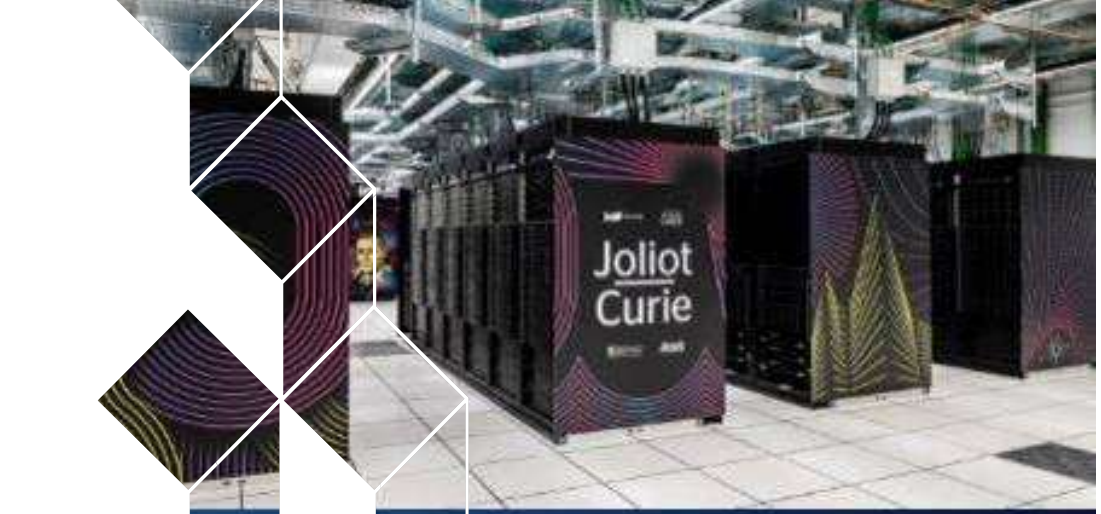


Evolution of High-Performance Computing Architectures

Patrick Carribault

Research Director & Fellow

patrick.carribault@cea.fr

TOP500 June 2024

Rank	Country	System	Cores	Rmax	Rpeak	Power
1	United States	Frontier	8,699,904	1,206.00	1,714.81	22,786
2	United States	Aurora	9,264,128	1,012.00	1,980.01	38,698
3	United States	Eagle	2,073,600	561.20	846.84	
4	Japan	Fugaku	7,630,848	442.01	537.21	29,899
5	Finland	Lumi	2,752,704	379.70	531.51	7,107

- **Exaflop/s!!! (2 systems so far…)**
  - Millions of cores
  - Power consumption between 20 and 40 MWatts
- Compute architecture based on accelerators
  - Main suppliers: Intel, AMD, NVIDIA
- **Current architectures…**
- **Hardware challenges**

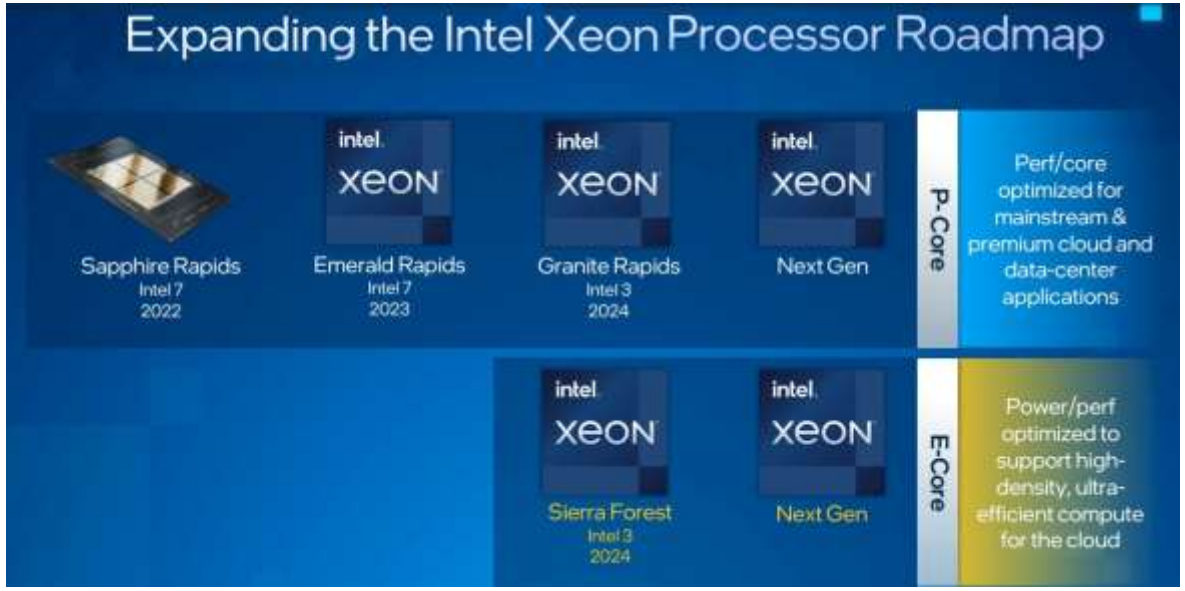
Source: RIKEN

Current Architectures - Intel

Source: Intel

Current Architectures

- **CPU**
  - Xeon: X86, AVX512-based instruction set
  - Xeon Max: SPR+HBM
- **GPU**
  - First « real » high-performance generation
  - Ponte Vecchio (PVC)



