OriginAI

Leveraging a comprehensive reference architecture for AI that significantly reduces time-to-insight

PENGUIN COMPUTING

Contents

Solution-at-a-Glance

Features

- NVIDIA DGX A100 Server with NVIDIA DeepOps Management Software
- Mellanox HDR Infiniband Network by NVIDIA
- Penguin Computing ActiveData Solution with WekaIO WekaFS

Benefits

- Reduce time-to-insight.
- Maximize the performance and utility of high-value AI systems.
- Ensure increased productivity and highest ROI.
- Support scalability and flexibility.
- Ensure data security, resilience, and governance without compromising performance.
- Meet the demanding requirements of AI and analytics models.

The Challenge

Increasingly, organizations are recognizing the business potential artificial intelligence (AI), machine learning (ML), and deep learning (DL) represent; business derived from AI is expected to reach $3.9 trillion by 2022 (Gartner, 2018). As organizations move to adopt these technologies, CIOs are faced with the challenge of implementing complex new systems that are performant, bespoke, and secure.

In this highly specialized endeavor, the odds for success are stacked against CIOs: it is predicted that 85% of AI projects will ultimately not deliver for their organizations (Gartner, 2018). In order to be successful, every component of the AI infrastructure must be expertly tuned to support an organization's unique AI workload. Challenged by critical gaps in expertise, organizations are frequently unable to optimize, resulting in data path bottlenecks that render expensive resources idle, leading to lost productivity and increased time to insight. Adding to the complexity, the speed at which AI evolves demands flexible, scalable architectures that can keep pace as technology advances.

For organizations that are new to high performance computing (HPC) and high performance data analytics (HPDA) systems, developing AI infrastructure can be costly and time consuming, taking an estimated two to three years of internal research to bring AI-based products to market. In fact, 78% of AI projects are stalled even before they can be deployed (Dimensional Research, 2019).

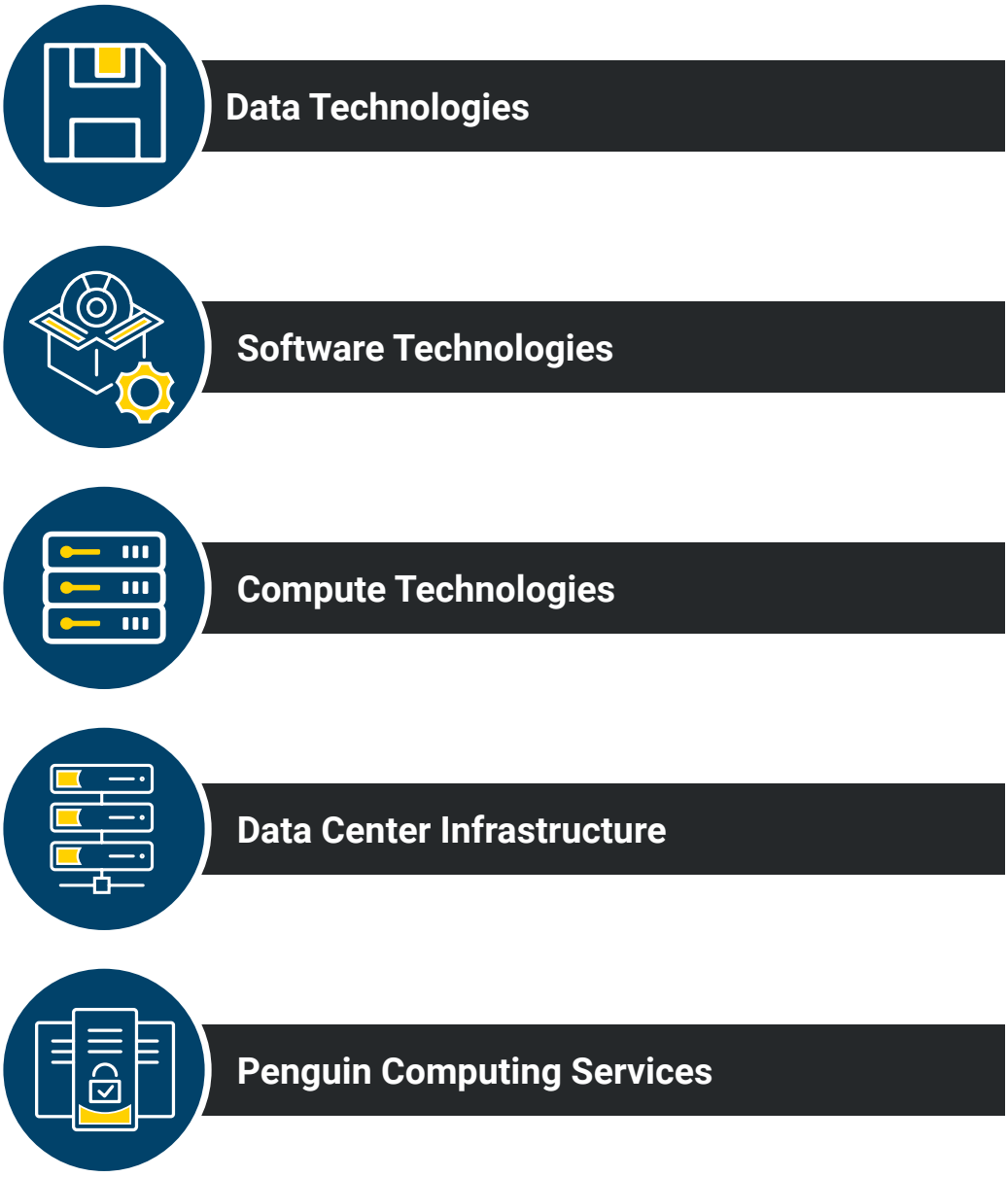
Realizing a need for a comprehensive AI solution that encompasses architectural design, hardware and software services, hosting, and deployment, CIOs are increasingly seeking the expertise of single solution providers.

