CERN openlab II

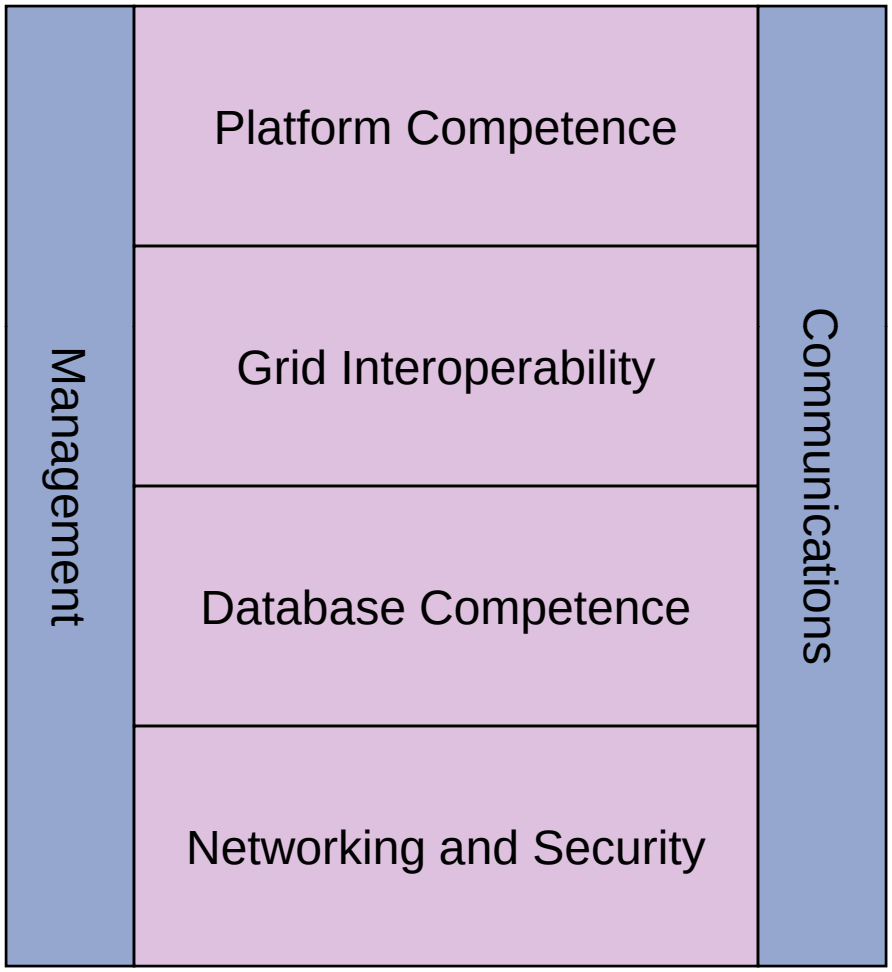
Summary of Technical Achievements

Sverre Jarp, 24 April 2008

CERN openlab CTO

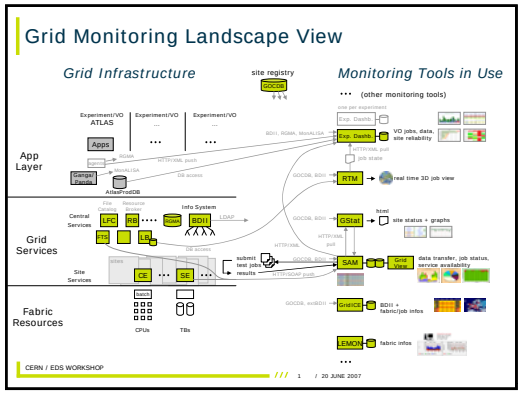
sverre.jarp@cern.ch

openlab II structure

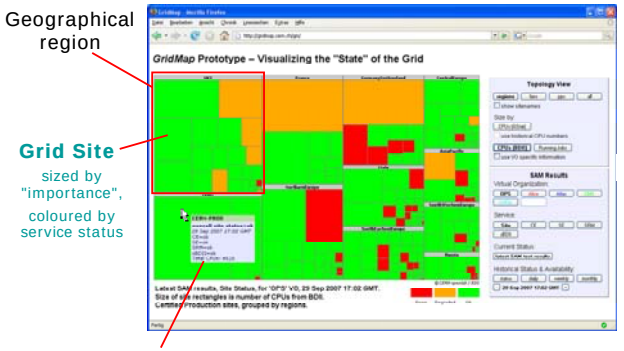


Grid Monitoring

- Analysis – Current Situation
  - Grid Monitoring Landscape
    Q2 2007, CERN openlab / EDS Workshop



