STAR & Virtualization, looking beyond: Integrating Scientific, Grid, and Cloud Computing Infrastructures

Jérôme LAURET, STAR Experiment
Presenter: Mark L. Green, Tech-X Corporation
Overview

• History, motivations, examples …
  – Why we should all pay attention to this?

• Past & current efforts

• Looking beyond
  – Integrating Scientific, Grid, and Cloud Computing Infrastructures

• Conclusion
History, motivations
why VM?

- Simulation is easy … we can all do it
  - Self-contained event generation especially

- Beyond that, POW often speaks of user analysis and real data production
  - Highest resource demand …

- Realities:
  - Complex experimental application codes
    - STAR case: Developed over more than 10 years, by more than 100 scientists, comprises ~2 M lines of C++ and Fortran code
  - Require complex, customized environments
    - Rely on the right combination of compiler versions and available libraries
    - Dynamically load external libraries depending on the task to be performed (system or third parties: ROOT, mysql, libxml, …)
  - Environment validation
    - To ensure reproducibility and result uniformity across environments is essential
    - Regression tests cannot be done on all OS flavors (due to simple manpower considerations)

- Green+blue = what we would do exclusively on dedicated sites (Grid or local submission)
- 5%-10% of the red = what we could safely do on non-STAR OSG site
- All bright red = what we can do on EC2 today …

STAR resource plan – 09-15

- Year
- CPU (K$/2k)
- Misc external (user)
- Simulation needs
- BNL user analysis (UA)
- BNL calibrations
- Reco total

OSH All-Hand 2009/03 - Virtualization Workshop
History, motivations
why VM?

• Solutions?
  – Smart guy’s comment: redesign your application
    • STAR code already in relative tight control (peer reviews, annual reviews)
    • But sure, please help … and provide the same results of the past 10 years Physics publications
  – Naive solution: package dependencies tools, compile on the fly, pilot jobs, pull models
    • Compilation from scratch takes a long time
      – ~ 8 hours for STAR on a standard dual core Linux node
      – Full knowledge is sparse + there is no single ‘make’ ‘make install’
    • Validation time is NOT included, external factors may influence results
      – Science CANNOT suffer risk
      – Would need to do this ALL the time / waste of resources (test suite in STAR has 25 jobs of 2-4 hours each)

• The only sensible solution: use Virtual Machine
  – Bring a WELL-KNOWN environment with your job
  – Use a pre-VALIDATED experimental software stack
  – Minimize experiment team’s efforts
  – Side advantages
    • Excellent enforcement, performance isolation
    • Good isolation for user analysis

Cloud: EC2 or Nimbus

BNL
History, motivations  
why all should pay attention to …

1. Truly Opportunistic use of resources
   • Surprisingly NOT a STAR problem: Largest stakeholders use 5%-10% non-dedicated resources only
   • Simply NOT possible on a wide scale
     • leveraging each other’s resources or spare cycle exists on paper only. Ex:
       • STAR use of FermiGrid resources in 2007
       • Useful but validation and all took ~ 3 months to sort out OS and environment differences and have a go
       • Overhead is simply too steep
   • ALL experiments with large framework are using DEDICATED sites with pre-installed software

2. Economics of clouds puzzling
   – On demand, pay-as-you-go and just-as-you need, … clouds are growing their offered services
   – 200 CPUs “High-CPU Extra Large Linux” and a 1 TB usage a day (storage, transfer in & out) => 125 k$/months
     • Interesting to compare with existing facility cost …

3. VM was part of OSG’s long term vision POW
   – OSG Project description 2006/05 stated (page 17)
     • E.3 Phase III and IV: Months 37-48 and 49-60: “Create virtual-machine-based sandbox for safe testing on an extended Grid and easy install and teardown of lab resources.”
   – 2006+37 months … is pretty much … NOW!
   – Workshop + CEDPS SciDAC partnership with STAR and work on-target
   – Time to think of the next steps too …
Past & current efforts

- STAR Interested in VM since 2006/2007
  - Discussion started in summer 2007 (CEDPS)
  - First testing was as soon as for CHEP 2007 (CHEP/OSG boot
    - Had run Hijing event generator simulations on EC2.
    - Measured 10% CPU overhead in running in VM comparing to regular
      CPU: negligible impact compared to benefits

- 2008, decent Monte-Carlo full event simulation and
  reconstruction run on EC2 was encouraging
  - Environment
    - Xen image with OSG 0.6.0 CE on SL 4.4
    - Xen image with OSG 0.6.0 WN on SL 4.4
    - Use Globus Workspaces to deploy gatekeeper and worker nodes
      on EC2
    - Launch 100 node cluster in ~ 30 min.
  - Experimental software stack
    - Full STAR software environment on SL4.4 fully validated
    - Overall workload ½ a day install + a few hours for validation
  - Full event reconstruction, NOT only event generation was then done

- 2008 – also performed an IO testing test with standard nodes
  - Results: very poor performance – 5 MB/sec saturation
  - Result independent of tools (tried gsfiftp, srm-copy)
  - IO nodes more expensive – did not test there
Efforts cont.+ pre-conclusion

• 2008/2009 – new attempt to run large production
  – Updated STAR software, was as painless as usual
  – Software deployed, updated & validated with ultimate confidence – all results came to the exact decimal as similar node/OS at BNL
  – Attempted to run and …
    • Many problems with Grid stack

Most recent lesson (re)learn:
• Grid software stack and cryptic error messages still there after N-years of development / troubleshooting is painful
  – CEDPS troubleshooting anyone?
  – Tech-X UCM?
• Virtualization is NOT the “sole” silver-bullet for Grid success …
• VM is “a” step forward providing an essential short-cut to fast “hassle free” deployment of validated Experimental software stack
looking beyond:

• The flexibility offered by a well-designed virtualization infrastructure possesses the capability to deliver significant benefits on several different dimensions at once. These benefits are as follow:
  
  – **Efficiency:** Multiple virtual machines can run on a given server. If performance becomes an issue, it is easy to relocate one of the virtual machines to an underused server. Server utilization can often increase from 10% to 60%-plus.

