DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

HPC Status and Roadmap

Focus on R&D at CEA

Patrick Carribault

HPC Fellow

patrick.carribault@cea.fr

CEA, DAM, DIF, F-91297 Arpajon

OUTLINE

- HPC Status
  - Supercomputing Ranking
  - HPC History
  - HPC Roadmap
  - French/Europe Focus
  - Main Challenges
- R&D on HPC Software Stack
  - Open-source HPC development at CEA
  - Focus on parallel runtime: MPC
- Beyond Exascale
- Conclusion

HPC STATUS

HISTORY, CURRENT MACHINES AND FUTURE

PARALLEL COMPUTING

- Tentative definition

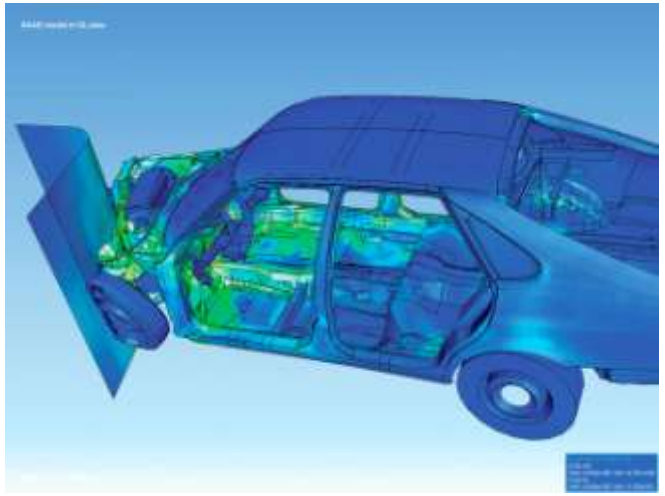
Ability to exploit multiple compute units at the same time to solve a problem

- Involve various domains

Car industry
Chemistry
Bio-informatics
Energy

- Related to different « problems »

Car security
Molecule interaction/reaction
Plane behavior w/ bad weather
Weather forecast
...



PARALLEL COMPUTING

- Why not more experimenting?

- Some (non-exhaustive) reasons
  - Problems too large
  - Too complex
  - Too expansive
  - Not possible to experiment

- Solution: rely on computational power

- From simple servers to supercomputers
  - Clusters of regular servers
  - Different approaches depending on application domains

- High-Performance Computing + subdomains
  - Big data
  - Data analytics
  - HPDA

- How to classify such machines?

SUPERCOMPUTER CLASSIFICATION

- Small application to compare machines
  - Benchmark or *miniapp* or *proxyapp*
  - Results ➔ metrics able to compare machines

- Example: **Top500**
  - Rank machines according to the computational power on regular codes
  - Homepage: http://www.top500.org

- Benchmark: `Linpack`
  - Linear solver based on linear algebra
  - Relies on performance of DGEMM
  - Towards HPCG (Conjugate Gradient)

TOP500

- List of 500 most powerful machines
  - Measure mainly the computational power
  - According to Linpack results

- Updated twice a year
  - June/July: ISC conference in Germany
  - November: SC conference in US

- Machine Information
  - Main info (rank, site)
  - System (name and short description)
  - Number of cores
  - Performance (`Rmax`, `Rpeak`)
  - Power

- Notes
  - Performance in Tflops/s (10^{12} floating-point operations per second)
  - Difference between max performance (Rmax) and Linpack result (Rpeak)
  - Power measured in kW

TOP500 JUNE 2022 (#1 TO #5)

Rank	Country	System	Cores	Rmax	Rpeak	Power
1	United States	Frontier	8,730,112	1,102.00	1,685.65	21,100
2	Japan	Fugaku	7,630,848	442,010.0	537,212.0	29,899
3	Finland	Lumi	1,110,144	151.90	214.35	2,942
4	United States	Summit	2,414,592	148,600.0	200,794.9	10,096
5	United States	Sierra	1,572,480	94,640.0	125,712.0	7,438

Source: RIKEN

TOP500 ANALYSIS

- First comments

- Exaflop/s!!!
  - Exaflop/s = 10^{18} floating-point operations per second
- Machines with lot of *cores*
  - Several millions!
- Power consumption up to almost 30 Mwatts
- Top 10 exhibits different system architectures (homogeneous & heterogeneous)

- Deeper analysis

- Big difference between *Rmax* and *Rpeak*
- Big difference between *Rmax* and *Power*

- Ordering based on power efficiency: **Green500**

- Sort supercomputers according to the ratio power consumption / `Linpack` performance

GREEN500 JUNE 2022

R	Top500	System	Cores	Rmax	Power (kW)	Gflops/W
1	29	Frontier-TDS	120,832	19.20	309	62.684
2	1	Frontier	8,730,112	1,102.00	21,100	52.227
3	3	LUMI	1,110,144	151.90	2,942	51.629
4	10	Adastra	319,072	46.10	921	50.028
5	326	MN-3	1,664	2.18	53	40.901

Source: HPCWire

GREEN500 ANALYSIS

- Main ordering

- First machine is not the most *powerful*
  - Rank 29 in Top500 → subpart of Top500 supercomputer
- Small machines may be power efficient
- Rely on accelerators & specific architectures (mainly derived from GPU)

- Top500 and Green500 limits

- `Linpack` is a very specific benchmark
- Regular computation (mainly linear algebra)
- Few communications/synchronization between parallel units

- Need different benchmarks to classify supercomputers

- Most powerful machines on irregular codes: Graph500
- Based on graph traversal
- GTEPS: Billions of edges traversed per second
  - → Fugaku (and subpart of this machine) is top-ranked
  - → Same for LUMI

CONCLUSION ON RANKING

- Multiple ranking methods
  - Correspond to various needs
  - Highlight different architectures
- Where do the differences come from?
  - Various domains of applications
  - Depends on the target users
  - Impact on the design choices
  - Difference machine architectures
  - Processors, memory, network…
- How did we end up with such current lists?
  - A little bit of HPC/architecture history…

