FP7 Support Action - European Exascale Software Initiative

DG Information Society and the unit e-Infrastructures

EESI System Software WG 4.2

Franck Cappello, INRIA and University of Illinois

Bernd Mohr, JSC

François Bodin, Marc Bull, Toni Cortés, Jesus Labarta, Jacques C. Lafoucriere, David Lecomber, Thomas Ludwig, Simon McIntosh-Smith, Jean-François Méhaut, Matthias Müller, Raymond Namyst, Olivier Richard, Pascale Rossé, Karl Solchenbach, Vladimir Voevodin, Felix Wolf

WG4.2 Scope and Challenges

- Software Eco-system: between Applications and Hardware
  - Programming models, Compilers, Runtime, Tools, OS, I/O, System management
  - WG4.2 does not cover Framework, Numerical libraries, Workflow, Scientific data management, Visualization → other WGs
- Many technical challenges on all layers of the software stack
  - Scalability
  - Heterogeneity
  - Power
  - Errors, Faults, Failures
  - Reducing as much as possible Overhead, Jitter, Noise, etc.
- A trend toward community software

WG4.2 TOPICS

- System Software
  - Operating Systems, System Management
  - Job and Resource Manager
  - Runtime Systems
  - I/O systems
- Development Environments
  - Programming Models, Compilers
  - Debuggers
  - Performance tools
  - Correctness tools
- Crosscutting Dimensions
  - Resilience
  - Power management

WG4.2 Chairs, Leaders and Experts

Chairs:

Franck Cappello	INRIA&UIUC,	1 FR,	**Resilience and FT**
Benrd Mohr,	JSC,	1 DE,	Performance tools
Jesus Labarta	BSC	1 ES	**Programming models**
Marc Bull	EPCC	1 UK	OpenMP / PGAS
François Bodin	CAPS	2/V FR	Programming/Compiler for GPUs
Raymond Namyst	LABRI Bordeaux,	3 FR	MPI&OpenMP **Runtime**, GPUs
Jean-François Méhaut	INRIA Grenoble	4 FR	Performance Modeling/Apps.
Matthias Müller	TU Dresden	2 DE	**Validation/correctness Checking**
Felix Wolf	GRS	3 DE	**Performance Tools**
David Lecomber	ALINEA	2/V UK	**Parallel debugger**
Simon McIntosh-Smith	U. Bristol	3 UK	Computer Architecture & FT
Vladimir Voevodin	MSU	1 RU	Performance tools
Thomas Ludwig	DKRZ	4 DE	**Power management**
Olivier Richard	INRIA	5 FR	**Job and Resource Manager**
Jacques C. Lafoucriere	CEA	6 FR	**I/O, File system**
Toni Cortés	BSC	2 ES	I/O, Storage
Pascale Rossé	BULL	V FR	**OS, System management**
Karl. Solchenbach	Intel	V BE	All

WG4.2 EESI Roadmap content for Sys-soft eco system

- Starting point: IESP roadmap, Exascale studies, Soft. used in HPC centers
- Description of the scientific and technical perimeter
- Social benefits, societal, environmental and economical impact
- S
- A ty,
  P
- E
- S
- N
- P
- Existing funded projects and funding agencies
- *Timeline, needs of HR, provisional costs,*
- Needs for experimental platform: size, reconfigurable?, dedicated?
- What software in an Exascale stack and with what level of responsibility

Next slides → only a small subset of the current results
sort of executive summaries

The roadmap with timeline, HR and cost
is not yet completed

Recommendations for EU may evolve

WG4.2 Programming model

Chair: Jesus Labarta, BSC

- Programming models
  - Clean separation and interaction between application developer and system (hardware and software)
  - Improve productivity, Significantly reduce maintenance costs
  - **Incremental parallelization will reduce development costs**
- Main issues at Exascale:
  - Concurrency, Asynchrony, Malleability, Address spaces and locality, Hierarchy, I/O, Modularity, interoperability, Productivity/portability, Incremental path,
  - **But also more decoupling between programmers and machines,**
  - Cross cutting: Power Management, Fault tolerance, Performance
- Recommendations for EU
  - Existing developments with demonstrated potential and willing to play a role in the future exascale software environment: HMPP, StarSs (OmpSs)
  - Important contributions to standards: MPI/OpenMP
  - **Programming models that provide expressiveness, incremental portability and performance** will boost performance, visibility and impact of applications having early access to them