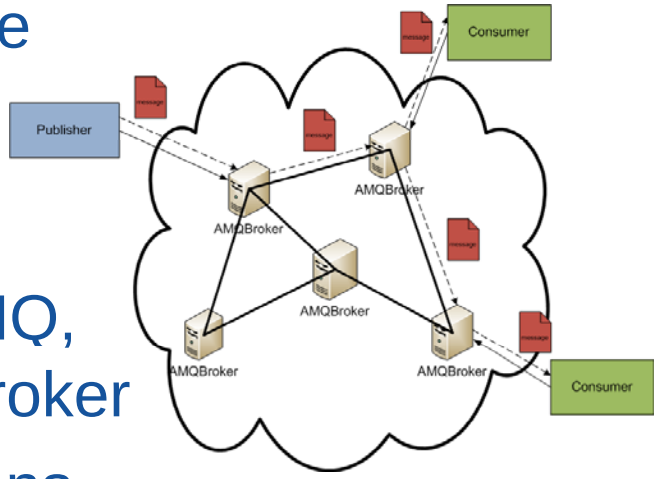
- New Monitoring Management Views
  - Developed *GridMap* Prototype
    Q3 2007
  - Presented at EGEE'07
    Oct 2007, Demo+Talks
  - Documentation, Releases
    Q4 2007
  - Variants: ServiceMap, ...
    Q1 2008



Live link: http://gridmap.cern.ch

MSG

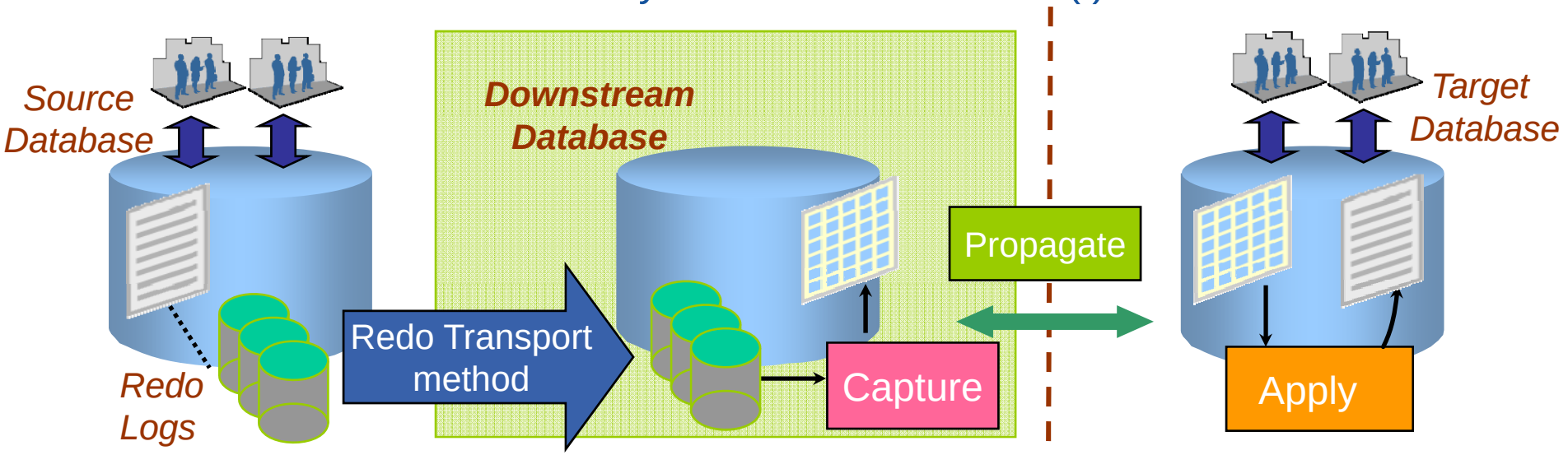
- MSG: ‘Messaging System for the Grid’
  - Objective: Integrate different monitoring tools using a reliable infrastructure



- Work started Sep 07
  - Extensive testing of ActiveMQ, an open-source message broker
  - Prototype of different solutions (mainly Python)
  - Currently OSG and Gridview production data is being published and consumed

