PAVILION, VMWARE AND NVIDIA VIRTUAL GPU SOLUTION BRIEF

PAVILION

Transforming Enterprises with Performant Virtualized GPUs

Benefits

- Certified by VMware® and Proven with NVIDIA®
- Unmatched price/performance for GPU storage
- Share, aggregate, and automate GPUs to maximize efficiency and Investment
- Proven NVMe-oF speed and reliability

Features

- HyperParallel Block, File & Object Data Platform
- World's Fastest NVMe-oF Windows Clients
- Security & Acceleration for Containers & VMs
- Many-Controller Architecture for Live Migration and DR
- Unmatched performance density for GPUDirect

Organizations achieve superior GPU efficiencies with Pavilion and VMware® for optimized GPU sharing, aggregation and NVIDIA™ Magnum GPU Direct®

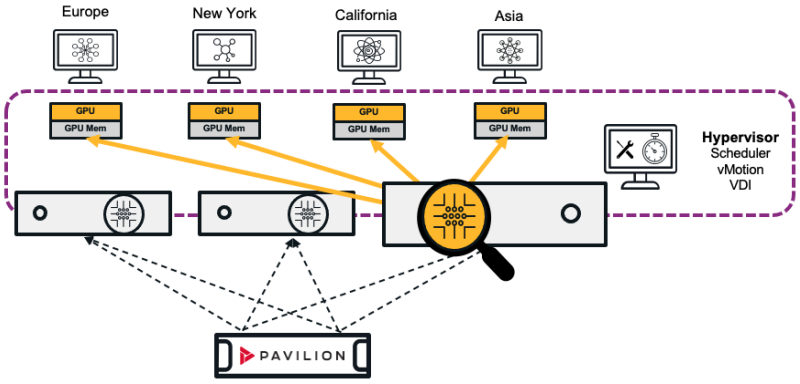
The steady progress of GPU technology is delivering groundbreaking AI/ML/DL results for Supercomputing, Enterprise HPC, Federal Agencies, Manufacturing, Media and Entertainment and many other market segments. Now implementations are expanding in size and scope revealing opportunities to virtualize GPU environments for optimal efficiency and return on investment.

Certified by VMware for NVMe-oF and validated by NVIDIA, customers can leverage Pavilion's HyperParallel Data Platform in virtual GPU environments to achieve greater productivity for global workforces and speed AI training and inference in the data center and at the edge while performing comprehensive data analytics and machine learning with unmatched in performance, density, scalability, and flexibility for on-prem and hybrid cloud storage.

Using NVIDIA AI Enterprise software running on VMware vSphere 7, Update 2, customer workloads access NVIDIA's CUDA applications, AI frameworks, pre-trained models, and software development kits. With Pavilion, they share resources and improve GPU utilization while managing the entire environment under Tanzu and ESXi using the latest A100 Tensor Core GPUs.

Shared GPUs for Remote Workforce Optimization

Here's an example of GPU resource sharing for a "follow-the-sun" workflow. Each user has access to the GPU for rendering at any time, and live migration using vMotion assures zero downtime as different users access GPU resources.

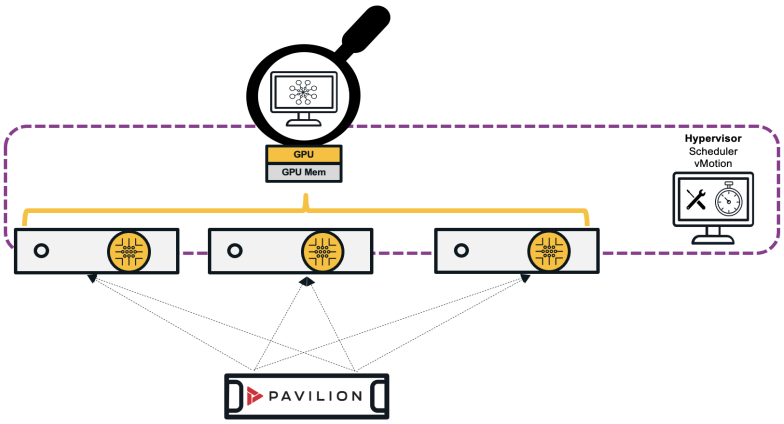


Pavilion storage delivers sub-microsecond Virtual Desktops for GPU sharing at an unprecedented scale. Use block storage with 3rd party file systems or simplify block, file, and object protocols in a HyperParallel data platform that scales up and out.

Migrations of users to various GPU resources are seamless with the Pavilion multi-controller architecture. VMware Tanzu orchestrates everything. Any VDI user can share access to a GPU, and there is no noticeable latency from Pavilion's data storage layer, even as the number of clients increases.

Aggregated GPUs for Heavy AI/ML Workloads

Here's an example of aggregating GPUs to a single user for maximizing resources during off-hours or when several GPUs are idle:



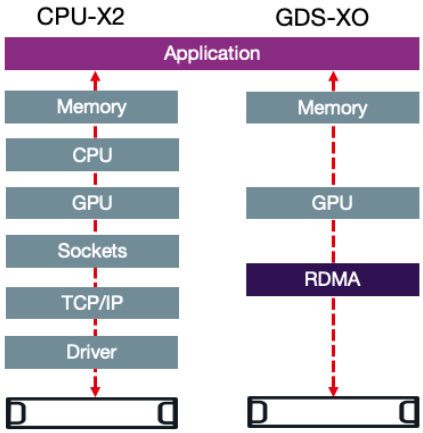
In this example, vMotion has migrated all users off all GPUs and aggregated multiple GPUs to a power user. The power user can run complex AI Inference, rendering, or other data pipelines on a scheduled and orchestrated basis.

