Implementation and Applications of Massive-Scale Computing

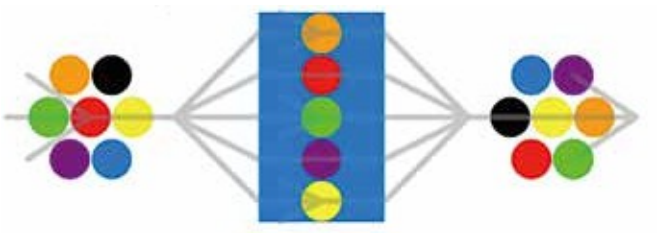
Author: Liu Tianhai

International Business Machines Corporation (IBM) developed the supercomputer "Blue Gene/P"

Human computational capabilities are ill-suited for repetitive and complex computational tasks. Thus, humanity created computers to compensate for these limitations and has persistently pursued ever-greater computational power. Today's single-core CPUs can stably perform over 3 billion operations per second, yet they have reached the limits of chip manufacturing technology. Strength lies in numbers. Based on this fundamental understanding, to achieve faster computing speeds, we have pursued two approaches: packing more cores into a single CPU chip and connecting multiple computers to form a supercomputer. Why does this lead to faster calculations? Consider this example.

A teacher asks students to calculate the sum of numbers from 1 to 100 without using any formulas—just simple addition.

addition. The teacher specified two approaches: First, one student calculates the sum alone. Second, the task is divided equally among 50 students. Each student first calculates the sum of two consecutive numbers. The resulting 50 intermediate results are then distributed among 25 students. The 25 new intermediate results are further divided among 12 students, and so on until the final sum is obtained. Assuming each addition takes 1 minute, the first method takes 99 minutes, while the second method takes 7 minutes plus the time required for the 7 task distributions. As long as the time spent distributing tasks is sufficiently small (less than 13 minutes), this approach achieves a significant speedup. In the field of large-scale computing, this is known as parallel scalability—the collaborative aggregation of computational resources.



Supercomputers achieve faster processing through parallel computing—breaking problems into multiple segments that are simultaneously processed by separate processors, with results ultimately integrated.

This example illustrates a simple application of massively scaled computing. However, real-world applications involve far greater computational workloads—reaching trillions of trillions of operations—and more complex computations. Instead of addition within 100, they solve intricate partial differential equations. For instance, a partial differential equation describing the sagging shape of a fixed-end iron chain under gravity would have taken even Isaac Newton, one of calculus' inventors, several days to compute with low precision. A supercomputer, however, might achieve high-precision results (numbers within $\pm 1.^{79\times10^{308}}$accurate to 16 decimal places) in mere minutes. It is precisely this high computational power that enables us to explore the world and predict the future faster and more accurately.

By now, you might be curious about what this "monster" actually looks like. Let's dive into the world of supercomputers.

Speed is the ultimate advantage—Supercomputers

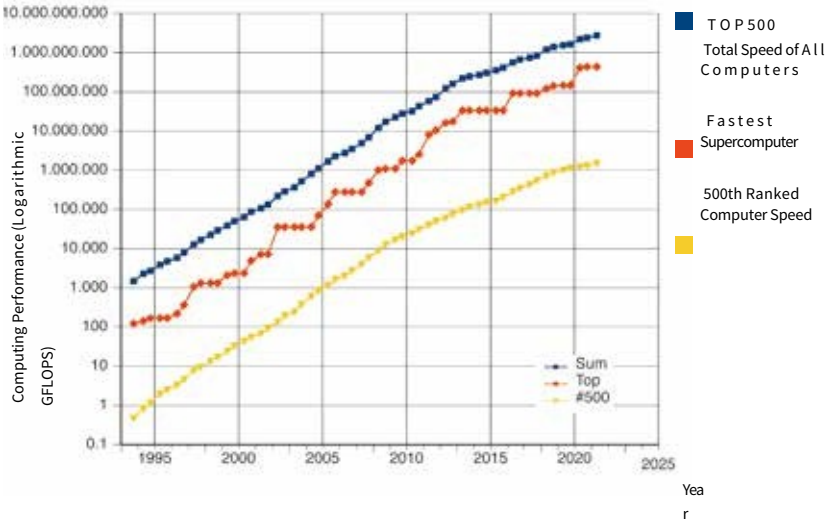
A supercomputer (Super Computer, abbreviated as supercomputing or high-performance computing) represents an implementation of ultra-large-scale computing. At its most fundamental level, it refers to the aggregation of computational resources to deliver higher computational power (often termed computing power). This power is typically harnessed to solve large-scale or complex problems in defense, science, or engineering—such as analyzing high-risk scenarios with elevated accident rates that pose significant threats to life safety.

The industry employs supercomputers to replace manual labor in tasks such as underground coal mining, high-altitude operations, blasting work, and the processing and analysis of petroleum exploration data.

