HEPiX Large System SIG Report

Platforms For Physics

Linux

Grid Operations Experience

Matthias Schröder, IT/PS

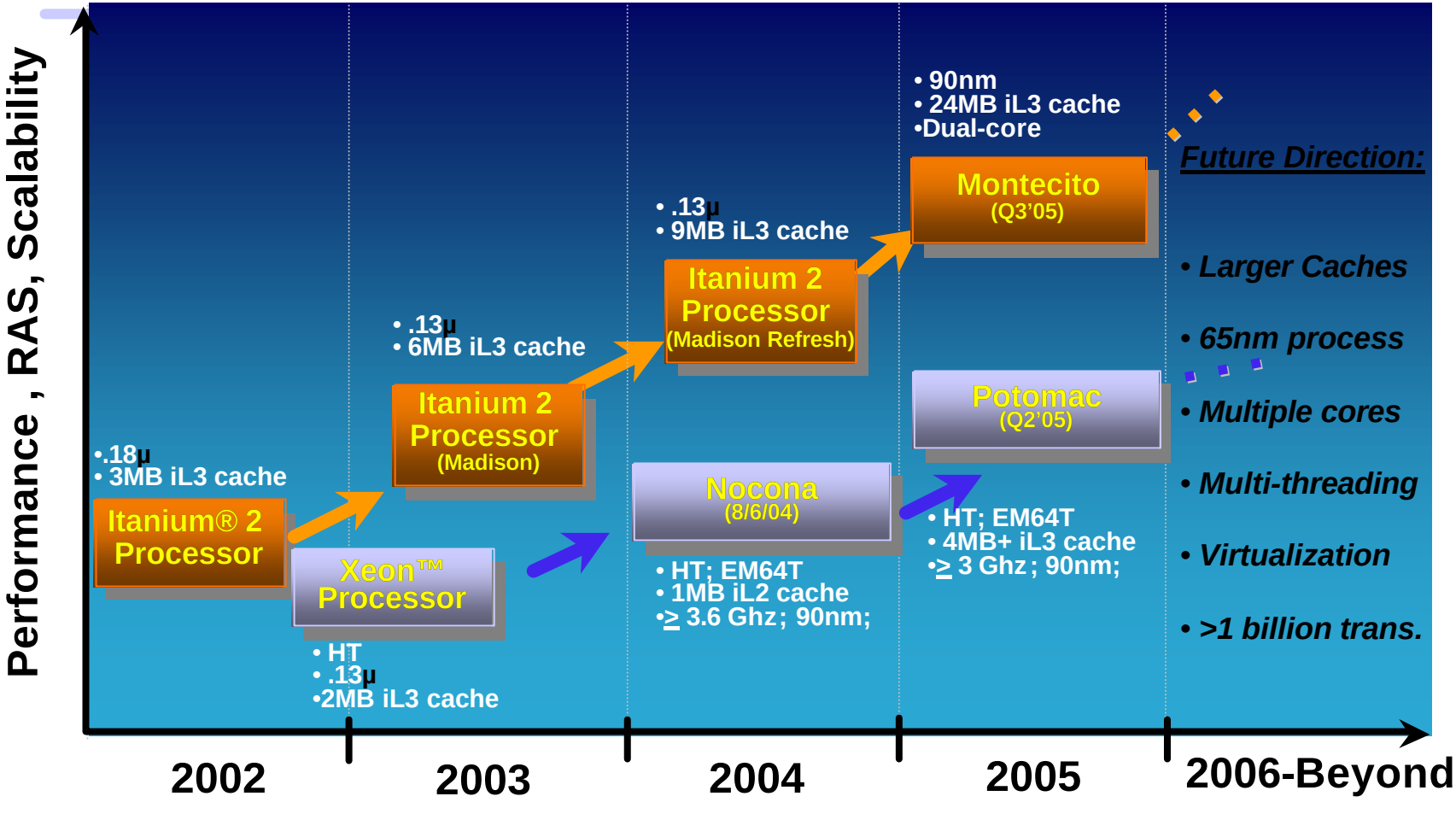
Platforms For Physics

- Intel Processor Roadmap
- 64 bit Extensions
- Performance AMD/Intel
- Cooling Issues
- Huge Memory System
- Mac G5 Cluster

