Dell AI Infrastructure Updates

STEVE MCDOWELL, CHIEF ANALYST
10/21/24

CONTEXT

Dell Technologies made significant strides in AI infrastructure with its Integrated Rack 7000 (IR7000) launch and associated platforms for AI and HPC. The announcements introduce enhancements in computing density, power efficiency, and data management, catering specifically to AI workloads.

This Research Note details the technical features of Dell's latest offerings, including the IR7000, new PowerEdge servers, PowerScale storage innovations, and data management tools. It evaluates their impact on the competitive landscape.

DELL INTEGRATED RACK 7000

The Dell Integrated Rack 7000 (IR7000) is a high-density, Open Compute Project (OCP) standards-based rack designed to meet the computing and power demands of large-scale AI and HPC deployments. It is engineered to offer superior scalability, power efficiency, and cooling capabilities for next-generation AI workloads and infrastructure requirements.

KEY TECHNICAL FEATURES

1. **High Density Design**:
    - The IR7000 is a 21-inch rack optimized for maximizing the density of CPUs and GPUs. It supports cutting-edge architectures and future technology generations and is designed specifically to handle high-performance computing setups for AI and large-scale data processing.
2. **Enhanced Cooling**:
    - The rack is purpose-built for liquid cooling, supporting advanced cooling systems capable of dissipating up to 480KW of heat. This

makes it ideal for handling the thermal output of future high-power CPUs and GPUs.

- It captures nearly 100% of the generated heat, enhancing cooling efficiency and overall energy savings.

3. **Scalability and Future-Proofing**:
   - The IR7000 features wider and taller server sleds to accommodate next-gen, larger CPU and GPU architectures and ensure compatibility with evolving hardware.
   - The rack is designed for multi-generational use, allowing organizations to future-proof their AI and HPC investments.
4. **Sustainability and Energy Efficiency**:
   - The IR7000's integrated power and cooling systems provide superior energy efficiency, making it ideal for organizations focused on sustainable computing.
   - Its ability to accommodate liquid cooling not only reduces the overall data center footprint but also minimizes the energy required for cooling.
5. **Flexibility in Networking**:
   - The rack supports Dell and third-party off-the-shelf networking, giving organizations flexibility in choosing their preferred networking infrastructure.
6. **Seamless Integration**:
   - The IR7000 integrates with Dell's Integrated Rack Scalable Systems (IRSS), offering a fully integrated plug-and-play infrastructure that simplifies deployment and optimizes performance for AI workloads.

USE CASES

The IR7000 is particularly suited for industries and organizations requiring high-density computing environments, such as:

- Large-scale AI training and inferencing tasks.
- HPC workloads.

- Research institutions and enterprises working with intensive computational tasks.
- Data centers focus on sustainable, energy-efficient operations.

DELL POWEREDGE XE9712

The Dell PowerEdge XE9712 is Dell's new high-performance server platform for large-scale AI workloads, particularly for tasks like training LLMs and real-time inferencing.

It's part of Dell's AI Factory solutions, providing dense GPU acceleration and advanced compute capabilities optimized for AI and HPC. Below are its detailed technical features and capabilities:

KEY TECHNICAL FEATURES

1. **High GPU Density**:
    - The PowerEdge XE9712 is designed to maximize GPU density, making it capable of handling the most demanding AI tasks. It supports up to 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs in a single rack-scale design.
    - The 72-GPU NVLink domain allows all GPUs in the system to function as a unified GPU, significantly boosting scalability and performance for AI applications.
2. **Optimized for Large-Scale AI**:
    - This server is specifically designed for trillion-parameter LLM training and inferencing, offering up to 30x faster real-time performance for these tasks.
    - Its rack-scale architecture enables efficient handling of the massive compute and memory requirements typical in AI workloads like natural language processing, computer vision, and deep learning.
3. **Liquid Cooling Efficiency**:
    - The XE9712 incorporates liquid-cooled NVIDIA GB200 NVL72 GPUs, providing significant thermal management improvements over traditional air-cooled systems. These GPUs are up to 25x more

power-efficient than their air-cooled counterparts, making the system highly efficient for power-intensive AI training.

