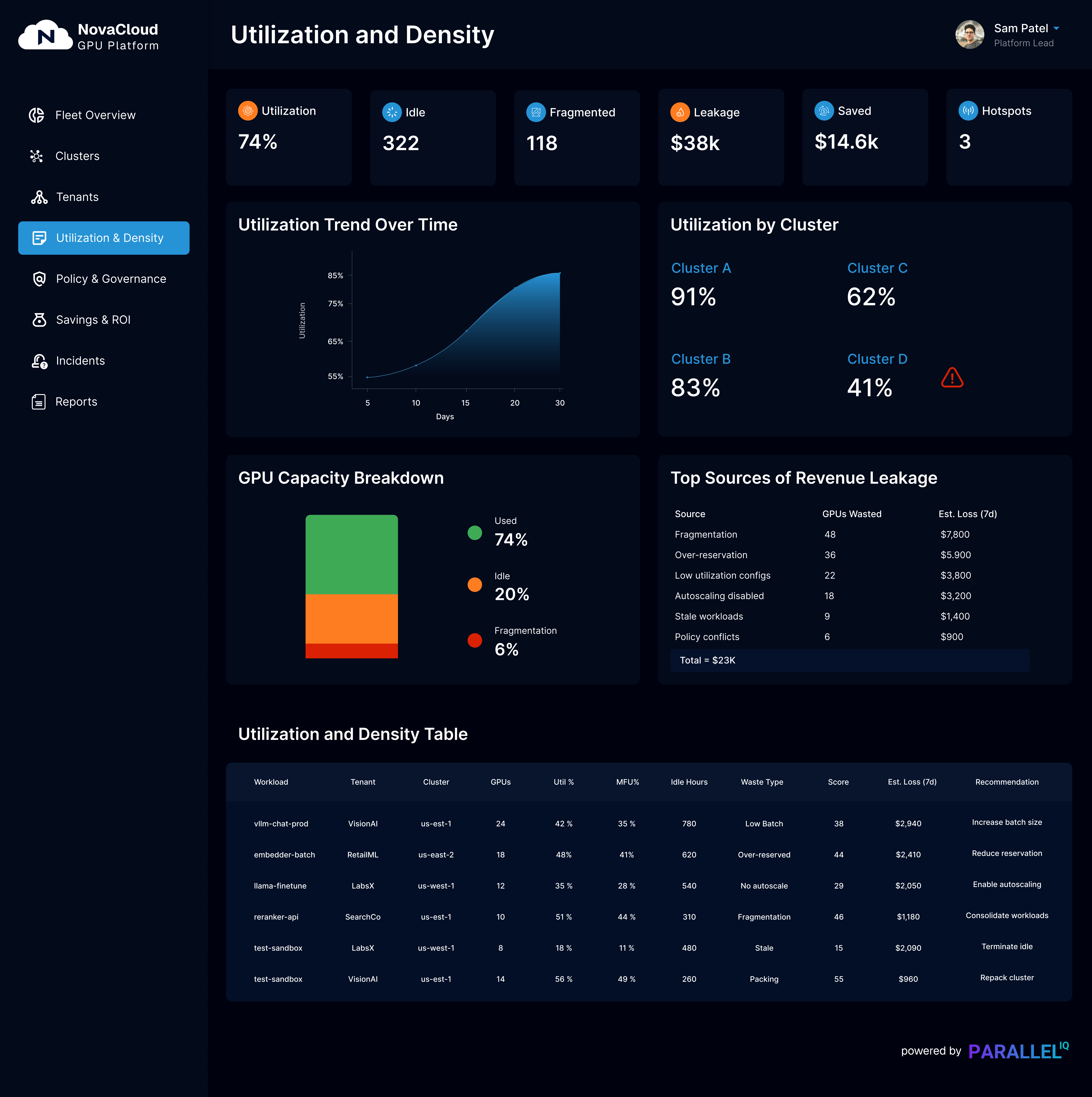

NovaCloud
GPU Platform

Utilization and Density

Sam Patel
Platform Lead

- Fleet Overview
- Clusters
- Tenants
- Utilization & Density
- Policy & Governance
- Savings & ROI
- Incidents
- Reports

Utilization	Idle	Fragmented	Leakage	Saved	Hotspots
74%	322	118	$38k	$14.6k	3

Utilization Trend Over Time

Utilization by Cluster

Cluster A 91%	Cluster C 62%
Cluster B 83%	Cluster D 41% ⚠

GPU Capacity Breakdown

● Used 74%

● Idle 20%

● Fragmentation 6%

Top Sources of Revenue Leakage

Source	GPUs Wasted	Est. Loss (7d)
Fragmentation	48	$7,800
Over-reservation	36	$5.900
Low utilization configs	22	$3,800
Autoscaling disabled	18	$3,200
Stale workloads	9	$1,400
Policy conflicts	6	$900
Total = $23K		

Utilization and Density Table

Workload	Tenant	Cluster	GPUs	Util %	MFU%	Idle Hours	Waste Type	Score	Est. Loss (7d)	Recommendation
vllm-chat-prod	VisionAI	us-est-1	24	42 %	35 %	780	Low Batch	38	$2,940	Increase batch size
embedder-batch	RetailML	us-east-2	18	48%	41%	620	Over-reserved	44	$2,410	Reduce reservation
llama-finetune	LabsX	us-west-1	12	35 %	28 %	540	No autoscale	29	$2,050	Enable autoscaling
reranker-api	SearchCo	us-est-1	10	51 %	44 %	310	Fragmentation	46	$1,180	Consolidate workloads
test-sandbox	LabsX	us-west-1	8	18 %	11 %	480	Stale	15	$2,090	Terminate idle
test-sandbox	VisionAI	us-est-1	14	56 %	49 %	260	Packing	55	$960	Repack cluster

powered by PARALLEL IQ