Intel Processor Roadmap [King]

- Intel will maintain Xeon and Itanium lines
- Xeon expected to follow Moore's law
- Itanium expected to outpace Moore's law
  - Itaniums performance
- Xeon better price/performance than Itanium
  - But gap is expected to narrow with time, close by 2007?
- Itanium should be first to get multi-cores
  - …and stay ahead of Xeon
    - 2 cores in 2005
    - 4 cores in 2007 (target is 8 cores in 2007)

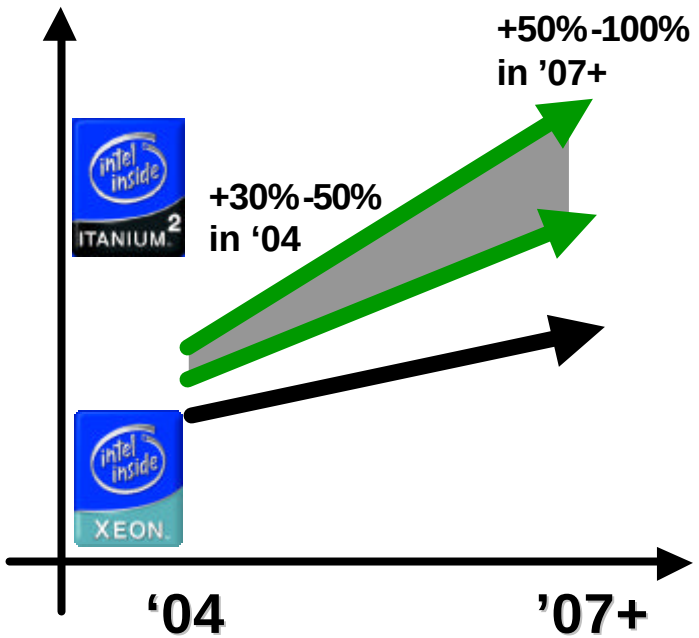
Intel® Server Processor Roadmap



Intel® Server Processor Platforms

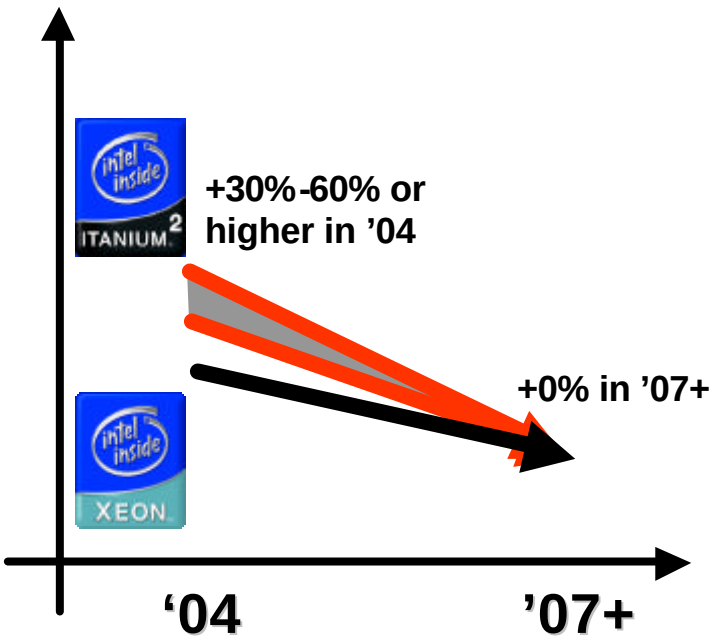
Performance [1]

On track to deliver 1.5-2X better performance than Intel® Xeon™ processor



Platform Cost [2]

While achieving platform cost parity via common platform infrastructure



[1] **Data based on Intel projections.**

[2] **'04 Price based on comparable OEM systems, HW only for enterprise and technical computing applications.**

Source : www.ioncomputers.com, www.dell.com

ION SR4004, (4) Xeon processors 2.8 GHz, 2MB cache, 24 GB system memory, 36 GB HDD, no OS - $34,616

