Avesha

Inferencing-as-a-Service:

AI Model Deployment at Scale

Unlock the true potential of AI with our Inferencing-as-a-Service platform. Deploy AI models at scale with ease and efficiency. Our solution is designed to tackle the growing demands of AI inference workloads.

The Growing Market Opportunity

Avesha

Exponential Growth

AI inference workloads are experiencing exponential growth.

Key Challenges

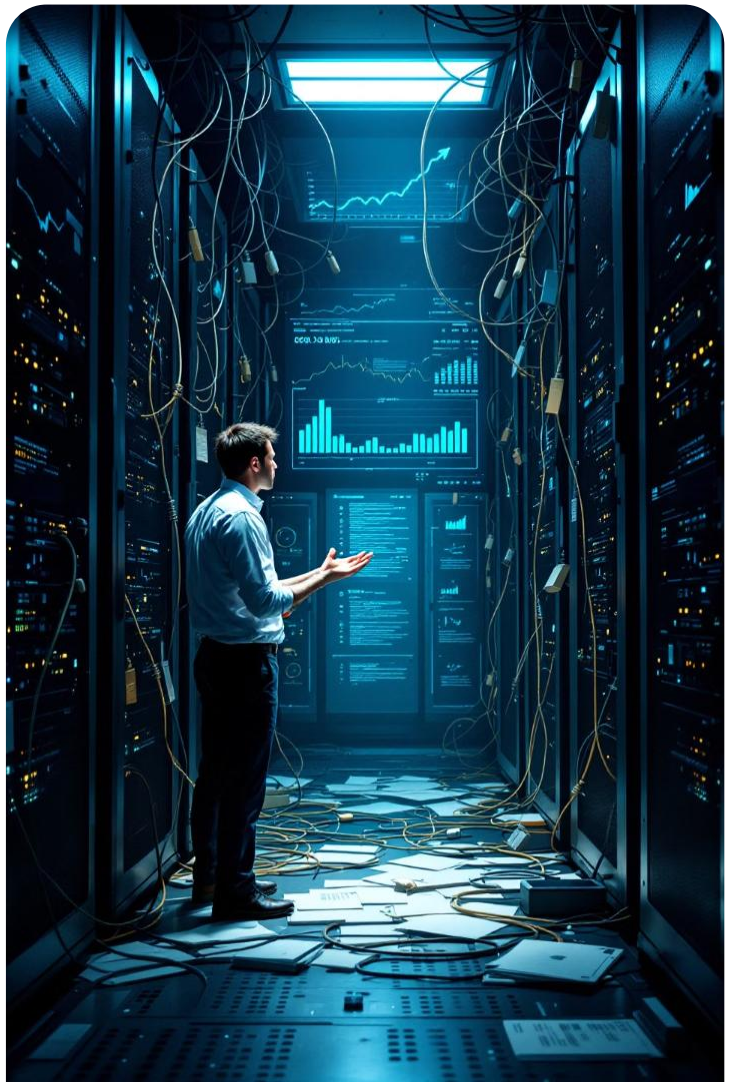
Managing GPU costs, latency, and scalability remains a challenge.

The Need

There is a growing need for on-demand, cost-efficient AI inferencing solutions.

The High Cost of AI Model Deployment

Avesha



Expensive & Inefficient

AI model deployment is often expensive and inefficient.

GPU Underutilization

GPU underutilization leads to unnecessary costs.

Lack of Flexibility

Traditional cloud inference lacks the necessary flexibility.

Our Solution:

On-Demand GPU Orchestration

Avesha

Unified Platform

Manage GPUs across on-premise and cloud environments.

On-Demand

Providing on-demand GPU orchestration for AI inference.

Intelligent Optimization

Optimizing workloads to lower costs and improve performance.

How Our Inferencing-as-a-Service Works

Avesha

1 **Connect GPUs**

Seamless integration with on-prem & cloud IaaS.

2 **Deploy AI Models**

Supports TensorFlow, PyTorch, ONNX, and more.

3 **Optimize Workloads**

Automated scaling & cost-efficient inference.

4 **Scale Dynamically**

Adapt to changing demands without manual intervention.

Key Benefits of Our Service

Avesha

Scalability

Auto-scale AI workloads dynamically and efficiently.

Cost Efficiency

Reduce GPU waste with pay-per-inference usage.

Performance

Optimize latency with smart GPU orchestration.

Security

Benefit from enterprise-grade encryption & compliance.

Our Competitive Advantages

1. **Multi-Cloud Flexibility**
   Offers multi-cloud & on-prem flexibility.

2. **Optimized Utilization**
   Optimizes GPU utilization for maximum efficiency.

3. **Faster Deployment**
   Enables faster model deployment cycles.

4. **Lower TCO**
   Results in a lower Total Cost of Ownership (TCO).

Diverse Industry Use Cases

Avesha

Finance

Fraud detection, high-frequency trading.

Healthcare

Medical imaging, accelerating drug discovery.

