

Service Definition

Sovereign AI Compute Management SaaS

1. Service Overview

Sovereign AI Compute Management SaaS is an infrastructure software service that enables UK public sector organisations to securely deploy, manage, and scale containerised workloads, including AI, data analytics, and research applications, across sovereign, multi-cloud, and hybrid environments. The service provides a unified Kubernetes-based control plane for managing compute, GPU, networking, and platform services while maintaining data sovereignty, regulatory compliance, and operational resilience.

2. Intended Users

This service is designed for:

- UK public sector bodies
- Research institutions and universities
- Regulated organisations requiring sovereign AI and GPU compute
- Technical teams operating Kubernetes, container platforms, or AI workloads

Users typically include platform engineers, DevOps teams, data scientists, and IT operations staff.

3. Service Scope and Functionality

The service provides:

- A managed control plane for Kubernetes and container infrastructure
- Multi-cloud and hybrid cluster management across on-premise, UK-hosted, and approved public cloud environments
- Secure GPU and AI workload orchestration
- Integrated networking, identity, and access management
- Centralised monitoring, logging, and audit capabilities

The service focuses on infrastructure management and orchestration. Customers retain responsibility for their applications, data, and workloads under a shared responsibility model.



4. Key Features

- Kubernetes-based container orchestration platform
 - Multi-cloud and hybrid cluster management
 - Secure network mesh for sovereignty and resilience
 - Integrated DevOps and MLOps tooling
 - AI/ML platform support including Kubeflow and model lifecycle management
 - Marketplace for Helm charts, operators, and AI models
 - Cost, performance, and carbon-aware optimisation
 - Role-based access control with identity federation
-

5. Service Benefits

- Enables sovereign and compliant AI compute operations
 - Reduces vendor lock-in through multi-cloud portability
 - Accelerates deployment of AI and data workloads
 - Improves resilience through distributed architectures
 - Simplifies complex infrastructure management
 - Provides predictable operational controls for regulated environments
-

6. Deployment and Architecture

The service supports:

- Public cloud deployments
- Private cloud deployments
- Community cloud deployments
- Hybrid cloud and on-premise deployments



Customers connect their environments to the Stack8s control plane using secure networking and identity federation. All deployments use containerised, declarative infrastructure.

7. Security and Compliance

Security controls align with UK cloud security principles and recognised best practices:

- Encryption of data in transit using TLS 1.2 or above
- Encryption of data at rest
- Role-based access control and least-privilege access
- Multi-factor authentication and identity federation via Auth0
- Regular penetration testing by CHECK-approved providers
- Centralised audit logging retained for at least 12 months

Data storage and processing locations are known and configurable, with support for UK and EEA data residency.

8. Availability and Resilience

The service is designed for high availability and resilience using:

- Distributed Kubernetes control planes
- Automated failover and self-healing
- Multi-zone and multi-cloud deployment options
- Monitoring, alerting, and incident response procedures

Standard availability targets start at 99.5% monthly uptime, with higher guarantees available by agreement.

9. Onboarding and Service Management

Onboarding includes:

- Guided technical onboarding
- Secure environment configuration



- Training sessions and documentation
- Support for integrating existing cloud or on-premise infrastructure

The service is accessed via a secure web interface and APIs.

10. Support

Support is provided by UK-based engineers and includes:

- Online ticketing and email support
 - Optional enhanced and managed support tiers
 - Incident response and escalation procedures
 - Service status dashboard and outage notifications
-

11. Data Import, Export, and Exit

Customers can export their data at any time via UI or API, including:

- Kubernetes configurations and manifests
- Logs, metrics, and audit data
- AI models and datasets
- User and access configurations

At contract end, customers retain full control of their data and can complete self-service exports without vendor lock-in.

12. Free Trial

A time-limited free trial is available.

The trial includes access to core platform functionality and documentation but excludes production workloads, sustained compute capacity, and dedicated support.

13. Service Constraints



Planned maintenance occurs weekly with advance notice. Emergency maintenance may be required for security or stability. The service SLA applies to the Stack8s control plane and managed components, excluding third-party infrastructure providers.

14. Service Responsibility Model

Stack8s is responsible for securing and operating the platform.

Customers are responsible for:

- Their applications and data
 - Workload configuration
 - Compliance within their deployed environments
-

End of Service Definition

If you want, I can next:

- Convert this into **PDF/A-ready formatting**
- Tighten language further for **CCS reviewer preferences**
- Cross-check it line-by-line against the **G-Cloud supplier guide**
- Or draft the **Terms & Conditions** to match this definition