bugs and special configurations necessary to achieve scale were reported back to HTCondor and glideinWMS teams.
    - Most are rolled into the latest HTCondor release; only about 4-5 system tunings are needed. OSG is pushing to make everything scale “out of the box”.
    - The difference between OSG-measured performance and the CMS-observed performance has been whittled down to differences in *hardware* and *VO-specific* applications.
    - We will retest applications in ~3 months to validate improvements we suggested.
- **SOFTWARE**: Throughout the next 6 months, we will continue to:
    - Support important pieces of abandoned software (GUMS, BeStMan).
    - Retire software to decrease size of software stack.
    - Continue rollout of HTCondor-CE.

Extra Slides

Major OSG Software Components

- From https://twiki.grid.iu.edu/bin/view/SoftwareTeam/ComponentOverview, with some updates/additions
- **RED**: Criticality of software. What happens if it does not function?
  - **1** = LHC cannot produce any physics without it;
  - **2** = LHC physics throughput is degraded;
  - **3** = Accounting or monitoring for LHC is degraded;
  - **4** = LHC unaffected.
- **BLUE**: Readiness for Run 2.
  - **1** = Not ready for Run 2.
  - **2** = Prepared for Run 2, but software is orphaned or abandoned by developers.
  - **3** = Modest concerns about sustainability or slated for retirement for Run 2.
  - **4** = No risks identified.

Major OSG Software Components

- Computing / site:
    - GRAM Gatekeeper: **1**, **3**. To be retired spring 2016.
    - HTCondor-CE: **1**, **4**.
    - GridFTP: **1**, **4**.
    - Gratia Probes: **3**, **3**.
    - GUMS: **2**, **2**. Supported by OSG. Small developments for stakeholder requests.
    - Worker node client tools: **2**, **2**. Covers many components; file transfer tools (lcg-utils, srmcp, srm-client) and VOMS client are biggest concern.
- Storage:
    - GridFTP: **1**, **3**.
    - HDFS: **1**, **3**.
    - BestMan: **1**, **2**. Supported by OSG. No future developments foreseen.
    - Xrootd: **1**, **4**.

Major OSG Software Components, Continued

- Monitoring, Accounting, and Information services:
    - RSV: **3**, **4**. OSG developed/maintained.
    - osg-info-services/GIP: **2.5**, **3**. To be retired. OSG maintained.
    - Gratia Probes: **3**, **3**
- VO Services:
    - VOMS: **1**, **3**. Several releases behind, but no requests for upgrades.
    - glideinWMS frontend: **2**, **4**. Many discussions in the past about OSG VO issues with this; CMS use case in better shape.
- Software for OSG-operated multi-VO services.
    - glidenWMS factory: **1**, **4**
    - Gratia collector: **3**, **3**