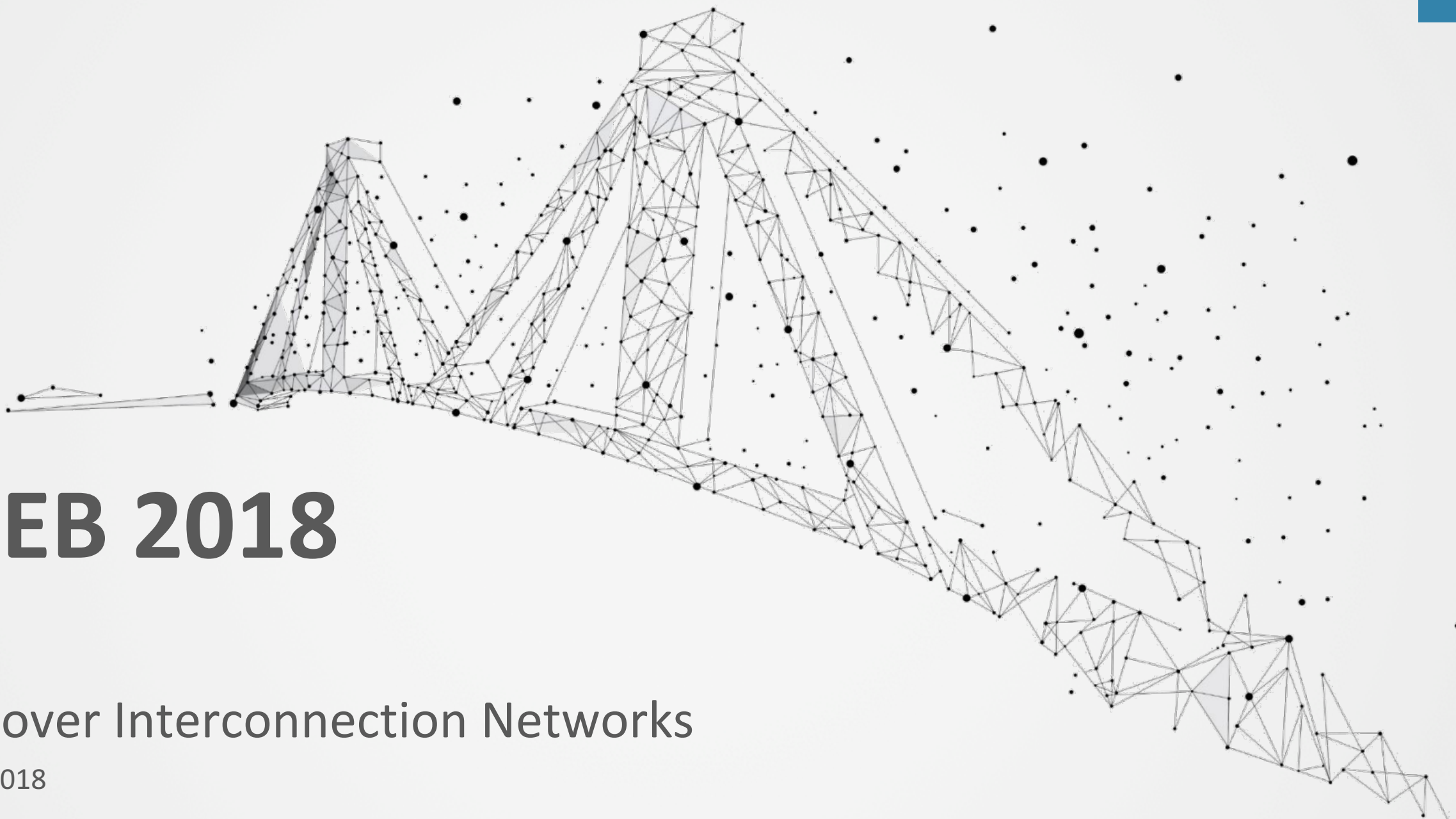




HiPINEB 2018 Panel

AI and ML over Interconnection Networks

Eitan Zahavi, Feb 2018

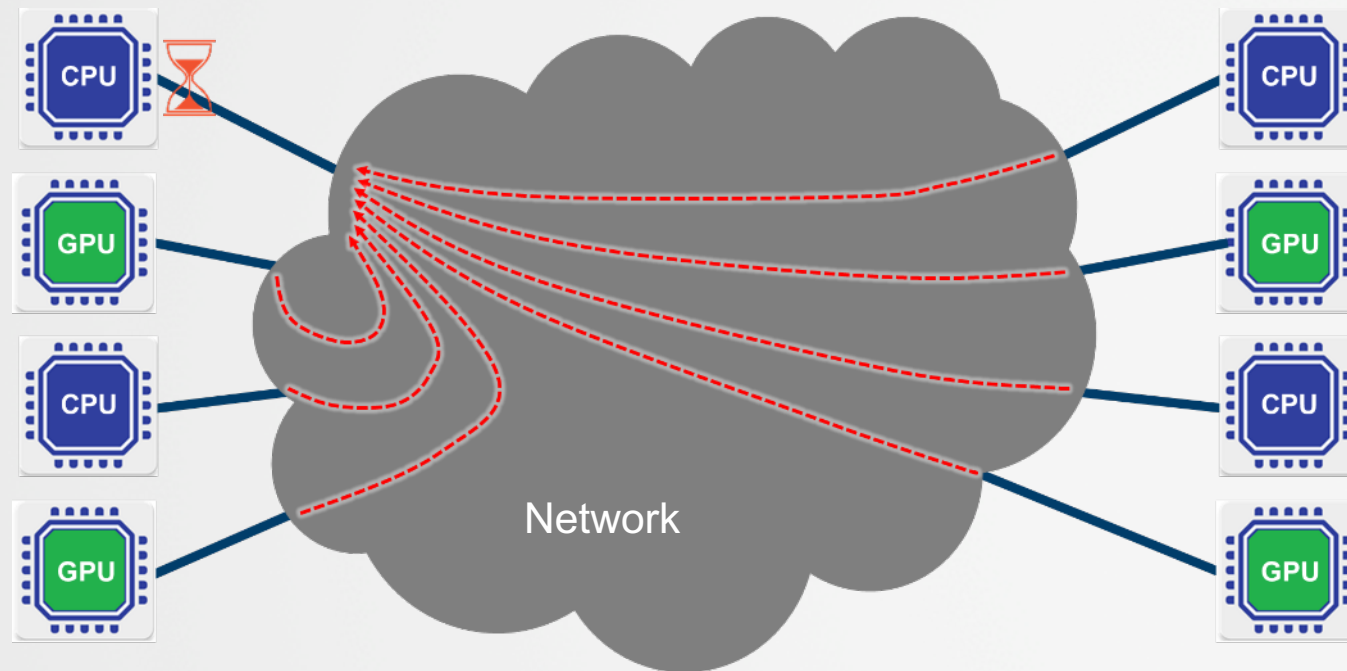
Question 1:

