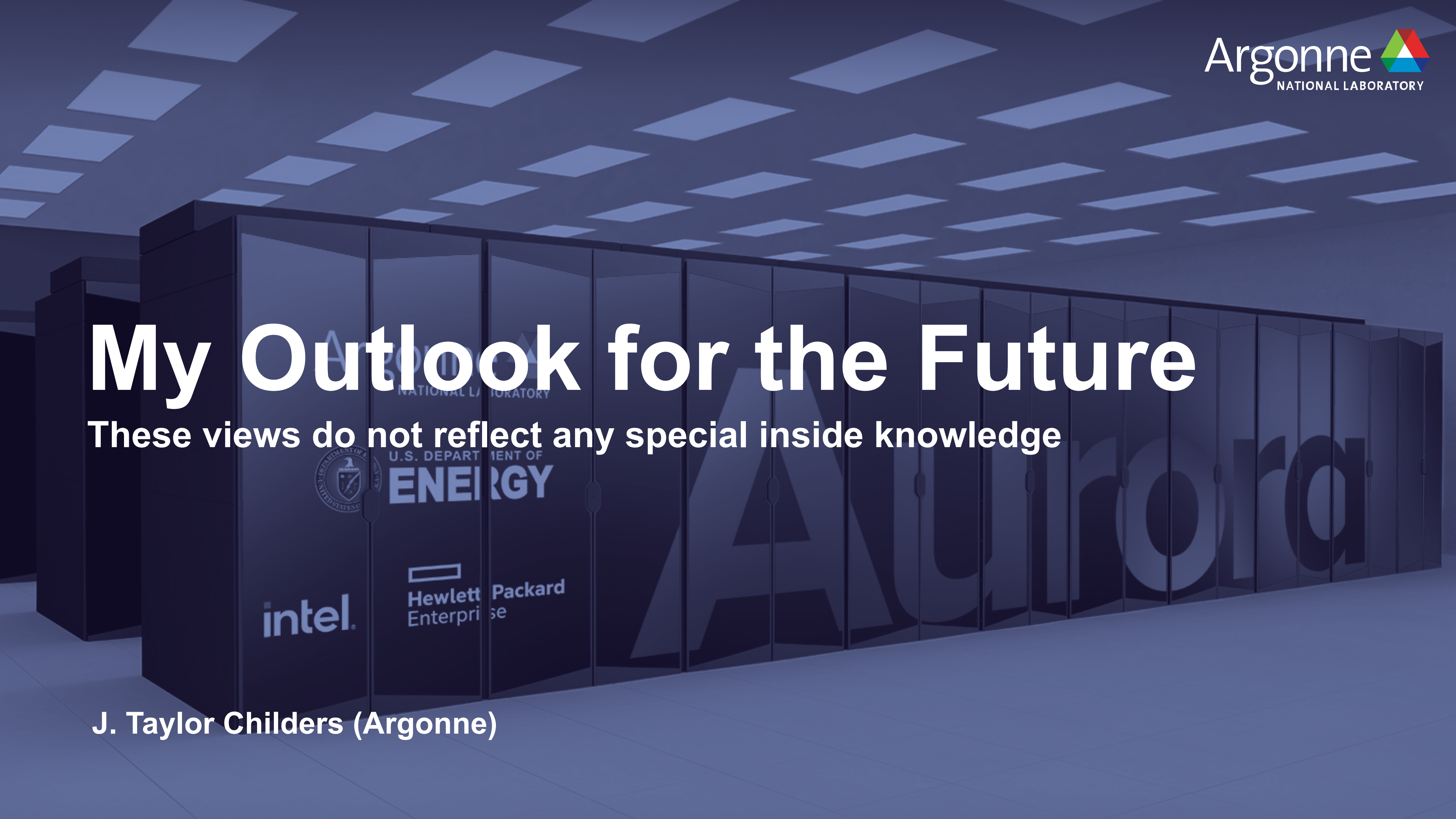
My Outlook for the Future

