REPORT ON THE OAK RIDGE NATIONAL LABORATORY'S FRONTIER SYSTEM

Tech Report No. ICL-UT-22-05

Jack Dongarra
University of Tennessee, Knoxville
Oak Ridge National Laboratory
University of Manchester

and

Al Geist
Oak Ridge National Laboratory

May 30, 2022

Prepared by
UNIVERSITY OF TENNESSEE,
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE,
INNOVATIVE COMPUTING LABORATORY

Overview

The Frontier compute system was designed and built by HPE, Cray, and AMD and installed in 2021 at Oak Ridge National Laboratory (ORNL), located in Oak Ridge, Tennessee. It represents the first of three Exascale computer system being developed for the US Department of Energy (DOE). The other two are the Aurora computer to be delivered to Argonne National Laboratory in 2022 or 2023 and the El Capitan computer to be delivered to Lawrence Livermore National Laboratory in 2023 (see Figure 1).

DOE's ORNL Campus

ORNL is a U.S. multiprogram science and technology national laboratory. It is the largest science and energy national laboratory in the Department of Energy (DOE) system (by size) and third largest by annual budget. Located in Oak Ridge, Tennessee, its scientific programs focus on advanced materials, neutron science, energy, high-performance computing, systems biology and national security, sometimes in partnership with the state of Tennessee, universities and other industries.

In 2004 ORNL was selected to be a primary site for the DOE Leadership Computing Facility (LCF). The mission of the LCF is to provide world-class computational resources and specialized services for the most computationally intensive global challenges and enable transforming discoveries in energy technologies, materials, environment, health, and basic science. The LCF has a highly competitive user allocation program (INCITE) open to researchers around the world to ensure the most important and timely computational challenges are chosen for the LCF each year. Over the past two decades, the LCF has been critical to advancing innovation in energy assurance, ecological sustainability, scientific discovery, and global security. The LCF has played a vital role in providing extreme-scale capabilities to assist in global events analysis and response as exemplified by the response to the Fukushima earthquake and tsunami, and the ongoing Covid-19 pandemic response.

As exemplified by its mission, the Oak Ridge Leadership Computing Facility has had a string of the world's top supercomputers, including Jaguar, Titan, Summit and now Frontier (see figure 2). In 2012 with the delivery of Titan, Oak Ridge spearheaded a shift in supercomputer architectures to a hybrid GPU/CPU node. Oak Ridge has continued this trend through Summit and Frontier, increasing the number of GPUs per CPU in each generation (see Figure 3). The GPU/CPU architecture has three big advantages for supercomputers. First, the raw computational performance of a GPU vs a CPU. Second, the power efficiency or flops/watt of a GPU, and third, the high-performance multi-precision operations in modern GPUs makes them very effective for machine learning and AI problems. Using mixed precision, Summit was the first computer to solve a full application at 2.3 Exaops [reference 2018 Gordon Bell award]

Development of Exascale hardware for Frontier and the other two U.S. Exascale systems started in 2012 with a DOE program called Fast Forward, which funded a half-dozen HPC vendors to explore innovations in chip design and system reliability and power efficiency. The Fast Forward

work by AMD allowed Frontier to break the 20 MW per Exaflop target. Frontier is at 14.5 MW per Exaflop for an operation like matrix multiplcation (DGEMM). Development of Exascale applications and software began in 2016 with the launch of the U.S. Exascale Computing Project (ECP).

Exascale Computing Project

The DOE Exascale Computing Project is a large national project that includes over 1,000 researchers from 15 labs, 70 universities, and 32 vendors to tackle exascale application development as well as software libraries and software technologies [cite kothe]. The ECP is a collaborative effort between the DOE Office of Science (DOE-SC) and the DOE National Nuclear Security Administration (NNSA).

ECP has three technical areas: Application Development, Software Technologies, and Hardware and Integration. Exascale applications are a foundational element of the ECP and are an important vehicle for delivery of consequential solutions and insight from exascale systems. The breadth of these applications runs the gamut: chemistry and materials; energy production and transmission; earth and space science; data analytics and optimization; and national security (see Figure 4). Applications are built on underlying software technologies that play an essential supporting role in application efficacy on a broad range of computing systems. An expanded and vertically integrated software stack is being developed to include advanced mathematical libraries and frameworks, extreme-scale programming environments and tools, and visualization libraries (see Figure 5). ECP activities ensure a capable exascale computing ecosystem by integrating exascale applications, software technologies and hardware innovations into the DOE HPC facilities. The project also supports US HPC vendor R&D focused on innovative architectures for competitive exascale system designs.