Interconnection networks vs. Eth

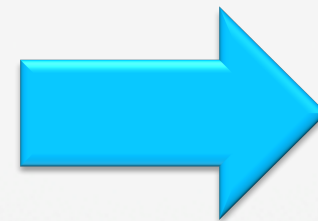


InfiniBand/RoCE Offloading vs. On-loading Ethernet

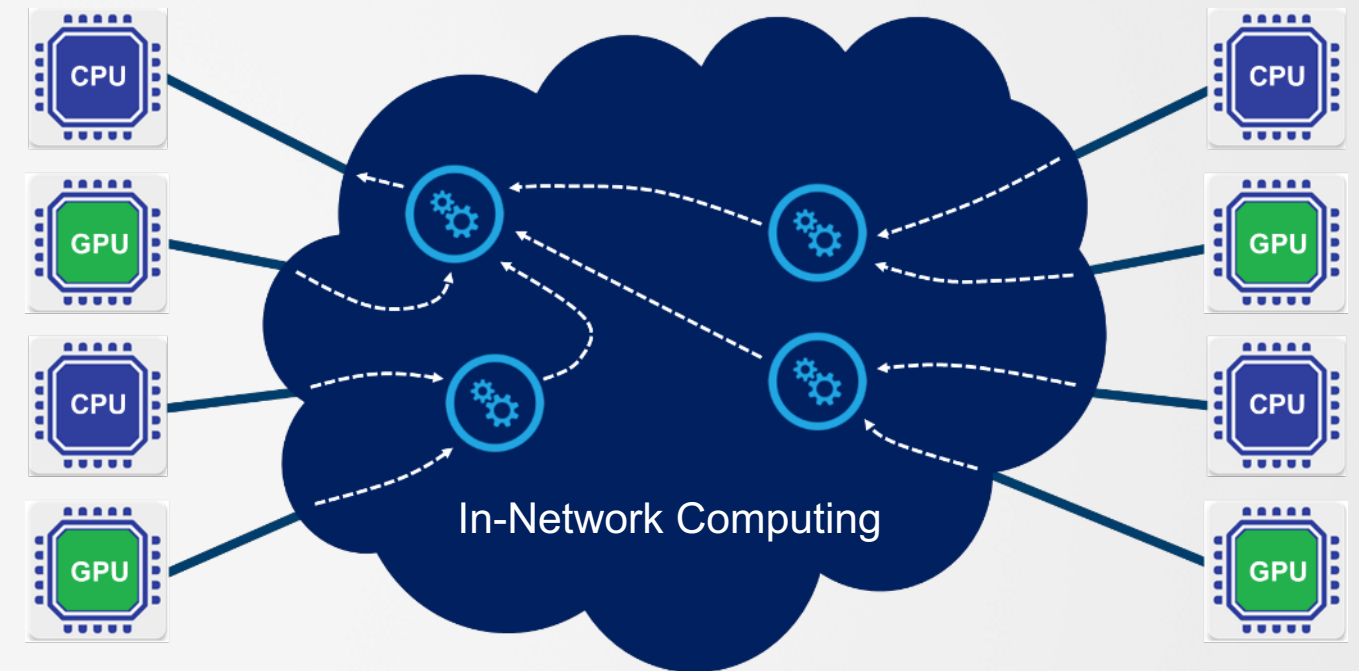
CPU-Centric (Onload)



Onloading Communications Latencies
of 30-40us



Data-Centric (Offload)



