IDC

SPOTLIGHT

Sponsored by: HPE and Intel

Businesses are on the cusp of a major paradigm shift to unlock endless opportunity and groundbreaking advances. As distributed computing, artificial intelligence, and sustainable and secure computing come together, they are changing when, where, and how quickly businesses turn data into intelligence.

Digital Strategies in the Era of AI

February 2024

Written by: Ashish Nadkarni, Group Vice President and General Manager, Worldwide Infrastructure and BuyerView Research

Introduction

Digital transformation — a technology-based business strategy — has brought about a slew of organizational, process, and technological changes, many of which are disruptive. These changes are leading to massive and permanent shifts in how the enterprise operates, creates value, and positions itself as a differentiated entity to its employees, customers, and stakeholders. While organizations will continue to innovate, the time has come for many to embrace the next era of digital transformation.

AT A GLANCE

WHAT'S IMPORTANT

- Businesses are producing increasing quantities of data. In 2020, 64ZB of data was produced. In 2025, 181ZB will be created annually.
- Businesses must develop an IT strategy to gain rapid and deep insights from data, regardless of where and how it is generated.
- The AI strategy must take a full-stack approach to AI training and inferencing workloads, both at the core and at the edge.
- Investing in the right infrastructure vendor with a proven stack, a comprehensive solutions portfolio, and a vast ecosystem enables businesses to gain consistent insights and do so at scale.

Digital infrastructure (DI) is a critical enabler of business innovation and success. The immediate consequence of digital transformation is data — and lots of it. Today's businesses are producing increasing quantities of data. IDC's Global DataSphere study finds that in 2020, 64ZB of data was produced. By 2025, about 181ZB will be generated each year, a threefold increase in just five years. All this data generation is a direct result of rapid digital transformation.

But companies are not just generating data for the sake of doing so. They want the ability to gain real-time or near-real-time insights from these data sets in direct support of their digital transformation initiatives. Further, this insatiable thirst for insights also pushes businesses to explore new ways to generate and compute data. Those that unlock the power of data, while protecting it and creating meaningful impact from intelligent insights in a compliant manner, will be leaders in their respective industries.

AI and analytics are thus top drivers for the digital enterprise. Successful AI and analytics initiatives are linked to business outcomes. Deploying AI and analytics at scale requires well-planned investments in digital infrastructure. It places increasing demands for more fit-for-purpose computing platforms at the core and edge, optimized for data-intensive workloads, security, connectivity, and scale. DI supporting the scale and complexity of AI and analytics requires advanced automation and observability for infrastructure and operations teams.

Business Differentiation in the Era of AI

Businesses are on the cusp of a major paradigm shift. Rapid technological changes mean the use of new approaches like AI can unlock even deeper and more actionable insights from data. This shift will expand opportunities, breakthroughs, and innovation. It also means businesses must take advantage of newer scaling approaches for computationally intensive workloads to fundamentally change when, where, and how quickly they turn data into intelligence. However, adopting these technologies also brings additional risks like data security as well as privacy and regulatory requirements.

Generally speaking, AI and analytics workloads fall into two categories:

- **Training** requires building new models and/or modifying existing models. It involves supervised or unsupervised learning usually in a batch or iterative mode. Use cases include generative AI, computer vision training, retraining, tuning, or optimizing large language models (LLMs), scientific discovery, and financial market modeling. The time to results can be hours, days, or weeks. This process usually requires large clusters of highly performant servers and/or supercomputers. Enterprises also are using fine-tuning, retrieval augmented generation (RAG), and prompt engineering to connect LLMs to their data.
- **Inferencing** requires running trained and optimized models with corporate data, with real-time or near-real-time execution. Use cases include computer vision, natural language processing (NLP), fraud detection, predictive maintenance, and medical imaging. Time to results can be microseconds, milliseconds, or seconds in the worst case. It requires standalone servers that can be deployed in the datacenter or at the edge.

Use Case: Computer Vision AI at the Edge

Computer vision uses AI and machine learning (ML) to derive meaningful information from anything visual (digital images, videos, and schematics) and to take actions or make recommendations based on that information. Computer vision trains machines to perform human functions but uses AI for increased speed, scale, and precision; with the learning capabilities of AI, it can improve outcomes over time.

Computer vision is used in industries ranging from energy and utilities to manufacturing and automotive, with new use cases added each year. For example, companies can use ML to teach computers placed in manufacturing plants to flag defects and other issues during the assembly process.

