

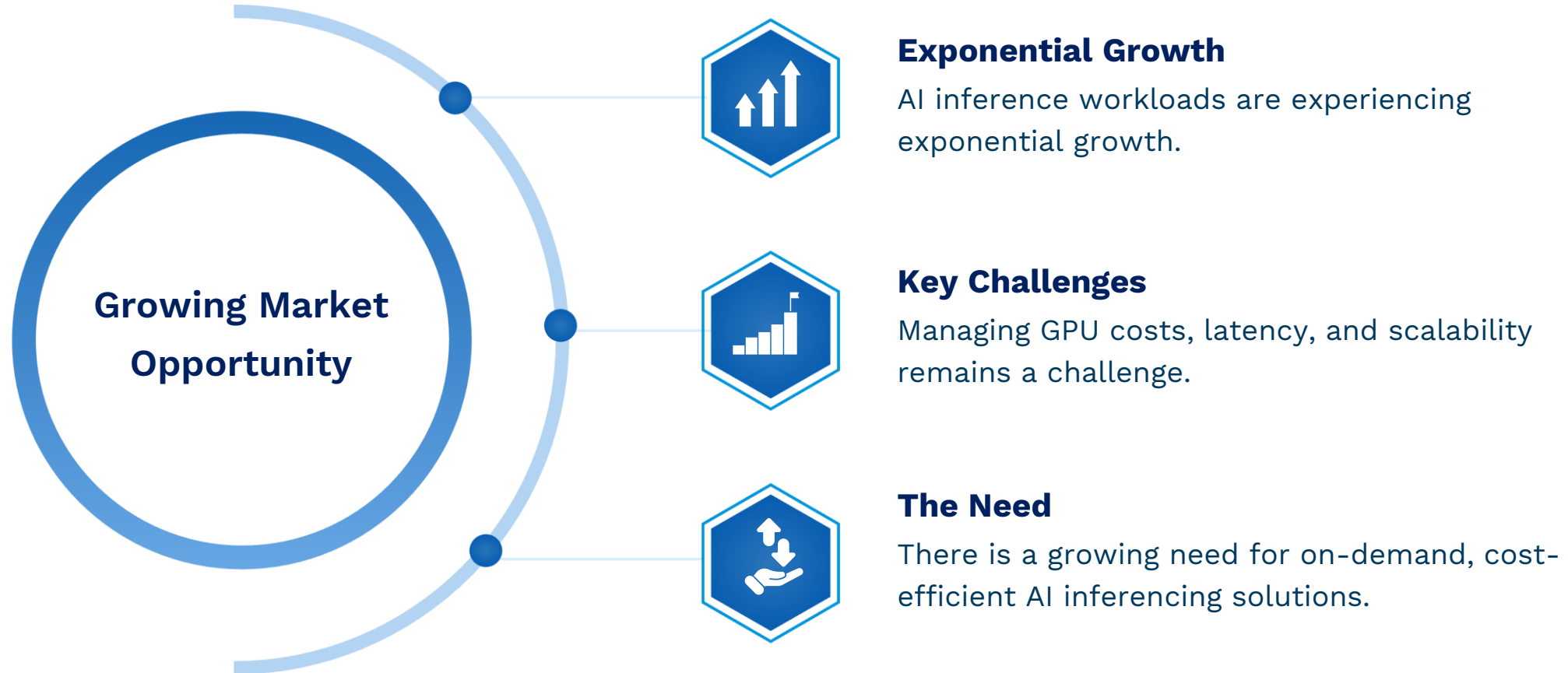


Inferencing-as-a-Service: AI Model Deployment at Scale

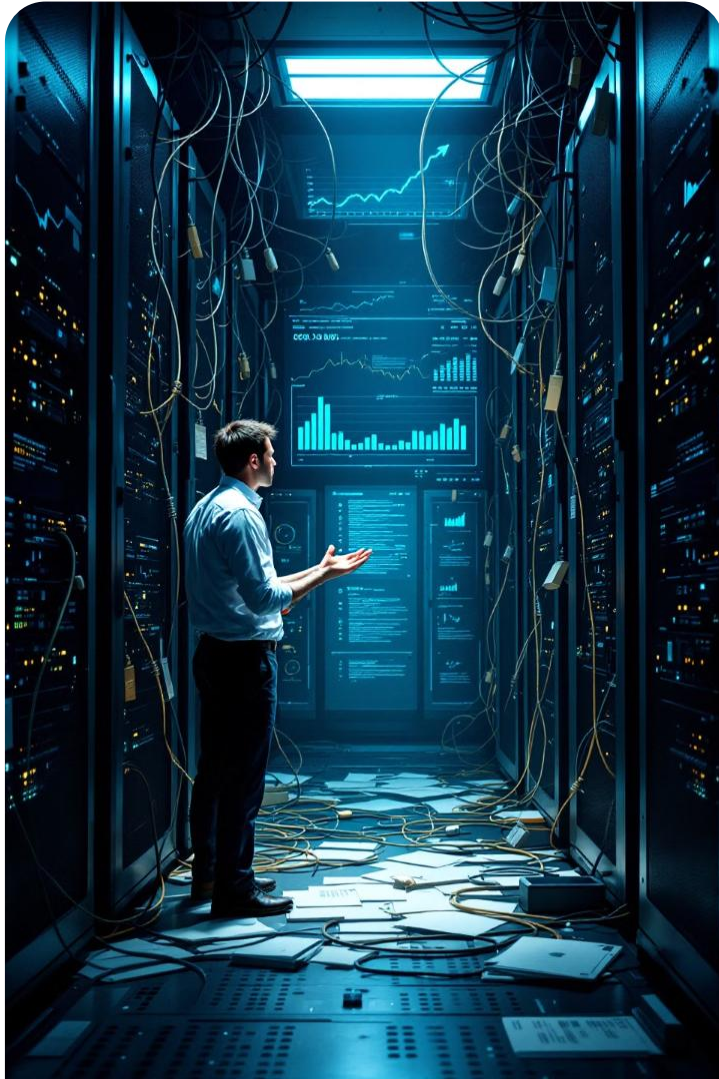
Unlock the true potential of AI with our Inferencing-as-a-Service platform. Deploy AI models at scale with ease and efficiency. Our solution is designed to tackle the growing demands of AI inference workloads.



The Growing Market Opportunity



The High Cost of AI Model Deployment