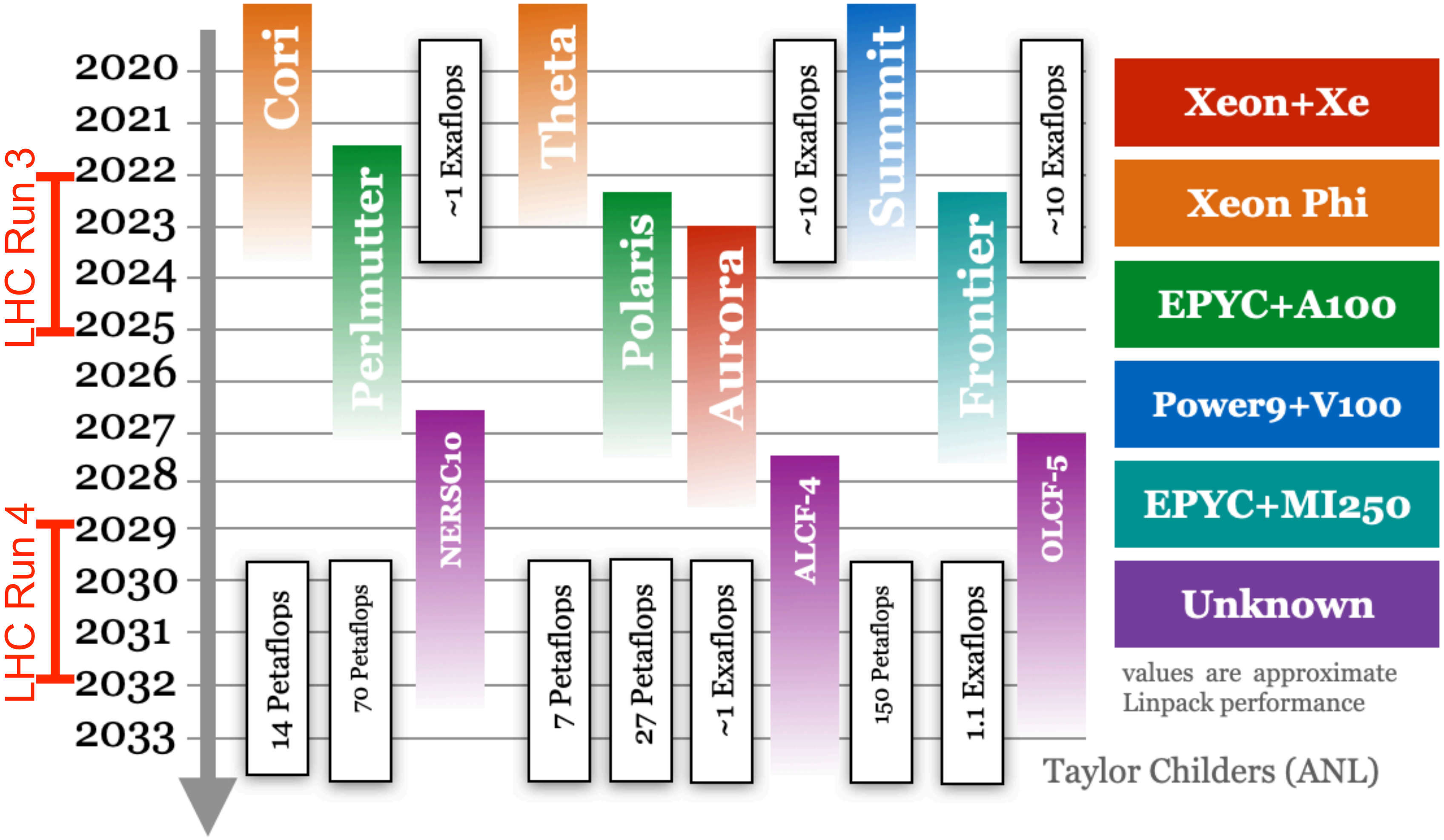
These views do not reflect any special inside knowledge



