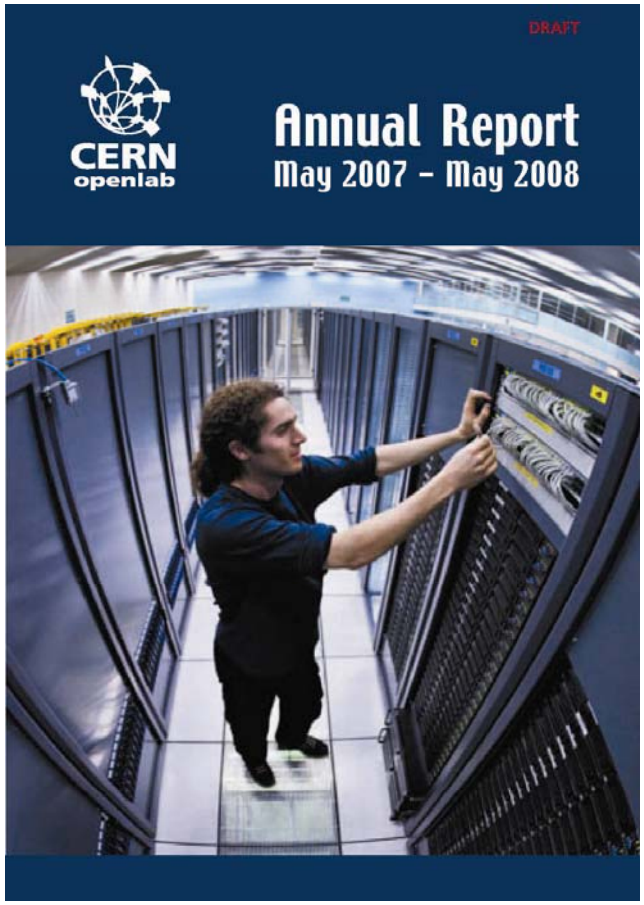
A review of the current technical activities in the CERN openlab

Sverre Jarp, 7 May 2008
CERN openlab CTO
sverre.jarp@cern.ch

Agenda

- General
- Grid Interoperability
- Database Competence
- Network and Security
- Platform Competence
- Summer student programme
- Conclusions

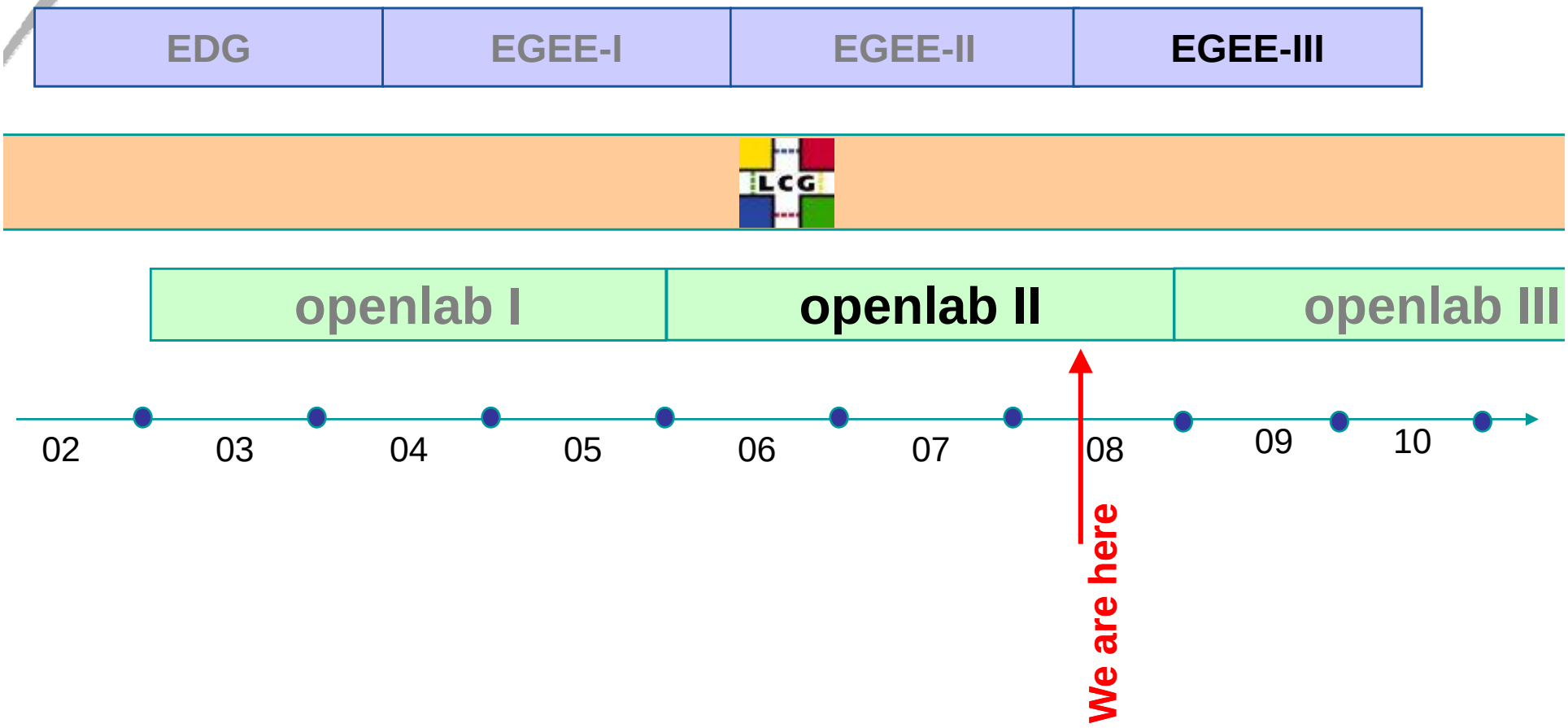


Please note that there are too many activities in openlab to mention them all. (See our Annual Report for more exact information)

General


CERN
openlab
www.cern.ch/openlab
PARTNERS
hp
invent
intel
ORACLE
CONTRIBUTORS
EDS

openlab time line



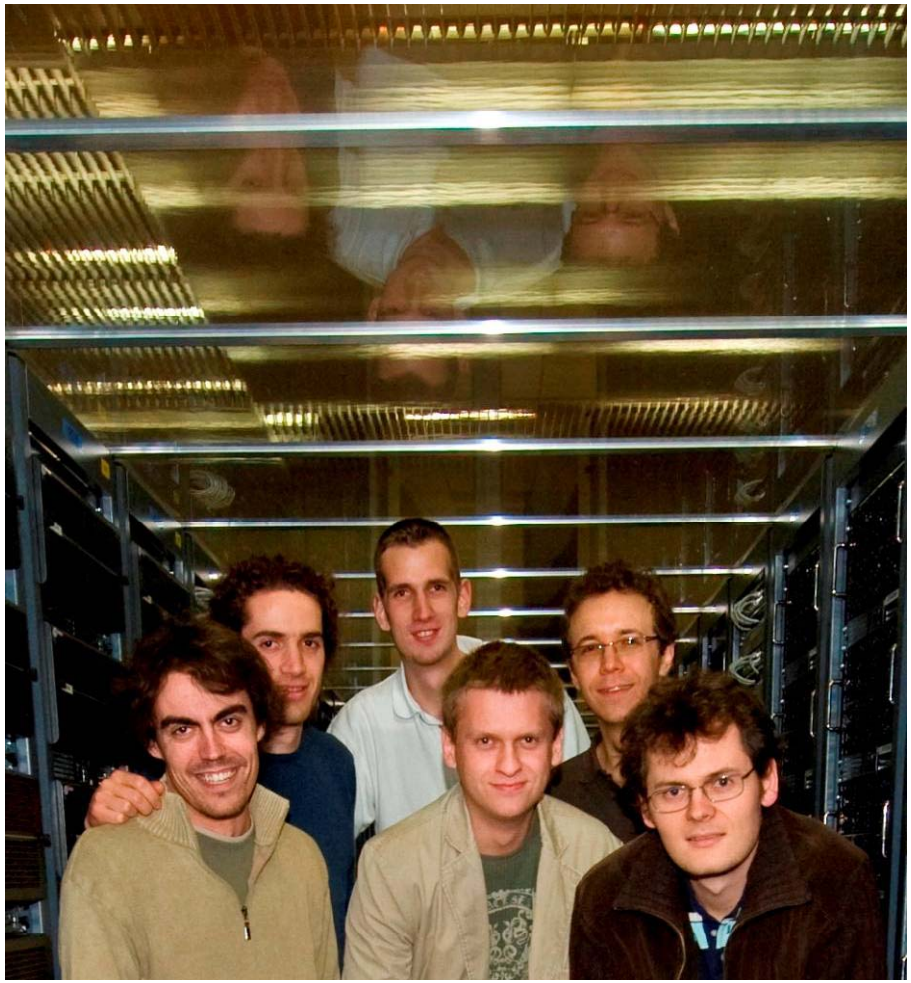
Contributions in openlab II

- **From Partners**
  - In-cash resources (for hiring)
  - Company staff
  - In-kind resources
- **From CERN**
  - Technical environment
  - Technical manpower
  - Supervisory resources
  - Management resources
  - Communications resources