China's supercomputing development began relatively early. Since the birth of the Galaxy-1 in 1983, over the past 40 years, a series of supercomputers have been developed, including the Galaxy, Tianhe, Dawning, KD, Sunway, and Shenteng. Among these, the Sunway TaihuLight, developed in 2016 using domestically developed chips, held the world's top supercomputing position for two years. Located beside Lake Taihu in Wuxi, the Sunway TaihuLight comprises 48 standard cabinets (40 compute cabinets and 8 network cabinets). Each cabinet measures 2 meters high, 0.8 meters wide, and 1.4 meters deep—roughly the size of a double-door refrigerator—occupying a total area of 605 square meters, equivalent to the size of a tennis court. It utilizes Shenwei CPUs (, 1.5 GHz) manufactured using a 28nm process. Each CPU contains 260 cores, totaling approximately 10.65 million cores. With 1.3 million GB of memory, it can perform about 900 petaflops (900 billion billion double-precision floating-point operations) per second. The total cost was 1.8 billion RMB, and its power consumption is 14 megawatts. The electricity required to run it for one hour is enough to power an average Chinese household for 10 years. China now holds 173 of the world's 500 fastest supercomputers, followed by the United States with 149.

The fastest supercomputer in 2021 is Japan's "Fugaku" in Kobe. (Fugaku), comprising 432 standard racks, occupies 2,000 square meters and cost a total of 6.4 billion RMB. Fugaku utilizes AMD 7nm process technology with Fujitsu-designed 2.2 GHz CPUs ( ) featuring 48 cores per CPU, totaling 7.63 million cores. It boasts 5 million GB of memory and achieves 44 quadrillion double-precision floating-point operations per second, though its power consumption reaches a massive 30 megawatts.

Next, we will introduce the organizational structure and fundamental hardware/software of mainstream supercomputing systems worldwide. Specific details such as internal CPU parallelism, memory parallelism, high-speed interconnects, kernel optimization, and compilation optimization will be addressed later.



The international TOP500 organization is the authoritative body publishing rankings of globally installed supercomputer systems, jointly compiled by supercomputing experts from the United States and Germany. Data released by TOP500 demonstrates that the comprehensive performance of supercomputers is growing rapidly year by year (Credit: Top500.org)

Floating-Point Operations

The decimal numbers we commonly use are stored in computers as floating-point numbers, typically represented in scientific notation. Floating-point computation involves arithmetic operations on these numbers. Due to the inability to represent values precisely, approximations are used, leading to two formats: single-precision and double-precision floating-point numbers. Double-precision floating-point numbers use 64 bits to store a single value, representing absolute values rangingfrom -2^{1024} to 2^{1024}. (FLOPS) is commonly used to measure computational speed.

Cluster

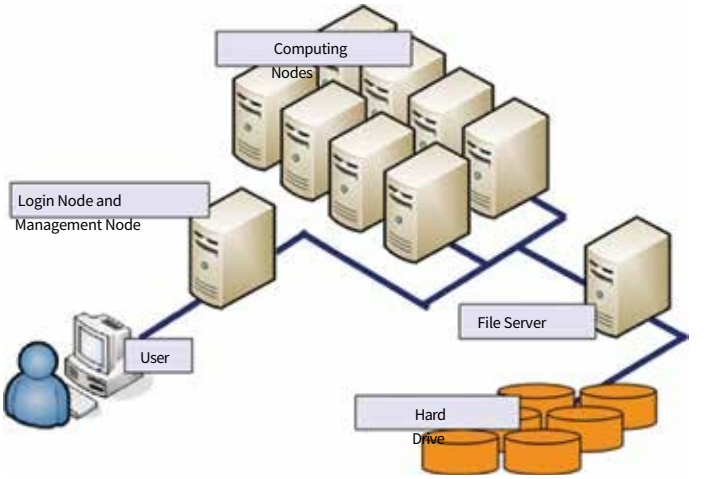
90% of supercomputers in the TOP500 list adopt a cluster architecture. A cluster is a parallel computing system that uses ordinary microcomputers or high-end commercial machines as computing nodes, interconnected via high-speed networks. Communication between nodes occurs through sending and receiving messages, such as receiving computational tasks and returning results. The latency of message passing and the rationality of task allocation fundamentally determine the computational performance of the entire cluster system. Since each computing node possesses independent memory, this distributed memory structure grants clusters high scalability—meaning we can adjust the supercomputer's computational power by increasing or decreasing the number of computing nodes.

Clusters can be categorized into five types of devices.

(1) Login Node. The login node serves as the gateway for ordinary users to access the supercomputer. Users typically log in to this node to compile, submit, and initiate overall computational tasks. Login nodes generally do not require high computational power and may consist of just a few ordinary servers. A computational task may need to be copied bit by bit from hundreds of hard drives provided by users to the login node.

(2) Management nodes. Management nodes interconnect and control computers within the cluster, performing operations such as power cycling and upgrading operating systems on compute nodes. Management nodes also have low computational demands. Management and login nodes may reside on the same hardware, with administrators and regular users possessing distinct permissions, allowing only operations within their authorized scope.

(3) Computing Nodes. Computing nodes form the computational core of the entire cluster. Once a computational task is initiated, nodes communicate to distribute tasks among themselves and dynamically adjust subsequent task allocations during processing. If a node performs exceptionally slowly and other nodes are waiting for its results, the task is simultaneously reassigned to faster nodes. Regardless of which node completes first, the computational



The vast majority of supercomputers adopt a cluster architecture. Users access the computing cluster through a login node, submitting tasks to the management node. The management node decomposes the tasks, allocates them efficiently to various compute nodes for parallel processing, then collects the computational results and stores them in the file system or on hard drives.

