Frontier – World's First ExaFLOPS Supercomputer*

V. Rajaraman

A race began around 2010 among many countries to build a supercomputer that would breach the ExaFLOPS (10^{18} floating-point operations per second) barrier, a formidable goal. Frontier, a supercomputer designed by a team of engineers from Cray-HPE and the Oak Ridge National Laboratory, achieved this goal in May 2022. It executed the HPLinpack benchmark program at a speed of 1.102 ExaFLOPS. In this article, I trace the events leading up to the design of Frontier and its architecture. I also describe various benchmarks used to compare the speeds of high-performance computers.



V. Rajaraman is at the Indian Institute of Science, Bengaluru. Several generations of scientists and engineers in India have learnt computer science using his lucidly written textbooks on programming and computer fundamentals. His current research interests are parallel computing and history of computing.

Introduction

Ever since supercomputers (now called high-performance computers (HPC)) were built, it was the ambition of computer scientists to build an HPC that would breach the PetaFLOPS barrier, namely, designing an HPC with a speed of 10^{15} floating-point operations per second (FLOPS). IBM built a computer named Roadrunner that breached the PetaFLOPS barrier in 2008. When this was achieved, the goal of HPC designers was to breach the next barrier – ExaFLOPS – a thousand times PetaFLOPS, that is 10^{18} (one followed by 18 zeros or a quintillion) FLOPS. (If each person in the world adds two numbers in 1 second, it will take them 4 years to add what an ExaFLOP computer would add in 1 second!). The challenge was to build such an HPC which will consume reasonable power. Many countries joined the race to build an ExaFLOPS computer. Among them were the US, the UK, Germany, Japan, and China. Building an ExaFLOPS computer was considered important as such machines would be able to solve

Keywords
Exascale computers, Frontier HPC architecture, HPC benchmarks, GPGPUs in HPC.

*Vol.28, No.4, DOI: https://doi.org/10.1007/s12045-023-1583-7

Building an ExaFLOPS computer was considered important as such machines would be able to solve many problems such as short-term weather prediction, aid research in materials science at the molecular level, predict earthquakes, design personalised drugs, design new drugs for treating cancer, study stellar phenomenon, find effective vaccines for new viruses, and a myriad of other hitherto intractable problems.

many problems such as short-term weather prediction, aid research in materials science at the molecular level, predict earthquakes, design personalised drugs, design new drugs for treating cancer, study stellar phenomenon, find effective vaccines for new viruses, and a myriad of other hitherto intractable problems. An ExaFLOPS computer would be able to reduce the time to obtain solutions to such problems from weeks to hours. Computer scientists also wanted to combine modelling and simulation with big data analytics and machine learning using ExaFLOPS HPCs which, it was felt, would lead to new insights and inventions.

As building the first ExaFLOPS computer would demonstrate the technological leadership of a country, the US government decided to allocate resources to ensure that it was the first to achieve this goal.

On 29 July 2015, US President Barack Obama issued an executive order titled "National Strategic Computing Initiative" [1] with the following primary goals:

1. Accelerate the delivery of an ExaFLOPS computer.

2. Unify technologies used in modelling and simulation and data analytics.

3. Forge a path to continue increasing the speed of HPCs even after the current semiconductor technology reaches its performance limit.

4. Continue research and development in HPC by supporting work on relevant areas such as networking technology, workflow, algorithums, and software.

5. Ensure accessibility to HPCs for researchers and develop human resources in HPC and related areas.

6. Promote public-private collaboration so that the results of the research and development are shared with industries and academia.

The Exascale Computer Project

As a follow-up, President Obama assigned the major task of fulfilling the strategic objectives to the Department of Energy (DOE) of the US government. In 2016, DOE started a programme called the Exascale Computing Project (ECP) [2]. The National Nuclear Safety Administration (NNSA) and the Office of Science (OS) were given the responsibility of carrying out the project. NNSA was primarily responsible for nuclear weapons development and maintenance and the Office of Science for funding basic science and engineering research in HPC development and applications. Exascale Computer Project was planned as a seven-year project with the goals of advancing not only the hardware art to build an exascale computer but also developing the system and applications software that would enable the optimal use of exascale computers.