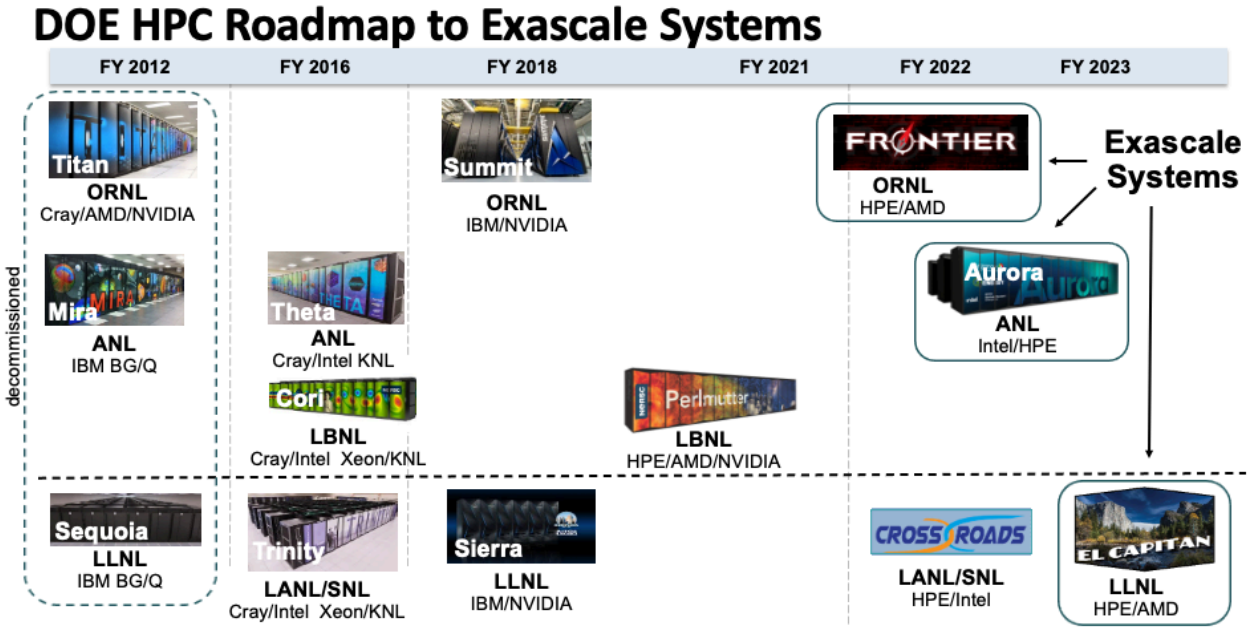

Figure 1: DOE HPC Roadmap to Exascale Systems

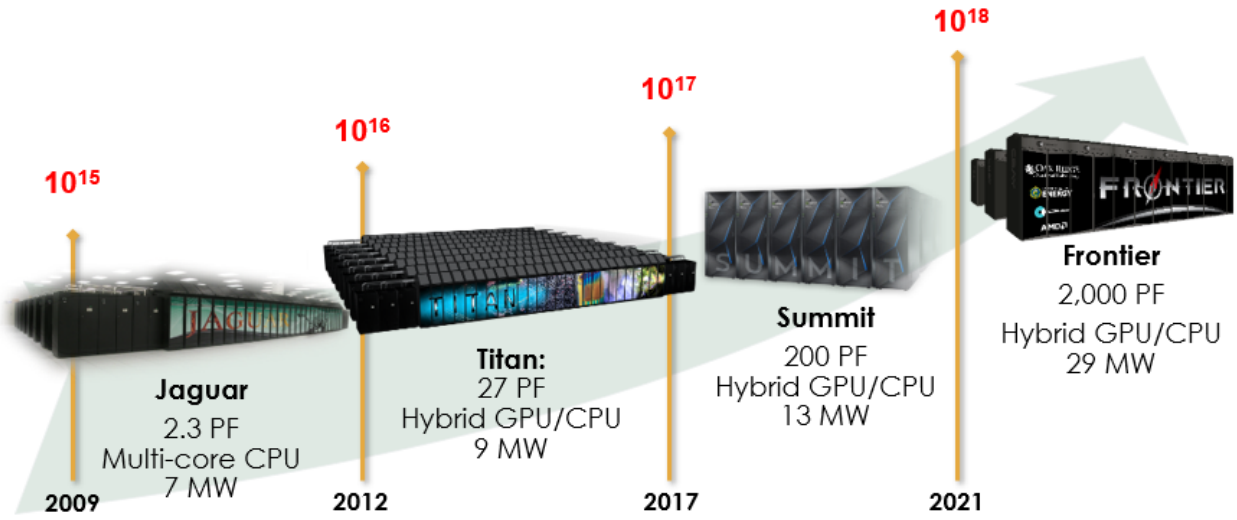

Figure 2: Four Generations of Supercomputers at ORNL

System	Titan (2012)	Summit (2017)	Frontier (2021)
Peak	27 PF	200 PF	2 EF
# nodes	18,688	4,608	9,408
Node	1 AMD Opteron CPU 1 NVIDIA Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 AMD EPYC CPU 4 AMD Radeon Instinct GPUs
Memory	0.6 PB DDR3 + 0.1 PB GDDR	2.4 PB DDR4 + 0.4 HBM + 7.4 PB On-node storage	4.6 PB DDR4 + 4.6 PB HBM2e + 36 PB On-node storage, 66 TB/s Read 62 TB/s Write
On-node interconnect	PCI Gen2 No coherence across the node	NVIDIA NVLINK Coherent memory across the node	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Cray Gemini network 6.4 GB/s	Mellanox Dual-port EDR IB 25 GB/s	Four-port Slingshot network 100 GB/s
Topology	3D Torus	Non-blocking Fat Tree	Dragonfly
Storage	32 PB, 1 TB/s, Lustre Filesystem	250 PB, 2.5 TB/s, IBM Spectrum Scale™ with GPFS™	695 PB HDD+11 PB Flash Performance Tier, 9.4 TB/s and 10 PB Metadata Flash. Lustre
Power	9 MW	13 MW	29 MW

