From TERA to EXA 1
Supercomputers for extreme scale simulation

Didier JUVIN
HPC & Simulation Program Manager

OCTOBER 2017
HPC is a key capacity for the CEA missions

- Nuclear Energy
- Defense Simulation Program
- Medical Imagery
- Astrophysics
- Climate modeling
- Personalized Medicine
- Predictive Material
Three complementary coordinated pillars, in symbiosis with the French and European policy.

**Technologies**

Develop a globally competitive technology value chain.

**Applications**

Expand the limits in modelisation and complex multiphysics simulation.

- **Research**
  - Maison de la Simulation

- **Industry**
  - Ter@tec

- **Defense**
  - Defense Simulation Program

**Infrastructures**

- **Research**
  - Curie

- **Industry**
  - Cobalt X3

- **Defense**
  - CCRT
  - TERA1000: Xeon and Xeon PHI configurations.
HPC is a major stake for the competitiveness of companies

Mastering and exploitation of HPC technology will induce **2 to 3% annual GDP increase in France**

Mastering HPC technologies is vital for data-intensive critical application: a national sovereignty stake

In 2014, President François Hollande decided to launch an action plan to master HPC technology: The Supercomputer National Industrial Plan

Objectives of the plan:
- To acquire exascale computing capability and develop the necessary HPC ecosystem all the way from hardware and system software to application
- To give France the industrial capability required to design and build exaflop-class computers in a sustainably competitive way by 2020

CEA in charge of the organization of the exaflops R&D program for 2021 to push HPC French industry in the Top 3 leader worldwide
R&D program : a 2 step process

Technological challenges :

• High energy efficient general architecture
• Interconnect a huge number of nodes and memory with low latency
• Mastering the data at exascale
• To waranty a high level of resilience & availability in operation
• To develop new programming and execution models able to exploit massive parallelism

Solutions developed in a sustainable and competitive objective (cost of acquisition, exploitation, ..)

Phase 1 (2013 – 2015) : Design of a pre-exascale industrial system (up to 100 Pflops)

Fully achieved ➔ Bullion, Sequana

Phase 2 (2016 – 2020) : Design of a exascale industrial system for 2020
Key Performance Indicators

Indicator #1: Number of SC at the TOP 500

Indicator #2: Energy Efficiency

Indicator #3: Market attractiveness

Indicator #4: Maturation of technologies

Target:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gflops/W</td>
<td>2</td>
<td>5</td>
<td>6-7</td>
<td>6-7</td>
<td>14</td>
<td>14</td>
<td>28</td>
</tr>
</tbody>
</table>

Increase in turnover
International market penetration
Increase in the R&D self-financing ratio

Achievement of the technical milestones on time
R&D phase 1 successfully completed
SEQUANA : a new generation of supercomputers

Of the world’s best technologies

- **Platform**: Sequana (~1 Pflops / system – 288 nodes), DLC
- **Integration density**: 3 KNL nodes / board
- **A high efficient interconnect**: Bull eXascale Interconnect
- **Energy efficient**: > 6 GFlops/W
- **Architecture**: efficient « huge data » management
- **Pre-exascale software stack**

SEQUANA, a pre exascale system (up to 100 Pflops)

Tera1000 is the serial #1 of the R&D phase 1 program results
<table>
<thead>
<tr>
<th><strong># of Compute Nodes</strong></th>
<th><strong>8000</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Memory per node</strong></td>
<td><strong>128 GByte</strong></td>
</tr>
<tr>
<td><strong>Fast Memory size per node</strong></td>
<td><strong>16 GByte</strong></td>
</tr>
<tr>
<td><strong>Interconnect</strong></td>
<td><strong>BXI</strong></td>
</tr>
<tr>
<td><strong># proxy I/O nodes</strong></td>
<td><strong>150</strong></td>
</tr>
<tr>
<td><strong>Flash storage</strong></td>
<td><strong>4 Pbyte, 2 TB/s</strong></td>
</tr>
<tr>
<td><strong>Electrical consumption</strong></td>
<td><strong>4 MW</strong></td>
</tr>
</tbody>
</table>
Key Performance Indicators : Phase 1 status

Indicator #1 :
20 Bull supercomputer are ranking in the TOP 500
# 2 worldwide for the number of systems > 1 Pflops in operation

Indicator #2 :
SEQUANA/KNL equipped with BXI interconnect
achieves a performance factor of 6.5 Gflops/W

Indicator #3 :
Exascale program has a very positive impact on
ATOS sales and image

Indicator #4 :
Market attractiveness

Every milestones defined in the convention has been reached
on time
Impact of the R&D program: Market attractiveness

**Bullion**: 8 to 16 CPU, up to 24 TB of memory

**Bull Sequana X1000**
Already deployed in the field

**SERVEURS BULLION**
Bullion: 8 to 16 CPU, up to 24 TB of memory

**DELL EMC distributes Atos Bullion**

- Announcement made on Aug 22, 2017
Technological context

- **Evolution in the microelectronics domain**
  - Moore’s Law slows down
  - Power density limitation
  ➔ A theoretical peak computing power less and less usable in practice
- **Evolution of the positioning of techno providers**
  - INTEL expands its positioning in the value chain
  ➔ Alternative track to be evaluated (ARM, AMD, ...)