J. Taylor Childers (Argonne)

US HPC Outlook

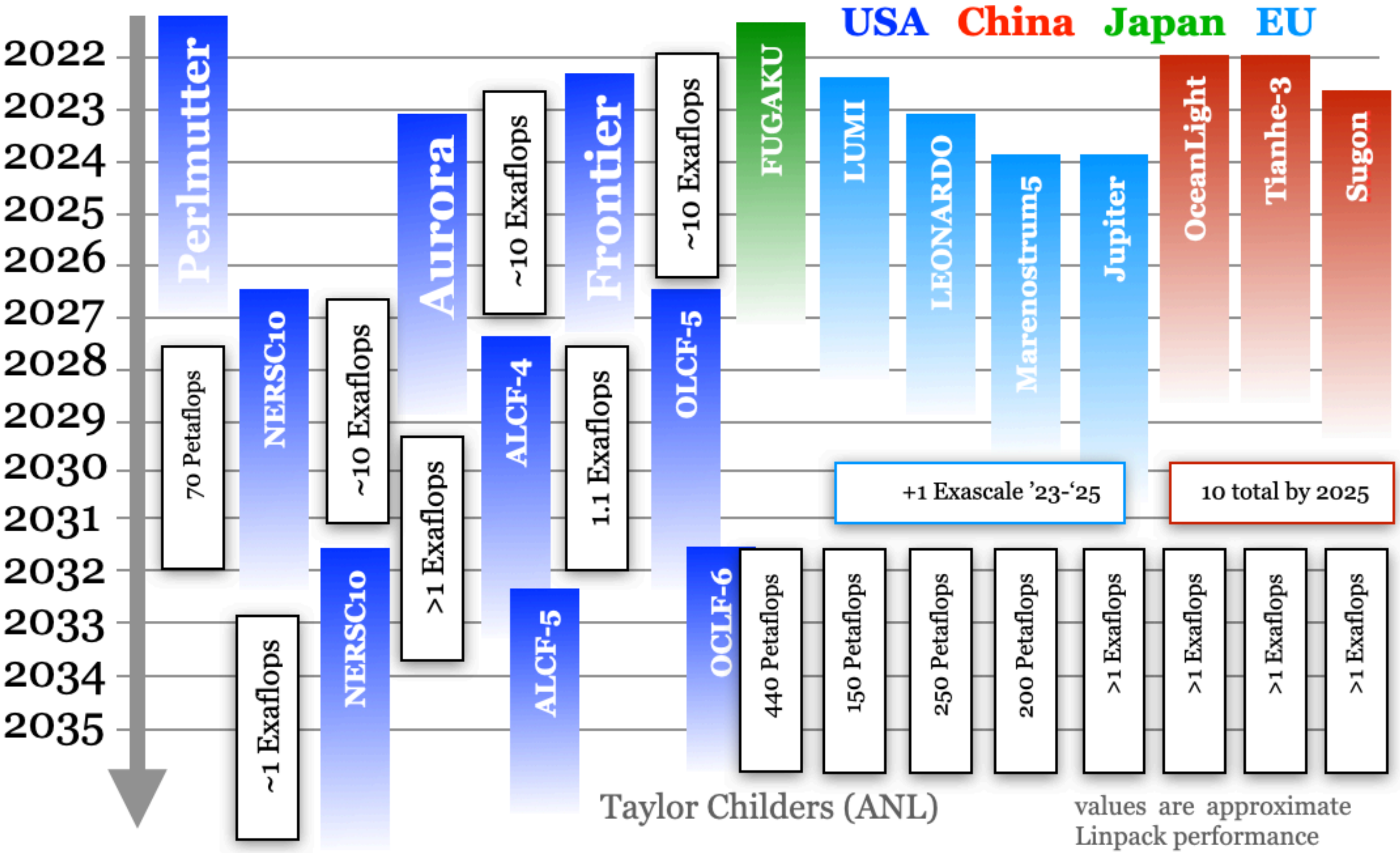
- We are at the cusp of Exascale in the US HPC arena.
- DOE has deployed Frontier and Aurora will come in the next months.
- Next generation machines will be online at all DOE HPC facilities during the HL-LHC turn on.
- Currently operating HPCs with Intel, NVidia, and AMD.



Taylor Childers (ANL)

Wider HPC Outlook

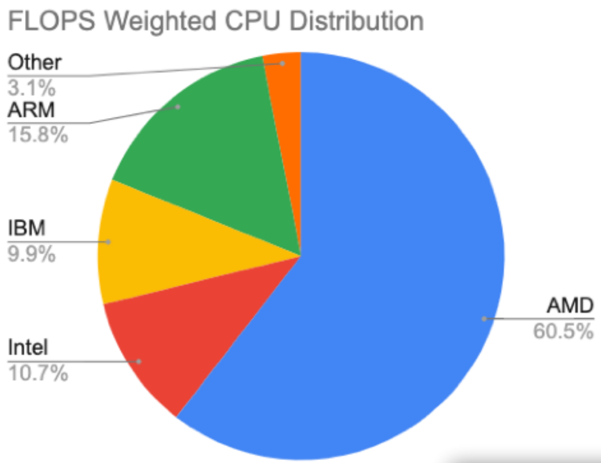
- Most regions are investing in “sovereign” technologies.
- EuroHPC JU:
  - investing in ARM then shifting to RISC-V
  - though recent foundry investment from Intel may shift this plan
- Japan investing in ARM via Fujitsu chips
- China already has 3 Exascale machines on the ground with home grown tech.
- Expect to see more options later in the decade.