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



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



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



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



Finance

Fraud detection, high-frequency trading.



Healthcare

Medical imaging, accelerating drug discovery.



Retail

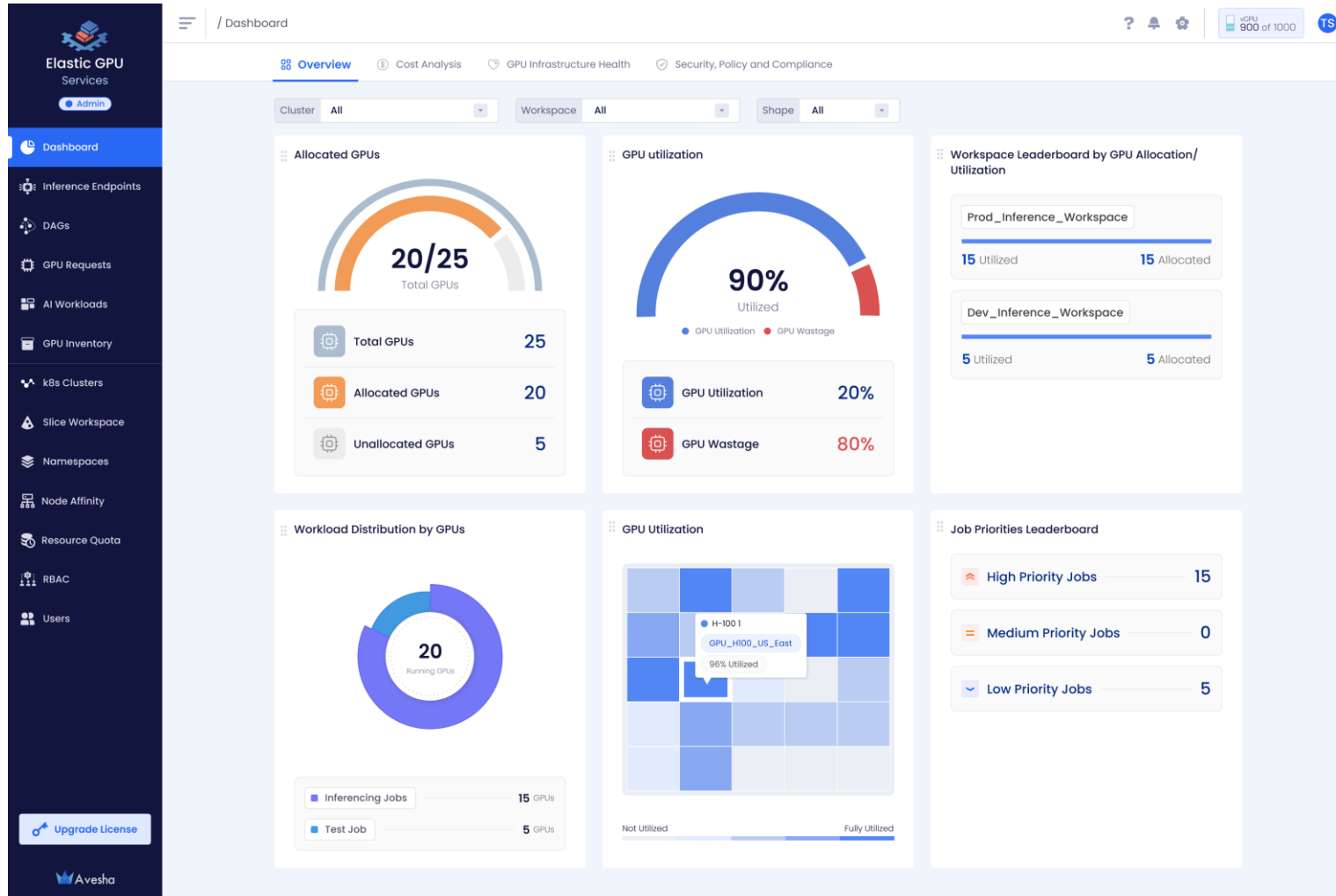
AI-powered customer insights, demand forecasting.



