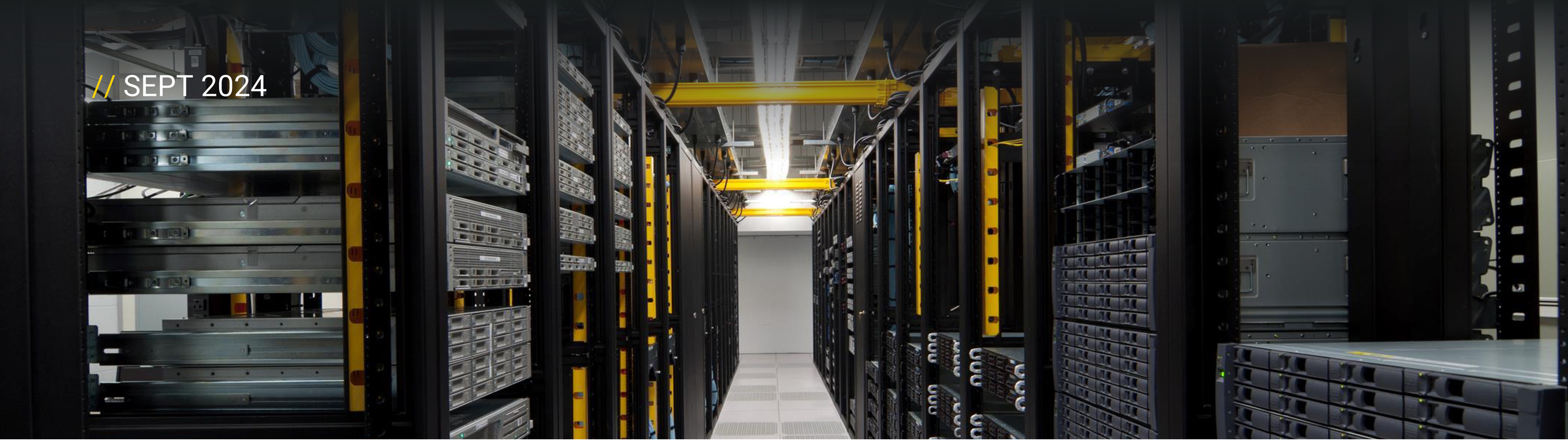


// SEPT 2024

PENGUIN SOLUTIONS

The Easy Button For Training Compute

PENGUIN™ SOLUTIONS

Tesla without Penguin AI Factory Partnership

> “Training compute should soon not be so much of a limiting factor.
>
> Very difficult bringing the 10k H100 cluster online btw. Similar experience to bringing our now 16k A100 cluster online.
>
> Uptime & performance are low at first, then improve with lots of work by Tesla & Nvidia.”

Elon Musk @elonmusk
30 August 2023

Actual Failure Rates Published by Meta

- Meta publishes report on 54-day Llama 3 Model on 16,384 Nvidia H100 80GB GPUs on July 23, 2024
- 220 unexpected interruptions recorded were caused by issues with GPUs or their onboard HBM3 memory
- Only two CPU failures were recorded during the 54-day training period
- This averages to one GPU failure every 3 hours
- Using the same failure rate on a 128 GPU cluster equates to one GPU failure every 16 days
- The SXM board is the FRU for the AI servers with 8 GPUs per board
- Mathematically in 256 days from turning the 128 GPU AI cluster on you could have ZERO operational GPUs
- There were actually 419 unexpected interruptions, we generously only used 220 for these calculations

Meta report – The Llama 3 Herd of Models

“

The complexity and potential failure scenarios of 16K GPU training surpass those of much larger CPU clusters that we have operated.

Moreover, the synchronous nature of training makes it less fault-tolerant – **a single GPU failure may require a restart of the entire job**.