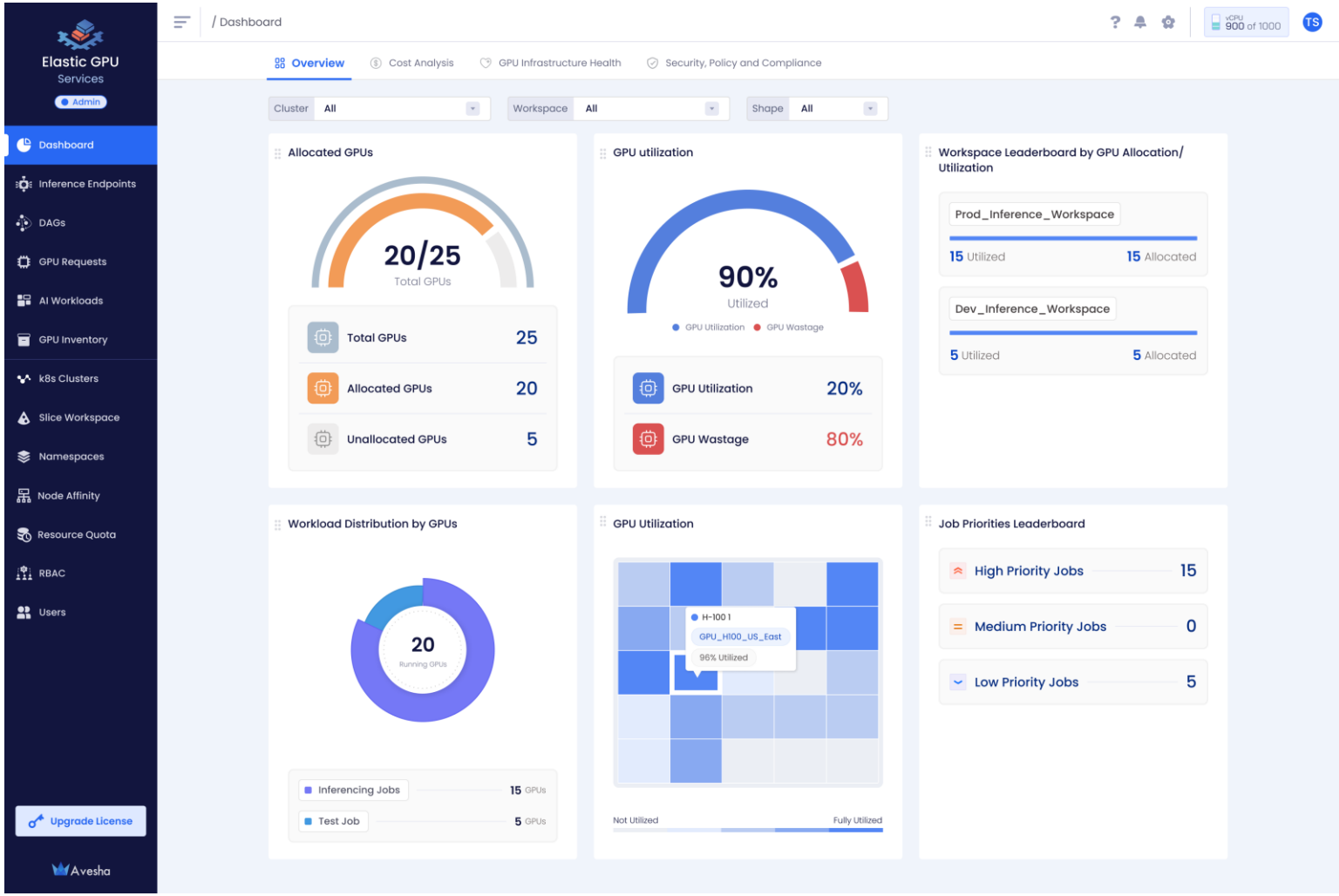
Retail

AI-powered customer insights, demand forecasting.