Figure 3: Specifications for ORNL's Titan, Summit, and Frontier Systems

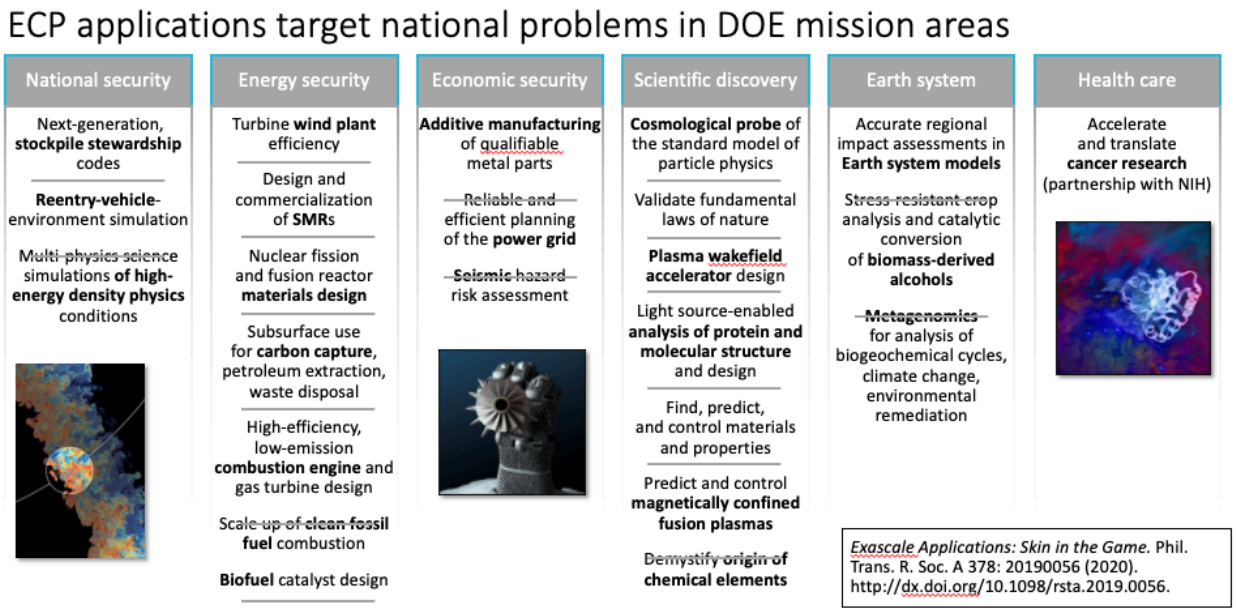

Figure 4: DOE ECP Mission Areas Applications Targeting National Problems

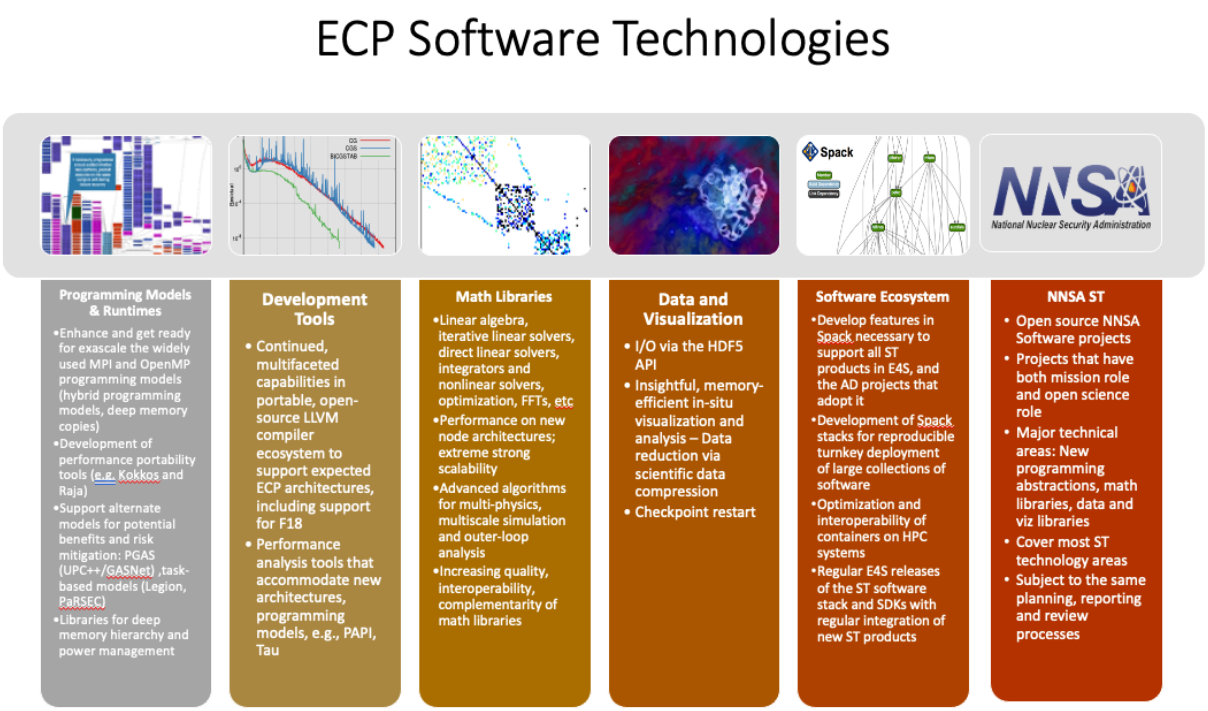

Figure 5: ECP Software Technologies

Frontier System Configuration

The Frontier is based on HPE Cray EX supercomputer and Slingshot interconnect. The Frontier machine includes a total of 9,408 nodes in just 74 cabinets. 73 cabinets with 128 nodes and one partially full cabinet with 64 nodes.

The Frontier node is called the HPE Cray EX235a, containing one customized 3rd Gen AMD EPYC processor and 4 AMD Instinct MI250X GPU accelerators. The amount of memory per node is 1024 GB. The 4 GPUs and 1 CPU are fully connected, i.e., each GPU has an XGMI (AMD Infinity Fabric) link to each of the other GPUs and a link to the CPU. (see Figure 6)

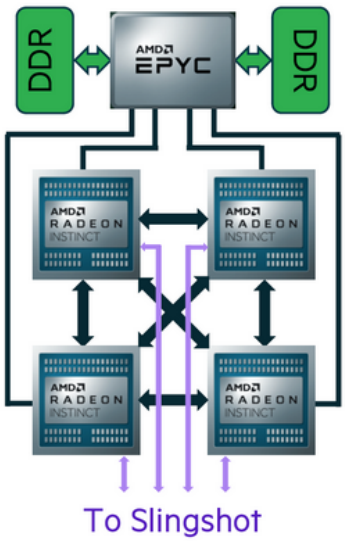

Figure6: Frontier Node containing 1 – AMD EPYC 7A53 CPU & 4 - AMD Instinct MI250X GPUs

System Characteristics	
Node	1 - AMD EPYC 7A53 CPU & 4 - AMD Instinct MI250X GPUs
Nodes per Blade	2
Nodes per Cabinet	128
Nodes per System	9408
Total system memory:	9.2 PB (4.6 PB HBMe2 + 4.6 PB DDR4)
Total on-node NVM	37 PB (66 TB/s read, 62 TB/s write)
Frontier Storage	706 PB (695 PB disk + 11 PB SSD (9.4 TB/s)
Memory Bandwidth between HBM2e and each GPU	3,200 GB/s (3.2 TB/s)
Memory Bandwidth between DDR4 and the CPU	205 GB/s

Bandwidth between CPU and GPU	288 GB/s

The AMD EPYC CPU was customized in two ways for Frontier. First, XGMI links were added to allow coherent shared memory across the CPU and GPUs on Frontier node. ORNL's previous supercomputer Summit has coherent shared memory across the node and the requirement was that this feature also be on Frontier. Second, extra PCIe lanes were needed to allow for 4 TB of on-node Non-Volatile Memory (NVM) on each Frontier node. This local storage gives Frontier an impressive 11 billion IOPS rate for AI and ML and provides node local IO performance of 66 TB/s Read and 62 TB/s Write.

The AMD EPYC CPU has 512 GB of DDR4 memory. It has 64 cores with a nominal cycle time of 2.0 GHz. and can issue 4 instructions per cycle, and can complete 16 64-bit floating point operations per cycle per core. (In terms of FP64 performance, that 64 cores x 16 ops/cycle x 2 GHz equals 2.048 TFlop/s.) Frontier CPU characteristics:

- AMD EPYC 64 core CPU
- 2 Tflop/s peak 64-bit operations
- # of Threads 128
- Base Clock 2.0 GHz
- Total L3 Cache 256MB
- TDP 280W
- CPU Socket SP3
- Any number of cores can boost to 3.5 GHz, but the CPU is not allowed to exceed 280W.
- Per Socket Mem BW 204.8 GB/s
- PCI Express Version PCIe 4.0 x128
- System Memory Type: DDR4
- Memory Channels: 8

Each AMD Instinct MI250X GPU accelerator has 220 cores, and 128 GB HBMe2 memory, 3.2 TB/s memory bandwidth. AMD Instinct MI250X GPU characteristics [hpcw]:

- 128 GB HBMe2 per GPU (512 GB per node at 3.2TB/s)
- 560W per GPU
- 7 nm technology
-
- 64-bit, double-precision FP64: 47.9 TFLOP/s
- 64-bit, double precision matrix FP64: 256 flops/cycle, 95.7 TFLOP/s
- 64-bit, double precision vector FP64: 128 flops/cycle, 47.9 TFLOP/s
-
- 32-bit, single precision FP: 47.9 TFLOP/s
- 32-bit, single precision matrix FP32: 256 flops/cycle, 95.7 TFLOP/s
- 32-bit, single precision vector FP32: 128 flops/cycle, 47.9 TFLOP/s
-
- 16-bit, IEEE half precision matrix FP16: 1024 flops/cycle, 383 TFLOP/s
- 16-bit, Bfloat half precision matrix BF16: 1024 flops/cycle, 383 TFLOP/s
-

- 8-bit integer (AI Inference): 276 Tops
- 8-bit integer: 1024 ops/cycle

An advantage in the Frontier design is that the communication NIC(s) are not hung off the CPU as was done with Summit and Titan. Hanging the NIC off the CPU creates a bottleneck for the data coming from the fast GPUs by forcing it to go over the CPU-GPU link. (see Figure 11)

In Frontier, every GPU has a direct link to its own NIC. A Frontier node has 4 NICS each connected to its own GPU. As a result, the fast HBM memory and data on the GPUs are much closer to network. An issue is if the CPU wants to send a message it has to route out through one of the 4 GPUs. There is a direct message bypass thru the GPU for this so the GPUs are not interrupted.

Frontier Interconnect

Frontier's Slingshot11 network provides adaptive routing, congestion management and quality of service. Frontier is configured with a three-hop dragonfly topology(Figure 7), but Slingshot supports any number of topologies such as flattened butterflies and fat trees. The use of the dragonfly topology is largely motivated by cost. It does so through the reduction of long global channels. The fewer the long optical cables, the less expensive the system is. HPE claims that up to 90% of the cables in the system are inexpensive copper cables with only 10% optical.

The interconnect provides a bisection bandwidth of 540 TB/s. Each compute node has four Cassini NICs, each providing 200 Gb/s (25 GB/s) for a total outbound injection bandwidth of 800 Gb/s (100 GB/s). The interconnect connects all the compute and storage cabinets. i.e. the compute nodes, I/O nodes, and front-end nodes.