”

https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

Meta
23 July 2024

Meta with Penguin AI Factory Partnership

> 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

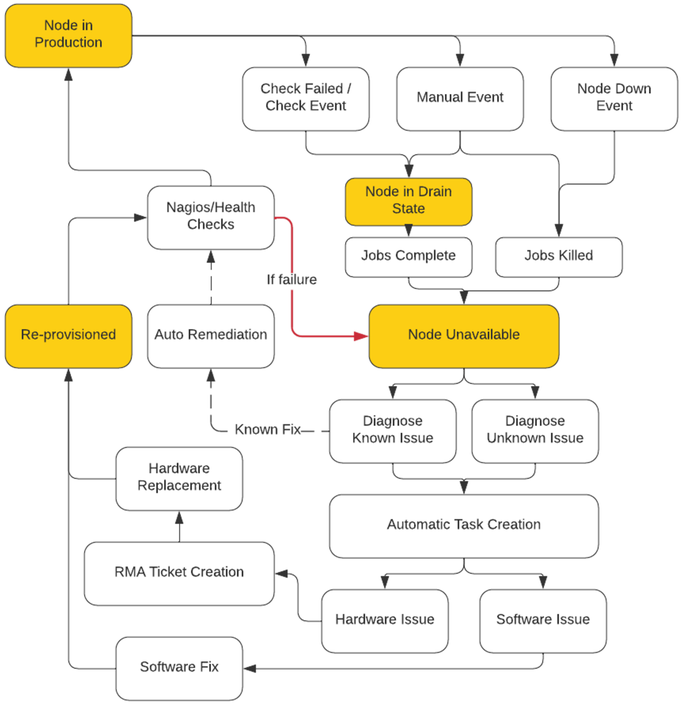
18 May 2023

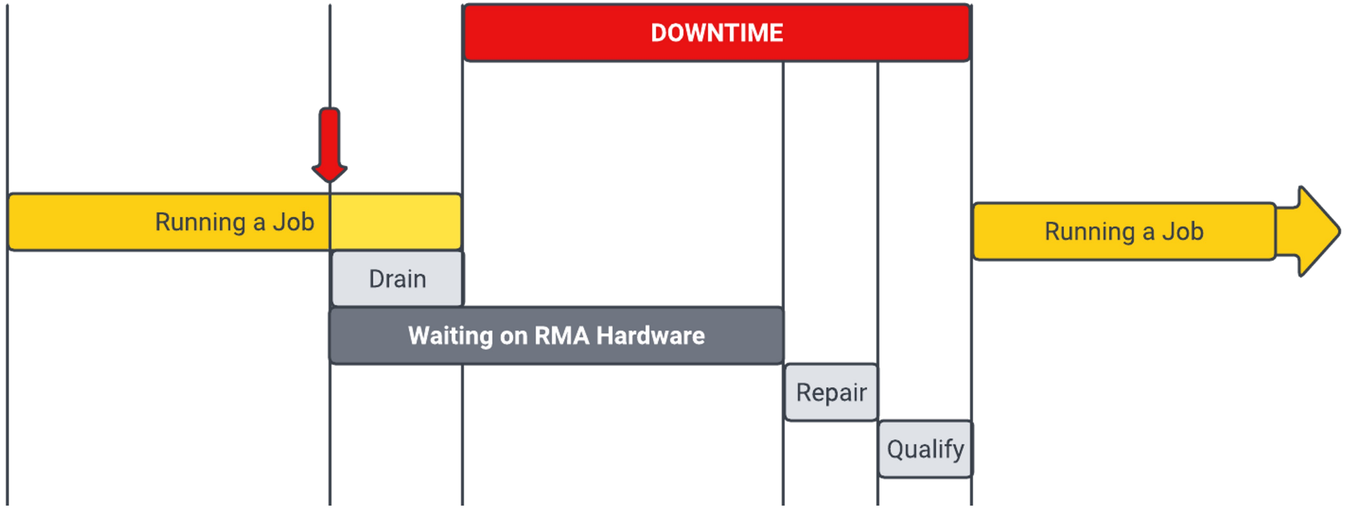
How the Easy Button was built

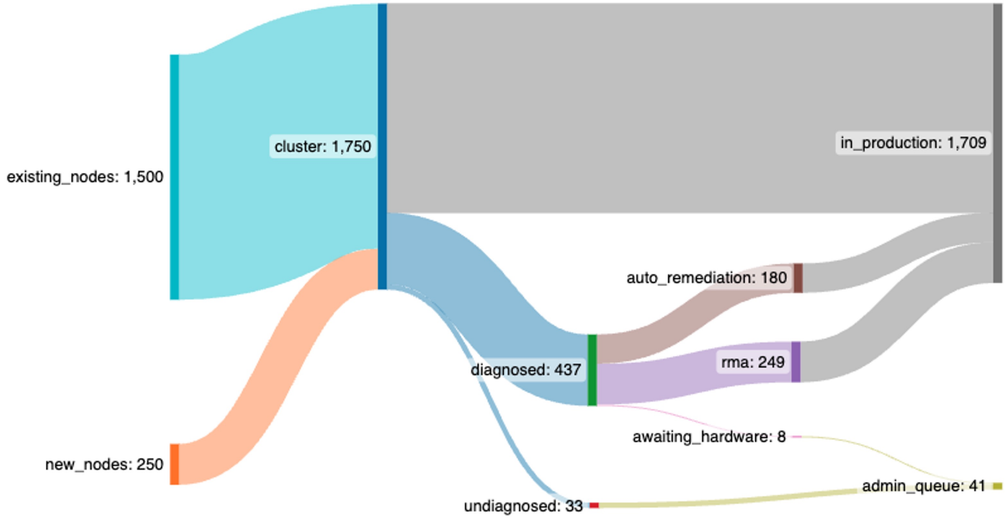
- 7.5 years ago Penguin wins a RFP for a company by the name of Facebook
- After winning the contract, each subsequent year requests were made for additional functionality
- Phase 1 request was for improved telemetry
- Phase 2 request was for RCA using the additional data collected
- Phase 3 request was for Human out of the Loop remediation
- Phase 4 request for further process improvements to increase availability
- Phase 5 request was for managing multiple GPU-clusters at different physical locations using the same control plane
- This accelerated timeline was possible because of the scale we had access to during this process

Automating GPU Infrastructure Management

- Human Out-of-the-Loop: Automation wherever feasible
- Automated triage and remediation of known node failures
  - Streamline repair and re-qualification of hardware prior to returning to production.
  - Reduce downtime with early detection and RMA of failed hardware
- Rapid deployment of updates and configuration changes
- Automated detection of wiring and InfiniBand issues
- GPU and network performance monitoring
- Flexible templates to meet customer needs