Taylor Childers (ANL)

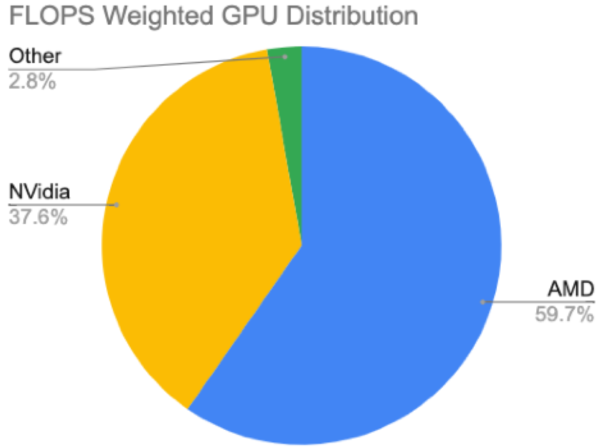
values are approximate Linpack performance

June 2022 Architecture Distribution



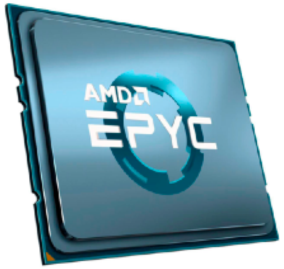
Taken from Top500
> 10 PetaFLOPs
~50 machines

