

# Case Study: Deploying GenAI Infrastructure with GPU Servers, OpenShift & NVIDIA Networking in an IT Software Development Organization

**A comprehensive examination of transformative AI infrastructure deployment in the Indian IT sector**



# Setting the Stage: Why GenAI Infrastructure Matters



## Massive Compute Power

**GenAI workloads demand unprecedented parallel processing capabilities and GPU acceleration for training complex neural networks**



## Low Latency Networking

**High-speed interconnects essential for distributed training across multiple GPU nodes without bottlenecks**



## Scalable Orchestration

**Dynamic workload management and resource allocation to handle varying AI development cycles efficiently**

# The Organisation & Challenge



## Company Profile

**Mid-sized Indian IT software development company with 2,000+ employees, specialising in enterprise solutions and seeking to embed GenAI capabilities into product development pipelines**

## Infrastructure Limitations

**Existing infrastructure lacked GPU scale and AI-optimised networking, creating significant bottlenecks in model training and inference operations**

## Strategic Requirement

**Needed a scalable, secure, and manageable platform to accelerate AI model development whilst ensuring enterprise-grade reliability and performance**

# Technology Stack Overview



## GPOU Servers

**High-density GPU clusters specifically optimised for AI/ML workloads, delivering unmatched parallel compute performance with advanced cooling and power management systems**



## OpenShift Platform

**Enterprise Kubernetes platform enabling containerised AI workloads with automated scaling, security policies, and comprehensive lifecycle management capabilities**



## NVIDIA Networking

**High-throughput, low-latency InfiniBand and NVLink interconnects ensuring rapid data movement across GPUs with minimal communication overhead**



# Deployment Strategy & Architecture

01

---

## Infrastructure Foundation

**GPOU GPU servers integrated into existing data centre with enhanced power and cooling infrastructure to support high-density compute requirements**

03

---

## Network Optimisation

**NVIDIA networking components installed to connect GPU nodes with high bandwidth and minimal latency for optimal distributed computing performance**

02

---

## Orchestration Layer

**OpenShift deployed as the primary orchestration platform for containerised AI workloads with automated resource allocation and workload scheduling**

04

---

## Hybrid Readiness

**Cloud-ready design architecture enabling burst scaling with public cloud GPU resources for peak workload demands and disaster recovery scenarios**

# Overcoming Key Challenges



## Bottleneck Elimination

**NVIDIA InfiniBand reduced GPU-to-GPU communication latency by 40%, significantly accelerating distributed training workflows and model convergence times**



## Dynamic Scalability

**OpenShift's auto-scaling capabilities enabled seamless handling of fluctuating AI workloads without manual intervention, optimising resource utilisation**



## Security & Compliance

**OpenShift's built-in security guardrails combined with GPU server hardware encryption ensured data privacy and regulatory compliance requirements**



## Operational Excellence

**Centralised monitoring and management systems reduced AI infrastructure downtime to under 2%, ensuring consistent development productivity**

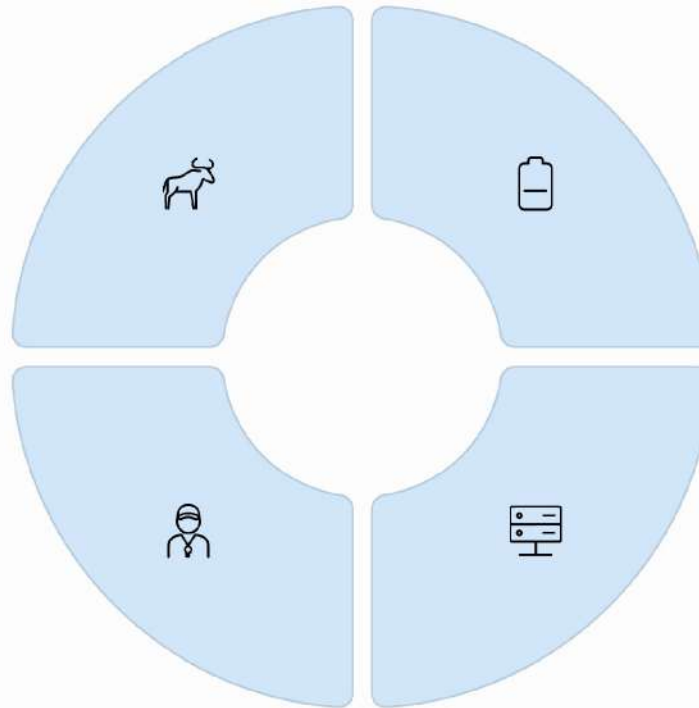
# Strategic Partnerships & Expertise

## NVIDIA Collaboration

**Hardware expertise and networking optimisation guidance for maximum GPU utilisation and performance tuning**

## Skills Development

**Internal upskilling programmes trained 50+ engineers on AI infrastructure management and DevOps practices**



## OpenShift Support

**Rapid container orchestration deployment with expert tuning and configuration optimisation services**

## GPOU Partnership

**Tailored hardware configurations specifically optimised for GenAI workloads and enterprise requirements**

# Impact & Results

**30%**

**Training Speed Increase**

**Reduction in time-to-market for AI-powered features through accelerated model development**

**25%**

**Developer Productivity**

**Boost through integrated AI tooling and streamlined workflows on OpenShift platform**

**98.5%**

**Infrastructure Uptime**

**Supporting 24/7 AI development cycles with minimal disruption to critical workloads**

**15+**

**GenAI Use Cases**

**Deployed across development, testing, and customer support automation processes**

**The transformation delivered measurable improvements across all key performance indicators, establishing a robust foundation for continued AI innovation and competitive advantage in the market.**

# Visualising the Transformation

*GPOU GPU clusters interconnected via NVIDIA InfiniBand networking*



*OpenShift orchestration managing AI workloads in containers*

## Architecture Highlights

- **High-speed data flow between GPU nodes**
- **Container orchestration with automatic scaling**
- **Integrated monitoring and management layers**
- **Hybrid cloud connectivity for burst capacity**



# Building the AI-First Future

## Key Achievements

- Empowered organisation to lead AI innovation in software development
- Created scalable, secure, high-performance platform
- Unlocked new productivity and operational efficiencies
- Established blueprint for Indian IT firms targeting GenAI at scale

**Next Steps:** Expanding hybrid cloud integration and exploring agentic AI workloads for autonomous software engineering capabilities