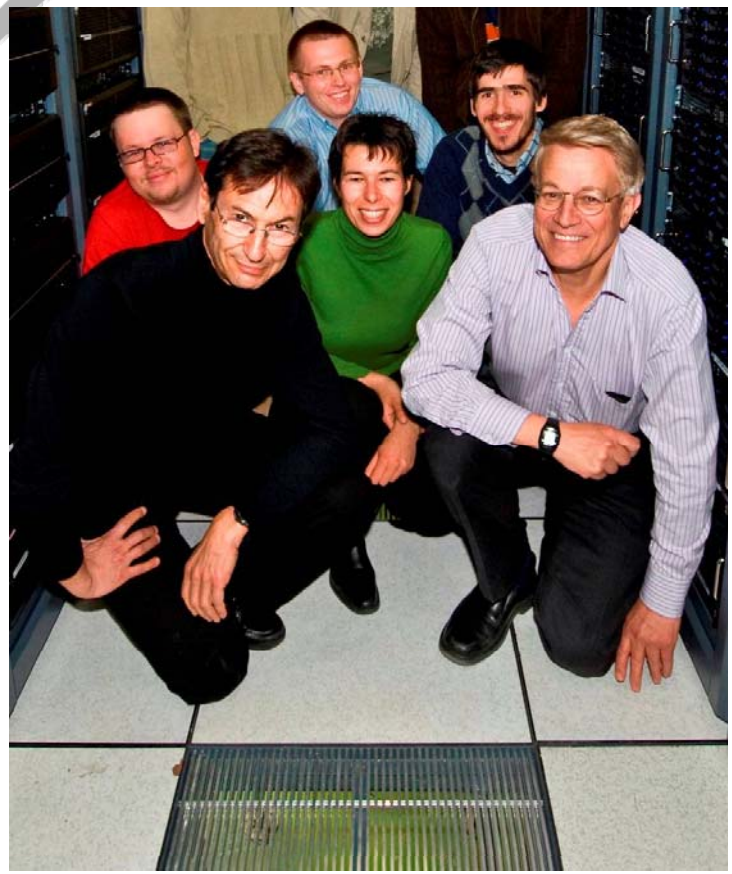
openlab Technical Team

CERN openlab Staff and Fellows
12

Gyorgy Balazs	Student (CERN)
Havard Bjerke	Fellow (Intel)
Daniel Filipe	Fellow (EDS)
Xavier Gréhant	Fellow (HP)
Milosz Hulboj	Fellow (ProCurve)
Ryszard Jurga	Staff (ProCurve)
Andreas Hirstius	Staff (Intel)
Andrzej Nowak	Fellow (EU/CERN)
Eva Dafonte Perez	Staff (Oracle)
José M. D. Perez	Fellow (HP)
Anton Topurov	Fellow (Oracle)
Dawid Wojcik	Fellow (Oracle)



CERN manpower contribution

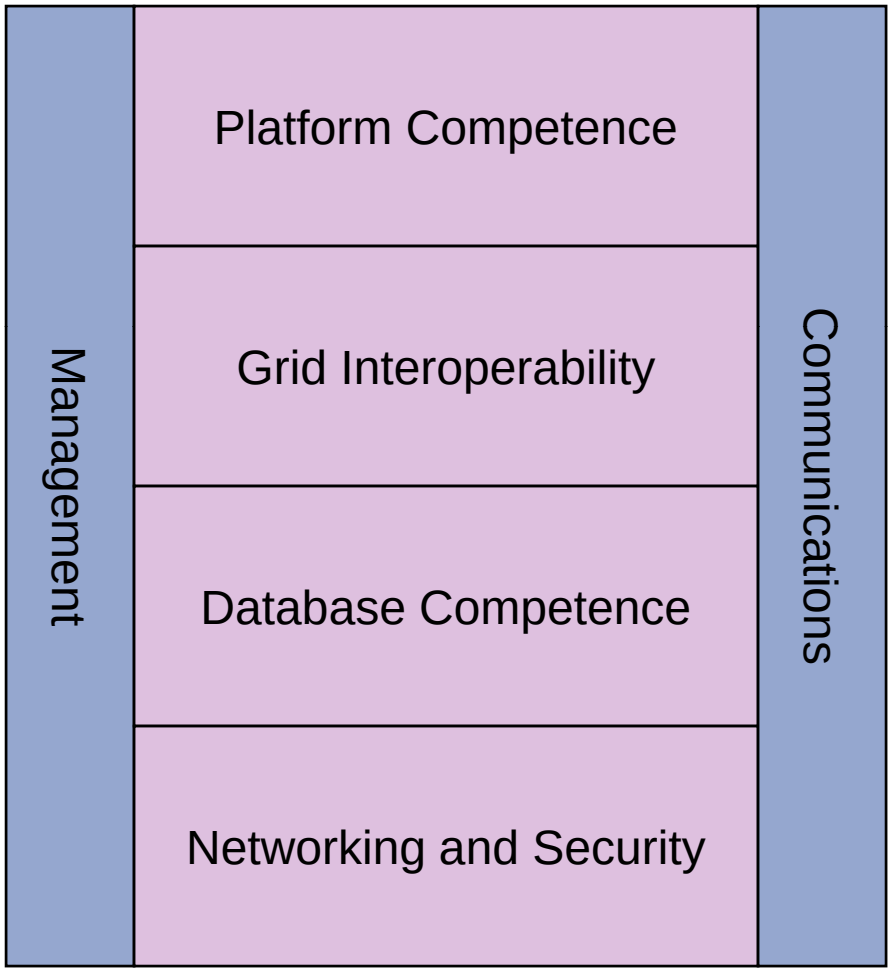


IT staff involved (part/full time)

16

Ian Bird	LCG group leader
Dirk Düllmann	DM group
François Flückiger	openlab manager
David Foster	CS group leader
Maria Girone	DM group
Eric Grancher	DES group
François Grey	Comm. Team
Denise Heagerty	Security section
Sverre Jarp	openlab CTO
J-M Jouanigot	CS group
Chris Lampert	DES group
Mats Möller	DES group leader
Alberto Pace	DM Group leader
Séverine Pizzera	Admin. Assistant
Markus Schulz	GD Group leader
Jamie Shiers	GS Group leader
and others …..	

openlab II structure

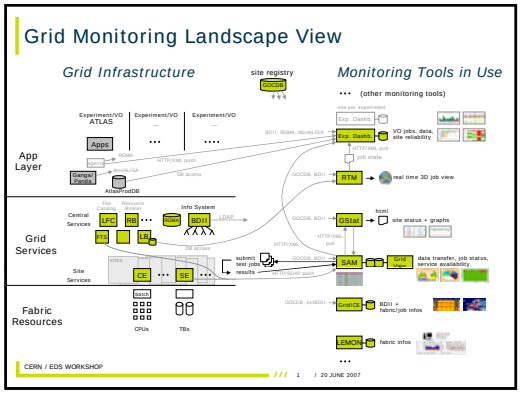


GIC

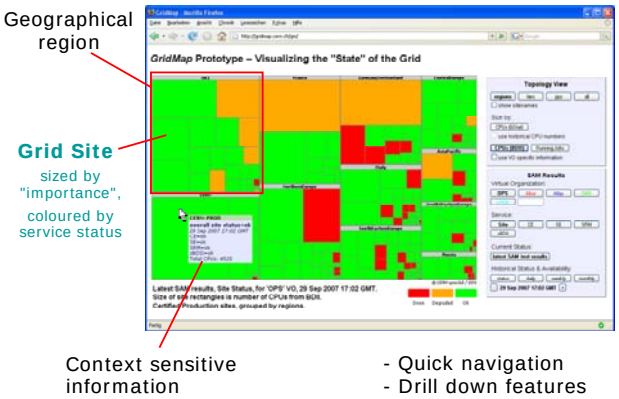

CERN
openlab
www.cern.ch/openlab
PARTNERS
hp
invent
intel
ORACLE
CONTRIBUTORS
EDS

Grid Monitoring

- Started with an analysis of the (then) current situation
  - Grid Monitoring Landscape
    Q2 2007, CERN openlab / EDS Workshop



- New Monitoring Management Views
  - Developed *GridMap* Prototype
  - Presented at EGEE'07
  - Documentation, Releases
  - Variants: ServiceMap, ...



