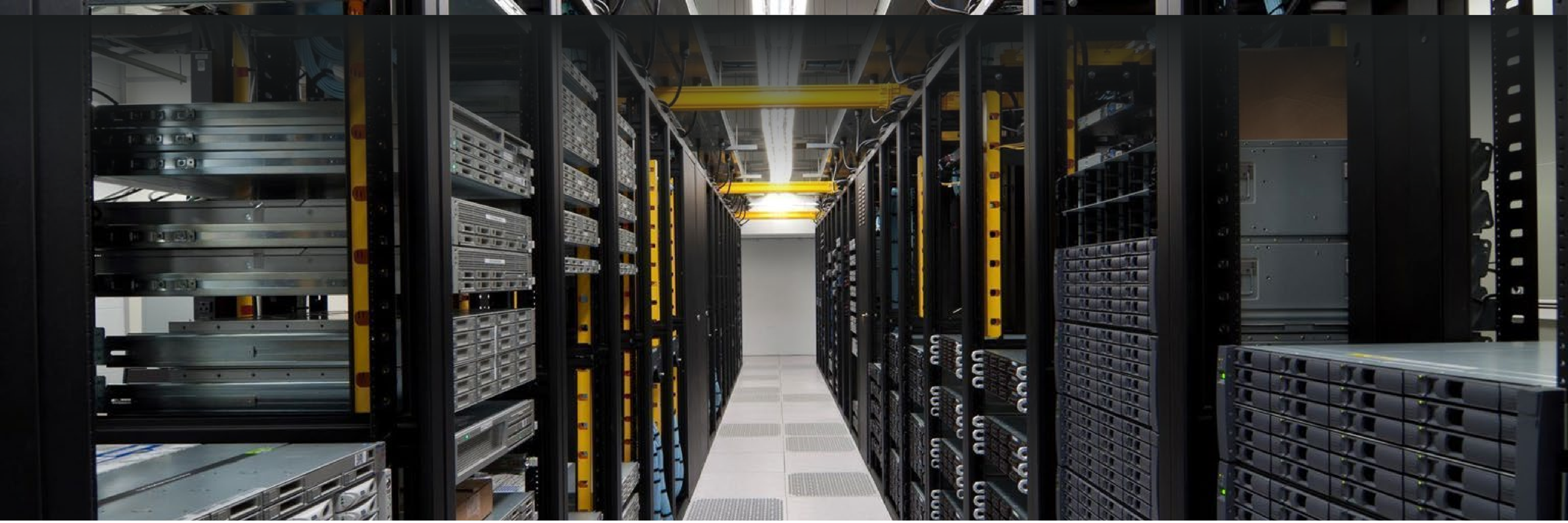


# PENGUIN SOLUTIONS

## AI and Accelerated Computing Infrastructures at Scale

PENGUIN™ SOLUTIONS

# AGENDA

1. About Penguin
   YOUR TRUSTED ADVISOR FOR AI FACTORIES
2. Delivering AI Factory at Scale
   CHALLENGES AND SOLUTIONS
3. Penguin’s Scyld Suite
   YOUR AI FACTORY UNIFIED CONTROL PLANE
4. Your End-to-End AI Factory Journey
   GUIDANCE AND FLEXIBILITY

PENGUIN™ SOLUTIONS

# About Penguin

Penguin Solutions, an SGH company, designs, builds, deploys, and manages AI and accelerated computing infrastructures at scale.

**Tailored:** We deliver highly tuned solutions that are designed to significantly improve results.

**Proven:** We have over 24 years of HPC experience and have deployed over 50,000 GPUs in partnership with leaders in AI.

**Innovative:** We continually incorporate the best elements and technologies.

**24 YEARS OF EXPERIENCE**
AI/HPC

**BROAD MARKET EXPERTISE**
Commercial, Fed, DoD

**AI CLUSTER MANAGEMENT**
Scyld Suite

**NVIDIA CERTIFIED**
DGX Managed Service Provider

**SGH**
NASDAQ

**FLEXIBILITY**
On Prem, Cloud, and Hybrid

PENGUIN SOLUTIONS

# Experts at AI Factory Infrastructure

Enterprise infrastructure designed to support advanced workloads with optimal performance and stability at scale

AI FACTORY

User Workloads

AI & HPC Applications

RStudio · ANSYS · Parabricks · jupyter · SIEMENS · Keras · GROMACS · WRF · Kubeflow · TensorFlow