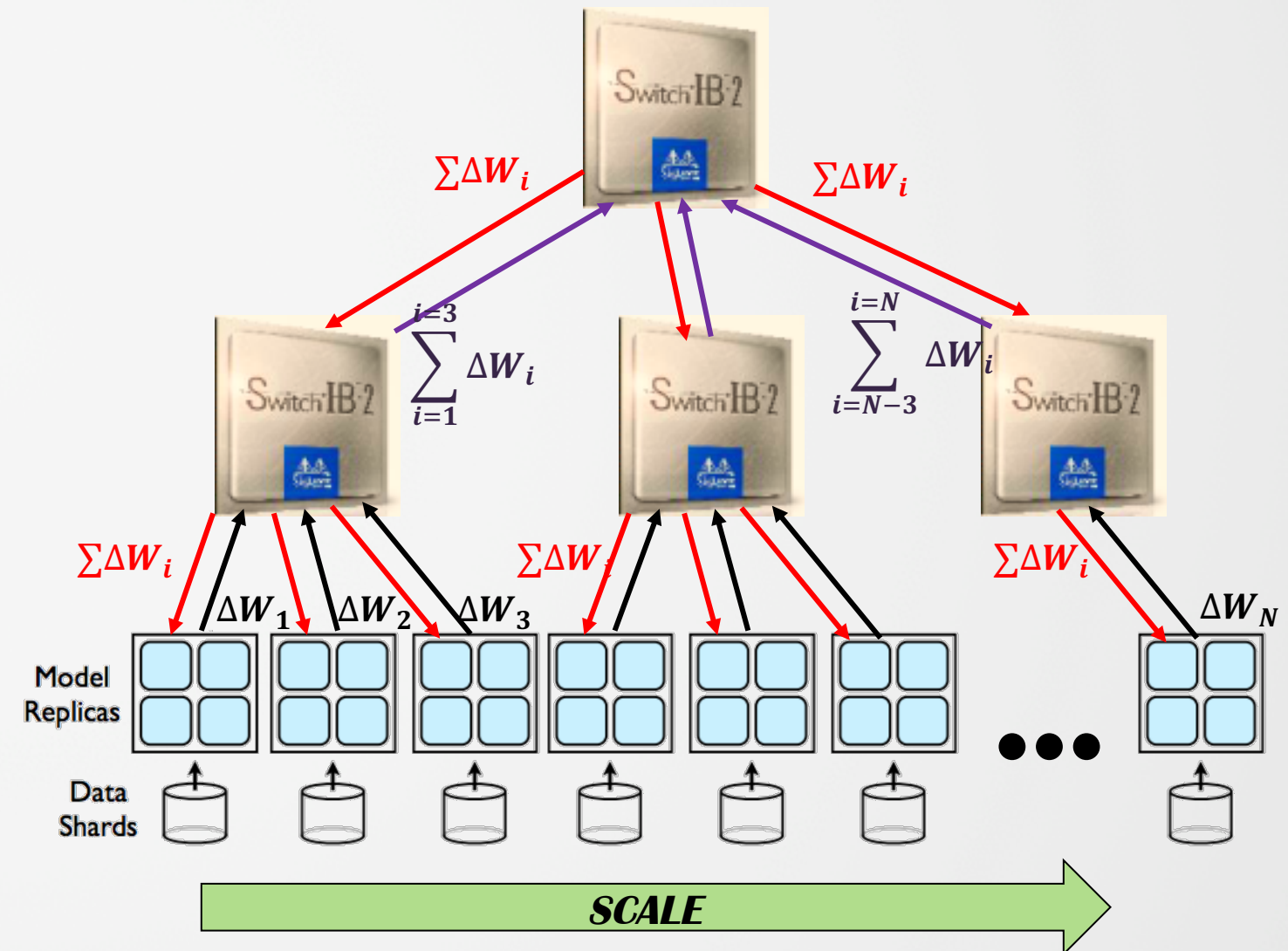
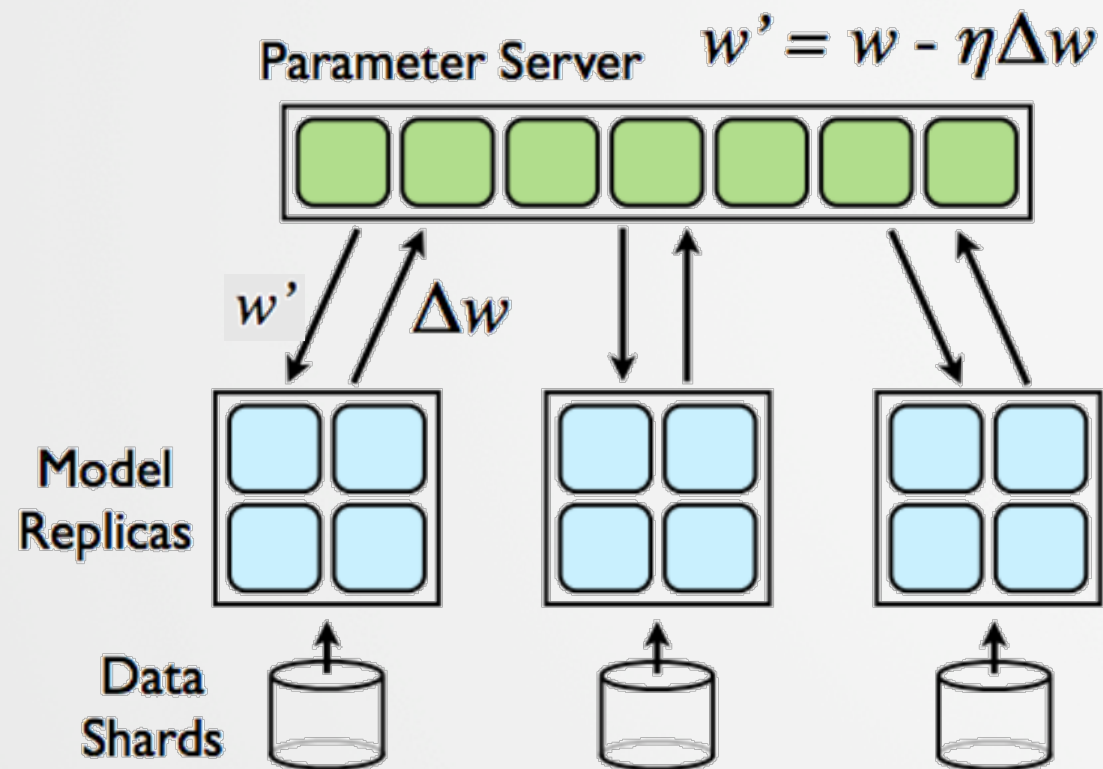
HPC / Machine Learning
Communications Latencies of 3-4us