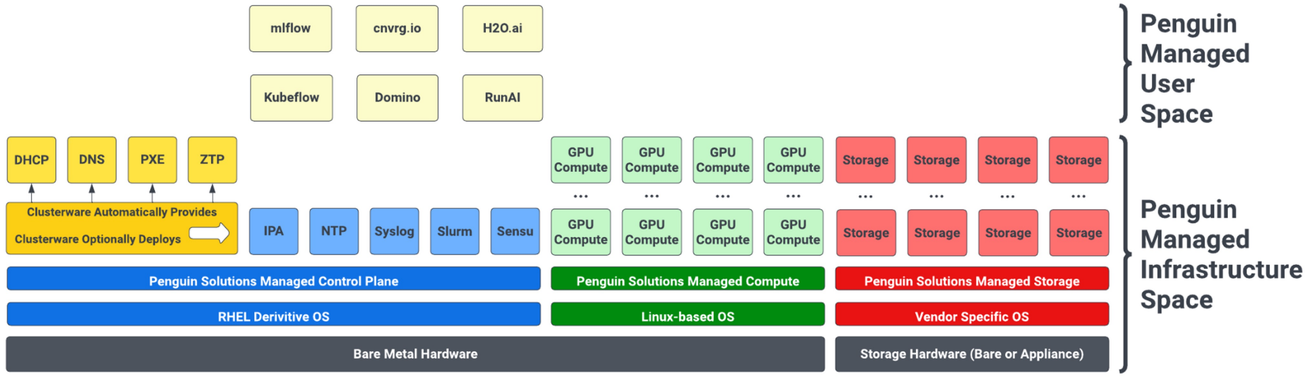


PENGUIN SOLUTIONS

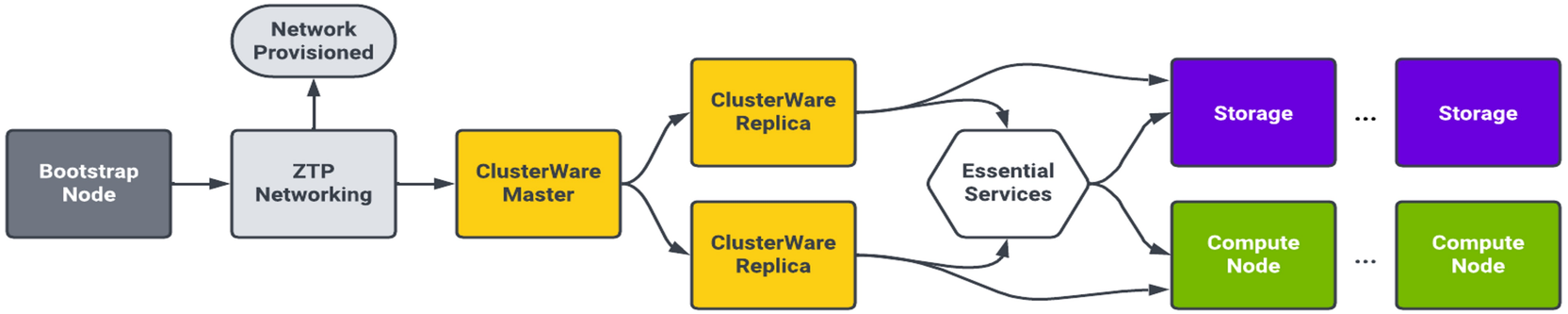
Automated GPU Infrastructure Deployments

"The state of *everything* should be defined in code."

Automating and managing infrastructure through code *instead of manual processes*.