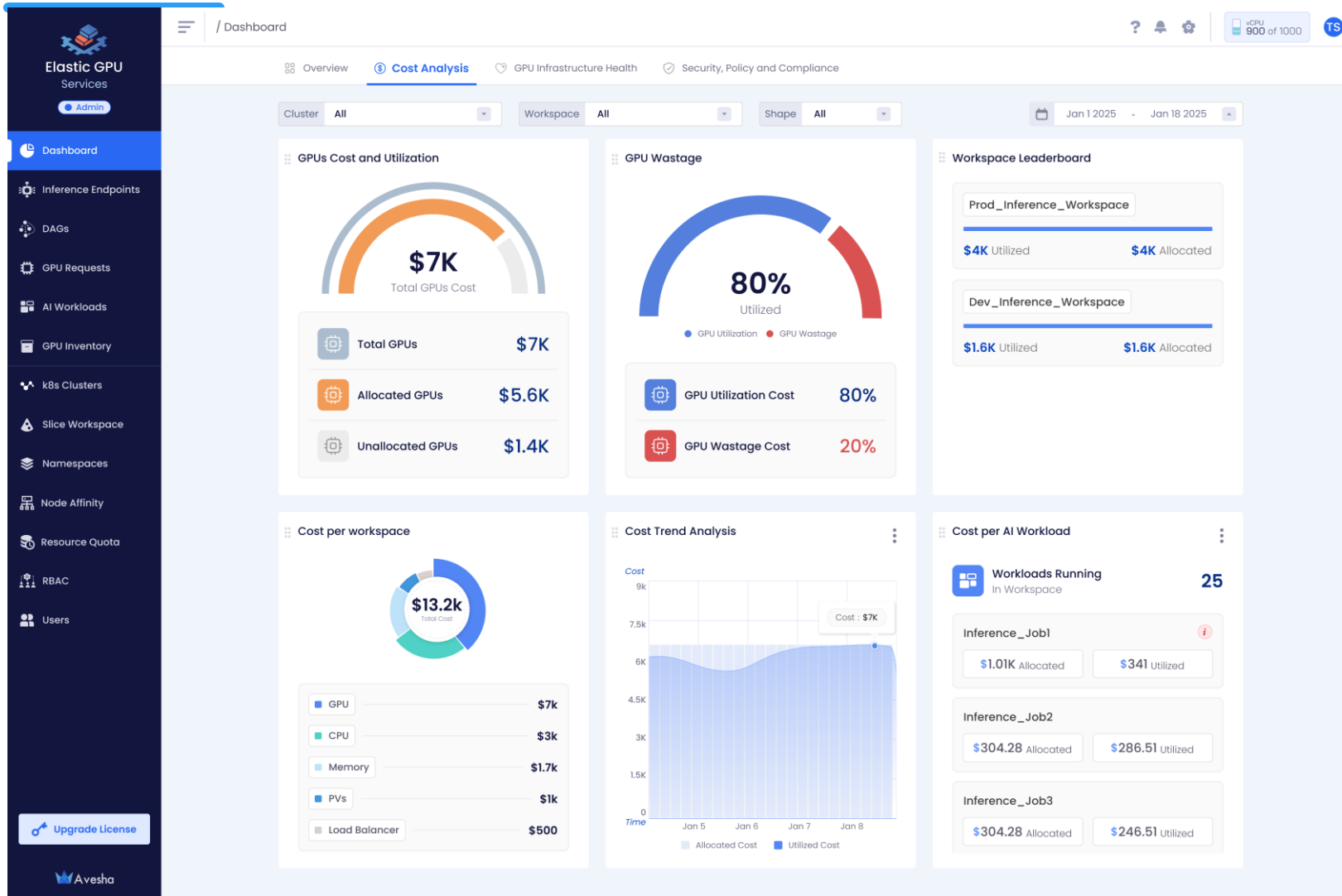
IoT & Manufacturing

Predictive maintenance, anomaly detection.