The results are then transmitted to waiting nodes.

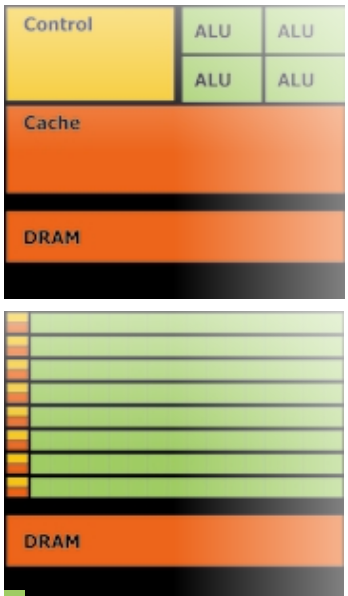
(4) Network Devices. Network devices include high-speed switches and interconnect cables for node connectivity. Most modern supercomputers utilize InfiniBand dedicated network cables. A typical 12x-link network achieves speeds of 290 gigabits per second—58 times faster than 5G networks.

(5) I/O devices and storage devices, i.e., supercomputer peripherals and fiber-optic or SCSI-based hard drives, are used to store computational data and results.

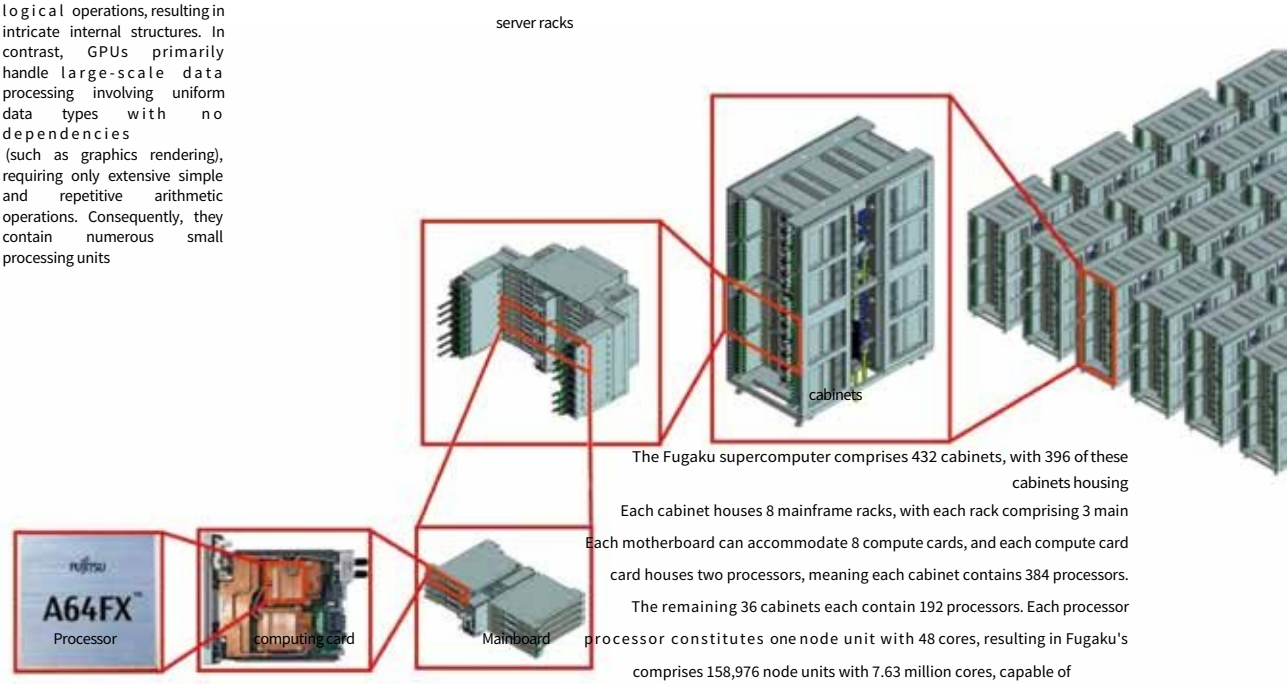
Hardware

The hardware composition of supercomputer compute nodes is fundamentally similar to that of personal computers, adhering to the von Neumann architecture. Beyond essential CPUs, supercomputers utilize GPUs within compute nodes to significantly enhance computational efficiency.

The CPU is a general-purpose processor that powers most computational tasks, while the GPU was originally developed for rendering graphics on screens. Rendering doesn't require logical operations but demands massive amounts of simple, repetitive arithmetic calculations. This single-purpose computation can be perfectly executed using very small integrated circuits, which is why GPU cores are significantly smaller than those in CPUs. To make the rendering process more efficient, more cores are packed into the GPU. For instance, the Nvidia GeForce GTX 980 Ti boasts 2,816 cores, while contemporary CPUs possess only 32 cores. If a CPU is likened to a PhD proficient in all types of computation, then a GPU resembles 2,000 elementary students who can only perform addition. Compared to a PhD, calculating 2,000 simple addition problems is undoubtedly faster when handled simultaneously by 2,000 students. This is why supercomputers commonly employ



CPUs and GPUs feature distinct architectures tailored to different application scenarios. CPUs require high versatility to process diverse data types and perform complex logical operations, resulting in intricate internal structures. In contrast, GPUs primarily handle large-scale data processing involving uniform data types with no dependencies (such as graphics rendering), requiring only extensive simple and repetitive arithmetic operations. Consequently, they contain numerous small processing units



The Fugaku supercomputer comprises 432 cabinets, with 396 of these cabinets housing

Each cabinet houses 8 mainframe racks, with each rack comprising 3 main

Each motherboard can accommodate 8 compute cards, and each compute card

card houses two processors, meaning each cabinet contains 384 processors.