The challenge was to build an exascale computer that would consume reasonable power as power dissipated increases with speed. The power budget was set at 20 MW, and this was not an easy goal. The plan started with numerous workshops to explore areas for collaborative research between academia, research laboratories, and industries. These workshops aimed to examine the benefits of exascale computers, their potential applications, and the challenges of building the complete ecosystem (high-speed compute nodes, interconnection network to connect compute nodes, system software stack, and scalable applications) for the exascale era. Exascale Computer Project is a collaboration of the six premier US government laboratories – Argonne National Laboratory (ANL), Los Alamos National Laboratory (LANL), Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), Oak Ridge National Laboratory (ORNL), and Sandia National Laboratory (SNL). Initially, three exascale computers were planned. The first called "Aurora" was to be designed by Intel and ANL and located at Argonne, the second called "Frontier", to be designed by Cray-HPE and ORNL and located at Oakridge, and the third called "El Capitan" designed by Cray-HPE and LLNL and located at Livermore.

The plan started with numerous workshops to explore areas for collaborative research between academia, research laboratories, and industries. These workshops aimed to examine the benefits of exascale computers, their potential applications, and the challenges of building the complete ecosystem (high-speed compute nodes, interconnection network to connect compute nodes, system software stack, and scalable applications) for the exascale era.

Aurora and Frontier were to be used primarily for research in the frontier areas of science and technology and El Capitan for weapons research in addition to scientific uses. Intel ran into difficulties in fabricating the multi-core microprocessor to be used by Aurora and was delayed. The Cray-HPE system at ORNL was the first to be completed and broke the exascale barrier in May 2022, achieving a speed of 1.102 ExaFLOPS executing a double-precision (64-bits) Linpack benchmark program and a peak speed of 1.68 ExaFLOPS using double-precision operands. It topped the June 2022 TOP500 [3] list (See *Box* 1) and its speed of 1.102 ExaFLOPS was more than double that of the next computer in the list, Fugaku of Japan, whose speed was 442 PetaFLOPS. On the HPL-MXP benchmark that uses mixed precision programs (32-bits, 24-bits, and 16-bits) suitable for machine learning programs, Frontier's speed was 7.94 ExaFLOPS. Besides this, it was also the second most power-efficient computer in the world in the June 2022 Green500 list with a power efficiency score of 52.23 GFLOPS per watt.

Box 1. HPC Benchmarks

Benchmarks are computer programs designed to measure the performance of computers. Ideally, a customer planning to buy an HPC would collect a representative set of programs from prospective users and request competing vendors to run them on their HPC and give the results and the time taken to get the results. This is used to compare the performance of HPCs. This method is individualised, costly, and takes time. One would like to have a publicly available measure of the performance of HPCs in the market. At the International Supercomputer Conference held at Mannheim in Germany in 1992, the general chairman of the Conference, (Late) Professor Hans Meuer of the University of Mannheim, broached this topic and proposed that a publicly available list of the fastest HPCs in the world be published. Professor Jack Dongarra of the University of Tennessee, USA, was well known for steering a group to standardise the Message Passing Interface (MPI) for message-passing parallel computers and a benchmark program written in Fortran he had developed called the Linpack benchmark to evaluate the performance of supercomputers. This program solves a linear system of equations $\mathbf{A\ x = b}$, where $\mathbf{A}$ is a dense matrix with random numbers as coefficients, $\mathbf{x}$ is a vector of variables, and $\mathbf{b}$ is a vector of constants, using double-precision (64-bits) calculation. Dongarra chose this problem as it occurs frequently in scientific and engineering computations. The size of the matrix is variable. However, nowadays, 1000 *times* 1000 matrix and HPLinpack, a version

Contd.

