Monitoring GPU Utilization

Le Mai Weakley, Scott Michael, Abhinav Thota, Ben Fulton, Laura Huber, and Matt Kusz

Research Technologies, Indiana University

UNIVERSITY INFORMATION TECHNOLOGY SERVICES

Research Technologies is a division of University Information Technology Services and a center in the Pervasive Technology Institute at Indiana University.

Motivation

Research Technologies (RT) division at Indiana University (IU) operates the HPC systems for IU researchers

Systems operated by RT have had some substantial component that is equipped with accelerators (specifically NVIDIA GPUs)

Gamut of IU user behavior is wide - not necessarily large scale users

Important to learn how and to what extent these resources are being utilized

Desire to understand what users are capable of instead of what the machine is capable of

What will be covered in this talk

We will cover the data collection framework, some analysis of the data, a short overview of other works and some plans for future work.

Caveat Emptor: this is the approach we have applied, certainly other sites have different tricks up their sleeves and we're here for it and would be interested in learning from the community as well

Story time!

Big Red II, a Cray XE6/XK7 hybrid system (2013), first large scale GPU-based system

Tracking of GPU usage on the Big Red II via Cray-provided RUR data integrated with the scheduler output from PBS/Torque

The RUR system was not particularly well documented

Data were largely used to determine whether or not a job requesting a GPU used it

In 2020 RT deploys a set of Slurm-scheduled Lenovo and HPE Apollo nodes with a combined total of 16 P100 GPUs and 104 V100 GPUs on Carbonate

Required a new data collection framework; inspiration found in CSCS presentation on resource tracking on their Slurm-scheduled Cray system

Current Implementation

Python script using NVML is run as part of Slurm epilog for every job

Process records are pulled from GPU ring buffer (buffer will hold maximum of 4,000 records) and removed from the card

JSON records are stored in the Slurm AdminComment field

Slurm data are copied to historical database on daily basis

GPU tracking framework implemented on Big Red 200 (HPE Cray EX) at end of 2022

Code: https://github.com/IUResearchApplications/gpu_accouting

Other works

We conducted a survey to learn about what tracking tools other sites were using and how the data collected was being used

Various sites are tracking utilization in some manner (nvidia-smi, NVML, nvtop, DCGM)

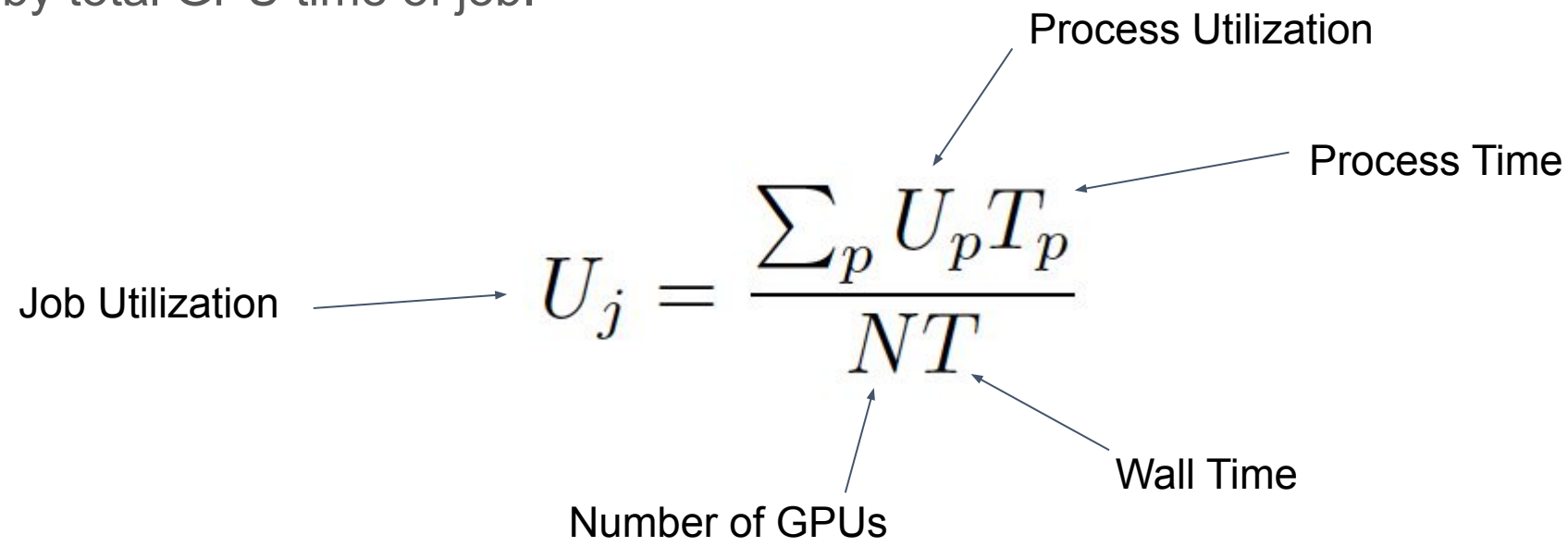
Standard practice to measure and bill for GPU time consumed by jobs but not a lot of response on actual and extent of utilization

Let’s look at some Data!

