CERN-LCGAPP-2010-03
Final - 02/12/2010

Whole-Node Task Force Proposal

Introduction

The current model by which HEP applications utilize multi-core CPU's is very simplistic. We exploit event level processing parallelism and launch one independent application process per core, each processing independent sets of events. The local batch schedulers are then configured correspondingly to schedule independently (and incoherently) individual "jobs" on each core.

This simple method has been very effective and has clear benefits. However, as the number of cores/CPU increases with each new CPU generation, including continual increases in the total amount of physical memory needed per box, an ever increasing number of incoherent readers and writers (to local disk and/or remote storage) will not scale sufficiently well in memory usage.

In order to more efficiently use multicore CPU's, the LHC experiments have been developing multicore-aware applications capable of exploiting more than a single core via multi-processing and/or multi-threading.

One outcome of the "2nd Workshop on adapting applications and computing services to multi-core and virtualization[1]" was a request from the experiments to have access to the whole node rather than just single cores. This would enable the use of these multi-process and/or multi-threaded applications and permit maximum optimization in the use of the entire node's resources, at the same time, taking the responsibility that the node is fully utilized.

It is expected that switching to "whole node" scheduling instead of single core will require end-to-end changes from the submission framework to the local batch configuration and Grid middleware. In addition it is expected that memory and CPU accounting and monitoring will need to be adapted and handling of larger files will be necessary for the larger parallel jobs.

We thus propose a "Whole-Node Task Force" involving the experiments, grid service providers and sites, with the following mandate.

Mandate

- Follow the "commissioning" of multi-core jobs being carried by the experiments. Facilitate discussion between the experiments and the computing resource providers to understand the specific problems encountered, realistic requirements and ingredients for an eventual "whole-node" deployment.
- Understand what, when and by whom changes need to be made in the full job submission chain. The idea is to consider the complete chain starting from the job submission user interface tools, and finalizing to the concrete configuration of the batch systems at the Grid sites.
- Prepare a roadmap for the deployment of end-to-end whole-node job submission taking into account the real status of the experiment applications.

---

[1] http://indico.cern.ch/conferenceDisplay.py?confId=89681

- Propose new resource accounting and monitoring for multi-core jobs. The traditional accounting schemas are probably no longer adequate when the experiment allocates the complete node and takes responsibility of make proper use of all the resources. New methods will need to be discussed and agreed.

Task Force Composition

We propose the following composition for the task force:

- One or two people representing each of the LHC experiments (one LHCb and ALICE and two ATLA and CMS)
- One person representing each of the CERN-IT groups relevant for this Task Force (e.g. ES, GT, PES)
- One person representing the LCG Applications Area.
- One representative of each major computing site interested in participating to the Task Force.

Timescale

The goal is to draw conclusions in about 6 months, by May, 2011.