IoT & Manufacturing

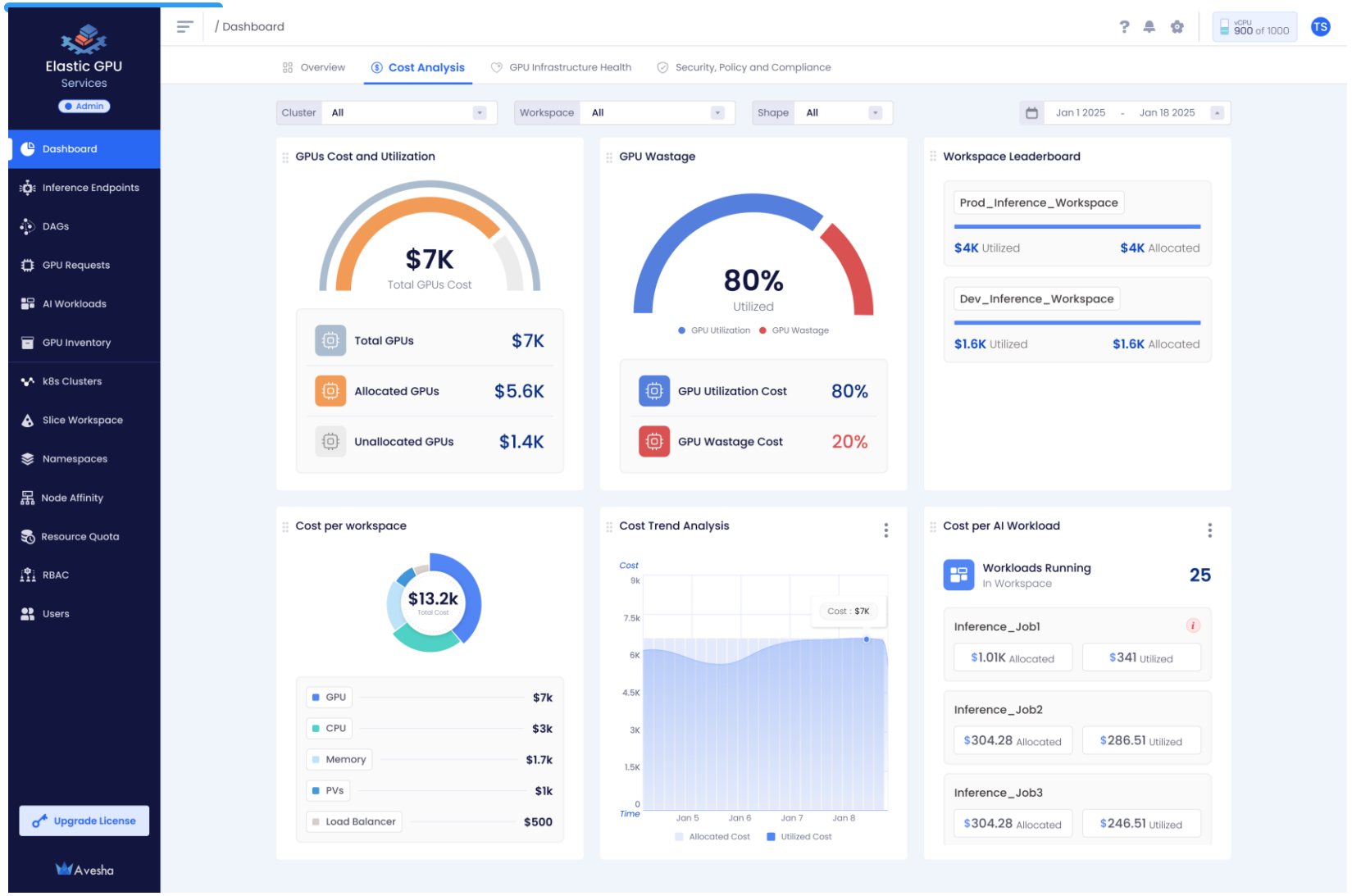
Predictive maintenance, anomaly detection.

