

HIGH-PERFORMANCE COMPUTING IN THE CLOUD: IS YOUR BUSINESS READY TO REAP THE BENEFITS?

FROST & SULLIVAN VISUAL WHITEPAPER

CONTENTS

- 3** The Democratization of Data
- 4** Defining HPC
- 5** Artificial Intelligence (AI) and ML Drive HPC Use Cases
- 6** Disparate Data Types and Needs Shape HPC Workloads
- 7** HPC Applications Require a High-performance Network
- 8** The Last Word

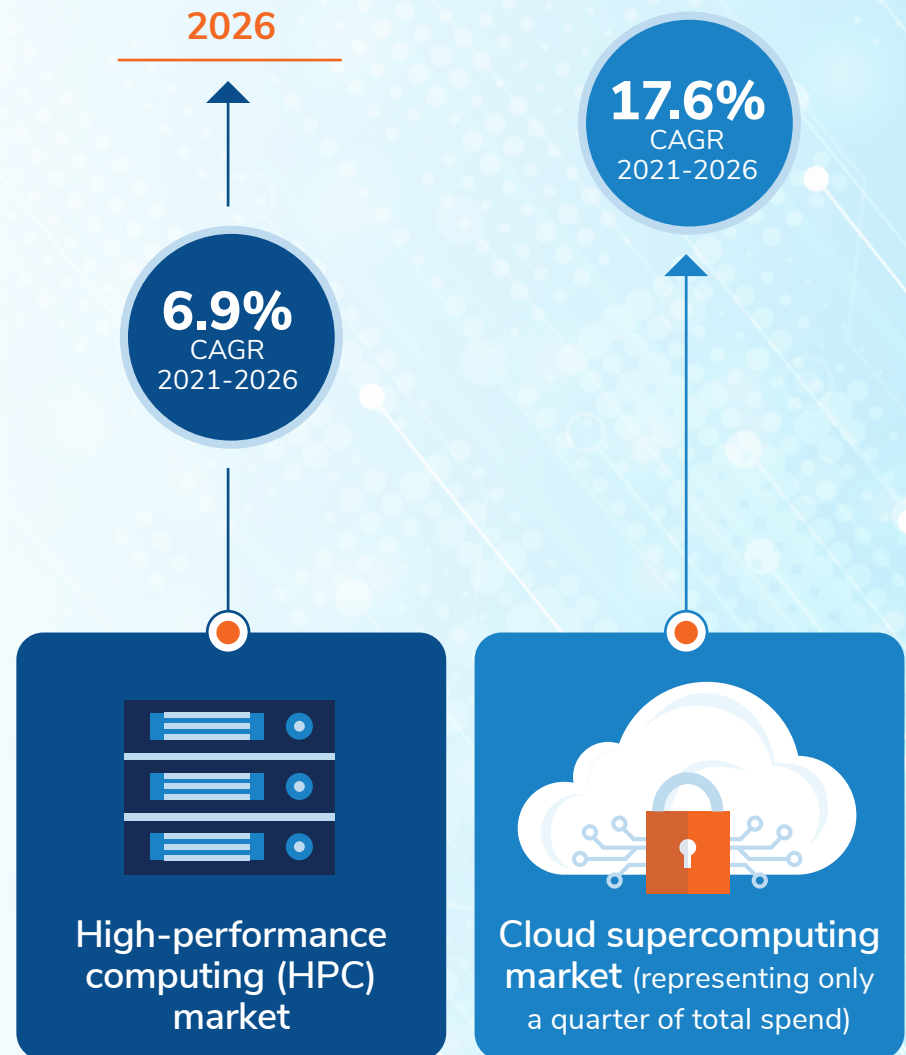
The Democratization of Data

Supercomputing is inching downmarket. Although large government and research institutions were once the main large-scale processing users, it is now within reach of businesses and organizations looking to create value and insights from vast data volumes.

The cloud is a major supercomputer market driver, and according to industry research, the on-premises high-performance computing (HPC) market is growing at a compound annual growth rate (CAGR) of 6.9% through 2026. The cloud supercomputing market—currently representing only 27% of total spend—is growing at a 17.6% CAGR.¹ The cloud model offers nearly limitless resources for processing and storage across a network of cloud centers, making cloud supercomputing attractive and achievable for businesses that are unable or unwilling to build and maintain the immense facilities required for on-premises HPC.

