A PROTOTYPE OF A VIRTUAL ANALYSIS FACILITY

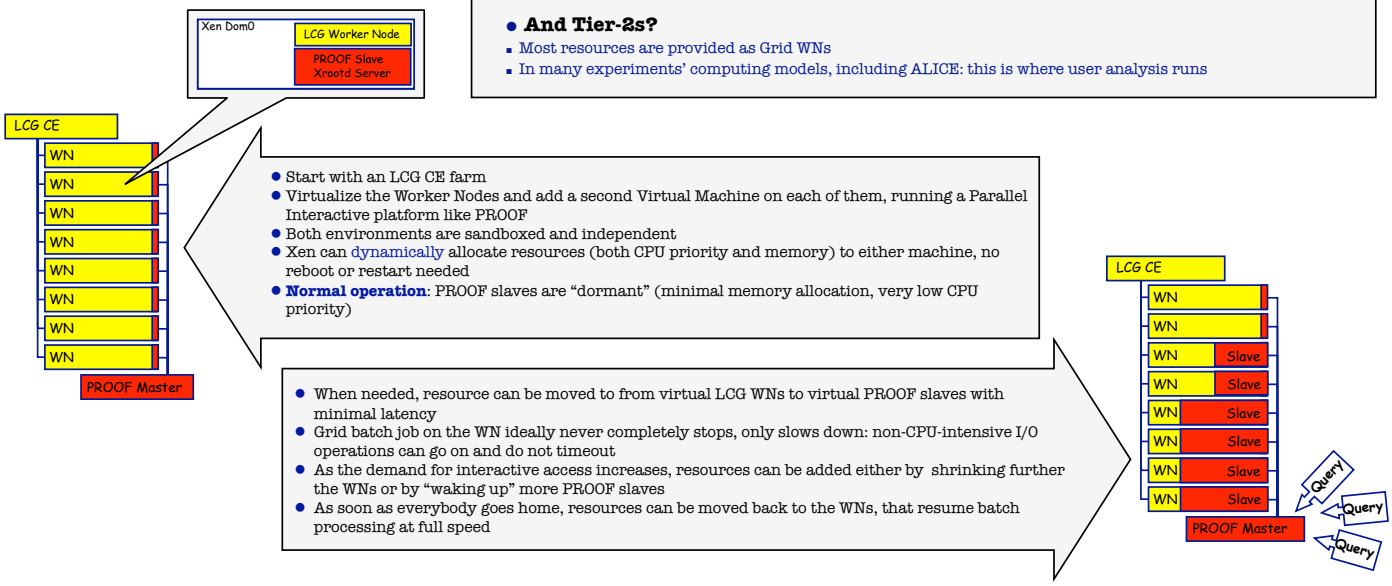
Stefano Bagnasco, Dario Berzano, Stefano Lusso, Massimo Masera

Istituto Nazionale di Fisica Nucleare, Torino and Dip. di Fisica Sperimentale, University of Torino

Abstract - Current Grid deployments for LHC computing (namely the WLCG infrastructure) do not allow efficient parallel interactive processing of data. In order to allow physicists to interactively access subsets of data (e.g. for algorithm tuning and 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 may afford to build such facilities at the expense of their Grid farms, or exploit the large number of jobs finishing at any given time to quickly collect a number of nodes to temporarily allocate for interactive work, this is likely not to be true for smaller Tier-2s 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 LCG Grid Farm. Using the Xen paravirtualisation hypervisor, it is then possible to dynamically move resources from the batch instance to the interactive one when needed, minimizing latencies and wasted resources. We present the status of the prototype being developed, and some experience from the very first users.

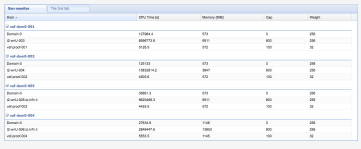
INTERACTIVE ANALYSIS IN THE TIERED MODEL

- **At Tier-1s**
  - Large Computer Centres with thousands of CPUs
  - Feasible to take some out of the Grid infrastructure to build a PROOF-based Analysis Facility
  - Or may even be possible to "drain" jobs and switch to interactive mode quickly as soon as jobs finish and WNs become available
- **At Tier-3s**
  - Very small number of CPUs
  - Probably not even a Grid site, at least with gLite middleware: use PROOF for parallel processing
- **And Tier-2s?**
  - Most resources are provided as Grid WNs
  - In many experiments' computing models, including ALICE: this is where user analysis runs



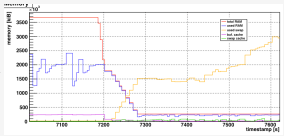
THE PROTOTYPE

- **Hardware**
  - 4x HP ProLiant 360DL, dual quad-core, plus one head node for access, management and monitoring
  - Separate physical 146GB SAS disk for each virtual machine performance isolation
  - Private network with NAT to outside world (currently including storage)
- **Software**
  - Linux CentOS 5.1 with kernel 2.6.18-53 on dom0
  - Xen 3.0
  - gLite 3.1 Worker Node suite
  - PROOF/Scalla 20090217-0500
  - Custom monitoring & management tools

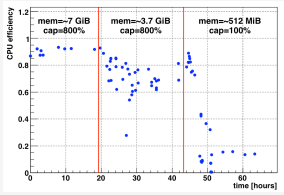


MOVING RESOURCES

- Moving resources from the Wn to the PROOF slave, as seen by the WN
- About 1.5 minutes to complete transition (but the slave can be started almost immediately)
- Swap usage continues to increase afterwards

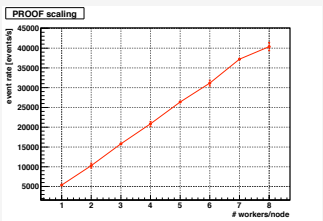


- CPU efficiency (CPU time/wall clock time) for regular ALICE MC production/reconstruction jobs, in three different resource configurations
- Jobs become increasingly I/O-bound as swap activity increases
- No abnormal job terminations observed (above what is seen in regular WNs)



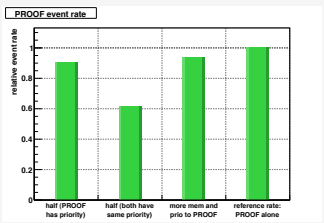
FIRST RESULTS

- **Main advantage: no big development needed, all tools are production-grade already**
- Prototype deployment did not require any development



- Parallel interactive analysis on PROOF nodes
- Regular load on WNs (mix of memory-hungry ALICE production jobs, ALICE user jobs and some jobs from other VOs)
- PROOF scaling OK (saturates only when # of workers = # of CPUs)

- PROOF processing rate in several resource configuration
- Separate physical disks ensure minimal impact from swapping
- System behave as expected: Grid jobs go on running, even though very slowly, and resources are efficiently exploited by PROOF



➜ **Contacts:** [bagnasco|berzano|masera|lusso]@to.infn.it