



FP7 Support Action - European Exascale Software Initiative

DG Information Society and the unit e-Infrastructures



EESI Final Conference

Exascale key hardware and software
technological challenges and their impacts

Speaker: **Bernd Mohr**, Juelich

Typical HPC Hardware and Software Universe



Scientific Applications

Numerical Libraries, Solvers and Algorithms

System Software Eco-system

HPC System (including compute nodes, interconnect, storage)

Scientific
Software
Engineering

An Opportunity for Europe



- Europe is strong in software and algorithms
 - HPC applications (2011 IDC survey):
83% developed in Europe, 66% of IP in Europe

- Europe is experienced in executing large international projects
 - CERN, ITER, PRACE, ...

- Europe can play an important role in designing and constructing the necessary system software, programming environment and tools

- International collaboration is indispensable
 - Because of the scale of investment
 - And because applications teams and research communities are often already international

- Europe and Japan emphasize the importance to include both academic and industrial needs

- US, Japan and Europe already cooperate

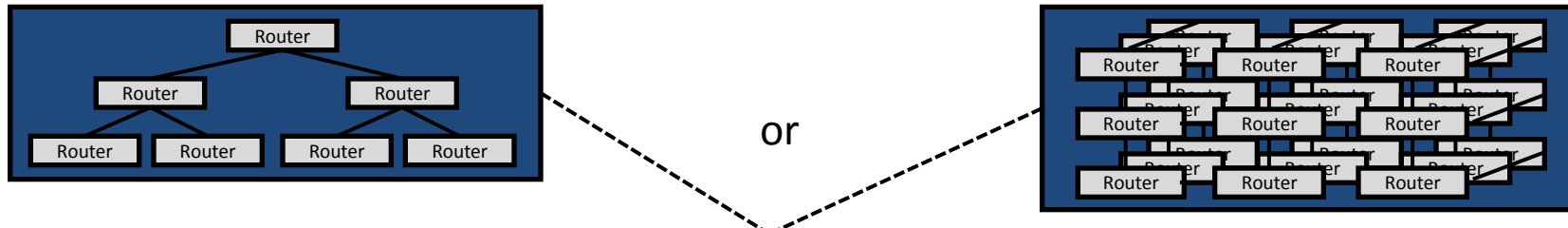
- Russia and China also announced Exascale efforts

- Europe is key player in some HPC software areas
 - Programming models and runtime systems
 - Performance and debugging tools
 - System design (mobile, network, energy efficiency)
 - Simulation frameworks/coupling tools, meshing tools
- Europe has some of the best scientists in applied mathematics
 - Many existing scientific libraries either developed in Europe or significant European input to US-led/international projects
- In some of these domains Europe has shown its capacity to support long term and costly efforts

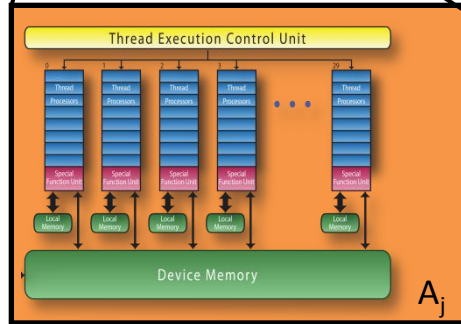
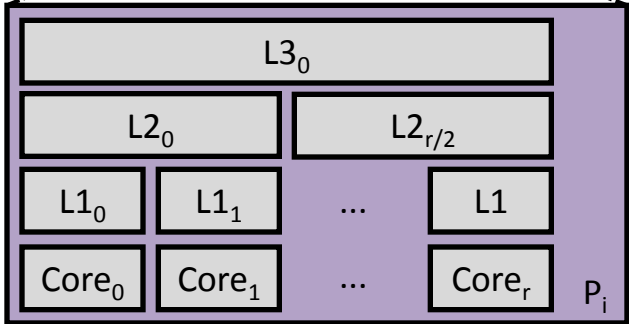
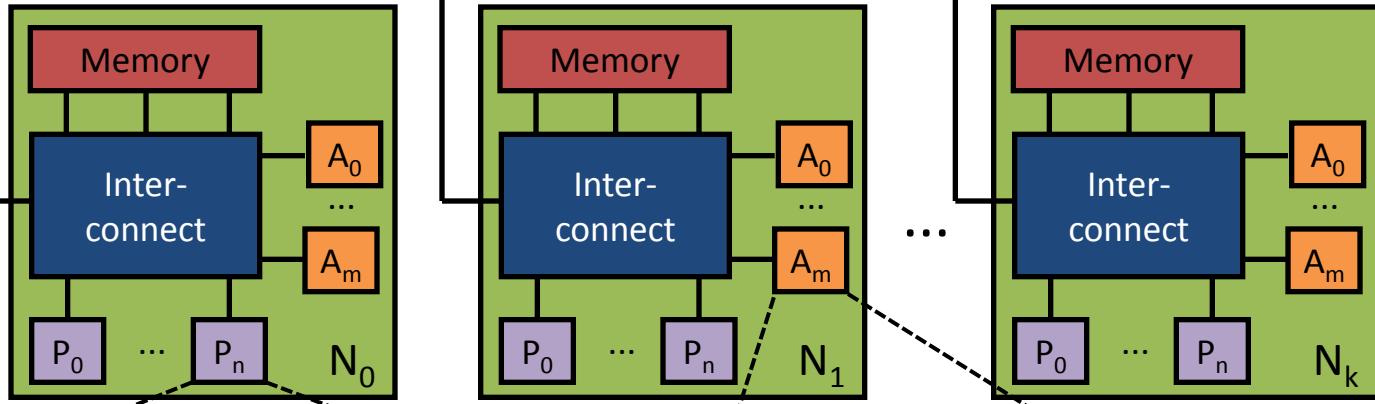
Exascale Key Hardware and Software

