pepperdata

Optimize GPU Spend and Efficiency at Scale

- Match GPU Supply with Demand
- Maximize GPU Efficiency and Utilization
- Dramatically Reduce Costs

Match GPU Supply with Demand

In today's data-intensive world, Graphics Processing Units (GPUs) have become indispensable for running AI and other compute-intensive workloads. Efficient GPU management is critical for preventing overprovisioning, maximizing access to limited GPU resources, and ensuring SLAs are met with minimal cost.

Pepperdata Demand Optimization provides GPU platform managers with a holistic understanding of GPU supply and demand in their environment. Pepperdata empowers platform managers to make data-driven decisions about shifting GPU workload demand based on schedule or GPU type. Instead of reacting to requests from across the company, platform managers can now proactively address imbalances between GPU demand and availability. Pepperdata transforms GPU resource management from reactive and ad hoc to proactive and data driven.

Pepperdata Benefits

- **Strategic demand shifting:** Equipped with a deep understanding of GPU supply and demand, platform managers can move less critical workloads to off-peak hours or other available GPU types based on application requirements. Pepperdata shows not only current but also pending GPU workload demand.
- **Improved GPU resource allocation:** Pepperdata equips platform managers with the information they need to ensure that GPUs are consistently operating at optimal capacity.
- **Reduced overprovisioning:** With accurate insights into true demand from Pepperdata, platform managers can avoid unnecessary GPU scaleups in the cloud or additional capital expenditures on prem.
- **Optimized cloud vendor contracts:** With a clear understanding of their GPU requirements, platform managers can optimize their cloud vendor contracts and avoid paying for unnecessary resources.

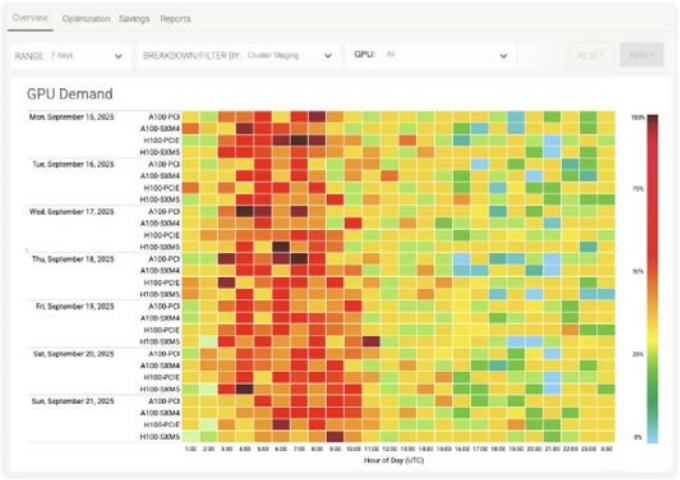

Figure 1: Pepperdata provides GPU platform managers with a holistic understanding of their infrastructure.

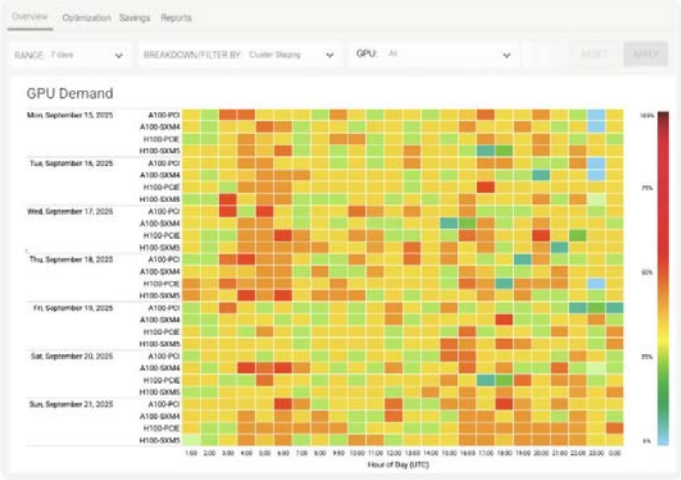

Figure 2: Equipped with this information, platform managers can match GPU demand with available resources.

Automatically Optimize Efficiency and Utilization Across Your GPU Fleet

GPUs are designed for highly parallel workloads. However, many workloads—particularly those performing AI inference—do not require the compute power of a full, high-end GPU. When a relatively small workload locks up the resources of an entire GPU, the result is underutilization and wasted resources.

GPU slicing—partitioning a GPU into smaller units for use by a single application—may seem like a promising solution to this challenge, but it has proven prohibitively difficult for many organizations due to its manual, time-consuming nature.

Pepperdata Resource Optimization maximizes utilization and significantly reduces GPU cost by leveraging NVIDIA's Multi-Instance GPU (MIG) feature. Pepperdata automatically partitions single GPUs into secure, independent GPU slices, creating three GPU slice pools for workload placement:

1. **Full GPUs**: Dedicated for demanding workloads that require an entire GPU.
2. **½ GPUs (2 MIG slices per GPU):** For medium workloads that only require half of a GPU.
3. **⅓ GPUs (3 MIG slices per GPU):** Ideal for lighter workloads that fit within a third of a GPU.

Pepperdata continuously monitors GPU usage and demand. Based on this real-time data, Pepperdata dynamically adjusts the capacity of each pool, scaling up or down as needed to prevent underutilization and bottlenecks. Pepperdata then intelligently assigns workloads to the most appropriate GPU partitions, learning from historical usage patterns to refine these assignments over time.

