bitfusion

VMware and Bitfusion Offer Elastic AI Infrastructure Seamlessly Integrated into your vSphere-based Cloud

Solution Brief

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

Transformation to AI Changes Cloud and Enterprise Virtualization

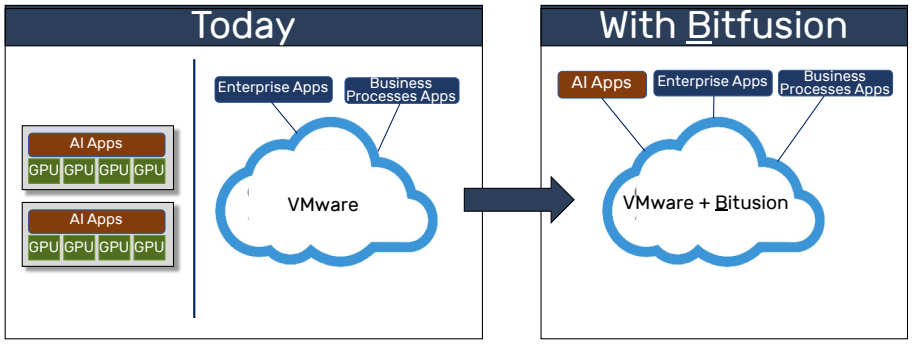
Organizations are quickly embracing Artificial Intelligence (AI), Machine Learning and Deep Learning to open new opportunities and accelerate business growth. AI workloads, however, require massive compute power which has led to the proliferation of GPU acceleration in addition to traditional CPU power. 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.

AT A GLANCE

Bitfusion FlexDirect software delivers remote and virtual access to any GPU accelerator in your network, from any VMware vSphere-based virtual machine. With Bitfusion, organizations can extend the productivity, agility and powerful utilization of compute, storage, networking gained with vSphere to GPU accelerators running artificial intelligence, machine learning and deep learning workloads. Now the same vSphere economics that applied to servers and storage, applies to GPU accelerators, be it in your private cloud or in a public cloud.

Bitfusion on VMware vSphere for Elastic GPU Virtualization

Bitfusion on VMware vSphere 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 Bitfusion, GPU accelerators are now part of a common infrastructure resource pool and available for use by any VM in the vSphere-based cloud.

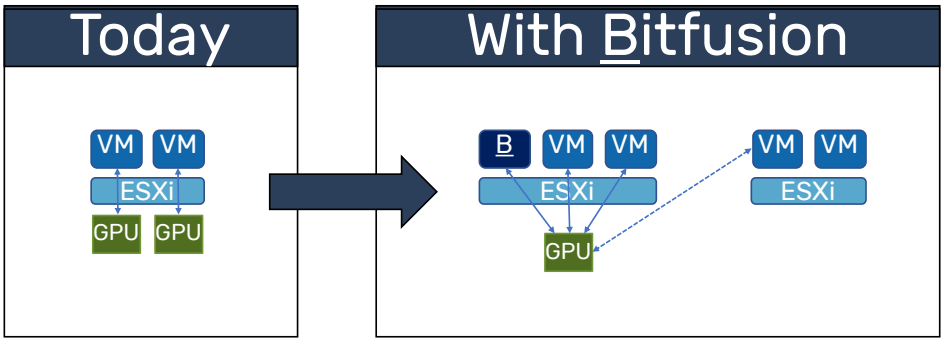


How It Works

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

Virtualization Extended to GPU Accelerated Servers

With the new Bitfusion and VMware 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.



BENEFITS

- Shared pool of GPUs can be assigned to any VM (across the network)
- GPUs attached and detached based on real-time workloads needs (elastic)
- Fractional GPUs can be partitioned based on the demand for AI resources
- VMs run remotely in any flexible configuration, with no need to be bundled with physical GPU servers
- Accelerated development process as demanding workloads can get allocation from common GPU pools
- Optimized Capex and Opex as GPUs are treated as shared pool and assigned per organization priorities
- Maximum business agility and high availability as VMs run on compute server, which are physically separated from GPU accelerators

Single Platform

Compute, Storage, Network, and now GPUs are part of the enterprise VMware vSphere-based cloud. Organizations can scale the operations with policies and business logic (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 the workloads, with known schedule and utilization plan. For example, GPU resources from Department A which completed intensive training and development schedule, can be reassigned to Department B which now experiences peak demand for GPUs for urgent AI project.

Operate in Multi Cloud

With VMware and Bitfusion's solution, IT organization do not need to make hard choices between public cloud, private cloud, or on-premises; the combined solution (Bitfusion and vSphere) will work in any environment.

Learn More

Visit www.bitfusion.io to get more information on Elastic GPUs
Download a trial version of Bitfusion elastic GPU software at https://bitfusion.io/product/flexdirect/
Information on industry migration at http://blog.bitfusion.io/bitfusion-in-2018-01
Bitfusion product information at https://docs.bitfusion.io/docs