The remaining 36 cabinets each contain 192 processors. Each processor

processor constitutes one node unit with 48 cores, resulting in Fugaku's

comprises 158,976 node units with 7.63 million cores, capable of

performing 44

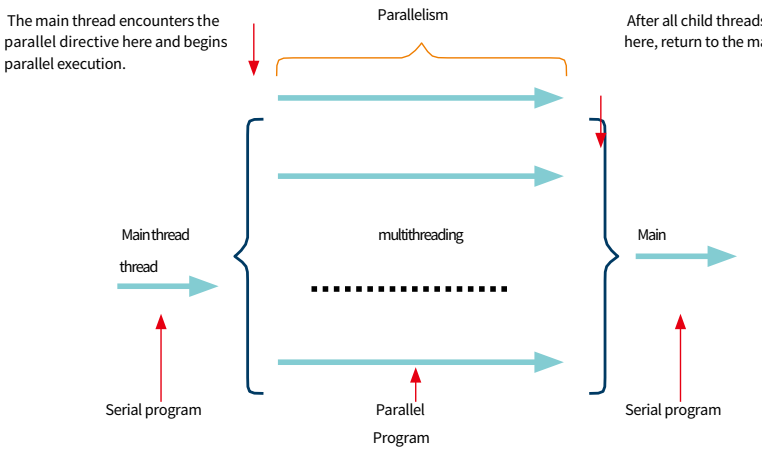
quadrillion double-precision floating-point operations per second

GPUs serve as accelerators to speed up large volumes of simple arithmetic operations.

Software

While supercomputers boast formidable hardware capabilities, software plays a crucial role in unlocking their full potential. Regarding operating systems, nearly all supercomputers utilize Linux, owing to its inherent open-source philosophy and modular design. Much like building with blocks, developers can easily integrate new driver modules to support emerging hardware. Second, for programming languages, C and C++ are the mainstream choices for developing high-performance applications. Finally, with parallel programming frameworks, developers can focus on solving specialized problems and implementing application functionality. The frameworks handle the efficient allocation of computational tasks and coordination among compute nodes.

We will examine a C++ program that sums 100 numbers using the OpenMP parallel programming framework. OpenMP is a C++ parallel programming framework that enables ordinary code to be automatically converted into parallel programs with minimal modifications. Specifically, it analyzes the OpenMP compilation directive `#pragma`, automatically converting code blocks defined by the directive into parallel threads for execution, and finally merging the thread results. The `num_threads` clause controls the number of threads created, specifying how many threads are allocated to complete the computational task. The `parallel for` directive generates a parallel domain and distributes computational tasks among multiple threads, thereby accelerating the execution speed. (For more advanced control directives, refer to the official OpenMP documentation.) In the example, the computational task is distributed across 10 threads. Upon completion, each thread passes its result to the next computational node, ultimately yielding the sum of 100 numbers. This program's execution time is approximately one-fifth that of the non-OpenMP version.



OpenMP is a C++ parallel programming framework that enables ordinary code to be converted into parallel code with minimal modifications.

modifications to automatically transform it into parallel code.

Program. That is, only a single main thread exists at the start. When parallel computation is required, multiple branch threads are spawned to execute parallel tasks. After parallel code execution completes, the branch threads

converge, and control flow is returned to the single main thread.

Of course, more threads don't necessarily mean higher efficiency. As mentioned earlier, low latency between data nodes and efficient task allocation also play decisive roles. Interested readers are encouraged to experiment themselves.

```
/* Import OpenMP header file to use OpenMP directives in code */
#include <omp.h>
/* Program execution begins at the main function and ends at the last curly brace. Statements within curly braces are executed sequentially. */
  * statements are executed sequentially */
int main()
{
/* Define a variable sum to store the calculation result, initialized to 0 */
    int sum = 0;
/* When the compiler encounters the #pragma directive, it hands off compilation to
 OpenMP to handle the for statement block, then resumes control after processing */
    #pragma omp parallel for num_threads(10) reduction(+:sum)
/*OpenMP creates and compiles a new for loop block for each thread*/
        for(int i=1; i<=100; i++)
        {
/*Summation*/
        sum = sum + i;
        }
/* Print result to console */
    printf("Sum of first 100 natural numbers is %d\n", sum);
/* Program terminates, returning 0 indicates no errors occurred */
    return 0;
}
```

A real supercomputing application does not require business logic processing. When handling diverse transactions, its code may not be extensive, but its algorithms are highly complex and must be rigorously optimized to maximize supercomputer performance.

What was once reserved for the elite now enters the homes of ordinary people—Hyperscale Data Centers

Supercomputers primarily serve computational tasks for defense, government, and research institutions, while private or corporate applications face certain barriers. Advances in supercomputing technology have driven the establishment of hyperscale data centers (HDCs), making direct access to hyperscale computing feasible for individuals and private organizations.