TECHNOLOGICAL CHALLENGES

Technical State of the Art



Network or Switch



Exascale:
#nodes $\rightarrow 10^3 - 10^4$!
#cores $\rightarrow 10^5 - 10^6$!

- Currently two main directions
 - “Massively Parallel Processing” System
 - Weakly heterogeneous
 - 1,000,000 nodes of 1,000 core processor
 - IBM BlueGene (US) or Fujitsu K (JP) like architecture
 - “HPC Cluster with Accelerators”
 - Highly heterogeneous
 - 100,000 nodes of 10,000 core processor with acceleration
 - Tianhe-1A (CN), TSUNAME 2.0 (JP), Titan (US) like machine

Potential System Architecture Targets



| System attributes | 2010 | "2015" | | "2018" | | Difference 2010-18 |
|----------------------------|-----------|---------------------|----------|---------------------|-----------|--------------------|
| System peak | 2 Pflop/s | 200 Pflop/s | | 1 Eflop/sec | | O(1000) |
| Power | 6 MW | 15 MW | | ~20 MW | | |
| System memory | 0.3 PB | 5 PB | | 32-64 PB | | O(100) |
| Node performance | 125 GF | 0.5 TF | 7 TF | 1 TF | 10 TF | O(10) – O(100) |
| Node memory BW | 25 GB/s | 0.1 TB/sec | 1 TB/sec | 0.4 TB/sec | 4 TB/sec | O(100) |
| Node concurrency | 12 | O(100) | O(1,000) | O(1,000) | O(10,000) | O(100) – O(1000) |
| Total Concurrency | 225,000 | O(10 ⁸) | | O(10 ⁹) | | O(10,000) |
| Total Node Interconnect BW | 1.5 GB/s | 20 GB/sec | | 200 GB/sec | | O(100) |
| MTTI | days | O(1day) | | O(1 day) | | - O(10) |

- Huge numbers of cores
 - ⇒ Fault tolerance and total power consumption become major issues
- Tightening memory/bandwidth bottleneck
 - Decreasing memory-per-core and byte-to-flop ratios
- Huge numbers of cores + memory bottleneck
 - ⇒ Even higher number of threads needed to hide latency
 - ⇒ Extreme concurrency and hierarchical approaches required
- Extreme concurrency + heterogeneity
 - ⇒ Programmability?
 - ⇒ Efficiency?
- I/O and data intensive computing become more important

- Special Exascale requirements and high costs lead to
 - One-of-a-kind architectures / systems
 - Debugging and testing at customer site
 - Higher ratio of community-developed open-source SW
 - Even in areas traditionally implemented and sold by vendor like operating and file systems, system management, batch systems, development tools
 - Collaborative Co-design needed involving
 - HW vendors
 - System software and tool developers
 - Application programmers

Revolution or Evolution?



- ❑ Some experts are convinced that only a radically new way of programming will allow efficient programming of Exascale architectures
- ❑ Others point out huge number of existing HPC software packages and high cost of adaptation to new programming models and architectures
 - Smooth transition path need
 - Possible path:
 - ⇒ MPI
 - ⇒ MPI + pragma-based tasking on CPU and accelerators
 - ⇒ MPI + multi-thread programming model + accelerator programming model

Exascale Key Hardware and Software

IMPACTS

- Hardware and system software are **enabling** technology for Exascale applications

 - Basic programming environment and performance, validation and debugging tools increase efficiency of applications and productivity of programmers
 - Reduced time to solution
 - More and/or higher quality results in the same time
- Multiplying** impacts of applications

- Large companies and SMEs with a niche business model will grasp the added-value
 - Energy-efficiency (e.g. cooling)
 - Fault tolerant and error correcting hardware parts
 - Integrators for one-of-a-kind systems

- Complex programming will create higher need for consulting services
 - Porting
 - Performance optimization and tuning
 - Training

- Research centres that will host the most advanced computing capability will attract talents to build and disseminate knowledge
 - Only one (or very few) Exascale systems centres because of
 - High costs due to system size and complexity
 - Power requirements
 - Exascale computing centres will need to operate more like other large-scale scientific instruments
 - No longer feasible to move data to/from researcher home organization
- Unprecedented level of cooperation and coordination will enable us to deliver the value

- Exaflop also means
 - Petaflop systems in one box
 - For 200K EUR and 20 KW

- Huge impact for those, academic, industrial, large and small structures, including SMEs, that will be able to take advantage of “Exascale” technology