Slingshot diverges from prior interconnects in that it embraces Ethernet as the baseline interconnect. The Slingshot switch first operates using the standard Ethernet protocols but will then try to negotiate the advanced 'HPC Ethernet' features when a connected device supports them. The intention here is to allow the advanced HPC Ethernet features to work within the network of devices (e.g., other Slingshot switches) that supports it while being completely interoperable with Ethernet devices that do not.

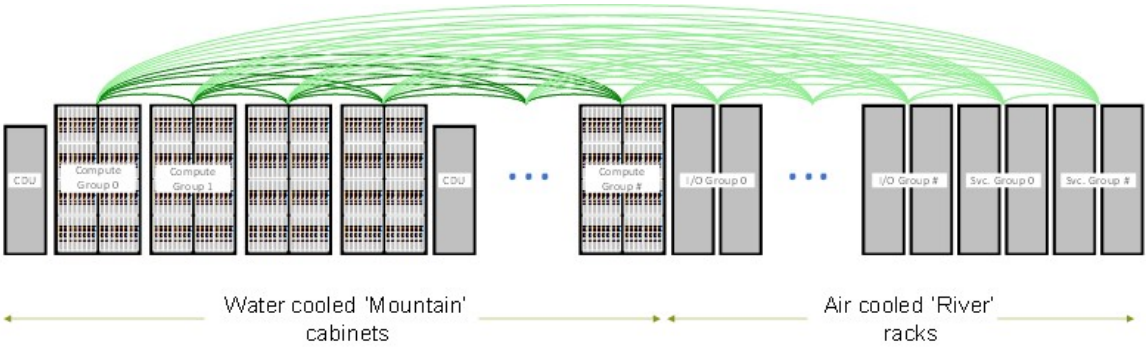

Figure7: Cabinet Interconnect

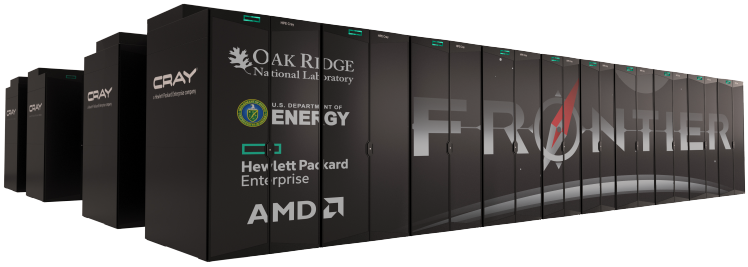

Figure 8: Artist Rendering of DOE's Frontier System

Figure 9: Frontier's Front 12 cabinets

Figure 10: Frontier (with the doors off) has 3 computer cabinets for every cooling unit (CDU)

Front of the cabinet shows all water cooling lines. Red and Blue lines are the hot and cold water lines. The silver marks above the cooling lines are the removal handle for each node.

Figure 11: Read View with Doors Off; Rear of cabinet has power and all the interconnect wiring.

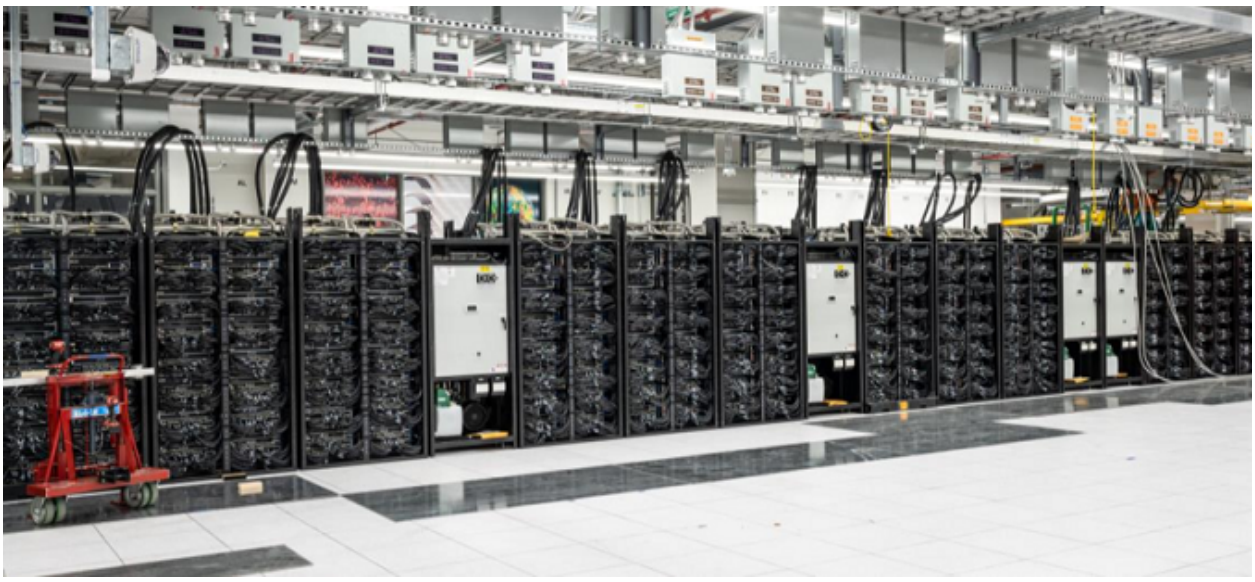

Figure 12: One row of the Frontier System as Installed at DOE's ORNL

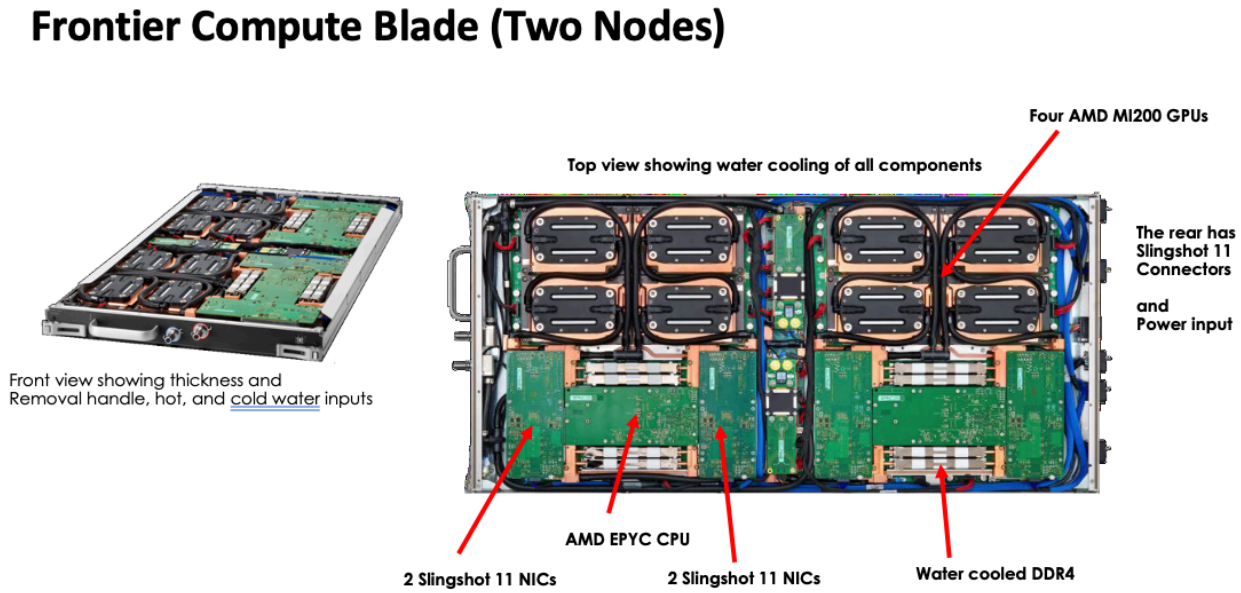

Figure13: Frontier Compute Blade made up of 2 nodes

Figure14: One Rack Containing 64 Blades (128 nodes)