78% of AI projects are stalled even before they can be deployed.

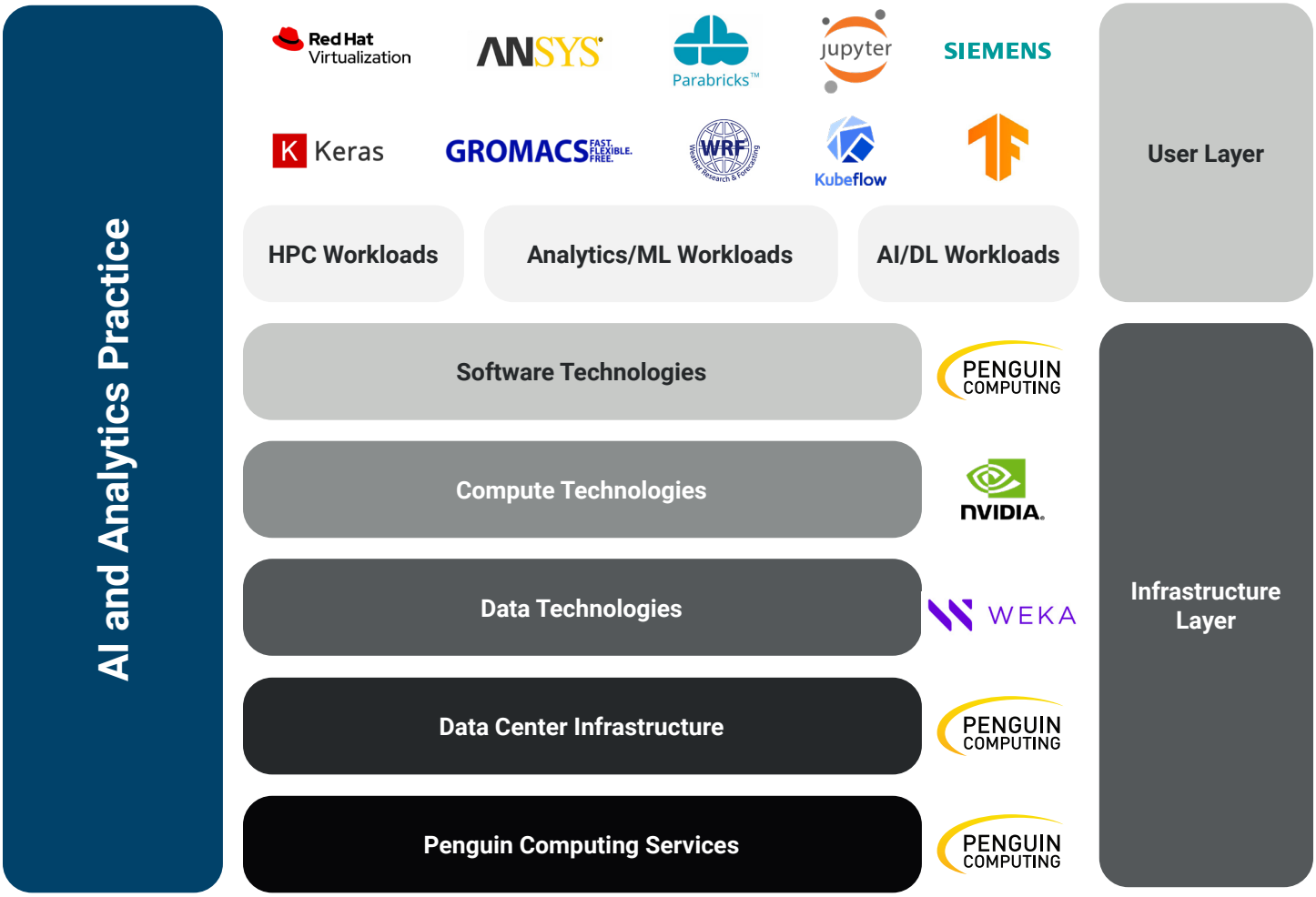
Penguin Computing OriginAI

Penguin Computing partnered with NVIDIA Corporation, WekaIO, Inc., and Mellanox Technologies to create OriginAI™, a comprehensive, end-to-end solution for datacenter AI that breaks down software, server/switch hardware, data storage and governance, and infrastructure barriers that cause AI projects to fail or stall. A full-service consultancy, Penguin Computing's Analytics Practice acts as a single point of reference for hardware, software, architectural design, hosting, and management, enabling organizations to focus their time and resources on the business and human