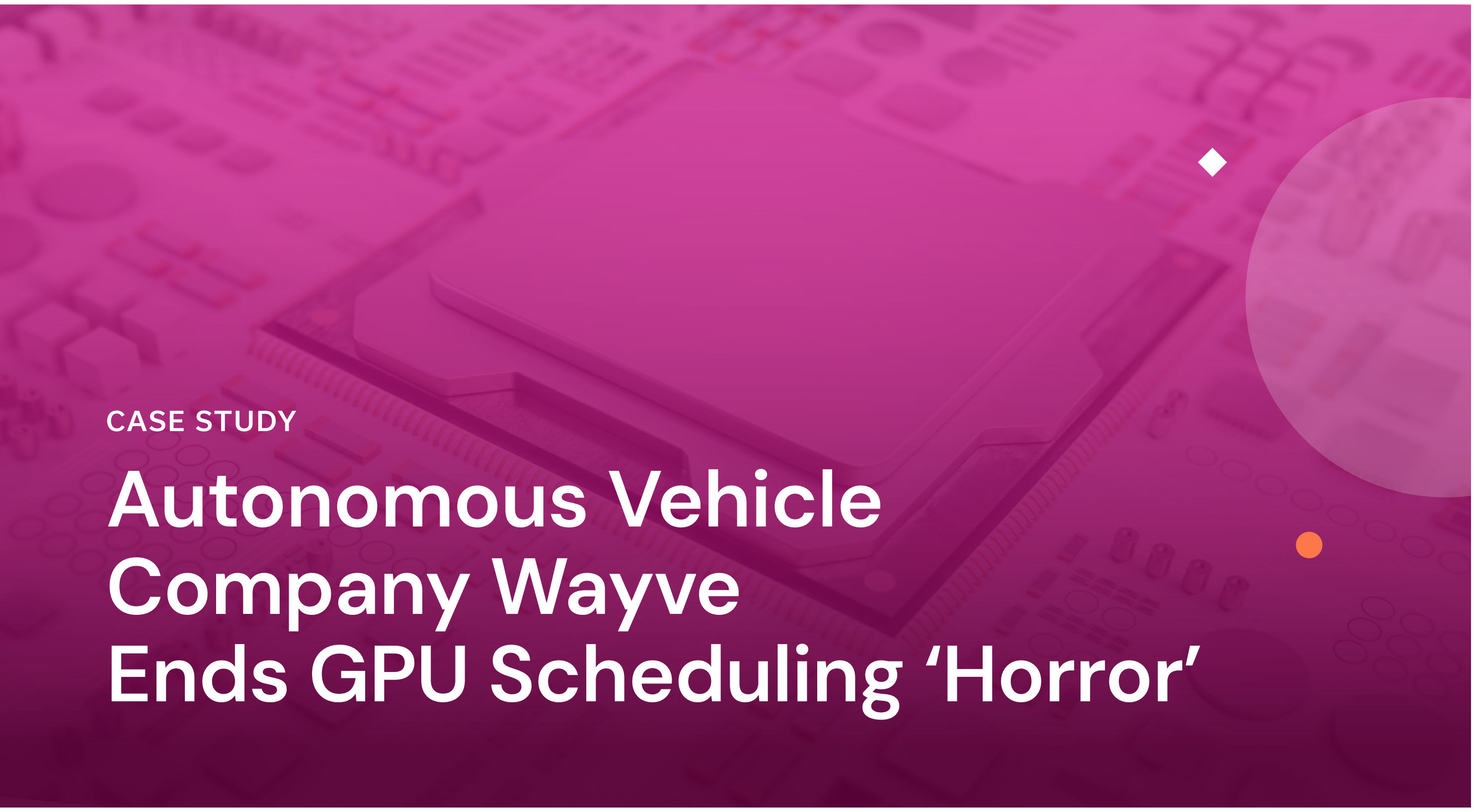


CASE STUDY

Autonomous Vehicle Company Wayve Ends GPU Scheduling 'Horror'

By moving from spreadsheets to Run:ai, Wayve goes from <45% to >80% utilization of GPU resources

About Wayve

Wayve is a London-based company developing artificial intelligence software for self-driving cars. Wayve's embodied AI technology does not depend on costly sensing and instead focuses on greater intelligence, for better autonomous driving in dense urban areas.

