

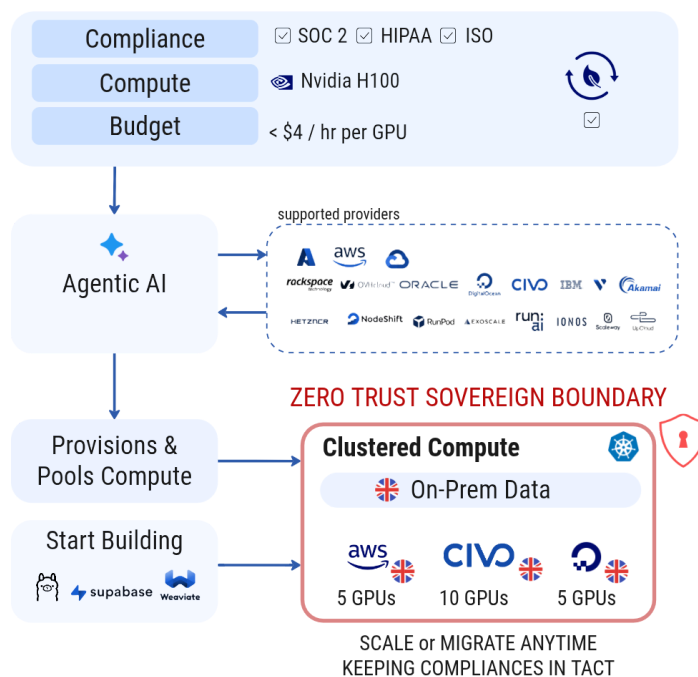
Sovereign AI Compute Management SaaS

G-Cloud 15 Service Definition

1. Service description

Sovereign AI Compute Management SaaS is an infrastructure software service that provides secure, sovereign, multi-cloud AI and GPU compute management for public sector and regulated organisations. It enables customers to deploy, manage, and scale containerised workloads—including AI/ML, data analytics, and research applications—across on-premise environments, UK-hosted infrastructure, and approved public cloud providers under a unified control plane.

The platform offers an AI/ML and container orchestration layer built on hardened Kubernetes, combined with integrated DevOps and MLOps capabilities. Organisations can standardise how they run GPU and CPU workloads across heterogeneous infrastructure while enforcing data sovereignty, cost controls, and security policies that align with NCSC and CIS guidance. Access is via a secure web-based management portal and APIs/Terraform, supporting infrastructure-as-code workflows and CI/CD integration.



2. Service features and benefits

2.1 Service features



- **AI/ML and container orchestration platform**
Provides a Kubernetes-based orchestration environment optimised for AI/ML, analytics, and containerised applications across on-premise and cloud infrastructure. Users can deploy, scale, and manage workloads consistently regardless of underlying provider, with support for GPU and CPU resources.
- **Military-grade Kubernetes with NCSC and CIS alignment**
The Kubernetes platform is hardened according to recognised best practices and aligned with NCSC guidance and CIS Benchmarks. This includes secure configuration of control plane and nodes, network policies, and access controls suitable for sensitive public sector workloads.
- **Dynamic cost, performance, and carbon footprint optimisation**
The service incorporates mechanisms to optimise where and how workloads run based on cost, performance, and environmental factors. This allows organisations to balance budgets, SLAs, and sustainability objectives by utilising the most appropriate sovereign infrastructure at any given time.
- **Certified data security and information assurance**
Platform security controls, monitoring, and operational processes are designed to support high assurance environments, with clear separation of responsibilities between stack8s and customers. This underpins secure handling of data and workloads across the multi-cloud and hybrid estate.
- **Secure networking mesh for data sovereignty and resilience**
A secure network mesh connects on-premise environments and sovereign cloud providers, enforcing jurisdictional boundaries and minimising unnecessary egress. This mesh supports resilient, low-latency connectivity for distributed workloads while maintaining data sovereignty guarantees.
- **Unified control plane for multi-cloud resource governance**
A single control plane provides visibility and governance across all clusters and environments, including resource utilisation, policies, and access. Administrators can define and enforce standards for security, placement, and capacity across on-prem, UK-hosted, and approved public clouds.
- **Integrated DevOps and MLOps**
The platform is designed to integrate with CI/CD pipelines and MLOps tooling so that infrastructure provisioning, application deployment, and ML workflows can be automated end-to-end. APIs and Terraform interfaces support infrastructure-as-code, while integrations with tools such as Kubeflow and databases provide a complete AI stack.
- **LLM marketplace, Database as a Service, and Kubeflow as a Service**
An LLM marketplace exposes curated AI models that can be deployed onto the managed clusters, while managed database services and Kubeflow as a Service supply persistent data and MLOps capabilities directly within the same sovereign platform. This enables organisations to run

AI workloads, including large language models, alongside their data and pipelines under unified governance.