Reduce Network Load and Improve Responsiveness

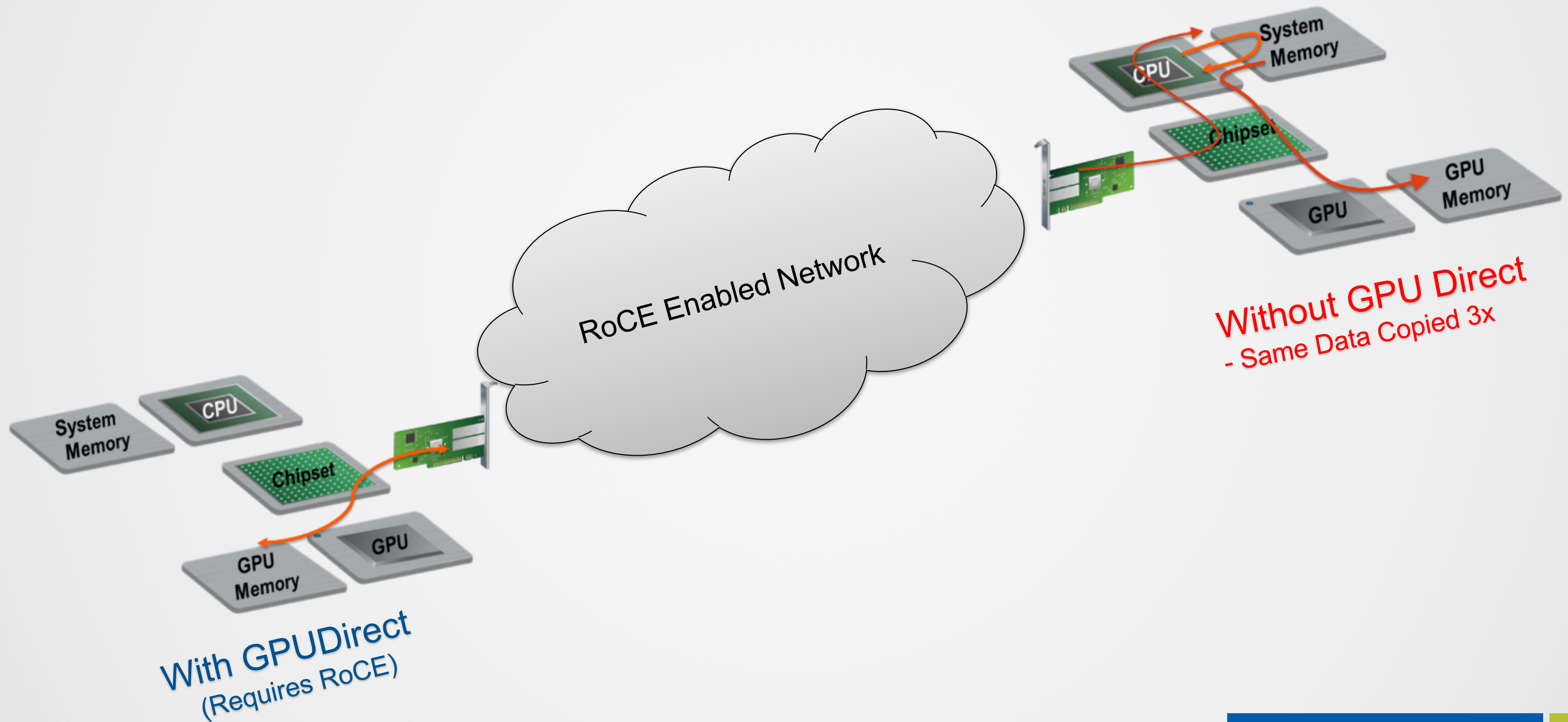
ML Offloading for Parallel Learning

- Bottleneck created by Parameter Server

- SHARP Reduces Network Load and Latency



GPUDirect™ RDMA Technology



ML/AI Parallelism via GPUDirect RDMA

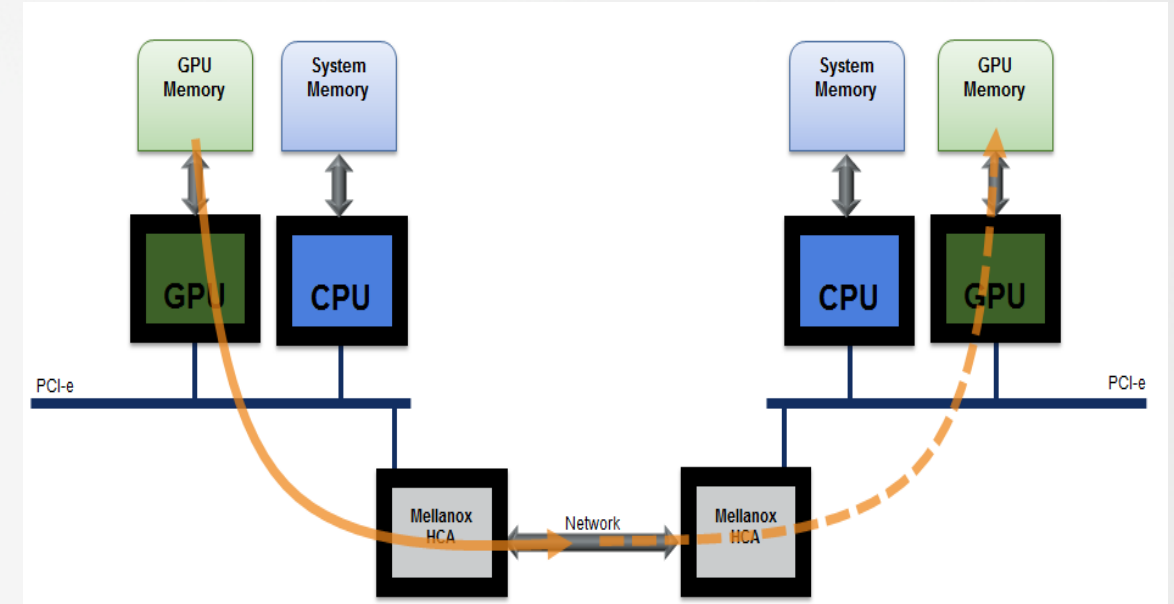
10X

Higher Performance
with GPUDirect™

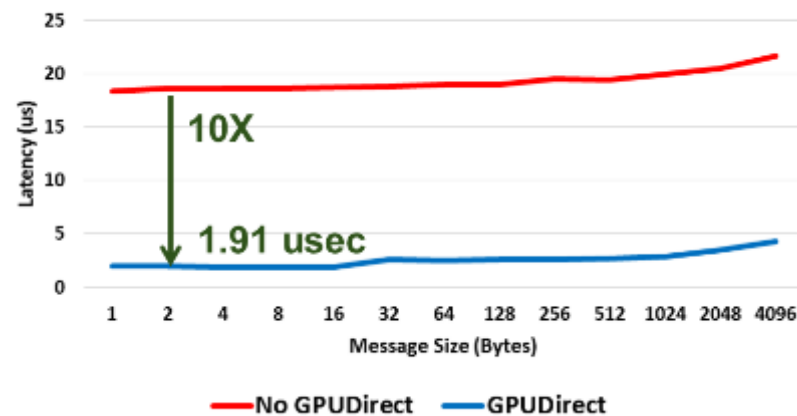


Source: Prof. DK Panda

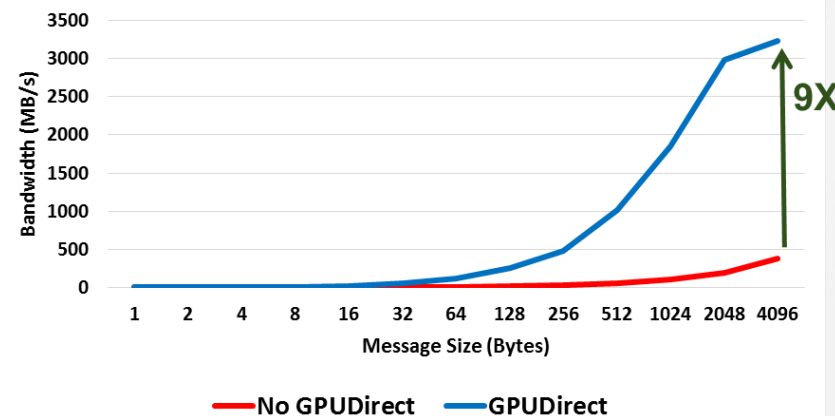
GPUDirect™ RDMA, GPUDirect™ ASYNC



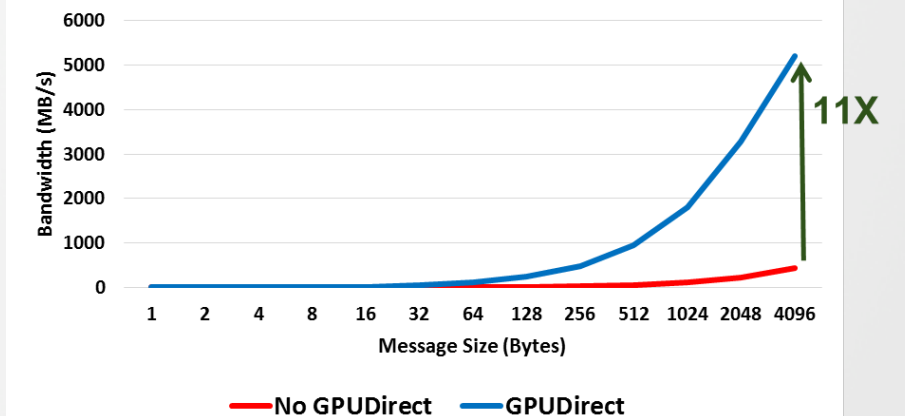