ION I2X4, (4) Itanium 2 processors 1.5 GHz, 6 MB cache, 24 GB system memory, 36 GB HDD, no OS - $44, 950

Dell PowerEdge 2650, (2) Xeon processors 3.2 GHz, 1 MB cache, 4 GB system memory, 36 GB HDD, no OS - $6,143

Dell PowerEdge 3250, (2) Itanium 2 processors 1.4 GHz, 1.5 MB cache, 4 GB system memory, 36 GB HDD, no OS - $9,499

64 bit Extensions [King]

- 64 bit extensions allow addressing larger memory
  - 40 bits physical, 48 bits virtual memory
- Full 4 GB address space, easy to use with legacy apps
- Require adapted OS
  - Microsoft
    - (planned for 1H05)
  - Redhat
    - EL 3 update 2
  - Suse
- Have to watch for availability of needed drivers & compilers

64 bit Extensions AMD/Intel [Wiesand]

- DESY looked at 64 bit extension systems
  - AMD64 / Xeon EM64T
- Ran performance tests
  - Root, Sieglinde, Pythia, FORM
  - Compared build times and run times
    - Please see http://www.rhic.bnl.gov/hepix/talks/041021am/wiesand.pdf for details
  - 64 bit mode much better performance than 32 bit
    - Up to 30%
  - Opteron better performance than Xeon
    - Even more pronounced for second CPU (memory access?)

64 bit Extensions – Issues [Wiesand]

- All libs have to be recompiled for 64 bit mode
  - Cernlib is not
    - Blocks many legacy apps from profiting from performance boost
- FP registers are 64 bit again
  - Not 80 bit as in x87 units
- 64 bit Linux systems allow using 32 bit apps
  - 32 bit libs live under ../lib
  - 64 bit libs live under ../lib64
    - …with some exceptions…
  - Mixed installation sometimes a pain
- 64 bit ext's allow us to prepare code for 64 bit era

IA64 Servers [Hirstius/Horvath/Iven]

- Test setup for LHC Tier 0 data rates
  - A.Hirstius and A.Horvath, IT/ADC
  - In framework of openlab
  - Total data rate estimated at 55GB/s
- Target for 2005: 500MB/s disc to disc, sustained
- Requires multiple 1 Gb/s or 1 10 Gb/s connection
  - 10 Gb/s connection more expensive, but closer to reality
- Few 10Gb connections available from CERN

IA64 Servers [Hirstius/Horvath/Iven]

- Disc to memory results:
  - Internal data rates (ms+ss)
    - ~770MB/s read
    - ~350MB/s write
  - 10Gb back-to-back (ms+ss)
    - ~710MB/s read (NIC is the limitation)
    - ~350MB/s write
  - 10Gb to California (ss)
    - **~660MB/s** read :-)) (~1300GB in 2000s)
- Similar tests using Infiniband have started

Cooling Issues [Alef]

- High density systems produce lots of heat per m^2
- In extreme cases setup at limit of air cooling
- Potential solution: water cooling
  - Cool CPU directly
  - Cool the rack
- Karlsruhe went for water cooled racks from Knürr
- Currently using 30 racks
  - 25 more on order
- Experience very good
  - No significant problems
  - Noise level reduced significantly

Mac G5 Cluster [Boeheim]

- Requested from Astrophysics group at SLAC
  - Already used as desktops
  - Nice visualisation features
- 2 file servers, 2 interactive servers, 10 compute nodes
- Hardware installation very easy
  - Nice setup fitting in 19” racks
- No remote BIOS management and no power mgt.
- BSD based OS, but many confusing config details
  - Passwd file exists, but not used
- Non-case sensitive HFS+ file system
  - Makefile and makefile are the same

Huge Memory System [Wachsmann/Mount]

- Major computing challenge: sparse access to objects in large datasets
- Going through huge datasets dominated by latency
  - True for tape and disc
  - Going several times through whole dataset painful
- Loading whole dataset into memory would ease ‘skimming off’ interesting events
- Huge memory machines would allow this
  - Cluster of fast CPUs, huge memory, high bandwidth netw
  - New techniques for memory access needed
- Interesting not only for HEP