Hyperscale data centers boast world-class computational power and highly efficient resource utilization.

including high-performance computers (servers) capable of storing massive amounts of data, network equipment, and communication facilities, enabling large-scale, efficient computing. When we use digital services like WeChat Pay, data centers perform numerous demanding tasks behind the scenes to ensure seamless operation.

- Research Group, a strategic intelligence firm specializing in IT and cloud industries, recently examined hyperscale data centers (HDCs) worldwide. By the end of 2021, HDC capacity had doubled over four years, with the number growing to 659 facilities. U.S. enterprises account for half of these, particularly Microsoft, Amazon, and Google, which collectively operate over half of the current total. Naturally, carrier capital expenditures continue to set new records—reaching $150 billion over the past four quarters.

Hyperscale data centers also represent an implementation of hyperscale computing. Like supercomputers, they utilize cluster architectures but differ in several key aspects.

(1) Supercomputers prioritize high computational power per node, with minimal expectation for frequent, large-scale data transfers between nodes. HDCs emphasize massive data storage and transmission, relatively deemphasizing computational power. Consequently, supercomputers feature numerous high-end nodes, while HDCs predominantly use standard PCs equipped with distributed file systems and databases to deliver massive data storage capacity.

(2) Supercomputers serve national-level computational tasks in defense, science, technology, and industry, while HDC primarily addresses big data operations. Daily, we generate approximately 2.5 zettabytes of data. With the increasing prevalence of IoT (Internet of Things), this data creation rate will grow exponentially. Additionally, national-level supercomputing tasks cannot be conducted on commercial-grade systems due to confidentiality concerns.

With the rapid annual growth in data volume, there is a need for servers with more powerful data processing capabilities. Grid computing is a cost-effective computing model that delivers exceptional data processing power. It utilizes the internet to organize computers scattered across different geographic locations into a virtual supercomputer. Each participating computer is a node, and the entire computation is a grid composed of thousands of nodes working together to solve complex problems that cannot be addressed by local resources alone.

on the HDC.

(3) Traditional supercomputing requires multiple layers of approval. Even after obtaining usage rights, users must wait in queues. Moreover, once approved, computational resource specifications (such as processing power configurations) cannot be adjusted on demand. This often leaves time-sensitive computational tasks in a precarious situation. HDC offers significant convenience. If resources are available, users can apply anytime and automatically gain access. Alternatively, computational resources can be reserved in advance, providing flexible usage options. Furthermore, cloud computing enables HDCs to adopt flexible resource allocation strategies, allowing users to allocate computational resources on demand.

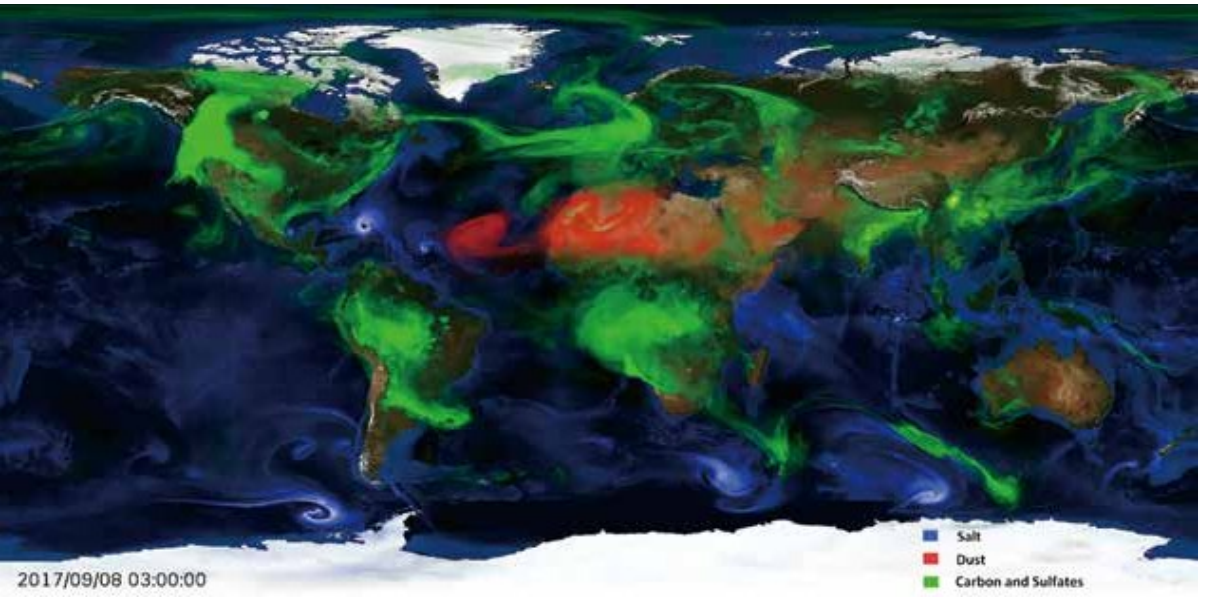
Every talent has its purpose

Numerical Weather Prediction

We now frequently check weather forecasts via mobile apps, especially the highly accurate numerical weather predictions that tell us the exact time and probability of rain. This capability relies on supercomputers simulating future weather patterns. Consider a city's numerical weather forecast: First comes data collection. Weather stations scattered across the city gather meteorological data every few minutes—including wind direction, wind speed, temperature, and humidity at each location. Simultaneously, various meteorological data are provided by satellites, weather balloons, ocean buoys, and other sources. This massive volume of weather data is fed into supercomputers, which contain a series of partial differential equations derived by meteorologists from the patterns of atmospheric movement. The supercomputers solve these equations using known weather change data, calculating short-, medium-, and long-term weather trends at different pressure levels at specified intervals. These trends are then visualized as weather maps, such as a 6-hourly cumulative precipitation distribution map showing 2-hourly totals.