GPU-InfiniBand-GPU Latency



GPU-InfiniBand-GPU Throughput (Uni-Dir)



GPU-InfiniBand-GPU Throughput (Bi-Dir)

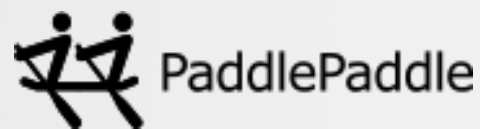


Question 2:

Interconnection Networks and ML/AI

RoCE Unlocks the Power of Artificial Intelligence

AI Software Frameworks



Leading AI Solutions at Scale



8x **V100**



128x **DGX-1'**



27,000x **V100**



60%
Higher ROI

50%
Lower CapEx
& OpEx

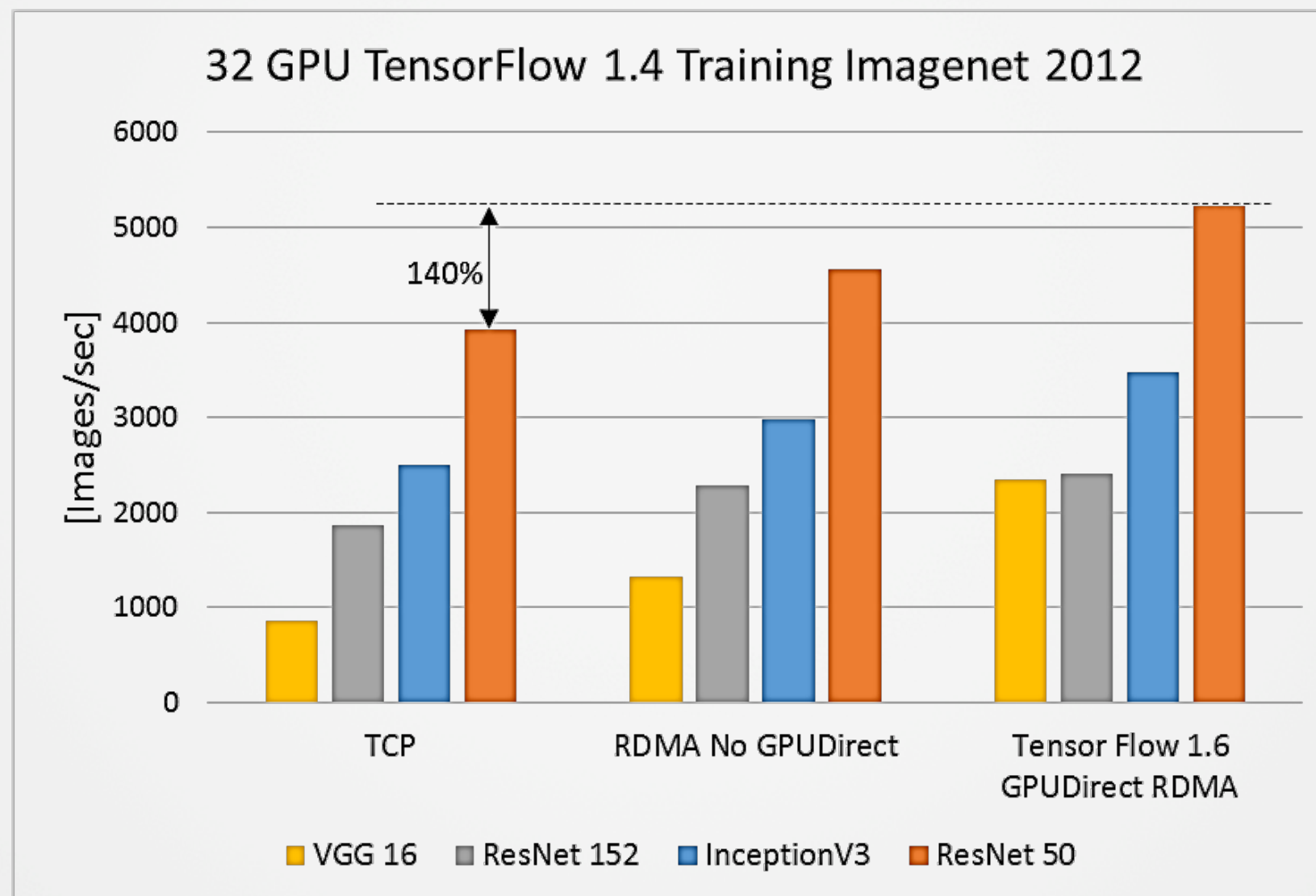
Question 3:

How Network Bandwidth Impacts ML?



Its not the BW but what you do with it!