Data

Analyzed Carbonate data from July 2020 - March 2023 and BR200 data from December 2022 through March of 2023 (295,470 Jobs)

Defined GPU utilization as the sum across all processes of time-weighted utilization divided by total GPU time of job.

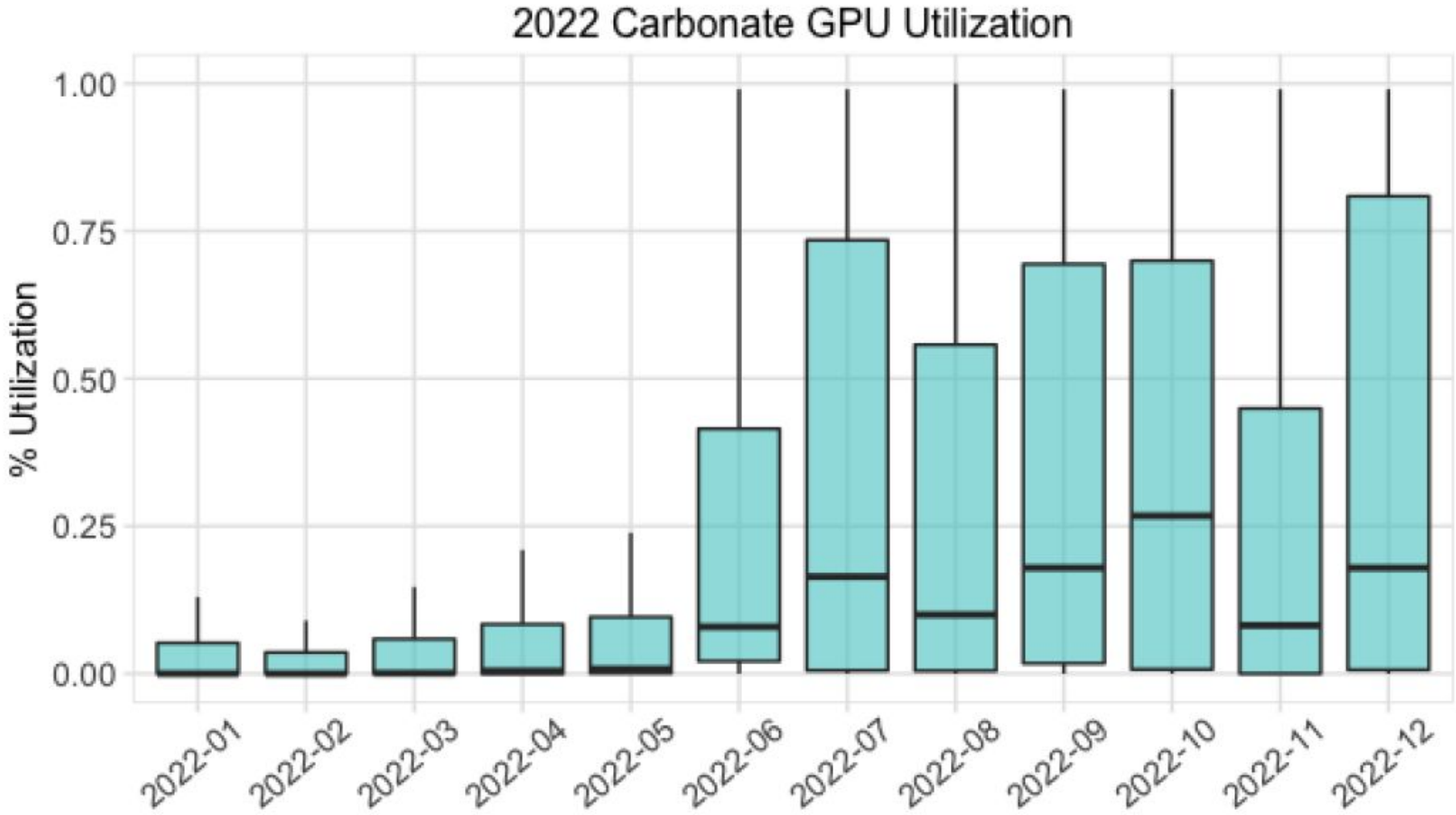


Observations

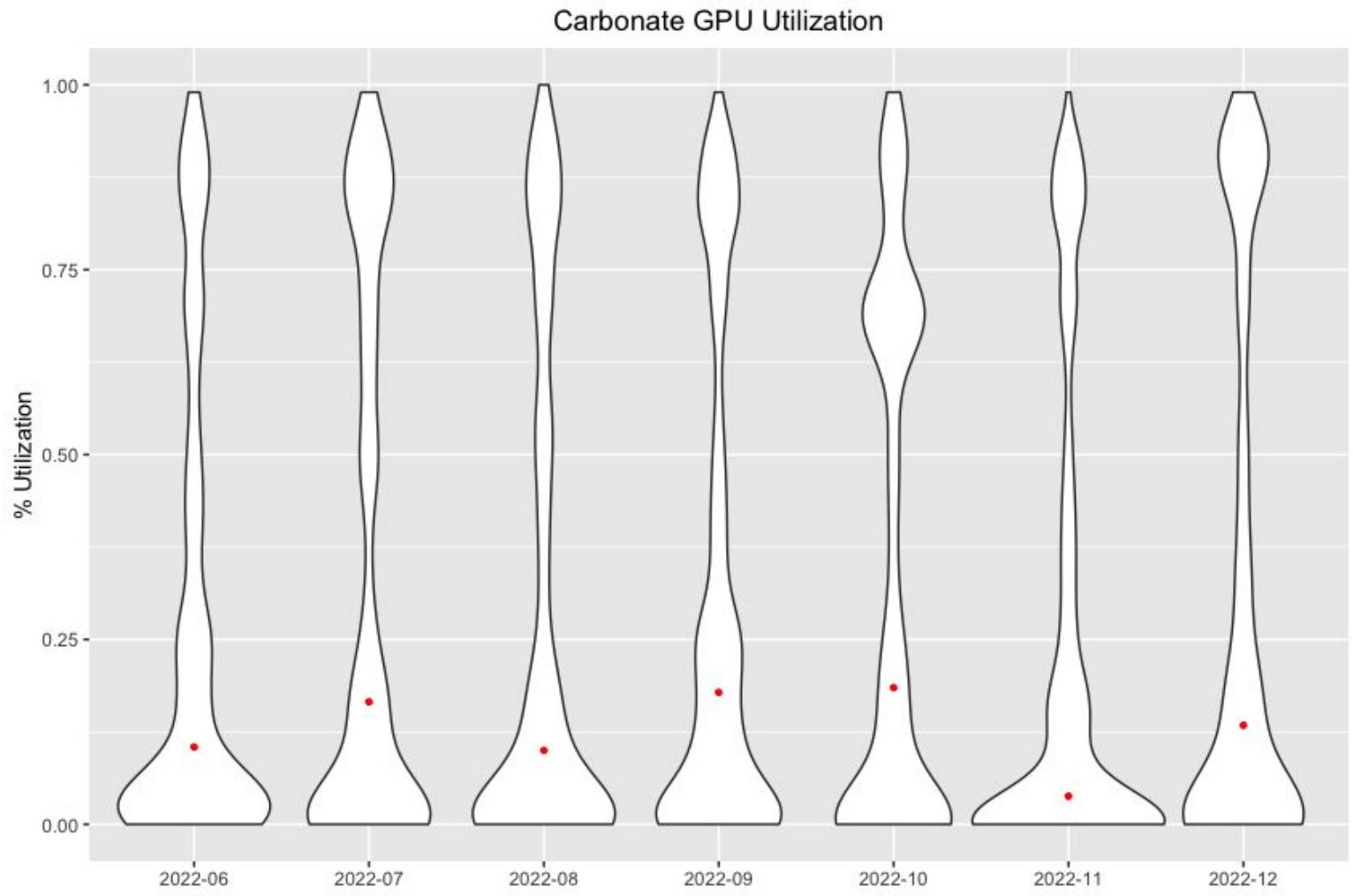
We noted:

- the number of users requesting GPUs is increasing
- there's evidence that suggests that users are optimizing their GPU workflows over time
- many jobs are requesting 1-4 cards and these jobs see higher utilization than jobs requesting more than 4 cards
- prevalence and popularity of AI workflows

