UMassAmherst | Manning College of Information & Computer Sciences
Center for Data Science & Artificial Intelligence

UNITY
W: https://unity.rc.umass.edu
E: hpc@umass.edu

Unity GPU Efficiency Analytics Suite

Adobe

Ardavan Bozorgi, Ayush Ravi Chandran,
Tan Le, Nitya Karthik Aryasomajula,
and Christopher Odoom

Summary

- Developed a **data-driven GPU analytics framework** to detect usage patterns, inefficiencies, and improvement opportunities.
- Empowers Unity to deliver **actionable recommendations** to optimize GPU allocation, cut costs, and boost performance in HPC environments.
- Unity intends on utilizing this tool to enhance their outreach process.

Project Goals

- GPUs are **scarce and expensive**, yet often **underutilized** in HPC environments.
- A SLURM jobs database is continuously updated with job submissions, GPU memory usage, and requested GPU types.
- Goal: How can we **quantify underutilization**, and how can we **automate the identification of users** and **generate user reports**?

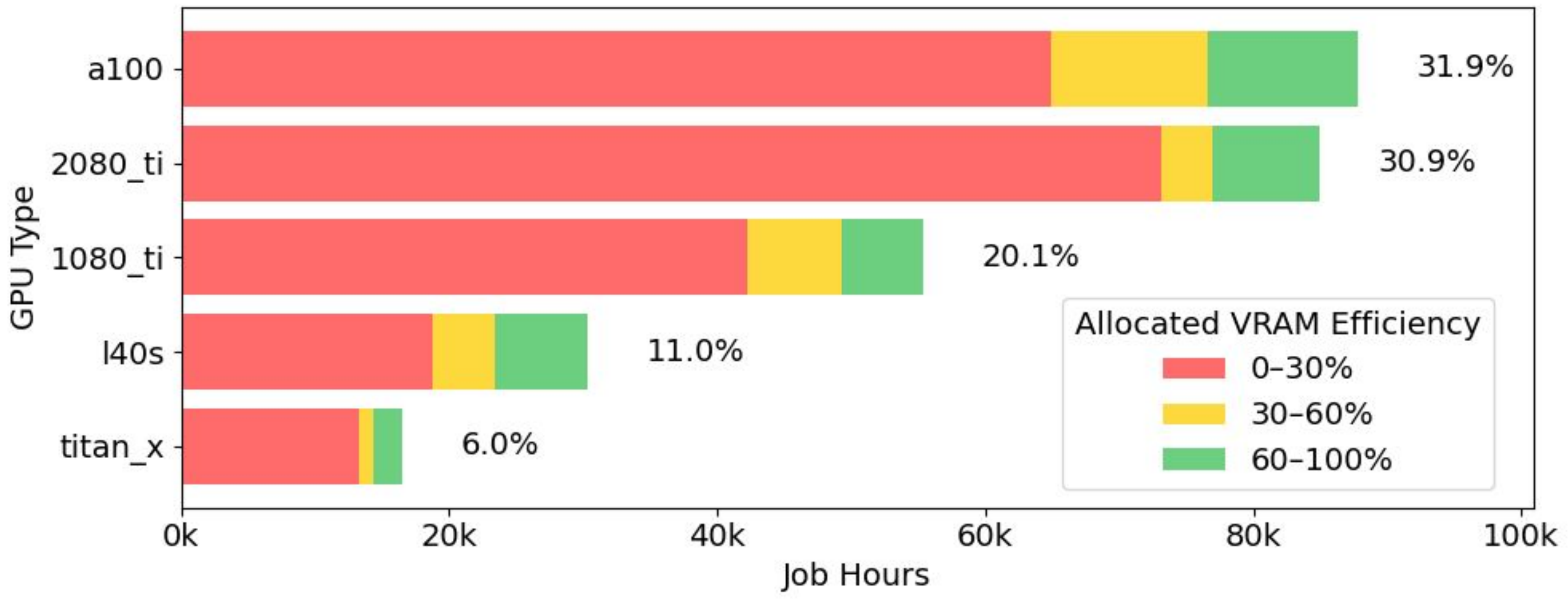


Figure 1: Total Job Durations per GPU Type

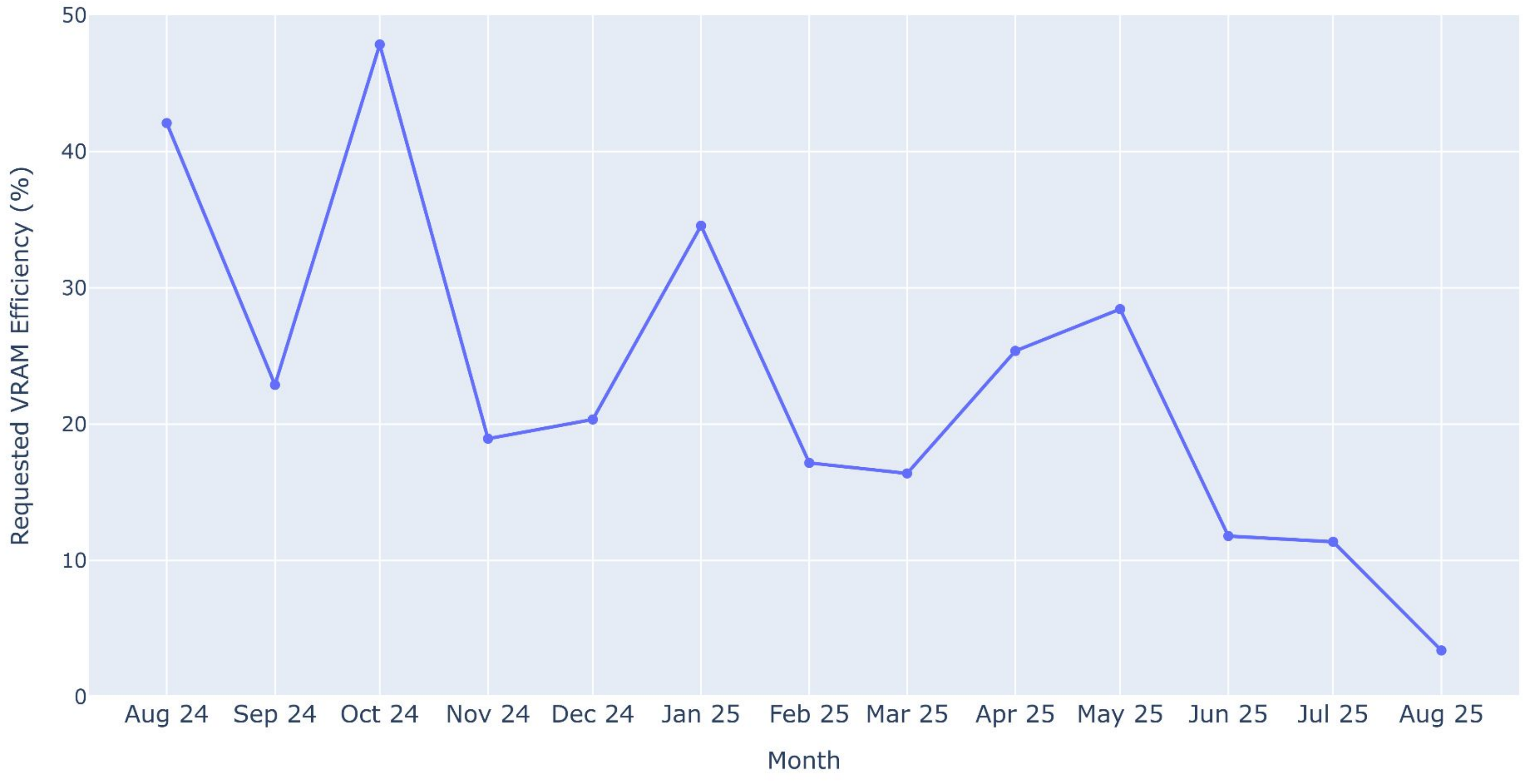
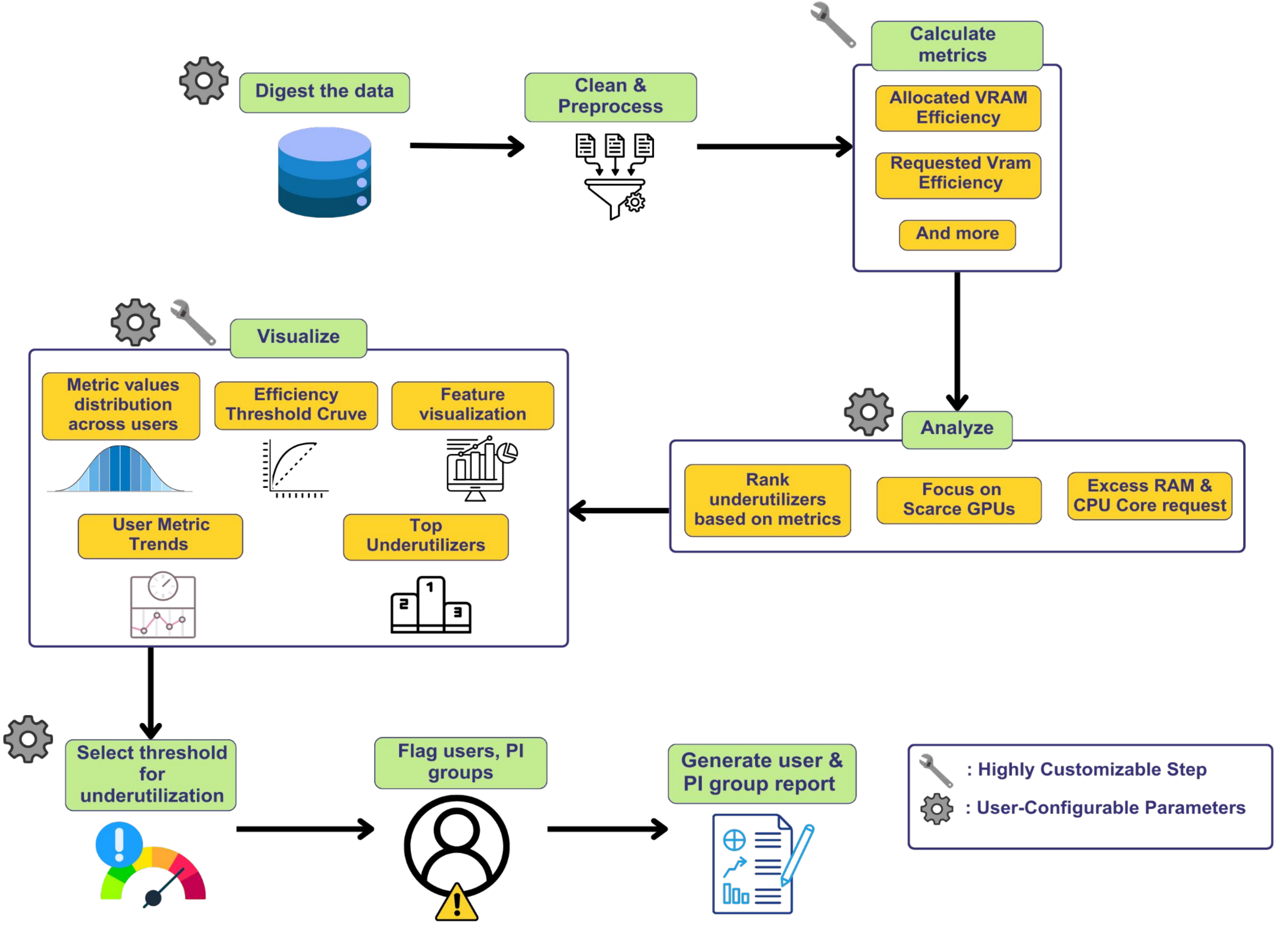