Supercomputers process vast climate datasets to calculate short-, medium-, and long-term atmospheric pressure variations. Pictured is the global aerosol forecast generated by the Goddard Earth Observing System Forward Processing (GEOS FP) model, which runs every 6 hours daily, incorporating approximately 4 million observational data points per run.

(Credit: NCCS)



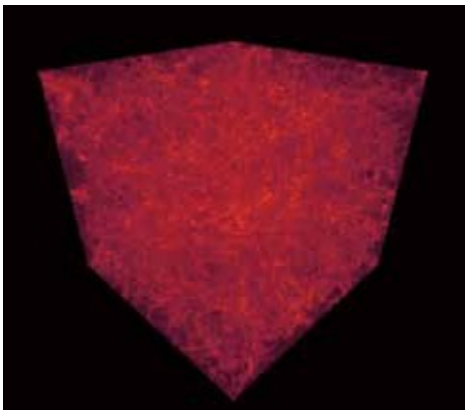
China's Changsha Supercomputing Center provides computational platform support for numerical weather forecasting to a meteorological bureau in a central Chinese province. Following operational deployment, the bureau's numerical forecasting capabilities have significantly improved: horizontal resolution has increased from 20 kilometers to 4 kilometers, and heavy rainfall resolution has improved from 37 kilometers to 15 kilometers. This brings us one step closer to predicting phenomena like "rain not falling across a river" and "winds differing within ten miles."

Cosmic Evolution

Dark matter constitutes approximately 84% of all matter in the universe, with the remaining 16% being the familiar ordinary matter we know. Dark matter is invisible and intangible, even impervious to electromagnetic waves, yet it exerts a profound influence on galaxy formation. Scientists have discovered that dark matter interacts solely through gravity. After neglecting other forces exerted by the small amount of ordinary matter, N-body simulations provide an ideal method for tracing the motion and distribution patterns of dark matter in the universe. To analyze the role of dark matter in the cosmic transition from chaos to the formation of specific structures, the National Astronomical Observatories of China (NAOC) and the Supercomputing Center of the Chinese Academy of Sciences (CAS) collaborated on a thousand-core scale simulation using the ShenTeng 7000 supercomputer. They represented the vast number of dark matter particles in an early cosmic region using N identical point masses within a three-dimensional box. These particles were assigned initial positions and velocities based on cosmological models. By tracking the gravitational motion of these N particles, scientists could "observe" the final structures formed by dark matter particles. In the video generated from the simulation results, each frame represents the state of the universe at a specific point in time. Bright spots indicate locations where galaxies have formed. These massive bright spots are connected by numerous filamentary and sheet-like structures, and the regions enclosed by these filaments or sheets are called voids. If all dark matter were visible to us, the universe would resemble the network-like structure depicted in the image. Without supercomputers, it would be extremely difficult to visualize all of this.

Industrial Engineering Simulation

When high-speed trains exceed 350 kilometers per hour, aerodynamic noise becomes the primary source of disturbance. To enhance passenger comfort and mitigate noise impacts on residents along railways, the Chinese Academy of Sciences' Supercomputing Center and Institute of Mechanics collaborated. Using numerical simulation combined with theoretical and experimental research, they predicted and evaluated aerodynamic noise from high-speed trains, proposing optimized noise-reduction designs for critical train components.

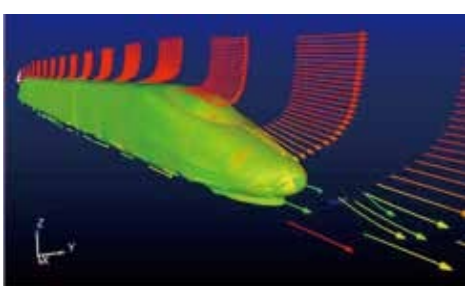


Supercomputing can simulate the evolution of the universe dominated by dark matter. The image depicts the cosmic landscape at a specific moment, with bright spots indicating regions where galaxies have formed. Scan the video QR code to observe how the universe gradually transitions from chaos to specific structures under the influence of dark matter. This visualization aids astronomers in studying phenomena of scientific interest.