Box 2. ***Contd.***

written for parallel computers are used. (Vendors may use the maximum size of the matrix optimized for their HPC). The speed of 500 HPCs running this benchmark in descending order was first published in June 1993 using data available from the Internet. It was later formalised, and vendors or users who wanted to be listed were asked to submit HPLinpack data on their computers. The list is published in June and November each year. (See top500.org). Besides the name of the computer and where it is installed, it gives the maximum number of double-precision (64-bit operands) floating-point operations per second (nowadays as PetaFLOPS) performed by the HPC solving the HPLinpack benchmark. It also lists the theoretical peak floating-point speed of the HPC and the power dissipated by it. If a vendor or a user does not provide the data, the computer is not listed.

The HPLinpack (HPL) benchmark is criticised as it gives a higher weightage to the CPU speed and not enough weightage to the memory bandwidth and the interconnection network. Consequently, an additional benchmark known as the HPCG (high-performance conjugate gradient) was proposed in 2013 by Dongarra and his co-workers. It uses sparse matrices and iterative solution methods that often occur in solving sets of non-linear partial differential equations. This benchmark is better at simulating data access patterns of applications and gives a better weightage to the speed of memory access and interconnection systems of current HPCs. Since 2013, the HPCG speed is also listed in the TOP500 list.

Currently, HPCs are used not only for solving numeric-intensive science and engineering problems but also problems related to the training of multi-layer neural networks that use 32-bit and 16-bit data. A benchmark called the HPL-MXP (mixed precision) is currently proposed. It relaxes the requirement of 64 bits throughout the computation of the HPL benchmark and allows the use of lower precision (32 or 16 bits) in relevant steps of a computation. Thus the HPL-MXP values will be higher than the values obtained by the HPL benchmark. The HPL-MXP benchmark results are listed at `www.hpl-mxp.org.`

HPCs have been power guzzlers. Given the current concern of reducing global warming, designers have paid more attention to reducing the power consumption of HPCs. Thus vendors of HPCs are requested to provide another metric that gives the GFLOPS per watt and is published in the TOPS500 listing as Green500. The most power-efficient computer is at the top of the Green500 list.

The Architecture of the Frontier HPC [4]

Each computing node of the Frontier HPC consists of an AMD EPYC "Trento" 64-core processor with a 2 GHz clock to perform general computation (CPU) with four AMD Radeon Instinct Mi250X general purpose graphics processing units (GPGPU) to perform high-speed floating-point operations. The GPGPU chip

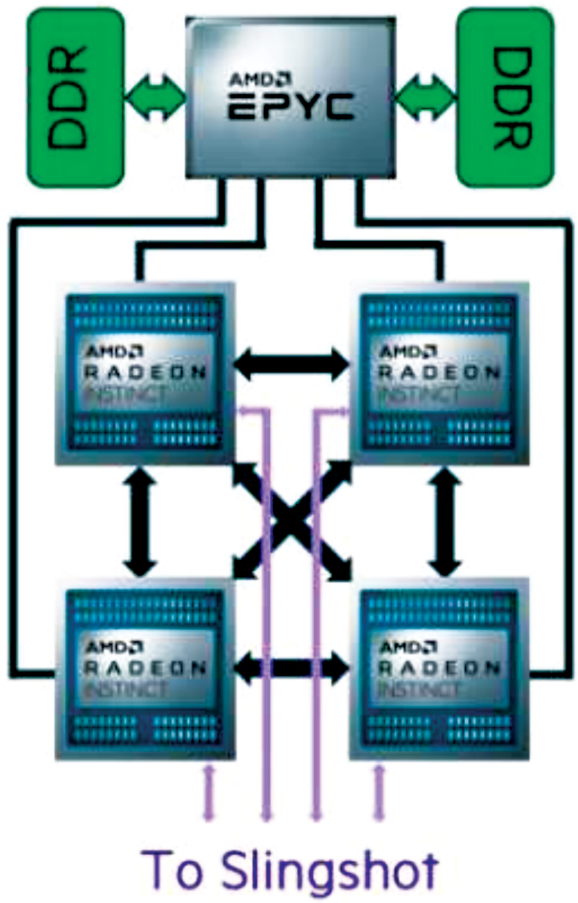

Figure 1. A Compute Node of Frontier HPC. (Figure courtesy: HPE).