EGS Dashboard - Overview



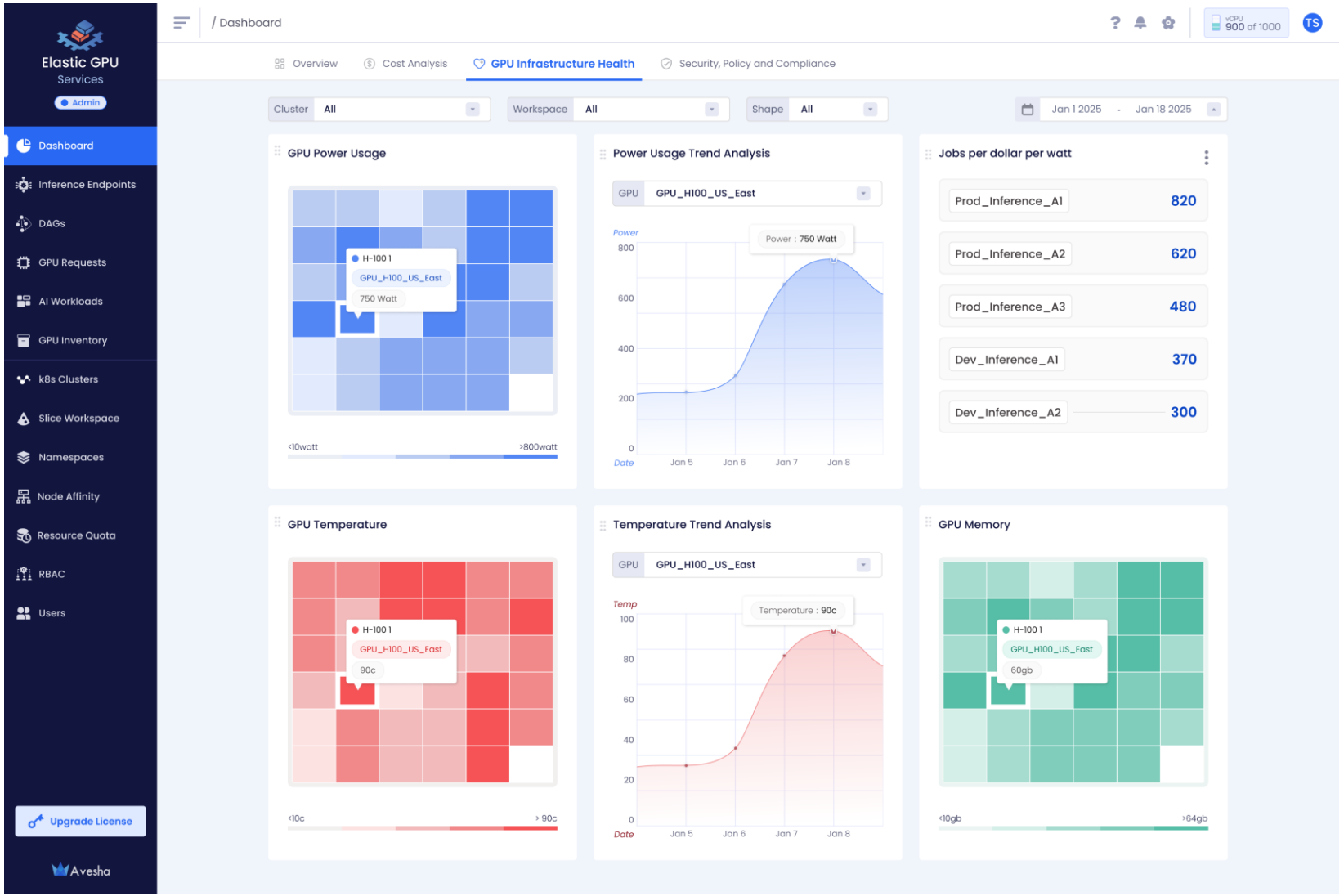
EGS Dashboard – Cost Analysis

Avesha

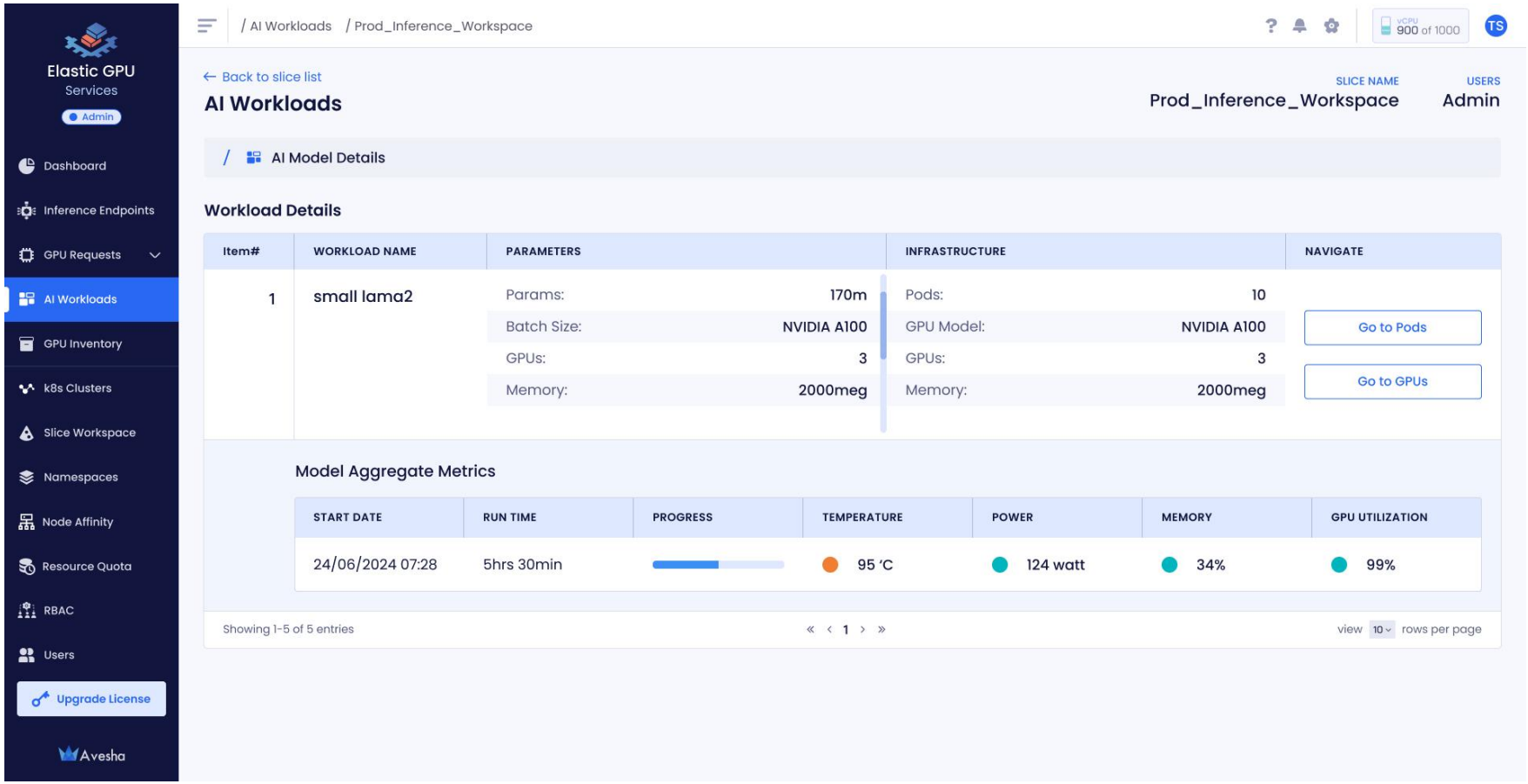