4. **Performance Scalability**:
    - The PowerEdge XE9712's architecture allows for modular scalability, enabling organizations to scale their AI infrastructure based on evolving needs.
5. **AI Workload Optimization**:
    - The server is optimized for superior performance for various AI workloads, including LLM training, inferencing, and deep learning. Its support for the NVIDIA Grace and Blackwell architectures ensures compatibility with the latest advancements in AI processing, making it ideal for real-time AI applications.
6. **Networking and I/O**:
    - The XE9712 is built with high-speed networking capabilities. It integrates NVIDIA NVLink to interconnect GPUs and provides extremely low-latency and high-bandwidth data transfers across the system.

USE CASES

The PowerEdge XE9712 is well-suited for:

- **AI Training at Scale**: Large organizations and AI research labs require massive computational power to train trillion-parameter models.
- **LLM Deployment**: Enterprises deploying large-scale natural language models for real-time applications, such as AI-based assistants, chatbots, or translation services.
- **High-Performance Computing**: Industries such as healthcare, automotive, and finance, where real-time data analysis and decision-making are key.

DELL POWEREDGE M7725

The Dell PowerEdge M7725 is a high-performance server designed for dense computing environments, optimized for AI, HPC, and research-intensive workloads. It is particularly suited for organizations requiring significant

computational power, such as government, fintech, and academic institutions. The server is built to maximize compute density, serviceability, and energy efficiency, leveraging advanced CPU architectures and cooling technologies.

KEY TECHNICAL FEATURES

1. **Compute Density**:
   - The PowerEdge M7725 delivers exceptional compute density with 64 or 72 two-socket nodes per rack, supporting 5th Gen AMD EPYC CPUs. This configuration enables 24K to 27K cores per rack, providing substantial processing power for compute-intensive tasks.
2. **Advanced Cooling Technologies**:
   - The server is designed with both Direct Liquid Cooling (DLC) and air cooling options, enhancing its flexibility for a variety of deployment scenarios.
     - Direct Liquid Cooling is applied directly to the CPUs, ensuring optimal thermal management, which is critical for high-density compute environments.
     - For systems requiring air cooling, the server features a quick-connect mechanism that integrates with the rack's cooling infrastructure, ensuring efficient heat dissipation even under heavy workloads.
3. **Serviceability and High-Speed I/O**:
   - The Front IO slots provide high-speed connectivity and facilitate easy access for servicing and upgrades. They ensure seamless connectivity to demanding applications, enabling rapid data transfers and low-latency operations.
   - This design improves the ease of maintenance and enhances system serviceability, minimizing downtime during component replacement or upgrades.
4. **AI and HPC Optimization**:
   - The PowerEdge M7725 is optimized for research, government, and other high-demand sectors that require advanced AI and HPC capabilities. Its dense compute architecture and high-speed connectivity make it well-suited for workloads like machine learning, simulations, and real-time data analysis.
5. **Energy Efficiency and Sustainability**:
   - The server is engineered to be energy-efficient, and the integration of direct liquid cooling reduces energy consumption and enables more sustainable deployments.

- Its compact form factor also contributes to energy savings by allowing more compute power per rack, reducing the data center's overall energy requirements.

6. **AMD EPYC CPU Architecture**:
   - Powered by 5th Gen AMD EPYC processors, the M7725 leverages AMD's high-core-count architecture to deliver significant performance improvements for multi-threaded applications. The server can be configured to maximize core density and memory bandwidth, which is crucial for applications that require extensive parallel processing.

USE CASES

The PowerEdge M7725 is particularly well-suited for organizations and sectors that require high-performance compute capabilities and efficient data handling:

- **Research and Academic Institutions**: Supporting scientific simulations, big data analysis, and AI research.
- **Government and Defense**: Handling large-scale simulations, encryption, and real-time decision-making processes.
- **Financial Services (FinTech)**: Processing massive datasets for real-time trading, risk analysis, and fraud detection.
- **Higher Education**: Powering data-intensive applications in computational research, artificial intelligence, and machine learning.