challenges in bringing AI projects to production. By providing organizations with a comprehensive solution, Penguin Computing radically reduces the time to insight from years to months.

OriginAI includes:



Penguin Computing OriginAI Components



Software Technologies

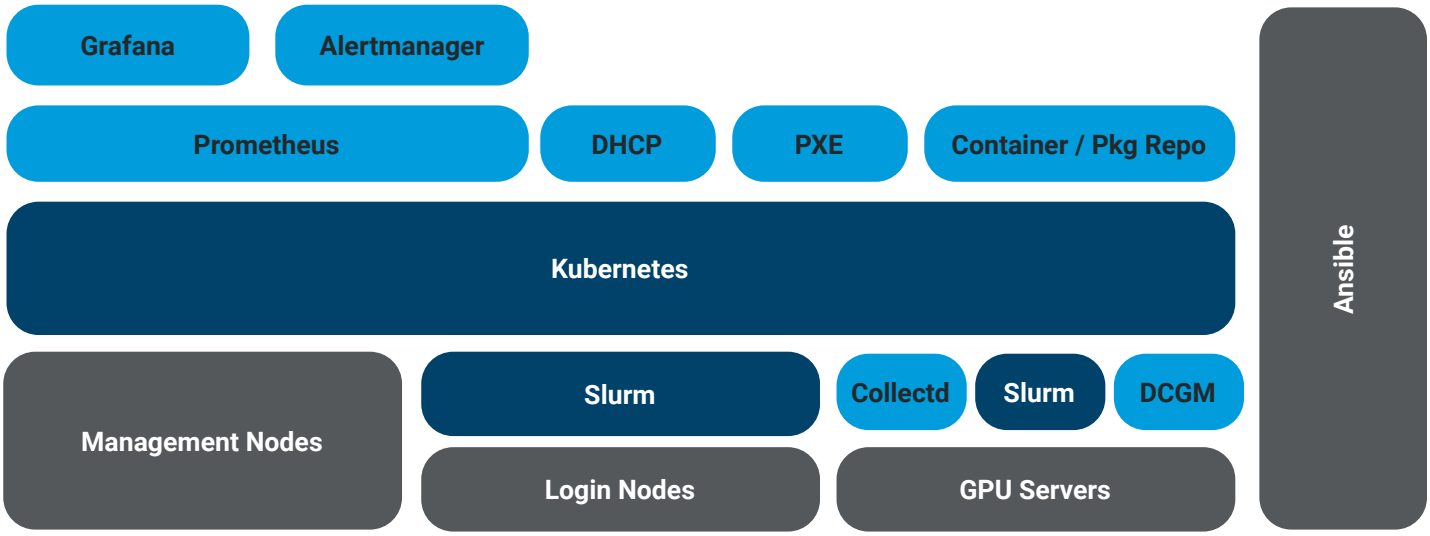
Penguin Computing OriginAI uses NVIDIA AI software to provide a high-performance DL training environment for large-scale, multi-user AI software development teams. It includes the DGX™ operating system (DGX OS), cluster management, orchestration tools and workload schedulers (DeepOps management software), NVIDIA libraries and frameworks, and optimized containers from the NGC container registry.