HPC Building Block

- Compute nodes with
  - Dual-socket SPR
  - PVC accelerators
- Aurora efficiency: 26 Gflops/W
- 1 Eflops (HPL) for 38 MWatts

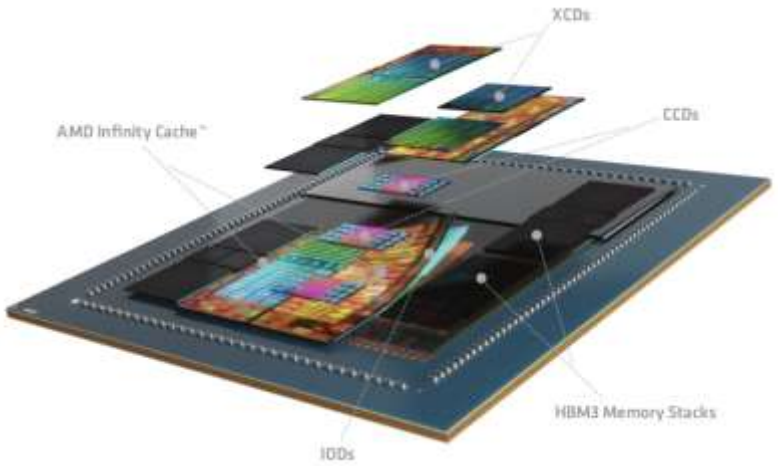
Next?

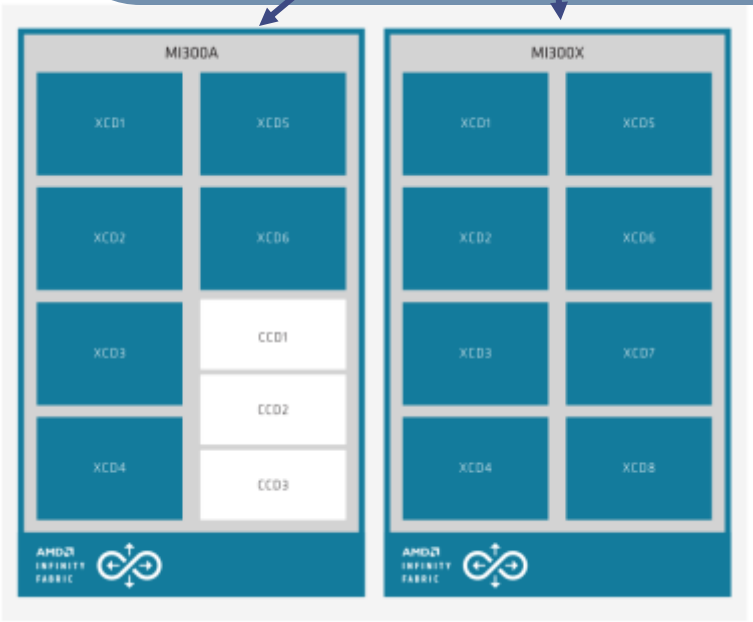
Current Architectures - AMD

Source: AMD

Current Architectures

- **CPU**
  - X86-based architecture
  - Increase number of cores,
  - Stable vector units
- **GPU**
  - 3rd generation of high-performance GPU (MI-series)
  - Discrete: MI300X
  - APU: MI300A





HPC Building Block

- MI300A APU
  - ¼ of CPU & ¾ of GPU
    - 24 CPU cores
    - 228 Compute Units
  - Shared memory (128 GB HBM3)
  - Seamless access to memory but small memory capacity + fixed ratio CPU/GPU
- El capitan Early Delivery: ~20 Pflops HPL for 512 APUs. Efficiency ~50-70 Gflops/W