Database downstream capture and Network Optimizations

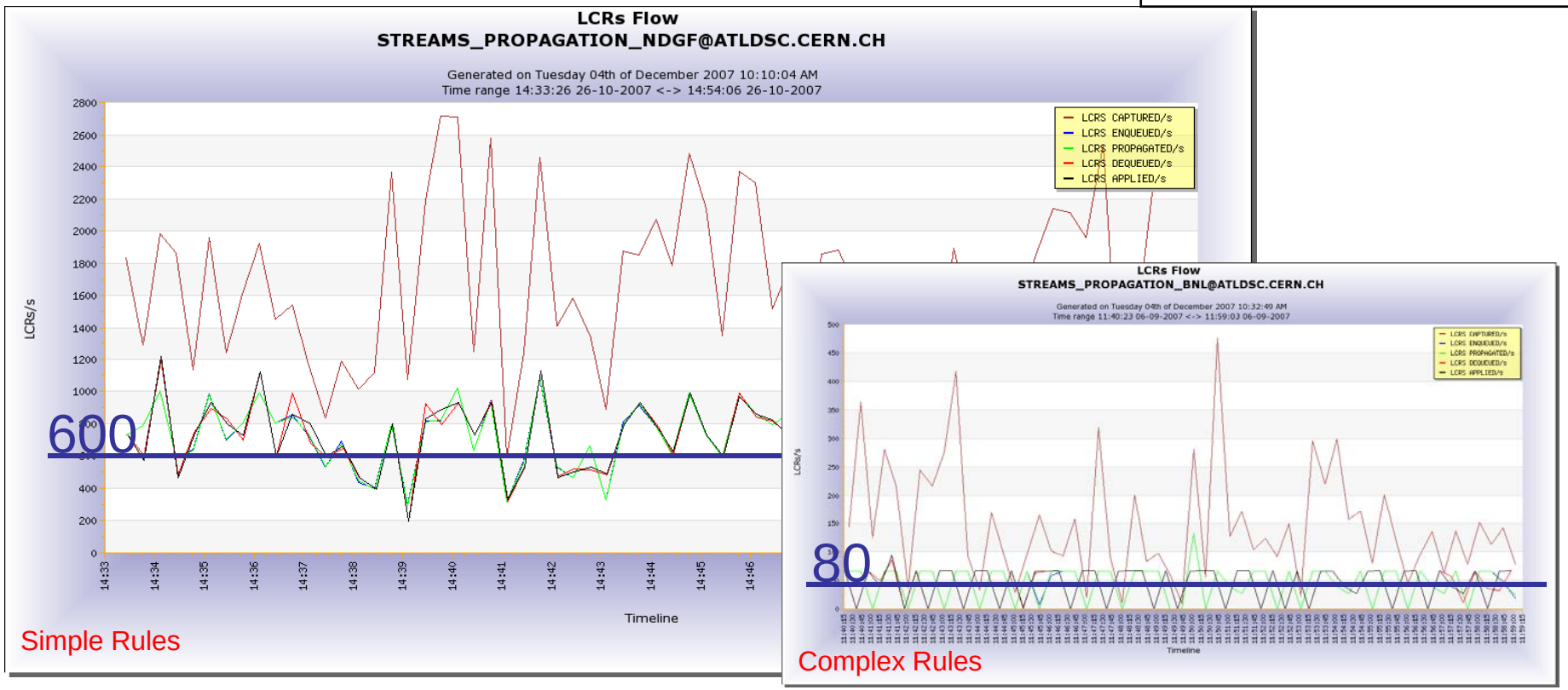
- Downstream capture to de-couple Tier 0 production databases from destination or network problems
  - source database availability is highest priority
- Optimizing redo log retention on downstream database to allow for sufficient re-synchronisation window
  - we use 5 days retention to avoid tape access
- TCP and Oracle protocol optimisations yielded significant throughput improvements (factor 10)
  - network latency to some sites 300 ms(!)



Oracle Streams Rules Optimizations

- **ATLAS** Streams Replication: filter tables by prefix
- Rules on the capture side caused more overhead than on the propagation side
- Oracle Streams complex rules: rules with conditions that include LIKE or NOT clauses or FUNCTIONS
- Complex rules converted to simple rules

LCR: Logical Change Record.

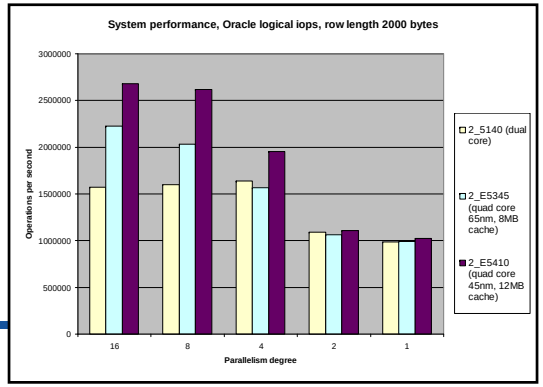


Oracle Streams Monitoring

- Requested features:
  - Streams topology
  - Status of streams connections
  - Error notifications
  - Streams performance (latency, throughput, etc.)
  - Other resources related to the streams performance (streams pool memory, redo generation)
- Architecture:
  - “strmmon” daemon written in Python
  - End-user web application http://oms3d.cern.ch:4889/streams/main
- 3D monitoring and alerting integrated with WLCG procedures and tools

Oracle RDBMS highlights

- Oracle RDBMS
  - Beta testing of 11g and 10.2.0.4
    - Workload Capture and Replay testing with PVSS and Castor Name Server workloads
    - IO Resource Manager Calibration testing
  - PVSS RAC scalability work continued, presented at UKOUG'07
  - Configuring and testing Oracle RAC in XEN virtualized environment
  - Performance testing on new quad core processors
  - 11g rpm testing and deployment



Oracle Enterprise Manager

- Oracle Enterprise Manager
  - Migration to high availability architecture on Linux & presentation at European EM user group
  - Upgrade to 10.2.0.4
  - Increased use of user defined metrics, custom reporting, and security policies
  - Big win: Databases monitored for backup activity - alert if time limit elapsed
  - Joint presentation with Configuration Management team at Oracle OpenWorld

CINBAD Achievements

Packet Sampling Studies

- Over 100 technical papers read and analysed
- Thorough Technical Report written

Understanding sFlow data sources

- Analysis of sFlow agents
- Simulation of sampling

CINBAD Achievements

Survey on various definitions of an anomaly

Survey of data acquisition at CERN

- Try to benefit from CERN experience in data acquisition

Scalable collector design

- Initial design of robust and scalable structure

10 Gb Networking

- With the first generation cards, we successfully prototyped high-throughput disk servers, but …
  - Very high cost
  - Reasonable throughput required jumbo-frames
    - MTU 9KB, rather than 1.5KB (Ethernet standard)
- Production disk servers (w/1Gb NICs) have now reached their throughput/capacity limit
- Today, we know that 2nd generation cards are much better
  - Native speed (9.49 Gbps) reached with standard MTU
  - Driver support native in Linux kernel
  - Reasonable cost, especially with CX4 cards

