bitfusion

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

Challenge

Development and deployment of AI and deep learning applications requires truly elastic platform supporting experimental testing to a production ready deployment. It does not fit well with current configuration practices and deployment models, which assume a static allocation of GPUs for each user or framework regardless of utilization, performance and scalability. Typically, GPU utilization throughout a work day, for even advanced users, averages below 15%. In large scale deployment, GPU servers cannot be shared or pooled, and would typically serve pointed and localized applications. It also forces a non-flexible configurations of compute GPUs.

Bitfusion FlexDirect solves the challenge

Bitfusion FlexDirect 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.

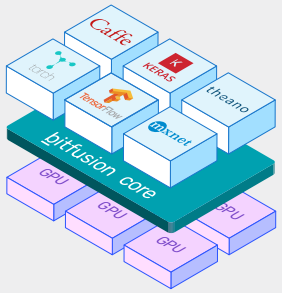
FlexDirect dramatically optimizes existing GPU solutions with 2-4X better utilization (which results in similar cost savings) and offers the ability to dynamically adjust compute resources from fractions of a GPU to many GPUs, with on-the-fly network attached GPUs from multiple systems.

FlexDirect delivers 5 dimensions of innovation

Remote attach over any network, dynamic attach, partial GPUs, multi-cloud and support for both CUDA and OpenCLI are the 5 pillars of FlexDirect. Combined together, it would deliver massive TCO savings, flexibility, IT productivity, and cloud economics in Enterprise IT.

Bitfusion FlexDirect is compatible with any environment

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



Connect Anywhere

FlexDirect connects any compute servers remotely, over Ethernet, Infinabad RDMA of RoCE network to GPU server pools.

Attach and Detach

FlexDirect attaches and detaches GPUs to workloads in real-time, offering unprecendented utilization of GPUs.

Slice GPUs

FlexDirect slices GPUs to virtual GPUs in any size allowing multiple workloads to run in parallel.

Work Anywhere

FlexDirect runs in userspace and proven to work in public cloud, private cloud, on-premise hardware, any hypervisor and containter.

CUDA & OpenCL Support

FlexDirect has extensions to support FPGAs and ASICs (any openCL compliant hardware).

> **"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

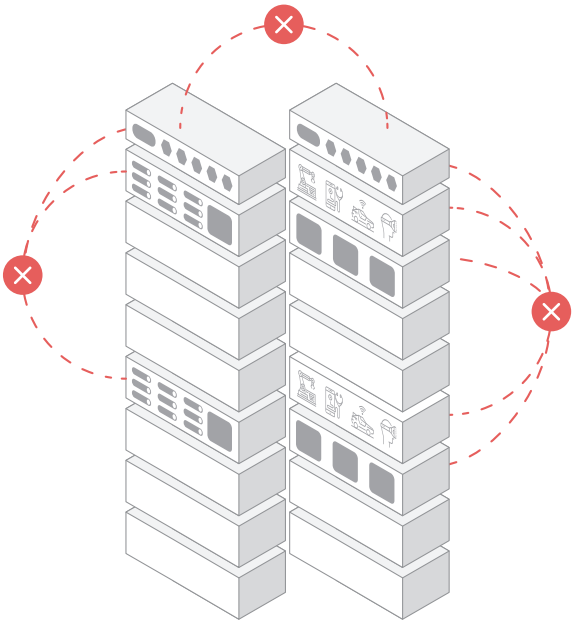
> **"It's clear that Bitfusion offers a powerful new virtualization technology to elastically manipulate compute resources, while also enabling a highly streamlined AI development experience."**
>
> **Bhavesh Patel**
> Dell EMC
> (GTC, 2017)

> **"We were able to put together a super node using Bitfusion's technology with 64 GPUs, which is really unheard of."**
>
> **Jerry Gutierrez**
> Global AI Solution Leader
> IBM Cloud

> **"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

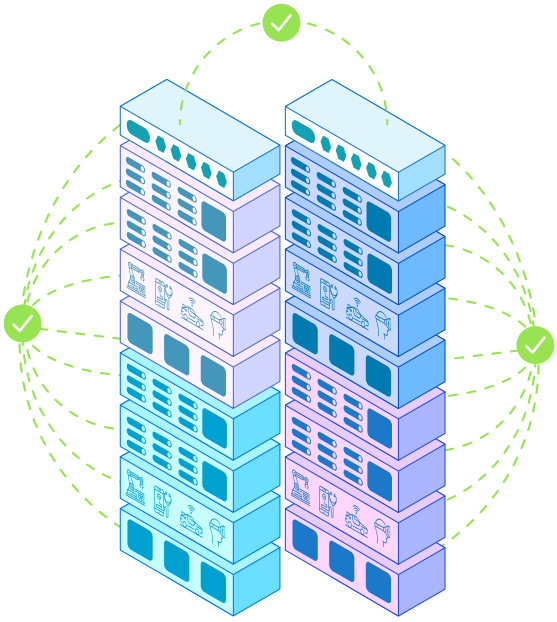
bitfusion

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



Current approach suffers from scaling and performance issues:

