INSIGHT AND ACTION

Getting Started on Your Next AI and HPC Project

PENGUIN™
SOLUTIONS

The introduction of ChatGPT in late 2022 and its explosive adoption throughout 2023 lit a fire in the marketplace. Artificial intelligence (AI) became mainstream and a necessity. That, in turn, drove the need for compute capacity to support these new HPC and AI initiatives in organizations of all sizes across all industries.

The resulting challenge in meeting this high demand is gaining access to clusters that can deliver on the compute requirements for these apps, which are complex and require a mix of processor types, workload accelerators, and high-performance storage and interconnect technology not commonly found in enterprise IT. In many cases, on-premises solutions must also be tightly integrated with cloud services.   Organizations must also consider if their IT departments are trained and prepared to deploy and manage such a new, complex environment.

Overprovisioning clusters is easy but comes at great expense. Installing systems that meet this demand are hard to optimize, scale, and manage on a day-to-day basis. As such, those undertaking such efforts must carefully plan for and address every step of the process – designing, building, deploying, and managing high-performance clusters for AI.

Issues to consider with scalable clusters for HPC and AI

One important point to note is that ideal clusters for new AI and HPC workloads are the first to combine GPU-based compute, InfiniBand networking, and high-speed storage. In the past, each of these elements was used at scale individually, but they were never brought together in large clusters. And that's not easy.

The way to do so successfully is to look at four key aspects of any new AI or HPC at-scale project – specifically, the design, build, deployment, and management phases. Let's look at the nuances of each phase:

Design: Businesses need specialized skills to design clusters that deliver the expected performance, security, stability, and scalability. Most businesses find they do not have the expertise or internal skills to do so, given that such clusters must incorporate elements (e.g., GPUs, accelerators, etc.) that in the past have not normally been part of enterprise IT systems. Another element to consider is the power and cooling requirements of HPC and AI clusters.

Build: Once the design is finalized, assembling a cluster requires additional unique skills. Businesses need expertise in cluster hardware integration. There is also a need for expertise in working with the software stack. The stack must be validated and optimized to reduce any compatibility issues.

Deploy: Clusters built for AI and HPC often have demanding and complex power and cooling requirements compared to traditional IT systems, and these must be properly integrated and optimized when deploying new clusters. Additionally, networking aspects must match the need to move data between storage and compute elements, as well as between GPUs. And with many businesses using their own data to train AI models, security and data privacy/protection must be addressed.

Manage: As with any cluster, those used for AI and HPC must be continuously managed so that they are highly available. Otherwise, critical workloads fail, and results are delayed. However, there are key differences with these clusters that compound management issues. First, AI and HPC workloads vary greatly from job to job and project to project. As a result, workload and performance optimization require persistent monitoring and adjustments. Second, AI and HPC clusters use special components that have unique failure signatures. Traditional tools might need to be modified to monitor and manage these elements properly.

Teaming with a technology partner

Businesses and organizations are in a perfect storm when it comes to meeting AI and HPC compute demands. They need systems that tightly integrate technologies (e.g., GPUs, high-speed interconnect, and high-speed storage) with each alone being unique and not widely used in enterprise IT. These systems also must be highly optimized to deliver the required performance in an economical manner. Such an undertaking requires special skills and expertise.

One way to approach this is to hire people with the needed industry knowledge and train existing staff in these areas. Like many other IT endeavors, an alternative is to work with a partner who brings real-world experience in working with the new technologies to deliver a suitable system.