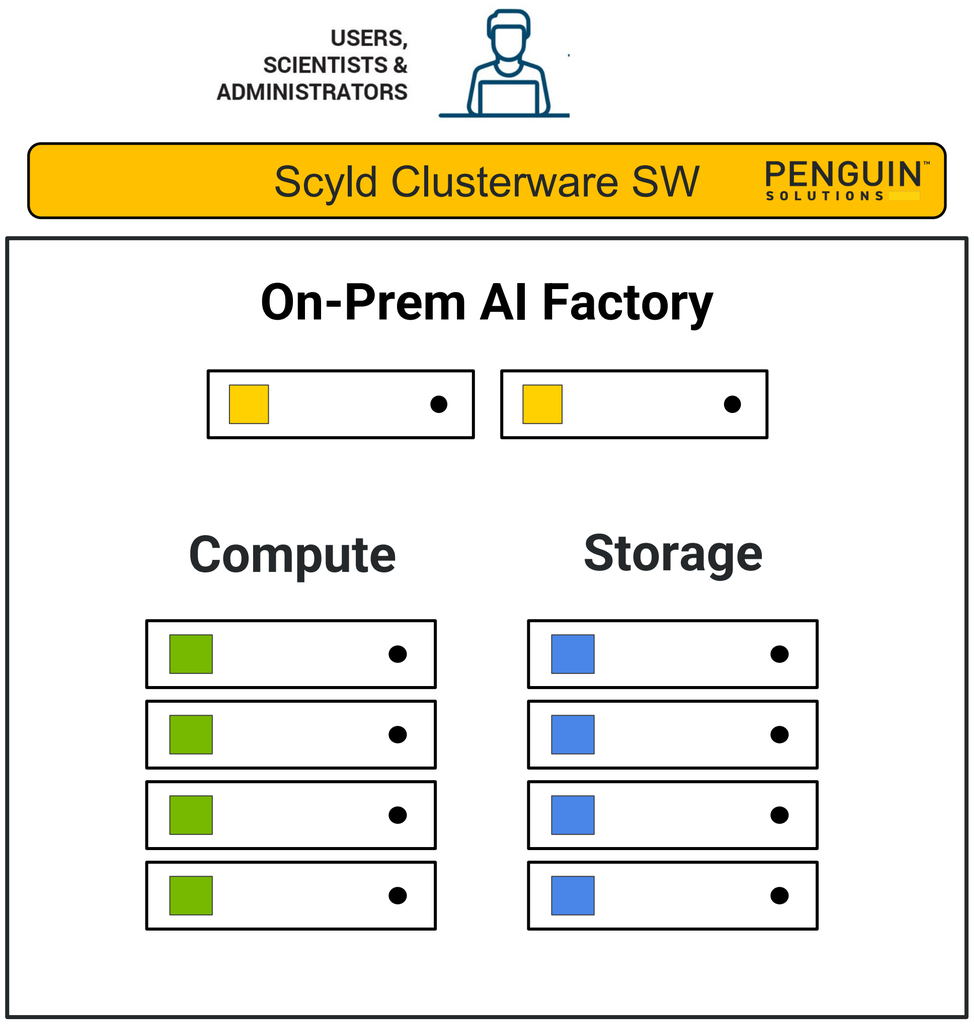
- Administrators spend time fixing problems instead of laboring over repetitive tasks.
- Automatically bringing nodes back into production after qualifying tests post-repair.