Grid Scheduling Survey

- X.Gréhant's PhD:
  - Synthesis on Grid Scheduling
  - VO management, resource access
    - EGEE, OSG, NorduGrid, Naregi, etc.
  - Direct scheduling in a VO
    - glideCAF, Cronus, GlideInWMS
    - AliEn2, DIRAC, Panda
    - DIANE
  - With the help of several grid developers at CERN
  - Submitted to the Journal of Supercomputing

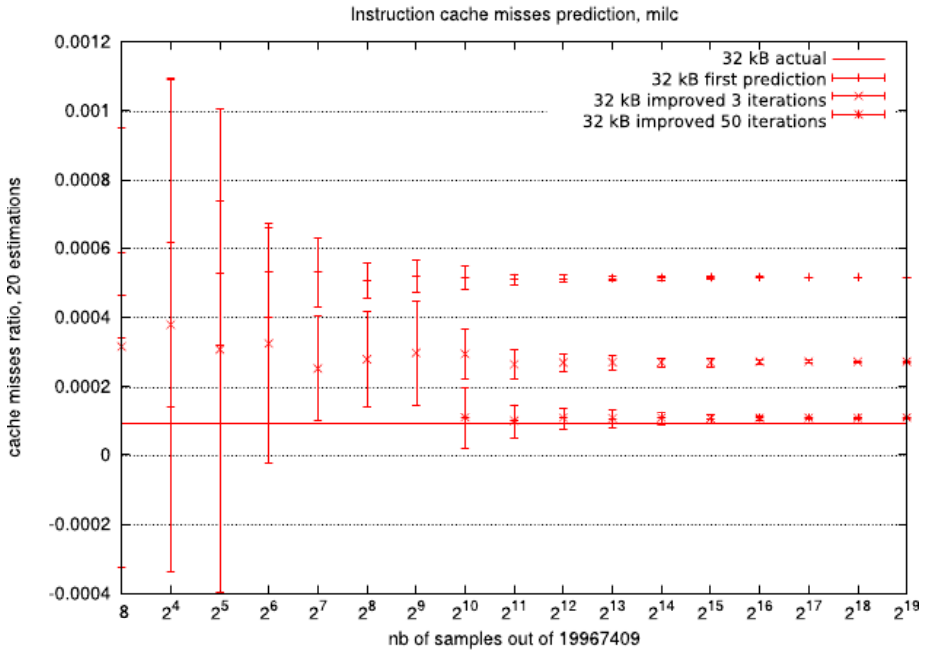
VO: Virtual Organization, federation of users.

Grid Resources Simulation

- Resource supply / consumption is heterogeneous
  - Benefits of careful allocation and migration?

- Design of a resource model
  - Performance prediction from microarchitecture to the grid level
  - Now evaluating a contribution in cache misses prediction

- Development of Levellab, a discrete-time simulator
  - Designed to compare the performance of grid schedulers
  - Realistic: lower-level resources accurately modelled



Grid Resources Deployment

- Resource availability is transient
  - Resilient service deployment
- Design of a P2P resource election mechanism
  - Decides where to (re-)deploy a service
- Development of SmartCitizens, based on SmartFrog