  – **Reliability:** When a physical machine crashes, the virtual machine(s) running on that computer can be re-started on other machines in the server farm. Cutover is fast, and data can be recovered easily if stored in a storage array. If a virtual machine is infected with a virus, rather than reinstalling each piece of software in sequence, the infected virtual machine can be replaced with a new copy of the pre-configured virtual machine and re-started almost immediately.

  – **Flexibility:** Automated management tools can be used to redistribute virtual machines across the server farm without human intervention, so more or fewer machines can be dedicated to a particular task as needs arise. For example, a surge in hits to a web server can be met by starting new virtual machines with web server software pre-installed.

  – **Testing:** Test environments can be as realistic as need be, and test images can be moved to production without changes, reducing installation errors when applications are deployed.

  – **Disaster recovery:** Easy to mirror current configuration of virtual machines over to a disaster recovery site in real-time, reducing the incidence of out-of-date configurations.
looking beyond:

- A Customizable Web Service for Efficient Access to Distributed Nuclear Physics Relational Databases (CWS4DB)
  - Phase II Nuclear Physics SBIR Award #: DE-FG02-07ER84757
  - Two specific tasks assigned to investigate on-demand capabilities:
    - (1) Utilize Nimbus for deploying and configuring STAR MySQL database servers and evaluate their performance;
    - (2) Provide a publicly accessible web service interface for accepting STAR root4star database queries
looking beyond:

- Scientific, Grid, and Cloud Computing have more similarities than differences.
  - With the advent of Grid Computing it was apparent that scientists struggled with the inherent complexity of the Grid.
  - One of the largest drawbacks was the lack of flexibility in the deployment and setup of complex and delicate experimental software stacks within work environments that are sensitive to scientific assurance of quality results.
  - Opportunistic usage of resources has been severely hindered by such difficulties and sharing of resources has not happened at the level needed to harvest the full power of a national Grid infrastructure.
  - By combining some of the aspects of Cloud Computing with current production grids we believe that this uncertain environment can be overcome with the use of Virtual Machines.
looking beyond:

• Cont’d:
  – Scientific applications can deploy their own predefined known environments within a production grid and thereby significantly eliminating the need for the discovery of “appropriate” environment suitable for their sensitive scientific applications.
  – Furthermore, there is a need for each Virtual Machine based application to be certified, updated, optimized, and deployed efficiently within a controlled environment by the application developer.
  – We believe the these goals can be achieved by leveraging the current grid infrastructure provided by Open Science Grid and extending the Nimbus open source infrastructure for remote deployment and management of virtual machine components.
  – We can do this by developing the Grid Cloud Computing Service (GCCS) for production grids.
looking beyond:

- Integrating Scientific, Grid, and Cloud Computing Infrastructures
  - Phase I Nuclear Physics SBIR Pending
  - Objectives:
    - (1) Deploy and manage heterogeneous application specific baseline STAR Nimbus On-Demand Application Resource (O-DAR) in a grid testbed.
    - (2) Integrate On-Demand Application Resources (O-DAR) within the Open Science Grid Integration Testbed.
looking beyond:

• (1) Deploy and manage heterogeneous application specific baseline STAR Nimbus On-Demand Application Resource (O-DAR) in a grid testbed.
  - We would like to ultimately determine how Virtual Machines integrate with the current Open Science Grid infrastructure where several different software, operating systems, and hardware resources exist.
  - We plan to do extensive testing of the testbed before moving forward to insure that we understand the virtual machine image management limitations, any security concerns, and overall performance and deployment of the testbed.
looking beyond:

• (2) Integrate On-Demand Application Resources (O-DAR) within the Open Science Grid Integration Testbed.
  
  – This is a new type of OSG virtual facility that can be used for cycle scavenging usage on hardware that is idle or migrated out of a production environment and might not even have OSG stack installed.
  
  – It can represent a lightweight method of deploying OSG worker nodes and building more capacity for scientific application usage.
looking beyond:

- (3) Investigate deployment of OSG Grid Cloud Computing Service (GCCS)
Conclusions

• Great progress has been made in past 18 months.
• The STAR, Tech-X, and Nimbus collaboration is proving to be fruitful. Problems still exist:
  – availability of service, data transfer bottlenecks, performance is often unpredictable, rate of scalability, I/O limitations, trouble shooting is very difficult, etc.
• We are still in the early stages of understanding the production quality of service and scalability issues with homogeneous hosts much less heterogeneous environments.
• There is a tremendous amount of front-loaded work involved in developing a viable virtualization environment tailored for demanding scientific applications.
• In the end, we believe virtualization matures it will payoff for scientific applications as its benefits out weigh the alternative approaches.