EGS Dashboard – GPU Infrastructure Health

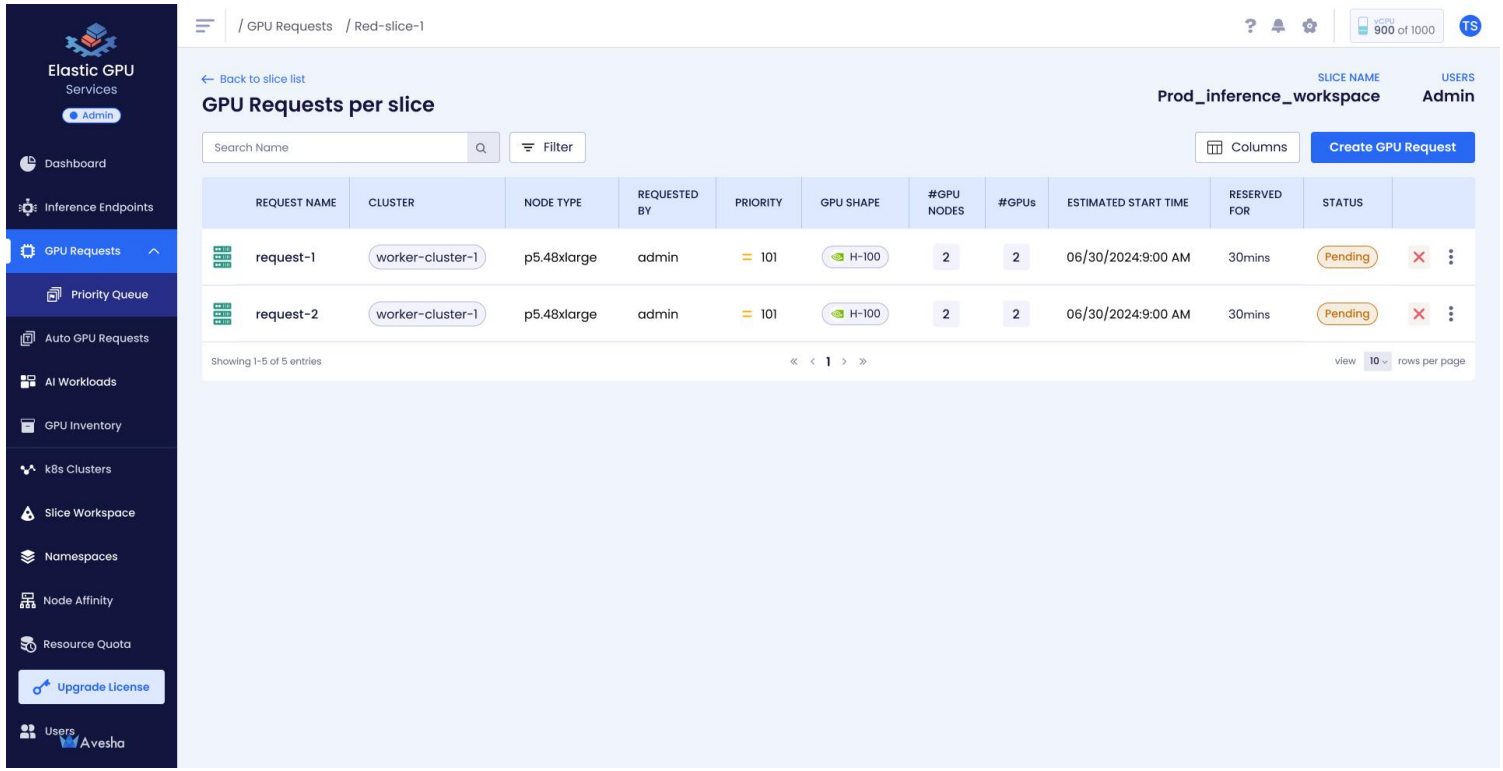
Avesha

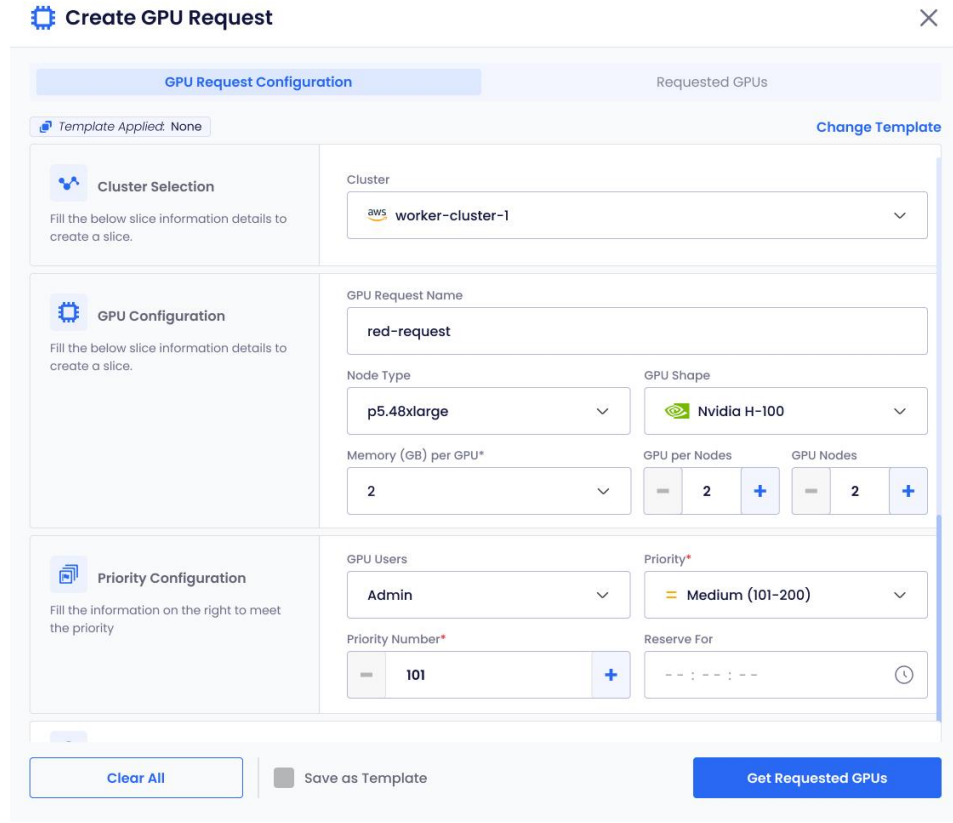