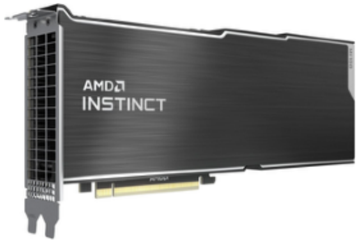
Frontier Dominates

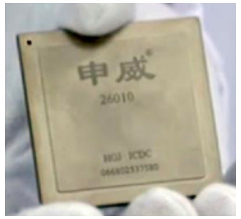


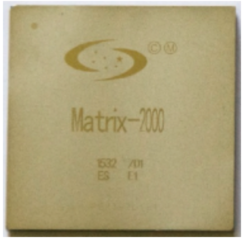
CPU

GPU



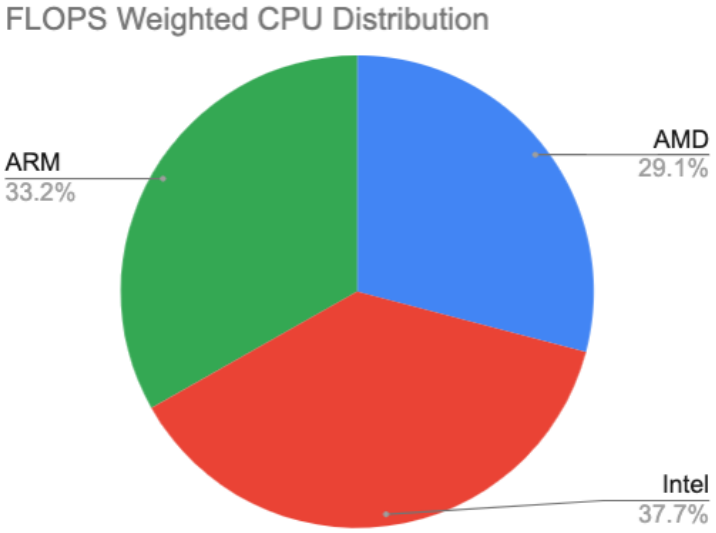


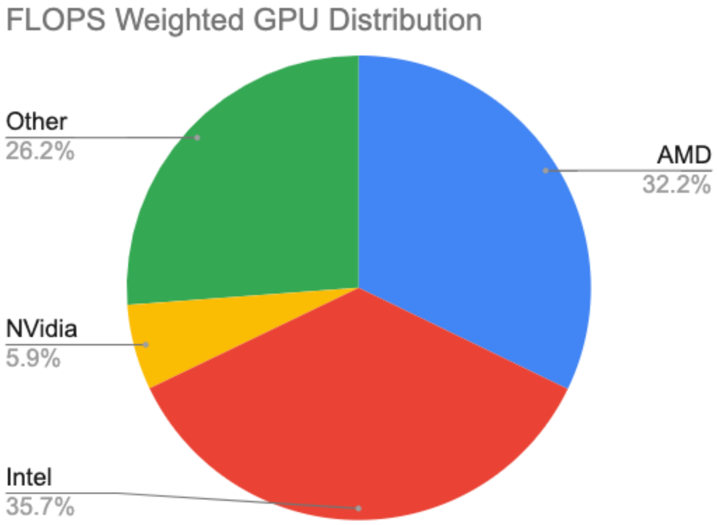


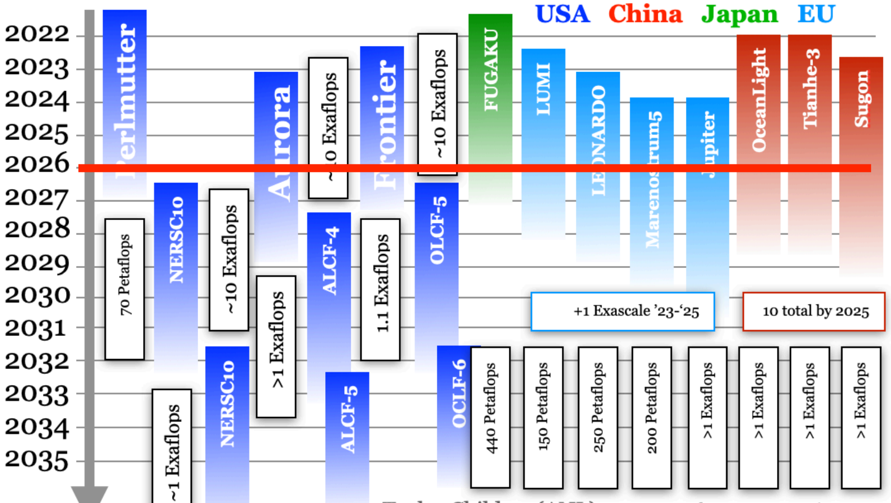


Approx. 2026 Architecture Distribution

- By 2026, US and Europe will have 2 Exascale machines, and China may be up to 10.
- I did not include the Chinese supercomputers in the plot.
- I included 2 European Exascale machines (who’s architecture has not been defined, but used these example specs:
    - Jupiter: ARM CPU, no GPU
    - EuroHPC #2: ARM CPU, RISC-V accelerator
- I could not find information on the Japanese program plans after Fugaku.





