Executive summary of the 2nd workshop on adapting applications and computing services to multi-core and virtualization

Background and Motivation

Two R&D projects in the PH department (under White-Paper Theme 3) started in early 2008 with strong collaborations between the LHC experiments and other existing structures (e.g. Openlab in IT):

- WP8 - Parallelization of Software Frameworks to exploit Multi-core Processors
- WP9 - Portable Analysis Environment using Virtualization Technology

The project 'kickoff' workshop took place on 14-16 April 2008, where the initial plans for both projects were laid down. Then, the first workshop on adapting applications and computing services to multi-core and virtualization took place one year ago, in which a number of follow up actions were identified.

Progress has been made in developing applications to exploit new hardware architectures and virtualization technology, and this imposes new requirements on the computing services provided by the local computer centers or by the Grids.

Goals

The goals of this workshop are to review the progress of a number of follow-up actions identified in the first workshop, to get new feedback from the experiments and set directions for the two R&D work packages.

Agenda

The agenda with all the material can be found at
http://indico.cern.ch/conferenceDisplay.py?confId=89681

Monday 21 June 2010
13:30->18:30 Welcome, R&D Status and "NEEDS" session
Tuesday 22 June 2010
09:00->12:30 "ACTIONS" Session
14:00->18:30 "FUTURE: Session
Final wrap-up and discussion

Summary of Experiments Requirements

ATLAS requirements (presented by Paolo Calafiura, Sebastien Binet):

- Need for a way to either schedule jobs taking over a complete computing node, or at least know how many slots out of the computing node one can take over.
- The current VOBox service level should be based on VMs.
- CernVM and CVMFS are priorities for ATLAS distributed analysis, and should be supported.
- Virtualization is no replacement for proper software distribution and configuration management, but it can make these tasks much easier and faster.

CMS requirements (presented by Peter Elmer):

- CMS would like to proceed with commissioning/deployment of multicore-aware applications in the next 6 months (by end 2010).
- They propose moving to "whole node" scheduling as part of this commissioning.
- They would like to see improvements to the storage systems and FTS to permit larger files, in particular by resolving the problems with from-scratch restarts after any errors.
- They would like to move beyond VSIZE/RSS for memory accounting (PSS proposal [http://elmer.web.cern.ch/elmer/memory.html], "whole node" accounting needed)
- They see virtualization as appropriate for (most) services, but not as a permanent aspect of high-throughput WN's

LHCb requirements (presented by Marco Clemencic):

- Multi-cored laptops and desktops can already use the parallel Gaudi and Athena. The Grid is not yet ready and therefore new policies must be put in place. Support is needed from the batch systems.
- CernVM should become a standard service supported by the batch systems and transparent for the users. A solid virtualization infrastructure is required to be usable and the cost in terms of performances must be taken into account.

ALICE experience and use cases (presented by Costin Grigoras):

- Excellent experience with VMs for services for compacting rack space and a good environment for building, testing and prototyping.
- Multitude of adopted virtualization platforms (with their positive and negative sides)
- Mastering storage from a VM is still an open issue, especially data servers
- Virtualization is generally accepted for services that are not I/O demanding, also not for DBs.

Main points from the workshop

1. Experiments requested **access to whole nodes rather than single cores**, with the experiments taking responsibility for ensuring that the node is fully utilized. The access to whole nodes would enable them to run the multi-threaded and multi-process applications that are being developed. In addition, they will be able to run their pilot frameworks with the possibility of managing the correct mix of jobs to best optimize the entire node resources.
2. The Grid implementation of whole-node scheduling would require **end-to-end changes** from the end-user submission framework to the local batch system configuration and including middleware. A number of additional issues were mentioned concerning accounting (memory and CPU), proper tools to monitoring and the handling of large files that will eventually result from larger parallel jobs.
3. Virtualization of Services is in general well accepted and the experience so far is very positive. In particular VOBox virtualization is imminent. The requirement for Experiments to Quattor-ize the VOBoxes in order to run applications needs to be re-visited.

4. Virtualization of Worker Nodes is mainly driven by the expected flexibility and efficiency. The LXCloud prototype at CERN-IT is starting to take shape and the results are very encouraging. Questions about the lifetime of the Virtual Machines and the need of an API to control them were raised.
5. Experiments expressed some concern about the performance loss on Virtual Machines, in particular for I/O operations to local disk.
6. Cluster Virtualization is also required for instantiating a more complex configuration than a single VM image (e.g. Proof cluster, production system, etc.)
7. A HEPiX document has been prepared establishing obligations for people providing virtual images. A skeleton schema including contextualization is being put in place. There has been general agreement that the experiment software is treated independently from the base operating system.
8. Experiments requested support for the CernVM File System (CVMFS) as the content delivery mechanism for adding software to a VM image after it has been instantiated.
9. LHCb and ATLAS requested IT department to host and provide 24*7 support for the CernVM infrastructure.
10. A mechanism is needed to track progress on the main points from this workshop. We propose the creation of the following ad hoc discussion meetings listed in the next section.

Spawned ad hoc meetings

1. Understanding VM performance. There is a need to organize a number of discussions and schedule real work to better understand the claimed loss of performance in I/O to local disk for some concrete applications.
2. Preparing the roadmap for the end-to-end whole-node job submission. Taking into account the real status of the experiments' applications, to understand what, when and by whom changes need to be done from the job submission tools to the back-end configuration of the batch systems.
3. Resource accounting and monitoring for virtual machines and multi-core jobs.