AI Infrastructure Engineering

Private AI, GPU clusters and LLM platforms | Senior engineering capacity, delivered white-label

TELEKOMPLEX People | Expertise | Results

Enterprises are moving AI from pilots to production, and many want it on infrastructure they control. That takes skills most delivery teams do not keep on staff: GPU clusters, high-speed interconnects, parallel storage and the operations discipline to keep them productive. Telekomplex extends your bench with senior engineers who have built and run these environments, so you can say yes to AI infrastructure work.

The 10 AI infrastructure services enterprises are asking for

01 Private LLMs & production inference
Model serving for open models, GPU sizing, access control and monitoring.

02 AI workload repatriation
Placement analysis, target design and migration for training, inference and data.

03 GPU cluster design & deployment
Compute, interconnect and storage design, bring-up, validation and handover.

04 GPU utilization & cost optimization
Scheduler policy, queue design and capacity reviews that raise real utilization.

05 Managed AI cluster operations
Day-2 operations: on-call escalation, upgrades, security maintenance and RCA.

06 Storage & data pipelines for AI
GPFS / Spectrum Scale, Lustre, NFS and object storage tuned for AI throughput.

07 AI infrastructure assessment & architecture
Current-state review, workload sizing, risks and a phased roadmap.

08 Private RAG platforms
Vector stores, model serving and data paths, deployed privately.

09 High-speed AI networking
Fabric design, tuning and troubleshooting for GPU-to-GPU traffic.

10 Fine-tuning environments
Multi-GPU environments your ML team can use to adapt open models.

Bench depth behind the service

- **25+ years** across Linux/Unix, HPC, enterprise storage, virtualization and cloud reliability.
- **Principal-level site reliability** experience inside a hyperscale cloud provider's object storage service, including team leadership and complex incident ownership.
- **HPC lead** for multiple large Linux clusters at a top-five global pharmaceutical company: Slurm, MPI, GPFS, NetApp/NFS and InfiniBand/RDMA.
- **Rack-to-production commissioning** of HPC systems: interconnects, Linux integration, tuning and production support.
- **Hands-on private AI:** builds and operates a multi-GPU Linux platform for local inference and retrieval-augmented generation.

How we engage

- A senior specialist for an escalation or a design review
- Capacity for a defined project
- Pre-sales support: architecture, sizing and PoCs
- Ongoing operations coverage, by the hour or by project

Technologies

- RHEL, SLES, Ubuntu; CUDA and multi-GPU inference
- Slurm, MPI, InfiniBand / RDMA
- GPFS / Spectrum Scale, Lustre, NetApp/NFS, ZFS, object storage
- Terraform, Ansible, Python, Bash
- VMware, Nutanix, KVM; OCI

Built for the channel

Your team stays in front of the customer. We work inside your tools and governance model, with no permanent headcount.

The right approach depends on your environment, your workloads and the tools your teams actually use. Kubernetes-based GPU scheduling and data-science work such as model fine-tuning are scoped per engagement.

Have an AI infrastructure opportunity on the table? Send us the scope.

sni@telekomplex.net | telekomplex.net