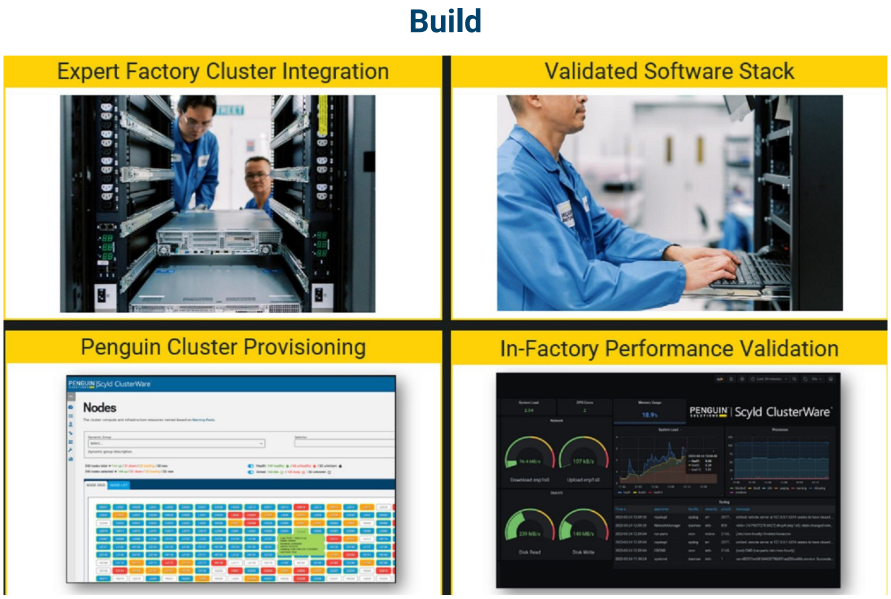
These are all areas where Penguin Solutions can help. We have long been known for our efficient HPC systems and proven record in designing and deploying cost-efficient HPC systems for extreme workloads. We now apply the same strategies to AI.

To that end, Penguin Solutions is applying its 25 years of HPC experience to designing, building, deploying, and managing AI factories, the supercomputing clusters that run sophisticated AI workloads. As the factory name implies, Penguin operationalizes the use of AI. We apply best practices and leverage our strong and long-term relationship with GPU, networking, and storage partners to build highly efficient and scalable AI systems for companies like Meta, which are leading the use of AI for business.

Going a layer deeper, Penguin Solutions can help in each of the four major stages discussed above, including:

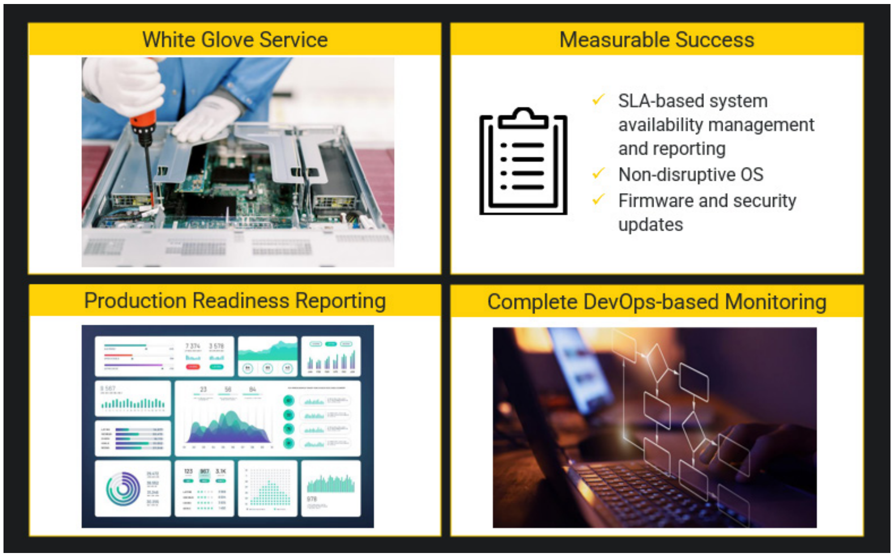
Design: An engagement with Penguin starts by reviewing a project's vision, evaluating where data will come from and how much data will be used, understanding the compute and storage requirements, and more. Penguin Solutions then uses proprietary software to design a system with the required performance, security, and scalability.