2.2 Service benefits

- **Guaranteed resilience and compliance through multi-cloud clustering**
Multi-cloud clustering and failover across sovereign providers support high availability and compliance with data residency requirements. Organisations gain resilience against outages in any single provider while maintaining control over where data and workloads run.
- **Accelerated time-to-value**
By providing a pre-integrated platform for AI, containers, and networking, the service significantly reduces the time needed to bring new workloads and services into production. Teams can focus on applications and models rather than assembling and operating infrastructure components.
- **Significant cost and carbon savings**
Dynamic optimisation across providers and environments, combined with visibility into utilisation and cost metrics, allows organisations to reduce unnecessary spend and improve energy efficiency. This supports both financial and sustainability objectives.
- **Elimination of vendor lock-in**
Workloads are deployed on open, Kubernetes-based infrastructure that spans multiple providers, reducing reliance on any single proprietary platform. This gives organisations greater flexibility in long-term sourcing and architecture decisions.
- **Access to UK-based specialist expertise**
Support and operations are delivered by UK-based engineers with experience in Kubernetes, GPU infrastructure, and multi-cloud/hybrid architectures. This supports public sector assurance needs and provides direct access to specialist knowledge.
- **Unbreakable security and data integrity**
Strong isolation between tenants, hardened Kubernetes, and secure networking mesh capabilities contribute to robust protection of workloads and data. Combined with monitoring and audit trails, this underpins trustworthy operation of critical AI and research applications.
- **Assured service continuity (SLA) and seamless workload migration**
The SLA focuses on the stack8s control plane and network mesh, with defined support response and resolution targets and access to professional services for complex migrations. Workloads can be moved between on-premise and sovereign cloud environments under the same governance model, simplifying migration and modernisation projects.

3. Service constraints

- **Planned and emergency maintenance windows**
Planned maintenance occurs weekly on Sundays between 00:00 and 05:00 GMT, with at least 7 days' advance notice. Emergency maintenance may be undertaken with a minimum of 4 hours' notice where urgent changes are required, potentially causing temporary service impact.
 - **Scope of support and shared security model**
Support is limited for customer-deployed Kubernetes objects, unsupported hardware, and non-native runtimes. Under the shared security model, stack8s secures and operates the platform (control plane and network mesh), while customers are responsible for securing their data, applications, and any custom workloads.
 - **SLA coverage and provider dependency**
The SLA covers the stack8s control plane and private network mesh but excludes the availability of third-party IaaS providers. Underlying cloud provider incidents may affect workloads until mitigated via the multi-cloud architecture.
 - **Data ingress/egress responsibilities and mesh optimisation**
Customers are responsible for data ingress and egress costs with underlying providers. The platform's private network mesh is designed to significantly reduce egress charges by optimising traffic paths between environments.
 - **Professional services for out-of-scope work**
A billable Professional Services team is available for ad-hoc support, complex migrations, and custom features not covered by the standard managed service.
-

4. Technical requirements

- **Client and identity prerequisites**
 - Latest stable web browser with HTTPS and JavaScript enabled to access the management portal.
 - Active stack8s account with appropriate security access and role configuration.
 - Customer identity provider must support SAML 2.0 or OAuth 2.0 for federated authentication.
- **Application packaging and runtime requirements**



- All applications must be packaged as OCI-compliant container images suitable for deployment on Kubernetes.
- **Network and connectivity requirements**
 - Outbound HTTPS (port 443) access from customer environments to the stack8s control plane is required.
 - Customer VPC/VNet must support VPN peering or equivalent secure connectivity for inclusion in the managed private network mesh.
 - Customer security policies must permit the required managed infrastructure connectivity.
- **Commercial and security prerequisites**
 - An active stack8s billing agreement is required for provisioning and running resources under the service.
 - Customers must install endpoint protection on all host environments under their control that participate in or connect to the managed infrastructure.