UNSTRUCTURED STORAGE INNOVATIONS

Dell's new Unstructured Storage Innovations are focused on enhancing AI application performance, simplifying global data management, and improving the efficiency of unstructured data handling for AI workloads. These innovations are designed to address the growing demands of AI and machine learning, particularly in terms of storage capacity, data management, and the speed of data access.

Key updates to Dell's unstructured data storage portfolio include improvements to Dell PowerScale, a scalable storage system, and the Dell Data Lakehouse platform, which offers robust data management capabilities for AI workflows.

Key Features of Dell's Unstructured Storage Innovations:

1. **Dell PowerScale Enhancements**:

- **NVIDIA DGX SuperPOD Certification**:
  - Dell PowerScale has become the first Ethernet storage system certified for NVIDIA DGX SuperPOD, a high-performance AI infrastructure platform. This certification ensures that PowerScale can deliver high-performance data storage and management for AI workloads integrated with NVIDIA's compute platforms.
- **Denser Storage**:
  - PowerScale now supports 61TB drives, significantly increasing storage capacity. This higher density allows customers to store larger datasets, which are crucial for training and fine-tuning AI models. The increase in storage efficiency helps reduce the overall data center footprint by 50%, making it more cost-effective and space-efficient for organizations managing large amounts of data.
- **Improved AI Performance**:
  - Front-end support for NVIDIA InfiniBand and 200GbE Ethernet adapters enhances data transfer speeds, increasing throughput by up to 63%. This improvement is crucial for AI workloads that require high-bandwidth, low-latency data access, enabling faster model training and real-time data processing.

2. **Dell Data Lakehouse Platform**:
   - **Metadata Management and Enhanced Discoverability**:
     - The Dell Data Lakehouse integrates advanced metadata management tools to improve data discoverability and insights. These tools allow users to access data faster and make smarter decisions by leveraging PowerScale metadata and the Dell Data Lakehouse for centralized data management.
     - A forthcoming open-source document loader for NVIDIA NeMo services RAG frameworks is aimed at speeding up data ingestion and reducing the compute and GPU costs associated with processing unstructured data in AI workloads.

- **Vectorization and Semantic Search**:
    - The Data Lakehouse is designed to support vector databases and semantic search capabilities. It allows AI models to perform more accurate queries and searches by understanding the meaning and context of data, not just its structure. This enables AI models to return more relevant and context-aware results, improving decision-making and AI-driven insights.

3. **Storage Capacity and Efficiency**:
    - 61TB QLC (Quad-Level Cell) SSDs have been introduced, nearly doubling the previous storage capacity (from 30.72TB) per drive. This increase supports larger AI models, enabling more extensive data training with reduced physical space and power consumption.

4. **Global Data Management and Flexibility**:
    - The Data Lakehouse platform offers new capabilities like disaster recovery, automated schema discovery, and comprehensive management APIs, which simplify the administration and scaling of data pipelines. This ensures high availability, easier management, and reduced downtime for critical AI applications.
    - The platform supports open table formats, such as Apache Iceberg, and will extend to support vector databases.

5. **AI-Specific Optimization Services**:
    - Dell offers Optimization Services for Data Cataloging and Implementation Services for Data Pipelines. These services focus on enhancing the accessibility and quality of data by providing discovery, organization, automation, and integration solutions for AI models.

ANALYSIS

Dell's announcement puts significant pressure on competitors such as HPE, NetApp, and IBM, which are also vying for leadership in the AI infrastructure space. By offering a complete stack—ranging from high-density compute with the IR7000 and PowerEdge servers to optimized storage with PowerScale and

the Data Lakehouse—Dell maintains its position as a one-stop shop for enterprise AI needs.

This most recent announcement sets Dell solutions at the center of an integrated, high-performance platform. Its holistic approach, covering compute, storage, and data management, gives it a competitive edge in the fast-evolving AI landscape, especially as enterprises seek more comprehensive and flexible AI infrastructure to meet their growing needs.