Figure 2: Average Requested VRAM Efficiency over time (All users)

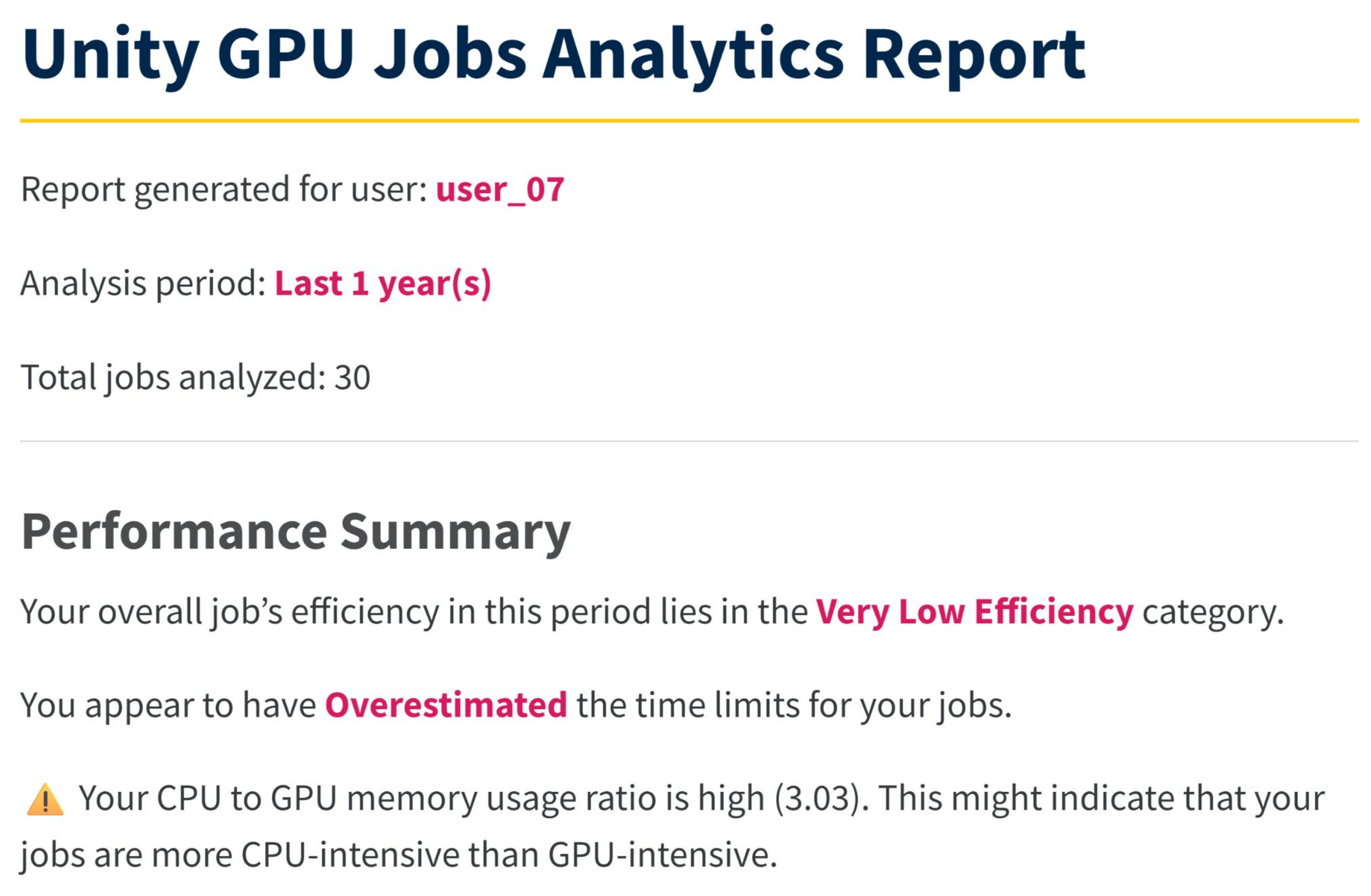
Methodology


Digest the data
Clean & Preprocess
Calculate metrics
Allocated VRAM Efficiency
Requested Vram Efficiency
And more
Analyze
Rank underutilizers based on metrics
Focus on Scarce GPUs
Excess RAM & CPU Core request
Visualize
Metric values distribution across users
Efficiency Threshold Cruve
Feature visualization
User Metric Trends
Top Underutilizers
1
2
3
Select threshold for underutilization
Flag users, PI groups
Generate user & PI group report
: Highly Customizable Step
: User-Configurable Parameters

Users may request a specific amount of VRAM, but SLURM can allocate more if no GPU with the exact capacity is available. We therefore define two metrics:

- **Allocated VRAM Efficiency**: Ratio of VRAM used to VRAM allocated by the scheduler.
- **Requested VRAM Efficiency**: Ratio of VRAM used to VRAM requested by the user.

Deliverables



Unity GPU Jobs Analytics Report

Report generated for user: **user_07**

Analysis period: **Last 1 year(s)**

Total jobs analyzed: 30

Performance Summary

Your overall job's efficiency in this period lies in the **Very Low Efficiency** category.

You appear to have **Overestimated** the time limits for your jobs.

⚠ Your CPU to GPU memory usage ratio is high (3.03). This might indicate that your jobs are more CPU-intensive than GPU-intensive.