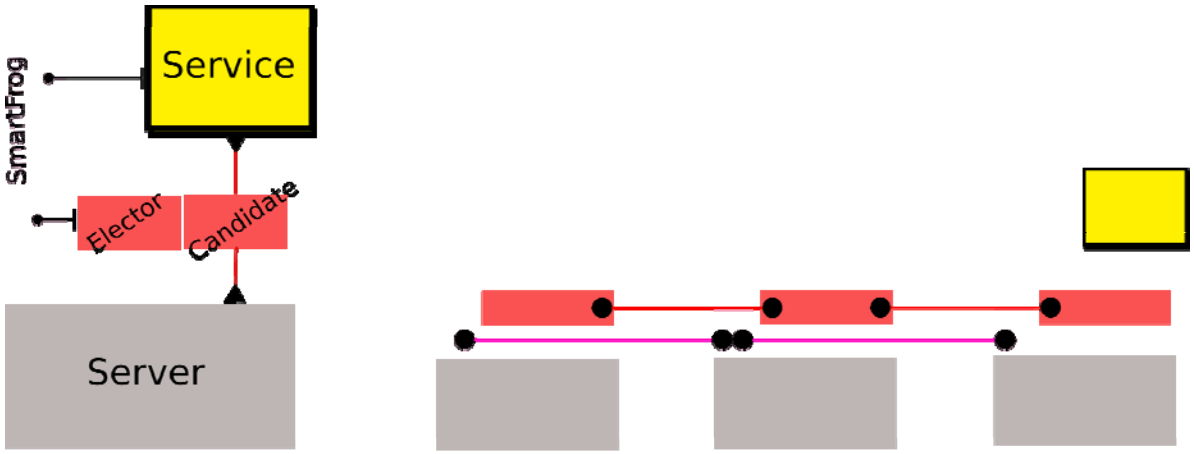


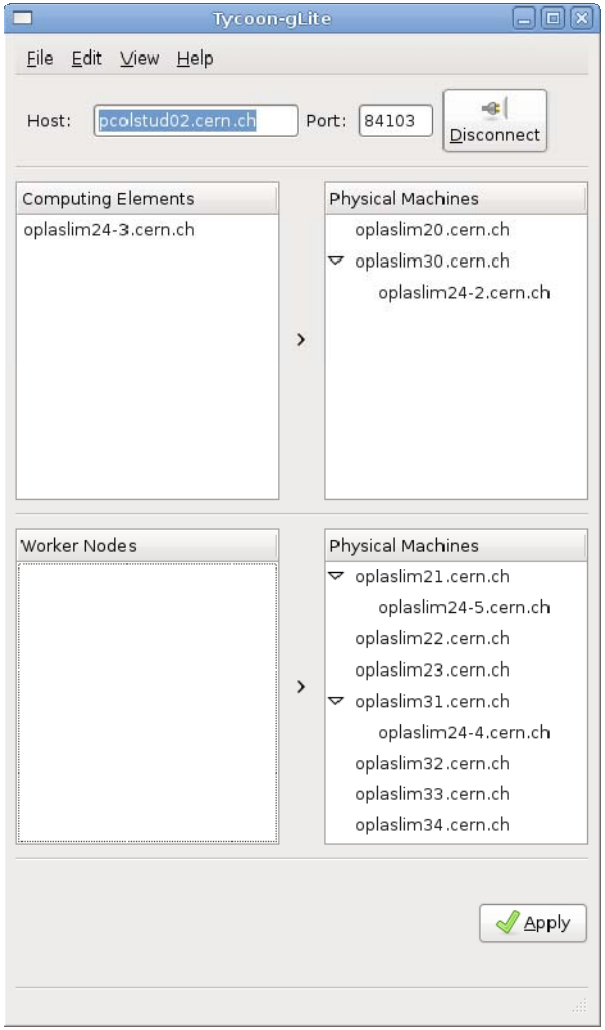
Figure: SmartCitizens Integration inside a node, and between nodes

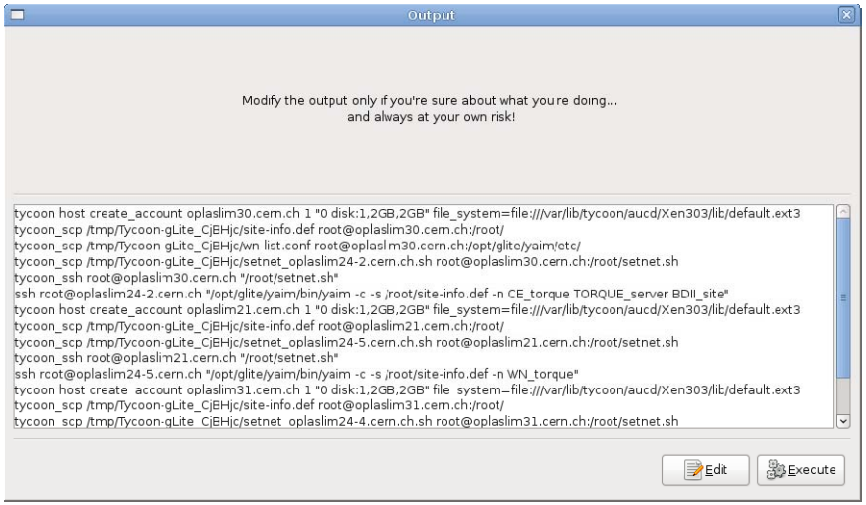
Tycoon summary

- A comprehensive technical report covering Tycoon activities in CERN openlab in 2007 and future plans for 2008 was produced for HP Labs (Palo Alto):
  - Collaborations (HP Labs, EGEE, BalticGrid)
  - Tycoon-gLite integration
  - Scalability tests
  - Issues concerning security and trust
  - Conference attendance
- Several modifications added to the Tycoon-gLite implementation