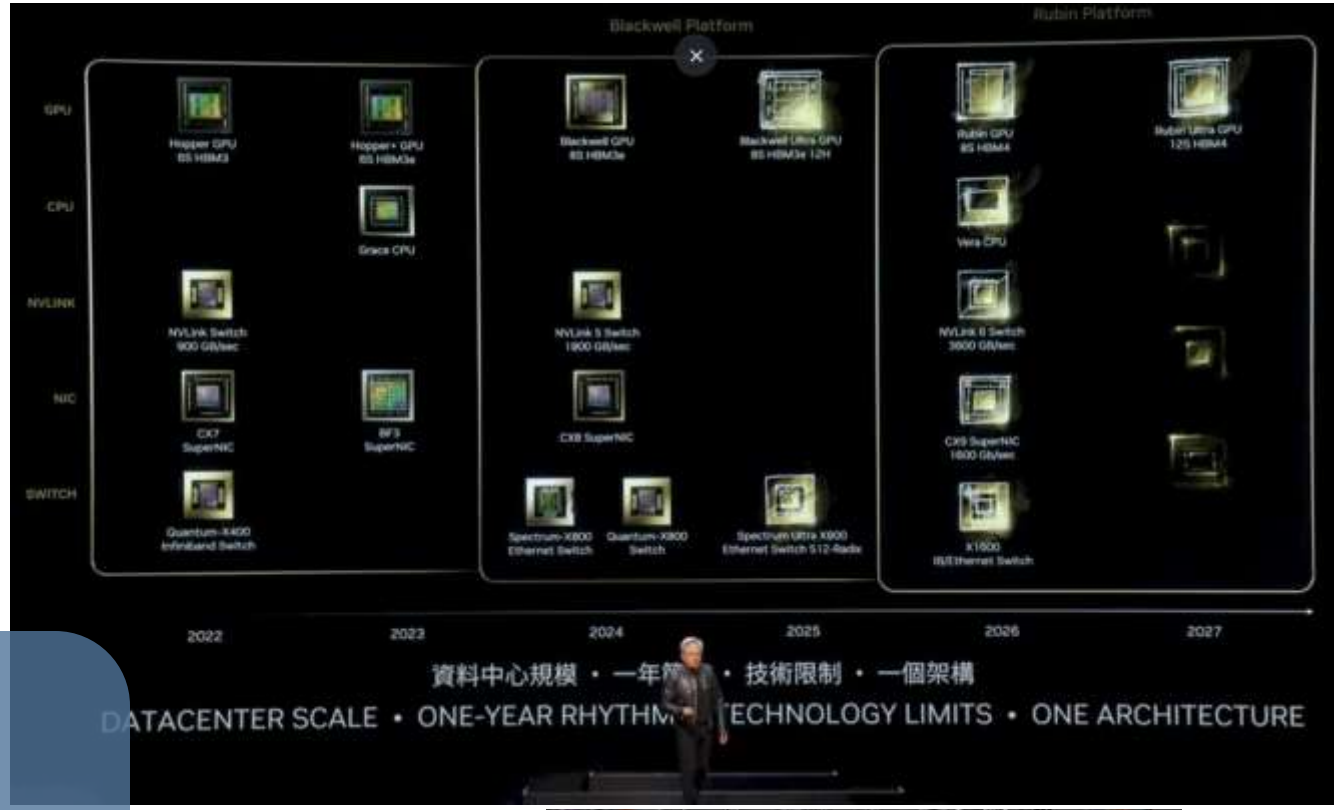
Next?

Current Architectures - NVIDIA

Source: NVIDIA

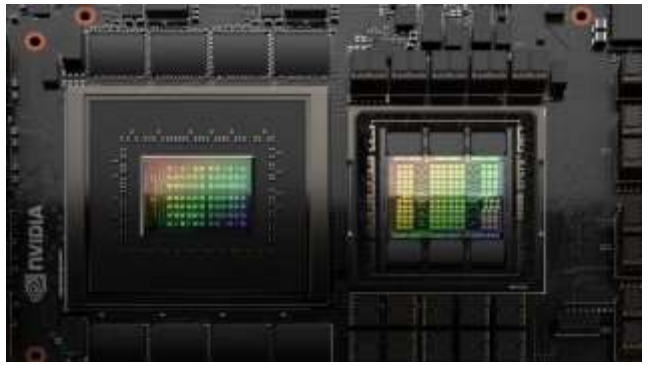
Current Architectures

- **CPU**
  - Grace: ARM-based architecture
  - >70 cores with SVE
  - LPDDR memory
- **GPU**
  - Hopper (2022)
  - Blackwell (2024)



HPC Building Block

- High pace!
- Approach of combined CPU/GPU
- Grace Hopper
  - Ratio 1:1 (1 CPU + 1 GPU)
  - Efficiency ~50-60 Gflops/W
- Grace Blackwell
  - Ratio 1:2 (1 CPU + 2 GPUs)



Next?