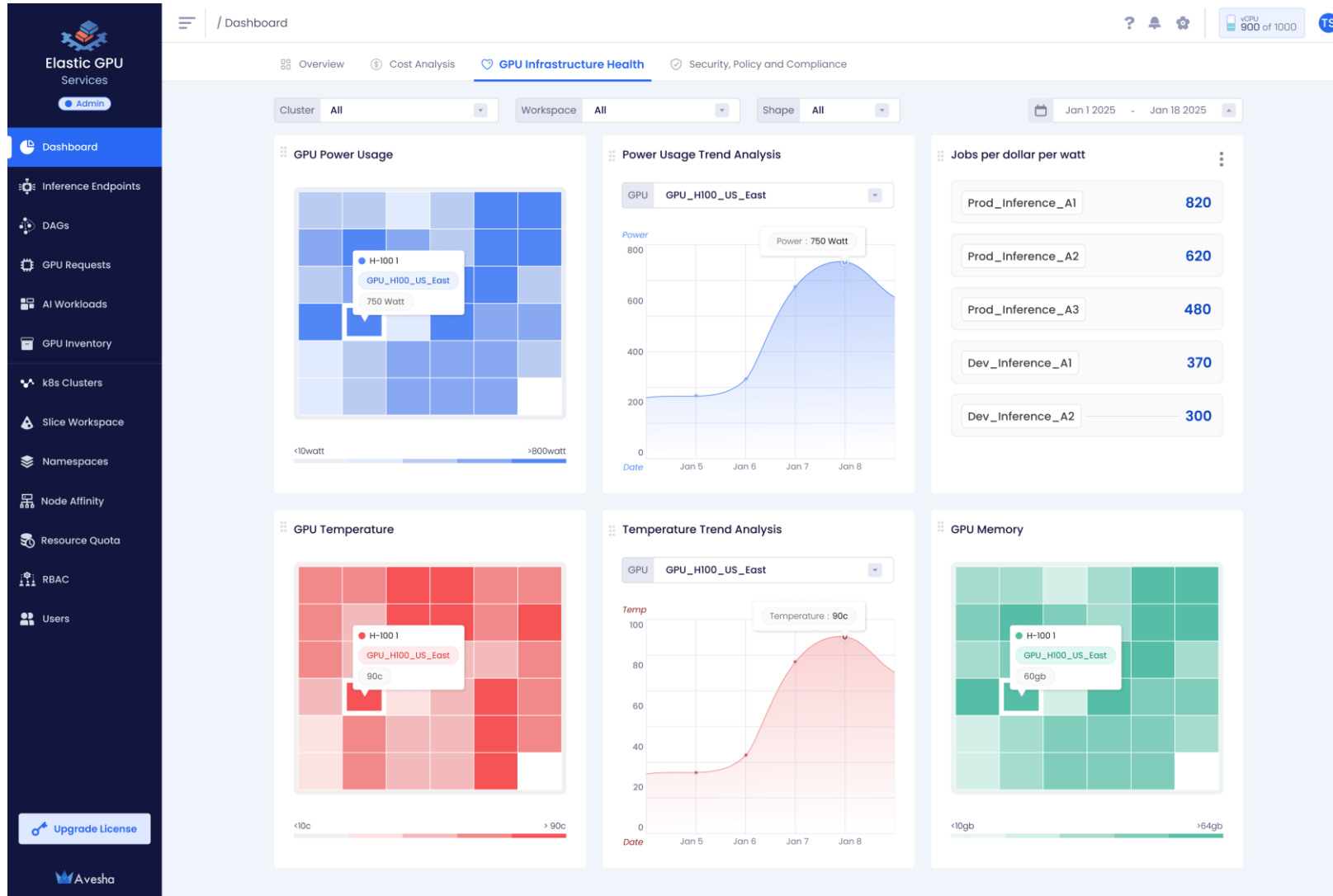
EGS Dashboard - Overview



EGS Dashboard – Cost Analysis



EGS Dashboard – GPU Infrastructure Health



EGS : Model and Infra Parameters



Elastic GPU Services
Admin

- Dashboard
- Inference Endpoints
- GPU Requests
- AI Workloads**
- GPU Inventory
- k8s Clusters
- Slice Workspace
- Namespaces
- Node Affinity
- Resource Quota
- RBAC
- Users

Upgrade License

Avesha

/ AI Workloads / Prod_Inference_Workspace

900 of 1000 TS

← Back to slice list

AI Workloads SLICE NAME: Prod_Inference_Workspace USERS: Admin

/ AI Model Details

Workload Details

Item#	WORKLOAD NAME	PARAMETERS	INFRASTRUCTURE	NAVIGATE
1	small lama2	Params: 170m Batch Size: NVIDIA A100 GPUs: 3 Memory: 2000meg	Pods: 10 GPU Model: NVIDIA A100 GPUs: 3 Memory: 2000meg	Go to Pods Go to GPUs

Model Aggregate Metrics

START DATE	RUN TIME	PROGRESS	TEMPERATURE	POWER	MEMORY	GPU UTILIZATION
24/06/2024 07:28	5hrs 30min	<div style="width: 50%;"></div>	95 °C	124 watt	34%	99%

Showing 1-5 of 5 entries << < 1 > >> view 10 rows per page

EGS : GPU Requests



Elastic GPU Services

Admin

Dashboard

Inference Endpoints

GPU Requests

Priority Queue

Auto GPU Requests