Background Fleet Learning Loop Consumes Many GPU Resources

Wayve's Fleet Learning Loop is their continuous cycle of data collection, curation, training of models, resimulation, and licensing models before deployment into the fleet. This loop is Wayve's 'driving school', taking the brains developed by researchers, then training and testing them at scale.

Wayve's primary GPU compute consumption comes from the Fleet Learning Loop production training. They train the product baseline with the full dataset over many epochs, and continually re-train as they collect new data through iterations of the fleet learning loop.

Challenges

Wayve came to Run:ai because of their extremely constrained GPU resources, and were hoping for help with scheduling issues as well. Once installed, it became clear that though nearly 100 percent of GPU resources were allocated to researchers, less than 45 percent of resources were utilized in the July-August time period when the testing was initially done. Because GPUs were statically assigned to researchers, when researchers were not using their assigned GPUs others could not access them, creating the illusion that GPUs for model training were at capacity even as many GPUs sat idle.

> "We were dealing with the horror of scheduling training models via spreadsheets, checking frequently to see who had which GPU and then seeing that a job had died because of a competing job."

Allocation VS. Utilization

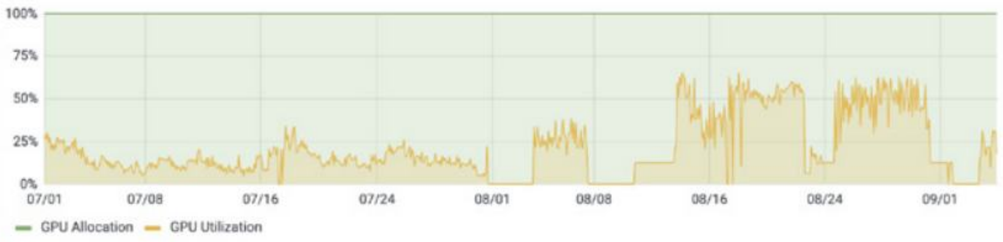
- **Allocation:** an amount or portion of a resource assigned to a particular recipient.
- **Utilization:** an amount or portion of a resource practically and effectively being used.

Wayve had 100% of their resources allocated, and yet less than 45% utilized at any given time.

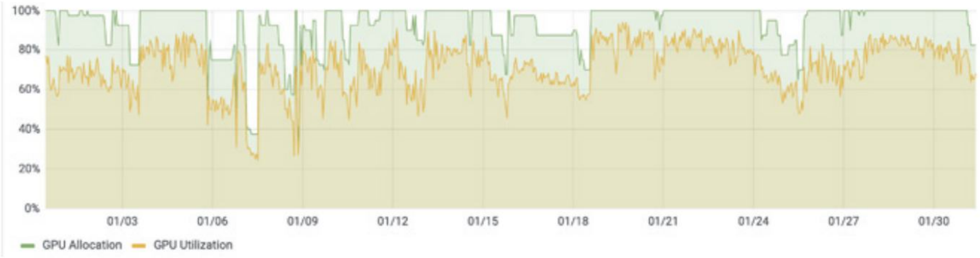
Solution – Advanced Scheduling with Run:ai

Run:ai tackled removing silos and eliminating static allocation of resources. Pools of shared GPUs were created allowing teams to access more GPUs, to run more workloads, and essentially be much more productive. Tens of jobs are submitted to the system by Wayve researchers every day, regardless of team, and jobs are queued and launched automatically by the Run:ai system when GPUs become available. Run:ai's dedicated batch scheduler, running on Kubernetes, enables crucial features for the management of DL workloads like advanced queuing and quotas, managing priorities and policies, automatic preemption, multi-node training, and more. It provides an elegant solution to simplify complex scheduling processes.

Before:
100% allocation,
~25% average utilization



After:
Efficient cluster
utilization of >80%


01/03
01/06
01/09
01/12
01/15
01/18
01/21
01/24
01/27
01/30
GPU Allocation
GPU Utilization