IDC LINK
Real-time IDC Research® opinion on industry news, trends and events

Penguin Solutions Announces Expanded OriginAI Solution to Accelerate AI Factory

June 25, 2024

By: Matthew Eastwood

IDC's Quick Take

Penguin Solutions, with decades of expertise in high-performance computing (HPC), artificial intelligence (AI), and the Internet of Things (IoT), is enhancing its OriginAI solution by deeper integration with NVIDIA technology. This expanded solution, supported by advanced cluster management software and expert services, aims to provide predictable performance, high throughput, and cost efficiency for complex, data-intensive workloads. Penguin's strategic focus on validated, predefined AI architectures aligns with industry trends to reduce deployment complexity and enhance manageability. As enterprises increasingly adopt generative AI (GenAI), Penguin addresses challenges such as scalability, cost, and compliance by emphasizing AI-ready infrastructure that balances performance and sustainability. Penguin's broad experience and collaboration with NVIDIA position the company as a key player in the competitive AI infrastructure market, meeting the growing demand for powerful and reliable AI solutions.

Product Announcement Highlights

Penguin Solutions provides technology services and solutions in high-performance computing, artificial intelligence, and the Internet of Things. Its offerings, which extend from edge computing to cloud applications, focus on integrating emerging technologies to streamline complex infrastructures and support digital strategies across a variety of market segments and industries. With more than 40 years of experience in the market, Penguin Solutions aims to help organizations improve performance and achieve quicker time-to-market outcomes.

Penguin Solutions has announced several enhancements to its OriginAI solution, incorporating NVIDIA technology into predefined AI architectures for improved implementation and performance. The expanded offering is supported by Penguin's advanced cluster management software and expert services, ensuring scalability and ease of deployment. The updates aim to deliver predictable performance, manage demanding workloads effectively, and optimize GPU utilization for cost efficiency, all backed by Penguin's experience and a proven history in high-performance computing and AI deployments. The announcement includes the following:

- **Expansion of the OriginAI solution:** Penguin Solutions announced the expansion of its OriginAI solution, now including validated, predefined AI architectures that incorporate NVIDIA technology, aimed at streamlining AI implementation and enhancing cluster performance. These architectures are based on 1-pod, 4-pod, and 16-pod configurations that can scale from 256 to more than 16,000 GPUs.
- **Backed by expertise and software:** The expanded OriginAI infrastructure is supported by Penguin's intelligent cluster management software and expert services, designed to facilitate scalable AI infrastructure deployment and manageability.

- **Predictable performance and return on investment (ROI):** The solution aims to provide predictable AI cluster performance and support customer return on investment for clusters ranging from hundreds to thousands of GPUs.
- **Infrastructure for demanding workloads:** OriginAI provides robust infrastructure solutions for critical workloads, integrating latest-generation hardware, advanced management software, and expert services in configurations scaling up to over 16,000 GPUs.
- **High performance and cost control:** The architectures are developed to optimize GPU performance and cost efficiency, ensuring high throughput and greater than 95% overall cluster efficiency.
- **Pre-deployment validation:** Penguin uses an in-factory burn-in and integration environment to validate AI cluster performance and ensure production readiness, confirming that customers receive the expected performance and ROI from the deployment.
- **Expertise and experience:** Penguin Solutions, with rich experience in HPC and having deployed over 75,000 GPUs, positions itself as a strategic partner for AI and HPC solutions, serving notable clients like Georgia Tech, Meta, Sandia Labs, and the U.S. Navy.

IDC's Point of View

Penguin Solutions has a long history of collaboration with NVIDIA, having previously been named HPC Preferred OEM Partner of the Year for its deployment of large-scale hybrid NVIDIA GPU-accelerated servers to high-profile customers. Together, the two companies have demonstrated the transformative power of GPUs in computational science, highlighting innovations such as earthquake visualizations with NVIDIA GPUs. This experience is foundational to Penguin's efforts to embrace AI workloads, particularly in the realm of generative AI.

IDC notes that selling an HPC infrastructure solution typically involves convincing decision-makers focused on maximizing computational power and efficiency for complex simulations and large-scale data processing. In contrast, selling an AI infrastructure solution requires addressing the needs of decision-makers who prioritize machine learning capabilities, data analytics, and real-time processing to drive innovative AI applications and insights. When selling AI infrastructure solutions, Penguin will need to engage with CIOs, CDOs, CTOs, data scientists, machine learning engineers, IT operations managers, business unit leaders, procurement officers, security and compliance officers, and AI and analytics program managers to address their diverse priorities and ensure the solution meets organizational goals, technical requirements, and regulatory standards.

Penguin Solutions' recent expansion of the OriginAI solution positions the company within the competitive landscape of enterprise AI infrastructure. Like other infrastructure providers, Penguin is intensifying its focus on AI through collaboration with NVIDIA to build relevance across cloud and edge environments. The company is enhancing its AI offerings by integrating NVIDIA technology and focusing on scalability and performance. Penguin is responding to the increasing demand for high-performance, reliable AI infrastructures that can support complex, data-intensive workloads across various deployment scenarios.

Penguin's focus on providing validated, predefined AI architectures with predictable performance and ROI aligns with industry trends toward solutions that reduce complexity and streamline AI deployment. This approach emphasizes flexibility, scalability, and enhanced management capabilities to cater to a wide range of enterprise needs. Such strategic alignments are critical as companies look to differentiate themselves in a market where enterprises are keen on minimizing technical debt and enhancing AI readiness.

Enterprises adopting GenAI face challenges such as scalability, cost, compliance, technical expertise, and sophisticated environmental management like cooling and control. To address these, Penguin's strategy aligns with IDC's framework for modern enterprise computing, emphasizing AI-ready infrastructure that balances performance, cost-effectiveness, and sustainability. Spending on AI-ready digital infrastructure is projected to reach $48 billion by 2027, driven by GenAI demands. Despite the adoption of GenAI technologies, organizations face excessive costs and inefficient data silos. To modernize systems and strategically place infrastructure across cloud, datacenter, and edge locations, businesses focus on innovations like advanced GPU acceleration; next-generation storage; high-performance computing; high-speed, low-latency networking; and sustainable datacenter technologies.

Overall, Penguin Solutions is carving out a position in a highly competitive market by focusing on performance optimization and deployment readiness, which is crucial for enterprises looking to rapidly scale their AI capabilities. By ensuring high cluster efficiency and leveraging extensive GPU deployments, Penguin aims to meet the growing needs for AI infrastructure that is both powerful and cost-effective, aligning with broader industry trends toward sustainable and efficient AI solutions.

Subscriptions Covered:
AI and Generative AI Infrastructure Stacks and Deployments