Build: Penguin approves, pre-configures and assembles a cluster before shipping and delivery. As part of that process, Penguin experts integrate all the components of the system. The in-factory work includes rack, cable, and burn-in testing. Additionally, we validate the software to avoid problems once the system is deployed.


Build
Expert Factory Cluster Integration
Validated Software Stack
Penguin Cluster Provisioning
In-Factory Performance Validation
PENGUIN | Scyld ClusterWare
Nodes
PENGUIN | Scyld ClusterWare

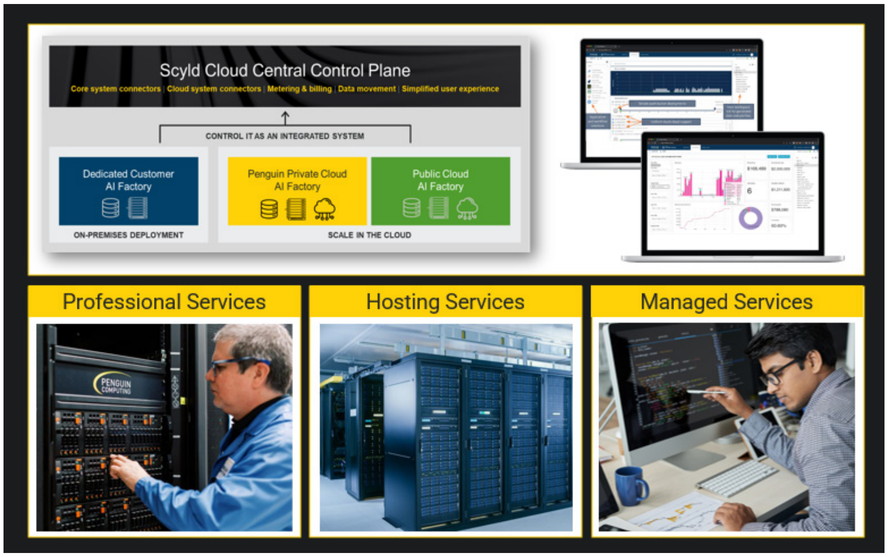
Deploy: Penguin Solutions delivers pre-built clusters, and then provides on-site help to ensure the clusters are working properly. We work with cooling and storage partners to verify operational performance. Once the system is physically in place, Penguin Solutions uses our own software, including Scyld ClusterWare and Scyld Cloud Workstation, to provision the solution.

Deploy


White Glove Service
Measurable Success
✓ SLA-based system availability management and reporting
✓ Non-disruptive OS
✓ Firmware and security updates
Production Readiness Reporting
Complete DevOps-based Monitoring

Manage: Penguin Solutions, a certified Nvidia DGX-Ready Managed Services provider, is one of the leading providers of AI factories with over 50,000 GPUs under management. But for most organizations, the complexity of supercomputers and cloud computing present serious budgetary and management challenges. That's why Penguin Solutions has developed our own software for cluster management and for supporting hybrid infrastructure, combining on-premises, private cloud, and public cloud environments.

Manage


Scyld Cloud Central Control Plane
Core system connectors | Cloud system connectors | Metering & billing | Data movement | Simplified user experience
CONTROL IT AS AN INTEGRATED SYSTEM
Dedicated Customer AI Factory
Penguin Private Cloud AI Factory
Public Cloud AI Factory
ON-PREMISES DEPLOYMENT
SCALE IN THE CLOUD
Professional Services
Hosting Services
Managed Services

A final word

Working with Penguin Solutions when starting or expanding AI and HPC efforts speeds up the time from concept to a working solution. And it frees existing staff to do other things, reducing the need to hire new staff with unique skills.

Most importantly, our proven AI factory solutions ensure optimized use of expensive compute resources, cost savings, and lower TCO.

Learn More

If interested in learning, please visit our website or contact a Penguin Solutions HPC AI expert.

PENGUIN™
SOLUTIONS