Solution brief

CPU-OPTIMIZED POWERHOUSE FOR THE ERA OF AGENTIC AI

HPE ProLiant Compute DL394 Gen12 with NVIDIA Vera CPUs

Preparing infrastructure for evolving AI paradigms

In today's fast-moving digital world, organizations are exploring new ways to harness data—whether it's navigating volatile financial markets, powering scientific breakthroughs, or streamlining manufacturing. As AI evolves from simple models to systems that reason, self-correct, and act independently—known as agentic AI—the hardware supporting these workloads must keep pace. These systems rely on continuous loops of decision-making, where the CPU, memory, and data buses work together seamlessly to generate insights, carry out actions, and learn from new information—all with minimal delay. For industries like finance, this means faster trading decisions, more accurate risk assessments, and more effective fraud detection. The challenge is that current servers struggle to handle the intense memory and processing demands of these advanced AI workflows, causing delays that can slow down critical operations. A new, easy-to-deploy solution—one that tackles memory bandwidth, low latency, and performance issues—is essential so organizations can focus on advancing agentic AI and reinforcement learning (RL) rather than wrestling with infrastructure challenges.

Agentic AI and the emerging role of the CPU

The first wave of AI was largely about training and running increasingly powerful models on GPUs. Today, a new generation of agentic AI is expanding the role of AI systems from generating answers to taking actions. Rather than simply responding to prompts, AI agents can plan tasks, use software tools, run code, analyze data, interact with applications, and evaluate their own results.

A useful analogy is to think of the AI model as a project manager and the agents as coworkers. The model reasons about what needs to be done and assigns tasks, but the actual work often happens outside the model itself. For example, an agent may generate Python code, open files, query databases, search for information, or create a spreadsheet. While GPUs remain critical for model inference, most of these actions run on CPUs.

This shift creates a new infrastructure challenge because modern AI systems increasingly operate in a continuous loop between GPU-based reasoning and CPU-based implementation:

— **Agents use tools and software environments.** Tasks such as running code, processing files, querying databases, and interacting with applications are primarily CPU workloads.

— **Sandboxes become essential.** Agent-generated code is often run inside isolated sandbox environments where it can be safely tested, validated, and refined before producing a final result.

— **RL increases implementation demands.** During RL training, models repeatedly interact with environments that evaluate outputs, compute rewards, and return feedback, much of which relies on CPU resources.

— **CPU latency directly impacts AI performance.** Many agent workflows are sequential: the model must wait for a tool call, code implementation, or evaluation to finish before it can take the next step. Slow CPU performance can therefore reduce overall system throughput and increase response times.

As agentic AI scales, CPU performance moves onto the critical path of AI infrastructure. Historically, CPUs were optimized for traditional enterprise workloads and virtualization, where maximizing total core count was often the primary goal. In contrast, agentic AI creates demand for fast, responsive environments that can launch, run, and evaluate thousands of concurrent tasks.

In the era of agentic AI, the CPU is no longer merely supporting the model—it is becoming a key determinant of how effectively AI systems can act, learn, and deliver real-world outcomes.

Optimized infrastructure for the future of agentic AI

To meet the evolving demands of agentic AI and complex data workloads, organizations require a server platform optimized for high-performance processing and rapid data movement. Expanding its industry-leading server portfolio with the introduction of the HPE ProLiant Compute DL394 Gen12, HPE now provides a solution for this new era of agentic AI. Engineered to deliver outstanding processing power and high memory bandwidth, the DL394 is ideal for sophisticated financial modeling, real-time decision-making, and RL applications. Powered by the NVIDIA® Vera CPU, this 2U server helps ensure deterministic performance with low latency, high efficiency, and increased throughput, enabling AI workloads to be orchestrated seamlessly across resources.

* Product configuration and appearance may vary

Figure 1. HPE ProLiant Compute DL394 Gen12

Accelerating agentic AI with the NVIDIA Vera CPU

At the heart of the HPE ProLiant Compute DL394 Gen12 is the NVIDIA Vera CPU, a processor specifically designed to address the emerging demands of agentic AI, RL, and large-scale AI infrastructure. Unlike traditional enterprise workloads, agentic AI systems create continuous cycles of reasoning, implementation, evaluation, and feedback. These workflows place unique demands on CPU resources, requiring high throughput, low latency, fast memory access, and predictable performance under sustained load.

The NVIDIA Vera CPU was engineered with these requirements in mind. It is designed to support the CPU-intensive activities that increasingly define modern AI systems, including tool invocation, orchestration, sandboxed code implementation, analytics pipelines, data processing, and software services that operate alongside accelerated AI models. According to NVIDIA, Vera is purpose-built for agentic AI workflows, helping keep AI pipelines moving efficiently while supporting both accelerated and CPU-centric workloads.

Several architectural innovations contribute to Vera's suitability for next-generation AI environments:

— **88 NVIDIA-designed Olympus CPU cores** provide high-performance processing for orchestration, runtime environments, analytics, and agent implementation workloads.

— **Spatial multithreading technology** enables each core to run two tasks simultaneously while maintaining consistent and predictable performance, an important capability for multitenant AI environments and large numbers of concurrent agents.

