



NCAR
OPERATED BY UCAR

ISS 2026 • UCAR

April 2026

VGAC: Predictive Queue Intelligence for GPU Cluster Observability

Follow along:

demo.vgac.cloud

Andrew Espira

Saint Peter's University, Department of Data Science



Scan

GPU queues are black boxes.



Researchers submit jobs — then wait, guess, and repeat.

73%

of ML engineers report
wasting time on queue uncertainty

\$2–4

per GPU-hour burned on
idle or mis-timed compute

0

production tools that tell you
why your job is stuck

The gap between cluster state and actionable guidance is the problem VGAC closes.

Today's Reality: Refresh. Check. Guess. Repeat.



```
$ queue -u researcher
```

```
JOBID PARTITION NAME STATE
```

```
294853 gpu-batch train-v3 PENDING
```

```
$ # Why is it pending?
```

```
# Nobody knows. Try again later.
```

What's missing:

- When will it start?
- Why is it waiting?
- What should I change?
- Is now a good time?
- Which partition is free?

01

WHY

"Why is my job stuck?"

See the real bottleneck — queue pressure, memory saturation, placement constraints. An answer, not another panel.

Queue breakdown · GPU allocation

02

WHEN

"When will it actually run?"

Calibrated wait-time predictions at submit time, updated as queue state changes. Plan your experiment, not your inbox.

AUC 0.756 · Real-time updates

03

HOW

"How do I make it faster?"

Data-driven recommendations — fewer GPUs, different partition, off-peak window. Grounded in what the cluster is actually doing.

Placement · Timing · Cost

Architecture: Three Integrated Layers.



① Collection Plane

- kube-state-metrics
- dcgm-exporter (GPU telemetry)
- Job lifecycle events
- Submit-time queue snapshot

② Prediction Service

- Scikit-learn classifiers
- Isotonic calibration (ECE-first)
- FastAPI · sub-10 ms latency
- ClickHouse time-series store

③ Policy Engine

- Risk bands Very Low → Very High
- Kubernetes admission webhook
- Slurm job_submit hook
- Grafana + REST API surfaces

AWS EKS

ParallelCluster

DCGM

Prometheus

Redis

Grafana

FastAPI

ClickHouse

Key Finding 1: Calibration Matters More Than Accuracy.



What is ECE?

Expected Calibration Error measures whether stated probabilities match observed frequencies.

"When VGAC says 70% chance of delay, jobs actually wait ~70% of the time."

High AUROC + poor ECE → Misleads threshold decisions

Calibrated ECE ≤ 0.05 → Trust advisory actions

0.756

AUC-ROC

EKS logistic regression

0.077

ECE

within advisory threshold <0.10

$<10\text{ms}$

Latency

sub-10ms prediction serving

582

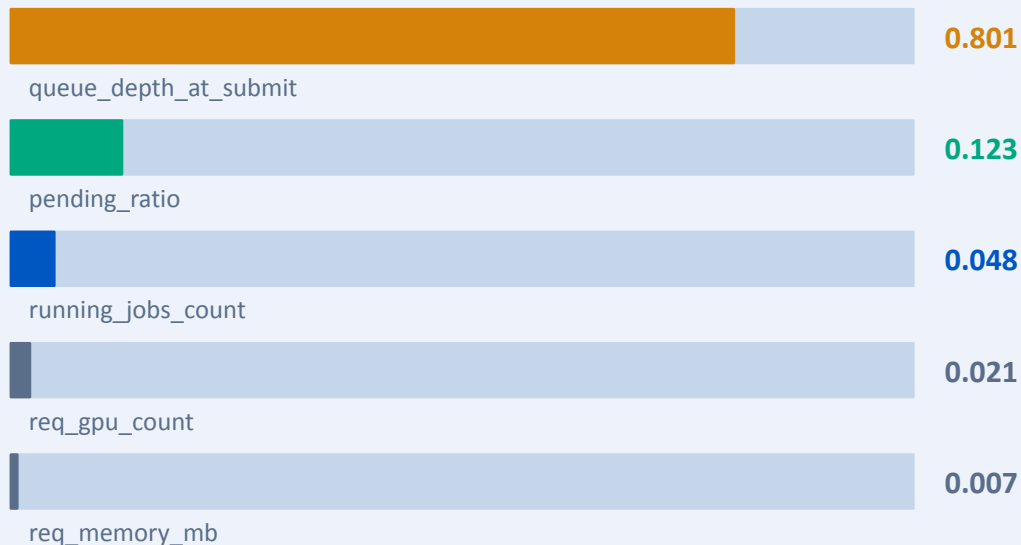
Jobs

production EKS lifecycle records

Key Finding 2: Queue Depth Dominates — Minimal Features Work.