Exascale Hardware and Software

KEY FINDINGS

Typical HPC Hardware and Software Universe



Scientific Applications

WG 4.4

Numerical Libraries, Solvers and Algorithms

WG 4.3

Scientific
Software
Engineering

System Software Eco-system

WG 4.2

HPC System (including compute nodes, interconnect, storage)

WG 4.1

The EESI System Software Roadmap



- Engaged 60 experts in four working groups
 - Produced a cartography of HPC vendor research activities worldwide and identified gaps
 - Produced an Exascale Software Roadmap including
 - Description of the scientific and technical issues + challenges
 - Societal, environmental and economic impacts
 - European strengths and weaknesses
 - Needs of education and training
 - Existing funded projects and collaborations
 - Timeline and costs
 - Investigated scientific software engineering issues

- HPC vendors are willing to collaborate with the EC and the European HPC community
- At least one European HPC system vendor has recently successfully demonstrated its expertise to build HPC system at the multi-Petascale performance level
- Europe still has all the necessary knowledge at hand to develop its own HPC technology stack

- ❑ Most HPC system software components have to be substantially adapted or newly developed to address Exascale challenges
- ❑ Application developers are reluctant to adopt new, not yet fully-accepted and standardized programming models
 - ⇒ A transition path for HPC applications to the Exascale programming models is required
- ❑ The different Exascale system architectures should not be visible at the level of the programming model
 - ⇒ The SW ecosystem must be portable
- ❑ Collaboration between HPC hardware and system vendors and European R&D laboratories and Academic researchers has to be ensured
 - ⇒ Co-design processes must be established

- Significant funding is essential in R&D areas where Europe has technology leadership and critical mass: programming models & runtimes, performance, validation and correctness tools.
 - ⇒ The key players should form alliances to define standards

- Establish cooperation programs with technology leaders in areas where European technology is not leading
 - ⇒ To ensure that system components and standards can be influenced from a European perspective
 - ⇒ To revitalize research in these important areas in Europe

- New fault-tolerant and scalable algorithms and their implementation in portable mathematical libraries, solvers and frameworks are a key prerequisite for developing efficient Exascale applications
 - ⇒ An Exascale Applied Mathematics Co-Design Centre should be established to coordinate and bundle European efforts for this important topic

- Consistency and completeness of the software stack as well as portability, robustness, and proper documentation has to be ensured
 - ⇒ An European Exascale Software Centre should be established to coordinate research, development, testing, and validation of HPC Exascale software ecosystem components developed in National and EU projects

- Co-design and development
 - Domain Scientists
 - Mathematical algorithm designers
 - Numerical computing experts
 - Computer scientists for parallelization and optimization
- Fault-tolerant and performance aware design and implementation from the beginning
 - Not as an afterthought or 2nd step
 - Might also include power saving controlling
- Increased importance of standards
 - Because of one-of-a-kind computing systems, high software development costs, and legacy software
 - Need to be backward-compatible with Tera- and Petascale systems

Costs System Software Eco-system



| Topic | Cost (PY) |
|-------------------------------------------|-------------|
| Parallel programming models and compiler | 568 |
| Runtime systems | 184 |
| Debugger | 42 |
| Correctness and validation tools | 62 |
| Performance tools | 168 |
| Performance modeling and simulation tools | 42 |
| Operating systems and systems management | 84 |
| Batch scheduler and resource manager | 42 |
| I/O and file systems | 100 |
| Resilience | 140 |
| Power management | 68 |
| TOTAL | 1500 |

Costs Mathematical Software



| Topic | Cost (PY) |
|-------------------------------------|-------------|
| Dense linear algebra | 226 |
| Graph partitioning | 70 |
| Sparse direct methods for $Ax=b$ | 111 |
| Sparse iterative methods for $Ax=b$ | 180 |
| Eigenvalue problems model reduction | 55 |
| Optimization | 215 |
| Control of complex systems | 88 |
| Structured and unstructured grids | 55 |
| TOTAL | 1000 |

Costs Scientific Software Engineering



| Topic | Cost (PY) |
|-----------------------------------------------|-------------|
| Domain Specific Languages | 218 |
| Coupling software and data exchange standards | 48 |
| Coupling and workflow technologies | 278 |
| Advanced mesh generation and partitioning | 278 |
| Flexible generic I/O layer | 98 |
| Computational steering | 278 |
| Checkpoint/restart to fast memory | 98 |
| API for error signaling and reporting | 98 |
| Fault-tolerant application support | 328 |
| Integration Development Environments | 278 |
| TOTAL | 2000 |

Costs for Timeframe 2012 - 2020



- Exascale system software and programming tools
 - 150 Meuro (1500 person years effort)
- Scalable algorithms, mathematical libraries and solvers
 - 100 Meuro (1000 person years effort)
- Application frameworks, workflows, visualisation and data management
 - 200 Meuro (2000 person years effort)
- Applied Mathematics Co-design Centre
 - 40 Meuro
- European Exascale Software Centre
 - 500 Meuro

THANK YOU