Use Case: Generative AI

Businesses are leveraging generative AI to gain deep, actionable, and real-time insights in a conversational, human-friendly manner. The models learn patterns from extensive analysis of proprietary and public domain data sets, thus transforming every internal and external facet of the business, including productivity, competitiveness, and customer engagement. With countless applications across various industries, generative AI is a necessary use case for several companies.

Use Case: Natural Language Processing AI

Natural language processing is the creation of systems or data that emulate human language. NLP seeks to build interfaces to accomplish useful tasks with humanlike responses. With the proper inferencing infrastructure, companies can develop and deploy end-to-end AI to power NLP for use cases such as speech, fraud detection, and predictive maintenance.

Investing in Fit-for-Purpose Computing Platforms for AI

Businesses starting off with AI and analytics initiatives must carefully choose the path they wish to take. Those with specific use cases and requirements for customization must investigate the use of fit-for-purpose platforms that complement their investments in software and services. IDC finds that businesses can increase the success rate of their AI initiatives by investing in a combination of on-premises infrastructure, cloud services, optimized inference engines (inherent in the software stack containers), and APIs to bring intelligence to business and enterprise applications. For the specific use cases outlined previously, an on-premises infrastructure yields better return on investment.

The selection criteria for a portfolio of fit-for-purpose hardware platforms must factor in a full-stack approach, in which the vendor has invested in designing, integrating, and optimizing the end-to-end infrastructure stack for AI — comprising accelerated computing hardware, enterprise AI software, and services — to deliver maximized performance, efficiency, and effectiveness of AI and analytics development infrastructure. This means IT organizations must:

- Invest in a purpose-built server for edge AI. These servers are optimized for analyzing video footage from cameras to detect and track objects, people, and behaviors for enhanced operations, security, and smart space experiences. Purpose-built servers suited for smart spaces will often include features such as ruggedization, edge-optimized onboard accelerators, and loss prevention security.
- Invest in platforms designed for generative AI. These servers are optimized for inferencing generative AI models to generate human-centric responses. Servers used for generative AI can accelerate the curation of new and existing content, thus mimicking real-world insights. They are built with features such as onboard accelerators and high-speed interconnects, enabling their use for performance-intensive applications.

Considering HPE

A modern digital strategy — which is much broader than IT strategy — requires businesses to invest in performant and highly distributed infrastructure. By partnering with proven technology vendors like Hewlett Packard Enterprise (HPE) and Intel, organizations can gain timely insights from complex data and process it into actionable information to save costs, grow revenue, and innovate.

HPE's Intel-powered portfolio of infrastructure products can help businesses implement AI at scale, in a secure and sustainable manner. HPE's datacenter and AI solutions coupled with Intel's embedded accelerator engines in their Xeon processors enable companies to turn data efficiently and securely into intelligence, in a manner that best suits the economics of their business.

HPE ProLiant Platforms Accelerate AI Inference Initiatives

Intel-based HPE ProLiant Gen11 servers deliver workload-optimized solutions that help organizations respond quickly to business needs and scale with growth. They are equipped with features that are optimized for AI/ML workloads:

- Increased core performance with instructions per clock cycle (IPC) uplift
- DDR5 memory for more performance, lower power requirements, and greater data consistency per server
- Enhanced processor-to-processor communications for performance, throughput, and efficiency in accelerated solutions

- New processor architectures that deliver higher frequencies to enable more transactions and iterations per cycle
- High core count and memory capacity with greater I/O capacity, plus PCIe 5.0 lanes with improved bandwidth for data-heavy workloads
- Support for software that delivers open scalable acceleration across processors, processors with onboard acceleration, and discrete accelerators

HPE ProLiant DL380a Gen11 with Up to Four GPUs

This server model is optimized for performance-intensive, accelerated computing workloads such as NLP and generative visual AI use cases. It is also optimized for running inference on LLMs. Key features of this server model are:

- Well-balanced I/O performance and capabilities across processors, which are powered by 4th and 5th Gen Intel Xeon Scalable Processors that support up to 64 cores at 350W and 24 DIMMs for DDR5 memory at speeds up to 5600MHz
- Support for up to 4 doublewide GPUs in a 2U server
- Advanced data transfer rates and higher network speeds from the PCIe Gen5 serial expansion bus, with up to 4 x 16 PCIe Gen5 and 2 OCP slots
- New HPE Integrated Lights-Out 6 (iLO 6) server management software that enables IT organizations to securely configure, monitor, and update HPE ProLiant Gen11 servers seamlessly, from anywhere
- Support for up to 3TB total DDR5 memory with 12 DIMM channels per processor that delivers increased performance and lower power requirements