(Some) Challenges

Compute Units

- Increase number of units
  - Need to expose more and more parallelism
- Deal with different kinds of units to limit energy consumption
  - Ratio CPU / Accelerators?
  - General purpose / dedicated or both?
  - GPU?, FPGA? Quantum?
  - New architectures (compute in memory)
- Floating-point precision
  - What about 64b units?
  - Need to deal with scalar/vector/matrix instructions

Memory

- Evolution of caches and coherency architecture
- More technologies (beyond DDR)
  - LPDDR
  - High-Bandwidth Memory (HBM)
  - Non-volatile memory (NVM)
- Extended memory levels
  - Links between CPU and accelerators
  - Support of unified memory? Coherent?
  - Disaggregated memory?

Network

- Increase in number of nodes
  - 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)
- Different networks?
  - Scale-up vs Scale-out networks
  - HPC vs IA workloads?

Challenges – Compute Units

Compute Units

- Increase number of units
  - Need to expose more and more parallelism
- Deal with different kinds of units to limit energy consumption
  - Ratio CPU / Accelerators?
  - General purpose / dedicated or both?
  - GPU?, FPGA? Quantum?
  - New architectures (compute in memory)
- Floating-point precision
  - What about 64b units?
  - Need to deal with scalar/vector/matrix instructions

Challenges – Compute Units

Compute Units

- Increase number of units
  - Need to expose more and more parallelism
- Deal with different kinds of units to limit energy consumption
  - Ratio CPU / Accelerators?
  - General purpose / dedicated or both?
  - GPU?, FPGA? Quantum?
  - New architectures (compute in memory)
- Floating-point precision
  - What about 64b units?
  - Need to deal with scalar/vector/matrix instructions

- **Increase number of units**
  - Number of cores in CPUs
  - Number of hierarchical units in GPUs
  - ➔ Impact on parallelism expressed in applications
- **Interest in matrix-based instruction**
  - Useful for some domains (e.g., IA)
  - Matrix-based instructions might not be generated by compilers
  - ➔ Usage in HPC applications?
- **Rely on units exposing reduced precision**
  - Deal with mixed and/or reduced precision
    - In library or with tools
    - At numerical level and/or computer science
  - Depends on application and workload

Challenges - Memory

Memory

- Evolution of caches and coherency architecture
- More technologies (beyond DDR)
  - LPDDR
  - High-Bandwidth Memory (HBM)
  - Non-volatile memory (NVM)
- Extended memory levels
  - Links between CPU and accelerators
  - Support of unified memory? Coherent?
  - Disaggregated memory?

Challenges - Memory

Memory

- Evolution of caches and coherency architecture
- More technologies (beyond DDR)
  - LPDDR
  - High-Bandwidth Memory (HBM)
  - Non-volatile memory (NVM)
- Extended memory levels
  - Links between CPU and accelerators
  - Support of unified memory? Coherent?
  - Disaggregated memory?

- **Evaluation of memory technologies**
  - Different characteristics (bandwidth, latency)
  - Possibility to mix different technologies inside compute nodes
  - Number of memory channels per socket slowly increases ➔ lower capacity?
  - ➔ impact on application for data locality/storage
- **Links between CPUs and GPUs**
  - Main trend is to put CPU and GPU closer
  - For example: NVIDIA Grace-Hopper or MI300A
  - But be careful about characteristics (latency / bandwidth) and coherency (IO, cache, …)
  - ➔ real need on application side

Challenges - Network

Network

- Increase in number of nodes
  - 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)
- Different networks?
  - Scale-up vs Scale-out networks
  - HPC vs IA workloads?

Challenges - Network

Network

- Increase in number of nodes
  - 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)
- Different networks?
  - Scale-up vs Scale-out networks
  - HPC vs IA workloads?

- **Network technologies**
  - Not many high-performance network available
  - One French/European solution: Eviden BXI
- **Main network features**
  - Number of nodes is important but number of NIC more!
  - Topology ➔ depends on communication pattern
  - Ratio compute units / NIC
- **Study scale-up & scale-out networks**
  - Network between GPUs mainly designed for AI so far
  - Possibility to leverage both networks for broader spectrum
    - E.g., with MPI?

Conclusion

- **TOP500**
  - Exaflops reached in 2022
  - 2 systems so far
  - Lots of core, different memory technologies, heterogeneous architectures
- **Current architectures**
  - Product lines include both regular CPU and GPUs
  - Main trend to increase number of cores and overall power consumption
  - But efficiency remains good, depending on operations
- **(Some) Challenges**
  - *Compute units:*
    - No major changes in vector length, Increase in number of cores
    - Deal with floating-point precision
    - Other architectures? RISC V
  - *Memory*
    - Different technologies (tradeoff between latency/bandwidth/capacity)
    - Possibility to have multiple technologies on the same node
  - *Network*
    - Topologies
    - Number of NIC per node, total number of nodes
    - Scale-up vs Scale-out

MERCI!

Questions?