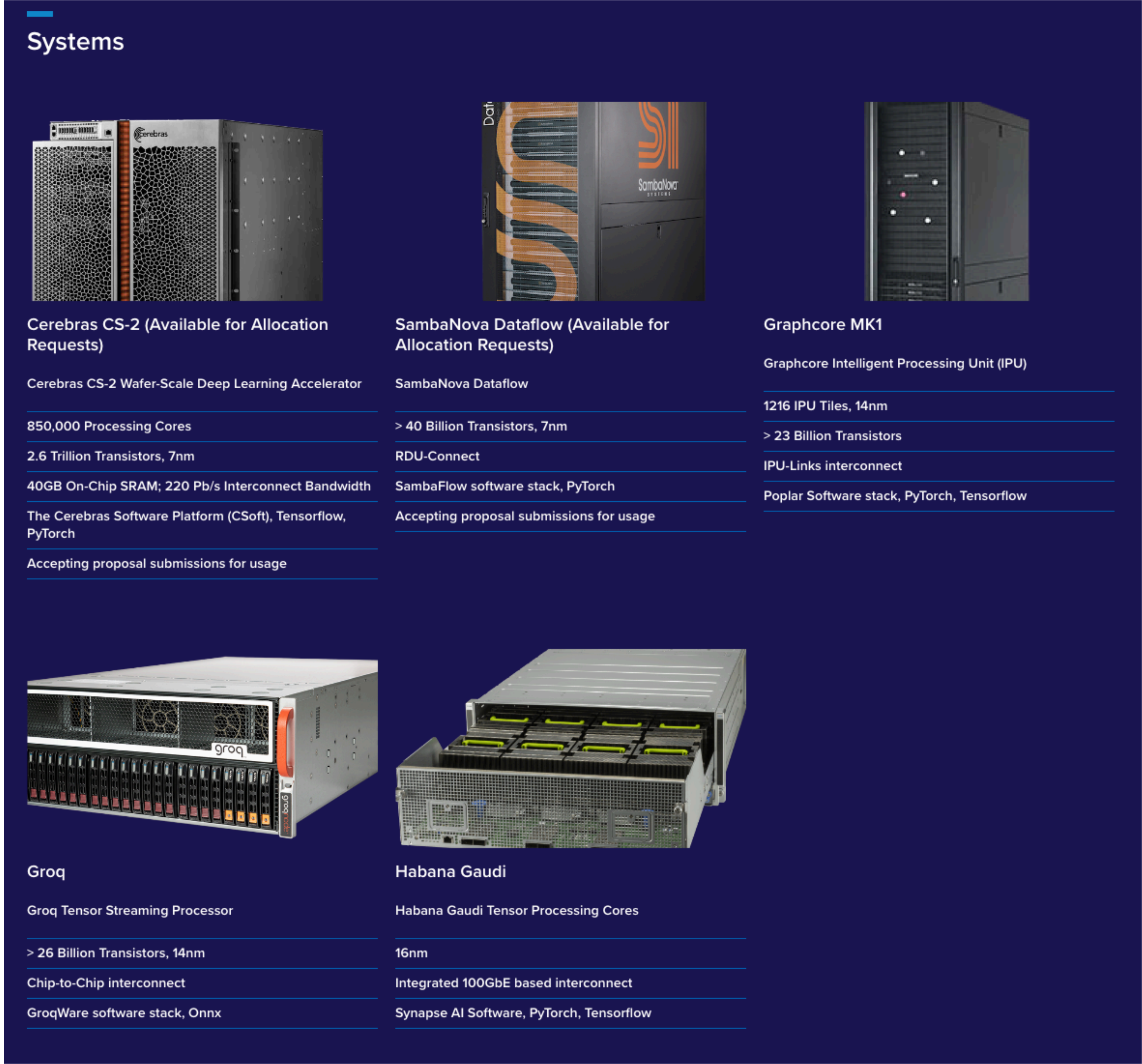


Taylor Childers (ANL)

values are approximate Linpack performance

Specialty Hardware

- DOE is currently working with many custom AI hardware vendors.
- ALCF hosts the DOE's AI Testbed: https://www.alcf.anl.gov/alcf-ai-testbed
- We are exploring the possibility of these systems being integrated as usable sidecars to future supercomputers.



Systems

Cerebras CS-2 (Available for Allocation Requests)

Cerebras CS-2 Wafer-Scale Deep Learning Accelerator

- 850,000 Processing Cores
- 2.6 Trillion Transistors, 7nm
- 40GB On-Chip SRAM; 220 Pb/s Interconnect Bandwidth
- The Cerebras Software Platform (CSoft), Tensorflow, PyTorch
- Accepting proposal submissions for usage

SambaNova Dataflow (Available for Allocation Requests)

SambaNova Dataflow

- > 40 Billion Transistors, 7nm
- RDU-Connect
- SambaFlow software stack, PyTorch
- Accepting proposal submissions for usage

Graphcore MK1

Graphcore Intelligent Processing Unit (IPU)

- 1216 IPU Tiles, 14nm
- > 23 Billion Transistors
- IPU-Links interconnect
- Poplar Software stack, PyTorch, Tensorflow

Groq

Groq Tensor Streaming Processor

- > 26 Billion Transistors, 14nm
- Chip-to-Chip interconnect
- GroqWare software stack, Onnx