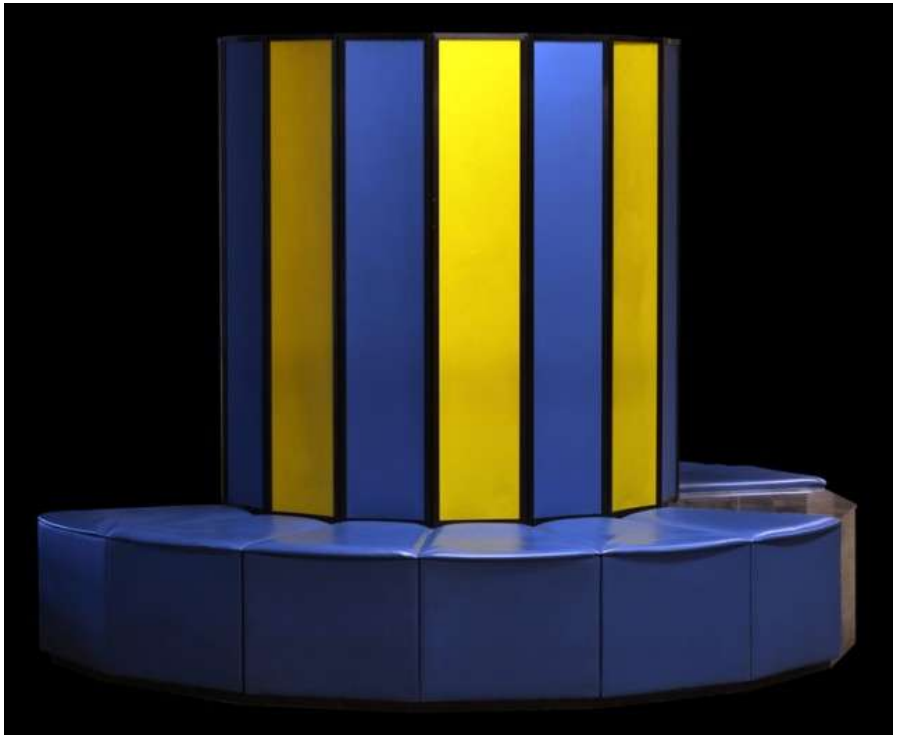
HPC HISTORY

- Cray 1
  - Built in 1976
  - Designed by
    - Seymour Cray
  - Cost: $5 - $8 million
  - Frequency: 80 MHz
  - Freon cooling

- Performance
  - **136 Mflops**

HPC HISTORY

- Cray XMP
  - Built in 1982
  - Up to 4 CPUs
  - Frequency: 105Mhz
  - Cost: $15 million
- Performance
  - 200 Mflops per CPU
  - **800 Mflops total!**



source: Extreme Tech

HPC HISTORY

- ASCI Red
  - Build in 1997
  - 6,000 CPUs
  - Intel Pentium Pros
    - Regular processors
  - Frequency: 200Mhz
  - Cost: $46 million

- Performance
  - **> 1 Tflops**
  - First Teraflop machine!

source: Extreme Tech

HPC HISTORY

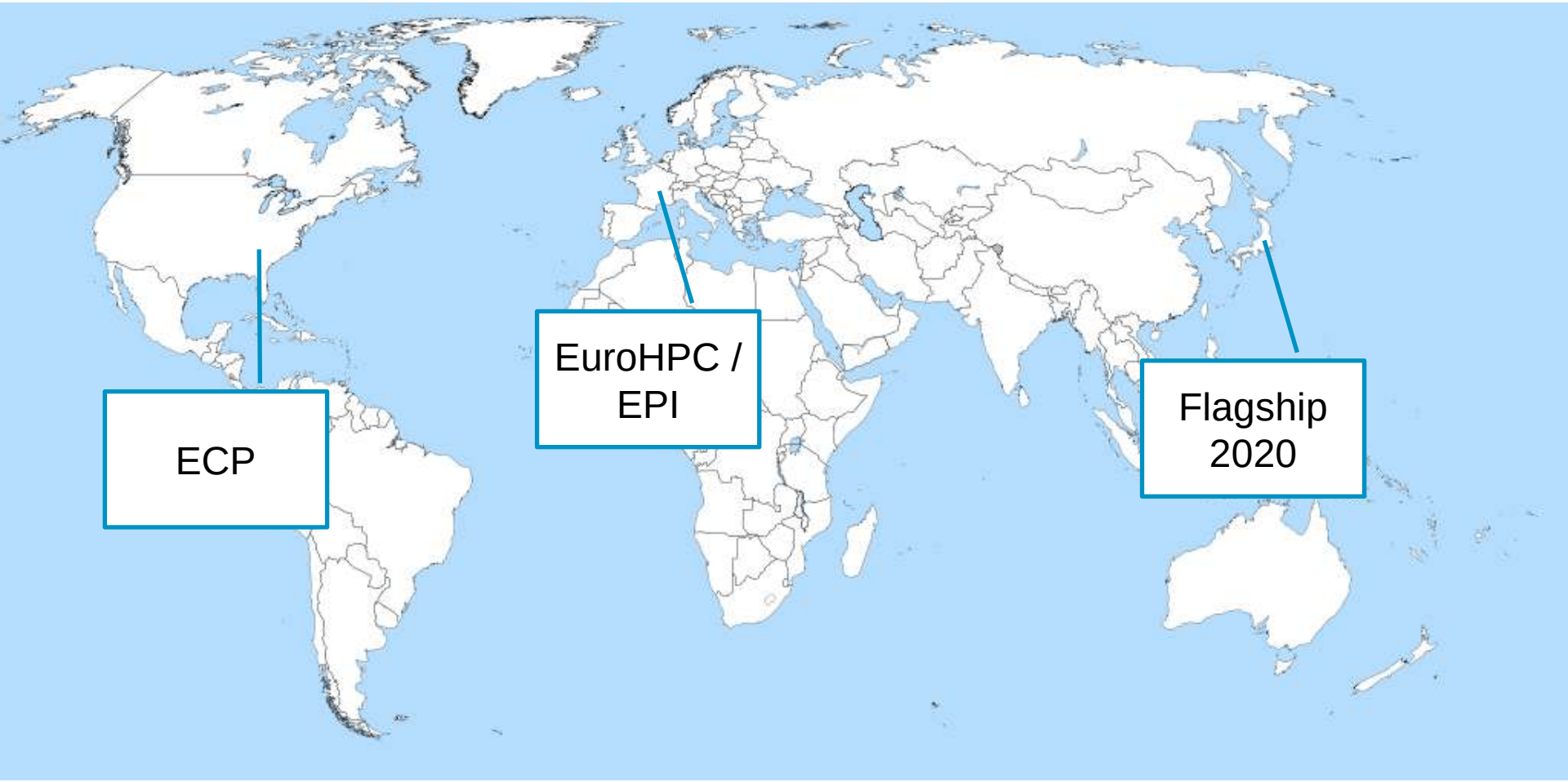
- IBM Roadrunner
  - Built in 2008
  - Hybrid
    - AMD Opteron
    - IBM PowerPC
  - Frequency:
    - 1.8GHz & 3.2GHz
  - Cost: $100 million

- Performance
  - **> 1 Pflops**
  - First Petaflop machine!

source: Wikipedia

HPC ROADMAP

- Exascale milstone reached in 2022 (ability to each 10^{18} FLOPS peak)
- Many hardware & software challenges to meet this deadline (➔ lot of R&D)



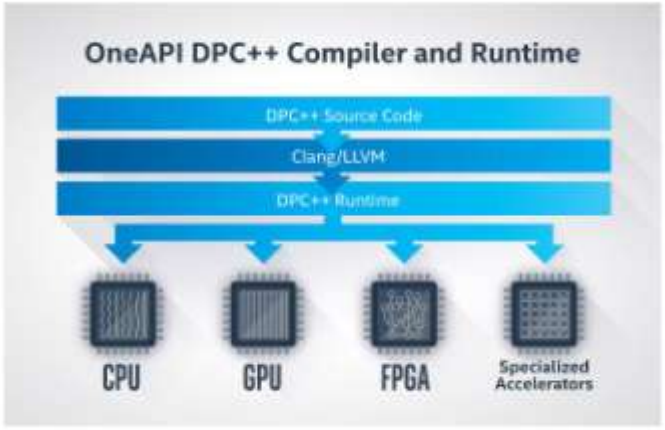
EXASCALE COMPUTING PROJECT (ECP)