EGS : Model and Infra Parameters

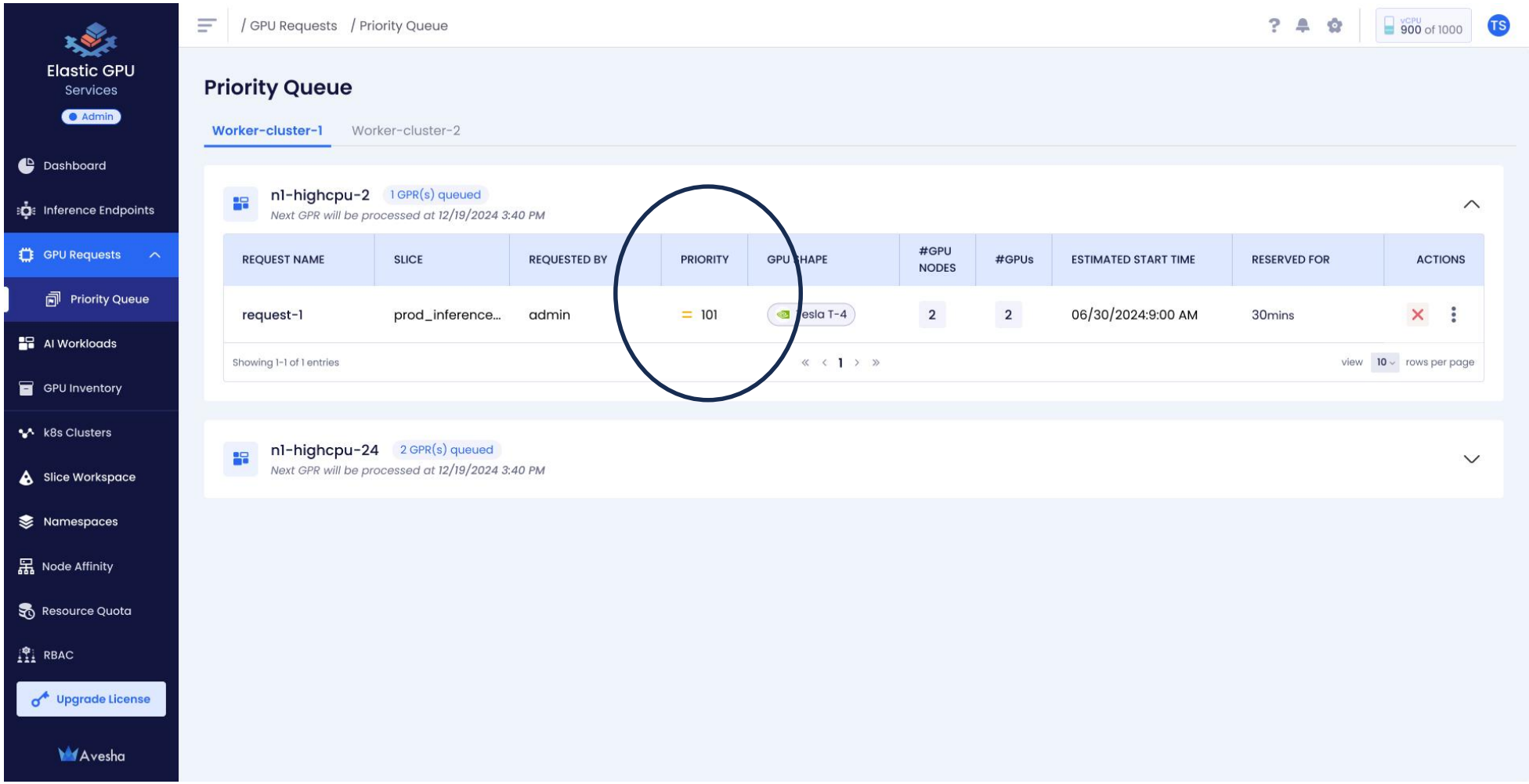


EGS : GPU Requests