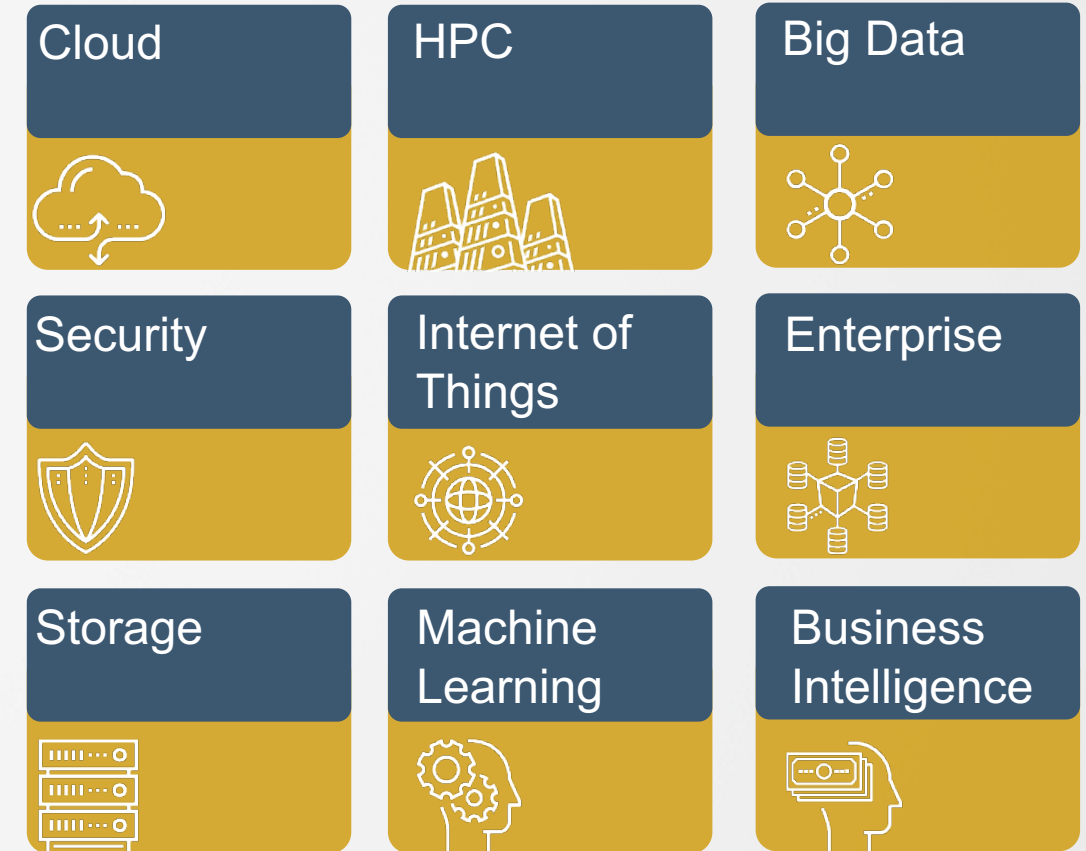
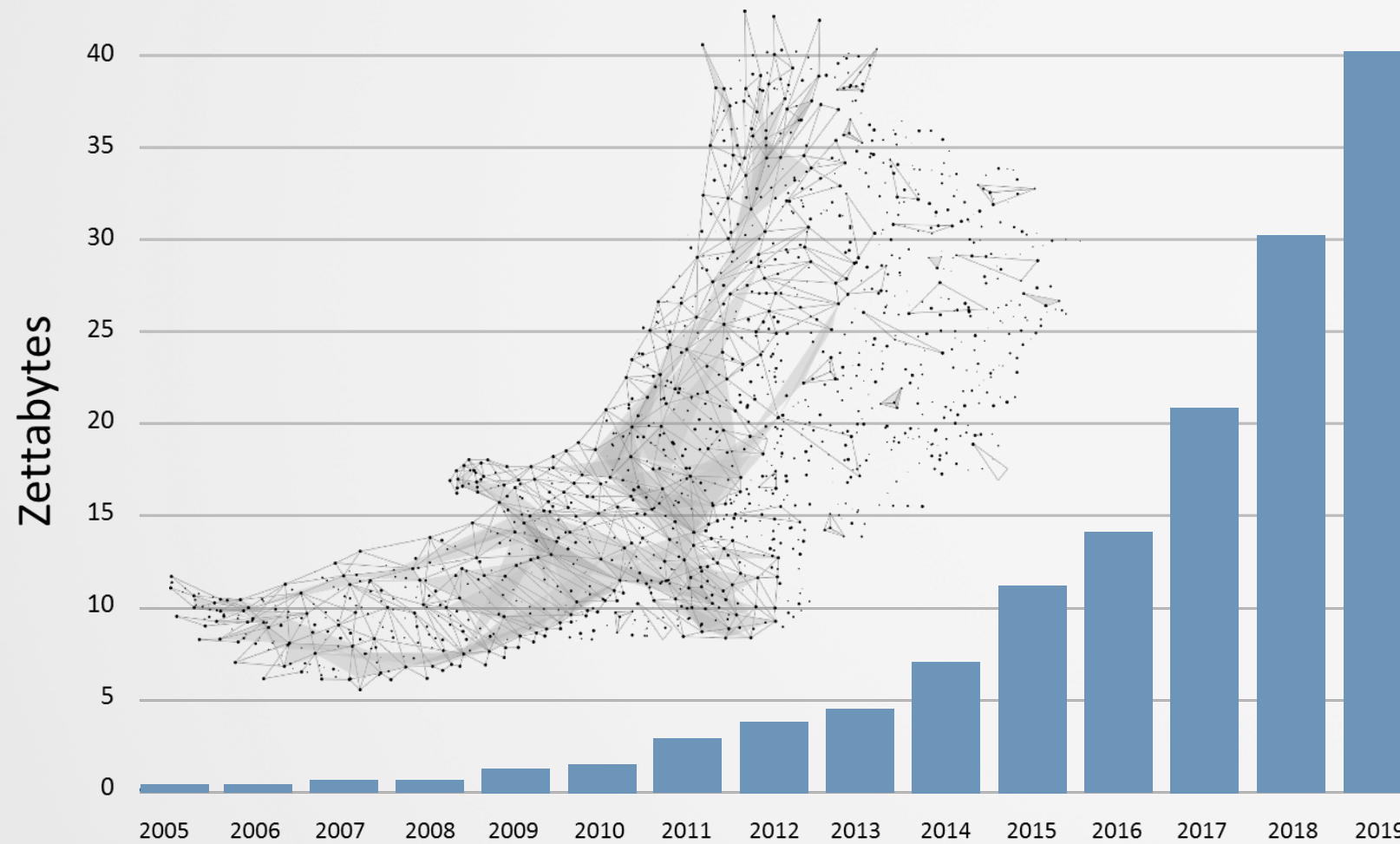
RDMA and GPUDirect makes the difference



Generally !

