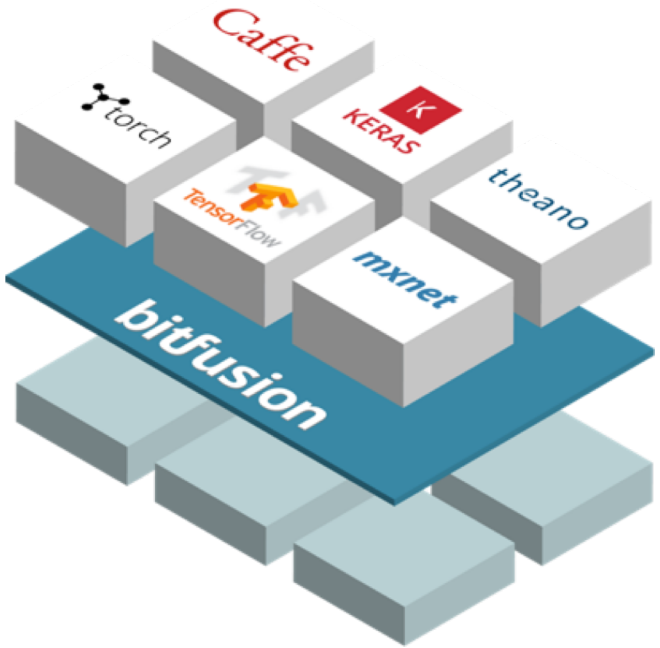
bitfusion

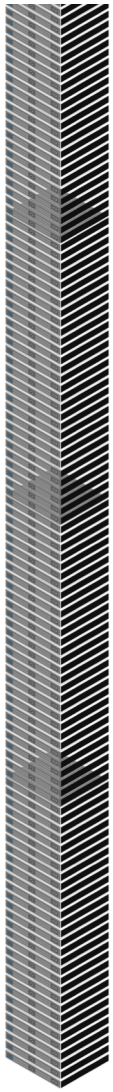
Elastic AI Platform

Jun-2018




Artificial Intelligence
Drives Markets over $100B

INFRASTRUCTURE

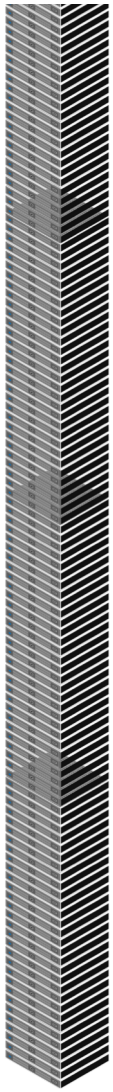


x86 Server

How many total x86 Servers (Globally) are deployed:

- ● >100M
- ○ 80M – 100M
- ○ 60M – 80M
- ○ 40M – 60M

200 : 1
R A T I O



x86 Server

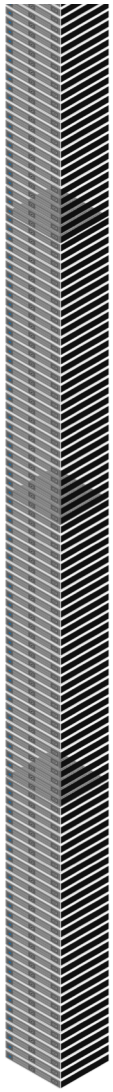
How many total x86 Servers (Globally) are deployed:

- ● >100M
- ○ 80M – 100M
- ○ 60M – 80M
- ○ 40M – 60M

How many total GPU Servers (Globally) are deployed:

- ○ >1M
- ○ 0.75M – 1M
- ○ 0.5M – 0.75M
- ● 0.25M – 0.5M

GPU Server



1 Billion man-year Software Eco System

$7,000 ASP

500w Typical Power

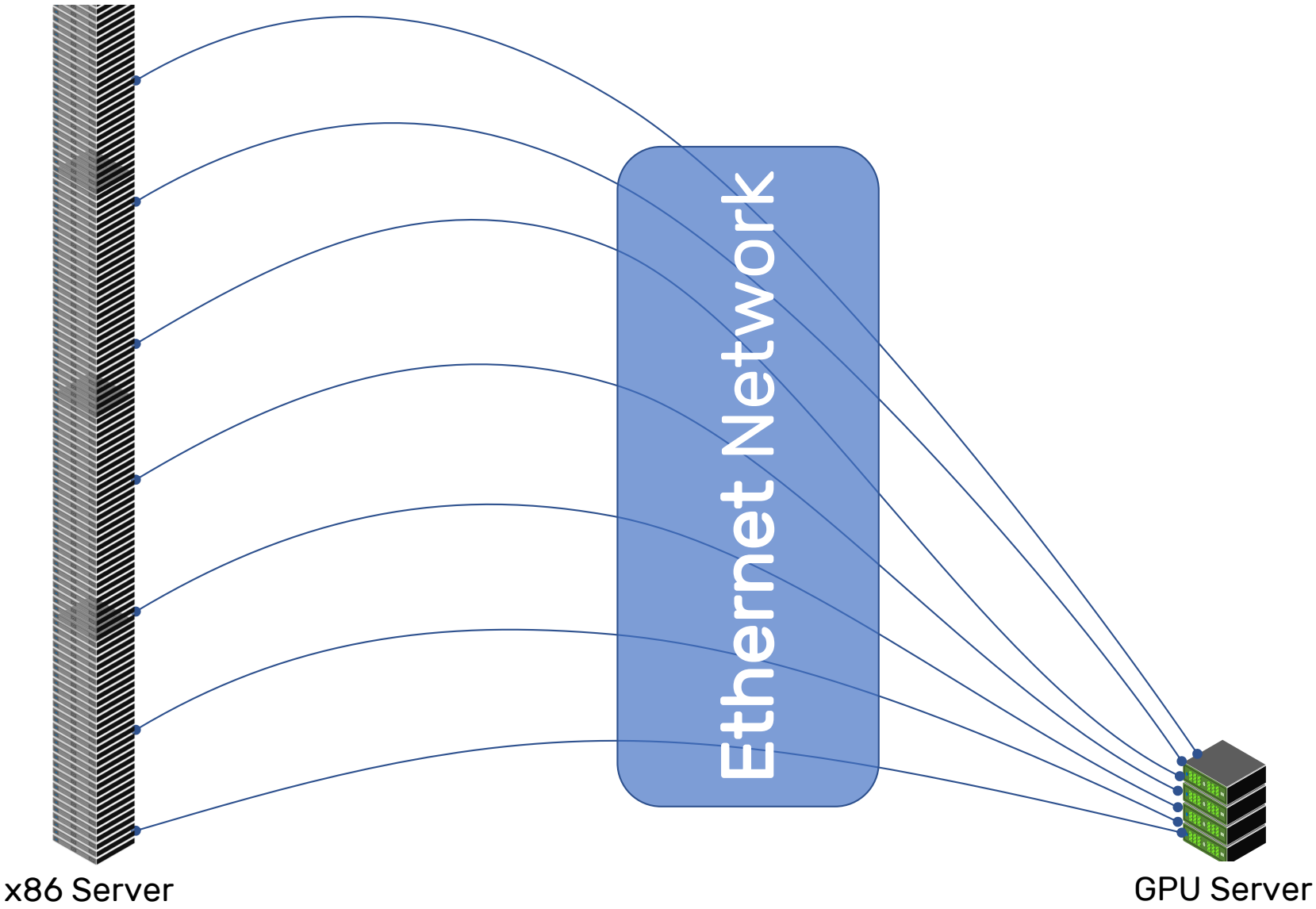
x86 Server

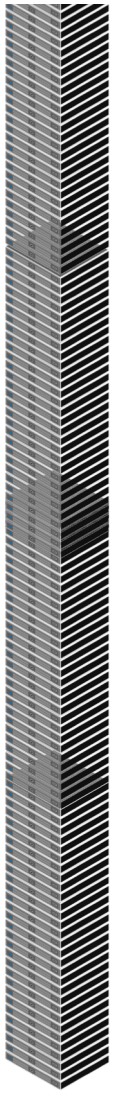
0.05 Million man-year Software Eco System