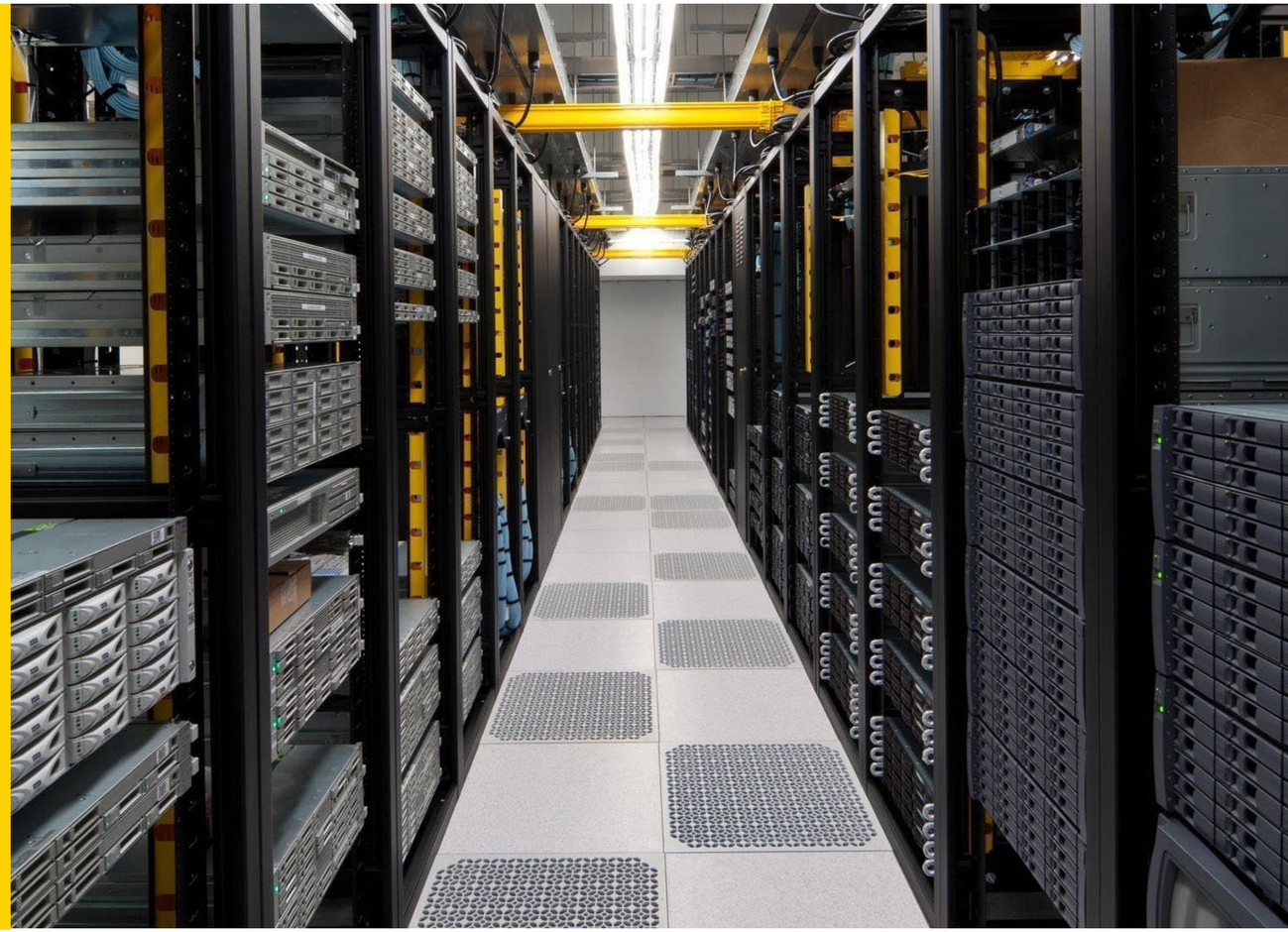
HPC Workloads | Analytics/ML Workloads | AI/DL Workloads

Platform Foundation

PENGUIN SOLUTIONS

Design | Build | Deploy | Manage

| Data Center | Cloud |
|---|---|
| Software Technologies | Software Technologies |
| Compute & GPU Technologies | Compute & GPU Technologies |
| Data Technologies | Data Technologies |
| Data Center Infrastructure | Secure Cloud Infrastructure |



