GPU cluster monitoring

Tom Rochette <tom.rochette@coreteks.org>

August 30, 2025 — 861fb9d0

1 Overview

In this article I list metrics and alerts one should have when monitoring a GPU cluster to ensure efficient utilization of resources.

GPU cluster monitoring is critical for organizations to optimally utilize the limited capacity they have. Without monitoring it is easy for users to leave jobs running that do not use GPU resources, or do not use them efficiently.
In some cases GPU clusters use certain technologies that require the users to provide images with specific libraries, and not including those dependencies can result in significantly worse compute performance.

2 Metrics

- Allocated GPUs
  - Used to determine who (or which project) has GPU allocated (i.e., currently assigned to a running workload)
- GPU utilization
  - Used to determine whether the GPU is partially or fully used, and if it is partially used, to potentially identify the causes
- GPU memory utilization
  - Used to determine if the GPU memory is partially or fully used
  - Used to identify out of memory issues and potential memory leaks
- InfiniBand receive/transmit bytes
  - Used to determine if a workload is making use of the technology
- Job launch wait duration
  - Used to determine when there's queueing of jobs due to compute being exhausted and how long it takes for jobs to start
- Job duration
  - Used to gather statistics about the type of workload running on the cluster in order to make informed decisions

3 Alerts

- Allocated GPUs are used
  - Used to detect jobs that may ask multiple GPUs but end up using 1 or only a few of them
- GPU utilization below threshold (<10%)
  - Used to detect workloads that do not make full use of the GPU or are allocated to an oversized GPU
- GPU utilization above threshold (>90%)
  - Used to detect when the GPU is saturated
- GPU utilization range above threshold (>25%)

  - Used to detect uneven distribution of GPU compute workload
- GPU memory utilization below threshold (<10%)
  - Used to detect workloads that do not make full use of the GPU or are allocated to an oversized GPU
- GPU memory utilization above threshold (>95%)
  - Used to detect when a job is about to run out of GPU memory
- InfiniBand receive/transmit > 0 when running multi-node workloads
  - Used to identify workloads that are not properly configured to use InfiniBand