- **Goals**
  - Maximizing the benefits from HPC for the US
  - Accelerating development of capable exascale computing ecosystem
  - 7-year project through 2023
- **Collaborative effort of DoE organizations**
  - Office of Science (DOE-SC)
  - National Nuclear Security Administration (NNSA)
  - Include US industry and Universities
- **Focus Areas**
  - Application development
  - Exascale systems
  - Hardware technology
  - Software technology

DOE EXASCALE SUPERCOMPUTERS

- Aurora

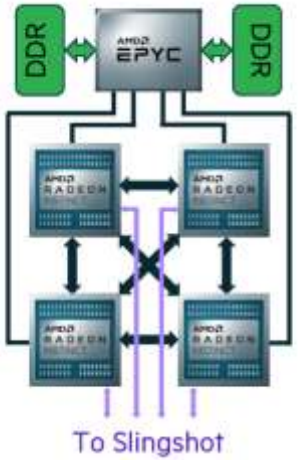
- Due in 2021 in Argonne National Lab
- Sustained performance > 1 EF DP
- Based on heterogeneous architecture
  - 2 Intel Xeon scalable processors (Sapphire Rapids),
  - 6 Xe arch-based GP-GPUs (Ponte Vecchio);
    - Execution Units (EU) into SubSlices (SS) and into Slices
  - Unified memory architecture across CPU & GPU
- Network: Cray Slingshot
- https://alcf.anl.gov/aurora



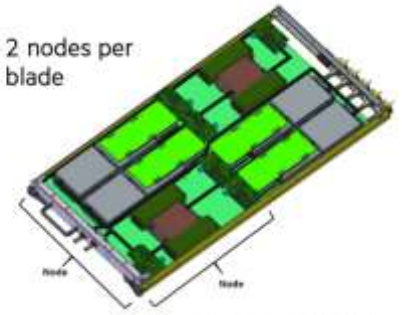
DOE EXASCALE SUPERCOMPUTERS

- **Frontier**

- Due in 2021 in Oak Ridge National Lab
- Peak performance > 1.5 EF
- Compute node
  - 1 HPC and AI Optimized AMD EPYC CPU
  - 4 Purpose Built AMD Radeon Instinct GPU
  - AMD Infinity Fabric Coherent memory across the node
- Network: Cray Slingshot
- https://www.olcf.ornl.gov/frontier/



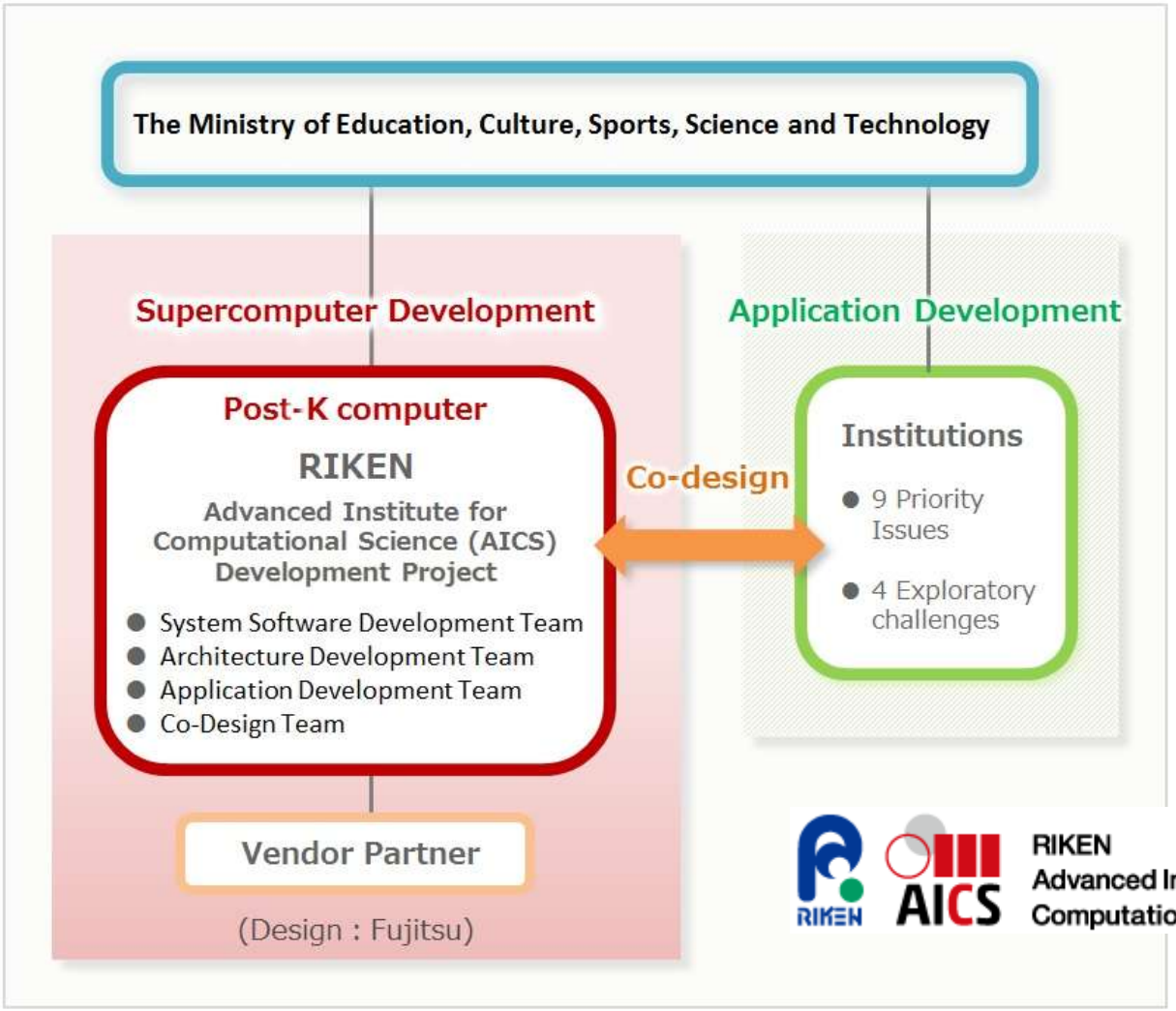
AMD GPU (ORNL)



