PROCEEDINGS OF SCIENCE

A prototype of a dynamically expandable Virtual Analysis Facility

Dario Berzano[1]
Dipartimento di Fisica Sperimentale, University of Torino and Istituto di Nazionale di Fisica Nucleare
Via Pietro Giuria, 1 Torino Italy
E-mail: dario.berzano@gmail.com

Stefano Bagnasco[2]
Istituto di Nazionale di Fisica Nucleare
Via Pietro Giuria, 1 Torino Italy
E-mail: bagnasco@to.infn.it

Stefano Lusso
Istituto di Nazionale di Fisica Nucleare
Via Pietro Giuria, 1 Torino Italy
E-mail: lusso@to.infn.it

Massimo Masera
Dipartimento di Fisica Sperimentale, University of Torino and Istituto di Nazionale di Fisica Nucleare
Via Pietro Giuria, 1 Torino Italy
E-mail: masera@to.infn.it

PoS(ACAT08)050

Current Grid deployments for LHC computing do not allow efficient parallel interactive processing of data. In order to allow physicists to access interactively subsets of data, for tasks like algorithm tuning or debugging before running over a full dataset, parallel Analysis Facilities based on PROOF have been deployed by the ALICE experiment at CERN and elsewhere. Whereas large Tier-1 centres can afford to build such facilities at the expense of their Grid computing clusters, this is likely not to be true for smaller Tier-2 centres. Leveraging on the virtualisation of highly performant multi-core machines, it is possible to build a fully virtual Analysis Facility on the same Worker Nodes that compose an existing Grid cluster. Using the Xen paravirtualisation hypervisor, it is then possible to dynamically move resources from the batch instance to the interactive one when needed. We present the status of the prototype being developed.

XII Advanced Computing and Analysis Techniques in Physics Research
Erice, Italy
3-7 November, 2008

[1] Speaker
[2] Corresponding author

1. Introduction

Interactive analysis on the Grid is far from having a production-grade infrastructure, even though several developments exist. For projects in contexts related to the activities described in this paper, see for example [1] for a medium-scale EU project and [2] for an analysis of limitations and perspectives of the current EGEE deployment. However, the need for interactive and very fast turn around facilities arises from several HEP-related activities, such as algorithm tuning, code debugging and quick data quality assessment.

In order to provide parallel interactive analysis resources to the experiment community, the ALICE experiment [3] deployed a Central Analysis Facility at CERN [4] based on PROOF [5], an extension of the ROOT system [6].

The ALICE experiment Computing Model [7] comprises three “tiers” of computing centres. The Tier-0 is CERN, Tier-1s are large regional computer centres with custodial storage capability (usually tape libraries) and thousands of job slots, whereas Tier-2s are smaller sites with disk-only storage and up to a few hundred job slots. In the Computing Model, Tier-0 and Tier-1 centres are mostly dedicated to centrally-managed scheduled activities, while “chaotic” end-user analysis takes place at Tier-2 centres. Most of the computing resources are provided as Grid Worker Nodes, part of the WLCG infrastructure [8]. Since small Tier-2 infrastructures cannot afford to dedicate statically a significant part of their resources to interactive data analysis, an alternative approach has to be pursued. One should envisage a way to dynamically allocate some of the resources to interactive work, at the same time avoiding the wasting of resources associated with waiting for a sufficient number of CPUs to become available as batch jobs finish. A possible solution to this issue, based on virtualization and dynamical resource allocation is presented in this paper; a recently deployed prototype, some tests run on it and some issues encountered are described.

While this specific Virtual Facility prototype is dedicated to the ALICE use case, it should be noted that the concept is completely generic and does not depend on the application.

2. A Virtual Analysis Facility Prototype

Virtualization provides a simple way of addressing the issue described in the introduction, by dedicating each physical machine to more than one application.

A prototype was built to assess the feasibility of a Virtual Facility on top of an existing Grid batch farm, with resources that can be moved dynamically between the two applications as a function of the demand for each of the services. The concept and prototype are described in this section.

2.1 Virtualization platform

The virtualization tool of choice is Xen 3.2, a widely used open-source virtualization platform and hypervisor originally developed at the University of Cambridge and now commercially distributed by Citrix Systems, Inc. [9]. The advantage of virtualization through Xen are twofold.