Start with nothing… *…end with a running cluster.*

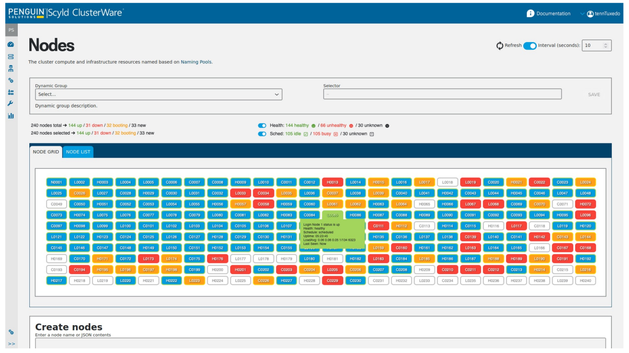
PENGUIN SOLUTIONS

Scyld Clusterware: AI Factory Enablement at Scale



Cluster Management Software that allows Administrators to efficiently manage hardware and software resources to get the most of your AI jobs.

- Tame Complexity
- Monitoring & Alerting
- Rapid Provisioning & Extensibility
- Supported Schedulers (Slurm, Kubernetes, etc.)
- Highly Secure



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

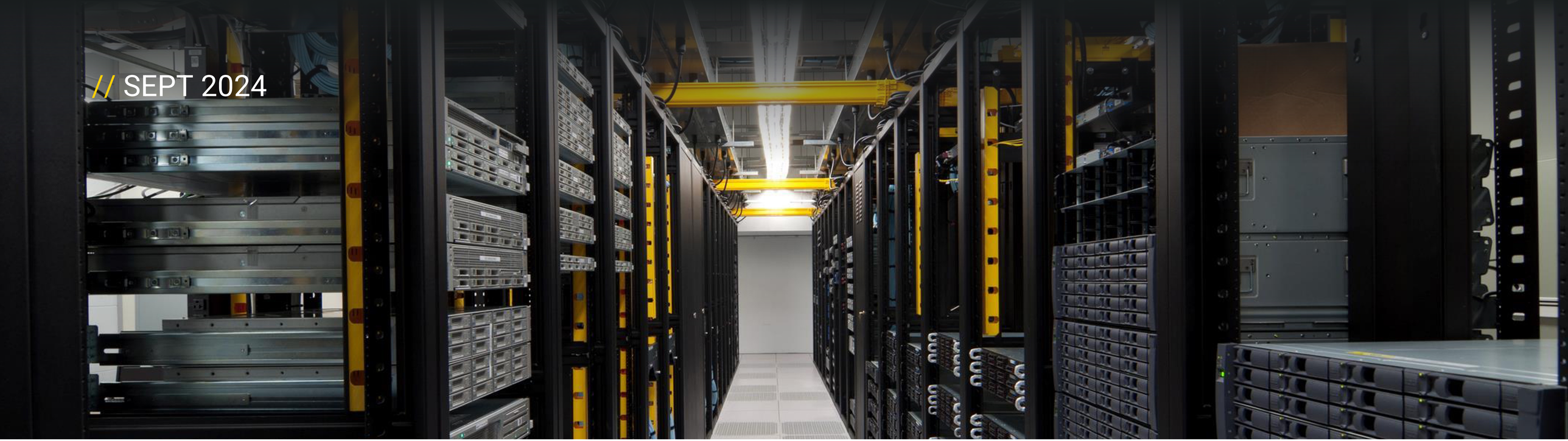
- Penguin Data Center
- Customer Data Center
- Power, Space, and Cooling Management
- As-a-Service Billing

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



// SEPT 2024

THANK YOU

WE APPRECIATE YOUR TIME

PENGUIN™ SOLUTIONS