Figure 15: Frontier 12 Cabinets Wide = 1526 Nodes

(Full Frontier System has six rows of 12 cabinets plus a seventh row of 3 cabinets)

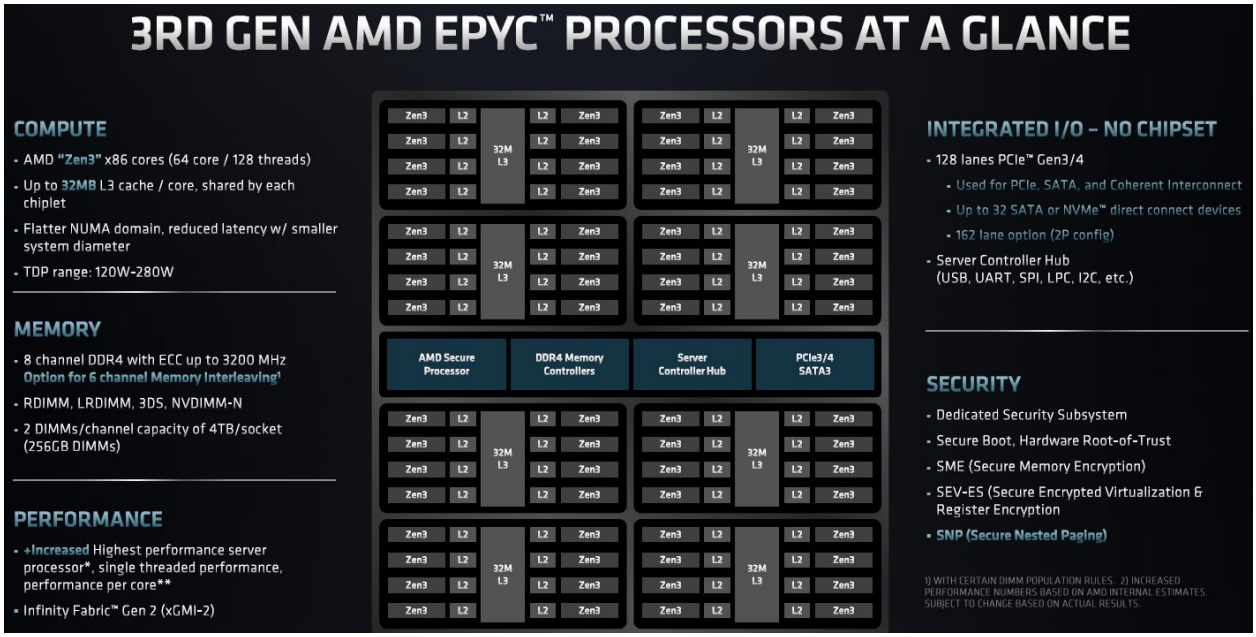

Figure 6**: AMD 3rd Generation EPYC Processor See: https://www.servethehome.com/wp-content/uploads/2021/03/AMD-EPYC-7003-Zen-3-SoC-at-a-Glance.jpg and https://en.wikichip.org/wiki/amd/epyc/7763**

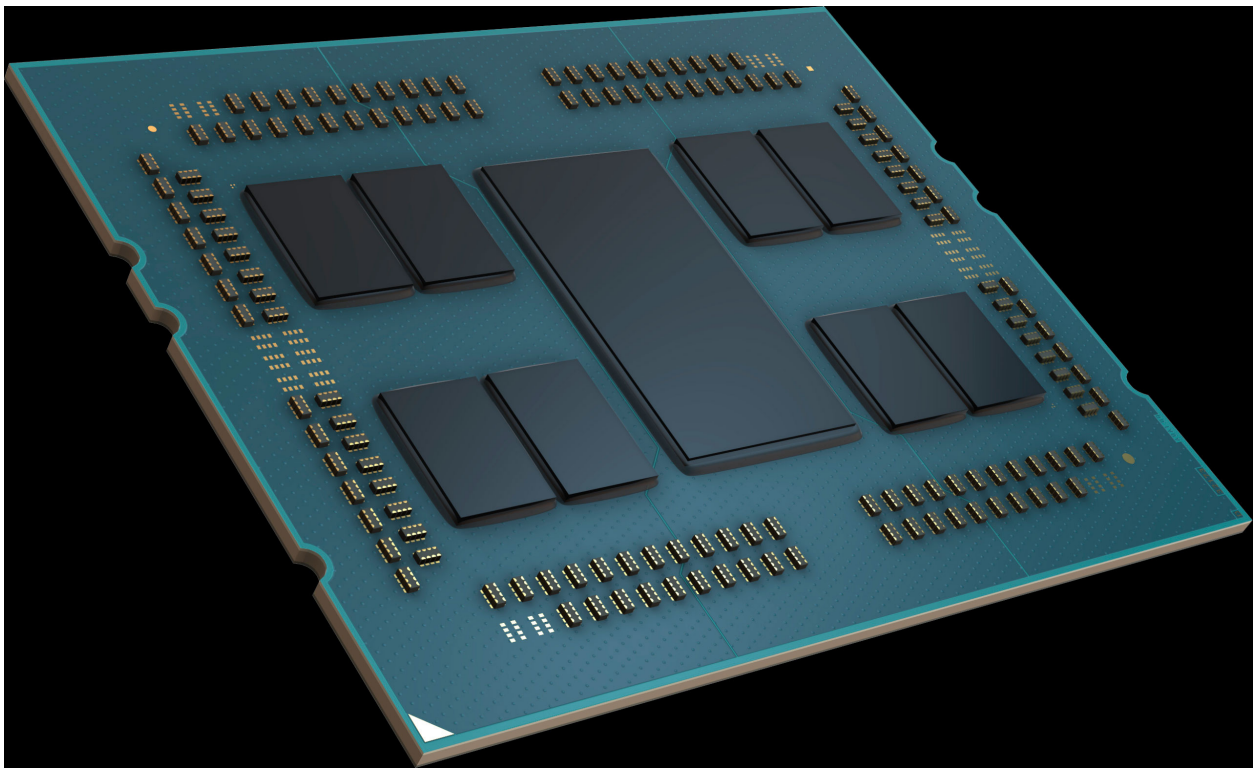

Figure 7: AMD EPYC 7A53 CPU

Storage System

The storage system consists of three primary layers:

- 1st Layer
  - Cache for global file system
  - Temporary file systems
    - Local file system for compute node
    - Shared file system for a job
- 2nd Layer
  - Lustre-based global file system

Frontier Storage Configuration

The Frontier storage was developed by ORNL and Cray to put together a new multi-tiered Cray Storage System that requires many fewer racks and optical cables than originally proposed.

Multitier I/O Subsystem	Read	Write
37 PB Node Local Storage	65.9 TB/s	62.1 TB/s
	11 Billion IOPS	
11 PB Performance Tier	9.4 TB/s	9.4 TB/s
652 PB Capacity tier	5.2 TB/s	4.4 TB/s
10 PB Metadata	2M Transactions per second	

There are two - 2 TB SSD NVM per node of local storage (Flash).