First, hardware-assisted paravirtualization approach adopted by Xen allows one to virtualize CPU resources with in principle no performance loss, at the relatively small price of running a special version of the Linux kernel that is aware of being run on a virtual machine.

Second but foremost, the Xen hypervisor is able to reallocate dynamically resources from one virtual machine to another without the need to stop or reboot the OS instance. The CPU priority and sharing across virtual instances can be changed at runtime, and since all domains share the same Credit Scheduler instance, scheduling is extremely efficient and flexible. What is probably even more important, also memory resources can be dynamically allocated and moved from one virtual instance to another relatively quickly, which is what differentiates this approach from simply using system-level tools to change the relative priority of processes running on the same machine.

In Xen jargon, the privileged domain that runs the scheduler and the hypervisor is called *dom0*, while the unprivileged domains that run the virtual instances are called *domU*s.

2.2 The Virtual Facility concept

In this application, each of the physical machines hosts two domUs. One runs the gLite 3.1 middleware suite for a Worker Node [11], while the other runs PROOF for parallel interactive work. The Grid node is fully included in the WLCG production infrastructure as a Worker Node of the INFN Torino site, which is an ALICE Tier-2 centre that supports also other Virtual Organizations. The PROOF node is attached to a PROOF master running on a head node, that acts also as a user interface onto which users log in to use the facility.

In the default configuration nearly all of the resources (CPU scheduling priority and physical memory) are attached to the Grid Worker Node, which appears, in this configuration, almost identical to the other WNs in the Grid farm. By simply using scripts from the head node it is possible to reassign some of the resources from the WN to the PROOF node. While resources become scarce for the processes running on the WN, they slow down and start swapping to disk, but should not crash. At the same time, the PROOF-running machine quickly becomes available for interactive use. Conversely, when it is not needed anymore, resources can be moved back to the WN, that resume processing at full speed.

An alternative solution could consist in fully suspending a virtual machine and then resuming the other one. Although this sounds appealing, primarily because neither memory erosion nor intensive disk swap usage occurs, most Grid jobs need to access data located on remote servers: suspending a virtual machine while a remote transaction is in progress could lead to job crash, unexpected results or even data loss. One could code jobs that are aware of transaction interruption, but existing simulation and reconstruction programs are mostly unaware of that, thus the better solution under a practical point of view seems to be the VM slowdown, not the suspension, even if under the computing point of view the latter is clearly more efficient.

2.3 The Prototype

The current deployment of the prototype comprises four HP ProLiant DL360, equipped with two quad-core Intel, 8 GB RAM (a memory upgrade is foreseen in the next future) and two high-speed SAS disks each. Each of the disks is dedicated to only one of the virtual machines, in order to minimize interference. This is needed since, when deprived of most of its memory, the WN would rely very heavily on swap space, thus generating large amounts of disk traffic.

Management, user access, monitoring and some other ancillary services are provided via a separate server, that acts as "head node" for the infrastructure. It should be noted that no "provisioning" is done in this prototype, i.e. the machines are statically configured and installed, and resources are moved only within each physical machine.

On the PROOF side, two different data access schemas are available. In one, similarly to what is done at the CERN CAF, data are replicated on a pool of local disks residing on the Virtual Facility machines. The whole of the disks is seen by the system as an xrootd pool [12], and data access is always local. In the second schema, data are accessed from a close Grid Storage Element through the xrootd protocol. Even though the first schema is clearly more performant[1], installing a sizeable amount of disk onto the blade servers commonly used as Worker Nodes may not be feasible, so the second one is explored in this paper.

3. Tests

Before actual deployment, feasibility has been assessed by measuring performance loss for typical High Energy Physics jobs in a virtualized environment. Subsequently, the performance of PROOF jobs with a few different memory and CPU sharing schemas have been tested. This section describes these tests and their results.

3.1 Feasibility checks

Relative performance of the non-virtualized and virtualized system was measured using SysBench [13], a simple and scalable modular benchmark tool for evaluating OS parameters, originally developed to test systems running the MySQL database. Three different tests were run: a primality test on the first 20000 integers to test CPU usage, a test with several concurrent threads writing 5GiB chunks of data in memory, and a similar I/O test to disk.

As expected, the tests showed that there is no loss in CPU efficiency, while disk and memory I/O performance is lower in the virtualized environment.