To provide additional functionality, the DeepOps management software includes third-party open-source applications and visualization tools recommended by NVIDIA which have been tested to work on DGX servers with the NVIDIA AI software stack.

NVIDIA DeepOps

The DeepOps management software is composed of various services running on the Kubernetes container orchestration framework for fault tolerance and high availability. Services are provided for network configuration (DHCP) and fully-automated DGX OS software provisioning over the network (PXE). The DGX OS software can be automatically re-installed on demand by the DeepOps management software.

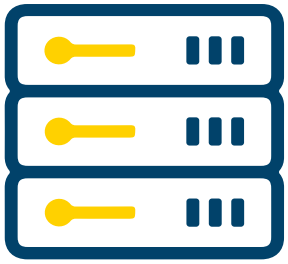
NVIDIA DeepOps Management Software



The DeepOps management software leverages the Ansible configuration management tool. Ansible roles are used to install Kubernetes on the management nodes, install additional software on the login and DGX systems, configure user accounts, configure external storage connections, and install Kubernetes and Slurm schedulers, as well as performing day-to-day maintenance tasks such as new software installation, software updates, and GPU driver upgrades.

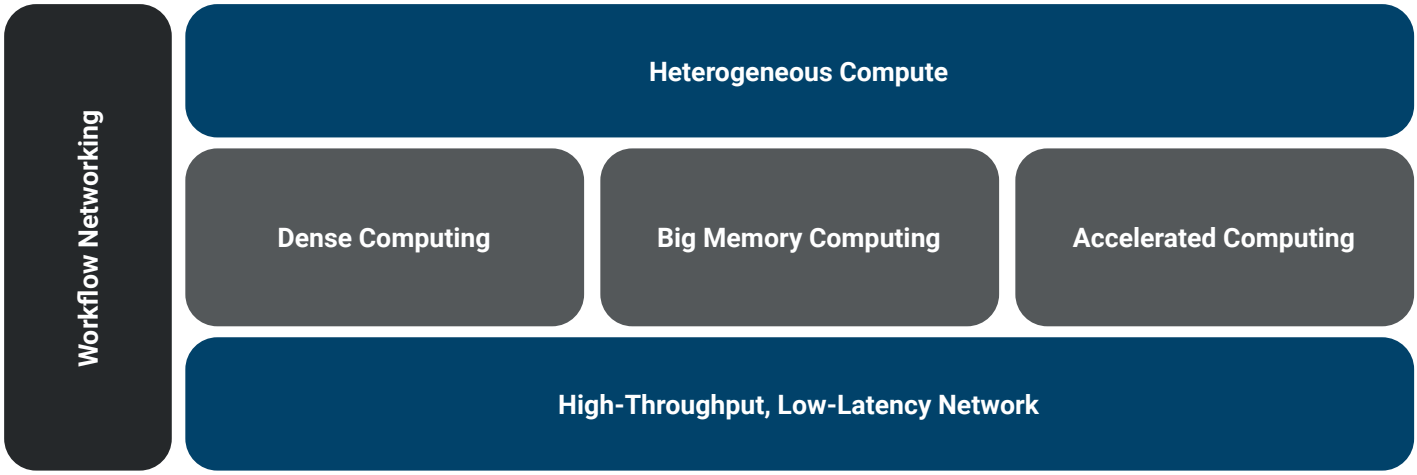
DeepOps monitoring utilizes Prometheus for server data collection and storage in a time-series database. Cluster-wide alerts are configured with Alertmanager, and system metrics are displayed using the Grafana web interface. For sites required to operate in an air-gapped environment or needing additional on-premises services, a local container registry mirroring NGC containers, as well as Ubuntu and Python package mirrors, can be run on the Kubernetes management layer to provide services to the cluster.