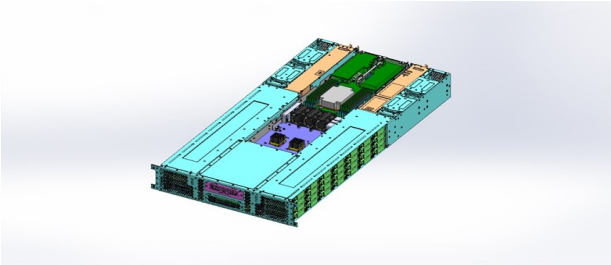

Figure 8: Gazelle SSD Storage board (Performance Tier and Metadata)

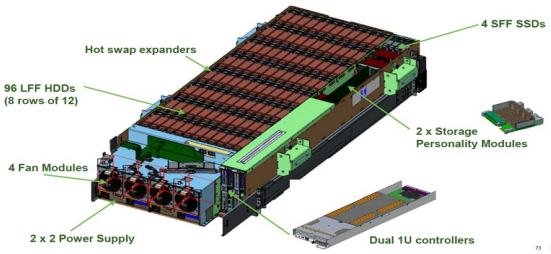

Figure 9: Moose HDD Storage board (Capacity Tier)

Peak Performance

	Frontier Normal Mode (2.0 GHz)	**Summit Computer**
Peak Double Precision (64 bit)	2.013 EFLOP/s	200 PFLOP/s
Peak Single Precision (32 bit)	2.446 EFLOP/s	400 PFLOP/s
Half Precision (16-bit float, IEEE standard)	11.21 EFLOP/s	3.3 EFLOP/s
Integer (8 bit)	3.9 EOP/s	1.1 EOP/s
Total Memory	4.85 PB	1.76 PB
Total Memory Bandwidth	163 PB/s	? TB/s

Figure 10: Peak Performance

High Performance LINPACK (HPL) Benchmark

The Frontier Linpack Benchmark shown in Figure 22 is for a run of the HPL benchmark program that achieved 1.102 Eflop/s out of a theoretical peak of 1.665 Eflop/s or an efficiency of 66% of theoretical peak performance. The run took a little under 2.5 hours to complete, with an average power consumption of 21.1 MW. The run used 9248 nodes and used only the 512 GB HBM on each node for a total of 4.7 PB.

Summary of HPL Benchmark run:

- HPL number = 1.102 Eflop/s

- 66% efficient (peak at 1.665 Eflop/s)
- Size of the matrix, n = 24,440,832 (4.77 PB)
- Logical process grid of pxq = 136 x 544
- Time to complete benchmark run: 8,828.74 seconds (2.45 hours)
- Average 21.3 MW
- 51.8 Gflops/W

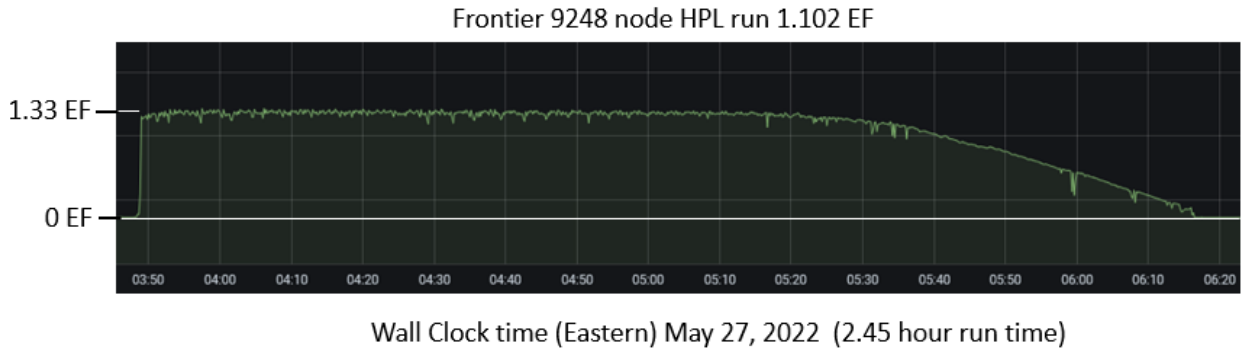

Figure 22: ***Performance profile for the HPL benchmark run.***

HPL-AI Benchmark

Frontier was able to achieve 6.86 Eflop/s running the HPL-AI benchmark program. The run used 9248 nodes and completed in 2 hours 11 minutes. The size of the matrix was, N = 27,852,800; The block size used was NB =2560. The PMAP used Node Grid - 2x4C with P = 272 and Q = 272.

The Software Stack

Frontier will support multiple compilers, programming models, and tools. Below are the compilers, programming languages and models, and additional tools that are available on Frontier.

Compilers	Cray PE
	AMD ROCm
	GCC
Programming Languages and Models Supported	C, C++, Fortran (for all compilers)
	OpenMP 5.x (Cray, AMD, and possibly GCC compilers)

	Cray MPI
	UPC (Cray and GCC compilers)
	Coarray Fortran, Coarrary C++ (Cray compilers)
	AMD HIP
	Chapel
	Global Arrays
	Charm++
	GASNet
	OpenSHMEM
System-level Programming Tools	CrayPat/Apprentice2
	Cray Reveal
	Open | SpeedShop
	TAU
	HPCToolkit
	Score-P
	VAMPIR
Node-level Programming Tools	GNU gprof
	PAPI
	AMD ROCProfiler
Debugging and Correctness Tools	ARM DDT
	Cray CCDB
	Stack Trace Analysis Tool
	Cray GDB4HPC
	gdb
	AMD ROCm debug service and GDB-MMI
	Cray Abnormal Termination Processing
GUI and Visualization APIs, I/O Libraries	X11
	Motif
	Qt
	NX, NeatX, or similiar
	NetCDF
	HDF5

Figure 23: Programming Environment

The Extreme-scale Scientific Software Stack[E4S] (E4S), supported by the DOE through the ECP, is a community effort to provide open source software packages for developing, deploying, and running scientific applications on HPC platforms. E4S aims to deliver a modular, interoperable, and deployable software stack based on the Spack package manager. E4S provides both source builds for native, bare-metal installations, as well as containers of a broad collection of software packages for secure, reproducible, container-based deployments. E4S exists to accelerate the development, deployment, and use of HPC software, lowering the barriers for HPC developers and users. E4S builds upon Software Development Kits (SDKs), which are collections of related software products and packages where coordination across package teams improves usability and practices, while fostering community growth among teams that develop similar and complementary capabilities.

E4S itself is a curated release of ECP Software Technology (ST) products built using Spack. Spack is an open-source package manager geared towards simplified, reproducible builds of the otherwise complicated dependency chains common in HPC software; Spack provides the ability to leverage existing compilers and runtime system libraries for native software installations. E4S currently includes over 50 unique ST products, spanning programming models and runtimes, math libraries, compilers, and tools for performance evaluation, data management, and visualization. E4S supports GPUs from multiple vendors and includes their supporting runtime libraries such as CUDA and ROCm. E4S also provides a validation test suite that helps build, execute, and validate these products.