Feature Importance



What this means:

- Queue depth alone captures 80% of predictive signal
- No DCGM or GPU-specific metrics required for MVP
- Deploy without deep infrastructure changes
- Add telemetry incrementally as cluster matures

Feature-first: instrument before you tune.

Results: Validation on Amazon EKS Cluster — 582 Jobs.



0.756

AUC-ROC

discrimination quality

0.077

ECE

≤0.10 advisory threshold

120s

SLO Threshold

P90 long-wait definition

48.5%

Positive Rate

balanced classes in EKS

Model comparison across environments:

Model	Dataset	AUROC	ECE	Tier
Logistic Regression	EKS (2 features)	0.806	0.054	Tier 2
Gradient Boosting	Slurm (17 features)	0.985	0.006	Tier 4 ✓
LR + Isotonic Cal.	EKS (calibrated)	0.756	0.077	Advisory

EKS → Slurm (cross-scheduler)

-8%

AUROC degradation

ranking ability partially transfers

22×

ECE increase

probabilities become meaningless

0.791

EKS→Slurm ECE

near anti-calibrated predictions

Implication:

Per-cluster recalibration is mandatory

Not optional — a universal calibrator is a fiction across heterogeneous GPU infrastructure.

AUROC-only evaluation is dangerous

A model that 'works' by AUROC produces probabilities that mislead every downstream decision.

PSI drift detection provides the safety net

Population Stability Index catches distribution shift before miscalibrated scores reach users.

Kubernetes

- 1 Validating admission webhook intercepts Pod/CREATE
- 2 Feature extractor queries API for current queue state
- 3 Calibrated risk score → annotation on pod spec
- 4 High-risk jobs receive placement alternatives
- 5 Tier 4: automated queue reassignment with override

Slurm / HPC

- 1 job_submit plugin intercepts sbatch / salloc
- 2 Queries DCGM exporter for GPU telemetry
- 3 Risk score computed in <10 ms — no overhead
- 4 Job comment annotated with risk band + suggestion
- 5 Tier 4: priority adjustment or submission hold

Staged rollout: Shadow → Advisory → Policy · Any SLO breach rolls back automatically

I Calibration over accuracy

ECE is the deployment gate, not AUROC. A probability users trust drives more operational value than a model that ranks marginally better but misleads decisions.

III Meet users where they are

Predictions surface via REST endpoints, K8s annotations, Slurm job comments, and Grafana — no new workflows or dashboards required.

II Minimal features, maximum signal

queue_depth_at_submit captures 80% of importance.
Deploy fast with two features; add DCGM telemetry as your cluster matures and data accumulates.

IV Safety valves at every tier

Shadow → advisory → policy gating with automatic rollback on any SLO breach. No automated action fires unless ECE remains below threshold for 30+ days.

Roadmap: Prototype to Production Reliability.



Conformal prediction intervals

Distribution-free coverage guarantees — 'your job starts in 45–120 s with 90% coverage'

SLO-based calibration drift monitoring

PSI drift detection triggers sliding-window recalibration before users notice degradation

Cross-cluster federated calibration

Share calibration statistics across institutions without sharing raw job data

DCGM longitudinal features

Thermal trends, ECC accumulation, power efficiency decay as long-horizon delay predictors

Live A/B advisory impact evaluation

Measure queue time reduction and user compliance rates vs a control group

Multi-scheduler generalization

Azure CycleCloud, PBS, LSF — unified schema designed, connectors in progress

Conclusion.



VGAC

Stop waiting. Start computing.

- ✓ Calibration quality determines trust — ECE is the deployment gate
- ✓ Queue depth captures 80% of signal — start simple, grow telemetry
- ✓ Predictions reach users through the tools they already use
- ✓ Per-cluster recalibration is mandatory in heterogeneous GPU infra

Scan to connect



Live Demo

demo.vgac.cloud



Connect

linkedin.com/in/andrew-espir-a

aespira@saintpeters.edu

Saint Peter's University · Department of Data Science

ISS 2026 · UCAR · Improving Scientific Software Conference