uses 6 nm technology, performs double-precision floating-point calculations as fast as single-precision calculations. Each GPGPU has 220 cores, and a core contains 64 streaming (or SIMD) units. AMD designed a custom-interconnect fabric to interconnect the CPU with the GPGPU that boosted the coherency between them, providing them with a unified view of shared data. Each GPGPU has 128 GB high bandwidth memory. As there are 4 GPGPUs per node, the aggregate GPGPU memory is 512 GB. The main CPU of a node also has its own main memory of 512 GB. Each computer node is provided with a 5 TB non-volatile flash memory to quickly feed data to the GPUs [4]. Two computing nodes are placed in a printed circuit board (called a blade) that can be

mounted in a rack. The blades are water-cooled. 64 blades are accommodated in a rack. The Frontier computer has 74 racks. Multiple 64-port switched Ethernet connections called HPE Slingshot, providing an aggregate bandwidth of 12.8 Tbits/sec, interconnect the blades. The Slingshot network interconnection system maximises throughput and minimises latency. It has adaptive routing, congestion control, and excellent quality of service (QoS). A "Dragonfly" topology system interconnects groups of blades, ensuring that there are no more than 3 hops to communicate between any two nodes of the Frontier [5]. The Frontier computer has $74 \times 128 = 9472$ AMD EPYC processors which work out to 606,208 processing cores. There are $4 \times 9472 = 37888$ GPGPUs, each with 220 cores. 145 kilometres of networking cables interconnect the computing nodes. The computer consumes 21 MW of power. It is estimated that it cost $600 million to build.

The Frontier computer has 74 racks. Multiple 64-port switched Ethernet connections called HPE Slingshot, providing an aggregate bandwidth of 12.8 Tbits/sec, interconnect the blades.

The Frontier computer also supports the world's largest and fastest parallel file system as its secondary storage. The system provides around 700 Petabytes of storage with a peak write speed of 5 Terabytes per second. (700 Petabytes can store 33 times the total data housed in the US library of congress, one of the largest libraries in the world!)

The Frontier computer also supports the world's largest and fastest parallel file system as its secondary storage. The system provides around 700 Petabytes of storage with a peak write speed of 5 Terabytes per second.



Figure 2. Frontier Exascale HPC. (Photo courtesy: Wikipedia).

Box 3. Why do Current HPCs use GPGPUs as Co-processors?

Most HPCs use a co-processor for performing floating-point computations as the floating-point speed of the main CPU is insufficient to achieve the PetaFLOPS range of speed. The first HPC to breach 1 PetaFLOPS speed achieved this speed by using a specially designed graphics co-processor, which was a modified version of a processor that was used in computer gaming consoles! You may wonder why a graphics processing unit (GPU) was used as a co-processor. Graphics processing units can perform matrix-based numerical computation at high speed and are amenable to both temporal and data parallelisation. In addition, current GPUs have additional cores called Tensor cores, which provide enhanced support for matrix multiply operations. GPUs use several parallel pipelines to simultaneously render many pixels on the screen of a computer. These pipelines are identical and the structure of the processors is uniform, easing the design and fabrication of the integrated circuits of the GPU. In 2007, a company named NVIDIA which was a manufacturer of graphics processing units (GPU), realised that they could be used also as a general purpose processor by making them programmable. NVIDIA designed a language extension to C, C++, and Fortran called CUDA (Compute Unified Device Architecture) to ease programming GPUs that effectively made GPUs into GPGPUs. GPGPUs nowadays have numerous cores (processing units). The GPGPU in Frontier HPC (AMD Radeon Instinct Mi250X) uses 220 cores, each with 64 SIMD streaming pipelines. It has a peak speed of 95.7 TeraFLOPS for 64-bit operands. It can also execute using lower precision operands 32-bit, 16-bit) which is an advantage in running algorithms to train deep neural networks. The peak speed with 16-bit operands is 383 TeraFLOPS.