Scan to watch the video

Scan to watch the video

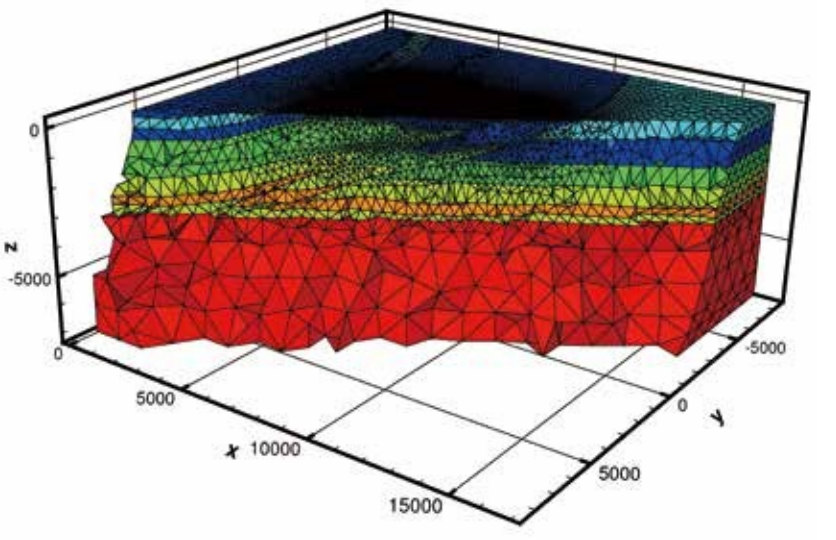
Supercomputers can predict and evaluate aerodynamic noise from high-speed trains, proposing noise reduction optimization designs for critical train components. The image shows simulated airflow patterns around a train during high-speed operation.



Geological Exploration

Modern geophysics leverages vast observational datasets and large-scale computational simulations to enhance our understanding of the Earth. Consequently, solid Earth modeling represents one of the largest applications of high-performance computing, spanning from seismic simulations to exploration data processing. For instance, when exploring the subsurface structure of a 20×20×20 cubic kilometer block (length×width×depth)—though actual exploration blocks are far larger—a survey team might manually generate controlled seismic waves at 4-kilometer intervals or other specified spacings. Surface receivers then capture reflected seismic waves from underground and store them on hard drives. These seismic waves include both P-waves and S-waves, generating massive data volumes—typically tens of terabytes. After collection, the hard drives containing this data are transported to a supercomputing center for modeling. Modeling involves dividing the subsurface structure into numerous grid cells, where each cell reflects or transmits seismic waves to neighboring cells until the simulated seismic waves match those collected in the field. To achieve sufficient model accuracy, the grid size must be sufficiently small—for example, a 50×50×20 cubic meter block yields 160 million grid cells. This requires the computer to solve for 160 million unknowns, which may be interdependent or independent. Specifically, it involves solving a linear system of equations with 160 million unknowns. This represents modeling for a single measurement set. Exploration teams frequently generate multiple artificial seismic events at different locations within the same block to collect new data. Consequently, a single model requires repeated computations to meet petroleum exploration requirements, making the computational workload and complexity immense.

Oil exploration companies maintain their own supercomputers to meet data analysis demands. As early as around 2000, CNPC began procuring Dawning 4000L supercomputers for petroleum exploration, and it now operates China's largest seismic data processing center.



Supercomputers can simulate the solid Earth, ranging from seismic modeling to exploration data processing. To achieve sufficient model accuracy, the subsurface structure is divided into numerous sufficiently small grids. For instance, a 50×50×20 cubic meter block contains 160 million grid cells, meaning each calculation requires solving hundreds of millions of unknowns. Since the model demands multiple computations to meet petroleum exploration requirements, the computational workload and complexity are immense.

Animation Rendering

In animated film production, animators first sketch characters and environments on computers. Based on the script, they then animate character expressions and movements while setting scenes. Each character's eyes, eyelashes, strands of hair, and clothing folds require their own dedicated video track. Similarly, every light source, window, and even leaf in the scene demands its own track. Animation rendering ultimately combines all video tracks at a single point in time to generate a single frame of 2D or 3D imagery. If the director disapproves of a rendered image or animation segment, we must "do another take"—adjusting character expressions, movements, or the scene before re-rendering. A one-minute animation segment comprises 1,440 frames (60 seconds × 24 frames), while a 4K ultra-high-definition film contains 8 million pixels
(4096×2160). If each frame takes 30 seconds to render (already fast for 4K films), that requires 12 hours of rendering time. Animation rendering is a task with no technical difficulty but immense workload and time consumption—precisely where supercomputers excel. Currently, animation rendering is mostly achieved through hyperscale data centers. From Journey to the West: The Return of the Monkey King and Big Fish & Begonia to Ne Zha and The White Snake 2: The Green Snake's Revenge, most domestic animation studios have adopted this Renderbus self-service rendering model, accumulating over 300 million core-hours of total rendering time.

Animation rendering involves combining all video tracks at a single point in time to generate a single frame of 2D or 3D imagery. While technically straightforward, this process demands immense labor and time, making it ideally suited for supercomputing in feature-length animation projects. The image above shows Blender's animation rendering interface. This open-source, cross-platform, all-in-one 3D animation software provides a comprehensive solution for creating animated shorts, covering modeling, animation, materials, rendering, audio processing, and video editing.