Kubernetes runs management services on management nodes. Slurm runs user workloads and is installed on the login node as well as the DGX compute nodes. Slurm provides advanced HPC-style batch scheduling features including multi-node scheduling.



Hardware Technologies

Penguin Computing partners with industry leaders, such as NVIDIA and WekaIO to design and build workload-specific computing platforms that accelerate time to insight. The design process includes in-system device-to-device bandwidth and resource optimizations and system-to-system communication optimizations. These design choices ensure that the underlying computing and networking infrastructure are optimized for the workloads that customers will run on them.



The NVIDIA DGX A100 is the universal system for all AI infrastructure, from analytics to training to inference. Combining the DGX A100 with the Mellanox HDR Infiniband Network enables in-network computing through the Co-Design Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ technology results in an order of magnitude of application performance improvements.

NVIDIA DGX A100 Server

- Universal system for all AI infrastructure, from analytics to training to inference
- 5 petaFLOPS of performance in a 6U form factor, setting a new bar for compute density
- Eight integrated A100 GPUs for unprecedented acceleration
- Optimized for NVIDIA CUDA-X™ software

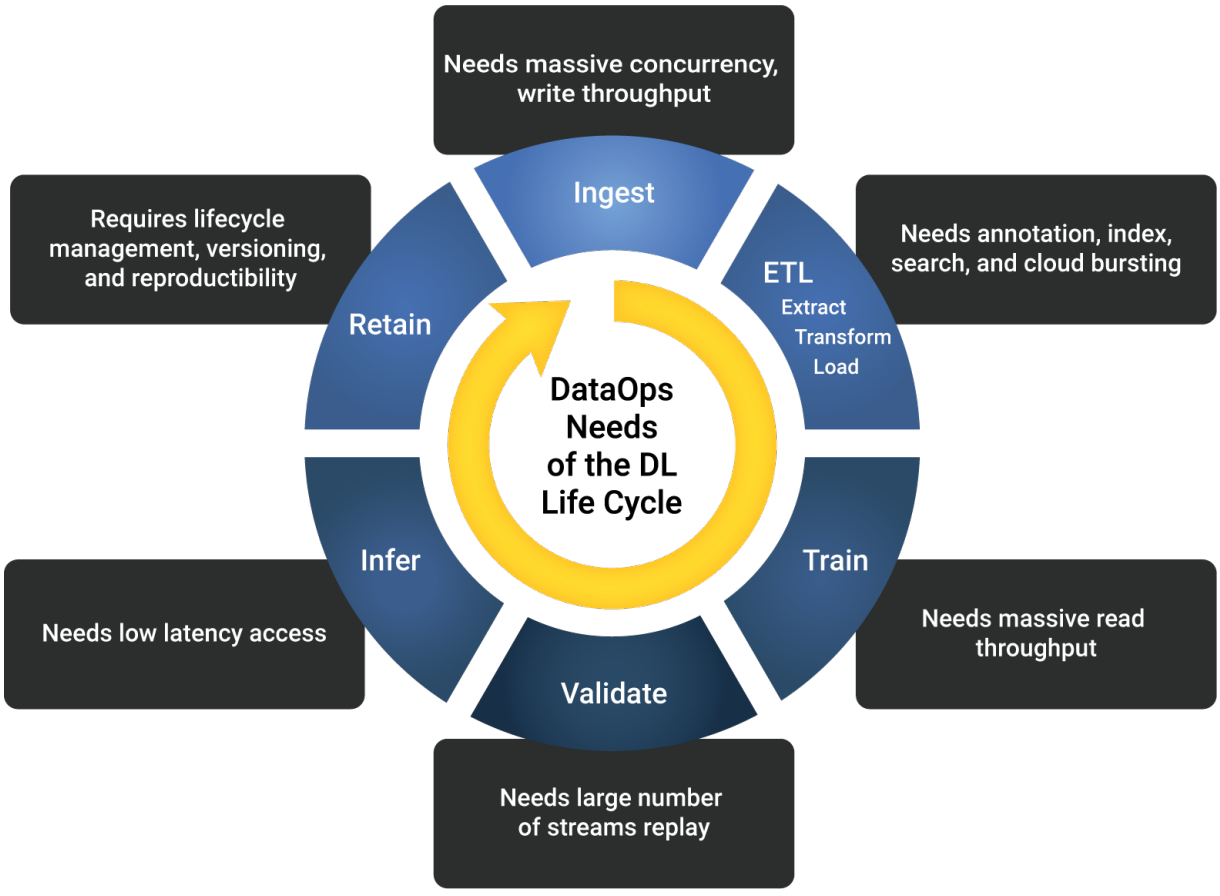
Mellanox QM8700 HDR Infiniband by NVIDIA