Tycoon-gLite integration

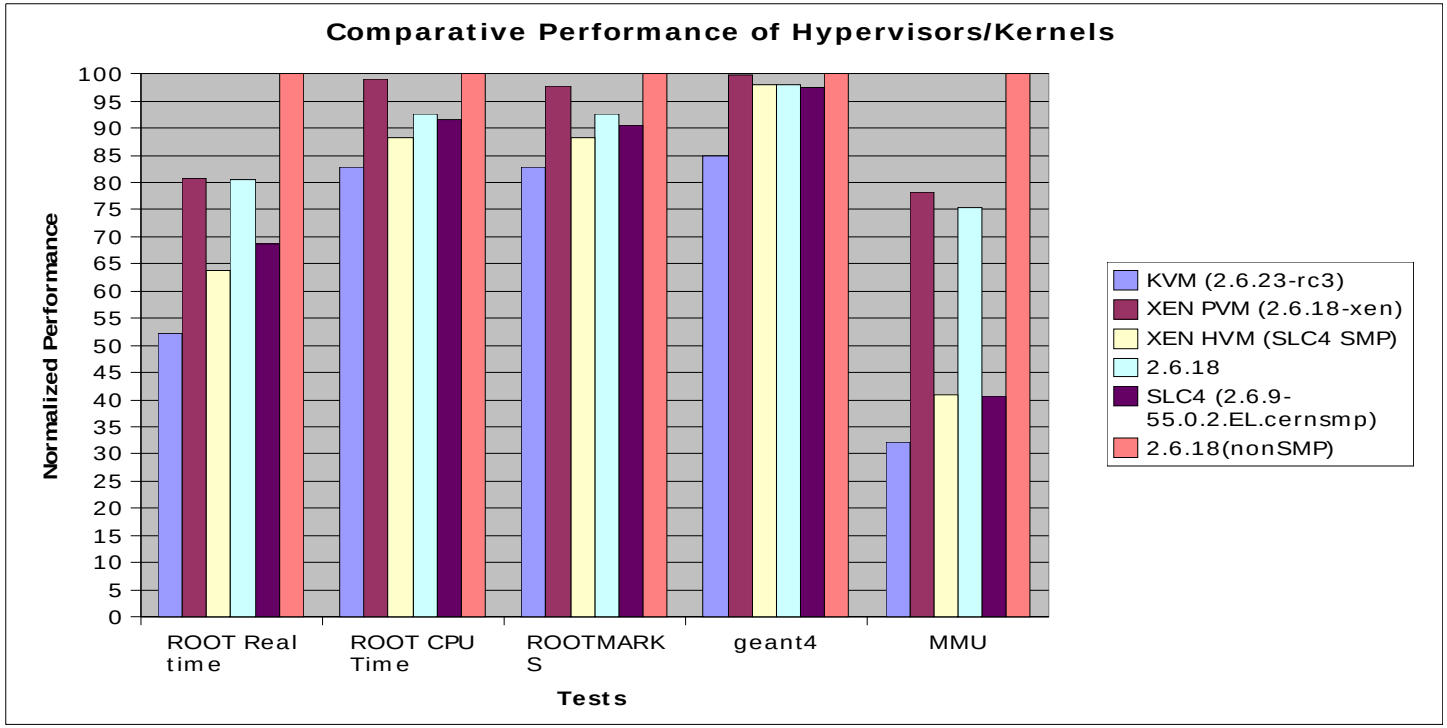
- The implementation was enhanced in order to:
  - Deploy different kinds of nodes more easily (i.e. Storage Elements)
  - Allow modification of output





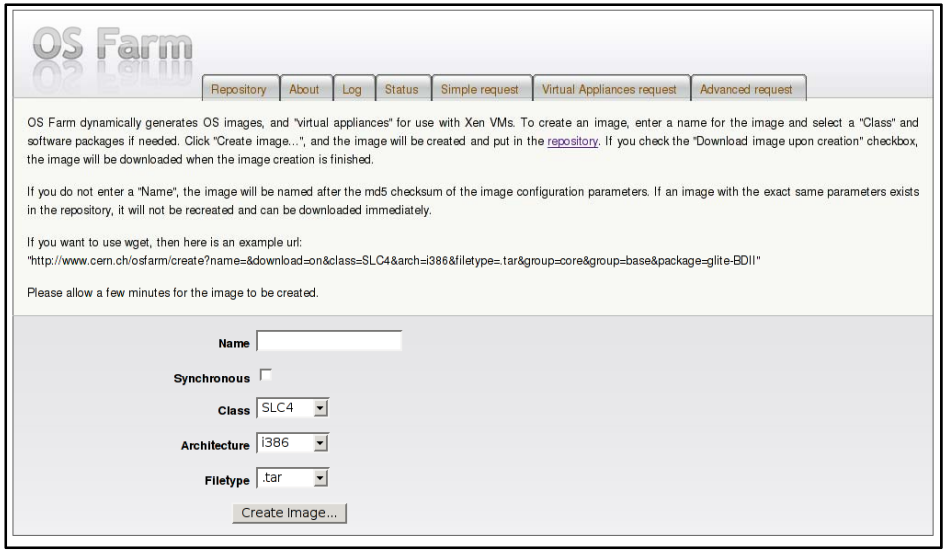
Xen (comparative benchmarks)

- Benchmarks run on para-virtualized and hardware-assisted virtualization platforms
  - point to strengths and weaknesses in hypervisors