$50,000 ASP

3000w Typical Power

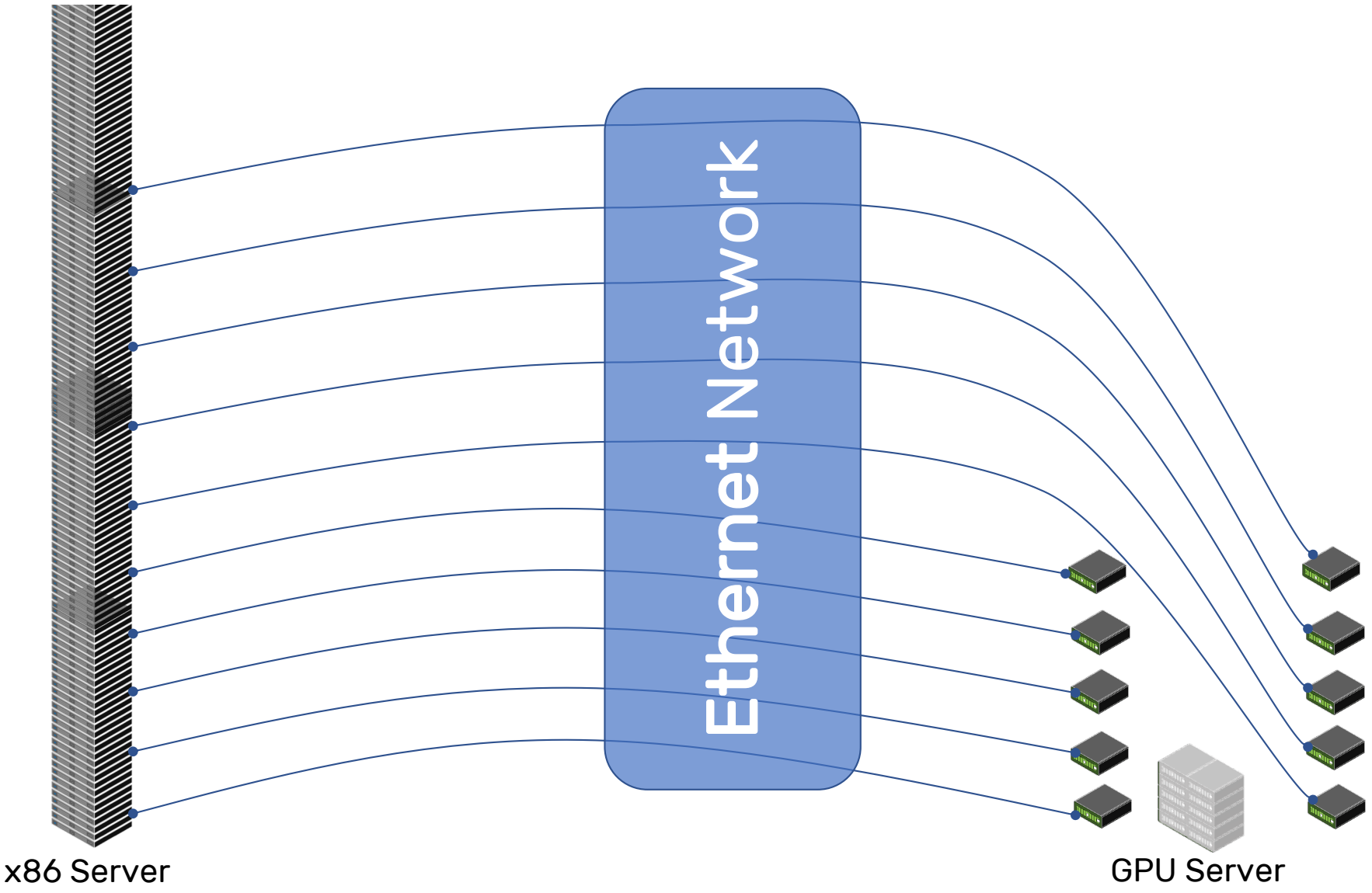
GPU Server

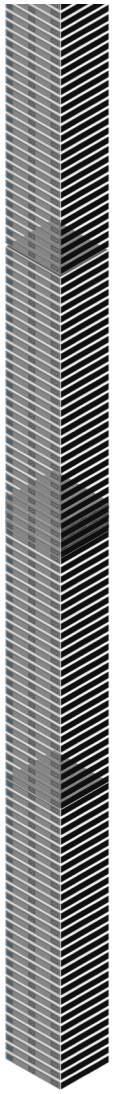

Ethernet Network
x86 Server
GPU Server



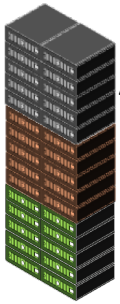
x86 Server

GPU Server


Ethernet Network
x86 Server
GPU Server



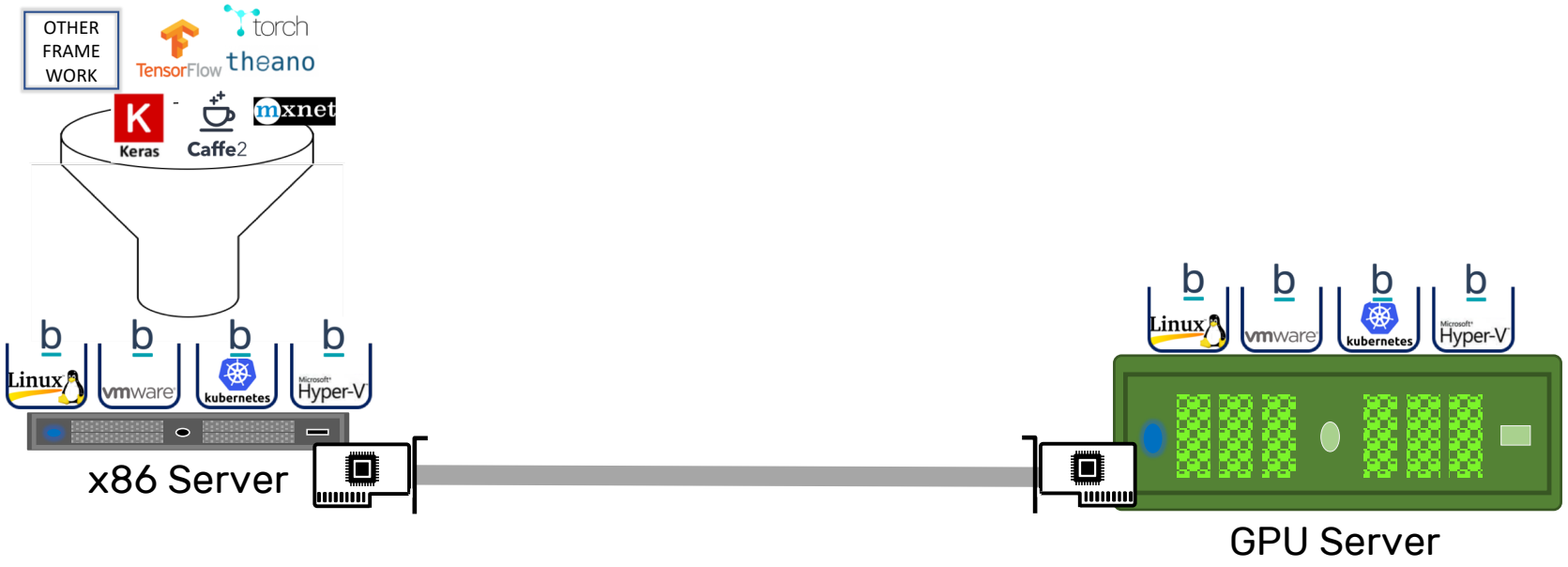
x86 Server

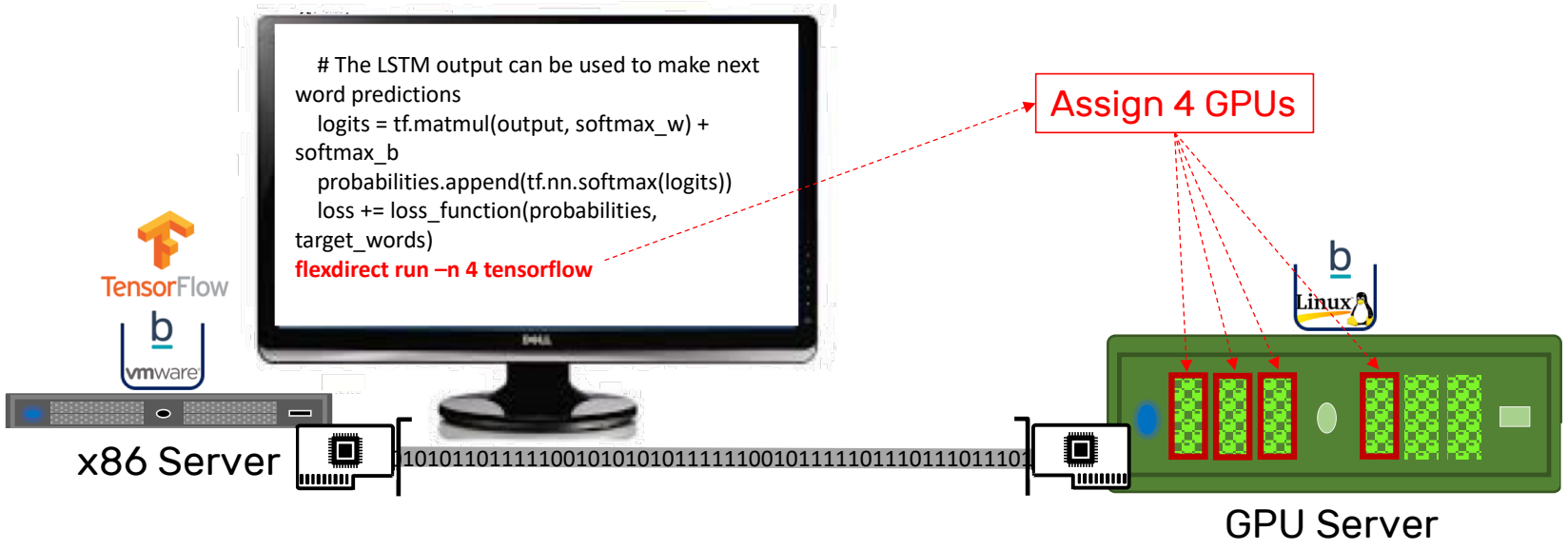


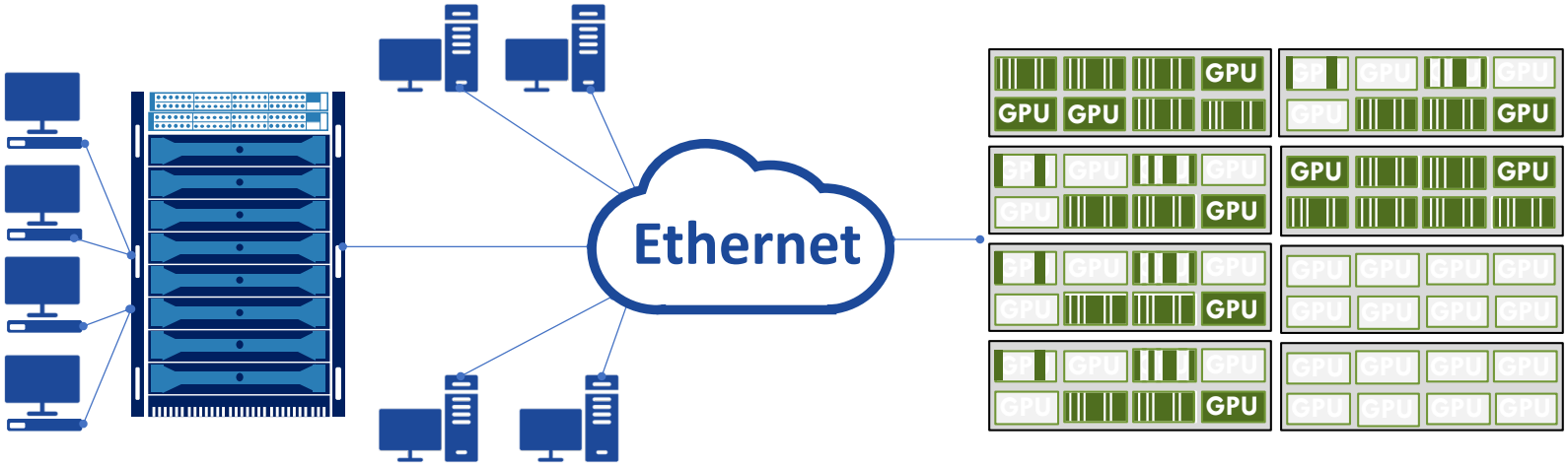
AI ASICs

FPGAs

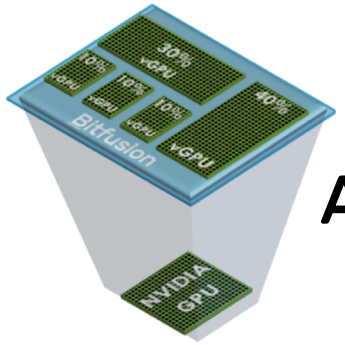
GPU Servers


OTHER
FRAME
WORK
TensorFlow
torch
theano