WG4.2 Runtime

Chair: Raymond Namyst, U. Bordeaux, INRIA

- Bridge the gap between underlying architecture and application requirements:
  - Scheduling, load balancing, Memory management, Communications, Sync.
  - Where accurate information is available about the actual power consumption of various hardware parts
- Main issues at Exascale:
  - Mastering heterogeneity: Unified/transparent accelerator models, Support for adaptive granularity, Fine grain parallelism, Scheduling for latency/bandwidth
  - Dealing with millions of cores/nodes: Scheduling, communication
  - **Supporting multiple programming models: MPI + threading model + accelerator**
  - Robustness: reconfigure itself when resources suddenly disappear.
- Recommendations for EU
  - Many European countries have a long-standing activity in runtime design
  - **Unified runtime system**, providing a unified API to deal with threads and lightweight tasks (together with their integration with MPI/PGAS communication systems)

WG4.2 Validation / Correctness

Chair: Matthias Müller, TU Dresden

- Tools and methods for validation and correctness checking:
  - Validate a program in accordance to a model/specification (ex: MUST for MPI)
  - Can detect many errors, Especially: portability and non-determinism related bugs
  - Exascale will drastically increase manifestation rates!
  - **Correctness & validation helps reducing the time to solution** (better productivity)
- Main issues at Exascale:
  - Scalability, Fault tolerance, Adaption to new paradigms, Integration into new paradigms, **Integration into debugging workflow**
- Recommendations for EU
  - MUST and Marmot collaborations: Supercomputing vendors, especially those involved in DARPAs HPCS program (Cray, IBM), Software companies like Rouge Wave (Totalview), National labs (LLNL, LANL, ORNL, ANL), Universities (Utah, Houston)
  - **Foster interaction of methods/groups** within and beyond Europe
  - **Horizon. interaction** of approaches/tools: compiler + runtime + classic debugger
  - **Vertical integration** of validation and correctness at all layers of software stack

WG4.2 Parallel Debugger

Chair: David Lecomber, ALINEA

- Any method to assist with the removal of software failures
  - Interactive tools, user tactics, Alinea DDT, Totalview Rogue Wave's, etc.
  - Primary focus on interactive tools that give user knowledge and control of program state and activity. **Limited to in single execution path.** Can be improved by static source code analysis.
  - **Some overlap with automatic Validation/Correctness tools**: Automated problem detection tools: MUST (TU-Dresden), or Umpire (LLNL).
- Main issues at Exascale:
  - Scale: responsiveness of debugger to user command and overhead), Architectural and Programming Model Unknowns (debugging support)
  - Crosscutting issues: Programmability, Resilience, I/O+Network, Reproducibility
- Recommendations for EU
  - **Continued financing/Sustainability** for Debugging tool developed in EU: DDT from Allinea, Globally only debugging product with Petascale ability. Collaborations with leading European research centres and initiatives and US
  - **Integration into debugging workflow with validation and correctness**

WG4.2 Performance Tools

Chair: Felix Wolf, GRS

- Diagnostic programs for performance optimization of applications
  - Ideally used in combination with performance modeling
  - Huge savings in terms of time and energy, Enables new scientific discovery: by calculating larger problem sizes.
- Main issues at Exascale:
  - Million-fold concurrency, Deeper memory hierarchies, More dynamic execution, **Limited reproducibility (asynchrony, faults and power managmt)**, **Limited bandwidth to extract performance data**, **Multiple programming models**
  - Cross cutting: resilience (tool should be FT), power, programmability
- Recommendations for EU
  - Many of the performance tools used in production today are made in Europe, **Europe has leading position in this area**
  - OPT (Allinea), ThreadSpotter (RogueWave), Vampir (GWT-TUD), Intel trace collector and analyzer originally from Europe
  - Popular Academic tools, Paraver/Dimemas (BSC) Scalasca (Jülich/Aachen)
  - Initiative for more integration with Score-P measurement system
  - **European tool builders can make significant contribution to exascale effort, but sustainability of funding must be ensured.**