Pavilion assures maximum Read/Write performance at scale for this type of workload. Traditional dual-controller AFA "islands" cannot scale up and out on-demand with thin provisioning to assure maximum efficiency and performance across multiple servers with multiple GPUs.

GPUDirect

A new technology from NVIDIA is Magnum IO GPUDirect Storage. Enabled by RDMA and RoCE, this technology is ideally suited for Pavilion HyperParallel Storage. GPUDirect eliminates the CPU from the data path, thereby feeding hungry GPUs faster.

On the left, you can see a traditional CPU-to-GPU pipeline. On the right is a GPUDirect pipeline.



GPUDirect has advantages.

In testing with NVIDIA's benchmarking tool gdsio, Pavilion's performance dominates the competition with read bandwidth >110% higher than competitors, write bandwidth >169% higher than competitors, and latencies reduced by as much as 73%. All of this with 40% to 67% fewer rack units.

In short, Pavilion has the pole position in GPU and GPUDirect storage performance.

Pavilion's validated benchmark testing for both file and block throughput is measured with a single Pavilion array and a single DGX A100. The final numbers are extrapolated to represent the performance of two Pavilion arrays in 8 rack units. With Pavilion performance scales linearly.

NVMe-oF for Windows

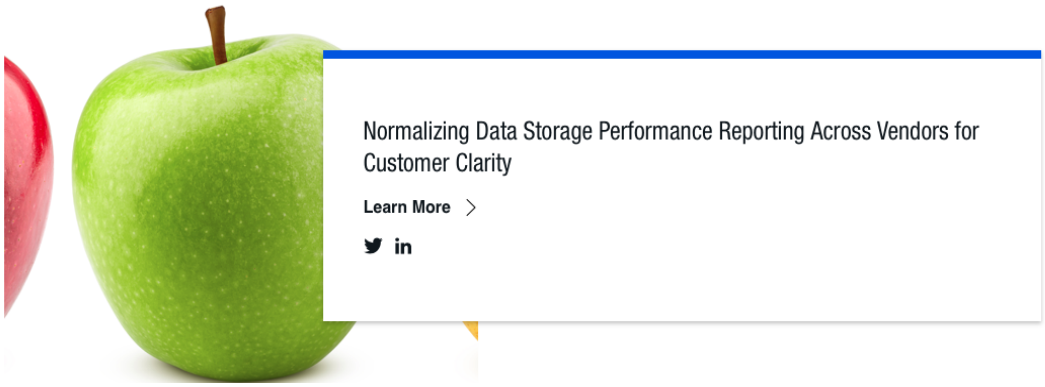
Pavilion also offers a custom NVMe-oF driver for Windows clients. NVMe-oF access over TCP or RoCE for Windows users is paramount and not available from OS or storage providers in certain situations. Pavilion offers drivers for Windows with support for RoCE or TCP for fast and easy access to servers with NVIDIA GPUs.

The Pavilion HyperParallel Data Platform™, powered by Pavilion HyperOS™, is ideally suited for GPU-centric computing. With up to 20 storage controllers, 2.2PB of NVMe capacity in just four rack units, this revolutionary storage system offers the most performant, dense, scalable, and flexible storage data platform in the universe.

Performance density

Performance density is one of the best measures for how different storage solutions compare to each other. The easiest way to do this is to break down the performance of each solution to how many IOPs or how much throughput is provided from each rack unit of space occupied. For example, the Pavilion HyperParallel Data Platform can deliver 120GB/s of read performance for block data with a 4RU system. This means that for each rack unit of space, Pavilion provides 30GB/s of read performance.

To learn more about performance density, please review this blog and storage vendor normalization tool:



Summary

Whether you are enhancing a current AI/ML/DL pipeline with more storage capacity or require greater performance to keep valuable GPUs fully saturated and utilized to the maximum, Pavilion's HyperParallel Data Platform offers the best-in-class block, file, and object throughput per rack unit with the lowest latencies of NVMe-oF storage alternatives.

With certification for VMware vSphere, 7 using RoCE, VMware, and NVIDIA, customers can confidently take advantage of the NVIDIA AI Enterprise Software Suite to share GPUs for optimal resource utilization and aggregate GPUs to accelerate AI/ML pipelines. As NVIDIA's Magnum IO GPUDirect Storage enters the mainstream, Pavilion is the unrivaled solution for customer choice and control of GPU efficiency.

About Pavilion

Pavilion shatters customer expectations and resulting organizational outcomes by revolutionizing data processing for modern AI/ML, HPC, Analytics, Enterprise Edge and other data-driven applications. The Pavilion HyperParallel Data Platform, powered by Pavilion HyperOS, delivers unmatched performance and density, ultra-low latency, unlimited scalability and flexibility, providing customers unprecedented choice and control. Learn why Fortune 500 companies and federal government agencies choose Pavilion. Visit www.pavilion.io or follow the company on LinkedIn.