HPE ProLiant DL320 Gen11 with Up to Four GPUs

This server model is purpose-built for AI workloads at the edge, which include computer vision AI use cases. Key features of this server model are:

- Powered by the 4th and 5th Gen Intel Xeon Scalable Processors that support up to 60 cores at 270W and 16 DIMMs of DDR5 memory up to 5,600MTps
- 16 DIMMs per processor for up to 2TB total DDR5 memory with increased memory bandwidth and performance and lower power requirements
- Advanced data transfer rates and higher network speeds from the PCIe Gen5 serial expansion bus, with up to 2 x 16 PCIe Gen5 and 1 OCP 3.0 slots
- New HPE Integrated Lights-Out 6 server management software that enables IT organizations to securely configure, monitor, and update HPE ProLiant Gen11 servers seamlessly, from anywhere
- Support for hot-pluggable or internal high-availability RAID1 NVMe M.2 boot options

Key Benefits of HPE ProLiant AI Solutions

HPE in collaboration with Intel provides a complete infrastructure portfolio designed to power strategic business imperatives and deliver consistent outcomes and business growth. HPE's ProLiant AI solutions offer the following benefits:

- **HPE's AI expertise.** The company's depth of experience is based on thousands of AI and data platform engagements and a robust partner ecosystem of AI experts. HPE can enable businesses to accelerate vision AI, speech AI, generative AI, and other AI workloads paired with high-performance CPU and GPUs. With HPE as a partner, businesses can fast-track modernization, gain faster insights, and maintain control over their AI initiatives.
- **Reduced complexity and cost.** Businesses can accelerate their AI initiatives to increase revenue, reduce support costs, and enhance market competitiveness. HPE's solutions increase flexibility, improve operational efficiency, and increase the innovation potential. With HPE GreenLake, businesses have access to a portfolio of cloud and as-a-service solutions that helps simplify and accelerate AI initiatives.
- **Workload-optimized and scalable computing solutions.** HPE's solutions can be readily deployed at the core and edge. Purpose-built HPE ProLiant servers are designed to support AI from development to deployment and thus to enable data to be converted into actionable insights faster and at scale. HPE can accelerate computer vision, NLP, and various AI workloads paired with high-performance GPUs from Intel and others.

Challenges and Opportunities for HPE

Businesses must have the choice of hardware to meet key business objectives for their AI initiatives, so vendors including HPE need to make their case with customers. Key opportunities when considering HPE's solutions are:

- **Enabling an edge-to-cloud computing continuum.** HPE's products are purpose-built and backed by a broad ecosystem of partners. This means IT teams will have fewer challenges around vendor lock-in, compatibility, and scalability. HPE hardware is optimized for high-performance, low-latency datacenter, network, and edge workloads.
- **Bringing AI everywhere.** Businesses can start adopting AI workloads on the same hardware used for other enterprise workloads. HPE's ProLiant platforms achieve high inference and training performance with built-in accelerators for Intel's Xeon Scalable Processors. With an open and full AI software stack, developers can use their tools of choice while increasing productivity and speeding development time.
- **Improving trust and sustainability.** HPE's platforms can help deliver platform-level power savings and advance sustainability goals. Silicon-based security technologies help deliver a trusted computing environment for businesses that deal with sensitive data and enable companies to gain a consistent security posture for workloads deployed across their infrastructure environments.

Conclusion

Businesses are being disrupted as they enter the era of intelligence. The convergence of a data-driven economy, human insight, and pervasive AI provides organizations an opportunity to grow at an unprecedented pace. Firms that unlock the power of data will become the leaders of the intelligence era. Those that do not risk being left behind. HPE's solutions are designed to help businesses realize the most value from data and thrive in the era of AI.

Businesses that unlock the power of data will become the leaders of the AI era.

About the Analyst

Ashish Nadkarni, *Group Vice President and General Manager, Worldwide Infrastructure and BuyerView Research*

Ashish Nadkarni is group vice president and general manager within IDC's worldwide infrastructure research organization. He leads two teams at IDC: One is focused on IDC's worldwide infrastructure research, and the other manages IDC's BuyerView products.

MESSAGE FROM THE SPONSOR

HPE ProLiant Gen11 is secure, efficient, optimized, and it's engineered for your hybrid world.

At HPE, we understand that in order for your IT initiatives to be successful, you need accelerated performance that can meet the needs of your existing and emerging AI workloads and deliver critical insights that drive tangible benefits. From edge to cloud, the right choice of compute — one that delivers a cloud operating experience built from the ground up with a fundamental foundation security approach — can set you apart from the competition.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com