- User Statistics
- Visualizations
- Performance Summary
- Personalized Recommendations
- Documentation

Key Insights

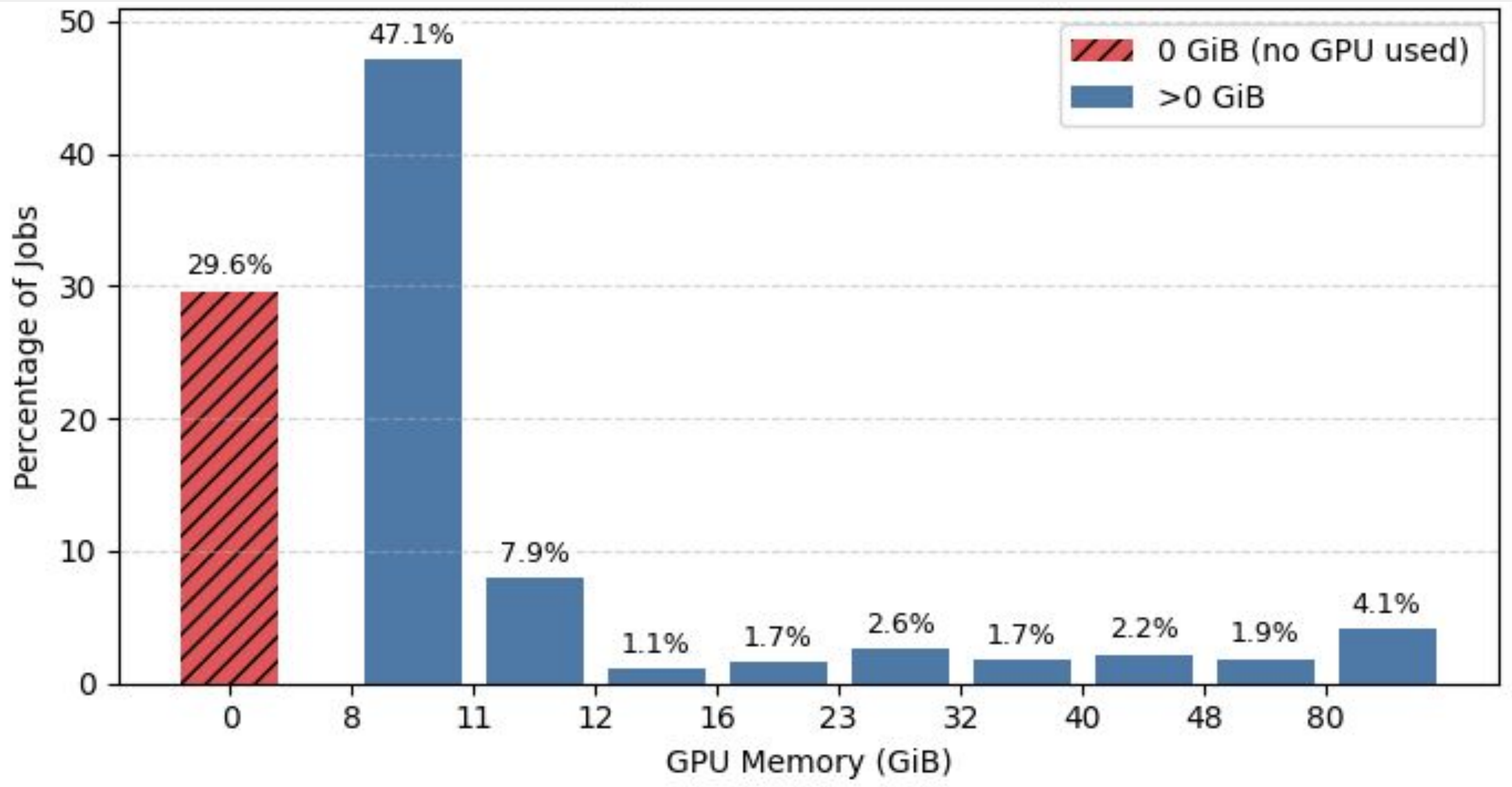


Figure 2: Histogram of GPU VRAM Usage

Only 6% of jobs request more than 48 GiB of VRAM, which necessitates the use of an A100 GPU. However, these A100 GPUs account for 34% of total GPU time, indicating that scarce resources are frequently consumed by jobs that may not strictly require them (see Figure 1).