- 16Tb/s of non-blocking bandwidth with sub 130ns port-to-port latency
- Forty 200Gb/s full bi-directional bandwidth ports
- Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) technology enables in-network computing to accelerate communications frameworks, resulting in order of magnitude application performance improvements

Data Technologies

As data-intensive workloads scale, it's critical to implement data-driven, software-defined architectures that meet the demands of large data sets. Optimized, accelerated data platforms promise an immediate and tangible solution for delivering discovery and insight from machine-generated data. These platforms combined with accelerated compute and the right software create a new storage category. This approach provides a data store that delivers an enterprise-ready, unified data platform that performs across the entire environment, while also providing essential data security, resilience, and governance. This type of platform is a requirement for data management in the era of AI.

Artificial Intelligence Data Life Cycle

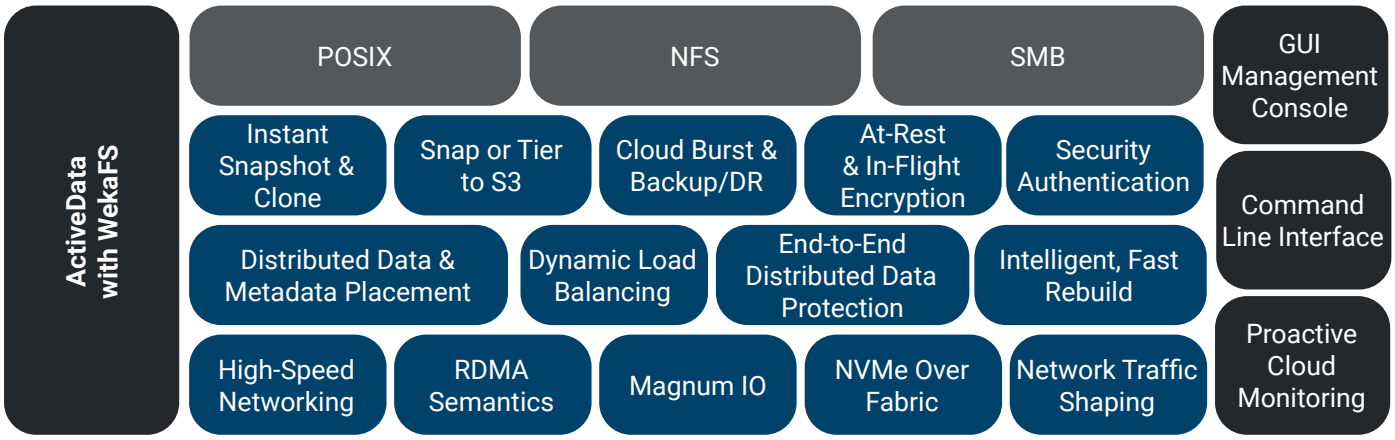


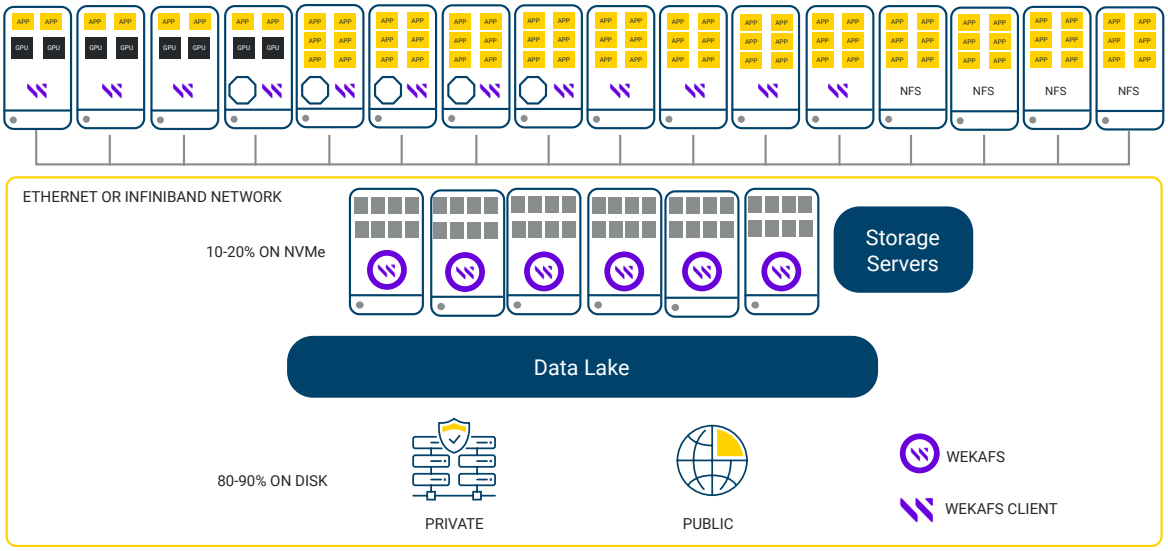
WekaIO WekaFS

Addressing storage IO requirements at each stage in the AI data pipeline, the Penguin Computing ActiveData™ solution with WekaFS delivers massive bandwidth for ingest and training, ultra-low latency for improved inferencing, and storage features to manage data workflows.

WekaFS is the world's fastest shared parallel file system. As the only shared storage solution that provides end-to-end data management for data-intensive AI workloads, WekaFS enables companies to scale performance across the GPU cluster.