WG4.2 I/O and File System

Chair: Jacques C. Lafoucriere, CEA

- Permanent storage: keep track of computing results for post processing and to start a new computation
  - Online (Disk, SSD, etc.) and Offline (tapes, disk based virtual tapes)
- Main issues at Exascale:
  - Extreme concurrency level (data, metadate locking), 1M disks, End2end integrity, too many transactions (contention), check usefulness of archived data
  - Storage plays a major role in Fault tolerance, NV memory → new failure scenarios
  - **Some HPC centers prefer Open Source FS** (or will be locked to a vendor)
- Recommendations for EU
  - 2 major EU companies offer storage solutions for Petascale computers
    - Xyratex: hardware and proprietary file system (Colibri)
    - Bull: hardware neutral integration of open source file systems (Lustre)
  - Significant set of small groups with good impact in the storage community around Europe. SCALUS as a 1st coordination initiative
  - Storage is a critical part of an Exascale solution
  - **EU must be a major contributor/partner to Open Source file system**

WG4.2 OS and System Management

Chair: Pascale Rossé, BULL

- OS and System management:
  - OS is a key element between hardware and runtime/application
  - Developing and debugging large scale HPC systems requires experts at all layers of the software stack, including OS.
- Main issues at Exascale:
  - Many cores, Low overhead, low noise, etc.
  - Scalability: Huge amount of statistics, log data, events VS Centralized tools
- Recommendations for EU
  - **Hardware initiative in Europe** around ARM, AMD Fusion, Nvidia Denver, etc. **should be complemented by an OS initiative**.
  - Sustain and develop further OS R&D for HPC in Europe
  - **Evolution rather than revolution**. Revolution would require too much effort for runtime/apps porting
  - **Standardization of event messages across sources, Scalable tools**
  - Automatic or assisted diagnostic, root cause analysis, etc.

WG4.2 Power Management

Chair: Thomas Ludwig, DKRZ

- Understand power consumption as a function of system usage:
  - Develop SW to control the HW mechanisms: SW in the OS, run-time system, etc.
  - Deploy next generation HW with energy saving mechanisms
  - A cost efficient and cost aware high performance computing is crucial for the competitiveness in science and engineering
- Main issues at Exascale:
  - Resilience (switching off/on components), Programmability (manual instrumentation), Performance optimization, Reproducibility (non-deterministic methods influence performance predictability and system noise)
  - **Power management API standardization**
- Recommendations for EU:
  - Trace tools: Vampir – Dresden, Scalasca – Jülich
  - Control daemons: Grid monitor – ParTec Munich, Germany, Power manager – Bull Cluster management France → Power capping, Power accounting)
  - **Leverage European leadership** in performance analysis tools, energy efficient hardware design and Know-how from embedded system

WG4.2 Resilience

Chair: Franck Cappello, INRIA and UIUC

- Fault tolerance is a multi-facets and crosscutting issue:
  - **No guarantee that all faults will be masked by hardware or tolerated by the applications efficiently** (hardware correction will generate noise, global restart)
- Main issues at Exascale:
  - #errors, #failures, Scale, Power consumption, Performance impact
- Recommendations for EU:
  - 1) Establish a fault-error-failure model,
  - 2) Develop automatic root cause discovery and failure prediction,
  - 3) **Extend the applicability of checkpoint-restart** (API, Partial RST, NV mem.),
  - 4) Need a inter software layer communication system for consistent fault-error-failure management (detection, notification, decision, etc.),
  - 5) **New fault tolerance paradigms** using non-volatile memory technologies
  - 6) **International Coordination**:
    - Common FT API (interface and semantic) and inter layer communication system for portability (to limit applications rewriting)
    - G8-ECS (France, Germany, USA, Japan, Canada), ARN-JST FP3C (France and Japan), INRIA-Illinois Joint-laboratory (France and USA).

WG4.2 Testbed needs in 2015 (draft)

	< 100 Petaflops enough	100 Petaflops needed
Dedicated/reconfigurable	-**Prog. Models and Runtime** (interaction of runtime with kernel scheduler) -**Performance tools**: (interaction with resilience / power components) -**OS** (scheduling, memory management) -Measure noise generation (most of the syst fotware)? -**I/O-File system** (may be done at lower scale) (need root access for reconfiguration) -**Job and resource manager** (need root access for reconfiguration)	-**Performance tools** (system level): measure and minimize overheads induced at the full scale by low-level performance monitoring infrastructure tightly integrated with OS -**Prog. Models and Runtime** (interaction of runtime with kernel scheduler)? -…
Production	-**Compilers** (node level) -**Programming models** (node level API) -**Performance tools** (node level) -**Power management**? -**Validation and correctness checking**?	-**Resilience** (FT protocols, ABFT, NFTA, execution state storage) -**Parallel debuggers**? (scalability test) -**Performance tools** (scalability of data collection and analysis) -**Runtime** (scalability test) -**Prog. models** (system level scalability) -**Performance modeling at scale**