The result is an automated GPU slice management solution that rightsizes workload placement to dramatically minimize resource waste for GPU-intensive workloads.

Pepperdata Benefits

- **More effective GPU capacity:** Pepperdata maximizes GPU compute and memory utilization, creating more effective capacity in the cloud or on prem.
- **Automatically increased throughput:** More available GPUs and increased utilization mean more workloads can run to completion with the same resources. Furthermore, smaller, less demanding workloads can complete quickly, without having to wait for a full GPU to become available.
- **Significant cost savings:** Pepperdata cuts costs by running more workloads on fewer GPUs.

Supported Technologies

- NVIDIA A100 and newer GPUs
- Cloud environments
  - Amazon EC2 Accelerated Computing Instances
  - Google Cloud GPUs
  - Microsoft Azure GPU instances
- On premises environments

Pepperdata Slices GPUs Automatically and Intelligently for Maximum Utilization

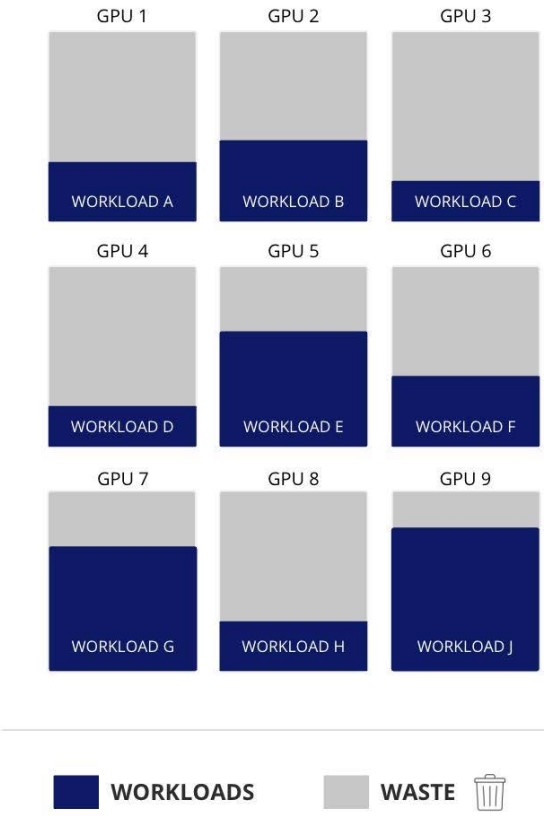

Figure 3: Before Pepperdata is enabled, GPUs operate a fraction of their full capacity. In the cloud, these resources must be paid for whether they are used or not. On prem, operating GPUs at only a fraction of their full capacity means valuable resources are being wasted.

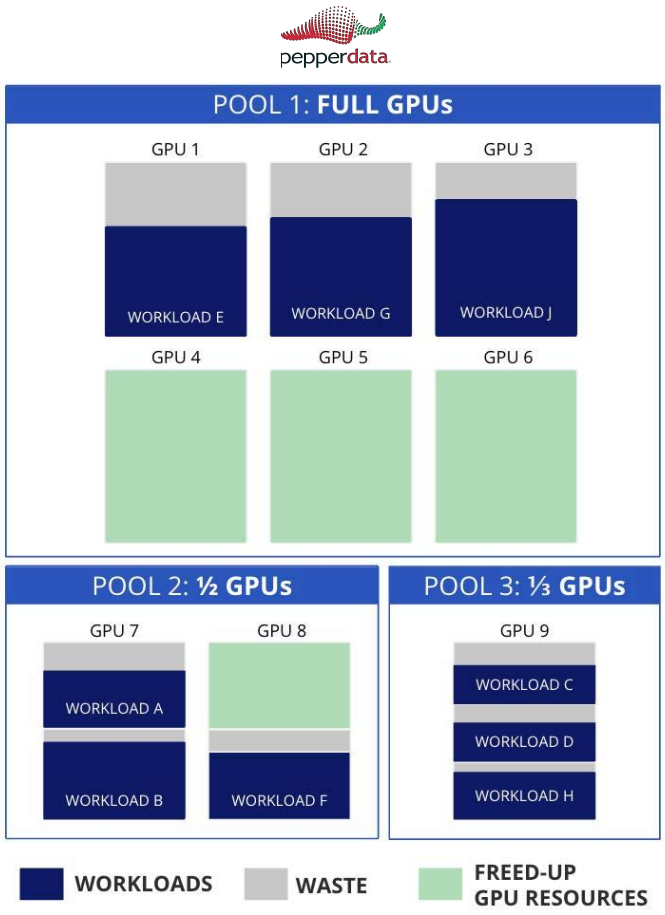

Figure 4: Pepperdata automatically slices GPUs intelligently so that they can be packed at maximum capacity, immediately improving efficiency and minimizing wasted GPU resources.

About Pepperdata

Deployed on over 30,000+ clusters, Pepperdata optimizes resources in some of the largest and most complex environments in the world, providing more pods per node in Kubernetes environments and more effective capacity in GPU environments. Since 2012 Pepperdata has helped companies ranging from startups and mid-sized ISVs to top enterprises such as Citibank, Autodesk, Magnite, Royal Bank of Canada, and members of the Fortune Five save over $250 million. For more information, visit www.pepperdata.com.

aws partner network
Select Technology Partner

AICPA SOC
Formerly SAS 70 Reports

100% ROI GUARANTEE

Pepperdata, Inc.
530 Lakeside Drive
Suite 170
Sunnyvale, CA 94085

Start a Free Trial
www.pepperdata.com

Send an Email
info@pepperdata.com