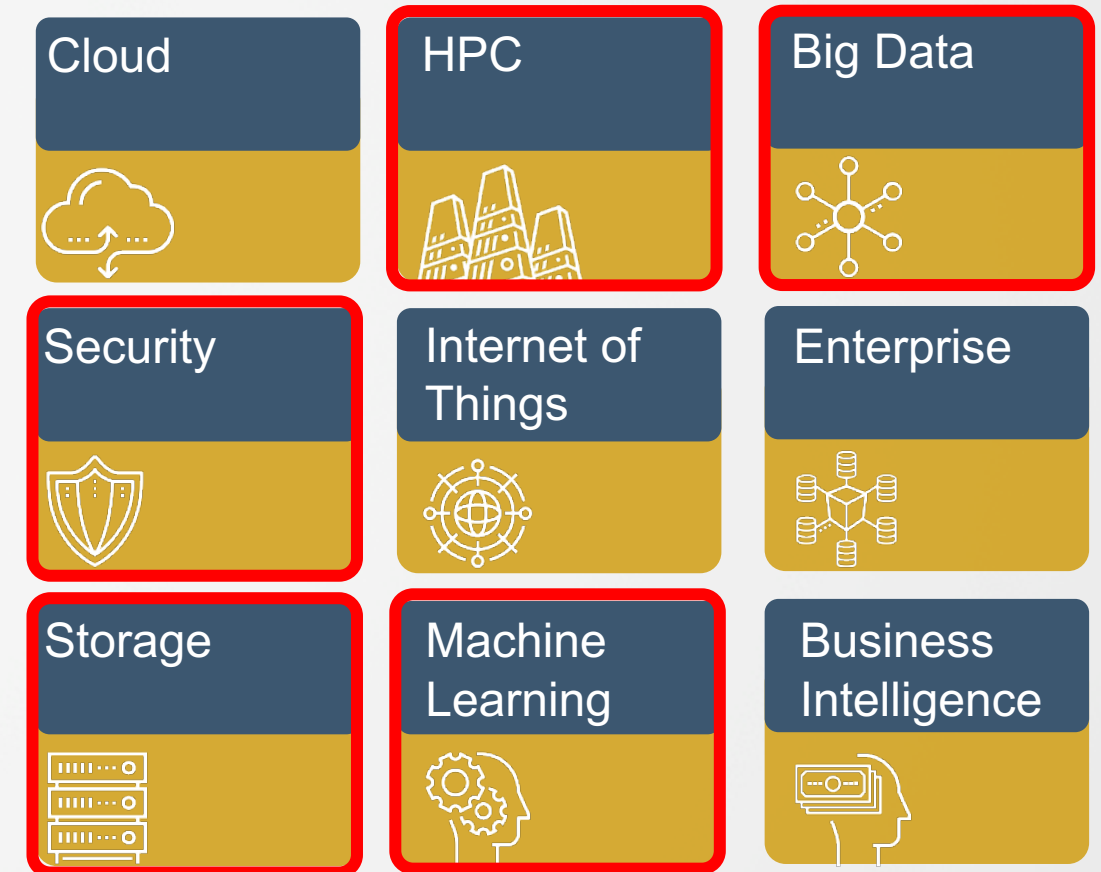
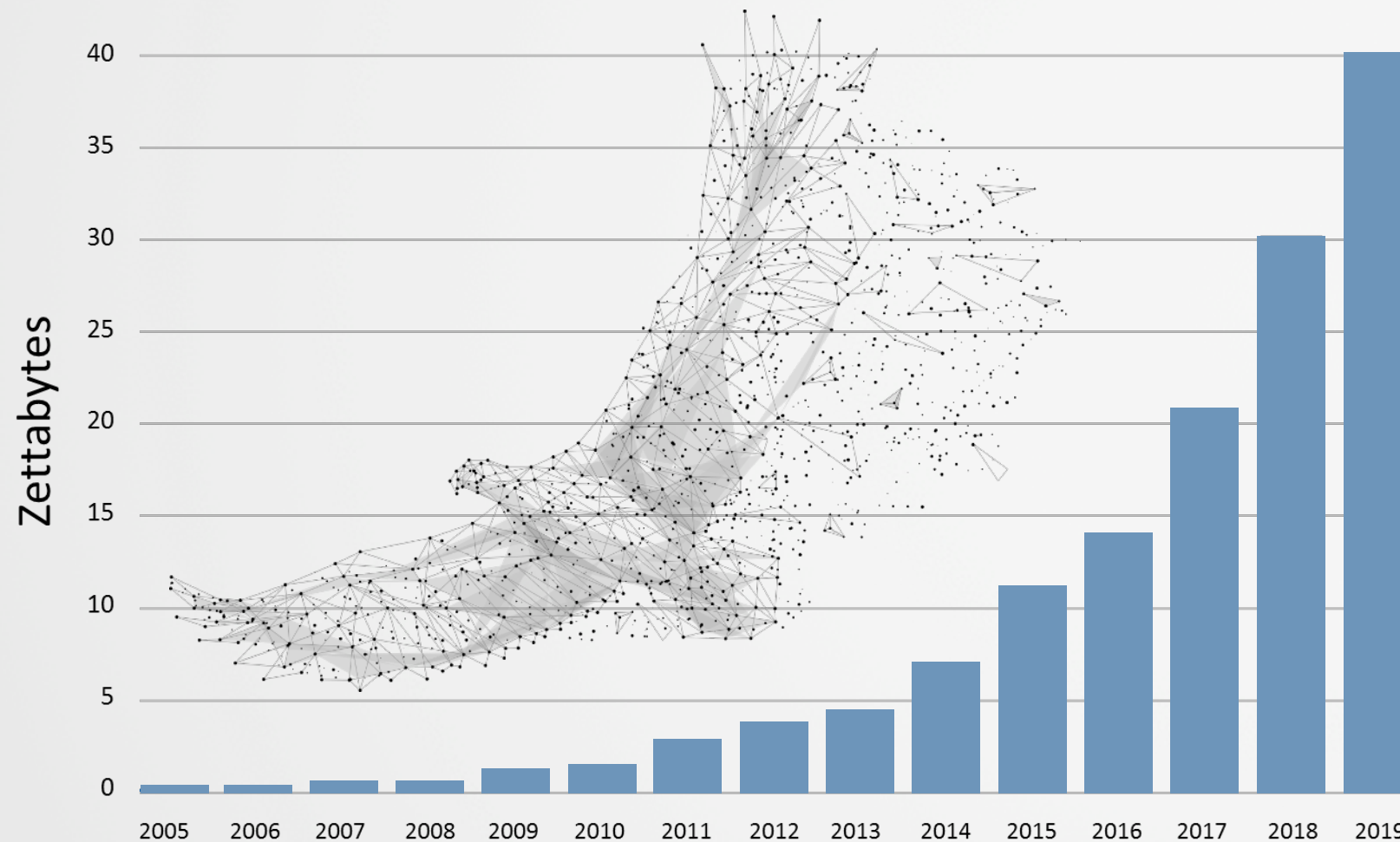
A large yellow square and a smaller blue square are positioned below the word 'Generally'.