# Penguin: Proven AI Factory Delivery at Scale

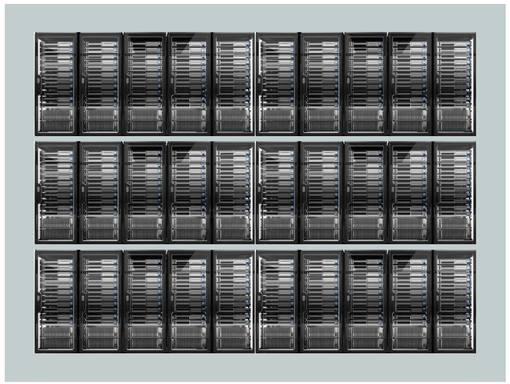
ChatGPT trained on

**10,000 GPUs**

OpenAI | Microsoft

Google A3 cluster

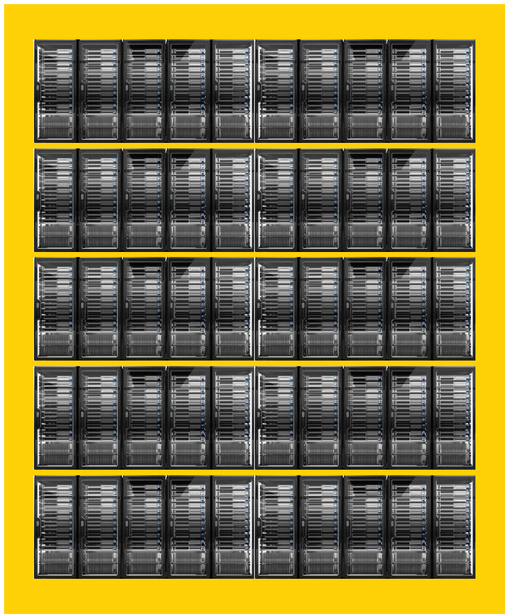
**26,000 GPUs**



Google

Penguin Manages

**50,000+ GPUs**



PENGUIN™ SOLUTIONS

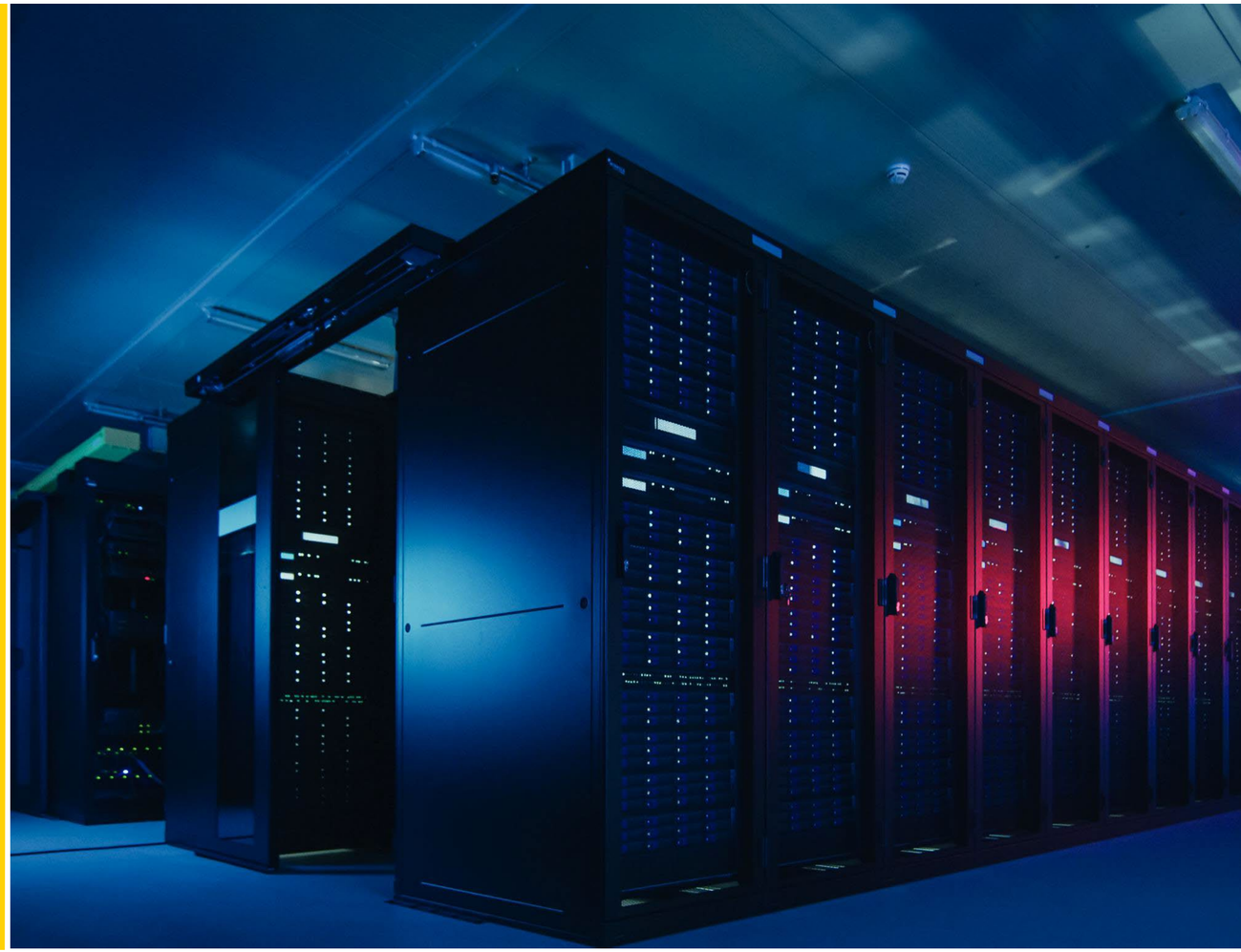
# Customer Profile — Ultrascale Company

## Business Challenge

- Urgently needed large-scale AI platform capable to support key product and innovation initiatives
