WHITE PAPER–MAY 2020

Elastic AI Infrastructure for Your vSphere-based Hybrid Cloud

vSphere Bitfusion Guides

vmware®

Table of Contents

AT A GLANCE

vSphere Bitfusion software delivers remote and virtual access to any GPU accelerator in your network, from any VMware vSphere-based virtual machine.

With vSphere Bitfusion, organizations can extend to GPU accelerators the productivity, agility, and powerful utilization of compute, storage, and networking gained with vSphere, so these benefits are available when running artificial intelligence, machine learning, and deep learning workloads.

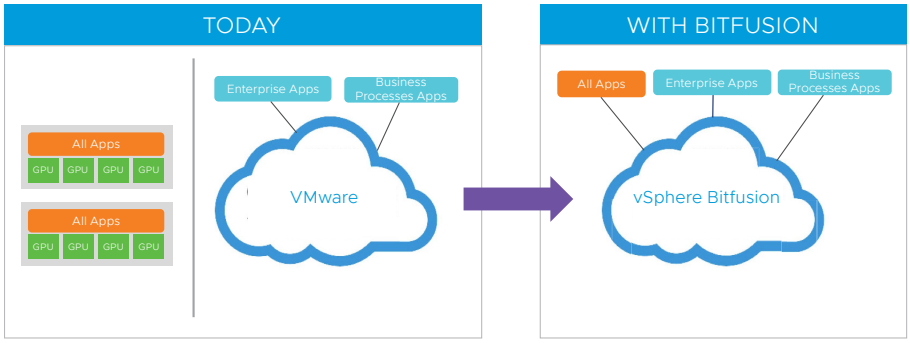
vSphere Bitfusion extends the power of VMware vSphere's virtualization technology to GPUs

Transformation to AI enhances Enterprise Virtualization

Organizations are quickly embracing Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning to open new opportunities and accelerate business growth. AI workloads, however, require massive compute power which has led to the proliferation of GPU acceleration, in addition to traditional CPU power. However, this new technology has created a break in the traditional data center architecture and amplified the problems of organizational silos, poor utilization, and lack of agility. The root cause is that GPU-accelerated servers became siloed, stand-alone assets. GPU servers reduce the agility gained by VMware vSphere®, as they are operated in separate IT "islands." Furthermore, they accelerate Capex and Opex spend, and slow data center modernization. vSphere Bitfusion enables administrators to offer sharing of full or partial GPUs to their users.

vSphere Bitfusion for Elastic GPU Virtualization

vSphere Bitfusion can be used for ML, AI, and HPC applications that benefit from hardware acceleration. One of the best resources presently on the market for hardware acceleration is a GPU. vSphere Bitfusion makes GPUs a first-class resource that can be abstracted, partitioned, automated, and shared — much like traditional compute resources. GPU accelerators can be partitioned into multiple, virtual GPUs of any size, and accessed remotely by VMs over the network. With vSphere Bitfusion, GPU accelerators are now part of a common infrastructure resource pool, and available for use.

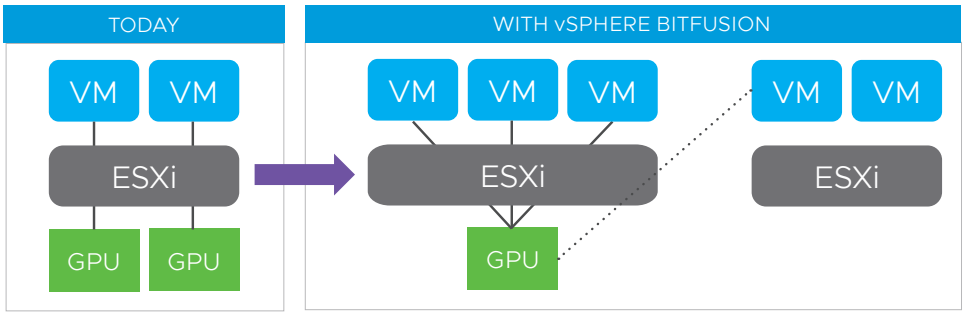


How It Works

vSphere Bitfusion client runs as userspace code within each VM instance, without any need to change the ESXi hypervisor, the kernel, or the AI applications. On the GPU-accelerated server, vSphere Bitfusion also runs as a transparent software layer in a VM and exposes the individual physical GPUs as a pooled resource to be consumed by VMs. vSphere Bitfusion will allocate GPU resources and dynamically attach them over the network. Upon completion of the AI runtime code, vSphere Bitfusion releases shared GPU resources back into the resource pool.

BENEFITS

- Share single, multiple, or partial GPUs
- Share pools of GPUs to any VM across the network
- Attach and detach GPUs based on real-time workload needs
- Partition GPUs based on the demand for AI resources
- Run VMs remotely in any flexible configuration — no local GPU required
- Accelerate development process as demanding workloads can get allocation from common GPU pools
- Optimize Capex and Opex as GPUs are treated as a shared pool and assigned per organization priorities
- Maximize business agility and high availability as VMs run on a compute server, which are physically separated from GPU accelerators
- Future-proof your environment by adding new accelerators over time (e.g., FPGAs, ASICs)



Virtualization Extended to GPU Accelerated Servers

With the new vSphere Bitfusion solution, GPUs are no longer a siloed, unconnected resource. Instead, they are a shared, virtualized pool of resources that can be accessed by any VM in the organization. Much like CPU and storage resources, GPU deployments now benefit from optimized utilization, reduced Capex and Opex, and accelerated development and deployment of R&D resources. These new benefits are extended to all data scientists and AI developers in the organization.

Single Platform

Compute, Storage, Network, and now GPUs are part of the enterprise VMware Hybrid Cloud. Organizations can scale the operations with policies and business logic (e.g., time-of-day policies, class of users, permission to access the top performance GPUs per user class, etc.) for AI developers. GPUs from different departments can be pooled to create bigger clusters to increase compute performance and infrastructure utilization.

Accelerated Development, Testing and Deployment

IT regains the ability to assign GPU resources based on organization business priorities and remotely pool together resources while attaching them in real-time to workloads, with known schedule and utilization plans. For example, GPU resources from Department A, which completed an intensive training and development schedule, can be reassigned to Department B, which now is experiencing peak demand for GPUs for an urgent AI project.

Learn More

Visit https://www.vmware.com/solutions/business-critical-apps/hardwareaccelerators-virtualization.html to get more information on Elastic GPUs, on industry migration, and vSphere Bitfusion.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 vmware.com
Item No: VMW-0520-2210_VMW_CPBU Technical White Papers_BitfusionDocs_19 Elastic AI Infrastructure Hybrid Cloud_1.3_MMT 5/20