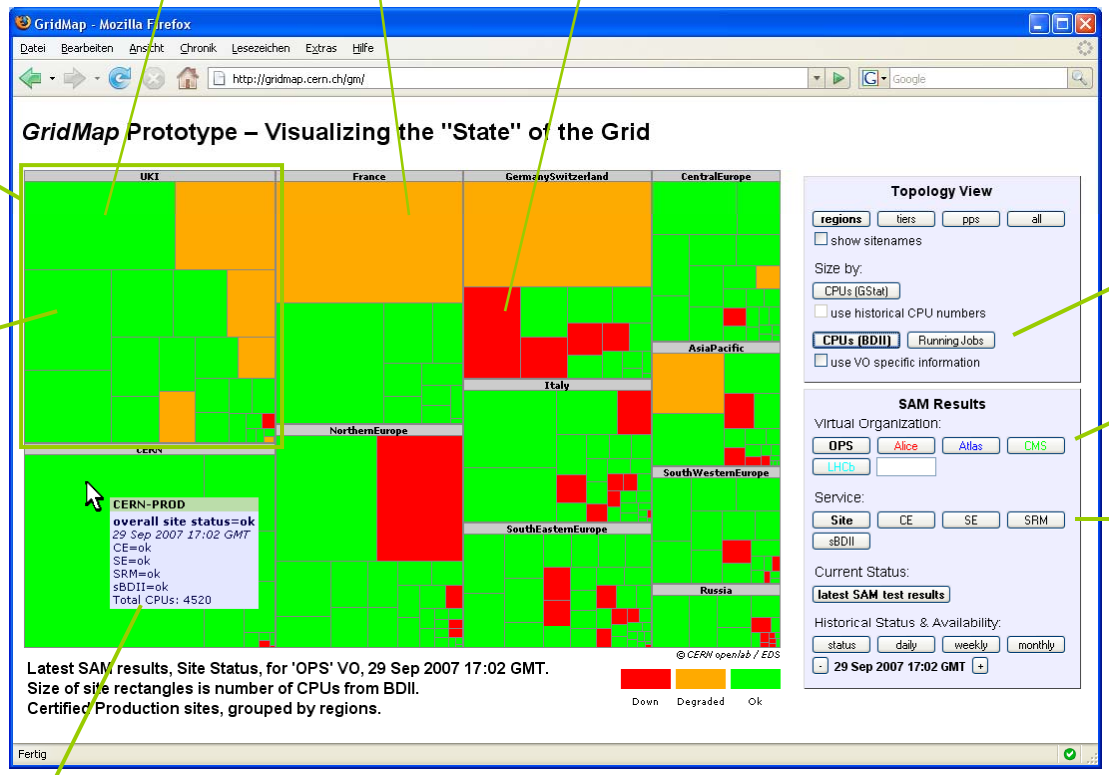
GRIDMAP in detail

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

Service status

ok degraded down

Geographical region



Grid Site

sized by "importance", coloured by service status

Size by ...

(e.g. #CPUs of the site, #running jobs, ...)

User specific views

Service type to be shown

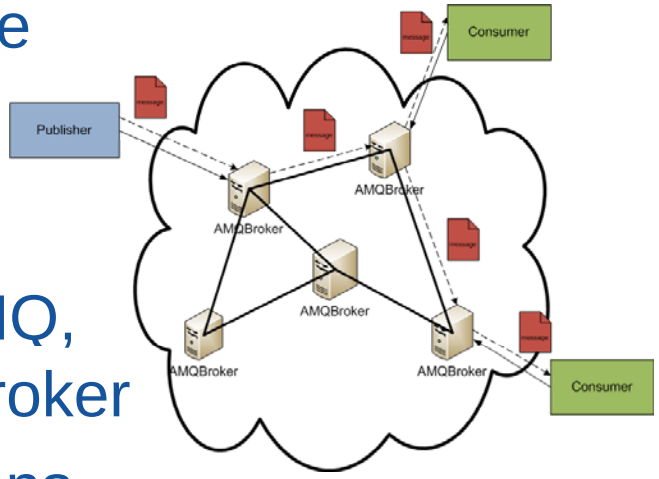
(CE=Compute Element, SE=Storage Element, ...)

Context sensitive detailed information

"Click" links to underlying tools
Drill down features

MSG

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

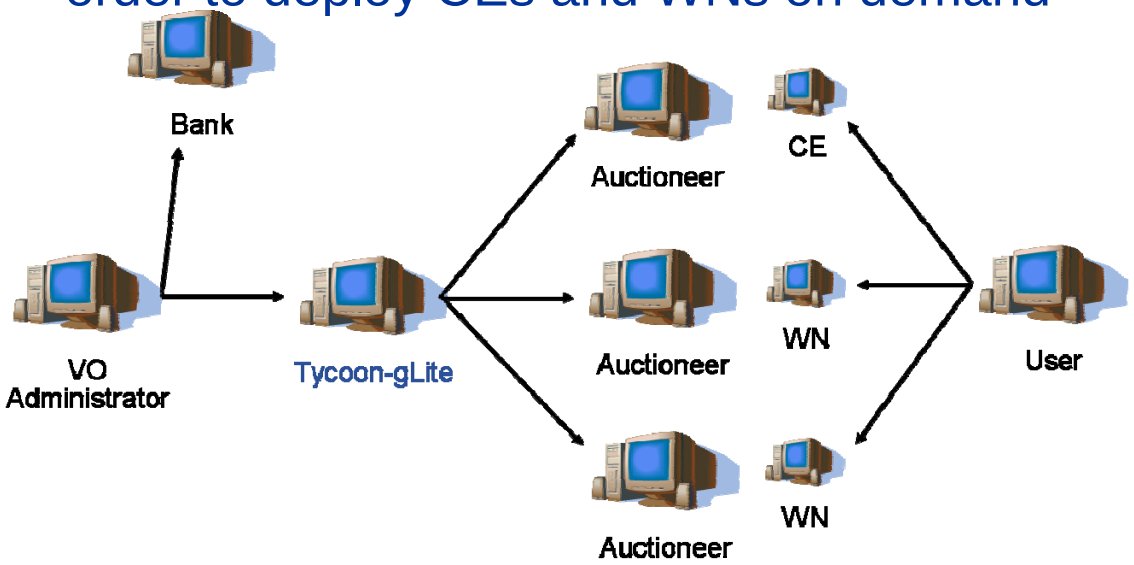


- Work started in Sept. 07
  - Extensive testing of ActiveMQ, an open-source message broker
  - Prototype of different solutions (mainly Python)
  - Currently OSG and Gridview production data is being published and consumed

HP Tycoon overview

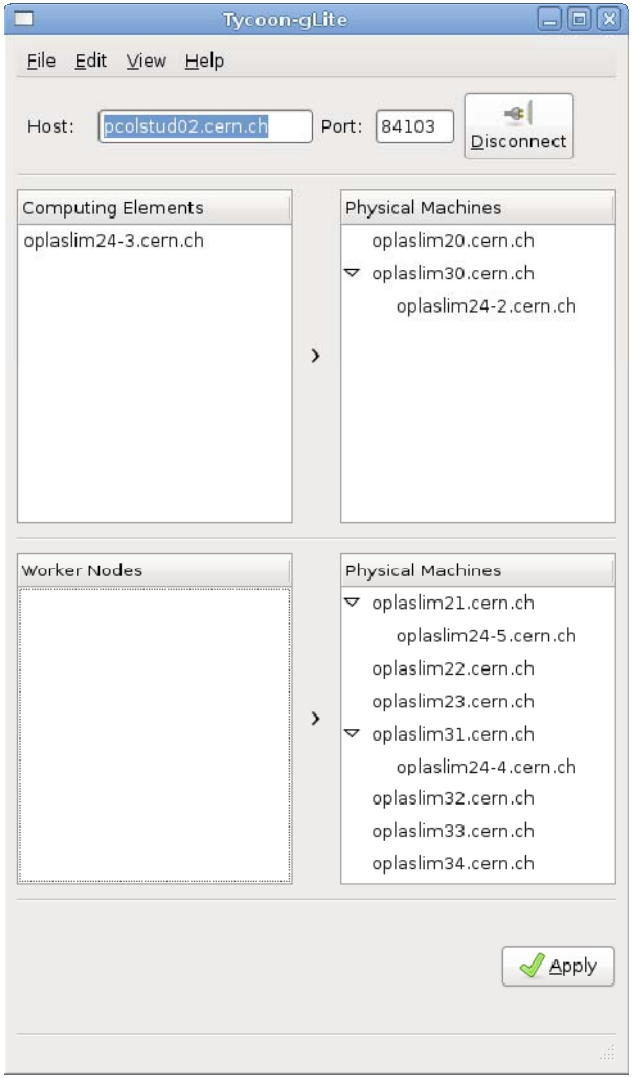
- A dynamic Grid infrastructure using a market-driven approach

- Joint development with HP Labs where:
  - HP Labs did the porting of Tycoon to SLC4 and recent versions of Xen
  - CERN openlab developed the integration with EGEE in order to deploy CEs and WNs on demand