- Co-development of a leading-edge technology solution
- DevOps needs for HPC/AI at scale

## Super Scale AI Infrastructure Platform

- Over 10,000 Nvidia A100 GPUs
- Hundreds of Petabytes of data storage capacity
- 200 Gb/s HDR InfiniBand per GPU
- exaFLOPS of mixed precision compute
- Delivered in 2022 with continued Penguin-provided managed services



PENGUIN SOLUTIONS

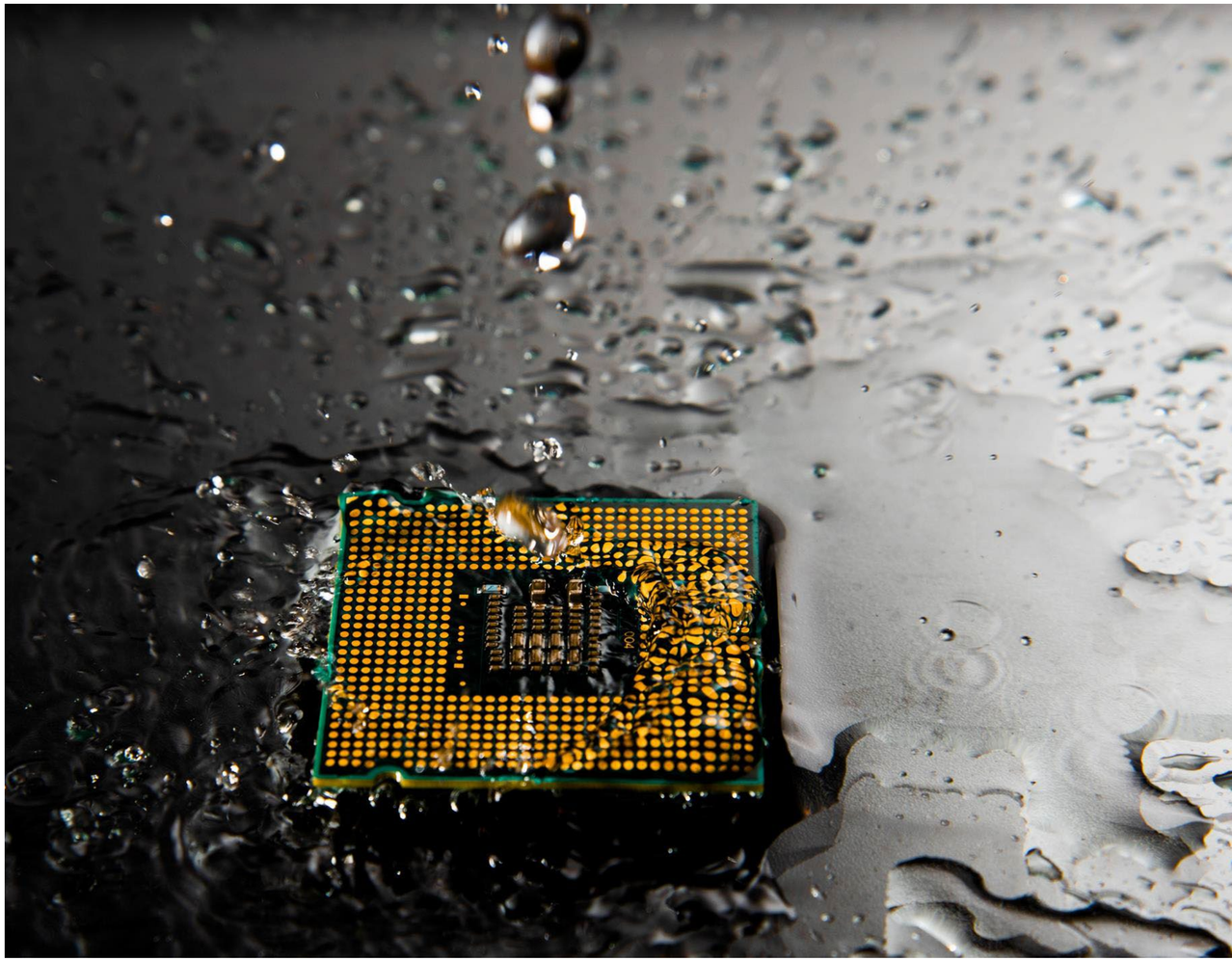
# Customer Profile — Global Energy Provider