- **El Capitan**

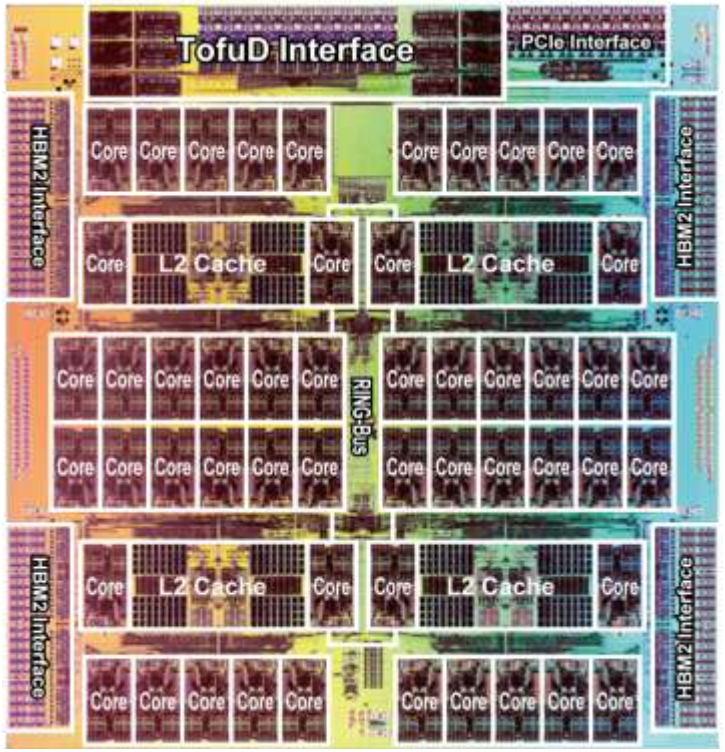
- Due in 2023 in Lawrence Livermore National Lab
- Peak performance > 2 EF
- Based on AMD Genoa (Zen 4) CPUs and Radeon

FLAGSHIP 2020 PROJECT (JAPAN)



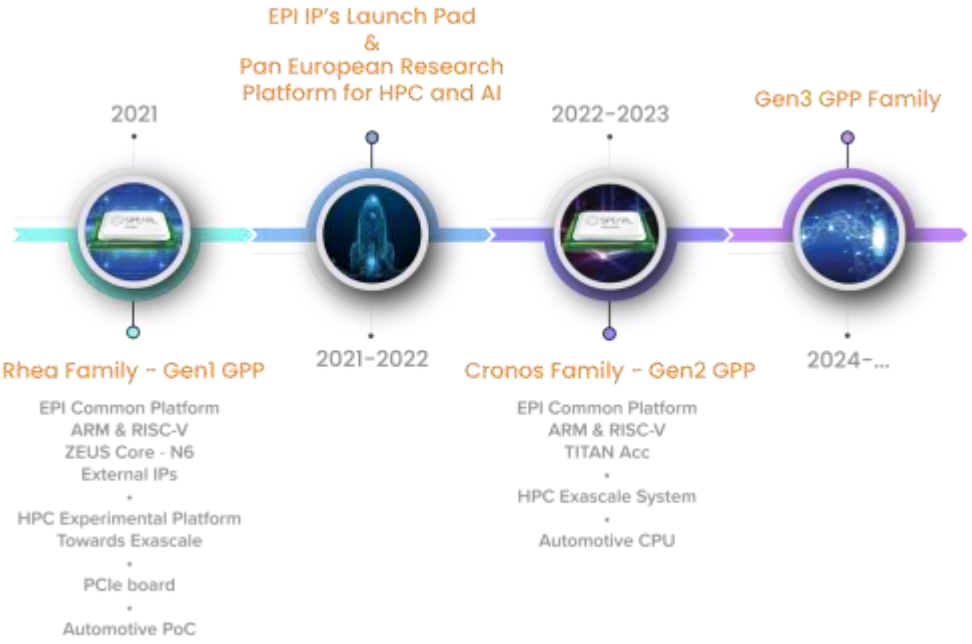
FUGAKU SUPERCOMPUTER

- Performance
  - Peak performance (2.0 GHz): 488 Petaflops
  - Peak performance (2.2 GHz): 537 Petaflops
- Architecture
  - 158,976 nodes
  - Interconnect: Tofu D
- Compute node
  - Armv8.2-A SVE 512bit
  - 48 cores for compute
  - 2 or 4 cores for OS activities
  - Memory: HBM2 32 GiB, 1024 GB/s



EUROPEAN PROCESSOR INITIATIVE (EPI)

- Goals
  - European independence in High Performance Computing Processor Technologies
  - EU Exascale machine based on EU processor by 2023
- Timeline



FRENCH STATUS

- Top500 June 2022: 5th country w/ 22 systems
  - 4.4% of systems
  - 3.8% of global performance

Rank	Site	System	Cores	Rmax	Rpeak	Power
10	CINES	Adastra	319,072	46.10	61.61	921
17	CEA	CEA-HF	810,240	23.24	31.76	4,959
33	Total	PANGEA III	291,024	17.86	25.02	1,367
45	CEA	Tera-1000-2	561,408	11.97	23.40	3,178
63	Meteo France	Taranis	294,912	8.19	10.32	1,672

FRENCH ECOSYSTEM

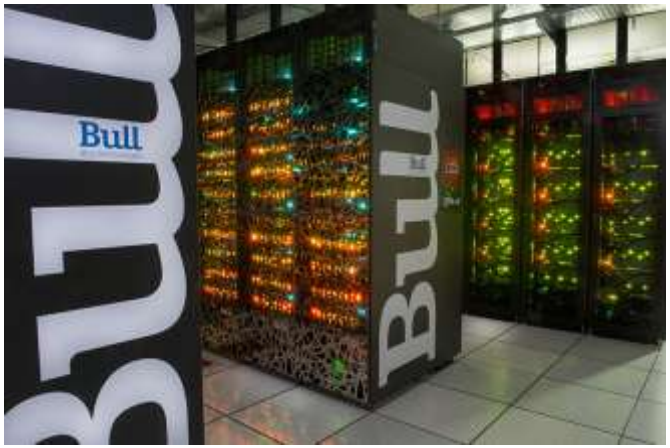
- Teratec
  - European pole of competence in high performance simulation
  - Technology, research, dissemination
  - Teaching & training

- Campus
  - Group multiple companies & research labs
  - Located in Bruyères-le-Châtel (on CEA campus)
  - Exascale Computing Research (Intel/CEA/UVSQ)
  - InHP@CT seminars
    - http://inhpact.hpcframework.paratools.com/

- Forum organized each year
  - Example: June 22-24, 2021 (virtual event)
  - Usually organized at Ecole Polytechnique
  - Presentations & Exhibition

FRENCH ECOSYSTEM

- Main French Vendor:
  - Bull Atos
- Inside Top500
  - 4th vendors
  - 42 systems (8.4%)
  - 5.8% of global performance
- Co-design with CEA



FRENCH VISION: BULL & CEA

- Co-design between Atos Bull & CEA
- Multiple machines inside Top500 made by BULL and hosted by CEA
- HPC at CEA
  - Mainly CEA/DAM (Bruyères-le-Châtel)
  - Different product lines

TERA

- Part of defense simulation program

- History

- Program started in 1996
- Predicted to set up 3 machines

- First machine: Tera 1 (HP/COMPAQ)

- 2,560 cores (Alpha CPU, 1 GHz)
- Quadrics interconnect
- Linpack performance: 3.18 Tflop/s
- Rank 4 in June 2002

- Second machine: Tera 10 (BULL)

- 8,704 cores (Intel Itanium 2, 1.6GHz)
- Quadrics interconnect
- Linpack performance: 42.9 Tflop/s
- Rank 5 in June 2006