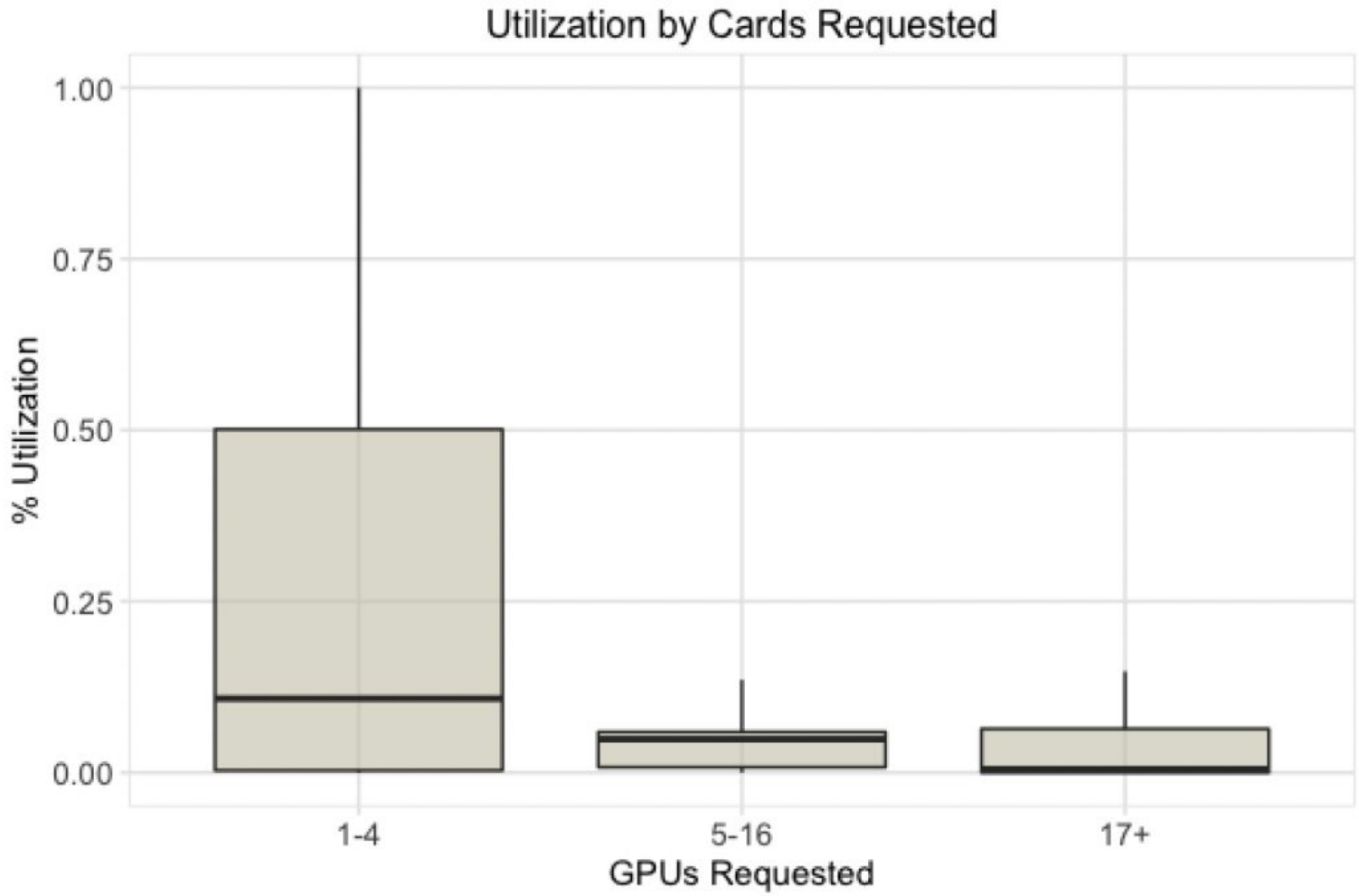
I REALLY NEED TO SEE THAT DATA
AS A BOX AND WHISKER PLOT
quickmeme.com


2022 Carbonate GPU Utilization
% Utilization
1.00
0.75
0.50
0.25
0.00
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11
2022-12

Carbonate GPU utilization, second half of 2022



Utilization percentage by number of cards requested



Workflows

To address the question of what kinds of workflows are utilizing the GPUs, we looked at 2022-2023 XALT data from Carbonate and BR200 jobs that utilize the GPUs