Methodological evolution

- **Simulation codes take more and more profit of AI based algorithms**
  - Uncertainty quantification
  - Data flow analysis
    ▪ Post-processing,
    ▪ Supercomputer behavior analysis (meta-data)
  - Complex system modelling
  ➔ Convergence HPC and Data Analytics (HPDA)
The 2\textsuperscript{nd} phase of the R&D SC National Plan

Phase 2 is designed to:

- Increase the national added value \textit{(objective: top 3 worldwide in 2021)}
- answer the technological challenges, taking into account the new context
- Exploit the achievements of the 1\textsuperscript{st} Phase

Challenges:

- **Complexity**: management of Millions of cores
- Maximize energy efficiency (target 28 Gflops/W)
- Design the low latency interconnect network
- Manage the huge data flow of I/O (Big Data)
- Mastering the resiliency

Design a HPC energy efficient node

Build a exaflopic class demonstrator for HPC & HPDA
Main topics of the 2\textsuperscript{nd} phase

1. Architecture of computing resources
2. Network interconnect
3. Data management
4. Software for resource management
5. Programming models and runtime
Challenge:
Design a flexible platform and architecture able to reach 1 Exaflops
⇒ Sequana 2 : a flexible architecture designed for the exascale

First results:
- Prototypes in final productization
  - Intel Xeon Skylake
  - Nvidia Pascal (GPU)
- Design of an ARM blade prototype
  - Joint effort with the Mont Blanc 3 project

Next steps:
- Increase node computing density
  - at SoC level
  - at system level (W/rack)
- Participate / Follow European Initiative (post Exanode, MontBlanc 2020,...)
2/ Interconnect

**Challenge:**
Dealing with millions of computing cores will put a high pressure on Interconnect

⇒ **Bull Extreme Interconnect (BXI):** 2 generations

- BXI V1 focused on Intel (Broadwell, Skylake & KNL)

**First results:**
- Linpack benchmark validated on 220 KNL nodes (entered in Nov. 2016 TOP 500)
- ASICs BXI V1 for Sequana available for production (TERA1000-2, …)
- Development of the BXI software stack: integrated into Bullx SCS
- Start of the Open BXI initiative

**Next steps**

- BXI V2 add new features
3/ Data management

**Challenge:**
Manage the huge flow of data produced by simulation codes
⇒ New storage architecture

**First Results:**
- Validation of a prototype of I/O delegation on Sequana
- I/O proxy to lighten the load on compute nodes

**Next steps**
Object storage ⇒ to go beyond classical parallel file systems
**Challenge:**
New approaches needed to administrate exascale systems
Manage in a highly secure manner a very large number of node (scaling of tools)
While being open to new usages (data analytics)

**First Results:**
Definition of a hierarchical highly secured solution for exascale systems to
- Install, Configure, Update, Monitor
Validation of a secured architecture for Sequana implementing secured access
- Study to implement end-to-end secure data
Study virtualization solutions in a production environment
- Towards a unified solution for HPC and Big Data (HPDA)

**Next steps**
Validate these new concepts for HPDA production at scale
5/ Programming models and runtime

Goal: provide the application developers a programming environment that

- Mask the complexity of large systems
- Help the developer to optimize the codes (legacy and new codes)

First Results:
- Study and selection of a tool set
  - Code analyser, profiling, debugging:
    - Definition of a development environment suitable for exascale programming
  - Runtime (MPC)
- Development of abstraction layers that hide complexity and heterogeneity of exascale platforms.

Next steps
- Increase participation to standard bodies
  - MPI
  - OpenMP
  - ...
- Develop/Optimize tools, Libraries & Run times
Integration and Validation at scale

OPEN BXI

Network interconnect

Co-design works

Platform Architecture

Programming models & Runtime

Data Management

Ressource Management

PCOCC

lustre File System

hadoop

Ganesha

MPC OpenMP

OpenMP

CentOS

Portals
Co-design is essential for exascale

- ATOS offers innovative integrated solutions, provides the industrial expertise
- CEA provides major contributions, the user expertise & application feedback
- A win/win collaboration

Hardware and Software are two sides of the same challenge

- The key to success: work closely together
  
  Within the company (hardware & software teams)
  In a short loop with end users (co-design)

ATOS & CEA are on the path to exascale