TERA

- Third machine: Tera 100 (Bull)
  - 140,000 cores (Intel Xeon Nehalem)
  - 4,300 compute nodes
  - IB QDR interconnect
  - Linpack performance: 1,050 TFlop/s
  - Rank 6 in November 2010

CURRENT TERA MACHINES

- Tera 1000-1
  - 70,172 cores (Intel Xeon Haswell)
  - IB FDR interconnect
  - Linpack performance: 1,871 Tflop/s
  - Rank 44 in June 2016

- Tera 1000-2
  - 561,408 cores (Intel Xeon Phi KNL)
  - Bull BXI interconnect
  - Linpack performance: 11,965.5 Tflop/s
  - Rank 14 in June 2018

- EXA 1
  - Supercomputer CEA-HF (17 @ Top500)
  - R&D for CEA-HE

CCRT

- Research and Technology Computing Center
  - *Centre de calcul pour la recherche et la technologie*

- French consortium
  - Started in 2003
  - Based on French academic & industry

- Goals
  - Provide High Performance Computing resources for large scientific computations
  - Foster a real synergy between research organizations, universities and industry
  - Promote exchanges and scientific collaboration between partners.

CCRT MACHINES

- Cobalt (Atos)

- Total: 39,816 compute cores (Intel Xeon Broadwell)
- Node w/ dual-socket (28 cores per node)
- IB EDR interconnect
- Rank 63 in June 2016
- 1.299 Pflops

- Topaze

- Announced in 2021 (not in Top500 yet)
- Open to Grand Challenges in June
- Based on AMD Milan processors
- Additional partition w/ NVIDIA A100 GPUs

PRACE

- Partnership for Advanced Computing in Europe

European Consortium

- 25 member countries
- 5 PRACE centers

BSC (Spain)
CINECA (Italy)
CSCS (Switzerland)
GCS (Germany)
GENCI (France)

- Currently

French machine Joliot Curie (TGCC, Bruyères-le-Chatel)

- Intel Skylake
- Intel KNL
- AMD Rome
- NVIDIA V100
- Fujitsu FX700

HARDWARE CHALLENGES

- Hardware evolution
  - Processors are building blocks of clusters
  - But one processor = cores + complex mechanisms
  - Clusters are made of many other components that are crucial for overall performance

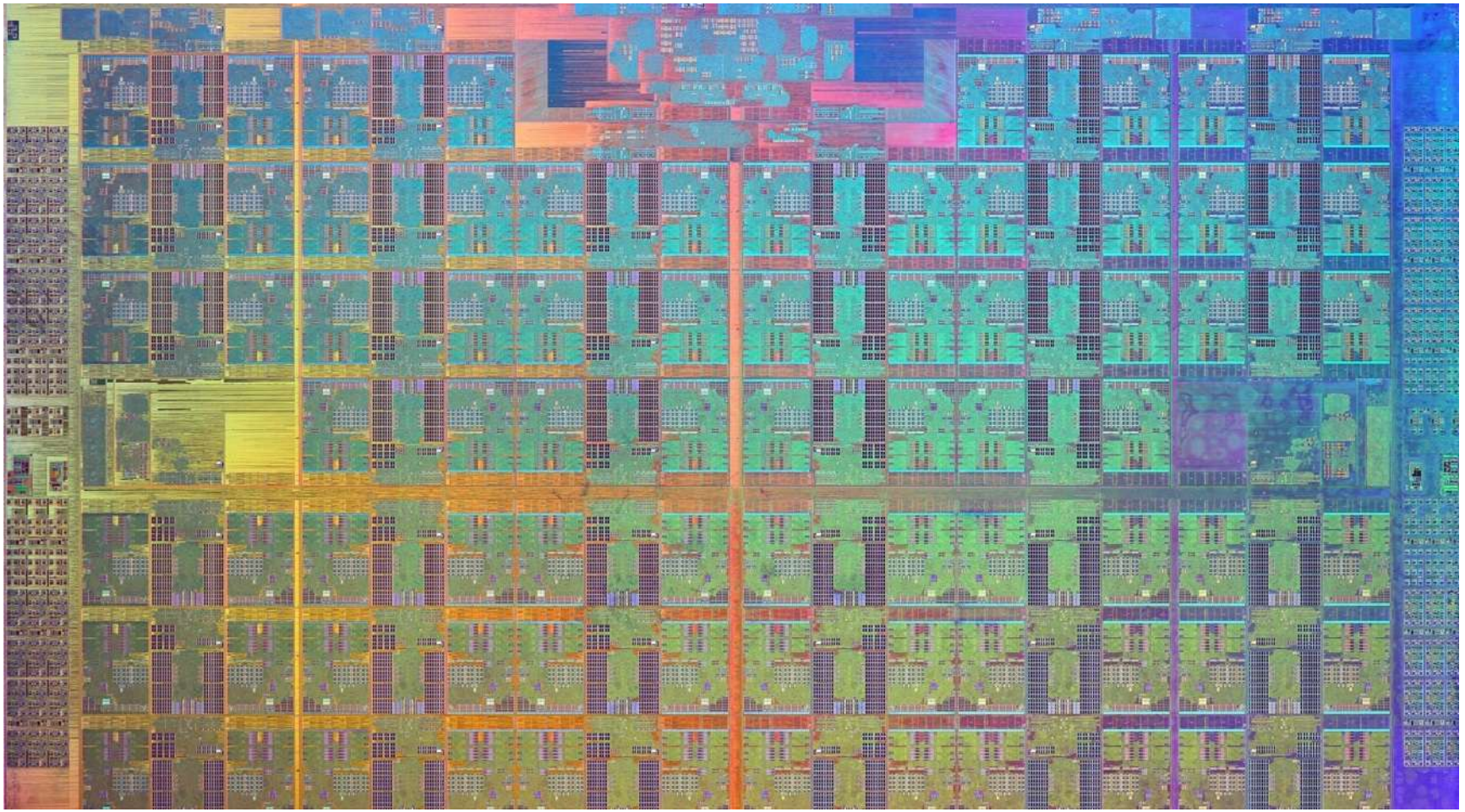
- List of major components
  - Processors
  - Memory
  - Network
  - Mother boards & nodes
  - …

- What are the challenges related to these components?