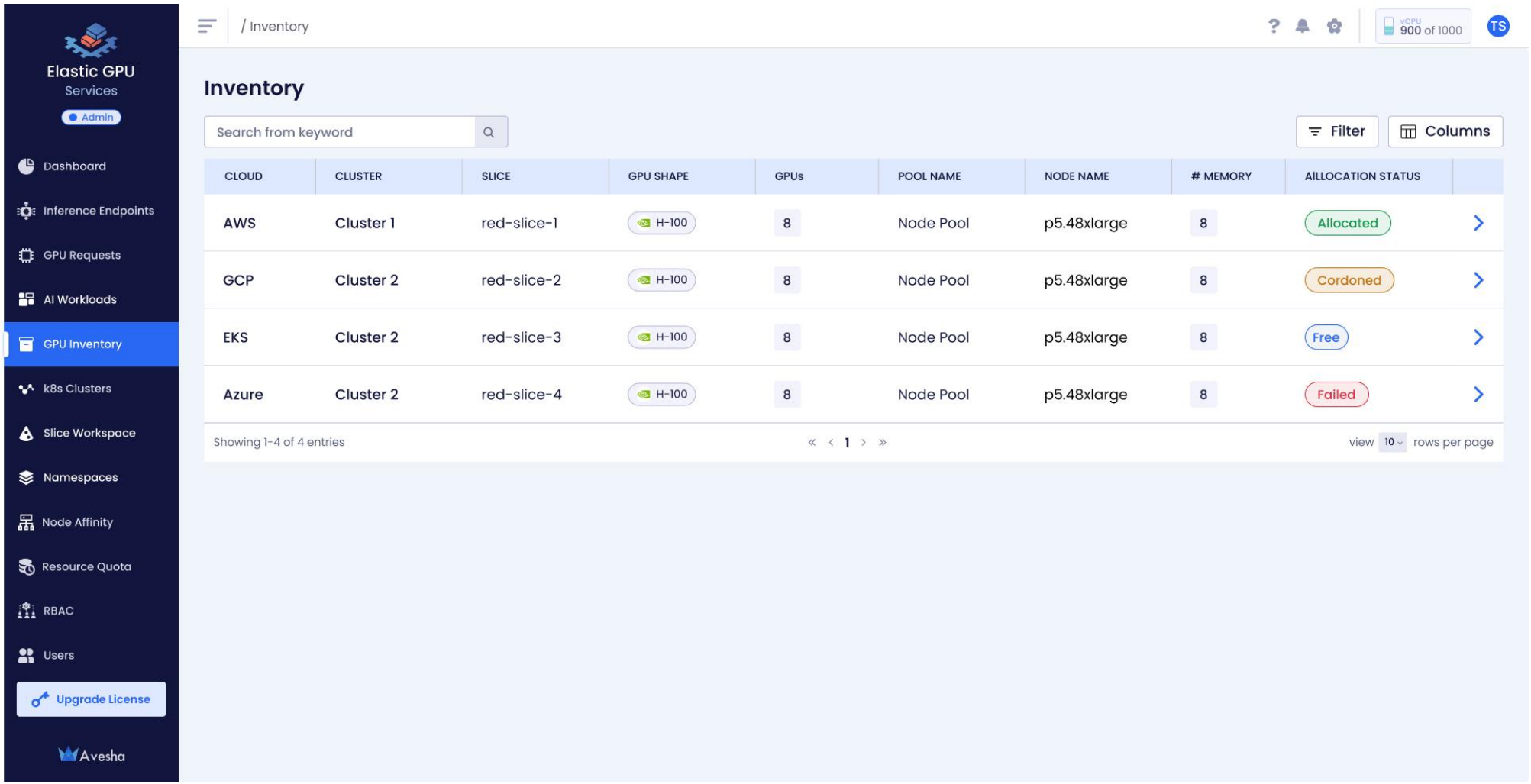


EGS : Priority Queue



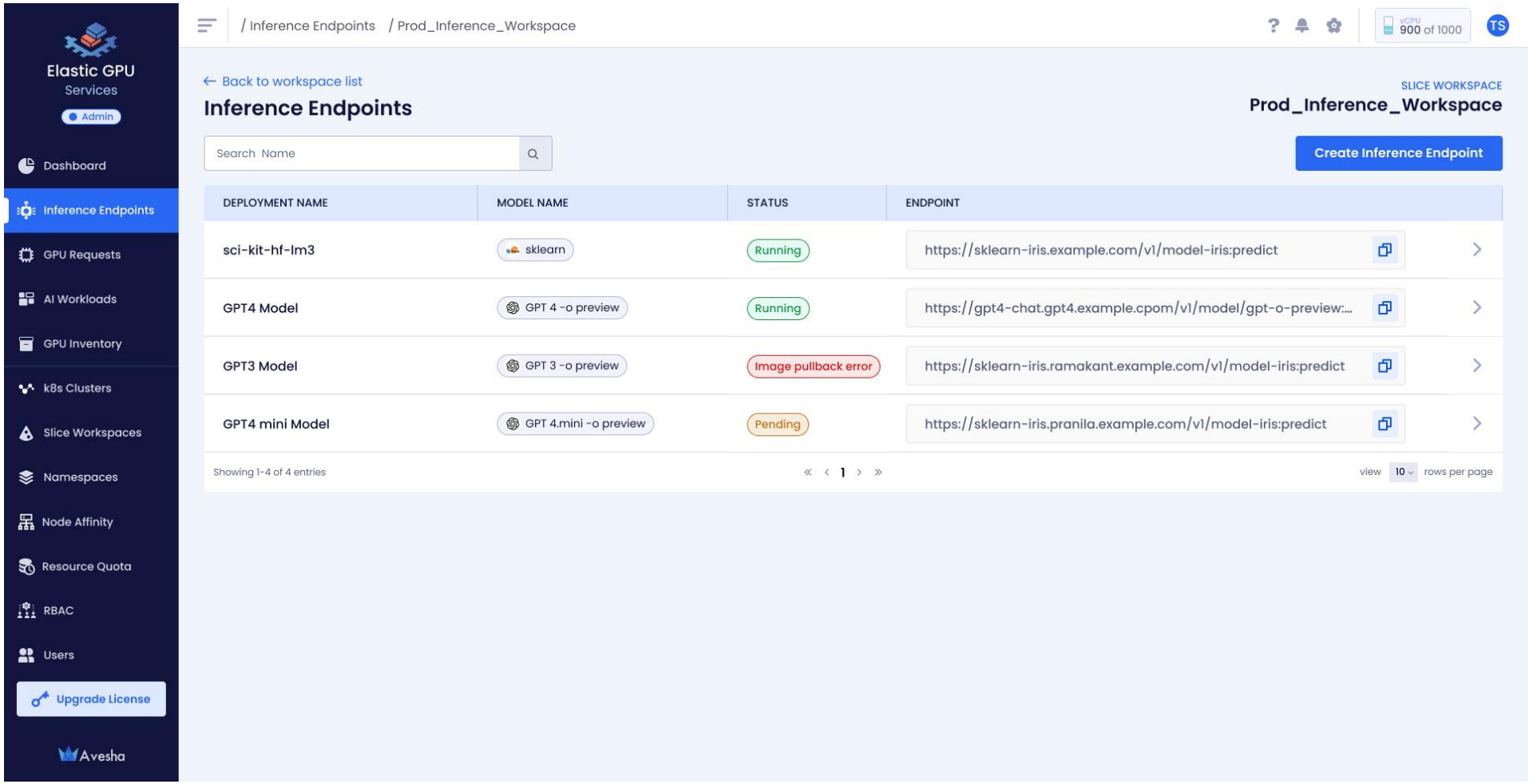
EGS : GPU Inventory

Avesha