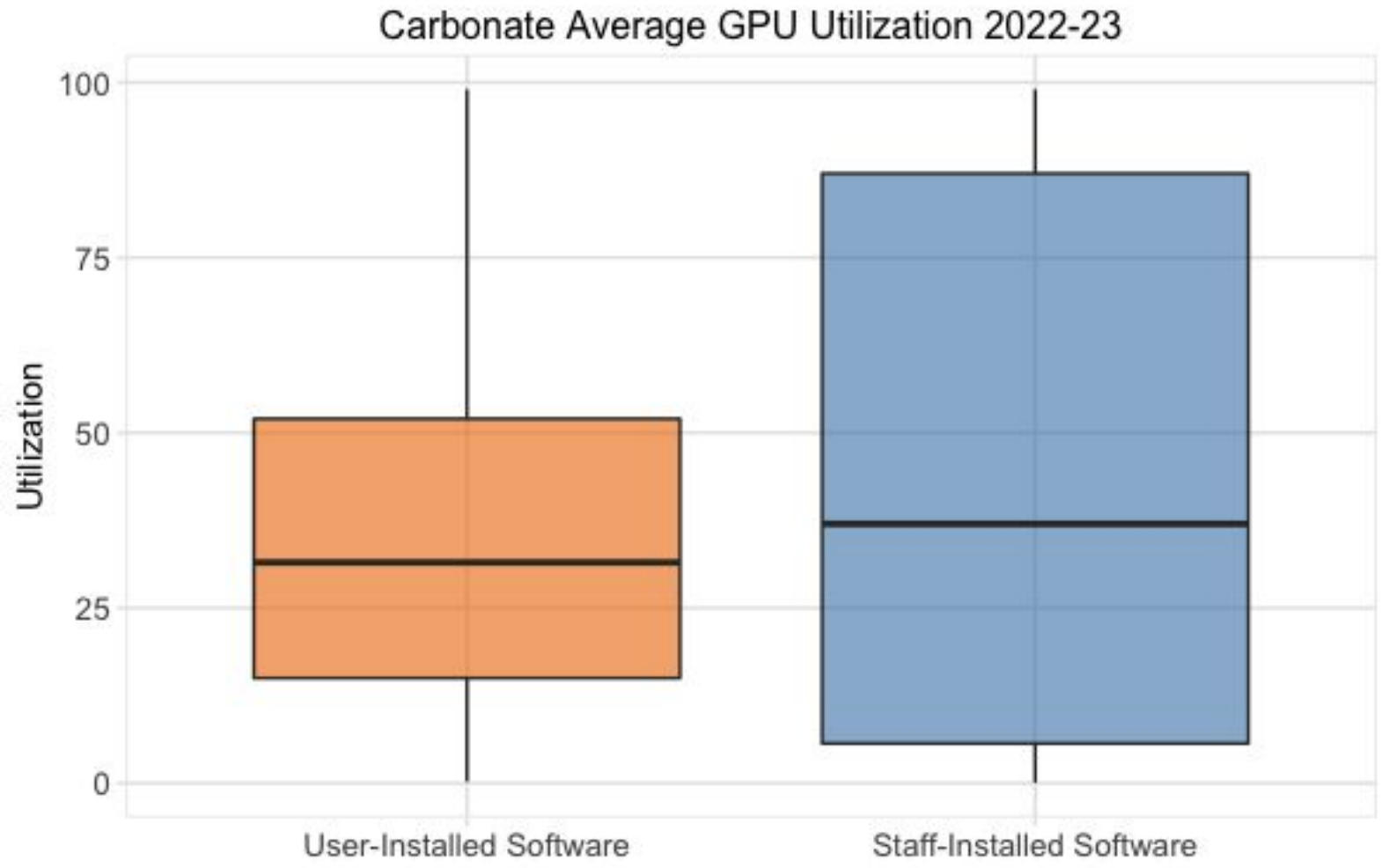
XALT is a tool from TACC that collects job-level data on libraries and executables users access in their jobs

Identified 645 applications from this data and determined if an application was staff- or user-installed and if it was part of an AI workflow or not.

We determined:

- 34% of applications using GPUs was staff-installed
- 53% of applications were part of an AI workflow

Utilization percentage: staff- vs. user-installed



Future work/hopes and dreams

Goals were to form a baseline understanding of how IU researchers are utilizing GPU resources. Our analysis shows a broad spectrum of utilization across our user base.

While the distribution of GPU utilization skews a little low, this does not discount the usefulness of these resources; the question still remains of “what is the desired level of GPU utilization for a job?”

This is certainly not the only way, we invite the community to let us know about their implementation and the analysis that they have done.

Many Thanks

Bret Hammond, Adrian Hosey, Matt Allen and Dave Hancock