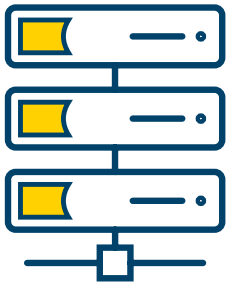
- Delivers unmatched performance at any scale while offering the same enterprise features and benefits of traditional storage
- Prevents GPU starvation, easily meeting the I/O requirements of the most demanding AI and analytics models
- Delivers 10x the performance of legacy network attached storage (NAS) systems and 3x the performance of local server storage





WekaFS Features

- Balanced flash storage and high-performance networking for predictable performance that scales
- Multi-workflow optimized storage with massive bandwidth for ingest and training, ultra-low latency for improved inferencing, and storage features to manage data workflows
- Can be deployed as a dedicated storage server (appliance model) or integrated into the application cluster (converged)
- Supports a hybrid cloud model, allowing enterprises to leverage on-demand public compute resources for cloud-bursting, remote backup, and disaster recovery
- Penguin Computing provides a single point of support for quick problem resolution



Data Center Infrastructure

OriginAI can be built using both a traditional 19” rack platform and a modern 21” OCP (Open Compute Project) platform. Traditional 19” rack infrastructures are supported in almost every data center worldwide and in a variety of dimensions. Modern 21” OCP rack infrastructures require data centers that can support the most demanding physical and power densities. Penguin Computing has partnered with leading data center facility pioneers who can support the demanding characteristics of today’s HPC platforms.