WG4.2 Contributions to an Exa Stack (draft,

- Programming models
  - **StarSs**: Node level programming model offering programmers a natural interface to an underlying data-flow execution model and aiming allowing the "same" source code to run on "any" target architecture. In particular, **OmpSs** implementation that integrates OpenMP and the StarSs concept (SMP, GPUs,…) and nicely integrates with MPI. http://www.bsc.es/smpsuperscalar and http://www.bsc.es/OmpSs, BSC
  - **HMPP**: Based on a set of directives, programming model designed to handle hardware accelerators, www.caps-entreprise.com/hmpp.html CAPS
  - **OpenHMPP** will be announced at SC11. Members of the consortium will be major HPC users, compiler companies such as Pathscale, INRIA
  - **T-Platforms Exascale API environments**: a new scalable version of MPI and a PGAS library with a new set of compilers and tools, Russia
  - European PGAS development called GPI (from Fraunhofer), www.gpi-site.org
- Runtime
  - **StarPU runtime system**, StarPU is a task scheduler exploiting multi-GPU multi-core platform efficiently, using adaptive performance models to carefully dispatch tasks over the available heterogeneous processing units. http://runtime.bordeaux.inria.fr/StarPU/
- Parallel Debugger
  - **DDT**: scalable parallel debugger for debugging MPI and multi-threaded codes, Alinea
- Validation/correctness
  - **MUST**: scalable Runtime Error Detection in MPI Programms, TU Dresden

WG4.2 Contributions to an Exa Stack (draft,

- Performance tools
  - **Scalasca**: Performance-analysis tool German Research School for Simulation Sciences and Juelich Supercomputing Centre
  - **Score-P**: Generic measurement infrastructure for several performance tools including Periscope, Scalasca, TAU, Vampir, Virtual Institute - High Productivity Supercomputing http://www.vi-hps.org
  - **BSC Integrated tools environment**: instrumentation (**Extrae**), visualization (**Paraver**), Simulation (**Dimemas**) and other analysis modules such as clustering, combined sampling and instrumentation, **Par** http://www.bsc.es , BSC
  - Persicope (Technical University of Munich) http://www.lrr.in.tum.de/~petkovve/psc/
  - ThreadSpotter (RogueWave) http://www.roguewave.com/products/threadspotter.aspx
  - Vampir (Technical University of Dresden) http://www.vampir.eu/

- OS and system management
  - **T-Platforms Clusterx** Exascale OS and system management: a highly scalable exascale-oriented OS and system software stack (highly scalable real-time monitoring and management system), Russia
  - **Clustershell**: start commands on the cluster nodes (in parallel, with a high scalability based on tree) and to get a usable output (summary, differences, ...) for human *sourceforge.net/projects/clustershell/, CEA*
  - **Ganesha:** "protect" the computing centre NFS servers from the load of the NFS clients sourceforge.net/projects/nfs-ganesha/, CEA
  - **Robinhood** : a FS monitor/space manager optimized for huge FS (PB today, EB in the future) sourceforge.net/projects/robinhood, CEA
  - ***INRIA OAR,*** *res. Job manager* ***, Kadeploy, f****ast and scalable deployment system****, Taktut:*** *parallel and scalable remote execution tool for cluster. http://mescal.imag.fr/software.php*

WG4.2 Roadmap to be completed by June

- Description of the scientific and technical perimeter
- Social benefits, societal, environmental and economical impact
- Scientific and technical hurdles
- Address cross cutting issues : Resilience, Power Mngt, Programmability, Perf optimization and Reproducibility of the results,
- European strengths and weaknesses in the worldwide competition
- Sources of competitiveness for Europe
- Needs of education and training
- Potential collaborations outside Europe
- Existing funded projects and funding agencies
- **Timeline, needs of HR, provisional costs,**
- Needs for experimental platform: size, reconfigurable?, dedicated?
- What software in an Exascale stack and with what level of responsibility