## Business Challenge

- Improve data center sustainability without compromising performance
- Optimize compute density, power utilization, and cooling
- Obtain end-to-end system build, deployment, and support

## Immersion Cooled AI Infrastructure

- Single-phase immersion cooling
- Dense NVIDIA A100 GPUs in Penguin AI servers
- Designed, delivered and supported by Penguin
- Reduced environmental impact from data center cooling
- Ongoing development and deployment of systems since 2021



PENGUIN™ SOLUTIONS

# Customer Profile — Federal System Integrator

## Business Challenge

- Hybrid Cloud solution needs for AI and HPC workloads
- Requirements from 5 separate federal agencies
- Desire for full DevOps-based managed services

## Designed AI & HPC Solution and Deployed in Colo

- Thousands of nodes of compute capability
- CPU, GPU and InfiniBand network infrastructure
- Multiple petabytes of data storage
- Penguin services provides all HW, SW & DevOps management
- Hundreds of users from diverse teams
- 25x compute capacity growth over 2.5 years

PENGUIN SOLUTIONS

Working in partnership **with our implementation partner, Penguin Computing, we improved our overall cluster management.** By the time we completed the second phase of building RSC, availability stayed above 95 percent on a consistent basis. This was no small feat given that we added a 10K GPU cluster while concurrently running multiple research projects.

We now have a template for building large GPU clusters that is repeatable and reliable.

Meta

18 May 2023

PENGUIN SOLUTIONS

# Challenges - AI Factories and Accelerated Computing at Scale

## COMPLEX DESIGN

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

## INTRICATE BUILD

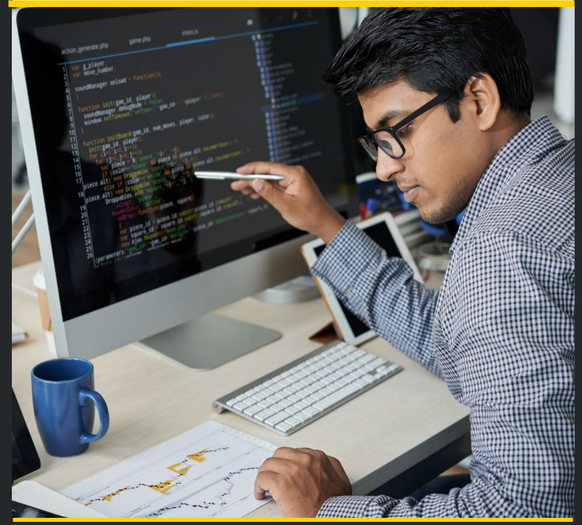
- Supply chain delays with specialty critical components
- Complex build process to achieve throughput
- Lack of software suite for deployment at scale


## LENGTHY DEPLOYMENT

- Complex on-site power, cooling, network, and security integration
- Need for pre-production performance and throughput validation
- Production cluster monitoring

## PRECISION MANAGEMENT

- Specialty components with unique failure signatures
- Performance issues drive significant financial impact
- User workloads interrupted and training time lost



PENGUIN SOLUTIONS

# Our Value - AI Factories and Accelerated Computing at Scale

## DESIGN

- **De-risking** investment through proven architectures
- **Experience** at scale
- Planning for **performance, security and scalability**