3.2 Prototype benchmarks

The most interesting tests are those run on the fully deployed system, with some load running on the virtual Worker Node and a PROOF parallel session on the virtual PROOF slave; the measured quantity here is the PROOF analysis event rate. We present here the first preliminary results of the tests, since the testing activity is still going on at the time of writing.

Such tests have been carried out in two different conditions: in the first, the WN was running a fake load generator, in the second the WN was included in the production Grid Farm of the Torino site, running standard production jobs form the Grid. While most of activity came from ALICE simulation jobs, there were also some mixed jobs from other VOs, mostly BioMed and LHCb. The result from the first batch of tests is summarized in fig. 1.

[1] This is mainly because PROOF is aware of data location and sends tasks close to their pre-staged data, so that data access is actually locat lo each PROOF salve and no network traffic occurs.

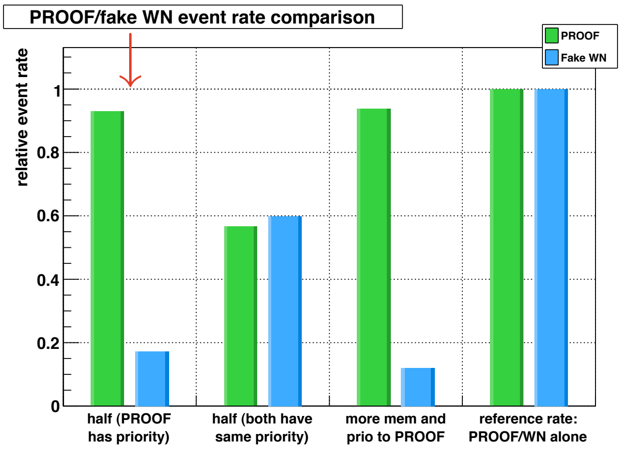

Figure 1. PROOF event rate comparison with a fake load on the virtual Worker Node.

With a load generator running on the WN side, three different conditions were tested:

1. Half of the memory to the WN, half to the PROOF slave, the PROOF slave has most of the CPU priority;
2. Same memory configuration, same CPU priority to both domUs;
3. Most of memory and CPU priority to the PROOF slave.

Since the PROOF analysis used for the tests is not very memory-demanding, it can be seen that the event rate depends only on the CPU scheduling priority, and that when sufficient resources are given to the PROOF slave, the loss of computing performance is again minimal.

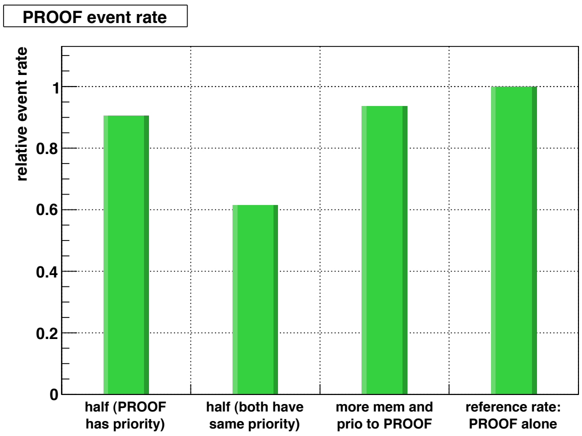

Figure 2. Proof event rate comparison with production Grid jobs running on the virtual Worker Node

When running with real jobs on the Worker Node, the results, as summarized in fig. 2, are not much different. The WN performance measurement here is not available because of the heterogeneous nature of the production jobs running on it.

By comparing fig. 1 and 2, it is important to note that the rates do not change significantly from a situation in which a purely CPU-bound load generator is running on the WN and a real

Grid job. Since most of the latter are ALICE jobs, that are extremely demanding in terms of memory and thus heavily swapping to disk in condition of memory scarcity, one can evince that Xen and the physical separation of disks are effective in insulating the performance of the two virtual machines.

As a qualitative estimate of the performance loss due to the reduction of available memory, in fig. 3 we plot the CPU time over Wall Clock time ratio for batch jobs running on the Worker Node, sometimes referred at as *CPU efficiency*. Jobs are taken into account for the plot if they have a duration longer than a given threshold (3 minutes) and a CPU ratio higher than 0.