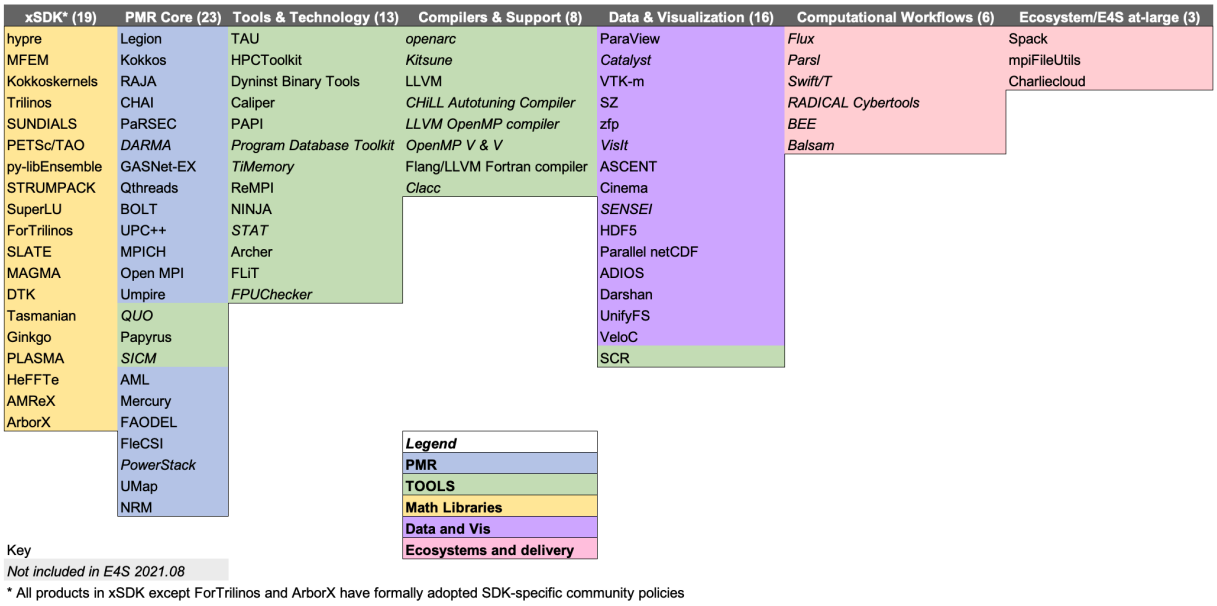

Figure 24: ECP Software Stack

Frontier users can easily use pre-installed packages and build packages based on Spack recipes.

Frontier will use the HPE Everest HPE Performance Cluster Manager (HPCM) System Software Stack for system configuration. HPCM is a fully featured cluster management suite that is responsible for the life cycle management of a HPC system. The software provides tools for switch management, image curation and provisioning, monitoring and health management, and cluster setup. HPCM is a more traditional cluster management suite of tools that has adopted some modern aspects of system monitoring and metric collecting through tools like Kafka, ELK, and Alerta. HPCM has a decade long history of supporting HPE and SGI clusters.

Applications Development

The ECP project is targeting 6 major application areas [kothe].

Chemistry and Materials

This area focuses on simulation capabilities that attempt to precisely describe the underlying properties of matter needed to optimize and control the design of new materials and energy technologies. These applications require the use of sophisticated models and algorithms to solve complex physics equations.

Co-design

These projects target crosscutting algorithmic methods that capture the most common patterns of computation and communication (known as motifs) in the ECP applications. The goal of the co-design activity is to integrate the rapidly developing software stack with emerging hardware technologies while developing software components that embody the most common application motifs.

Data Analytics and Optimization

This is an emerging area whose predictive capability is partially based on modern data analysis and machine learning techniques rather than strictly on approximate solutions to equations that state fundamental physical principles or reduced semiempirical models. This activity encompasses a broad range of research areas and techniques, some of which are only recently coming into maturity in the context of high-end simulation.

Earth and Space Science

The research in this area spans fundamental scientific questions, from the origin of the universe and chemical elements to planetary processes and interactions affecting life and longevity. These application areas treat phenomena where controlled and fine resolution data collection is extremely difficult or infeasible, and, in many cases, fundamental simulations are our best source of data to confirm scientific observations.

Energy

Energy applications focus on the modeling and simulation of existing and future technologies for the efficient and responsible production of energy to meet the growing needs of the United States. These applications generally require detailed modeling of complex facilities and multiple coupled physical processes. Their goal is to help overcome obstacles.

National Security

The focus of the National Security Applications is to deliver comprehensive science-based computational weapons applications able to provide, through effective exploitation of exascale HPC technologies, breakthrough modeling and simulation solutions that yield high-confidence insights into at least three currently intractable problems of interest to the NNSA Stockpile Stewardship Program (SSP).

ECP Applications

ECP applications, including both their technologies and solutions, will be far-reaching for decades to come, include:

- Predictive microstructural evolution of novel chemicals and materials for energy applications.
- Robust and selective design of catalysts an order of magnitude more efficient at temperatures hundreds of degrees lower.
- Accelerate the widespread adoption of additive manufacturing by enabling the routine fabrication of qualifiable metal alloy parts.
- Design next-generation quantum materials from first principles with predictive accuracy.
- Predict properties of light nuclei with less than 1% uncertainty from first principles.
- Harden wind plant design and layout against energy loss susceptibility, allowing higher penetration of wind energy.
- Demonstrate commercial-scale transformation energy technologies that curb fossil fuel plant CO2 emission by 2030.
- Accelerate the design and commercialization of small and micronuclear reactors.
- Provide a 'whole device' modelling capability for magnetically confined fusion plasmas required to design and operate ITER and future fusion reactors.
- Address fundamental science questions such as the origin of elements in the universe, the behavior of matter at extreme densities, the source of gravity waves; and demystify key unknowns in the dynamics of the universe (dark matter, dark energy and inflation).
- Reduce the current major uncertainties in earthquake hazard and risk assessments to ensure the safest and most cost-effective seismic designs.
- Reliably guide safe long-term consequential decisions about carbon storage and sequestration.
- Forecast, with confidence, water resource availability, food supply changes and severe weather probabilities in our complex earth system environment.
- Optimize power grid planning and secure operation with very high reliability within narrow operating voltage and frequency ranges.
- Develop treatment strategies and pre-clinical cancer drug response models and mechanisms for RAS/RAF-driven cancers.
- Discover, through metagenomics analysis, knowledge useful for environment remediation and the manufacture of novel chemicals and medicines.

- Dramatically cut the cost and size of advanced particle accelerators for various applications impacting our lives, from sterilizing food of toxic waste, implanting ions in semiconductors, developing new drugs or treating cancer.

Summary

The DOE ORNL Frontier system is very impressive with over 8.8 million cores and a peak performance of 2 EFLOP/s in 64-bit floating point for standard scientific computations and 11.2 EFLOP/s in IEEE 16-bit floating point for machine-learning applications. The Frontier system is almost three times (10×) faster than the system it replaces, Summit. The HPL benchmark result at 1.102 EFLOP/s, or 66% of theoretical peak performance, is also impressive with an efficiency of 52 GFLOP/s per watt. The ratio of floating-point operations per words from memory is 131 double-precision FLOP/s per Word transferred from HBM memory (214 TFLOPS/s per 13 TB/s or 214/1.62 flops/word), which shows a good balance for floating point operations to data transfer from memory. One would expect good performance on computational science problems and machine-learning applications.

Appendix A.

