INSIGHT AND ACTION

AI Factories and Creative GPU Utilization for AI

PENGUIN™ SOLUTIONS

Businesses have embraced the use of artificial intelligence (AI), and most plan to rapidly expand its use going forward. However, one major inhibitor and source of problems is that many AI efforts require compute infrastructures that make use of enormous numbers of GPUs. The problem is that such infrastructures are difficult to design and build. And most run inefficiently, with low GPU availability and utilization.

Numerous industry studies put these trends and the scope of the issue into perspective. A 2023 Forbes Advisor survey of 600 business owners using or planning to incorporate AI in business found that more than half of all respondents are using AI for customer service, cybersecurity, and fraud management. An earlier Accenture report on AI found that 84% of C-suite executives think leveraging AI will help them achieve their growth objectives. And some market research studies predict annual AI market growth rates to be in the 40% range for the next five years.

These studies and more may greatly under-represent the market as they were done before ChatGPT exploded onto the scene. ChatGPT introduced generative AI to the world, and again, businesses have been eager to use it. That same Forbes survey found that almost all (97%) business owners believe ChatGPT will help their business. To that point, generative AI is finding use by everyone, from marketers helping them write content to developers using it to write code.

The bottom line is that AI workloads are becoming the norm in businesses today.

Enter the AI Factory

Just as HPC went from the domain of academic computer centers and government labs and changed enterprise compute systems, AI's adoption needs new infrastructure.

The problem is that businesses are spending hundreds of millions of dollars on GPU-based systems, and they're only getting low (25% in some cases) performance and availability out of them.

The issue: The systems for AI are different from what's typically been used for HPC. Many businesses do not have the expertise, best practices, and more needed to design and deploy systems that efficiently deliver the needed compute power.

AI Factory Challenges

1 DESIGN COMPLEXITY

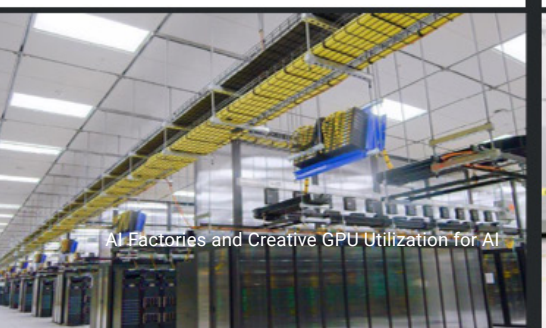
- New workload & architecture
- AI clusters are highly sensitive at scale
- Requirement to blend multiple networks & processor types

2 LENGTHY DEPLOYMENT

- Shortage of components
- Complex build & stabilization process to achieve throughput
- High prices

3 DELICATE OPERATIONS

- Unpredictable stability at scale
- Performance issues can drive significant financial impact
- Workloads interrupted and training time lost



AI Factory Challenges

That is an area where Penguin Solutions™ can help. We are long-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.

Specifically, Penguin Solutions applies more than 25 years of HPC experience to designing, building, deploying, and managing AI factories to operationalize the use of AI. We have applied best practices and leveraged our strong and long-term relationship with our technology partners to build highly efficient and massive AI systems for companies like Meta, which are leading the use of AI for business.

Additionally, in assembling these AI factories, we worked with the leading storage and networking partners to maximize the efficiency of each system's massive computing capacities. To that end, Penguin Solutions tested and helped select the optimal storage and most powerful networking solutions to meet the needs of each specific customer and their AI workloads.

Benefits of the AI factory approach

Working with Penguin, those deploying large systems for AI get much higher GPU utilization, thus maximizing their investment. Using our expertise in large-scale HPC systems, the AI systems can be deployed significantly faster than others in the industry.

Another benefit of the scalable infrastructure is less downtime and, thus, more availability. That is made possible thanks to the Penguin's best practices and experience deploying large HPC systems. And higher GPU utilization and availability lead to lower power utilization.

An example of the power of the Penguin Solutions AI factory approach is the work we have done with Meta. We provide AI-optimized architecture and managed services for the AI Research SuperCluster (RSC), Meta's cutting-edge AI supercomputer for AI research.

The initial system configuration called for more than 6,000 GPUs and 46 petabytes of cache storage. The final build makes use of more than 16,000 GPUs and 1 exabyte of storage. Penguin's hardware and software expertise brought together contributions from Penguin itself, a top GPU vendor, and Pure Storage. Together, the three partners supplied Meta with an optimized solution. For details on approach that was used for system deployment, visit this page on our website.

Working with Penguin Solutions, "we improved our overall cluster management. By the time we completed the second phase of building RSC, availability stayed above 95% on a consistent basis," according to a statement from Meta. "This was no small feat given that we added a 10K GPU cluster while concurrently running multiple research projects."

Perhaps the most telling statement about the approach, given the need to scale in the future, is "We now have a template for building large GPU clusters that is repeatable and reliable."

A final word on meeting the demands of AI workloads

The industry is at the very early stage of AI adoption. Already, businesses find they need special GPU-based infrastructure to run their workloads. Meeting the compute demands with such infrastructure has proven challenging. Most systems deliver very inefficient use and availability of those very expensive GPU resources.

Penguin Solutions has taken our expertise in designing and deploying large-scale HPC systems and applied that to AI. Our AI factory approach leverages its long-term relationship with CPU, GPU, networking, and storage partners to operationalize its AI systems. Those using them find they deliver exceptional GPU availability and high utilization rates.

Learn More

Learn more about Penguin Solution AI factories here.

https://www.penguinsolutions.com/computing/solutions/ai/

PENGUIN™ SOLUTIONS