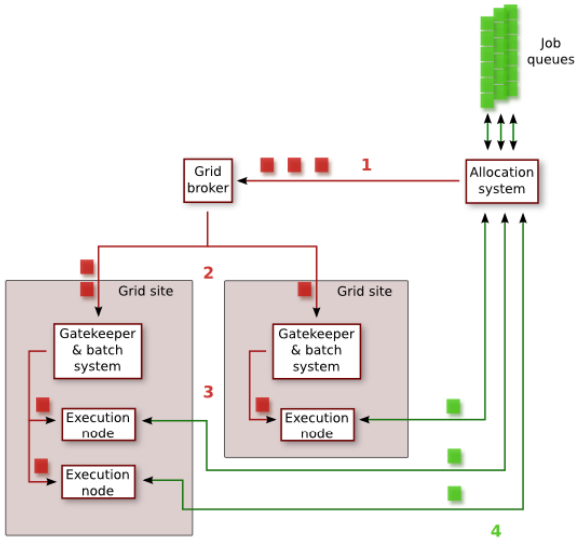
Tycoon-gLite integration

- Extensive scalability tests performed
- Some issues concerning “large-scale” security and trust
  - e.g. who runs the bank!
- Recently, the implementation was enhanced in order to:
  - Deploy different kinds of nodes more easily (i.e. Storage Elements)



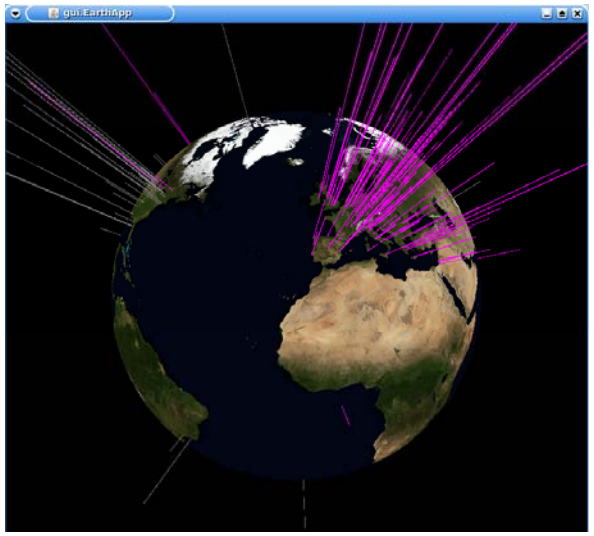
Grid Scheduling Survey

- X.Gréhant's PhD:
  - Synthesis on Grid Scheduling
  - In-depth analysis of VO management, resource access
    - EGEE, OSG, NorduGrid, Naregi, etc.
  - Direct scheduling in a VO
    - glideCAF, Cronus, GlideInWMS
    - AliEn2, DIRAC, Panda
    - DIANE
  - In collaboration with grid developers at CERN
  - Paper submitted to the Journal of Supercomputing



Grid Resource Simulation

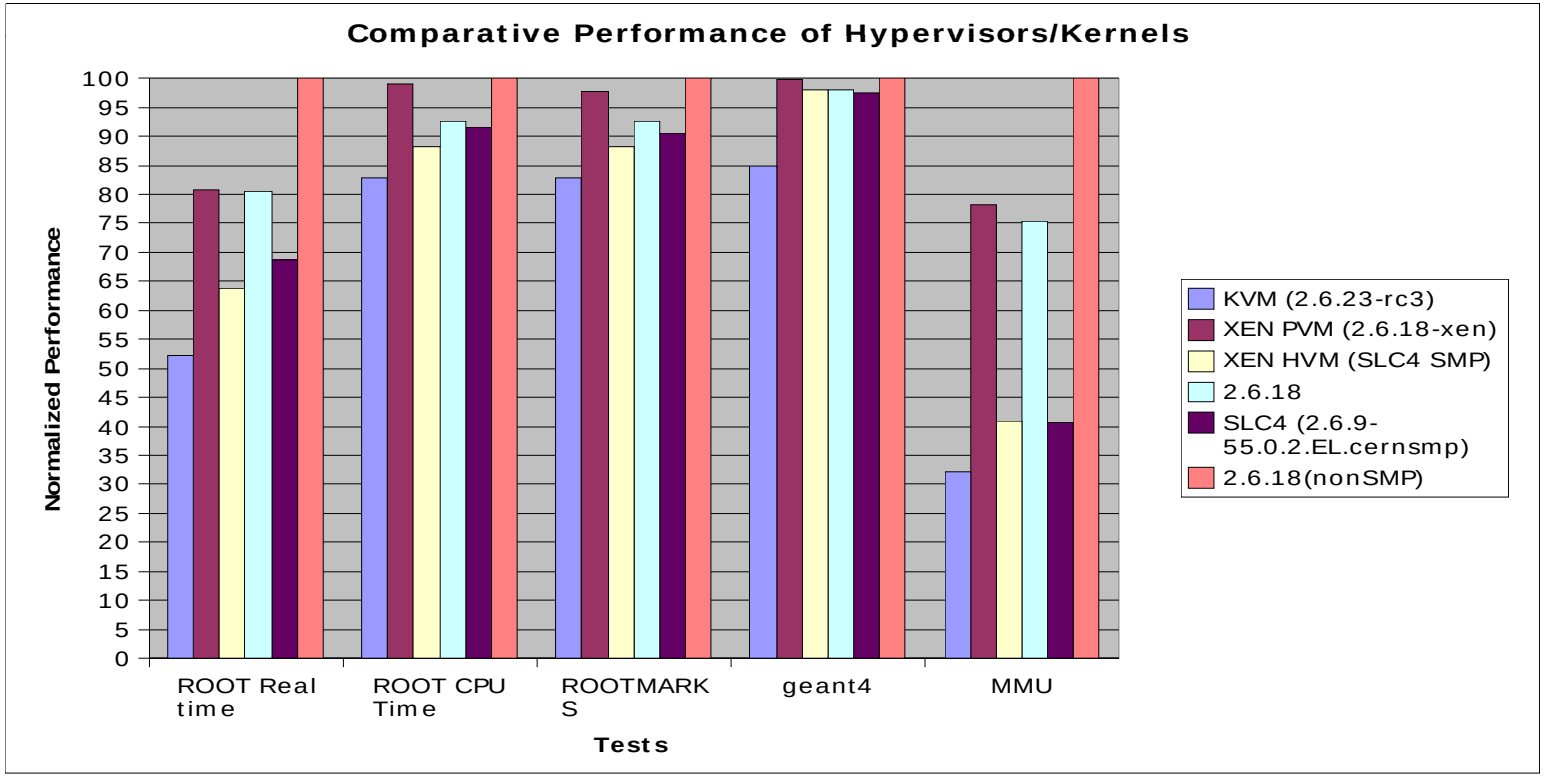
- Simulating such an environment
  - Level-lab:
    - Simulates the environment a VO gets on the grid
    - Evaluates performance of allocation algorithms
    - Development done jointly with summer student (2007)
  - Status
    - 3000 lines of code, 5000 of unit tests
    - Simple model working
    - Successive refinements in progress (job and resource profiles accuracy)



WLCG sites on Level-Lab visualization module

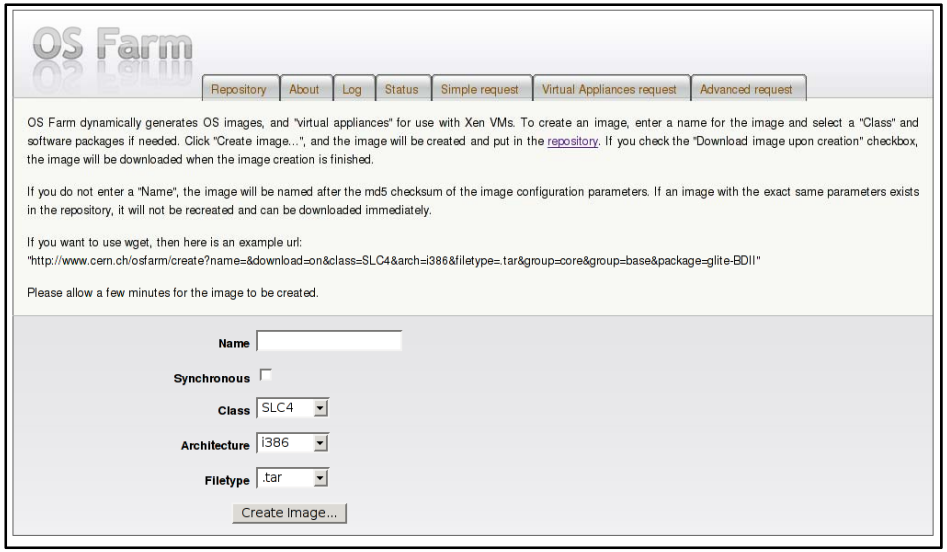
Xen (Virtualization benchmarks)