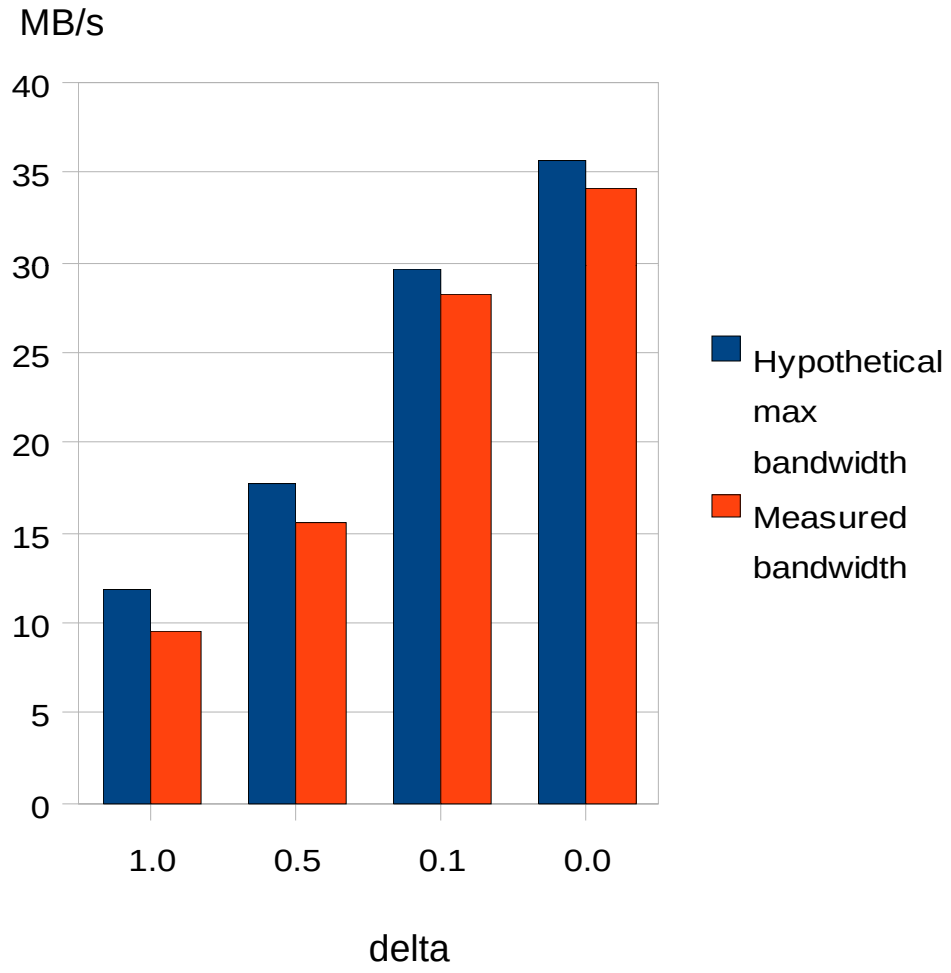
OS Farm (for Virtual Images)

- VM images generated using a layered cache
  - Core layer is instantaneous, using copy-on-write
  - Supports Debian and Red Hat based distributions
- Contextualization - customizes images according to deployment context
- Web service interface w/ example Java client
- XML image descriptions



Content Based Image Transfer (CBT)

- Most VM images are relatively similar
  - Transfer only the delta between images
- Efficiency close to hypothetical max (infinite CPU power)
- Integration with OS Farm

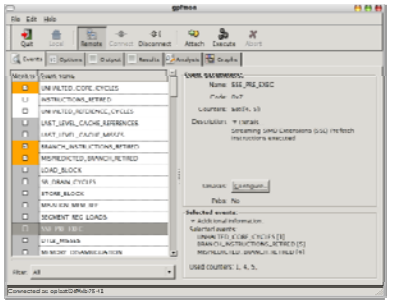


Multi-threading activities

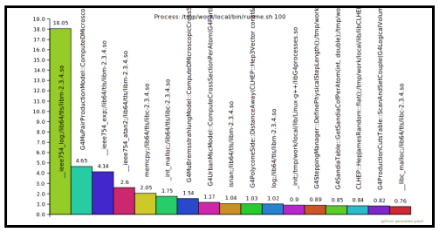
- Aim: Evangelize/teach parallel programming
- Two workshops arranged w/Intel teachers in 2007
  - 2 days, 5 lecturers, 45 participants, 20 people oversubscribed
  - Survey: 100% said expectations met
  - Next workshop: 29/30 May 2008
- Licenses for the Intel Threading Tools (and other SW products) made available
- Collaboration with PH/SFT research project
  - Geant4 parallelization prototype
  - Parallel minimization version (ROOT)

Performance Monitoring

- A joint project with S.Eranian/(ex-HP Labs)
- Aim: Ensure that his performance monitoring interface (*perfmon2* – originally developed for Itanium) gets integrated into the Linux kernel for use on ALL hardware platforms



- Our contributions:
  - Intense testing on Core 2 and Itanium
  - Increased sophistication in *pfmon* (user tool) for comprehensive symbol resolution
  - Graphical user interface: *gpfmon*



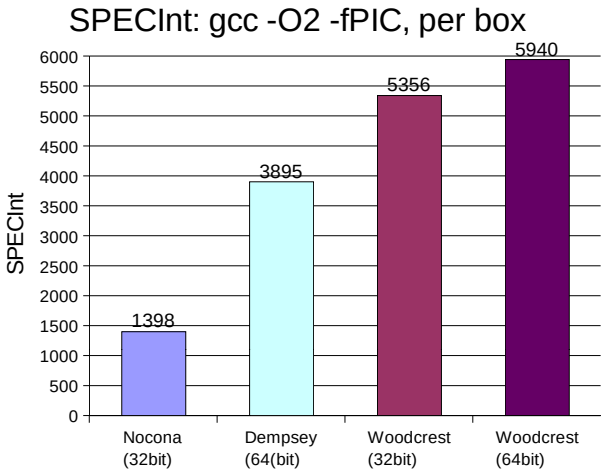
- *Also: Courses on architecture and performance*
  - *First one held on March 2008*