Keras
Caffe2
mxnet
Linux
vmware
kubernetes
Hyper-V
x86 Server
GPU Server


# The LSTM output can be used to make next word predictions
logits = tf.matmul(output, softmax_w) + softmax_b
probabilities.append(tf.nn.softmax(logits))
loss += loss_function(probabilities, target_words)
flexdirect run –n 4 tensorflow
Assign 4 GPUs
TensorFlow
vmware
Linux
x86 Server
GPU Server



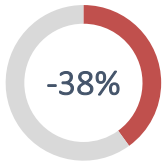
AI Attached Network



AI Hypervisor

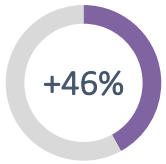
Industry's First Elastic AI Infrastructure

Lower Capex



Lower Opex

Productivity

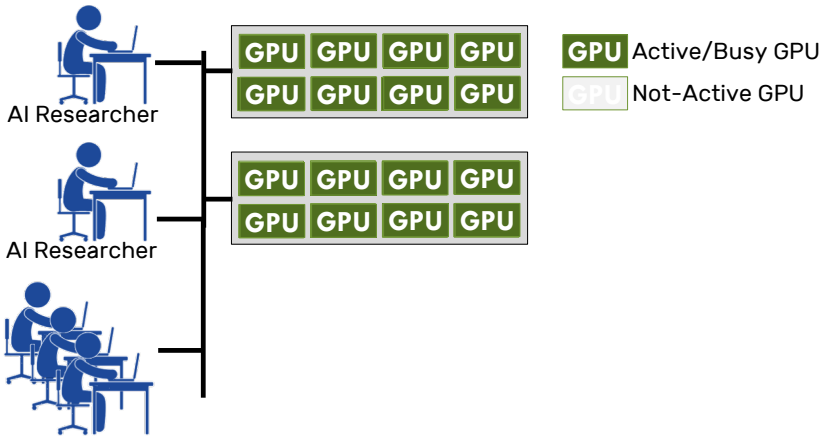


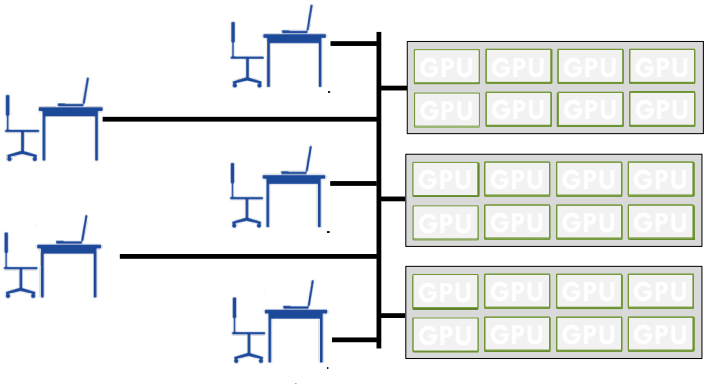
Flexibility

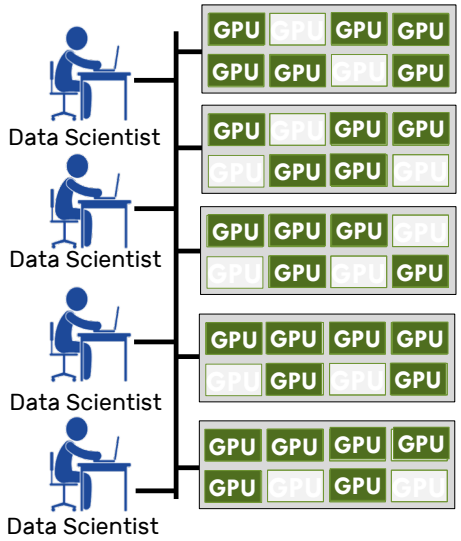
Agility

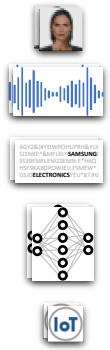
Virtual Elastic GPU cluster from all the scattered deployed GPU servers in the enterprise. Any user, framework and compute server (without GPU resources) can attach instantaneously to a remote fractional GPU, single GPU or group of GPUs in the virtual cluster, run the AI code, and detach.