PROCESSOR CHALLENGES

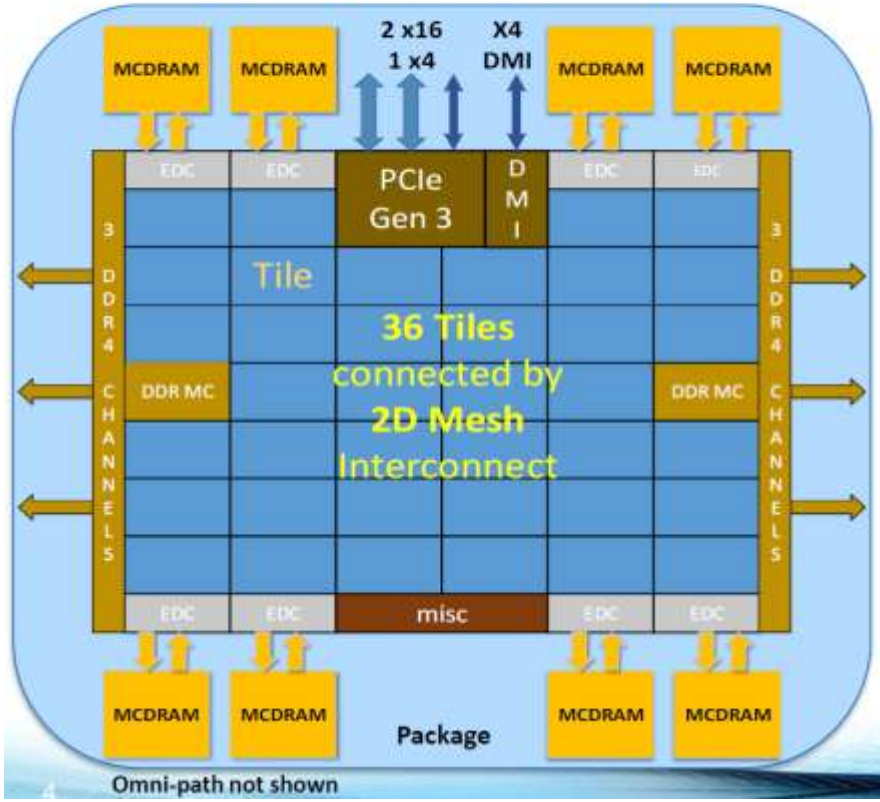
- Main trends
  - Increase number of cores
  - Larger compute units
  - General purpose or dedicated
- Increase in the number of cores
  - Per processor
  - Per nodes
- Evolution of compute units
  - Less microarchitectural mechanisms
  - Larger vector units
- General purpose or dedicated
  - Regular Intel Xeon multicore processors → Intel Haswell example
  - Intel Xeon Phi processors → Intel KNL example
  - GPGPU (NVIDIA, AMD, Intel) → NVIDIA V100, A100 and AMD MI100 examples

INTEL XEON PHI



INTEL XEON PHI

Knights Landing Overview



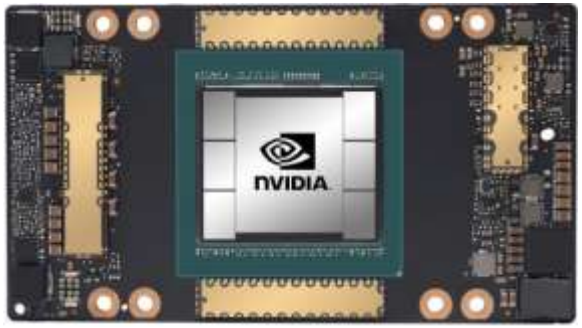
Chip: 36 Tiles interconnected by **2D Mesh**
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384GB
IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset
Node: 1-Socket only
Fabric: Omni-Path on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops
Scalar Perf: ~3x over Knights Corner
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when [illegible] used in the flat memory. Results have been estimated based on internal Intel analysis and [illegible] purposes only. Any difference in system [illegible] performance.

NVIDIA AMPERE ARCHITECTURE



NVIDIA AMPERE ARCHITECTURE

- Various compute units
  - Scalar/vector
  - Matrix
- FP64 performance
  - Less FP64 vector unit
  - 1 tensor core per warp scheduler
- Peak performance
  - Through matrix FMAs
  - Vector operations → half the peak performance

MEMORY SUBSYSTEM CHALLENGES

- Extended memory levels
- Evolution of caches
  - Still some private caches
  - May include scratchpad
  - Shared caches ➔ mesh-based coherency
- New memory levels
  - High-Bandwidth Memory (HBM)
  - Non-volatile memory (NVM)

NUMBER OF NODES

- Main trend

- Include challenges from processors and memory
- Increase in number of nodes

- Impacts

- Put the stress on network card (NIC)
  - Need to handle communication with more neighbors
- Imply new design for switches
  - Need to organize the network in specific topology (e.g., fat tree)

OUTLINE

- HPC Status
  - Supercomputing Ranking
  - HPC History
  - HPC Roadmap
  - French/Europe Focus
  - Main Challenges

- **R&D on HPC Software Stack**
  - **Open-source HPC development at CEA**
  - **Focus on parallel runtime: MPC**

- Beyond Exascale

- Conclusion

R&D ON HPC SOFTWARE STACK

R&D ON HPC SOFTWARE STACK

- HPC @ CEA/DAM

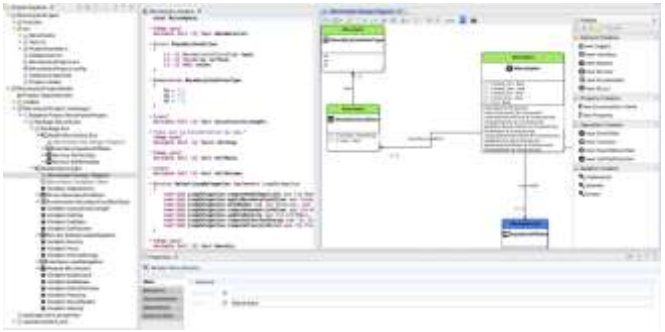
- R&D on hardware side
  - Co-design w/ major vendors: Atos/Bull, Intel, …
- R&D on software side
  - Open-source development
  - https://github.com/cea-hpc
  - https://hpcframework.com/
- Main description of HPC @ CEA/DAM
  - http://www-hpc.cea.fr/index-en.htm

- Sysadmin an low-level software

- **modules**
  - Environment Modules: provides dynamic modification of a user's environment
- **selFie** (Self and Light proFIling Engine)
  - Very light profiling tools for Linux commands and HPC codes
- **pcocc**
  - Run VMs on an HPC cluster

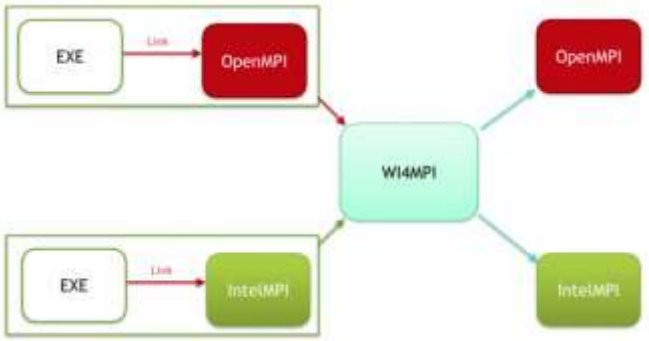
R&D ON HPC SOFTWARE STACK

- Parallel programming software (subset)



- **Nablab / Modane**
  - Full-fledged industrial environment for scientific computing and High Performance Computing

- **PCVS** (Parallel Computing -- Validation Suite)
  - Validation engine for Exascale project benchmarks



- **WI4MPI**
  - Translation framework between MPI constants and MPI objects from an MPI implementation to another one.