— **Up to 1.2 TB/s of LPDDR5X memory bandwidth** helps feed data-intensive AI, analytics, and RL workloads while supporting efficient movement of information throughout the system.

— **Advanced Arm®-compatible Olympus cores and NVIDIA Scalable Coherency Fabric (SCF)** are designed to provide strong single-thread performance, efficient data movement, and predictable throughput under demanding workloads.

By integrating the NVIDIA Vera CPU into the HPE ProLiant Compute DL394 Gen12, organizations gain a platform specifically designed for this new era of AI infrastructure—one where the speed of implementation, orchestration, and feedback is becoming just as important as the intelligence of the models themselves.

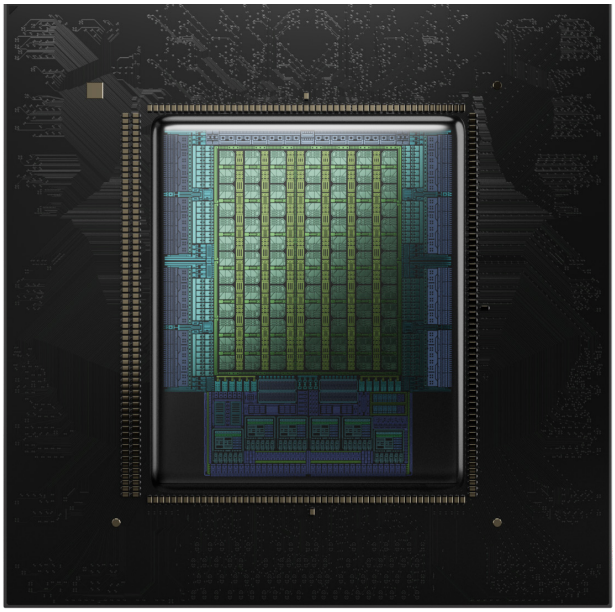

Figure 2. NVIDIA Vera CPU

HPE ProLiant Compute DL394 Gen12—Delivering a consistent, reliable HPE ProLiant experience for next-gen AI

While the NVIDIA Vera CPU provides the foundation required for agentic AI and RL workloads, enterprise deployment requires more than processing power alone. Organizations must also address security, manageability, operational efficiency, and lifecycle management across increasingly complex infrastructure environments.

The HPE ProLiant Compute DL394 Gen12 extends the benefits of the NVIDIA Vera architecture through the trusted HPE ProLiant experience that enterprises have trusted and relied upon for years. Designed for large-scale deployment and day-to-day operational simplicity, the platform combines advanced security capabilities, comprehensive infrastructure management tools, and intelligent automation to help organizations deploy and operate AI infrastructure with confidence.

From establishing a hardware-rooted chain of trust, to simplifying infrastructure management across the data center, to enabling proactive operations through AI-driven insights, the DL394 provides the enterprise-ready capabilities required to support next-generation AI environments. The following technologies play a key role in delivering this consistent and reliable HPE ProLiant experience:

Security built into the foundation—HPE iLO 7 with silicon root of trust, a robust security foundation embedded within the server hardware, helps ensure that every layer of firmware and software loads securely and is verified from the moment the server is powered on. This results in providing an unbreakable chain of trust, protecting servers against firmware attacks, unauthorized access, and tampering, thereby helping ensure the highest level of data integrity and system reliability.

Simplifying infrastructure operations with HPE OneView—HPE OneView works within the data center to help automate and streamline IT operations by providing a centralized interface to manage and monitor servers, storage, and networking devices. This helps ensure seamless integration and efficient management of diverse infrastructure environments.

AI-driven operations with HPE Compute Ops Management—HPE Compute Ops Management helps ensure smooth enterprise operations with automation and AI-driven insights from data center to edge, leveraging a unified management platform. Enable operators to react faster and gain greater control, from forecasting energy costs to managing a global server footprint. Boost productivity of IT staff by quickly pinpointing problem areas through dashboards, intelligent alerts, and a global map of all servers with health status and activity.

Conclusion: Built for the next era of AI infrastructure

As AI evolves from generating responses to completing actions, infrastructure requirements are changing just as rapidly. Agentic AI and RL introduce new demands on CPU performance, memory bandwidth, latency, and system responsiveness, making the CPU a critical component of overall AI system efficiency.

The HPE ProLiant Compute DL394 Gen12, powered by NVIDIA Vera CPUs, is designed to address these emerging requirements. By combining CPU architecture optimized for agentic AI workloads with the trusted security, management, and operational capabilities of the HPE ProLiant platform, organizations can deploy AI infrastructure that is ready for both today's workloads and tomorrow's innovations.

Whether supporting advanced financial modeling, large-scale RL, or next-generation agentic AI applications, the HPE ProLiant Compute DL394 Gen12 provides the performance, efficiency, and enterprise readiness needed to help organizations accelerate insights, automate decisions, and leverage new opportunities in the age of AI.

Learn more at

HPE ProLiant Compute DL394 Gen12

Visit HPE.com

Chat now

a00159684ENW

HEWLETT PACKARD ENTERPRISE

hpe.com