## BUILD

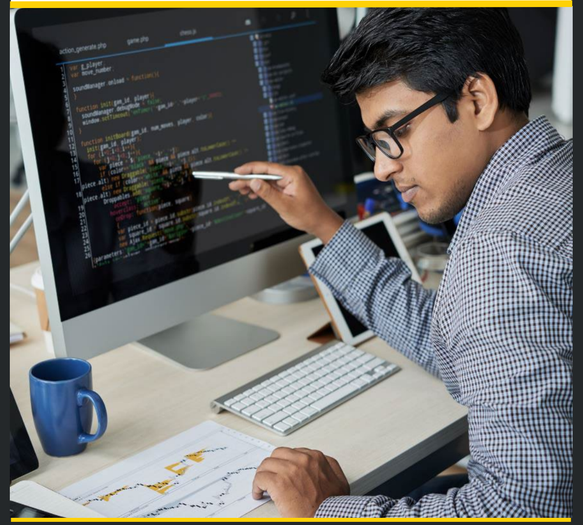
- **In-Factory rack, cable, and burn-in testing** enables smooth on-site deployments
- **Expert** cluster integration
- Validated **software stack** reduces compatibility issues


## DEPLOY

- Penguin software for **deep cluster health monitoring**
- End-to-End **Project Management** accelerates availability
- Planning for **performance, security and scalability**

## MANAGE

- Driving **uptime and throughput** for large scale environments
- **NVIDIA-certified** Managed Services Engineers
- **SLA-based** system management and reporting



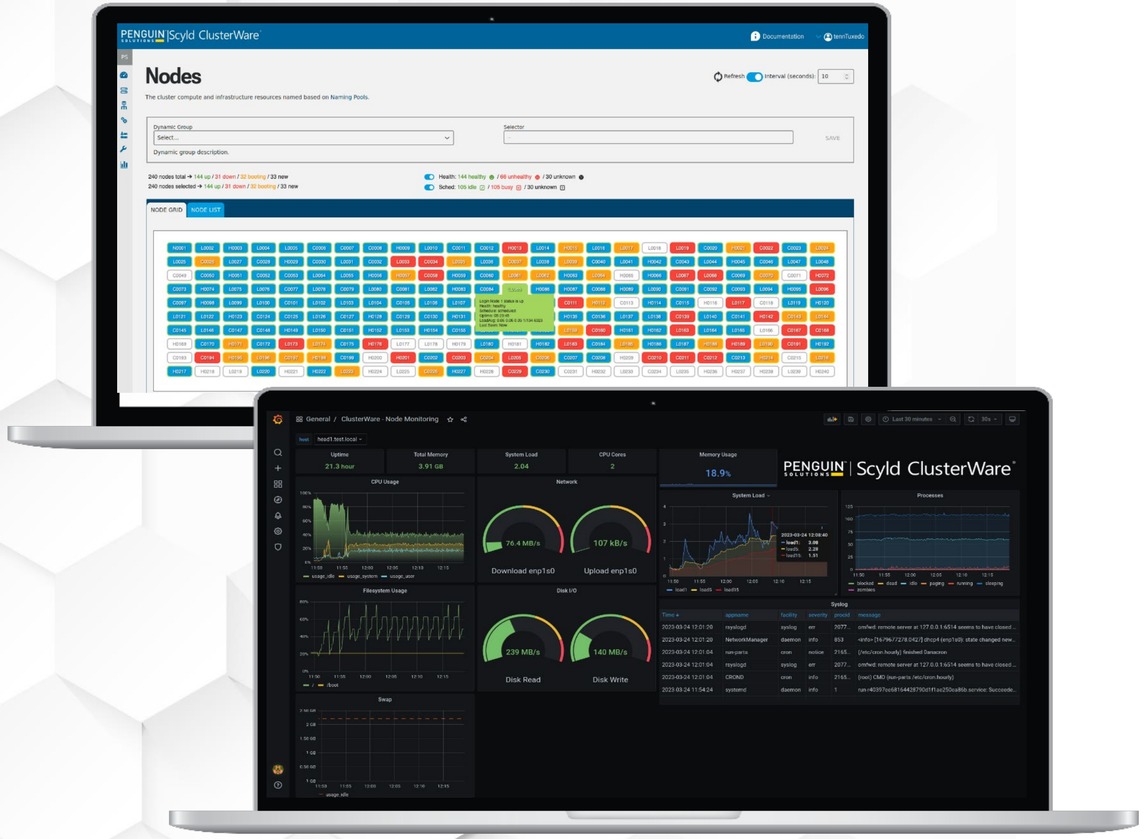
PENGUIN SOLUTIONS

# Scyld Clusterware

## AI Factory Enablement at Scale

Penguin-developed cluster management software that transforms raw hardware, network, and software resources into high-performance AI Factory platforms.

- Streamline complexity
- Rapid provisioning & extensibility
- Advanced workload scheduler support
- Cluster health monitoring & alerting
- Full DevOps integration
- Highly secure



# Scyld Cloud Central

## One-Stop Shop for Hybrid Deployment

Penguin software platform for hybrid cloud management that enables rapid system deployment in the cloud and easy integration of on-premises AI and HPC cluster capacity as solutions come online.