AI Workloads

GPU Inventory

k8s Clusters

Slice Workspace

Namespaces

Node Affinity

Resource Quota

Upgrade License

Users Avesha

/ GPU Requests / Red-slice-1

Back to slice list

GPU Requests per slice

Prod_inference_workspace Admin

Search Name Filter Columns Create GPU Request

REQUEST NAME	CLUSTER	NODE TYPE	REQUESTED BY	PRIORITY	GPU SHAPE	#GPU NODES	#GPUs	ESTIMATED START TIME	RESERVED FOR	STATUS
request-1	worker-cluster-1	p5.48xlarge	admin	= 101	H-100	2	2	06/30/2024:9:00 AM	30mins	Pending
request-2	worker-cluster-1	p5.48xlarge	admin	= 101	H-100	2	2	06/30/2024:9:00 AM	30mins	Pending

Showing 1-5 of 5 entries

Create GPU Request

GPU Request Configuration Requested GPUs

Template Applied: None Change Template

Cluster Selection

Fill the below slice information details to create a slice.

Cluster

aws worker-cluster-1

GPU Configuration

Fill the below slice information details to create a slice.

GPU Request Name

red-request

Node Type GPU Shape

p5.48xlarge Nvidia H-100

Memory (GB) per GPU* GPU per Nodes GPU Nodes

2 - 2 + - 2 +

Priority Configuration

Fill the information on the right to meet the priority

GPU Users Priority*

Admin Medium (101-200)