Exponential Data Growth Everywhere



Exponential Data Growth Everywhere

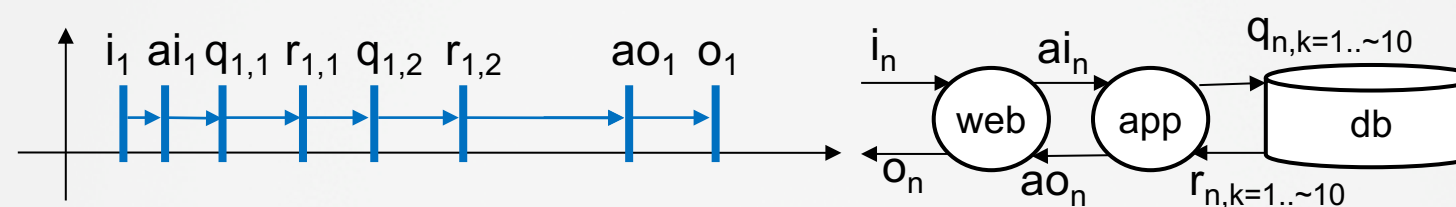
Parallel Applications Everywhere



Traffic Correlation of Parallel Applications

average network load

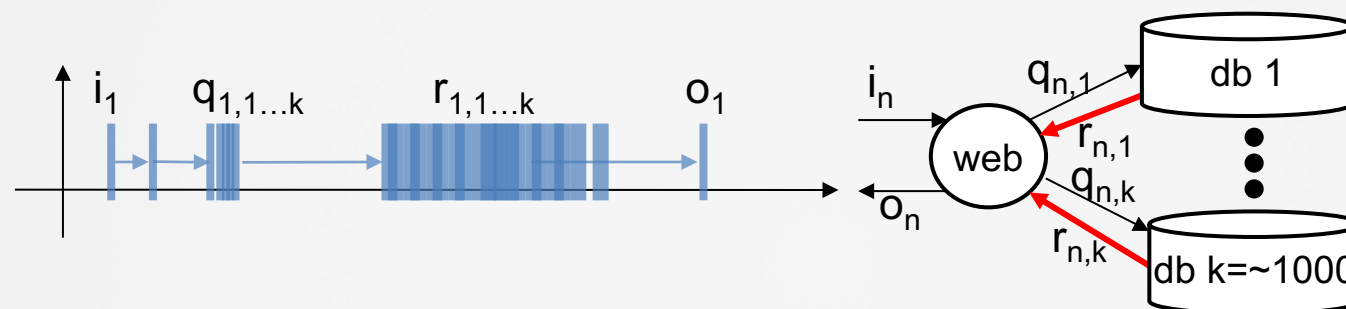
- Traditional 3-Tier Applications



Traffic Correlation of Parallel Applications

- Traditional 3-Tier Applications
- Big-Data distributed database

average network load
incast, latency tail



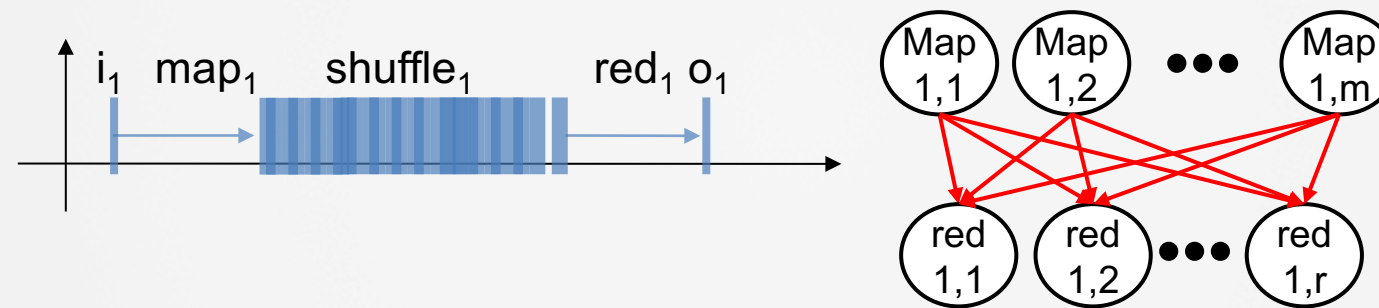
Traffic Correlation of Parallel Applications

- Traditional 3-Tier Applications
- Big-Data distributed database query
- MapReduce

average network load

incast, latency tail

incast, last reducer



Traffic Correlation of Parallel Applications

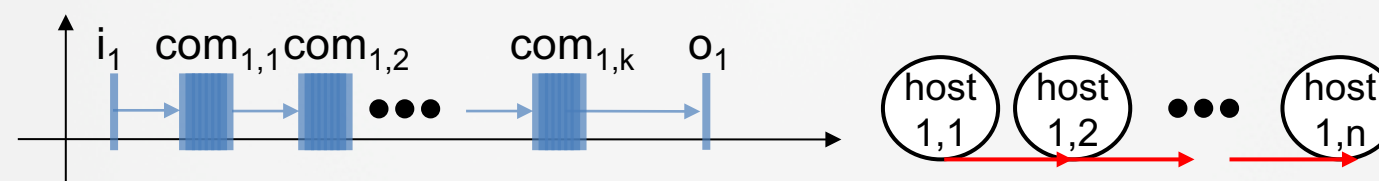
- Traditional 3-Tier Applications
- Big-Data distributed database query
- MapReduce
- Bulk Synchronous Paradigm programs (HPC)

average network load

incast, latency tail

incast, last reducer

increased runtime



Traffic Correlation of Parallel Applications

- Traditional 3-Tier Applications
- Big-Data distributed database query
- MapReduce
- Bulk Synchronous Paradigm programs

average network load

incast, latency tail

incast, last reducer

increased runtime

Parallel Applications result in Correlated Traffic
The end of statistical network sharing



Thank You