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



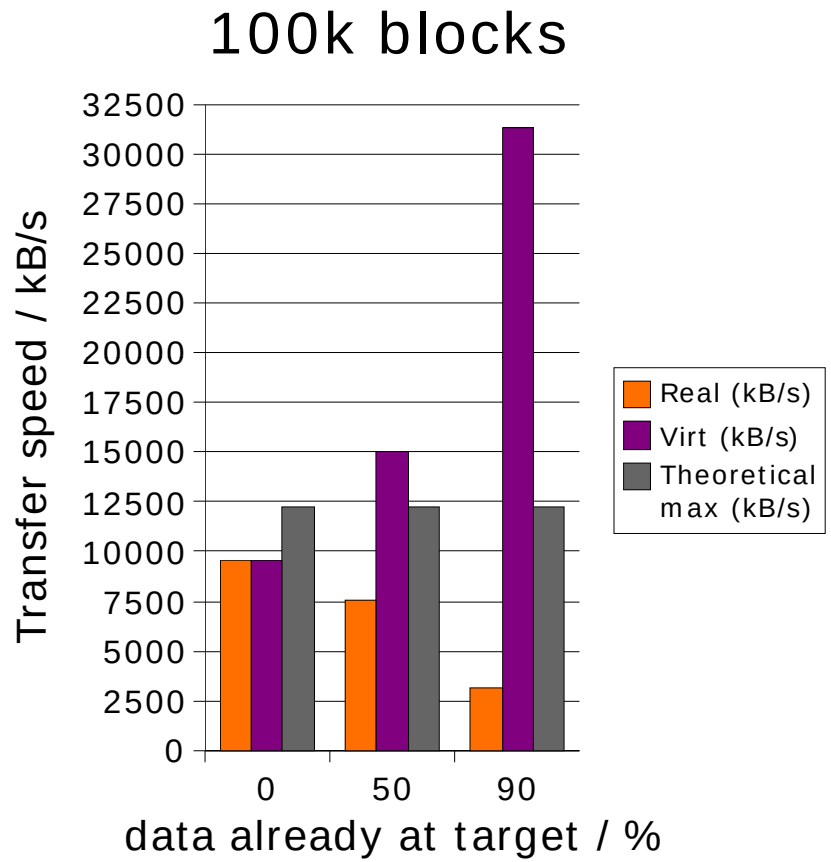
OS Farm (for Virtual Images)

- 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



H. K. F. Bjerke, D. Shiyachki, Andreas Unterkircher, Irfan Habib, Tools and Techniques for Managing Virtual, Machine Images, submitted to 3rd Workshop on Virtualization in High-Performance Cluster and Grid Computing (VHPC '08)

DBCC

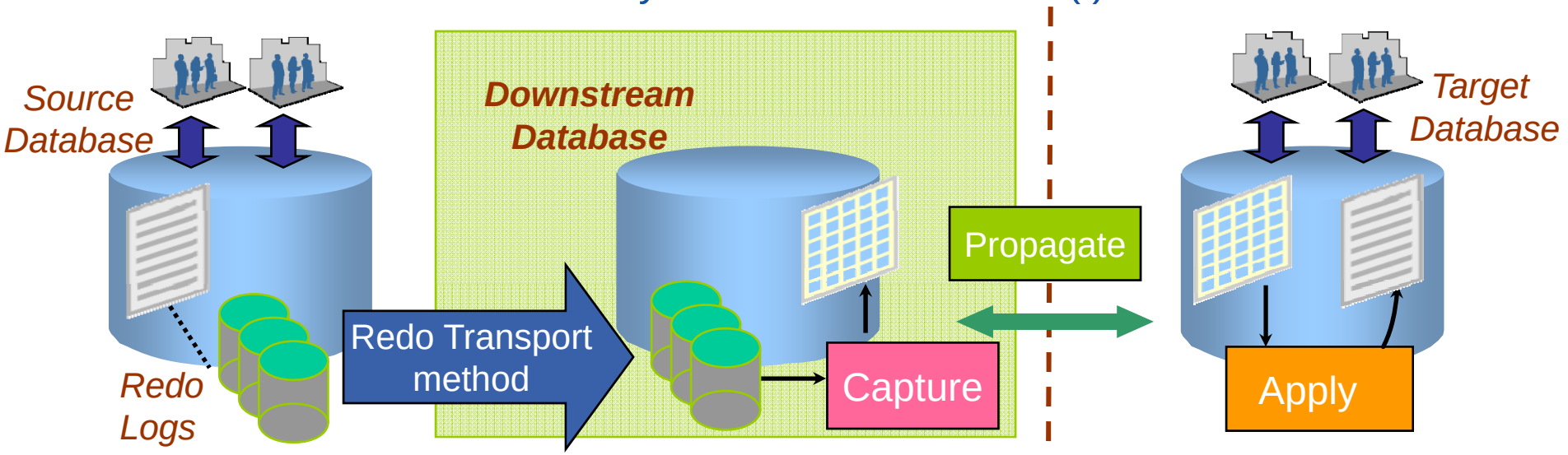

CERN
openlab
www.cern.ch/openlab
PARTNERS
hp
invent
intel
ORACLE
CONTRIBUTORS
EDS

Events and outreach highlights

- **Oracle OpenWorld, San Francisco, October 2007**
- Dirk Düllmann and Paul Otellini (Intel CEO) during keynote speech
- CERN presentations

Database downstream capture and network optimizations

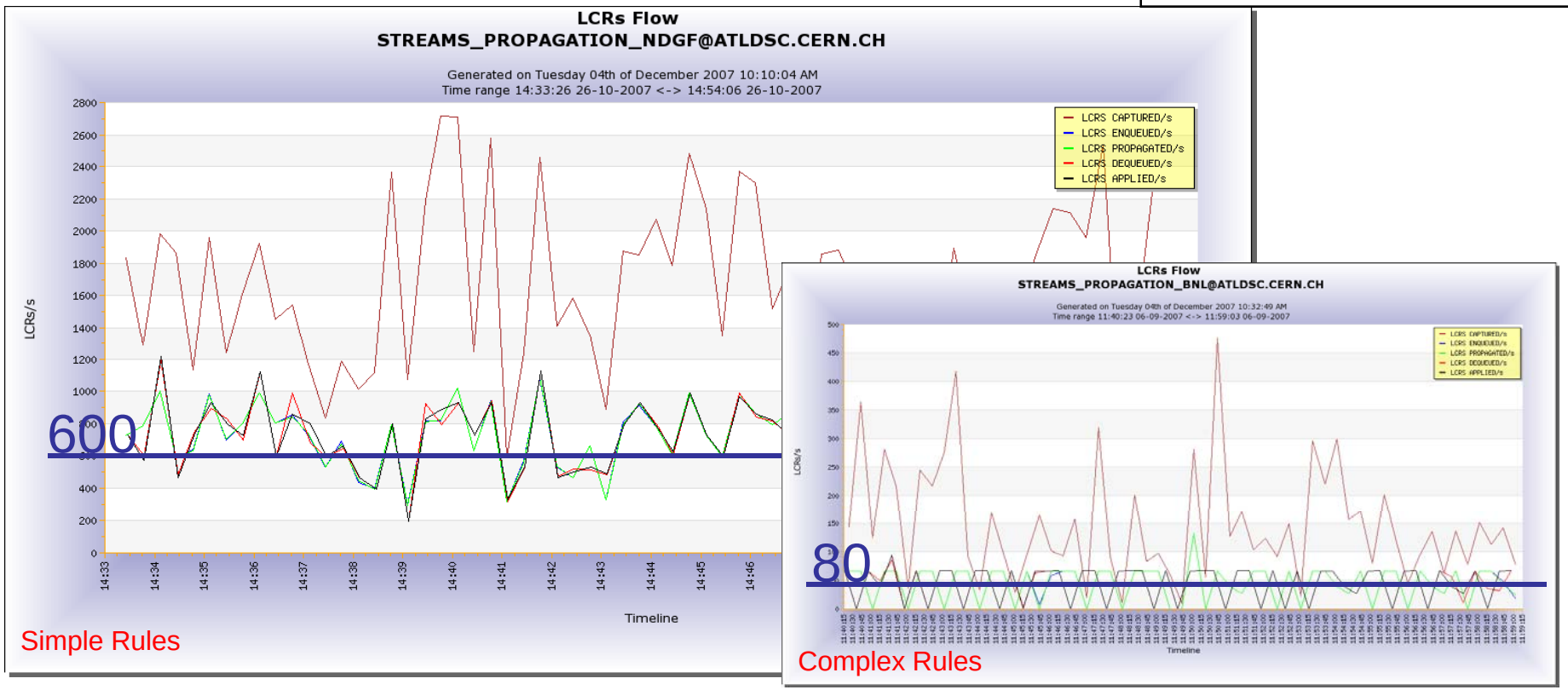
- Downstream capture to de-couple Tier 0 production databases from destination or network problems
  - source database availability is highest priority
- Optimizing redo log retention on downstream database to allow for sufficient re-synchronisation window
  - we use 5 days retention to avoid tape access
- 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