Software System of the Frontier HPC

The operating system of the Frontier HPC is based on Linux, which is modified to suit Frontier's architecture. The AMD Radeon RoCm open-source software is used for graphics. It has numerous compilers that include C, C++, and Fortran. It provides both a Cray message passing interface (MPI) and Open MPI for parallel programming. A mathematics library is available that includes BLAS (Basic Linear Algebra Subprograms), LAPACK (Linear Algebra Package) to solve linear systems of algebraic equations, an interactive refinement tool kit, etc. A list of available software can be found at `https://www.olcf.ornl.gov/frontier/`.

Epilogue

Oak Ridge National Laboratory set up the Oak Ridge Leadership Computing Facility (ORLF) as directed by the US DOE Office

of Science. ORLF aims to offer exascale computing resources to researchers from the government, academia, and industry who can convince a group of peers of the importance of the problem they want to solve and why they need an HPC to solve it. In preparation for the Frontier computer, which is expected to be released for general use in early 2023, a variety of projects have been selected to participate in the Frontier Centre for Accelerated Applications Readiness (CAAR) programme [6]. These projects combine simulation and data-intensive machine learning and require exascale performance. Initially, work will start using the existing IBM Summit HPC and will be scaled up to the Frontier HPC when it is released for general use. Some of the selected projects are: processing massive amounts of astrophysical data collected from the Milky Way to predict the behaviour of galaxies, human genome studies to explore genetic predisposition to Alzheimer's disease, addiction to opioids, autism, etc., numerical simulation of turbulent fluid flows in many fields, behaviour of physical processes within rock structures that have embedded fluids, simulate at the atomic level various electronic and magnetic properties of materials, nanoscale simulation to predict the behaviour of viruses to invent drugs and vaccines to fight them, study time-dependent nuclear reactions such as fusion and fission, and design plasma accelerators that would have several applications including in medical devices.

ORLF aims to offer exascale computing resources to researchers from the government, academia, and industry who can convince a group of peers of the importance of the problem they want to solve and why they need an HPC to solve it.

Acknowledgement

I thank Dr S. Balakrishnan, Prof. R. Govindarajan, and Prof. S. Ramani, and an anonymous referee for reviewing this article and for their valuable suggestions that improved it.

Suggested Reading

[1] National Strategic Computing Initiative, Issued by Barak Obama on July 29, 2015. https://obamawhitehouse.archives.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

[2] Exascale Computing Project, https://www.exascaleproject.org/about/

Address for Correspondence
Prof. V. Rajaraman
Supercomputer Education &
Research Centre
Indian Institute of Science
Bengaluru 560 012, India.
Email: rajaram@iisc.ac.in

[3] Tops 500, June 2022. https://www.top500.org/lists/top500/2022/06/

[4] Charles Q. Choi, Beneath the Hood of the First Exascale Computer, IEEE Spectrum, Volume 69, Issue 8, August 2022. https://spectrum.ieee.org/frontier-exascale-supercomputer. See also https://icl.utk.edu/files/publications/2022/icl-utk-1570-2022.pdf

[5] Reference for Dragonfly Topology. https://www.hpcwire.com/2019/07/15/super-connecting-the-supercomputers-innovations-through-network-topologies/ See also https://www.olcf.ornl.gov/frontier/

[6] Frontier Center for Accelerated Applications Readiness. https://www.olcf.ornl.gov/caar/frontier-caar/