- Hybrid Capability: Cloud and On-Premises
- Tailor cloud clusters to user needs
- Library of pre-built application workflows
- Integrated cost management and budget guardrails
- Elastic scaling for both data and compute

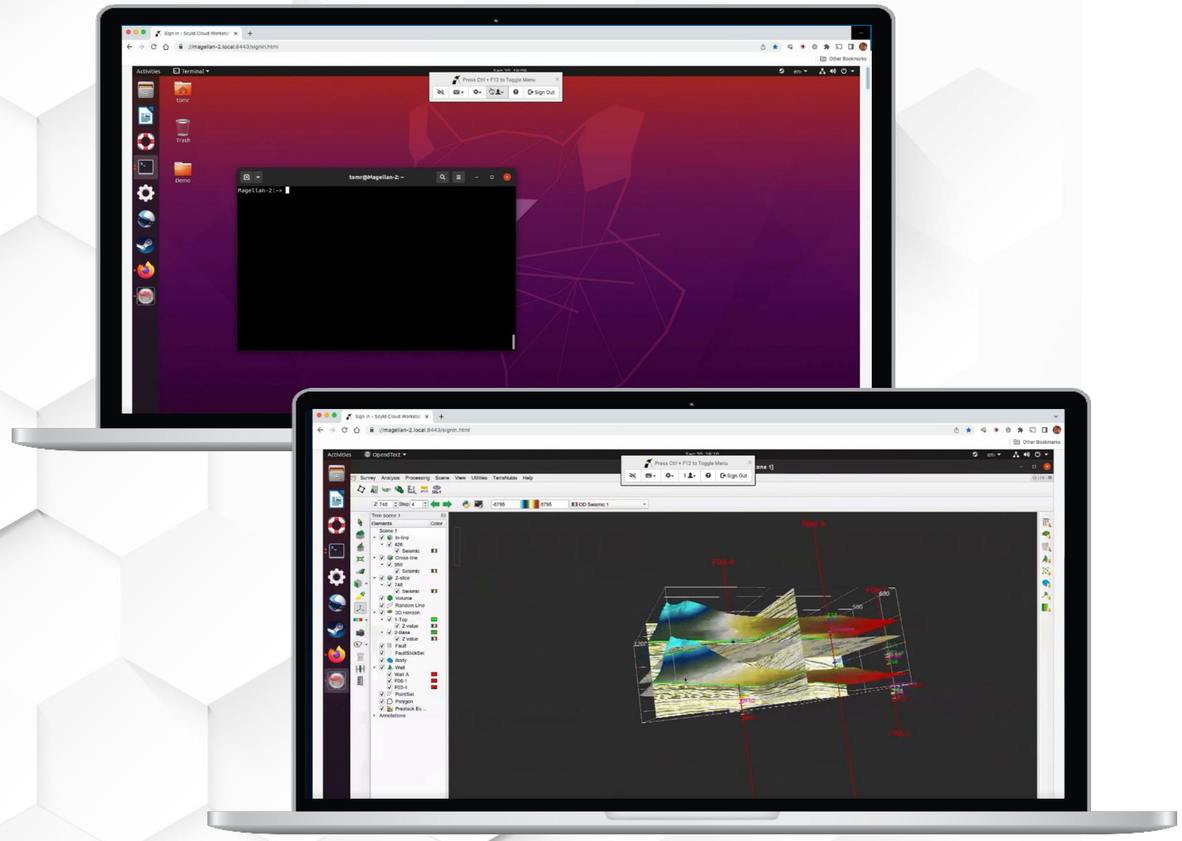

PENGUIN SOLUTIONS

# Scyld Cloud Workstation

## Secure Remote Application Access

Penguin software platform for secure virtual desktop infrastructure (VDI). High-fidelity access by users and data scientists to remote applications running in the cloud or in remote data centers.

- Remote user collaboration on shared desktops
- No client installation required
- Applications can access and manage massive data sets without moving data
- 60 frames per second video refresh
- Optimized for low-bandwidth connections



PENGUIN SOLUTIONS

# Complete Portfolio

## Design, Deployment, and Management Services


Certified NVIDIA DGX-Ready Managed Services Partner

## Design Services

**Workflow Design**

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

**Cloud Solution Design**

- Instance Type Selection
- CPU / GPU / Memory Options
- Network Bandwidth Requirements
- Software Compatibility
- Data Ingest and Egress
- On-Demand and Spot Instances

## Deployment Services

**Stand Up and Initialization**

- User, Group, Project Configuration
- Cluster Configuration
- Sample Deployment and Testing
- Job Performance Characterization