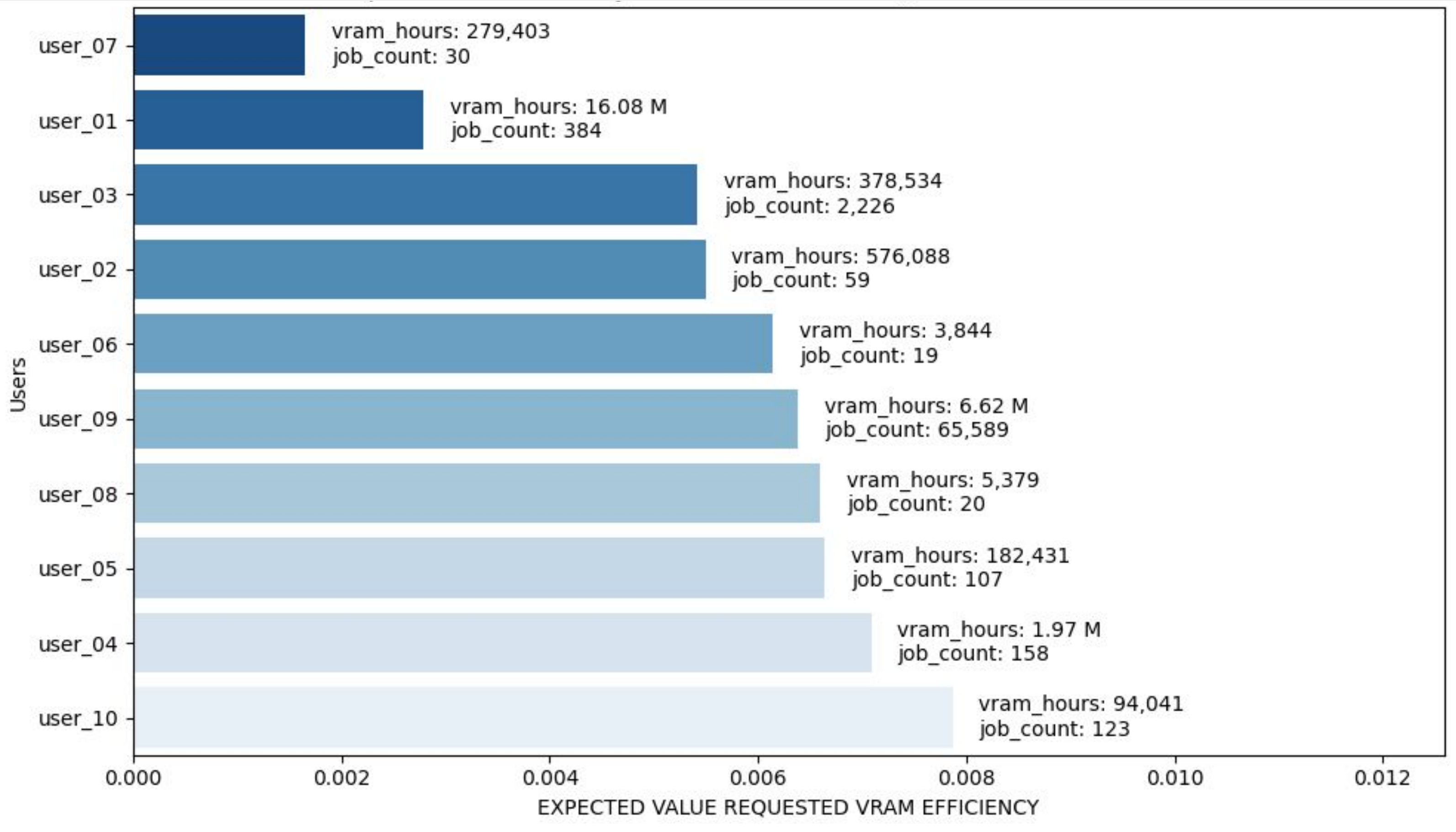


Figure 3: Inefficient Users Ranked by Requested VRAM Efficiency

We detect users with extremely poor VRAM efficiency- like those shown consuming millions of GPU-hours with near-zero utilization.

Github Repo: https://github.com/UnityHPC/ds4cg-job-analytics

Future Directions

- Future work entails adding clustering algorithms (K-Means, DBSCAN)
- Live tracking and automated report generation
- Develop metrics for Multi-GPU and Array Jobs

Partners: Unity is a multi-institution high-performance and AI computing platform leveraging the unique and collaborative resources at the MGHPCC. Unity is under active development and supports primarily research activities. Partnering institutions currently include UMass Amherst, University of Rhode Island, UMass Dartmouth, UMass Lowell, UMass Boston, and Mount Holyoke College.