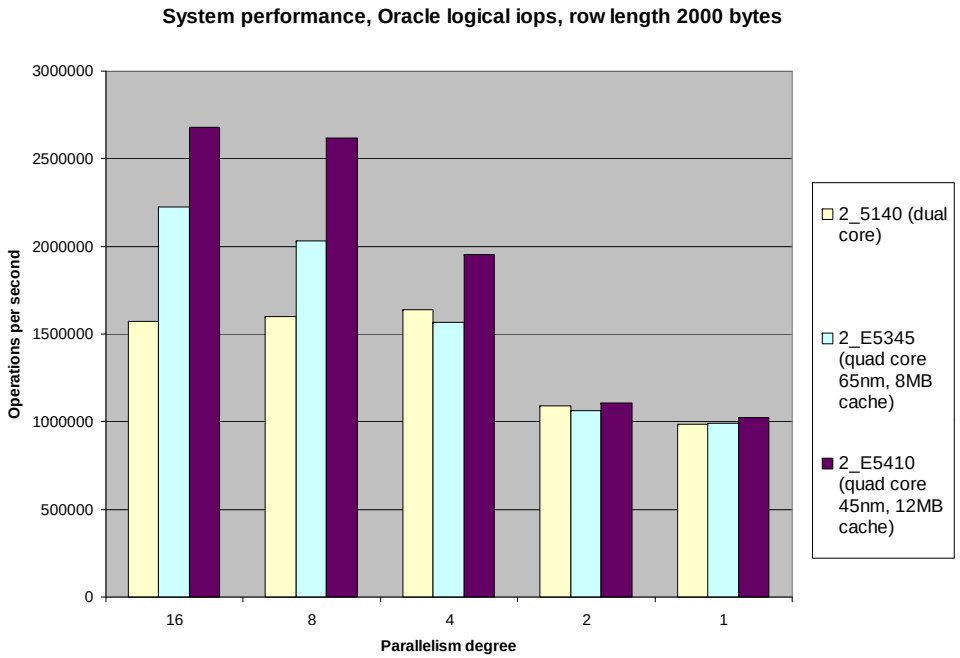
- Requested features:
  - Streams topology
  - Status of streams connections
  - Error notifications
  - Streams performance (latency, throughput, etc.)
  - Other resources related to the streams performance (streams pool memory, redo generation)
- Architecture:
  - “strmmon” daemon written in Python
  - 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 performance benchmarks

~15% more performance →



- Test and validate performance of new platforms

- Oracle RDBMS performance comparison between (all dual-socket platforms):
  - E5140 (2.33Ghz, 4MB cache, “Woodcrest” - DC), current deployment platform for CERN’s Linux RACs
  - E5345 (2.33Ghz, 8MB cache, “Clovertown” - QC)
  - E5410 (2.33Ghz, 12MB cache, “Harpertown” - QC)

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

A. Dechert, IT-Service Management at CERN and How It Can Be Improved by the Usage of Oracle Enterprise Manager, Diploma Thesis, Karlsruhe University of Applied Sciences 2008

Networking


CERN
openlab
www.cern.ch/openlab
PARTNERS
hp
invent
intel
ORACLE
CONTRIBUTORS
EDS

CINBAD project

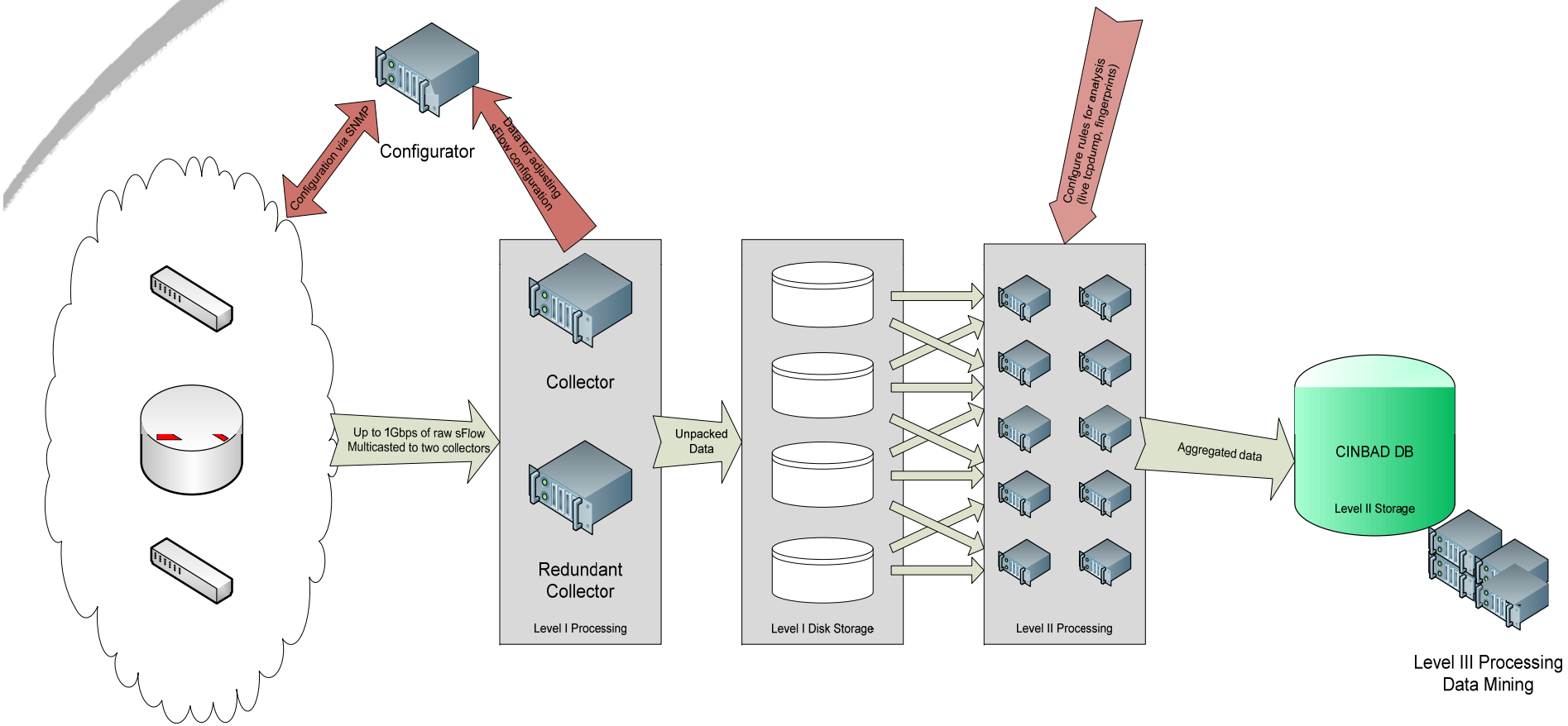
- Codename: “CINBAD”
  - **C**ERN **I**nvestigation of **N**etwork **B**ehaviour and **A**nomaly **D**etection
- Project Goal
  - “To ***understand the behaviour of large computer networks*** *(10’000+ nodes) in High Performance Computing or large Campus installations to be able to:*
    - *Detect traffic anomalies in the system*
    - *Be able to perform trend analysis*
    - *Automatically take counter measures*
    - *Provide post-mortem analysis facilities* “

CINBAD deliverables

- The project is tentatively divided into three phases, each with a particular set of investigation activities and deliverables:
  - **Data collection and network management**
  - **Data Analysis and algorithm development**
  - **Performance and scalability analysis**