- Increased cost with denser servers
- GPU density limited the 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 hyper-converged solution and network attached GPUs:

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

Bitfusion FlexDirect employs efficient runtime optimizations

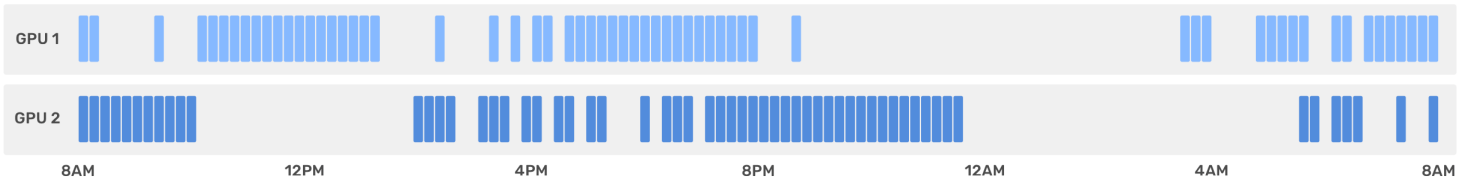
How is this achieved? The Bitfusion virtualization layer has several runtime optimizations to automatically adapt the best combination of transports: Host CPU Copies, PCIe, Ethernet, InfiniBand, GPU Direct RDMA to achieve superior results. In most cases, virtualized and remotely attached GPUs using Bitfusion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

Bitfusion vs. Alternatives

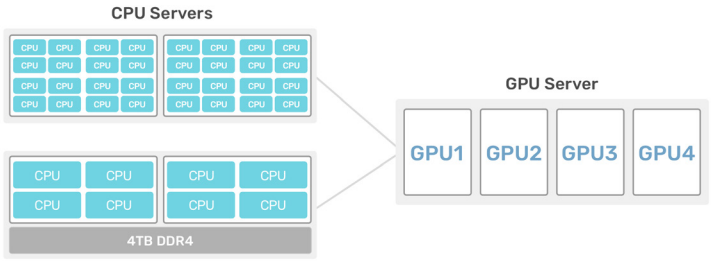
	Baseline	Bitfusion	Manual
Cost / User	1X	0.2X	0.5X
Security Isolation	N/A	Yes	Partial
Performance QoS	N/A	Yes	No
Setup Time	N/A	Minutes	Weeks
Ease of Management	Simple	Simple	Complex

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

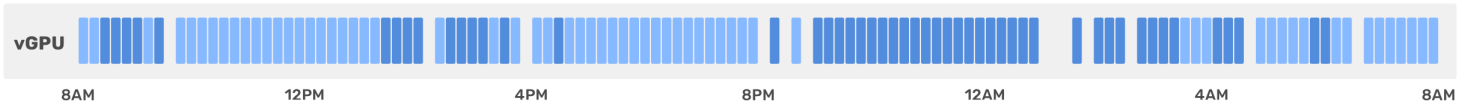
Bitfusion FlexDirect offers reduced TCO and increased utilization of expensive accelerators



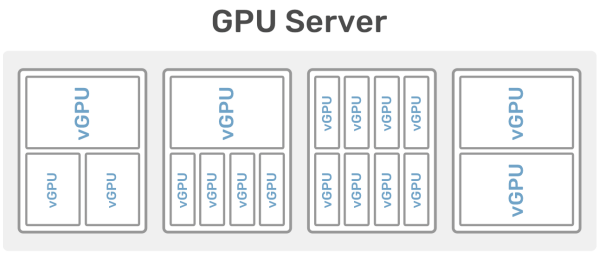
Not only does FlexDirect allow you to attach GPUs to any machine remotely, offering reduction in total cost of ownership, it also let's you slice a single GPU into multiple virtual GPUs of any size, providing increased performance along with increased utilization due to packing more workloads to run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



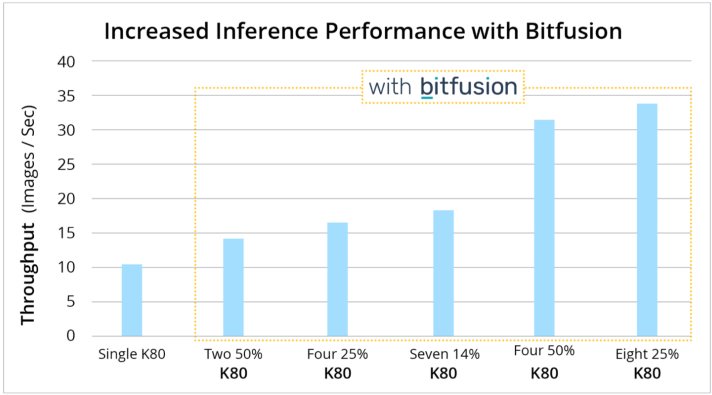
with bitfusion FlexDirect



FlexDirect allows you to take advantage of underutilized GPU compute cycles more efficiently 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.



Bitfusion FlexDirect improves the unit economics of use cases which may not take advantage of entire nodes and GPUs, 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. FlexDirect delivers high performance GPU instances with significantly lower costs and enables users to "right size" spend and capacity to various stages of development and testing.



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