Table A-1. Comparison with top machines on the TOP500

	ORNL Frontier	RIKEN Fugaku	ORNL Summit	Sunway TaihuLight	TianHe-2A
Theoretical Peak	2,000 PFLOP/s	514 PFLOP/s	200 PFLOP/s = (.54*2 CPU + 6*7 Accelerator)	125.4 PFLOP/s = CPEs + MPEs Cores per Node = 256 CPEs + 4 MPEs Supernode = 256 Nodes System = 160 Supernodes Cores = 260 * 256 * 160 = 10.6M	94.97 PFLOP/s = (7.52 CPU + 87.45 Accelerator) PFLOP/s
HPL Benchmark FLOP/s	1.102 EFLOP/s	415 PFLOP/s	149 PFLOP/s	93 PFLOP/s	13.987 PFLOP/s out of a theoretical peak of 21.86 PFLOP/s.
HPL % peak	66%	81%	74%	74.16%	63.98%
HPCG benchmark	Not run yet	13 PFLOP/s	2.92 PFLOP/s	.371 PFLOP/s	4096 nodes .0798 PFLOP/s
HPCG % peak	Not run yet	2.8%	1.5%	0.30%	0.365%
Compute nodes	9,408	152,064 (This is on 96% of the full system)	4608 = 256 cabinets * 18 nodes/cabinet	40,960	17,792
Node	Optimized 3rd Gen AMD EPYC (2 TFLOP/s, 64 cores) Plus 4 AMD Instinct MI250X GPUs / node (53 TFLOP/s, 220 cores each)	48 cores	2 IBM POWER9 CPUs 3.07 GHz plus 6 Nvidia V100 (.54 TFLOP/s each) Tesla GPUs / node (7 TFLOP/s each)	256 CPEs + 4 MPEs	2 – Intel Ivy Bridge (12 cores, 2.2 GHz) plus 2 Matrix-2000, 1.2 GHz)
Node peak performance	214 TFLOP/s (2 CPU + 4*53 GPU) TFLOP/s	3.4 TFLOP/s	43 TFLOP/s = (1.08 CPU + 42 GPU) TFLOP/s	12	5.3376 TFLOP/s = (2 × 211.2 CPU + 2 × 2.4576 Accelerator) TFLOP/s
Node memory	1024 GB	32 GB HMB2	32 GB CPU + 6 GB GPU	32 GB per node	64 GB CPU + 128 GB Accelerator
System memory	9,408 nodes*1,024 GB/node = 9.6 PB	4.85 PB	1.76 PB = 4608*600 GB of coherent memory (6×16 = 96 GB HBM2 plus 2×8×32 = 512 GB DDR4 SDRAM)	1.31 PB (32 GB × 40,960 nodes)	3.4 PB = 17,792 × (64 GB + 128 GB)
Configuration	9,408 nodes in 74 cabinets	158,976 nodes	256 Racks × 18 Nodes	Node peak performance is 3.06 TFLOP/s, or 11.7 GFLOP/s per core. 260 cores/node	2 Nodes per blade, 16 blades per frame and 4 frames per cabinet and 139 cabinets in the system.

				CPE: 8 FLOPs/core/cycle (1.45 GHz × 8 × 256 = 2.969 TFLOP/s) MPE (2 pipelines) 2 × 4 × 8 FLOPs/core/cycle (1.45 GHz × 1= 0.0928TFLOP/s) Node peak performance: 3.06 TFLOP/s 1 thread/core Nodes connected using PCI-E The topology is Sunway network. 256 nodes = a supernode (256 × 3.06 TFLOP/s = . 783 PFLOP/s) 160 supernodes make up the whole system (125.4PFLOP/s) The network system consists of three different levels, with the central switching network at the top, the super node network in the middle, and the resource-sharing network at the bottom. 4 SNs = cabinet Each cabinet ~3.164 PFLOP/s 256 nodes per SN 1,024 nodes (3 TFLOP/s each) per cabinet 40 cabinets ~125 PFLOP/s	
Total system	(64 CPU cores + 4*220 GPU cores) * 9408 nodes = 8,881,152 cores	7,630,848 = 158,976 * 48 cores	2,397,824 cores	10,649,600 cores = Node (260) × supernodes(256 nodes) × 160 supernodes	4,981,760 cores = (17,792 × 2 Ivy Bridge with 12 cores) + (2 × Matrix-2000 × 128)
Power (processors, memory, interconnect)	29 MW	28.33 MW (7.33*3.863)	11 MW	15.3 MW	16.9 MW for full system
Footprint	370 m^2 Each Cabinet weigh is 8,000 pounds without water.	1,920 m^2	520 m^2	605 m^2	400 m^2 (50 m^2/line*8) or 720 m^2 total room

Appendix B. References

[E4S] https://e4s-project.github.io/

[kothe] Kothe D, Lee S, Qualters I. 2019 Exascale computing in the United States. **Comput. Sci. Eng. 21**, 17–29. (doi:10.1109/MCSE.2018.2875366)

[kothe] Alexander Francis, Almgren Ann, Bell John, Bhattacharjee Amitava, Chen Jacqueline, Colella Phil, Daniel David, DeSlippe Jack, Diachin Lori, Draeger Erik, Dubey Anshu, Dunning Thom, Evans Thomas, Foster Ian, Francois Marianne, Germann Tim, Gordon Mark, Habib Salman, Halappanavar Mahantesh, Hamilton Steven, Hart William, (Henry) Huang Zhenyu, Hungerford Aimee, Kasen Daniel, Kent Paul R. C., Kolev Tzanio, Kothe Douglas B., Kronfeld Andreas, Luo Ye, Mackenzie Paul, McCallen David, Messer Bronson, Mniszewski Sue, Oehmen Chris, Perazzo Amedeo, Perez Danny, Richards David, Rider William J., Rieben Rob, Roche Kenneth, Siegel Andrew, Sprague Michael, Steefel Carl, Stevens Rick, Syamlal Madhava, Taylor Mark, Turner John, Vay Jean-Luc, Voter Artur F., Windus Theresa L. and Yelick Katherine, 2020, Exascale Applications: Skin in the Game, Phil. Trans. R. Soc. A. https://doi.org/10.1098/rsta.2019.0056

[top10] Lucas, Robert, Ang, James, Bergman, Keren, Borkar, Shekhar, Carlson, William, Carrington, Laura, Chiu, George, Colwell, Robert, Dally, William, Dongarra, Jack, Geist, Al, Haring, Rud, Hittinger, Jeffrey, Hoisie, Adolfy, Klein, Dean Micron, Kogge, Peter, Lethin, Richard, Sarkar, Vivek, Schreiber, Robert, Shalf, John, Sterling, Thomas, Stevens, Rick, Bashor, Jon, Brightwell, Ron, Coteus, Paul, Debenedictus, Erik, Hiller, Jon, Kim, K. H., Langston, Harper, Murphy, Richard Micron, Webster, Clayton, Wild, Stefan, Grider, Gary, Ross, Rob, Leyffer, Sven, and Laros III, James. *DOE Advanced Scientific Computing Advisory Subcommittee (ASCAC) Report: Top Ten Exascale Research Challenges*. United States: N. p., 2014. Web. doi:10.2172/1222713.

[hpcw] AMD Launches Milan-X CPU with 3D V-Cache and Multichip Instinct MI200 GPU By Tiffany Trader https://www.hpcwire.com/2021/11/08/amd-launches-milanx-cpu-with-v-cache-and-multichip-mi200-gpu/

https://www.olcf.ornl.gov/frontier/
https://technewsources.com/us-closes-in-on-exascale-frontier-supercomputer-installation-is-underway-hpe-cray/
https://www.redsharknews.com/technology-computing/item/7065-amd-announces-details-of-el-capitan-supercomputer-with-2-exaflops-of-compute-power
https://www.olcf.ornl.gov/frontier/
https://www.hpcwire.com/2021/11/08/amd-launches-milanx-cpu-with-v-cache-and-multichip-mi200-gpu/