- **MPC** (Multi-Processor Computing)

MPC
Multi - Processor Computing

Context

- Multi-Processor Computing (MPC) framework

- Runtime system and software stack for HPC
- Project started in 2003 at CEA/DAM (PhD work)
- Team as of Mid 2022 (CEA/DAM and ECR Lab)
  - 2.5 research scientists, 3 PhD students, 3 engineers
- Freely available at https://mpc.hpcframework.com (version 4.1.0)
- Contact: julien.jaeger@cea.fr

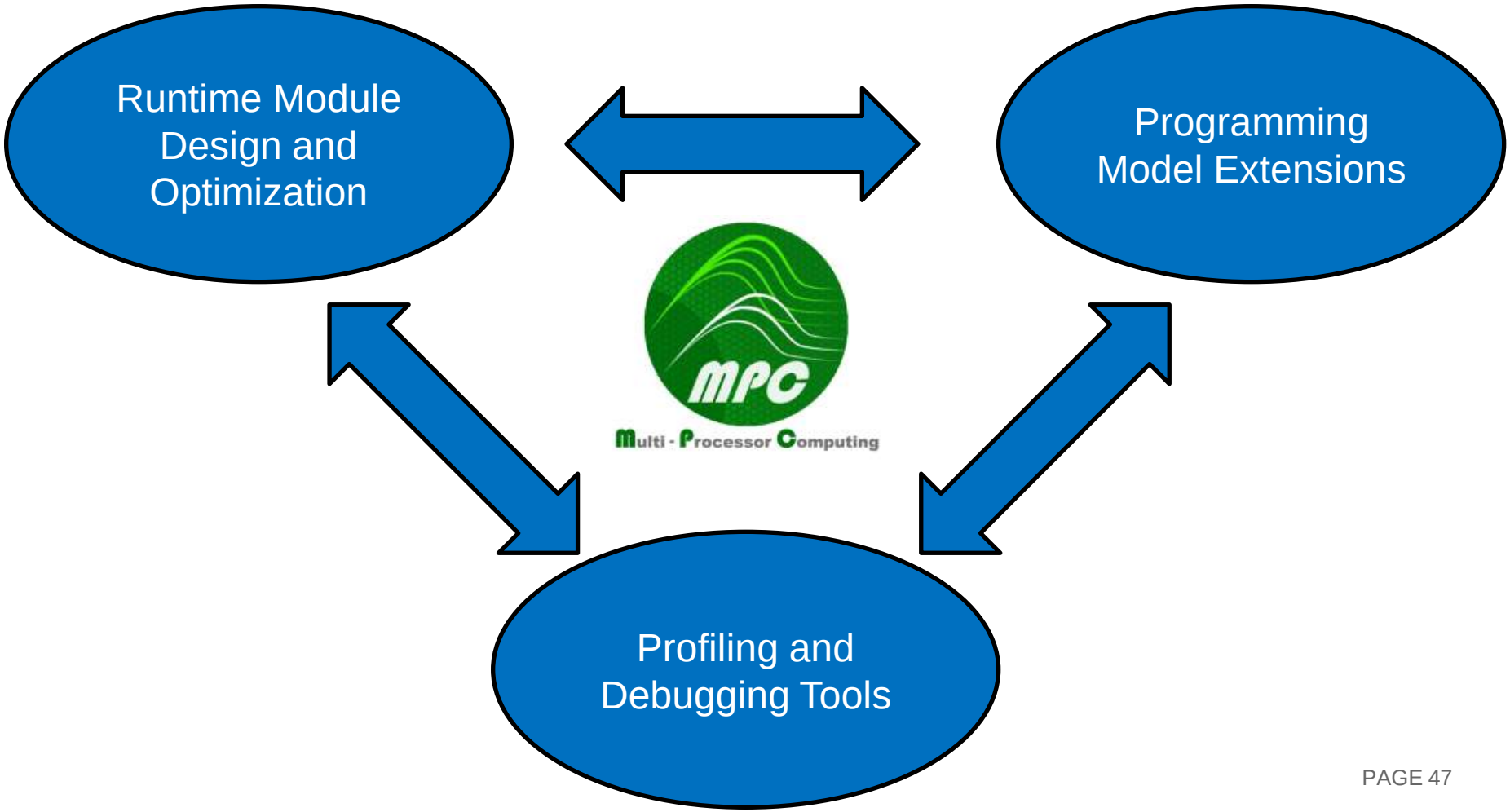
- Summary

- Unified parallel runtime for clusters of NUMA machines

- Main features

- Full MPI implementation
- Full OpenMP implementation
- Pthread compatibility
- NUMA-aware thread-aware memory allocator
- Debugger (patched GDB)
- Compiler (patched GCC and Intel compiler support)

Outline


Runtime Module Design and Optimization
Programming Model Extensions
MPC
Multi - Processor Computing
Profiling and Debugging Tools

