

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

### Background and Motivation

Two new R&D projects in PH department (under White Paper Theme 3) in January 2008.

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

The 'kickoff' workshop for this R&D initiatives took place on 14-16 April 2008 and the initial plans for both projects were laid down. Since proper resources are very limited, both projects work as 'collaborations' between LHC experiments and other existing structures (i.e. Openlab in IT).

Both R&D projects, and other initiatives from other groups, have started to deliver useful results for the experiments in exploiting multi-core and virtualization, and it is expected to continue in the near future.

The current computational services at CERN and at the Grid are not ready yet to support this new type of applications. Therefore, it was felt it is the right time to start thinking how to adapt or evolve the existing computing services to support this new requirements.

### Goals

The goals for the workshop have been:

- Familiarize with latest technologies and the current industry trends by inviting technology providers (vendors) to present what exists and what directions the computing industry is going.
- Understand applications from experiments and their new requirements introduced with virtualization and multi-core.
- Initial exploration of solutions adequate to current and future HEP applications

The idea was to be very open and unconstrained and to explore the full solution space with some strategically vision and not just minor adaptations to current services. This may imply moving away from current practices (like batch systems, resource scheduling, accounting)

### Agenda

The agenda with all the material can be found at <http://indico.cern.ch/conferenceDisplay.py?confId=56353>

Wednesday 24 June 2009

13:30->18:30 Welcome and Technology session

Thursday 25 June 2009

09:00->12:30 New Application Requirements

14:00->18:00 Computing Services

Friday 26 June 2009

09:00->11:00 Grid Services

11:15->12:30 Summary, Final Discussion and Wrap-up

## Summary

ATLAS requirements (presented by Paolo Calafiura):

- ATLAS is ready for the transition from multi-core and virtualization prototypes to production
- WLCG should support multi-core job submission
- CERN should add KSM to SLC5 (e.g. via deployment of RH5.4 Grid-wide)
- CernVM should transition from R&D project to official CERN platform
- Concerned about the number of different configurations for providing virtualisation solutions (combinations of hypervisors and host OSs)

CMS requirements:

- It should be possible to allocate a 'complete' node for a single large job (memory or CPU intensive)

LHCb requirements (presented by Thomas Ruf):

- Parallel processing should become fully supported for interactive and batch / Grid processing
- Virtualization seems to be the most efficient way to become independent of the local operating system. Should be very attractive for the Grid to decouple from the OS of the application and the worker node.

Request for special OS configurations and tool deployment

- Performance and other tools are now required by experiments – e.g. KSM, PERFMON, ...
- A test infrastructure is being prepared by IT Dept as a first step
- Further deployment needs to be addressed, e.g. on dedicated clusters

Could batch resources be dedicated for multi-core jobs?

- Overall resources need to be used efficiently
- Proposed to first test LSF functionality for mixed workload (e.g. reserving x cores on a single host)
- CPU should not be the only measure of batch resource utilization (e.g. Jobs may be memory or I/O bound). Should the accounting model be reviewed?
- Support is needed for running multi-core jobs Grid-wide (requirements may be covered by the EGEE MPI WG recommendations)

What is needed for sites to trust VM images?

- Confidence that image has not been 'tweaked'. General agreement that trust can be created for experiment images which follow clearly defined procedures for image creation.
- Traceability and integration with site logging. This may require to 'instrument' images with agreed monitoring utilities that will not interfere with the application itself.
- Sandboxing mechanism could be possible for 'less trusted' images (e.g. no network access for very old images not updated since a long time)
- An agreed (security) update mechanism

## Saving IP address space

- This problem has been identified at CERN where currently all WNs have public IP addresses (some sites already restrict public IP address usage, e.g. to headnodes)
- VMs will significantly increase the use of IP addresses
- Past experience showed private addresses were unacceptable
- Off-site traffic requirements for worker nodes needs to be understood
- Proxy servers can provide a gateway for some off-site traffic, e.g. http(s)
- Sites may be reluctant to run non-standard proxies, but known solutions (e.g. web proxies such as 'squid') should be ok

## Batch model

- Do we need a paradigm shift? – Not clear yet. More experience is needed.
- Move to a cloud-style reservation scheme? – pointed out the difficulty by physicists to know what they need tomorrow.
- Infrastructure as a Service? – Yes.
- Scheduling is still required to decide where to run a VM
- Split resources between traditional batch and supplied VM images?
- Run a batch scheduler inside a cluster of VMs?

## Is CPU usage the right accounting model?

- Need to take account of memory, I/O, infrastructure (power etc), ...
- Wall clock time + charge factor per CPU?
- Slot time?

## Follow up actions

1. **Transition of CernVM beyond the R&D phase.** We can distinguish two parts:
  - a) preparation of CernVM images (including CVMFS)
  - b) support for the infrastructure.We need to follow-up both a) and b) with IT and PH management.
2. **Test of 'parallel' jobs.** This can be done with existing infrastructure. ATLAS and WP8 should participate, in collaboration with IT/FIO, to understand the scheduling of parallel and sequential jobs on the same resources.
3. **Include the capability to run CernVM images** in the virtualized batch nodes initiatives at CERN and INFN .
4. **Prototype an 'lxcloud' solution for submitting jobs using the EC2/Nimbus API** (i.e. prototype a solution which can interoperate with other clouds).
5. **Deploy multi-core performance and monitoring tools** (e.g. KSM, PERFMON) both at CERN and at other Grid sites.
6. **Investigate scenarios for reducing the need for public IP addresses** on WNs.
7. **Establish procedures for creating *trusted* images** (e.g. CernVM) acceptable for Grid sites (i.e. sites agree to run such images).
8. **Provide input to initiatives for running multi-core jobs Grid-wide** (requirements may be covered by the EGEE MPI WG recommendations)