Linux

- Scientific Linux
- Experience with RedHat TAM

Scientific Linux

- Several sites decided to go for Scientific Linux
  - Fedora too short support times
  - RH Enterprise Linux too expensive
    - Except for specific applications on limited # of nodes
  - Shortcomings of other distributions
    - Require change of installation procedures
    - Support for third party apps uncertain
- Many now preparing or started conversion
  - Some nodes still on old Linux versions
  - Some nodes available under SL
- Sites preparing own systems for configuration mgt.
- FNAL have put major effort into providing SL

Scientific Linux

- CERN makes special release of SL
  - CERN add-ons
  - Distribution closer to RH than original SL
  - Other sites can use CERN distribution with or without CERN mods
- Original SL already contains handles for local tailoring

RedHat TAM

- Technical Account Management
- Allows close technical contact with RedHat
- RH contact person knows on site environment
  - Not always true for replacement…
- Get early information about RH plans & developments
- Response to requests for help rather fast
- Response to change requests much slower
- Help restricted to pure code from RH distribution
- Experience are mixed
    - SLAC happy with support
    - CERN still evaluating usefulness

Grid Operations Experiences

- Operational Models
- Monitoring and Accounting
- Incident Response/Security

General Operations Issues

- Not all sites created equal
  - Large sites
    - Might be part of many grids
    - Want flexible & powerful installation tools
    - Have staff for monitoring & support
    - Might discover issues with Tiers ‘below’
  - Small sites
    - Usually participate in one grid only
    - Want easy to use installation tools
    - Have little or no staff for monitoring & support
    - Are reluctant to allow intervention from Tier ‘above’

Communication between sites

- LCG developed hierarchical model with Regional Centers
  - Allows more flexibility and regional policies
  - GOC not supposed to contact sites directly
  - Adds latency in communication with sites
- In reality lots of direct communication with sites
  - Very resource consuming for GOC
- Unique problem tracking system for all levels & all sorts of problems needed
- Lack of true hierarchy requires cooperative approach

Data Challenges

- Large scale production effort of the LHC experiments
  - test and validate the computing models
  - produce needed simulated data
  - test experiments production frame works and software
  - test the provided grid middleware
  - test the services provided by LCG-2
- All experiments used LCG-2 for part of their production
- Daily ‘Grid certification’ important for success
- Many operational issues discovered
  - See Markus slides for full list…
- Grid middleware rather stable now
- Better tools needed for bulk handling of jobs

Monitoring

- LCG: now ~8000 CPUs & 96 TB discs on 83 sites
- Many local particularities of grid sites
  - complicates monitoring
  - LCG GOC established configuration database of sites
- A number of tools are used for monitoring
  - See http://www.rhic.bnl.gov/hepix/talks/041022am/kant.ppt & http://www.rhic.bnl.gov/hepix/talks/041022am/grundhoefer.pdf
  - Many don't have feature to store monitoring info
    - Relational Grid Monitoring Archiver stores their output
    - Information stored by R-GMA can be used for accounting

Incident Response Plans

- Open Science Grid
  - Has little or no control over physical resources
  - almost everything has to be done by the sites or the Vos
  - Sites security personnel will need to feel comfortable with grid use of resources
    - limited additional risks
    - local control over decisions
  - Centrally Provided:
    - List of site security points of contact
    - Secure email communications
    - Incident Tracking system
    - Functions of the Grid Operations Center (GOC)
    - Coordinate with other GOCs

Incident Response Plans - OSG

- Responsibility of local sites well defined
  - See http://www.rhic.bnl.gov/hepix/talks/041022am/cowles.ppt
- Incidents classified with 3 levels of severity
  - Potential to compromise grid infrastructure
  - Potential to compromise grid service or VO
  - Potential to compromise grid user
- Response Teams to be created to deal with serious incidents
- Procedure for handling incidents defined
- Plan (will be?) discussed with EGEE and LCG

- All glory goes to the authors of the original slides
- Special thanks to H.Meinhard and A.Silverman for their detailed trip reports
- All errors and omissions are my fault…