

stack8s – HPC Design with Kubernetes for AI/ML

G-Cloud 15 Service Definition

HPC Design with Kubernetes for AI/ML is a Lot 3 Cloud Support service that helps public sector organisations design and deploy high-performance computing infrastructure using Kubernetes for AI and machine learning workloads. It provides architecture design, GPU cluster configuration, Slurm integration, and containerised ML pipeline deployment to deliver scalable, cost-effective compute for model training and inference.

1. Service description

HPC Design with Kubernetes for AI/ML provides consultancy and implementation services to build GPU-accelerated Kubernetes clusters tailored for AI, ML, and data-intensive workloads. The service covers high-level HPC architecture design, detailed cluster configuration, integration of Slurm workload management, and deployment of ML pipelines and serving infrastructure.

It is aimed at organisations that need to run computationally intensive AI training and inference, but want the elasticity and automation of cloud-native platforms rather than traditional fixed HPC clusters. Engagements are delivered by engineers with experience in Kubernetes, GPU infrastructure, and MLOps, working closely with internal platform and data science teams.

2. Service features and benefits

2.1 Service features

- **HPC architecture design for AI/ML workloads** – Design of overall cluster topology, node types, GPU placement, and scheduling strategy for training and inference.
- **Kubernetes cluster setup for GPU-accelerated computing** – Installation and configuration of Kubernetes clusters optimised for GPU workloads in cloud or hybrid environments.
- **Slurm workload manager integration with Kubernetes** – Integration of Slurm with Kubernetes to orchestrate batch, MPI, and AI training jobs on shared GPU resources.

- **ML pipeline deployment using Kubeflow and MLflow** – Setup of Kubeflow pipelines and MLflow for experiment tracking to support reproducible ML workflows.
- **NVIDIA GPU Operator and device plugin configuration** – Configuration of GPU operators and device plugins to manage GPU drivers, runtime, and resource exposure to workloads.
- **High-speed storage and networking for AI training** – Design and configuration of storage (for example, high-throughput shared file systems or object storage) and high-bandwidth networking for data-intensive training jobs.
- **Model serving infrastructure with TensorFlow Serving and Triton** – Deployment of serving stacks for real-time or batch inference, including autoscaled model endpoints.
- **Distributed training setup with Horovod and Ray** – Configuration of frameworks such as Horovod and Ray for distributed model training across multiple GPUs and nodes.
- **Auto-scaling configuration for training and inference workloads** – Policies for horizontal and vertical scaling of GPU and CPU resources based on demand.
- **Observability and monitoring for ML operations** – Implementation of monitoring and logging for clusters, jobs, and ML pipelines to support MLOps practices.

2.2 Service benefits

- **Accelerated AI model training** through optimised GPU infrastructure and distributed training frameworks.
- **Reduced time-to-deployment for ML models** by providing a ready-to-use platform for pipelines and serving.
- **Dynamic scaling of compute resources** so GPU capacity grows and shrinks with workload demand.
- **Improved HPC throughput** via Slurm-orchestrated cloud resources that maximise utilisation.
- **Lower GPU costs** achieved through efficient scheduling, auto-scaling, and use of spot or pre-emptible instances where appropriate.
- **Reproducible ML experiments** using containerised workflows, Kubeflow pipelines, and MLflow experiment tracking.

- **Optimised compute spend** by aligning resource allocation and scaling policies with business priorities.
 - **Higher data-science productivity** as teams are freed from managing low-level infrastructure.
 - **High availability and resilience** through fault-tolerant cluster design and appropriate redundancy.
 - **Simplified MLOps** with integrated monitoring, logging, and standardised deployment patterns.
-

3. Service constraints

- Support and delivery are primarily remote; on-site engagements are available by agreement and may be scoped separately.
 - Customer environments must meet agreed minimum network, security, and access requirements to support GPU-enabled Kubernetes clusters.
 - Cluster deployments and migrations are scheduled within defined change windows agreed with the buyer.
 - Scope excludes unrelated application feature development that is not part of HPC or ML infrastructure design and deployment.
 - Successful delivery depends on timely access to subject-matter experts, including platform, security, and data-science stakeholders.
-

4. User support

- **Email and ticketing support:** Available, with responses to tickets and questions within four working hours Monday–Friday, 9am–5pm GMT; outside these hours responses are best-effort and triaged on the next business day.
- **Phone support:** Available 9am–5pm UK time, Monday to Friday for active engagements.
- **Web chat support:** Available 9am–5pm, seven days a week with an AI chatbot, on a platform designed to meet WCAG 2.2 AA accessibility standards (keyboard-only navigation, screen-reader checks, high-contrast verification).

HPC Design with Kubernetes for AI/ML includes a single standard support level for all engagements, delivered remotely by the consulting team and included within day-rate or fixed-price costs (no per-user or per-ticket charges). For larger programmes, an assigned technical account manager or lead cloud support engineer can act as a primary contact for planning, escalation, and service reviews, scoped and priced in the statement of work.