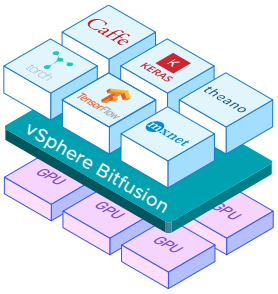
WHITE PAPER–MAY 2020

vSphere Bitfusion Provides an Elastic AI Infrastructure Disaggregation Platform for Cloud and On-Premise

vSphere Bitfusion Guides

vmware®

Table of Contents



CONNECT ANYWHERE

vSphere Bitfusion connects any compute servers remotely, over Ethernet, InfiniBand, or RDMA network to GPU server pools.

ATTACH AND DETACH

vSphere Bitfusion attaches and detaches GPUs to workloads in real-time, offering unprecedented utilization of GPUs.

SLICE GPUS

vSphere Bitfusion slices GPUs into virtual GPUs of any size, allowing multiple workloads to run in parallel.

WORK ANYWHERE

vSphere Bitfusion runs in user-space and is proven to work in cloud, on-premises hardware, and container environments.

Challenge

Development and deployment of AI and deep learning applications require a truly elastic platform that supports experimental testing through to production. These requirements are not met well with current configuration practices and deployment models, which assume static allocations of GPUs for users or frameworks regardless of GPU utilization, performance and scalability. Typically, workday GPU utilization, even for advanced users, averages below 15%. Large-scale deployments do not share or pool GPU servers, which typically serve point and localized applications. These deployments also force non-flexible configurations of compute GPUs.

vSphere Bitfusion Solves the Challenge

vSphere Bitfusion is a transparent virtualization layer combining multiple GPUs and CPUs into a single elastic compute cluster to support sharing, scaling, pooling, and management of compute resources.

vSphere Bitfusion dramatically optimizes existing GPU solutions with 2-4x better utilization (which results in similar cost savings) and offers the ability to dynamically adjust compute resources from fractions of a GPU to many GPUs, with on-the-fly network-attached GPUs from remote systems.

vSphere Bitfusion Delivers Four Dimensions of Innovation

The four pillars of vSphere Bitfusion are remote-attach over any network, dynamic attach, partial GPU, cloud. Combined, they deliver massive TCO savings, flexibility, IT productivity, and cloud economics in Enterprise IT.

vSphere Bitfusion Is Compatible with Any Environment

Requiring no operating system, hardware, or code changes, vSphere Bitfusion integrates seamlessly with existing bare-metal, virtual machine (VM), or containerized applications.

"Bitfusion's innovative technology fits right in with our reconfigurable cloud computing vision, and allows us to deliver superior market value."

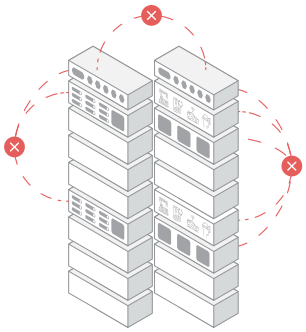
LEO REITER
CTO
NIMBIX

"It's clear that Bitfusion offers a powerful new virtualization technology to elastically manipulate compute resources, while also enabling a highly streamlined AI development experience."

BHAVESH PATEL
DELL EMC
(GTC, 2017)

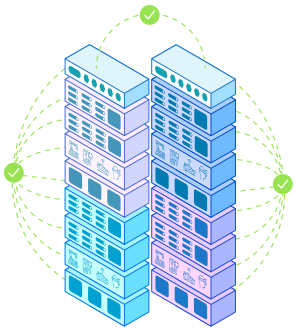
"With MapD plus Bitfusion running on IBM Cloud, we can deliver insights in milliseconds in real time, even over the biggest datasets."

TODD MOSTAK
CO-FOUNDER & CEO
MAPD



Current Approach Suffers from Scaling and Performance Issues:

- Increased cost with denser servers
- GPU density limited by physical dimensions and thermal constraints
- Power supply limits reduce rack density
- Top-of-the-rack bottleneck, limited scalability
- Limited multi-tenancy on GPU servers (limited CPU memory per user)
- Cannot support GPU applications with high storage, CPU, memory requirements



A Hyperconverged Solution and Network-attached GPUs:

- 50% less cost per GPU by using smaller GPU servers
- Scalable to multiple GPUs servers
- 4X more AI application throughput
- Supports GPU applications with high storage, CPU requirements
- Scalable — less global traffic
- Composable — add resources as you scale

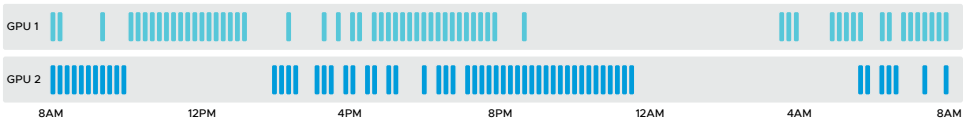
vSphere Bitfusion Employs Efficient Runtime Optimizations

How is this achieved? The vSphere Bitfusion virtualization layer has several runtime optimizations to automatically adapt the best combination of transports: Host CPU Copies, PCIe, Ethernet, InfiniBand, GP and RDMA to achieve superior results. In many cases, virtualized and remotely attached GPUs using vSphere Bitfusion approximate native GPU performance and efficiency across a variety of machine learning workloads.