## Hosting Services

**Co-Lo Hosting for Private Clouds**

- Penguin Data Center
- Customer Data Center
- Power, Space, and Cooling Management
- As-a-Service Billing

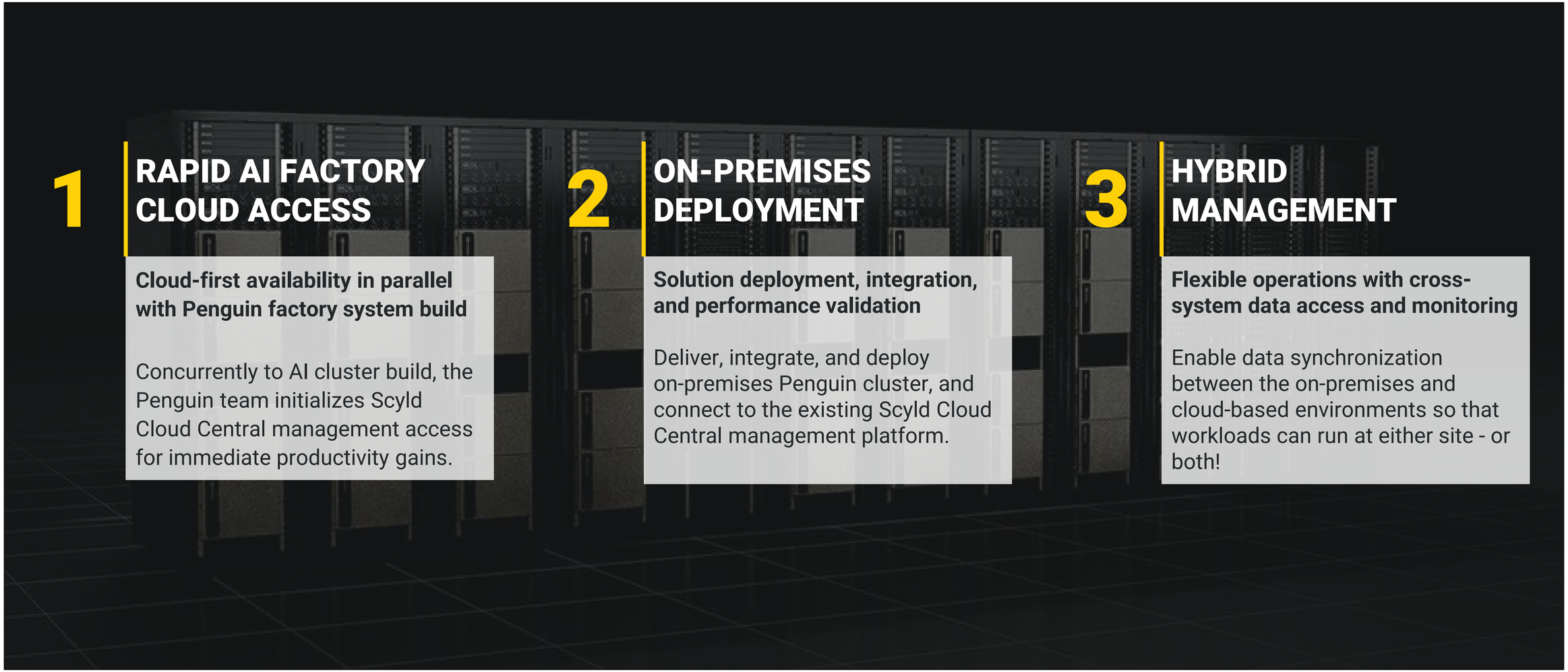
## Managed Services

**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 Cloud & Data Center Support

PENGUIN SOLUTIONS

# Guidance and Flexibility: End-to-End AI Factory Journey



## 1 RAPID AI FACTORY CLOUD ACCESS

**Cloud-first availability in parallel with Penguin factory system build**

Concurrently to AI cluster build, the Penguin team initializes Scyld Cloud Central management access for immediate productivity gains.

## 2 ON-PREMISES DEPLOYMENT

**Solution deployment, integration, and performance validation**

Deliver, integrate, and deploy on-premises Penguin cluster, and connect to the existing Scyld Cloud Central management platform.

## 3 HYBRID MANAGEMENT

**Flexible operations with cross-system data access and monitoring**

Enable data synchronization between the on-premises and cloud-based environments so that workloads can run at either site - or both!

PENGUIN SOLUTIONS

// WE CAN HELP YOU

# BE FIRST TO THE FUTURE

1999 // // 2023

PENGUIN SOLUTIONS