Priority Number* Reserve For

- 101 + - - - - -

Clear All Save as Template Get Requested GPUs

EGS : Priority Queue



Admin

Dashboard

Inference Endpoints

GPU Requests

Priority Queue

AI Workloads

GPU Inventory

k8s Clusters

Slice Workspace

Namespaces

Node Affinity

Resource Quota

RBAC

Upgrade License

Avesha

/ GPU Requests / Priority Queue

900 of 1000

Priority Queue

Worker-cluster-1 Worker-cluster-2

n1-highcpu-2 1 GPR(s) queued
Next GPR will be processed at 12/19/2024 3:40 PM

REQUEST NAME	SLICE	REQUESTED BY	PRIORITY	GPU SHAPE	#GPU NODES	#GPUs	ESTIMATED START TIME	RESERVED FOR	ACTIONS
request-1	prod_inference...	admin	101	tesla T-4	2	2	06/30/2024:9:00 AM	30mins	

Showing 1-1 of 1 entries

n1-highcpu-24 2 GPR(s) queued
Next GPR will be processed at 12/19/2024 3:40 PM

EGS : GPU Inventory

Elastic GPU Services

Admin

- Dashboard
- Inference Endpoints
- GPU Requests
- AI Workloads
- GPU Inventory**
- k8s Clusters
- Slice Workspace
- Namespaces
- Node Affinity
- Resource Quota
- RBAC
- Users

Upgrade License

Avesha

/ Inventory

VCPU 900 of 1000 TS

Inventory

Search from keyword

Filter Columns

CLOUD	CLUSTER	SLICE	GPU SHAPE	GPUs	POOL NAME	NODE NAME	# MEMORY	ALLOCATION STATUS	
AWS	Cluster 1	red-slice-1	H-100	8	Node Pool	p5.48xlarge	8	Allocated	>
GCP	Cluster 2	red-slice-2	H-100	8	Node Pool	p5.48xlarge	8	Cordoned	>
EKS	Cluster 2	red-slice-3	H-100	8	Node Pool	p5.48xlarge	8	Free	>
Azure	Cluster 2	red-slice-4	H-100	8	Node Pool	p5.48xlarge	8	Failed	>

Showing 1-4 of 4 entries

« < 1 > »

view 10 rows per page

EGS : Inference EndPoints



Admin

Dashboard

Inference Endpoints

GPU Requests

AI Workloads

GPU Inventory

k8s Clusters

Slice Workspaces

Namespaces

Node Affinity

Resource Quota

RBAC

Users

Upgrade License

Avesha

/ Inference Endpoints / Prod_Inference_Workspace

900 of 1000 TS

Back to workspace list

Inference Endpoints

Prod_Inference_Workspace

Create Inference Endpoint

DEPLOYMENT NAME	MODEL NAME	STATUS	ENDPOINT
sci-kit-hf-1m3	sklearn	Running	https://sklearn-iris.example.com/v1/model-iris:predict
GPT4 Model	GPT 4 -o preview	Running	https://gpt4-chat.gpt4.example.com/v1/model/gpt-o-preview...
GPT3 Model	GPT 3 -o preview	Image pullback error	https://sklearn-iris.ramakant.example.com/v1/model-iris:predict
GPT4 mini Model	GPT 4.mini -o preview	Pending	https://sklearn-iris.pranila.example.com/v1/model-iris:predict

Showing 1-4 of 4 entries

view 10 rows per page

EGS : Automatic GPU Request Templates



Elastic GPU Services

Admin

Dashboard

Inference Endpoints

GPU Requests

Auto GPU Requests

AI Workloads

GPU Inventory

k8s Clusters

Slice Workspace

Namespaces

Node Affinity

Resource Quota

RBAC

Upgrade License

Avesha

/ Auto GPU Requests / GPR Templates

vCPU 900 of 1000 TS

Auto GPU Requests

Slice Workspaces **GPR Templates**

+ Create GPR Template

Search from keyword

Filter Columns

NAME	CLUSTER	NODE TYPE	GPU SHAPE	MEMORY PER GPU	GPUs	GPU NODES	PRIORITY	STATUS	
red-template	aws worker-cluster-1	p5e.48xlarge	NVIDIA H-100	2	2	2	= 101	Ready	X ⋮
green-template	aws worker-cluster-1	p5e.48xlarge	NVIDIA H-100	2	2	2	= 101	Ready	X ⋮
blue-template	aws worker-cluster-1	p5e.48xlarge	NVIDIA H-100	2	2	2	= 101	Ready	X ⋮
yellow-template	aws worker-cluster-1	p5e.48xlarge	NVIDIA H-100	2	2	2	= 101	Ready	X ⋮

Showing 1-5 of 5 entries

view 10 rows per page



Avesha

**Thank
You,
Questions?**