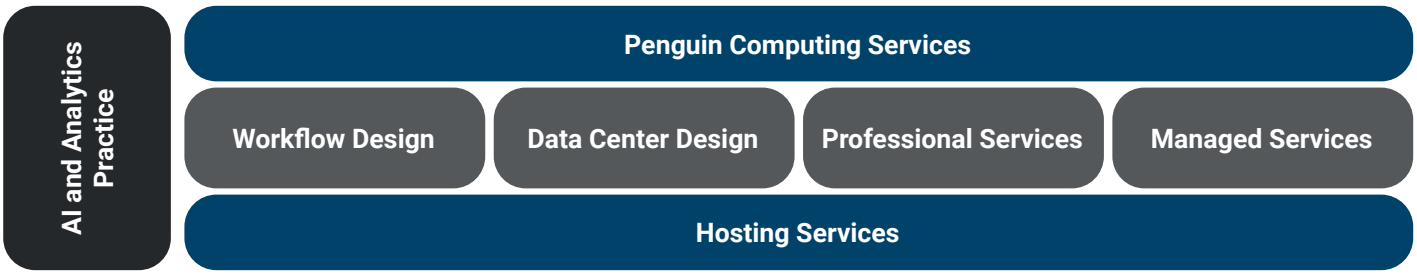
Power

OriginAI supports three-phase 50A or 60A, 208V, 277V, or 480V power options as well as A+B fully redundant power, or N+1 redundant power. 21” OCP also supports 12V or 48V power delivered directly to the servers, which enable much higher power density per rack.

Cooling

OriginAI can be air cooled with traditional HVAC equipment. Penguin Computing recommends using a combination of air cooling and liquid cooling when deploying OriginAI into a data center not designed for high-power equipment. Rear Door Heat Exchangers capture hot air exhaust at the rear of the rack, and can be deployed on most 19” and 21” rack infrastructures. OriginAI is also designed to integrate Direct-To-Chip cooling options that capture heat directly from the CPU block. This cooling solution removes 85% of server heat before it’s transferred into the air, and can be used in select 21” infrastructures.

Penguin Computing Services



Design Services

Workflow Design

- Software Orchestration
- Compute Performance
- Multi-Node Communication
- Data Storage and Data Tiering
- Data Ingest and Egest
- Environment Sizing

Data Center Design

- Rack and Floor Space
- Environment Scalability
- Maximum Power Consumption
- Power Phase Balance
- Efficient Cooling and Heat Removal
- Optimal Networking Topologies

Professional Services

Stand Up and Initialization

- System Burn-In Testing
- Racking and Cabling
- Software Installation & Tuning
- On-Site Deployment and Integration

Hosting Services

Data Center Hosting

- Penguin Data Center
- Customer Data Center
- Power, Space, and Cooling Management
- Monthly or Annual Billing (As-A-Service)

Managed Services

System Administration:

- Complete Hands-Off Experience
- Augment Existing IT Capabilities
- Collaborate with Penguin Support
- Tens to Thousands of Servers
- Terabytes to Exabytes of Data
- Multi Data Center Support

Penguin Computing OriginAI

Each DGX A100 rack within OriginAI:

- Supports up to six DGX A100 servers
- Provides a full non-blocking Mellanox HDR InfiniBand fabric
- Provides a GbE network for administration and IPMI remote systems management
- A fully populated DGX A100 rack weighs 1,265 lbs and requires 42 kW of power and cooling.

Penguin Computing OriginAI Rack Layout

Conclusion

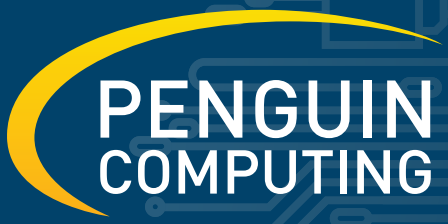
Penguin Computing OriginAI provides a single, secure, end-to-end solution that includes architectural design, hardware and software services, hosting, and deployment. Penguin Computing OriginAI frees private and public sector organizations from having to focus valuable time and human resources on creating an architecture from scratch, delivering cost savings, decreasing risk, and accelerating time to insight.

Penguin Computing can apply our decades of experience to create quality, integrated solutions for our clients. We offer a wide range of professional and managed services that can quickly bring your artificial intelligence, machine learning, and deep learning initiatives products to production.

Contact Us

Use this **form** or call Penguin Computing today at 1-888-736-4846 to find out how you can jump start your artificial intelligence, machine learning, and deep learning initiatives with a reference architecture that addresses software orchestration, compute infrastructure, Data, and infrastructure design to:

- Accelerate time to insight from years to months
- Maximize the performance and utility of high-value AI systems
- Ensure increased productivity and highest ROI



Expanding the world's vision of what is possible