However, it takes more than processing and storage to support a supercomputing application in the cloud. The cloud must securely transport massive amounts of data between cloud centers (and sometimes back to the source), and so organizations planning on running supercomputing applications must incorporate a supernetworking component—a network that is as high-bandwidth and massively scalable as the application it supports.

In this brief, we will review HPC definitions and data requirements and explore the role of the network in a successful cloud supercomputing initiative.



¹ Hyperion Research, 2022

Defining HPC

HPC refers to the aggregation of processing capability to support complex analytics and simulations on massive amounts of data measured in petabytes (PB) or even exabytes.

Supercomputing is a type of HPC that involves hundreds or thousands of individual processors working together on a task. The term supercomputer generally refers to a massively scalable hardware cluster designed to perform parallel processing.

Supercomputing requires specialized hardware, hosted in the cloud or on-premises, to perform complex algorithms on massive datasets simultaneously and quickly.

To support on-premises installations, commercial system manufacturers design hardware to handle massive data growth, scaling to hundreds of thousands of nodes while ensuring consistent performance. This approach can be costly and incur considerable capital expenditure as the application scales. Operating expenses can also skyrocket from hiring or engaging additional skilled resources to support growing infrastructure installations.



For cloud installations, public cloud providers offer specialized instance types designed to support compute-intensive or machine learning (ML)/neural networking workloads. In addition, providers may offer functionality to build, deploy, and manage HPC clusters. The public cloud option relieves organizations of the capital and operating expense burdens associated with on-premises installations, as the cloud service provider takes on responsibility for acquiring, installing, and maintaining the infrastructure.

Artificial Intelligence (AI) and ML Drive HPC Use Cases

If the public cloud is democratizing HPC infrastructure, cloud-based AI and ML functionality is democratizing HPC use cases. Commercial and public organizations of all sizes are turning to data to create value—for their use, on behalf of clients, or on behalf of society—and they are increasingly leveraging easily accessible and affordable cloud-based AI and ML tools to do so.

THE DIVERSE AND CREATIVE WAYS ORGANIZATIONS ARE USING AI AND ML INCLUDE:

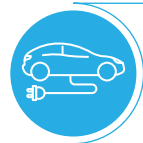


Modernizing financial services

(e.g., credit card fraud detection, real-time stock tracking, trading automation)

Diagnosing disease and prescribing treatment

(e.g., gene therapy)



Disrupting industries

(e.g., self-driving cars, chatbot-based services, self-guided technical support)

Exploring contingencies

(e.g., flight simulators, behavior predictions)

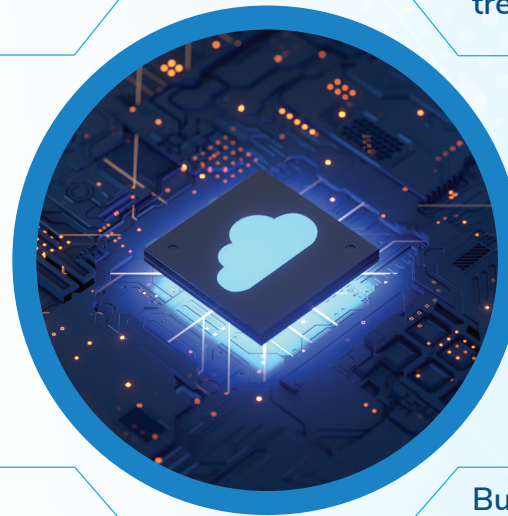


Processing environmental data

(e.g., seismic data processing from, deep learning statistical analysis from oil and gas exploration)

Building simulations for manufacturing designs

(e.g., automotive and airline industries, Photorealistic 3D rendering)



In each case, the AI/ML model requires massive amounts of data for ongoing training. As input data quantities and variables grow and the outcomes become more complex, the AI/ML model requires more processing power. When the volume, complexity, and required processing speed exceed normal computing capacity, it becomes HPC.