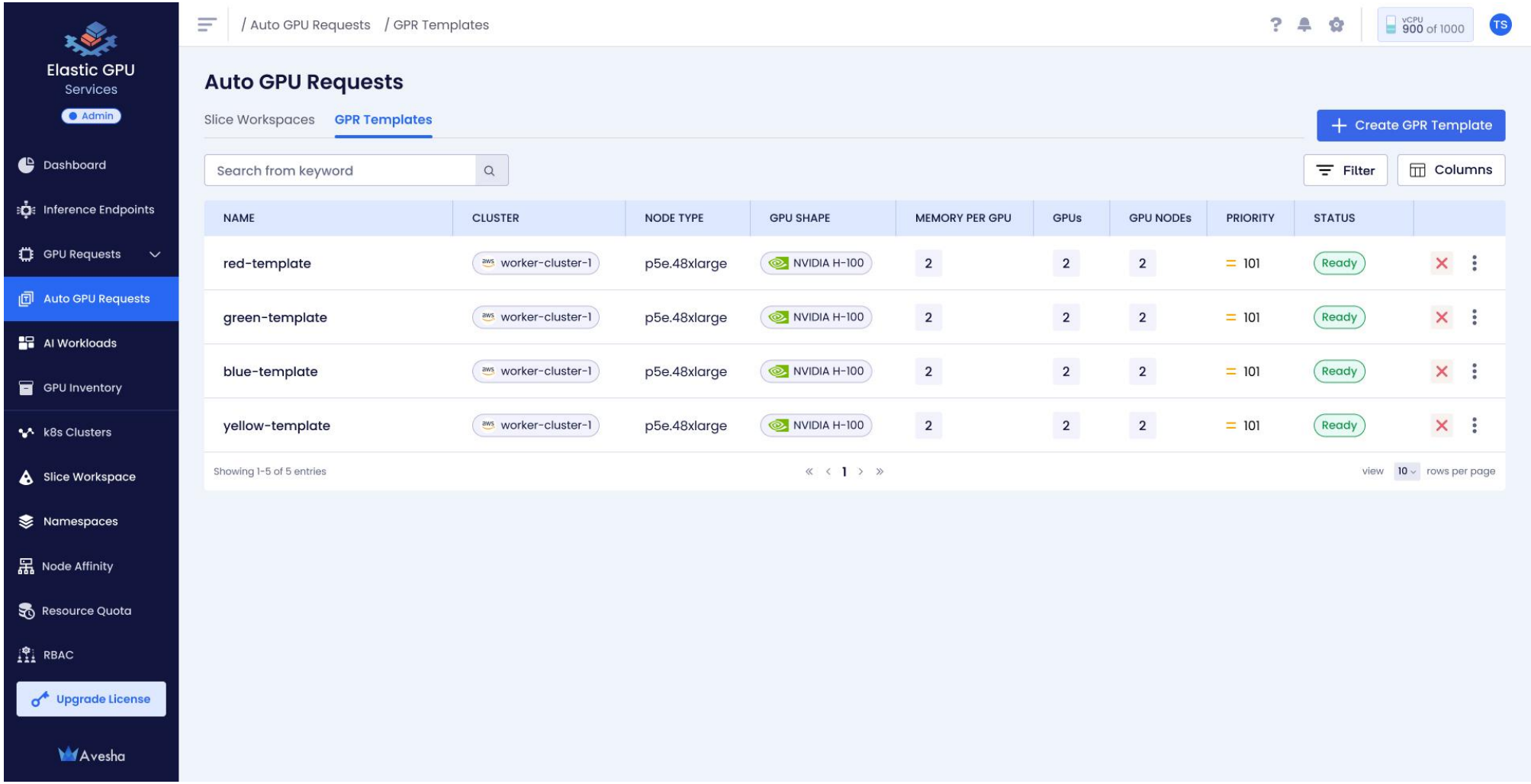
EGS : Inference EndPoints

Avesha



EGS : Automatic GPU Request Templates

Avesha



Thank You, Questions?