(Credit: Blender)

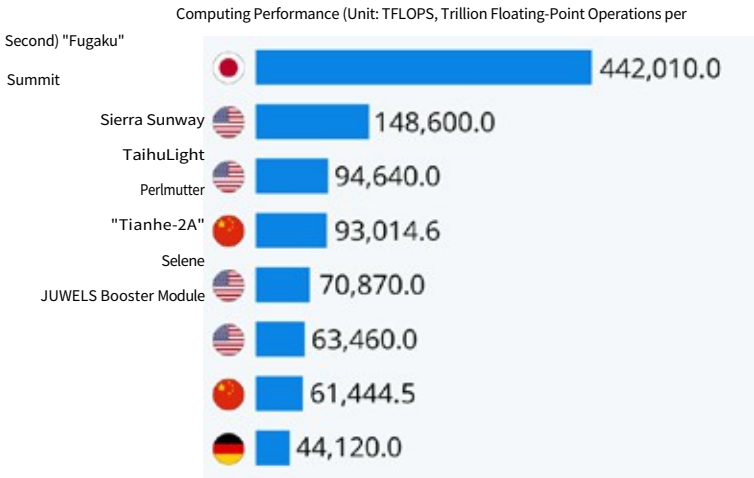
The tide is calm, the shores are wide; The wind is fair, the sail is set.

Future Supercomputing

According to expert research and relevant news reports, future trends in hyperscale computing include but are not limited to the following:

(1) Continuous Speed Enhancement. Computing power remains the primary objective for supercomputers. Currently, nearly all systems in the world's TOP500 supercomputer rankings achieve petaflop (P) level processing speeds (10^{15} operations per second). Nations are fiercely competing to develop exaflop (E) level supercomputers (10^{18} operations per second). Based on historical speed improvements, the TOP500 organization previously projected exaflop supercomputers would emerge around 2020. However, due to the economic impact of the global pandemic, the semiconductor industry has experienced slow growth over the past three years, even facing chip shortages. The arrival of exascale computers will be delayed. Reports indicate that China has already deployed two systems achieving exascale speeds. The U.S. National Strategic Computing Initiative (NSCI) plans to complete development of two exascale supercomputing systems by the end of 2023.

(2) Smaller size, lower energy consumption. The architecture of supercomputers has seen no significant changes in the past decade, with cluster-based systems remaining the mainstream architecture. While the underlying architecture remains unchanged, rather than relying solely on increasing computational resources to dramatically boost computing power, there is growing interest in the technological convergence brought by quantum computers and biological computers. Quantum computers excel at solving certain types of problems that are difficult for traditional computers to tackle. Biocomputers exhibit biological characteristics, such as leveraging inherent regulatory functions to automatically repair chip failures and mimicking human brain mechanisms. They operate 100,000 times faster than today's most advanced computers, consume only one-billionth the energy of conventional systems, and possess immense storage capacity.



Supercomputer Performance Rankings (Data as of November 2020): Japan's Fugaku is currently the world's fastest supercomputer, capable of performing 44 quadrillion double-precision floating-point operations per second—nearly three times the speed of the world's second-fastest supercomputer. (Credit: Top500.org/Statista)

East-West Computing Initiative

With the rapid development of the digital industry, China's data centers now span 5 million standard racks, delivering a computing power of 130 EFLOPS (130 quintillion floating-point operations per second). As digital technologies continue to permeate all sectors of the economy and society, the demand for computing power remains urgent, projected to grow at an annual rate exceeding 20%.

Currently, most of China's data centers are concentrated in the eastern regions. Continuing to develop large-scale data centers in the eastern areas, where land is scarce and resources are limited, has become increasingly challenging. However, the western regions possess abundant renewable resources, presenting an opportunity to develop data centers there to meet the computational demands from the east.

In February 2022, the national "East Data, West Computing" initiative was formally launched. It initially involves establishing eight national computing power hub nodes and ten national data center clusters in the Beijing-Tianjin-Hebei region, the Yangtze River Delta, the Guangdong-Hong Kong-Macao Greater Bay Area, Sichuan, Inner Mongolia, Guizhou, Gansu, and Ningxia. In "East Data, West Computing," "data" refers to information, while "computing" denotes processing power—the capability to handle data. This major initiative is hailed as the "South-to-North Water Diversion," "West-to-East Power Transmission," and "West-to-East Gas Pipeline" of the digital economy era. Its purpose is to optimize China's computing power supply structure through supply-side reforms.

By leveraging the western regions' affordable electricity, communication bandwidth, and land resources, the initiative reduces computing costs to more effectively support China's digital economy. Simultaneously, it stimulates western economic growth by fostering regional digital development through hub construction, driving related industries, and cultivating high-quality digital sectors.

Over the next five years, this initiative is projected to generate over 100 billion yuan annually in western industries, encompassing data center operations and maintenance, data center construction, servers, chips, system and database software, communication facilities, as well as future eastern industries such as data circulation, data infrastructure, and AI-driven data processing.

Abundant renewable energy and favorable climate offer significant potential for green data center development

Substantial user planning and strong application demand

The "East Data, West Computing" initiative, formally launched in 2022, establishes an integrated computing power network system combining data centers, cloud computing, and big data. It systematically redirects eastern computing demands to the west, leveraging western advantages in affordable electricity, communication bandwidth, and land resources to reduce computing costs. This optimizes data center deployment and more effectively supports China's digital economy development.

Author Profile: Liu Tianhai holds a Ph.D. in Theoretical Computer Science from Karlsruhe University of Technology, Germany. His professional experience includes positions at Darmstadt University of Technology's Computer Science Department, IBM Research in Böblingen, Germany, and aicas GmbH. His research focuses encompass testing and verification of complex software, embedded real-time virtual machines, telemetry data processing, smart grids/logistics, industrial IoT, and cloud computing.