Summary of Technical Achievements

Sverre Jarp, CERN openlab CTO

April 2\textsuperscript{nd} 2009

CERN openlab Board of Sponsors Meeting 2009
Both for openlab II and III: A set of Competence Centres

- Grid Interoperability Centre
- Database CC
- Networking/Security CC
- Platform CC

- Automation and Controls CC
- Database CC
- Networking/Security CC
- Platform CC
The secret of success:
- Fellows
- Staff
- Technical students
- Summer students

Solid investment by all partners, contributors and CERN
Presentations/Publications/Reports

Presentations:
- A. Nowak/CERN, High-throughput computing optimization issues at CERN, Bioinformatics in Torun, Torun, Poland, 14 June 2008
- H. Bjørke/CERN, High Throughput Computing for CERN’s Large Hadron Collider, ISCA, Beijing, China, 22 June 2008
- D. Rodrigues/CERN, Messaging System for the Grid, EGEE’08, Istanbul, Turkey, 24 September 2008
- S. Jarp/CERN, Faire face aux nouvelles architectures de processeurs : la physique des particules est-elle prête ?, JI’08, Obernai, France, 30 September 2008
- A. Topurov/CERN, CERN Experience with Virtualization of Oracle RAC with Native Xen and Oracle VM, TrivadisOpen, Zurich, Switzerland, 22 October 2008
- S. Jarp/CERN, Forget multicore! The future is manycore: An outlook to the explosion of parallelism likely to occur in the LHC era, ACAT’08, Erice, Italy, 6 Nov. 2008
- E. Grancher/CERN, Oracle and storage I/Os, explanations, experience at CERN and SSD tests, UKOUG conference, Birmingham, UK, 2 December 2008
- A. Topurov/CERN, CERN Experience with Virtualization of Oracle RAC with Native Xen and Oracle VM, UKOUG Conference, Birmingham, UK, 2 December 2008
- E. Grancher/CERN, Learning from failures, design errors, problematic recoveries and downtimes of Oracle databases, experience at CERN, UKOUG conference, Birmingham, UK, 3 December 2008
- L. Canali/CERN and D. Wojcik, Implementing ASM without HW RAID, a user’s experience, UKOUG Conference, Birmingham, UK, December 2008
- M. Girone/CERN, Distributed Database Services – a Fundamental Component of the WLCG Service for the LHC experiments – Experience and Outlook, CHEP’09, Prague, Czech republic, 21-27 March 2009

Publications:

CERN openlab Reports:
- N. Basha/Summer Student, CINBAD Investigation of Different Packet Filters, August 2008
- X. Dong/Summer Student, Multi-Threaded Geant4 with Shared Detector, August 2008
- P-L. Hémery/Summer Student, Improving Display and Customization of Timetable in Indico, August 2008
- W. A. Romero/Summer Student, Performance Monitoring of the Software Frameworks for LHC Experiments, August 2008
- A. D. Dumitru/Summer Student, Oracle RAC Virtualization, September 2008
It starts with the Platforms!

- As of October: 64 HP Blade Servers w/Intel 3.0 GHz Quad-core processors
  - Now, cornerstone of most of our activity, Performance Monitoring, Teaching, Benchmarking, Compiler Testing, etc.

- Itanium servers (also used by BE and EN/CV)

- Individual machines/boards/drives
  - Alpha-level Nehalem server; Atom N330 board
  - Dunnington (24-core system from HP) (short-term loan)
  - Desktop Nehalem i7 board; Solid State Drive X25-E drive
  - Production-level Nehalem server from E4

- Several Intel software tools for general usage at CERN
  - C/C++/Fortran compilers w/floating licenses
  - Thread Checker, Thread Profiler, VTUNE
It starts with the Platforms!

- As of October: 64 HP Blade Servers with Intel 3.0 GHz Quad-core processors.

- Now, cornerstone of most of our activity, Performance Monitoring, Teaching, Benchmarking, etc.

- Itanium servers (also used by BE and EN/CV)

- Individual machines/boards/drives
  - Alpha-level Nehalem server; Atom N330 board
  - Dunnington (24-core system from HP) (short-term loan)
  - Desktop Nehalem i7 board; Solid State Drive X25-E drive
  - Production-level Nehalem server from E4

- Several Intel software tools for general usage at CERN
  - C/C++/Fortran compilers w/floating licenses
  - Thread Checker, Thread Profiler, VTUNE

---

Svenne Jarp – CERN openlab BoS 2009
PCC activities (in more detail)

Summary list:

- Intel’s Energy whitepaper (issued at LHC start-up)
- Second Thermal Study (G.Balasz, Published Feb09)
- Atom N330 benchmark evaluation
  - Paper and CHEP09 presentation
- Solid Xeon benchmarking beta-programme
  - Harpertown, Dunnington, Nehalem, etc.
    - Results communicated directly to Intel
- Benchmarking repository w/HEP jobs from multiple domains
  - Initial contents shown in PCC Major Review, Sept08
- ALICE/CERN HLT (High-level trigger) benchmarks: Track Fitter & Track Finder
  - Many-core focus (together with Intel/Brühl team)
- Perfmon reports
  - Used in multiple environments, including HEPiX May meeting and HEPiX benchmarking Working Group; CHEP09 talk
Summary list (cont’d):

- **Compiler project**
  - Intel icc 11.0 and icc11.1; GNU g++ 4.3
  - Focus on comparisons icc versus g++ (Xeon and Itanium)
  - Autovectorization (new proposal from Brühl)

- **New language: C-throughput collaboration**
  - Early prototype version; Feedback directly to Intel’s Technology Group

- **CERN Technical Training (together w/Jeff Arnold)**
  - Computer Architecture and Performance Tuning *(Spring + Fall each year)*
  - Multithreaded programming *(Spring + Fall each year)*

- **Cross-fertilization with other CERN entities**
  - PH Multicore project, G4 team, ROOT team, ALICE HLT team, etc.

- **Solid State Drive study** *(Initial results published in January)*

- **10 Gbit Network Cards** *(Initial test results at BoS 2008)*

- **TOP500 run** *(as burn-in test for production servers)*
  - Listed #96 in June 08 list (ISC08); #186 in Nov.08 list (SC08)
- **PVSS (control system for LHC and experiments)**
  - Oracle archiver scalability
    - Target achieved: 150’000 changes per second

- **Database virtualisation**
  - Target is to make better use of available infrastructure, ease management, improve security
    - Worked on “Oracle VM” and management pack, successful evaluation and tests, Oracle press-release

- **Monitoring and security**
  - Audit, control, improve database security
  - Provide global management and empower CERN developers

- **Validation of Oracle’s high performance “database engine”**
  - Optimisation provides stability for very high data loading (Exadata)

_Sverre Jarp – CERN openlab BoS 2009_
PVSS (ETM/Siemens) is CERN’s chosen SCADA

- Target from experiment and LHC machine is ~150,000 changes per second (different workload)
  - Far higher than initial scalability

- Worked since 2006 on the Oracle archiver, in collaboration with Siemens, EN-ICE and IT-DM
  - Provided new architecture and new code
  - Siemens has now included the code in baseline code (PVSS 3.8)
  - Validated March 2009, performance target exceeded with new hardware
Database Virtualisation

- Target is ease of maintenance, lower cost
  - hardware, power, cooling and space
- Oracle VM tested, performance gain over Xen
- Press release introduction Oracle VM Management Pack
- Live migration (demonstrated at last major review)
- Being introduced for some services

- Storage: feed back into Enterprise Manager the storage evolution, analysis and pro-active actions.
Some of our workloads (data loading for accelerators) are data insertion intensive, for these the tablespace creation is a problem.

Exadata has a number of offload features, most well-known are row selection and column selection.

Successful tests organised with Oracle.

Validated the functionality and stability gains.

Sverre Jarp – CERN openlab BoS 2009
Oracle and the Physics Database Services

Reliable and resilient database services are fundamental to all functional areas in the WLCG Computing Model

- simulation, data acquisition, first pass reconstruction, data distribution, re-processing, analysis, etc.

**Oracle 10g provides the Key Technologies to the Physics Database Services:**

- **Oracle RAC/ASM for availability, scalability, flexibility and consolidation**
  - Building block architecture for the Distributed Database Services at CERN and Tier-1 sites

- **Oracle Streams for data distribution between CERN and Tier-1 sites**
  - PVSS, detector conditions and file bookkeeping:
    - key for data (re-)processing

- **Oracle Data Guard for critical DB data protection**
Oracle and the Physics Database Services

Oracle RAC/ASM for availability, scalability, flexibility and consolidation

Building block architecture for the Distributed Database Services at CERN and Tier-1 sites

Oracle Data Guard for critical databases data protection

PVSS, detector conditions and file bookkeeping: key for data (re-)processing

Sverre Jarp – CERN openlab BoS 2009
Major Areas of Work in 2008

- **RAC and ASM**
  - **Standardized** on coherent setups for LHC experiments online, offline and standby databases – minimize complexity and diversity
    - Oracle version (10.2.0.4, Red Hat EL4, x86, 64-bit)
  - Coherent tool for database and streams monitoring/alerts integrated and extended to display Tier-1 status.
    - Feedback to EM developers
    - Streams Enhancements now in new EM version 10.2.0.5

- **Streams Replication**
  - Downstream cluster re-organization needed to increase space for spilled Logical Change Records (LCR)
    - Larger time window for sites to be down without need of splitting them
  - Automatic Split & Merge procedures to isolate a site if it goes down for more than a few days
  - Use of transportable tablespaces for site re-synchronization
Major Areas of Work in 2008 (cont’d)