STATUS OF MPI- 3.1 IMPLEMENTATIONS

Status of MPI-3.1 Implementations

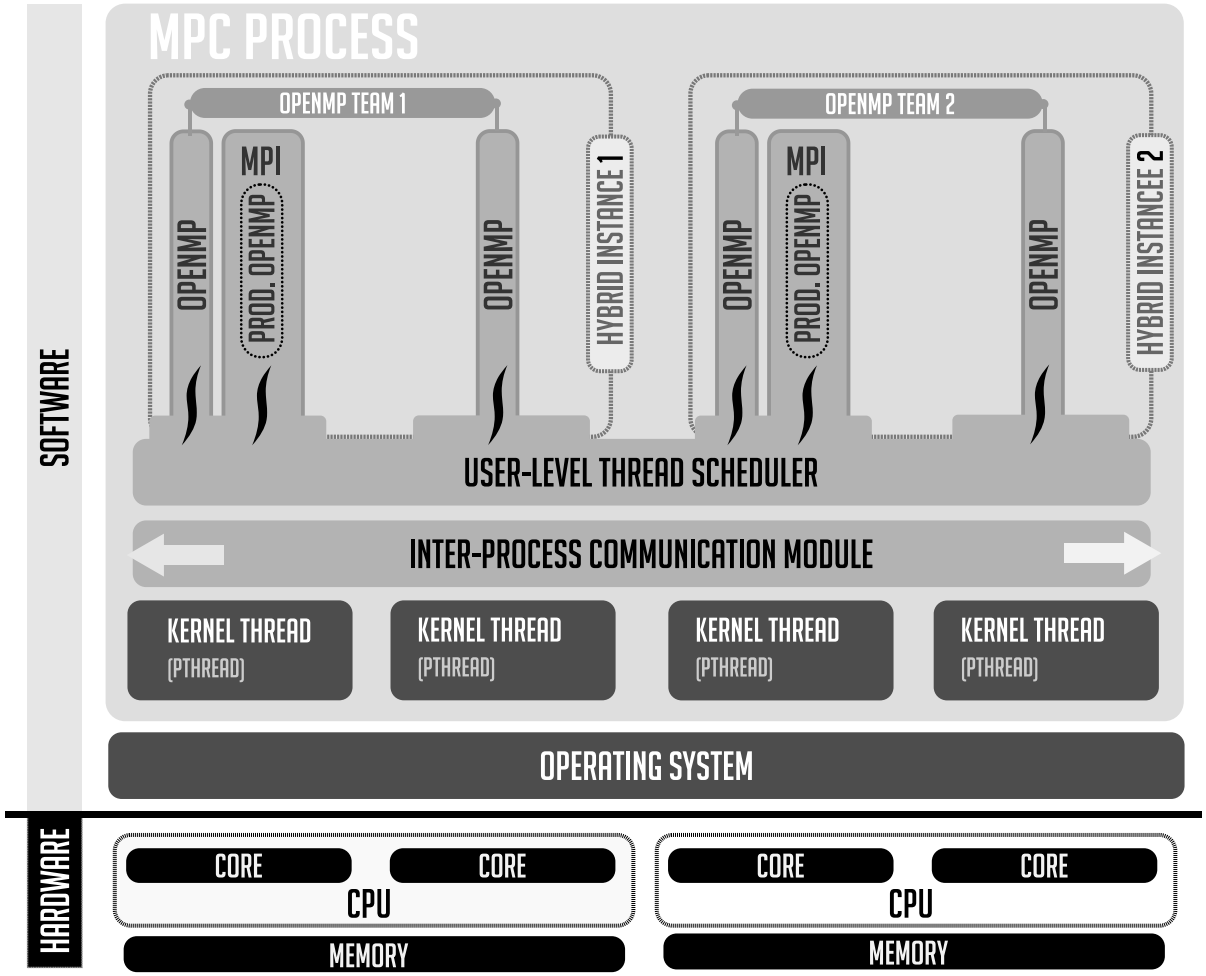
	MPICH	MVAPICH	Open MPI	Cray MPI	Tianhe MPI	Intel MPI	IBM BG/Q MPI [1]	IBM PE MPICH [2]	IBM Platform	SGI MPI	Fujitsu MPI	MS MPI	MPC	NEC MPI
NBC	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	(*)	✔	✔
Nbrhood collectives	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✘	✔	✔
RMA	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✘	Q2'17	✔
Shared memory	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✔	*	✔
Tools Interface	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	*	Q4'16	✔
Comm-creat group	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✘	*	✔
F08 Bindings	✔	✔	✔	✔	✔	✘	✔	✘	✘	✔	✘	✘	Q2'16	✔
New Datatypes	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✔	✔	✔
Large Counts	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✔	Q2'16	✔
Matched Probe	✔	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✔	Q2'16	✔
NBC I/O	✔	Q3'16	✔	✔	✘	✘	✘	✘	✘	✔	✘	✘	Q4'16	✔

Release dates are estimates and are subject to change at any time.
"✘" indicates no publicly announced plan to implement/support that feature.
Platform-specific restrictions might apply to the supported features

[1] Open Source but unsupported [2] No MPI_T variables exposed * Under development (*) Partly done

MPC HYBRID EXECUTION MODEL

- MPI/OpenMP integration
  - Automatic MPI task placement on the node
  - Automatic OpenMP thread placement
    - Topology inheritance
- Example
  - Node with 2 CPUs
  - 2 cores per CPU
  - 2 MPI tasks per node
  - Default: 2 OpenMP threads per team



ETLS: AUTOMATIC PRIVATIZATION

- Global variables

- Expected behavior: duplicated for each MPI task
- Issue with thread-based MPI: global variables shared by MPI tasks located on the same node

- Solution: Automatic privatization

- Automatically convert any MPI code for thread-based MPI compliance
  - Rely on Extended TLS (MPI level for global variables and OpenMP level for threadprivate variables)
- New option to C/C++/Fortran compiler: `-fmpc-privatize`
  - Require modifications of Front-end, Middle-end and Back-end
  - Completely transparent to the user
- Open source: available in MPC package

- Supported compilers

- GCC: patched GCC/G++/GFORTRAN shipped with MPC package
- Intel: compiler support for Xeon and MIC
  - Compilation flag for ICC, ICPC and IFORT : `-fmpc-privatize`
- PGI: future support

ETLS: AUTOMATIC PRIVATIZATION

- Official MPC support in Intel 15 compilers
  - Man page from icc/icpc/ifort

```
> man icc
...
Feature: Privatization of static data for the MPC unified parallel runtime
Requirement: Appropriate elements of the MultiProcessor Computing (MPC)
framework For more information, see http://mpc.sourceforge.net/
...
-fmpc-privatize (L*X only) / -fno-mpc-privatize (L*X only)
Enables or disables privatization of all static data for the MultiProcessor
Computing environment (MPC) unified parallel runtime.
Architecture Restriction: Only available on Intel(R) 64 architecture
Arguments: None
Default: -fno-mpc-privatize
The privatization of all static data for the MPC unified parallel runtime is
disabled.
Description:
This option enables or disables privatization of all static data for the
MultiProcessor Computing environment (MPC) unified parallel runtime.
Option -fmpc-privatize causes calls to extended thread-local-storage (TLS)
resolution, run-time routines that are not supported on standard Linux* OS
distributions.
This option requires installation of another product. For more information, see
Feature Requirements.
```

DEBUGGING

- Goal: tools to help application and feature debugging

- Static analysis [EuroMPI 13, IWOMP 14, EuroPar 15]

- Extend GCC compiler to analyze parallel application (MPI, OpenMP and MPI+OpenMP)
- PARCOACH platform

- Interactive debugging [MTAAP 10]

- Provide a generic framework to debug user-level thread
  - Evaluated on MPC, Marcel, GNUPth
- Provide a patched version of GDB
- Collaboration with Allinea DDT
  - MPC support in Allinea DDT 3.0

- Trace-based dynamic analysis [PSTI 13]

- Use traces to debug large-scale applications
- Crash-tolerant trace engine
- Parallel trace analyzer

OUTLINE

- HPC Status
  - Supercomputing Ranking
  - HPC History
  - HPC Roadmap
  - French/Europe Focus
  - Main Challenges
- R&D on HPC Software Stack
  - Open-source HPC development at CEA
  - Focus on parallel runtime: MPC
- **Beyond Exascale**
- Conclusion

BEYOND EXASCALE

Beyond Exascale

- Exascale machine

- Need to expose computational units w/ limited power consumption
- Heterogeneous computing
- Currently: GPUs (Nvidia, AMD, Intel)

- Evolution of heterogeneous computing

- Notion of accelerators
- Integration of discrete accelerators within same chip

- Possible directions

- More adapted/dedicated accelerators
  - ➔ FPGA
- More disruptive accelerators
  - ➔ QPU (Quantum Processing Unit)

CONCLUSION

CONCLUSION

- Exascale era
  - First Exascale machine in United States (DoE)
  - Based on heterogeneous systems
  - Co-design between vendor / computing center / academia

- CEA R&D
  - Several computing center
  - Co-design w/ Atos Bull (most recent machine Exa1-HF)
  - Internal R&D on software stack (mainly open source)
    - Example: MPC

- Beyond Exascale
  - Heterogeneous computing
  - Towards more integrated accelerators
  - Trend to programmable / adapted accelerator (e.g., FPGA)
  - Convergence of HPC – Quantum Computing

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM-Ile de France- Bruyères-le-Châtel 91297 Arpajon Cedex Cedex
T. +33 (0)1 69 26 40 00

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019