Compiler project

- Aim: Improved performance of jobs by influencing the back-end code generator
  - Based on our millions of lines of C++ source code
  - Also: Test suites for performance and regression testing
- 2008:
  - Target further improvements in execution time
  - Special emphasis on additional options on top of O2
  - Expand to more complex benchmarks
    - Multithreading/TBB + SSE
  - Compiler expert from Intel visiting (Sept./Oct.)
  - Compare Intel 11.0 beta with gcc 4.3.0
- Project is active since the start of openlab I
  - With particular strength in in-order execution

Benchmarking

- Aim: Identify most relevant (and convenient) benchmark for acquisitions
  - Currently: Parallel SPEC2000Int (based on gcc –O2 –fpic –threads)



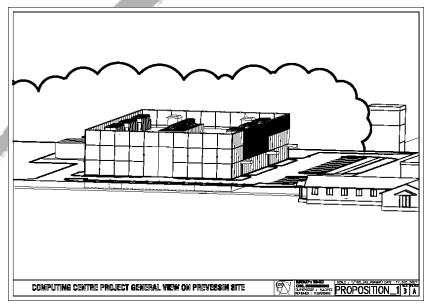
- Status:
  - Works well, but more modern benchmark suite needed
- Candidates:
  - All of SPECInt2006
  - C++ part of SPEC2006
  - CERN-specific codes

TOP500 runs

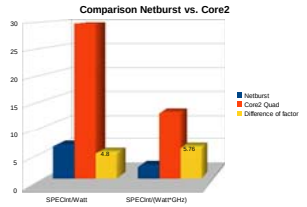
- Aim: Profit from the large acquisitions done for LHC to report the best possible number for TOP500
  - Also: Act as “burn-in” test for new systems
- Last Spring: 8.329 Tflops with 340 dual-core dual-socket servers
  - #115 in June 2007, #233 five months later (!)
- New submission for June:
  - 19.69 Tflops w/470 quad-core DS servers
- Working closely with Sergey Shalnov (Intel)
  - Using his “hybrid” version of High Performance Linpack

Thermal control

- Optimization of power/thermal efficiency in current Computer Centre
  - Enclosing cold aisles for better separation of cold/hot air
  - Add “thermal penalties” in all acquisitions



- Project for new facility
  - Understand all relevant issues, aim at 2.5 + 2.5 MW
  - Close contacts w/Michael Patterson/Intel



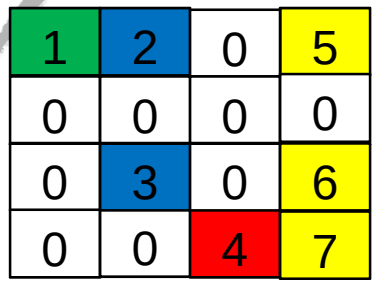
- Paper on power efficiency completed
- Project to understand thermal characteristics of each server component
  - Processors (frequencies and SKUs) ; Memory (type and size); Disks; I/O cards; Power supplies

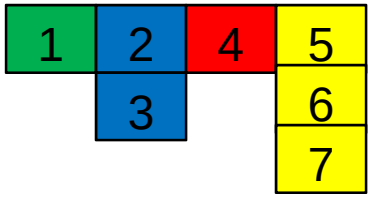
New processor activity

- Concerns both multi-core and many-core!

- Aim: Enable usage of all cores and reduce memory foot-print
  - Multi-core:
    - Get ready for Nehalem with Hyperthreading Technology
      - Up to 8 cores x 2 threads x N sockets
      - QPI and Integrated Memory Controller
      - Cost-effective MP servers

New language activity





- Started with visit and seminar by A.Ghuloum
  - Overview of Ct (Oct 2007)
- Now we are in the process of reviewing the specifications (v. 1.4)
  - Promising data parallel extension to C++
    - Need to understand how well Ct-kernels can be added to existing C++ frameworks
    - Also, which platforms are being targeted
- Waiting for first release
- Also here we are collaborating w/Intel, Brühl

Montecito upgrade

- Upgrade of 100 CPUs
  - Intel Itanium2 “Madison” to “Montecito” Dual-Core (1.6GHz)
  - Included upgrade of 50 HP mainboards!
- The Itanium cluster is quite extensively used for multiple activities
  - Two groups run parallel jobs using MPI based on Voltaire’s Infiniband switches
  - 20 servers: Computational Fluid Dynamics
  - 20-30 servers: Accelerator studies
- In addition:
  - Some systems in use by Procurve team
  - Some systems in use by IT’s security team
    - Correlating data from the CERN firewall
  - Several systems used by ourselves
    - Benchmarking, compiler testing, etc.

HP/Intel openlab Blade System

- All our testing and development require substantial x86 h/w resources
  - Agreed plan:
  - Install an expandable HP Blade System w/128 Xeon Harpertown processors
  - Great test bed for:
    - Benchmarking, Performance monitoring, Compiler testing, Virtualization tests, Grid testing, New processor simulation, New language testing, etc.
    - Also for hands-on during workshops and teaching.

Conclusion and outlook

- Together, we have produced more tangible results than ever before!
- Our key to success:
  - Highly qualified manpower
  - Constant demand for new solutions from within IT, from physics groups, EGEE and LCG.
    - Foster cross-project fertilization and collaboration (inside openlab)
    - Foster broad collaboration with all relevant external communities
- Thanks to our partners and contributors we continue to make great progress!