- **Data Guard for critical databases**
  - physical standby deployed for all the mission critical production databases on the online and offline database clusters prior to the LHC start-up

- **Limiting database downtime in the event of:**
  - Multi-point hardware failures
  - Logical and physical corruptions
  - Disasters
  - Hardware upgrades
  - Human errors
    - within configured redo apply lag (24 hours)

- **Ad-hoc testing of major schema upgrades or data reorganization on the standby**
CINBAD Achievements

- System for on-line collection and processing of the sFlow data has been implemented and tested with 500 HP switches and routers

- Encouraging results from initial data analysis
  - influence on CERN security policies

- Strong interest from different parties at CERN and HP/Procurve in the CINBAD project
CINBAD Achievements

- System for on-line collection and processing of sFlow data has been implemented and tested with 500 HP switches and routers.
- Encouraging results from initial data analysis.
- Strong interest from different parties at CERN and HP/Procurve in the CINBAD project.

Sverre Jarp – CERN openlab BoS 2009
CINBAD Achievements (details)

- **sFlow data collector has been designed, implemented and tested on a large scale**
  - leveraged CERN’s data storage and analysis know-how:
    - LHC data experts, Oracle experts
  - successfully tested last summer,
    - more than 1.5 Terabytes of data collected over a few days

- **Initial data analysis**
  - statistical approach
  - pattern based approach
    - using adapted Snort (Intrusion Detection System) with sampled data, appropriate traffic rules and signatures

- **Various network anomaly findings**
  - CERN security policy violations, e.g. p2p, icq (instant messaging)
  - Trojans, viruses
GridMap

- Interactive new monitoring visualization of the Grid
  - Introduced at EGEE'07 (Oct'07), v2 in Feb'08, v3 in Mar'09
  - Visual correlation of importance and availability status
  - Top-level live management views of EGEE and WLCG grids
  - Integrated with OSG sites

- Used in production by CERN to help manage the Grid
- Technology is reused for other applications at CERN and EDS
- Influential in other communities e.g. D4science project

http://gridmap.cern.ch
MSG ( Messaging System for the Grid )

- **Flexible, reliable and scalable messaging infrastructure**
  - Production service running for several months
- **Two ActiveMQ brokers (CERN and Croatia)**
  - > 440 topics; > 60 queues
  - > 240 subscriptions (>20 of them are durable)
  - > 950 enqueued messages per minute
  - File Based Persistence for reliable delivery
  - Failover pair
  - Two protocols available: STOMP and OpenWire
- **Testing Nagios bridges**
- **Offering support to different projects within the IT Grid groups**
- **Monitoring system for message brokers under heavy development (project started in mid-February)**
Monitoring system for message brokers

- Easy-to-use web interface for monitoring message broker activity

List of Topics Nagios Sites

(30/03/2009 - 19:00:02)

Sverre Jarp – CERN openlab BoS 2009
TYCOON: A market-based allocation system

- Project concluded after two years of investigations in openlab II
  - Close collaboration with HP Labs (Palo Alto), BalticGrid, and EGEE
  - Integration of Tycoon with gLite
    - Automatic deployment of Compute Elements and Worker Nodes
  - Multiple scalability tests performed
  - Tycoon experience presented at several EGEE conferences in 07 and 08
  - Reports with our experience
    - HP Labs, openlab Web site
  - Tycoon now used in HP’s Cloud Computing Initiative
Efficient and non-intrusive resource allocation in Grids

Three years of PhD studies in collaboration with HP Labs (Bristol)

Central point in thesis:
- Cost effectiveness of a given resource allocation
  - With several independent participants
  - Based on separation of supply and usage
- Key paper recently submitted to SMTPS’09
Projected signed last year

- Program of work: 1) PVSS  2) PLCs
- One staff and three fellows now in place
- First results will be reported by Siemens (today)
PVSS related program of work

- Open the PVSS development environment to Software Engineering
  - Source code management
    - CVS, Subversion
    - Panels, files and data
  - Configuration management
  - Improvement of debugging facilities
  - Toward a standard scripting language?

- PVSS deployment in large environments
  - Monitoring & deployment

- Security
  - Engineering & Operations
PLC related program of work

- **Security**
  - Definition of robustness & vulnerability tests
  - Hardening of automation devices
    (Operation and engineering perspectives)

- **Opening Step 7 to software engineering**
  - Source code management
  - 3rd party development tools

- **Deployment in large environment**
  - Step 7
  - Simatic Net
  - and others
Conclusions

- Excellent collaborations between partners and CERN teams
- In my eyes, an impressive set of contributions
  - from each of the multiple openlab teams
  - in most cases, the corresponding technologies are already deployed in production
    - Or, ready for wider deployment
- CERN openlab III starts on strong footing
  - Solid teams ready to invest effort into the agreed R&D domains
- I am optimistic that, also in openlab III, we will continue to deliver great results

Thanks to everybody who contributed to this slideset!