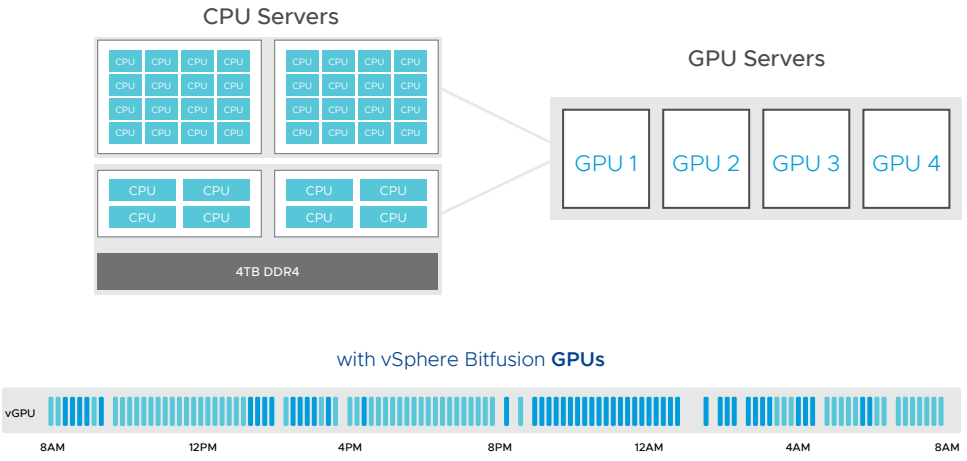
	BASELINE	vSPHERE BITFUSION	MANUAL
Cost/User	1x	0.2x	0.5x
Security Isolation	N/A	Yes	Partial
Performance QoS	N/A	Yes	No
Setup Time	N/A	Minutes	Weeks
Ease of Management	Simple	Simple	Complex

1. **Baseline:** Each application requires a dedicated physical GPU
2. **vSphere Bitfusion:** Many applications run on virtual GPUs mapped to a single physical GPU
3. **Others:** Manual container-only approaches

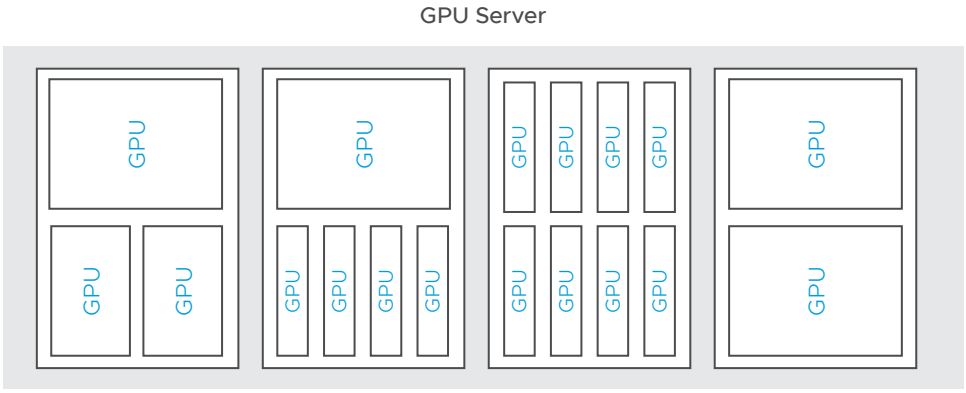
vSphere Bitfusion Offers Reduced TCO and Increased Utilization of Expensive Accelerators



Not only does vSphere Bitfusion allow you to attach GPUs to any machine remotely, offering reduction in total cost of ownership, it also lets you slice a single GPU into multiple virtual GPUs of any size, providing increased performance along with increased utilization due to packing more workloads to run in parallel on the same GPU.

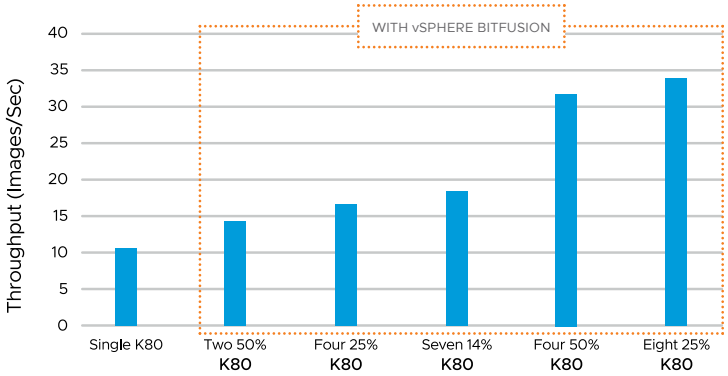


vSphere Bitfusion allows you to take advantage of underutilized GPU compute cycles by allowing real-time aggregation and disaggregation of GPUs. For instance, you can keep your workloads on CPU machines most of the time and remote-attach a GPU only when the workload needs a GPU, increasing utilization of GPUs by 2-4x.



vSphere Bitfusion improves the unit economics of use-cases which may not take advantage of entire nodes or GPUs, cases such as early testing and validation of machine learning algorithms. Fractional GPUs (as small as 1/20th of a GPU) can be assigned at runtime to support many more users than before on the same physical hardware. This affords fine-grained resource control without having to resort to a variety of lower-powered devices that would increase the scope and burden of infrastructure management. vSphere Bitfusion delivers high-performance GPU instances at significantly lower cost and enables users to "right-size" spend and capacity to various stages of development and testing.

Increased Inference Performance with vSphere Bitfusion



Use Case: Partial GPUs For Inference Workloads
Hardware: AWS EC2 r4.xlarge, p2.8xlarge (K80 GPUs)
Software: OS Ubuntu 16.04 LTS
Benchmark: Tensorflow 1.4.1 RNN Model
GPU: K80
Libraries: Cuda 9.0, CuDNN 7

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 Bitfusion Virtualization_1.5_MMT 5/20