M. M. Hulboj and R. E. Jurga
Packet Sampling for Network Monitoring, December 2007

CINBAD collector design



Highly Scalable Architecture

Rich database for investigative data mining

Intel 10 Gb networking

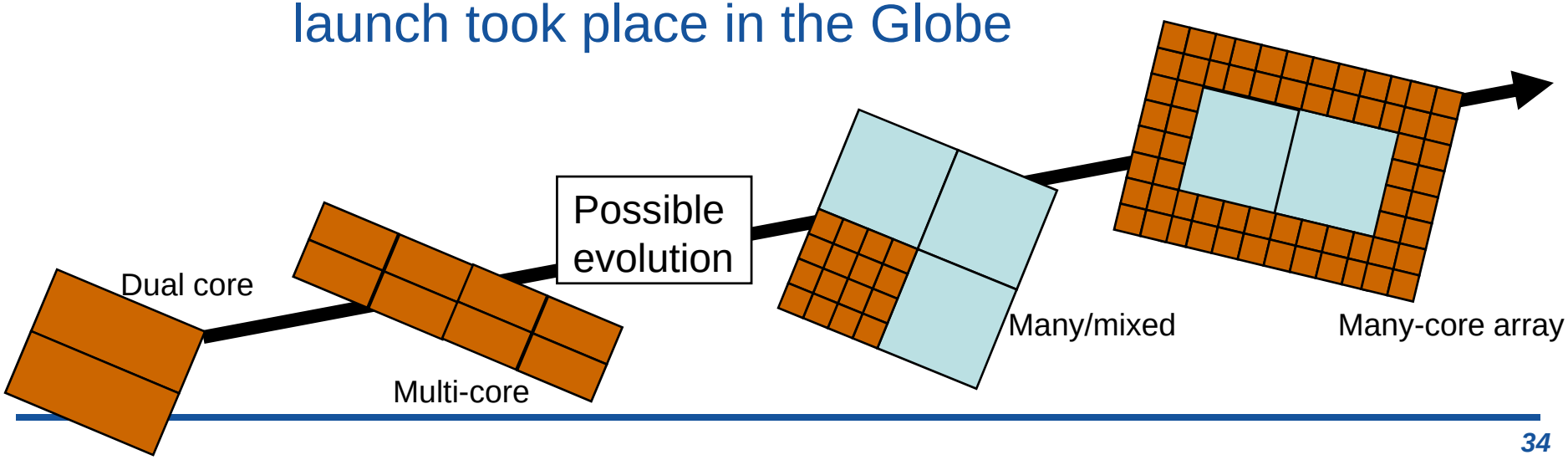
- With the first generation cards, we successfully prototyped high-throughput disk servers, but …
  - Very high cost
  - Reasonable throughput required jumbo-frames
    - MTU 9KB, rather than 1.5KB (Ethernet standard)
- Production disk servers (w/1Gb NICs) have now reached their throughput/capacity limit
- 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

PCC


CERN
openlab
www.cern.ch/openlab
PARTNERS
hp
invent
intel
ORACLE
CONTRIBUTORS
EDS

From Multi to Many

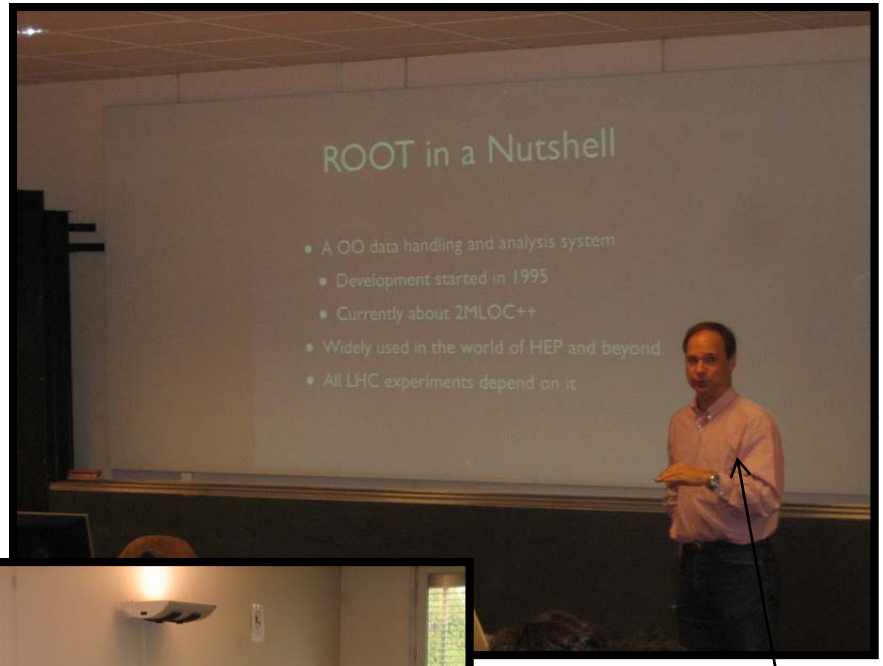
- The HEP “high throughput” computing model is ideally suited:
  - Independent processes can run on each core, provided that:
    - Main memory is added
    - Bandwidth to main memory remains reasonable
  - Testing, so far, has been very convincing
    - Woodcrest, Clovertown, Harpertown; Montecito
- In November 2006, Intel’s European Quad core launch took place in the Globe



Multi-threading activities

- Aim: Evangelize/teach parallel programming
- Two workshops arranged w/Intel in 2007
  - Topics: OpenMP, MPI, TBB, Intel tools
  - 2 days, 5 lecturers, 45 participants, 20 people oversubscribed
  - 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)

MT Workshop pictures

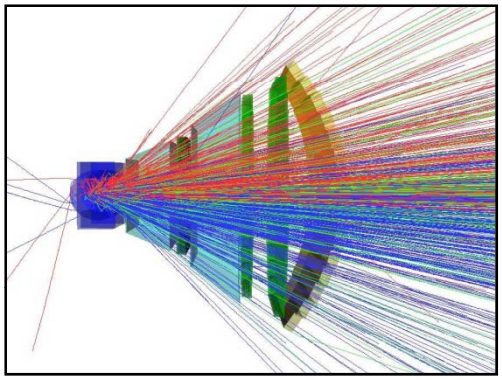


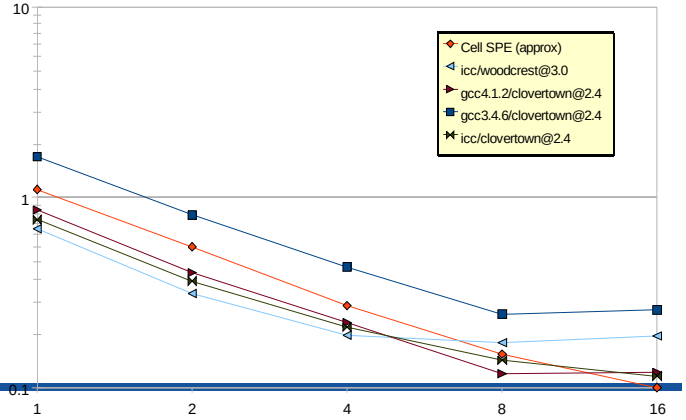
Herbert Cornelius and Hans-Joachim Plum from Intel

Fons Rademakers from CERN

Collaboration on parallelism

- CBM experiment's High Level Trigger Code
  - Originally ported to Cell processor
- Tracing particles in a magnetic field
  - Embarrassingly parallel code
- Re-optimization on Intel Core systems
  - Step 1: used SSE vectors instead of scalars
    - Operator overloading allows seamless change of data types, even between primitives (e.g. float) and classes
  - Step 2: added multithreading (via TBB)
    - Enable scaling with core count

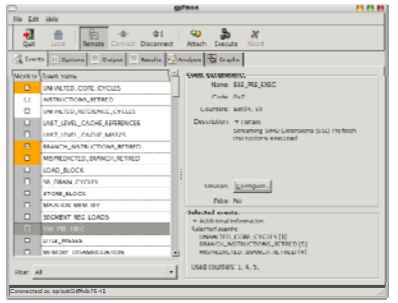




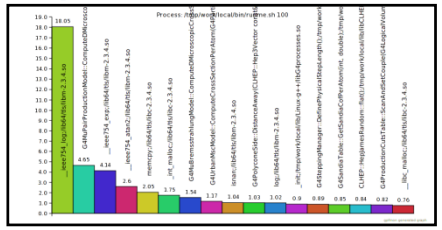
I.Kisel/GSI: "Fast SIMDized Kalman filter based track fit"
http://www-linux.gsi.de/~ikisel/reco/CBM/DOC-2007-Mar-127-1.pdf