Habana Gaudi

Habana Gaudi Tensor Processing Cores

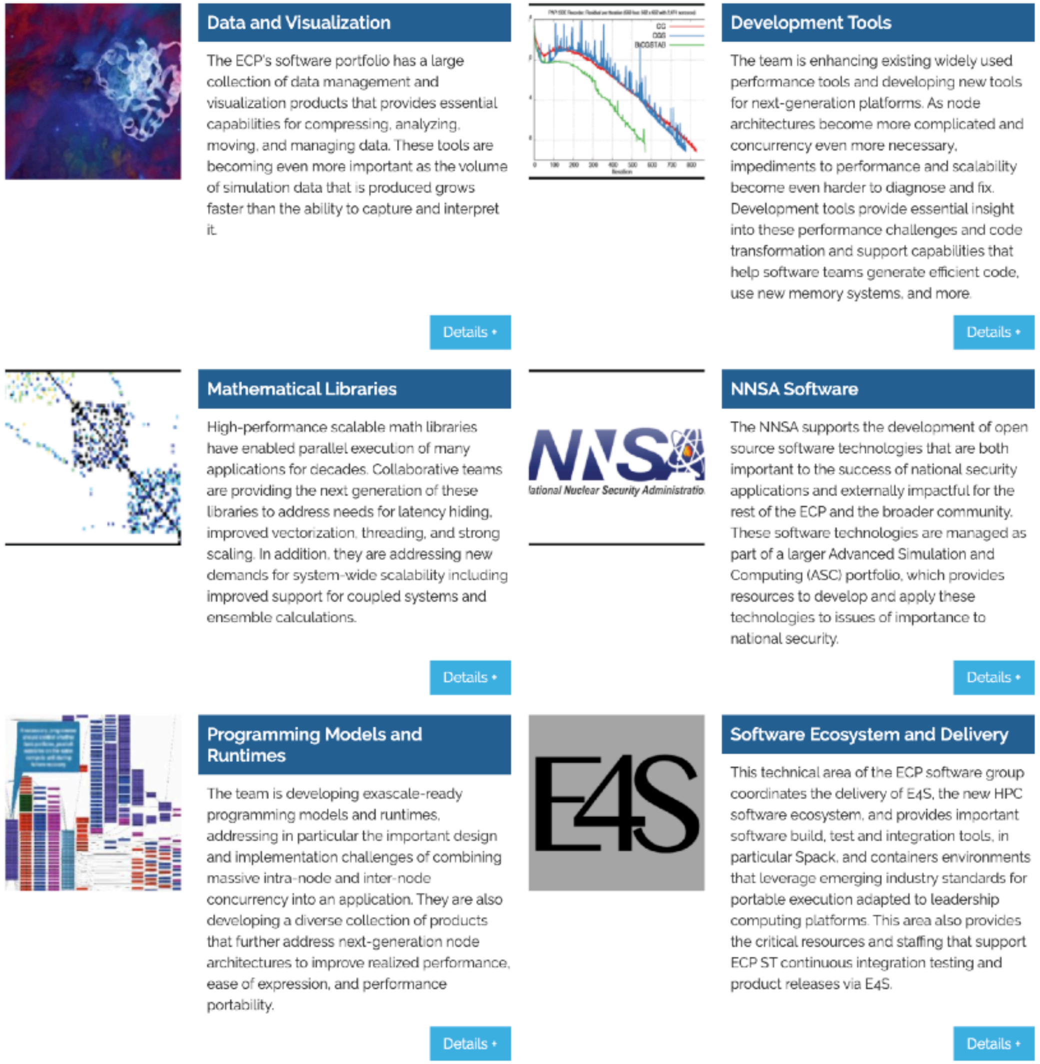
- 16nm
- Integrated 100GbE based interconnect
- Synapse AI Software, PyTorch, Tensorflow

AI4Science

- Exascale Computing Project: a multi-year funded project that laid the groundwork for deploying the US Exascale systems.
  - Included large efforts for software development, including large HPC community software.
- With Exascale machines having arrived, ECPs mission is coming to an end.
- Currently DOE-ASCR has an ongoing AI4Science process of community workshops which is modeled after how the ECP was initially developed.
- This recognizes that AI is going to be an impactful tool in the coming decade that we need to pivot to support and advance.
- If such a program is spun up, it will have important mission impact on LCF systems.

EXASCALE COMPUTING PROJECT

https://www.exascaleproject.org/



Data and Visualization

The ECP's software portfolio has a large collection of data management and visualization products that provides essential capabilities for compressing, analyzing, moving, and managing data. These tools are becoming even more important as the volume of simulation data that is produced grows faster than the ability to capture and interpret it.