Disparate Data Types and Needs Shape HPC Workloads

HPC workloads ingest massive volumes of data, and depending on the application or use cases, data may vary in format, pace, urgency, and security requirements.

FOR EXAMPLE:



Data may arrive in multiple formats, including bandwidth-hungry formats such as images or streaming media. A single HPC workload may use multiple types of data. For example, a diagnostics research organization may ingest images and tabular lab results to derive precision medicine solutions.



Data may arrive in a steady and predictable cadence (e.g., earth images used in geolocation apps) or come in spiky patterns (e.g., weather pattern alerts).



Data may require immediate processing and action (e.g., data derived from autonomous vehicles) or stored for future use (e.g., voluntary DNA databases).



Data may be subject to compliance regulations (e.g., private health or financial information) and/or considered proprietary intellectual property, sensitive to breach (e.g., media and entertainment files).



The volume and diversity of data—and the needs of each application—place a great burden on the networks transporting data to and within the cloud/data centers that perform the processing.

HPC Applications Require a High-performance Network

A data-intensive HPC workload may ingest upwards of 10 PB per day, process and return instructions instantaneously, and then forward the data for storage and additional processing in a remote cloud center.

Depending on the specific HPC application, data ingest and transfer scenarios may include:

- High volumes of data collected at the edge (e.g., via sensor-equipped Internet of Things [IoT] or Industrial IoT devices or endpoints) must shift to a cloud facility for processing, with results returned quickly.
- PB-scale data from multiple sources (e.g., smart infrastructure, weather stations) aggregate at a facility for cleansing and verification before shifting to the HPC cloud.
- Configuration of HPC workloads in private data centers for cloud-bursting (moving to a cloud HPC center) may occur when data volumes spike beyond the local center's capacity.
- An HPC workload may require multi-step processing of massive data volumes, with data exchanged among intermediaries based on processing outcomes.

Each scenario requires massive data volumes to securely, reliably, and quickly move across network connections.

Just as a standard enterprise server or public cloud cannot support the intense needs of an HPC workload, neither can traditional network services. Unfortunately, cloud professionals in most organizations do not have the expertise or experience to manage cloud HPC despite an ability to manage standard cloud workloads successfully. They may not be aware that as data volumes and AI-enabled services increase (e.g., simulation software), the underlying infrastructure needs to be reassessed. Their first indication of a problem may be poor app performance or lost data.

To maximize the value and performance of HPC workloads, you must ensure you have dedicated network links capable of supporting the unique requirements.

YOUR HIGH-PERFORMANCE NETWORK SHOULD DELIVER:



Very high bandwidth to rapidly move massive-scale ingest data into and out of HPC facilities



Guaranteed, consistent performance with end-to-end low latency and minimal packet loss, from every data source and for every data format that is transporting data for HPC purposes



Dedicated, direct connections to HPC centers, edge-to-cloud and cloud-to-cloud, enabling full visibility and management for crucial workloads



Highest levels of security to protect vital data end-to-end

The Last Word

Supercomputing workloads are no longer confined to stodgy research or educational institutions. Thanks to the cloud, any company with visionary insight can leverage massive-scale data and cloud-based tools—such as AI and ML—to transform industries and societies.

However, just as HPC cannot run on standard computing infrastructure, the workloads cannot utilize standard network infrastructure. Cloud HPC requires fast, secure, and reliable ingest of massive amounts of data into the cloud facility. Some workloads also require the same volumes to be transferred among cloud centers and back to the origin sites (edge or private data center). For essential circuits, organizations must rely on a network services partner that can offer a high-performance network—very high bandwidth, highly scalable, highly reliable, and far-reaching.

Do not let your network hold back the potential of your HPC workloads. With the right network partner, you will be prepared to derive maximum value and insight from your data-intensive HPC workloads.





Growth is a journey. We are your guide.

For over six decades, Frost & Sullivan has provided actionable insights to corporations, governments and investors, resulting in a stream of innovative growth opportunities that allow them to maximize their economic potential, navigate emerging Mega Trends and shape a future based on sustainable growth.

[Contact us: Start the discussion](#)