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*

S. Jarp et al.: Perfmon2: A leap forward in Performance Monitoring, CHEP2007, Sept. 2007

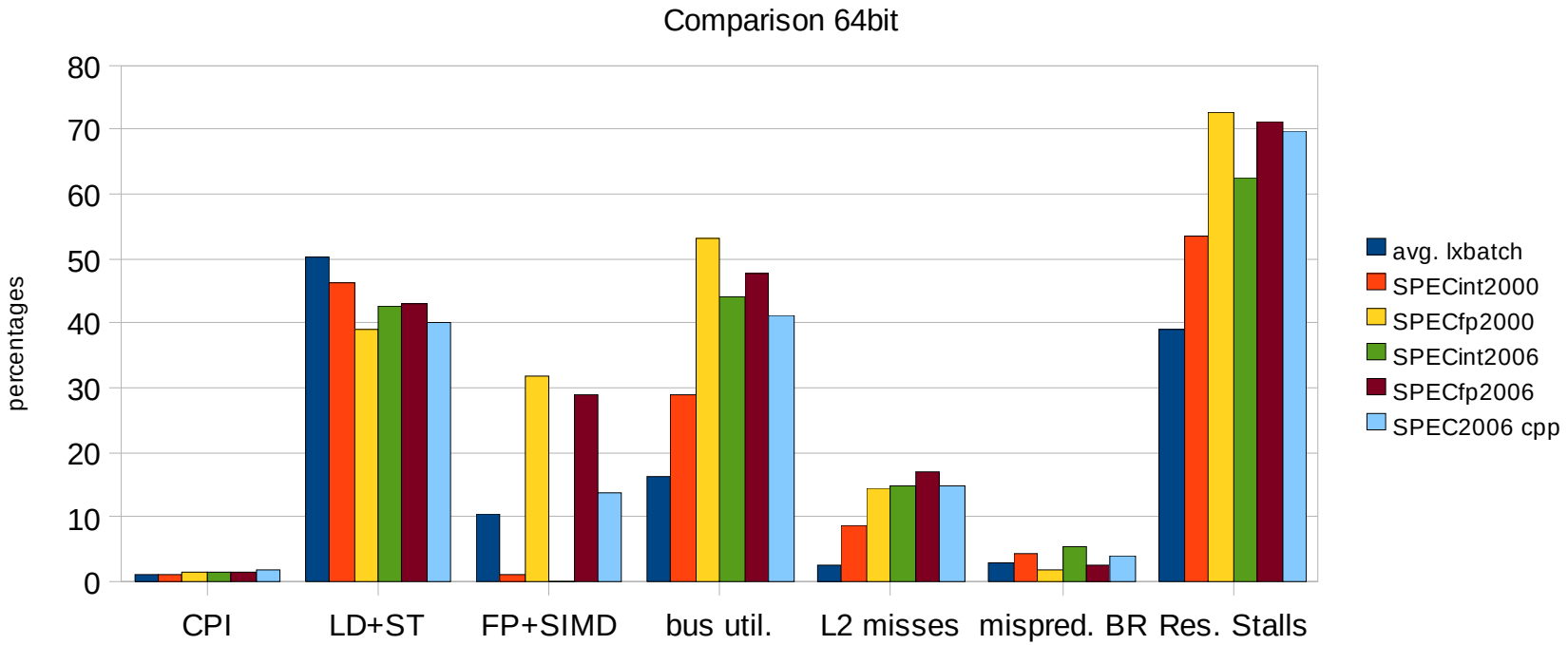
Recent pfmon example (ATLAS)

- ATLAS Athena 64-bit mode

```
# results for [7913<-[7907]] (/afs/cern.ch/sw/lcg/external/Python/2.5/slc4_amd64_gcc
/afs/cern.ch/atlas/software/builds/nightlies/dev/AtlasCore/rel_2/InstallArea/share/b
# total samples          : 13881105
# total buffer overflows : 6777
#
#                  event00
#     counts   %self    %cum          code addr symbol
     4356424  31.38%  31.38% 0x0000003061511930 __ieee754_log</lib64/tls/libm-2.3.4.
      976234   7.03%  38.42% 0x00002ac24924a9b0 G4MuPairProductionModel::ComputeDMic
double)</afs/cern.ch/atlas/offline/external/geant4/volume5/geant4.8.3.patch02.atlas0
      868803   6.26%  44.68% 0x00002ac2491f7bf0 G4VRangeToEnergyConverter::RangeLogS
const*, double const*, double, double, double, int)</afs/cern.ch/atlas/offline/exter
64/lib/Linux-g++/libG4processes.so>
      710397   5.12%  49.79% 0x000000306150e370 __ieee754_exp</lib64/tls/libm-2.3.4.
      613669   4.42%  54.21% 0x00002ac2491eba40 G4ProductionCutsTable::ScanAndSetCou
G4Region*)</afs/cern.ch/atlas/offline/external/geant4/volume5/geant4.8.3.patch02.atl
      397489   2.86%  57.08% 0x00002ac247da8650 G4PhysicsLogVector::FindBinLocation(
const</afs/cern.ch/atlas/offline/external/geant4/volume5/geant4.8.3.patch02.atlas02.
      367929   2.65%  59.73% 0x0000003061513470 __ieee754_log10</lib64/tls/libm-2.3.
```

Benchmarking

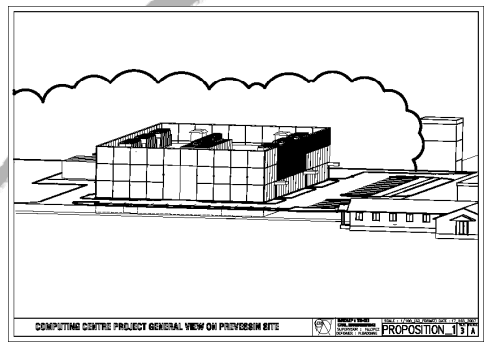
- Contributing to Benchmarking Working Group
  - Aim: Identify most relevant (and convenient) benchmark for acquisitions
    - Currently: Parallel SPEC2000Int (based on gcc –O2 –fPIC –threads)

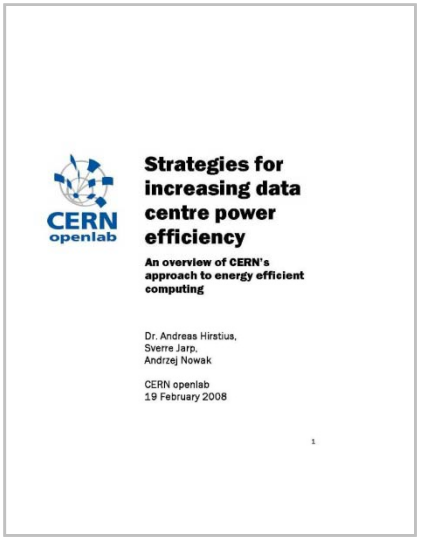


TOP500 runs

- Aim: Profit from the large acquisitions done for LHC to report the best possible number for TOP500
  - In reality: Perform a solid “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 Intel (Sergey Shalnov)
  - Using his “hybrid” version of High Performance Linpack

Thermal control




CERN openlab

Strategies for increasing data centre power efficiency

An overview of CERN's approach to energy efficient computing

Dr. Andreas Hirstius,
Sverre Jarp,
Andrzej Nowak

CERN openlab
19 February 2008

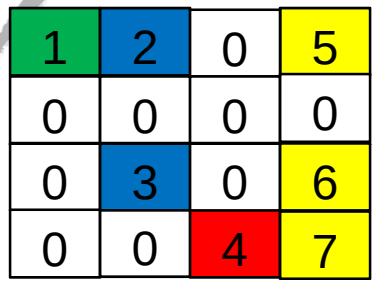
1

- Help optimize power/thermal efficiency in the CERN Computer Centre
  - Good collaboration with Michael Patterson, Intel's top expert
    - Enclosing cold aisles for better separation of cold/hot air
    - Add "thermal penalties" in all acquisitions
    - Collaborate on project for new facility
- 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

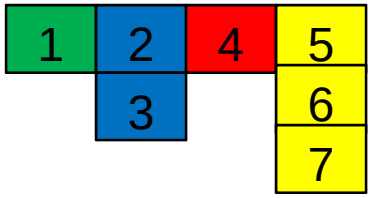
New processor activity

- 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 SMT technology
      - QPI and Integrated Memory Controller
      - DP: 4 cores x 2 threads x 2 sockets
      - Cost-effective (?) MP servers:
        - » 8 cores x 2 threads x 4 sockets
  - Will the HEP community start using HW threads?
  - Also: Study implications of AVX (Advanced Vector Extensions)
    - 256bits: Can HEP software make efficient use of four-vector operations??

New language activity



- Visit and seminar by A.Ghuloum/Intel
  - 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

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
  - 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

HP/Intel openlab Blade System

- All our testing and development require substantial x86 h/w resources
  - Next step:
  - Install an expandable HP Blade System w/128 Intel Xeon Harpertown processors
  - Great test bed for:
    - Benchmarking, Performance monitoring, Compiler testing, Virtualization tests, Grid testing, Simulator runs (AVX, etc.), New language testing, …..
    - Also for hands-on during workshops and teaching.

Summer students

Education – Summer student programme

S. Jarp is the new programme coordinator

12 students this year

Several co-funded by openlab partners

Summer Lecture series

CERN openlab student programme lecture series

July-August 2007

Lecture 1:	**Server hardware**	A. Hirstius/CERN
Lecture 2:	**Linux kernel**	J. Iven/CERN
Lecture 3:	**Oracle RDBMS**	B. Engsig/Oracle
Lecture 4:	**Computer Security**	S. Lopienski/CERN
Lecture 5:	**Compilers**	L. Pollock/ U. Delaware
Lecture 6:	**Benchmarking**	S.Jarp/CERN
Lecture 7:	**Networking**	M. Swany/ U. Delaware
Lecture 8:	**Virtualization**	H.Bjerke/CERN
Lecture 9:	**LCG**	L. Poncet/CERN
Lecture 10:	**gLite**	M. Schulz/CERN

Conclusion and outlook

- LHC Computing is complex and poses many challenges
- Rich openlab programme between partners and CERN IT groups
  - Important to ensure that both sides find value in the collaboration
    - Best when results have broad applicability
- Joint investment in (wo-)manpower is vital
- We are not perfect (by any means), but openlab is currently delivering more results than ever before
  - To the great satisfaction of everybody
- We are actively preparing openlab III