STAR experiment & Cloud computing

Jérôme LAURET
Presenter: Doug Olson

Provided questions

- Describe how your community uses infrastructure clouds
- What drew you to infrastructure clouds, i.e., what are the benefits for your community of using infrastructure clouds in order of significance?
- What are the challenges of using infrastructure clouds for your community, i.e., what makes them difficult to use in order of significance?
- How do infrastructure clouds compare to other options for outsourcing computation from the perspective of your community's needs?

How your community uses infrastructure clouds

STAR have had along history with Cloud computing

- Earliest try in 2007 (first real use for burst resource in 2008 – see The new Nimbus: first steps in the clouds)
- Many models / approaches tried and compared (see SciDAC 2010 workshop presentation and paper here).
  - Using native EC2 interfaces, Nimbus, Condor outside, Condor inside, Virtual Organization Cluster, PBS, Kestrel an IM mechanism to controlling jobs, …
- Used Magellan infrastructure (see The case of the missing proton spin and Magellan Tackles the Mysterious Proton Spin) at both NERSC and ANL
  - Autonomous “ant-farm” consumer/provider model developed based on past experience with Nimbus, Eucalyptus and OpenStack

How your community uses infrastructure clouds

- Uses included
  - Running simple Monte-Carlo simulations – **last minute burst of resources, elasticity of resources seek**
  - Running complex simulations (multi-stage with detector response) – **large demanding simulation requests** spanning years long, **compressed to "human** understandable" (**acceptable) time scale** for thesis result delivery
  - Near real-time data production – **priority resources for high stake / immediate feedback physics preview** ("W" physics preview and beam plan tuning)

- Planned additional usage
  - Easy software provisioning at remote sites (FTE consolidation with site scaling)
  - Long term software preservation for post RHIC EOL in the computing plan roadmap

What drew you to infrastructure clouds, i.e., what are the benefits for your community of using infrastructure clouds in order of significance?

- Grid infrastructures are distributed resources with great (initial) ideas but
  - Failed to provide resource elasticity for experiments with complex software stack due to complexity, heterogeneity and lack of troubleshooting features
    - Heterogeneity: configuration (OS, compiler, versions), lack of supporting base libraries (no FORtran library for example) or component libraries (xml, mysql, …) – possible to support with "army of FTE" but not likely to get from base funding.
    - Forced near all communities to pre-installed software at dedicated sites, with configurations tailored to the experiment (STAR is not exempt: PDSF is tailored to STAR needs)
  - Stable production level Physics delivery implies having to validate the experiment **software stack: validation is time costly**
    - "opportunistic" use of resources extremely doubtful as a concept within this fact alone

- Clouds in contrast
  - Has virtualization at its heart
  - Provide a way to **"can" our software in a self-consistent and self-sufficient container** – what we test and get once, we get everywhere
  - Virtualization is a way to **"normalize" the infrastructure** – terrific saving of workforce
  - Virtual Containers VERY attractive for software long term preservation
  - Light approach of Clouds tend to look-like "pull models" (bypassing Grid-RMS) and proven to work and scale
  - STAR demonstrated this many times (resources grabbed at Amazon, Wisconsin, Clemson, NERSC/Magellan, ANL/Magellan)

What are the challenges of using infrastructure clouds for your community, i.e., what makes them difficult to use in order of significance?

- Some (minor?) difficulties at this stage
  - Lack (or expensive / cost prohibitive) storage solutions on commercial clouds (National Lab Clouds would remove this + add proximity to move movement in/our of Mass Storage)
  - Operational support (who do I call if/when something goes wrong?)
  - Handling of VM images not always easy – Some stability in VM formats (or transformation tools) would be welcomed
  - Interoperability (so far, commands and approach are similar but can get out of hand as providers grow)
  - Possibilities to migrate jobs from Clouds to Clouds (previous bullet implied)
  - A consistent Global job monitoring (we developed our own)
  - A consistent AA(A) mechanism

- Note/Other – A true economic model likely needed
  - Beyond commercial cloud costs, higher priority based on cost may not be popular but the only way to "tame" the beast of ever-growing resource hunger from scientists without preventing the elastic nature
  - A "burn chip" mechanism would be good enough

How do infrastructure clouds compare to other options for outsourcing computation from the perspective of your community's needs?

- STAR's experience on Cloud (within Nx1,000 jobs) has been
  - Equal stability than Grid (QoS) like-OSG
  - Similar (or even better when no batch system is involved) job efficiencies as OSG
  - Commercial clouds have showed narrow network bandwidth availability (limits scope of work)
  - As many clouds used, as many support model (or no support / no guarantee) one need to deal with

- Infrastructure cloud Provides true elastic feature
  - Last minute high stake / high return physics possible
  - Students more motivated to work within understandable timelines (not several years of data production)

- Virtualization is essential for long term software preservation
  - This should not be under-estimated (EOL always come soon or later)
  - By using Clouds and building Virtual Containers, we have a "snapshot" of our software AND environment (repository of images + naming convention in STAR)
  - Note – Grid could have used Virtualization (technology was there long ago) – although in discussion for now 5 years, merging of the two worlds (still) possible

Thank you