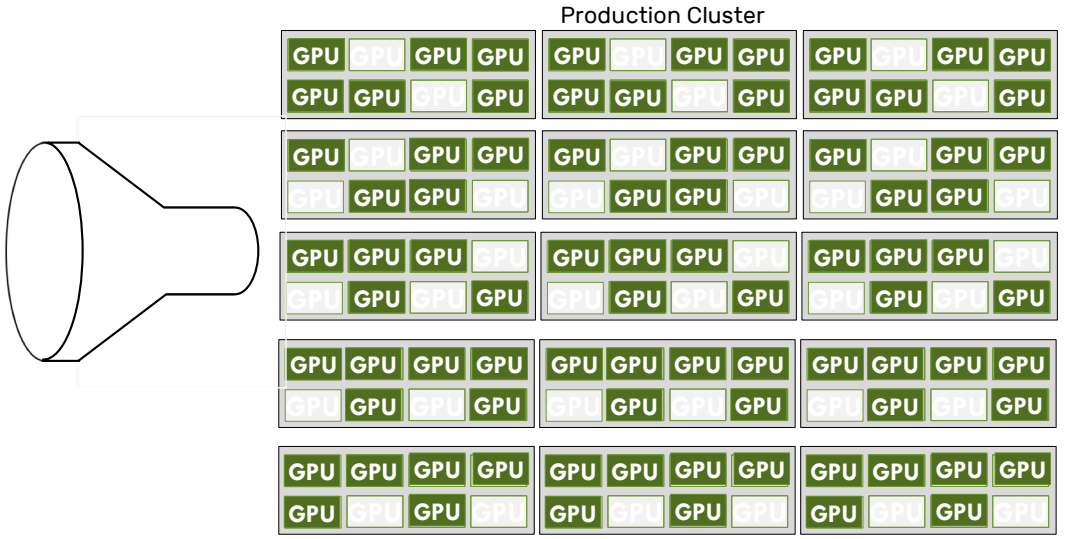
Bitfusion virtualize the capacity and the location of GPUs, and make them accessible to any compute machine in the network. The GPUs do not need to be viewed as physical entities, and 1/20, 1/7, 1/4, or any fraction of GPU can be assigned to a workload, on demand.

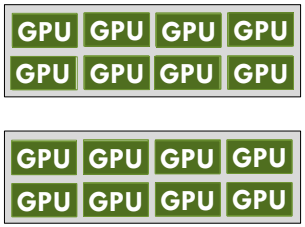

AI Researcher
AI Researcher
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU Active/Busy GPU
GPU Not-Active GPU

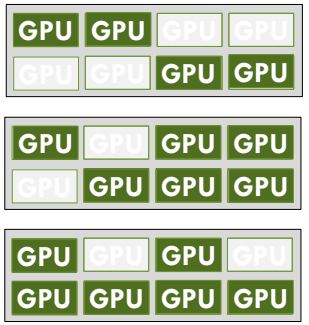

GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU

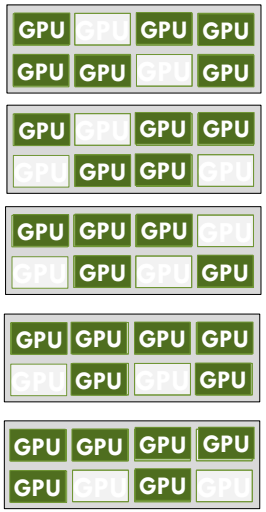

Data Scientist
Data Scientist
Data Scientist
Data Scientist
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU

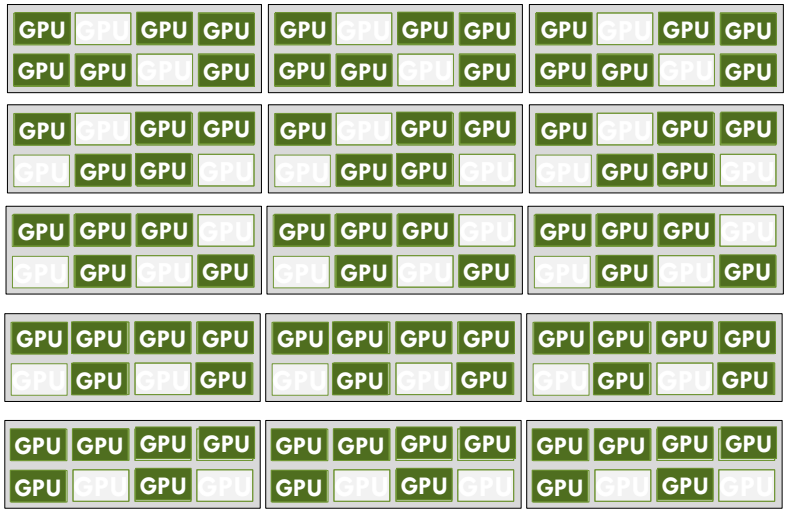

IoT

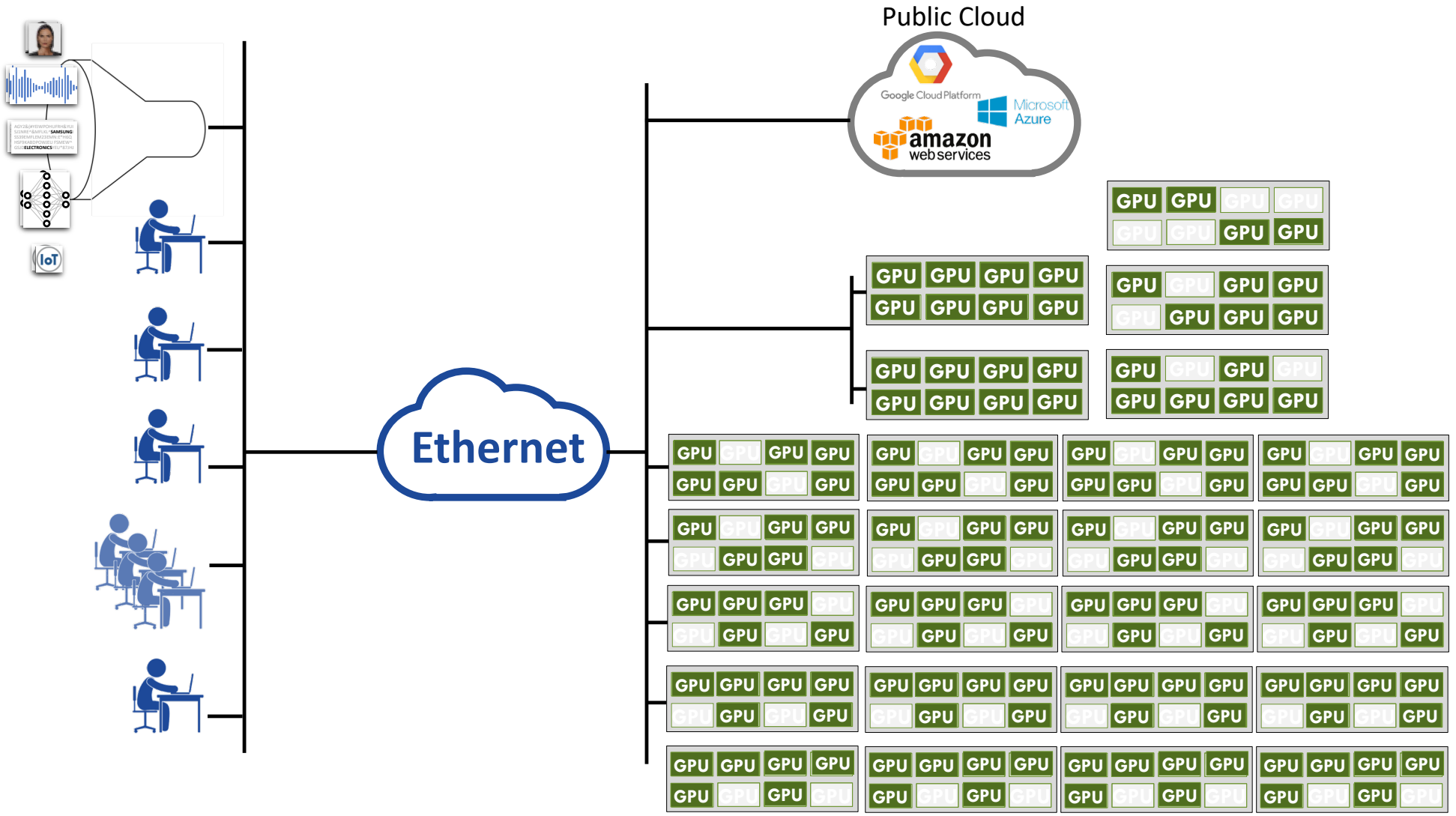

Production Cluster
GPU
GPU
GPU
GPU
GPU
GPU
GPU
GPU


GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU


GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU


GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU


GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU
GPU GPU GPU GPU


Public Cloud
Google Cloud Platform
Microsoft Azure
amazon web services
Ethernet
IoT
GPU