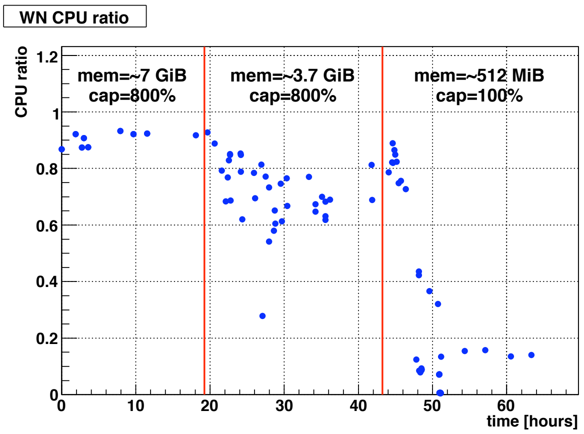

Figure 3. CPU time over Wall Clock time ratio for production Grid jobs running on the virtual Worker Node, in three different configurations.

At the points marked by the vertical lines, the running conditions are changed according to the following sequence:

1. 7 GiB of memory out of 8 and maximum CPU “cap[2]” (800%)
2. 3.7 GiB and 800% CPU “cap”
3. 0.5 GiB and 100% CPU “cap”

As expected, as swap usage increases because of physical memory scarcity jobs turn from being CPU-bound into I/O-bound. The rate of crashing jobs did not increase during the test.

4. Conclusions and outlook

To cater for the need of parallel interactive Analysis Facilities at medium-sized computer centres hosting WLCG Grid Sites, we developed the prototype of a system in which resources can be dynamically reassigned between two virtual clusters sharing the same physical hosts, one being a standard Grid Computing Element and the other an interactive PROOF cluster. The

[2] CPU priority is assigned by the Xen Credit Scheduler according to two tunable parameters, the “priority” proper and the “cap”, i.e. the maximum amount of CPU priority available even if there are idle CPU cycles. In this context, 800% means full access to all 8 cores of the physical machine.

early tests presented in this paper demonstrate the viability of the concept, based on the Xen virtualization platform, that allows reallocation of memory as well as CPU scheduling priority.

Along with more benchmarks and tests, next steps will include opening the facility to a small number of users to test the real-world usability of the system in different deployment configurations. Further developments include an improved management and monitoring interface, automatic reallocation of resources based on load and integration with existing Grid Resource Usage Accounting infrastructures.

References

[1] The Interactive European Grid Project: `http://www.i2g.eu` and references therein.

[2] C. Germain-Reanud, R. Texier, A. Osorio, C. Loomis, *Grid Scheduling for Interactive Analysis*, in *Challenges and Opportunities of HealthGrids*, proceedings of *Healthgrid 2006. Studies in Health Technology and Informatics* **120**:25-33 (2006)

[3] K Aamodt *et al.* (The ALICE Collaboration), *The ALICE experiment at the CERN LHC*, *JINST* **3** S08002 (2008)

[4] J.F. Grosse-Oetringhaus, *The CERN Analysis Facility — A PROOF Cluster for Day-one Physics Analysis, J. Phys.: Conf. Ser.* **119**:072017 (2008)

[5] M. Ballintijn *et al.* "Parallel interactive data analysis with PROOF", *Nucl. Instr. Meth.* **A 559**, 13-16 (2006)

[6] R. Brun and F. Rademakers, *ROOT - An Object Oriented Data Analysis Framework*, in proceedings of *AIHENP'96 Workshop*, Lausanne, Switzerland (1996). *Nucl. Inst. & Meth. in Phys. Res.* **A 389** (1997) 81-86. See also `http://root.cern.ch/.`

[7] The ALICE Collaboration, *ALICE Technical Design Report of the Computing*. CERN-LHCC-2005-018 (2005)

[8] `http://lcg.web.cern.ch/LCG/`

[9] `http://www.xen.org/` and `http://www.xensource.com/`

[10] P. R. Barnham *et al.*, *Xen 2002*, University of Cambridge Computer Laboratory Technical Report UCAM-CL-TR-553 (2003)

[11] `http://glite.web.cern.ch/glite/`

[12] A. Dorigo, P. Elmer, F. Furano and A. Hanushevsky *XROOTD/TXNetFile: a highly scalable architecture for data access in the ROOT environment* in proceedings of the *4th WSEAS international Conference on Telecommunications and informatics*, Prague, Czech Republic, 2005.

[13] `http://sysbench.sourceforge.net/`