WHITE PAPER–OCTOBER 2019

FlexDirect

Bitfusion Guide

vmware®

Table of Contents

FlexDirect GPU Virtualization

Bitfusion FlexDirect is a transparent virtualization layer combining multiple GPU and CPU systems into a single elastic compute cluster to support sharing, scaling, pooling and management of compute resources.

FlexDirect dramatically optimizes existing GPU solutions with two to four times 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 attachment of GPUs from multiple systems.

Overview

FlexDirect Delivers Several Dimensions of Innovation:

- FlexDirect connects any compute servers remotely, over Ethernet, Infiniband RDMA or RoCE networks to GPU server pools
- FlexDirect attaches and detaches GPUs to workloads in real time, offering unprecedented utilization of GPUs
- FlexDirect slices GPUs to virtual GPUs of any size, allowing multiple workloads to run in parallel
- FlexDirect runs in userspace and is proven to work in public cloud, private cloud, on-premise hardware environments, and with any hypervisor or container
- FlexDirect has extensions to support FPGAs and ASICs (any OpenCL compliant hardware)
- FlexDirect also allows you to manage the GPUs, maintaining policies for clients, idle timeouts, and monitoring GPU usage

FlexDirect has an analytics tool with which you can measure over time the GPU utilization on your cluster before trying any of our virtualization tools

FlexDirect Is Compatible with Any Environment

Requiring no operating system, hardware, or code changes, Bitfusion FlexDirect works with existing virtual machine (VM), Hypervisor or containerized applications to take full advantage of advanced GPU virtualization.

FlexDirect uses a client-server architecture where servers provide the GPU resources for the cluster, and clients are where end-user applications are run.

- **Bitfusion Application Instance (Client):** The machines where end users will be running their applications. It can be a GPU instance, but it is not required that it be.
- **Bitfusion GPU Instance (Server):** The machines which provide GPU resources to the cluster.

There are many flexible configurations which are possible using FlexDirect. However, the most common are: One-to-Many, Many-to-One, and Many-to-Many.

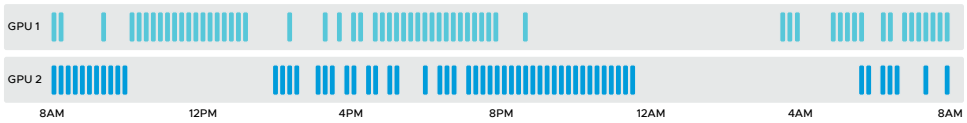
FlexDirect Employs Efficient Runtime Optimizations

FlexDirect has several runtime optimizations to automatically adapt the best combination of transports—Host CPU Copies, PCIe, Ethernet, InfiniBand, GPUDirect 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.

FlexDirect Offers Reduced TCO and Increased Utilization of Expensive Accelerators

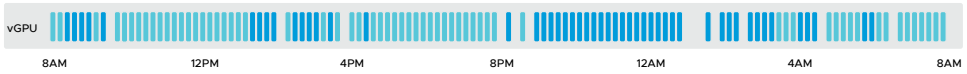
GPU utilization in an organization usually follows a trend like that below.

Today's Standard



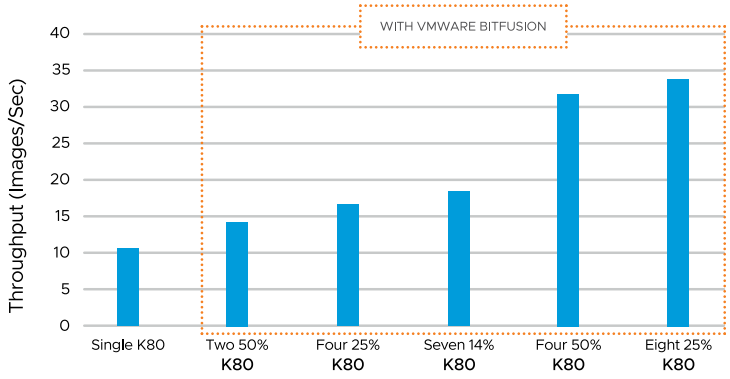
FlexDirect allows you to take advantage of underutilized GPU compute cycles more efficiently by allowing real-time aggregation and disaggregation of GPUs. For instance, you can keep your workloads on CPU machines most of the time and remote-attach a GPU only when the workload needs a GPU, increasing utilization of GPUs by two to four times.

with Bitfusion **vGPU**



Not only does FlexDirect 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 because of packing more workloads to run in parallel on the same GPU.

Increased Inference Performance with 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

FlexDirect improves the unit economics of use cases which may not take advantage of entire nodes and GPUs, such as early test 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" spending and capacity to various stages of development and testing.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 vmware.com