Any Application/Framework

Any OS/Hypervisor

Any Network

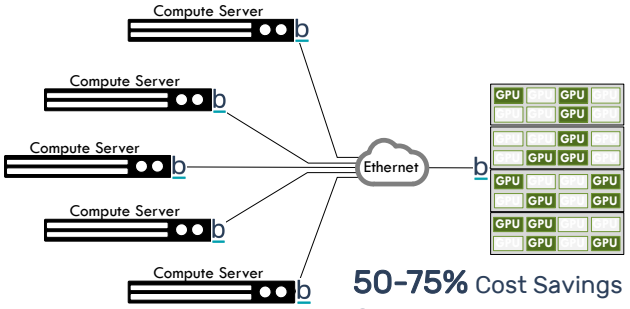
Any Cloud

Any Scale

bitfusion - The Elastic AI Company


INTEL
INTEL
INTEL
CPU
ASICS
FPGA
GPU
NVIDIA
Dave Simonds

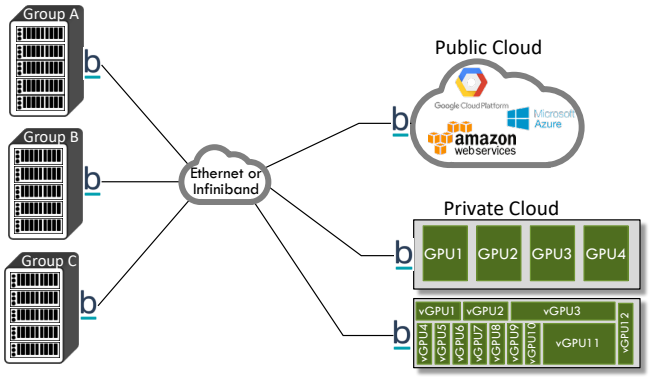
Dynamic and Remote Attach



50-75% Cost Savings
2-4x Better Utilization

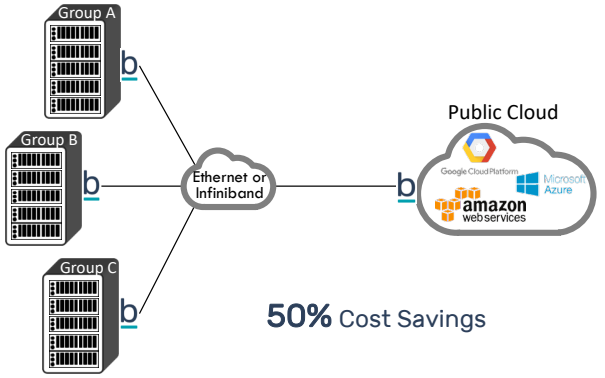
Bitfusion enables compute servers to attach dynamically at application runtime, only when GPU resources are needed to remote GPUs. Multiple Compute servers can attach dynamically to a single or many GPU server. Dynamically attach and detach allows Server consolidation, increase 2x-4x utilization of GPU (or serving 2x-4x more users)

Private & Public Cloud Support



Bitfusion runs in user space, and supports public and private cloud configurations. Many use cases, allow peak GPU demand to burst to the public cloud and extend GPU elasticity in private cloud to public cloud

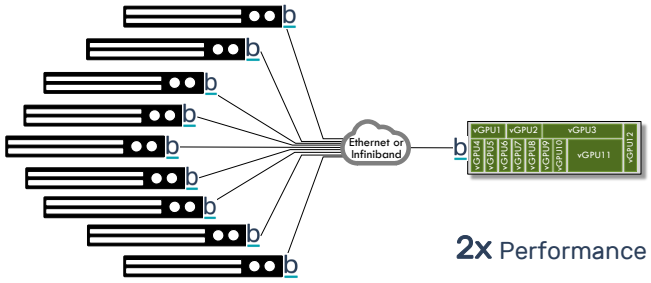
Optimization of Multi Cloud



50% Cost Savings

Bitfusion runs in user space, and support multi-cloud (public and private). Users can partition GPU server instances to many vGPU and serve more users and applications. Users can pair any compute and GPU instances.

Partial (Virtual) GPUs



2x Performance

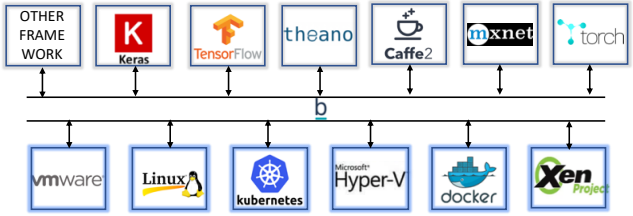
Bitfusion partition physical GPUs to partial virtual (not equal) GPUs, Remote (or local) compute servers can remotely attach to the partial GPUs dynamically. Partial GPUs have demonstrated highest throughput when workload don't consume all GPU memory and/or compute. Many inference applications have a good fit to this topology. Bitfusion can partition a GPU to non-equal parts to allow experimentation and optimization.

GPU Server Consolidation

4x Better Utilization

Bitfusion dynamic attach and detach, and auto-shutdown allows GPU server consolidation, and deliver 4x better utilization (and cost savings). It will translate to similar savings of Opex and Energy savings.

Any Hypervisor, Container or OS



No Management, Provisioning, Training or Administration Costs

Bitfusion is a transparent software layer that supports any upstream AI framework or container environment and any downstream hypervisor, OS or bare-metal. Bitfusion provides flexibility to the software environment while providing the cost, performance and utilization multipliers