Details •

Development Tools

The team is enhancing existing widely used performance tools and developing new tools for next-generation platforms. As node architectures become more complicated and concurrency even more necessary, impediments to performance and scalability become even harder to diagnose and fix. Development tools provide essential insight into these performance challenges and code transformation and support capabilities that help software teams generate efficient code, use new memory systems, and more.

Details •

Mathematical Libraries

High-performance scalable math libraries have enabled parallel execution of many applications for decades. Collaborative teams are providing the next generation of these libraries to address needs for latency hiding, improved vectorization, threading, and strong scaling. In addition, they are addressing new demands for system-wide scalability including improved support for coupled systems and ensemble calculations.

Details •

NNSA Software

The NNSA supports the development of open source software technologies that are both important to the success of national security applications and externally impactful for the rest of the ECP and the broader community. These software technologies are managed as part of a larger Advanced Simulation and Computing (ASC) portfolio, which provides resources to develop and apply these technologies to issues of importance to national security.

Details •

Programming Models and Runtimes

The team is developing exascale-ready programming models and runtimes, addressing in particular the important design and implementation challenges of combining massive intra-node and inter-node concurrency into an application. They are also developing a diverse collection of products that further address next-generation node architectures to improve realized performance, ease of expression, and performance portability.

Details •

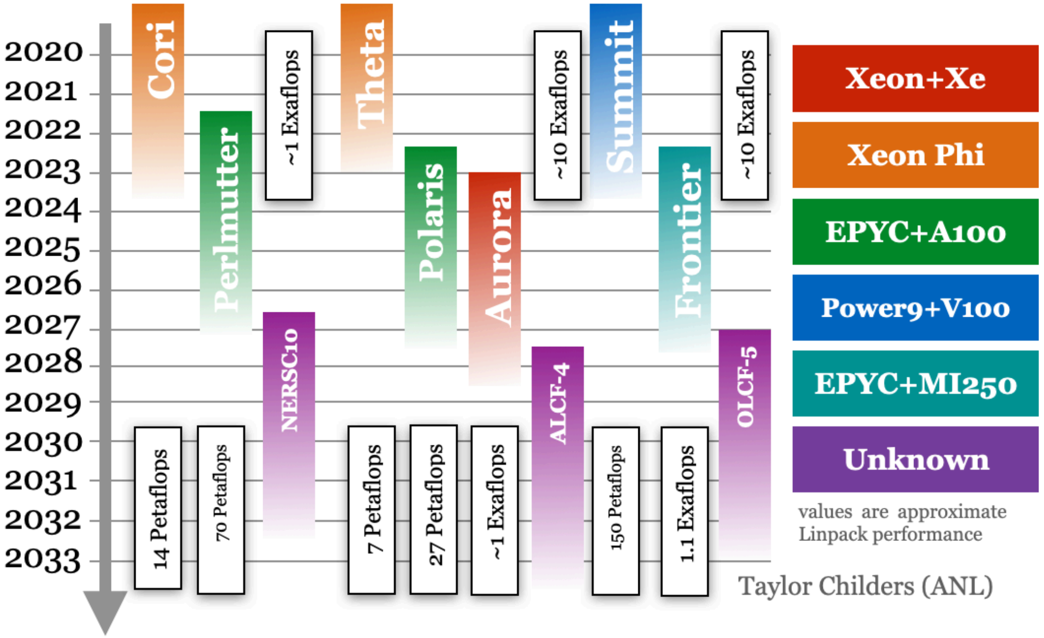
Software Ecosystem and Delivery

This technical area of the ECP software group coordinates the delivery of E4S, the new HPC software ecosystem, and provides important software build, test and integration tools, in particular Spack, and containers environments that leverage emerging industry standards for portable execution adapted to leadership computing platforms. This area also provides the critical resources and staffing that support ECP ST continuous integration testing and product releases via E4S.

Details •

Take Aways?

- Future of Architecture at HPC facilities will remain diverse if not grow more so:
  - Software implications: using portable frameworks will be a benefit (watching std::par and vocally demanding companies support it would be good)
- Current Exascale machines where largely decided before AI became important in DOE science, but will be a driver in the deployment of the next generation machines.
  - This could result in very similar looking ALCF-4, NERSC-10, and OLCF-5 systems.
  - It may also shifting things in unexpected ways.
- The future is hard to predict.



values are approximate Linpack performance

Taylor Childers (ANL)