Feeding Data Hungry GPUs with Networked Flash

Chris Lamb
VP Compute Software at NVIDIA
Michael Kagan
CTO at Mellanox
FEEDING DATA HUNGRY GPUS WITH NETWORKED FLASH

Chris Lamb, VP Compute Software, NVIDIA
TWO FORCES SHAPING COMPUTING

40 YEARS OF CPU TRENDS DATA

Single-threaded perf

1.5X per year

10^7

10^8

10^9

10^10


GPU Computing

1.1X per year

30-year era of Moore’s law is ending
Machines can learn with AI, but need massive computer power
Optimizing GPU systems, software, and algorithms is the path to 1,000X speed-up

ALEXNET: THE SPARK OF THE MODERN AI ERA
EXPONENTIAL GROWTH IN COMPUTING DEMAND

DATA SIZE GROWING

AI RESEARCH GROWING

AI MODEL COMPLEXITY GROWING

Source: IDC, GitHub, and OpenAI / NVIDIA
CUDA-X AI Addresses the End-to-end Development of AI | Optimized For Every AI Model and Framework
Available in Every Cloud and from Every Computer Maker
2.4 TB
Bisection Bandwidth
GPU OPTIMIZED SYSTEMS
With GPUDirect Storage
GPU OPTIMIZED DATA CENTERS
Clusters with GPUDirect Storage
CUFILE AND GPUDIRECT STORAGE

Architecture of the Stack

cuFile API
For applications

NVFS Driver API
For filesystem, block, and storage drivers

Application

cuFile API

CUDA

NVFS Driver

Filesystem Driver

Block IO Driver

Storage Driver
BANDWIDTH SCALES TRANSPARENTLY

Architectural Peak Bandwidth
EXTREME FILE IO BANDWIDTH

Up to 10x Bandwidth Per Core

FIO + GDS extension on EXT4 Filesystem - synchronous ordered IO with 64 job per device at queue depth 1
SCALED APPLICATION PERFORMANCE
Up to 20x Faster with GPU Accelerated TPC-H Query 4 Scaling on EXT4
FOR MORE INFORMATION

Join the GPUDirect Storage interest list in order to:
Provide feedback
Extend with other filesystems

Technical blog and link to sign up: https://devblogs.nvidia.com/gpudirect-storage/
Michael Kagan, CTO at Mellanox
Mellanox Networking Delivers the Solution
Industry Leading Ethernet & InfiniBand End-to-End
25, 40, 50, 56, 100, 200Gb/s (and soon 400Gb/s)
12.5GB/s of Storage to Every Tesla Pair

Internal NVMe SSD storage
25GB/s to Every Tesla Pair

External IB or Ethernet attached NVMe JBOFs
Unlimited High Performance Storage when Networked

**Plus: Storage Networking Advantages over Local Storage for GPUs**

- Unlimited capacity
- High Availability
- Higher Utilization
- Lower TCO
  - Less Power
  - Less Rack Space
  - Easier Management
  - Less Cost

DGX

Dual port 200Gb HCA or NICs

200Gb IB or Ethernet Switches

NVMe-oF All Flash Arrays
Now Storage & GPUs can Scale Independently

All because of Mellanox high performance ultra low latency storage networking
NVMe, NVMe-oF & RDMA Protocols

The Network and the Protocol MUST get faster.

Networked Storage

Storage Media

Storage
Protocol (SW)
Network

Protocol and Network

HDD SSD PM

0.01 1 100

Software
Target Software

Memory

RoCE/IB NIC

Hardware
NVMe Flash

SW-HW communication through work & completion queues in shared memory

© 2019 Mellanox Technologies | Confidential
RDMA Performance

Efficient Data Movement (RDMA)

Application
Buffer → Network → Buffer
Kernel Bypass
Protocol Offload

User
Application
Sockets → TCP/IP → Driver
RDMA
Hardware

Microsoft Storage Spaces Throughput

With RDMA
- 2x Better Bandwidth
- Half the Latency
- 33% Lower CPU

See MS demo: https://www.youtube.com/watch?v=u8ZYhUjSUol

No RDMA
100GbE
RoCE
Industry Wide RDMA Adoption

NFS
SMB (CIFS)
File
SMB Direct
NFSoRDMA
RDMA
RPM
Persistent Memory

iSCSI
iSER
NVMe-OF
Block

Ceph over RDMA
Object
Ceph
NVMe-oF Performance with RDMA

- More Bandwidth
- Less Latency
- Less CPU

- 5M IOPs, 4K block side
- ~3usec latency
- 0.01% CPU utilization
Offloads

- Compression
- Security
- Others...

10x Improvement With Offload

- Hardware Compression
- Software Compression

Encryption Allocated Cores

- 75% CPU Overhead

Only 16% CPU Overhead
Solutions

Pure Storage FlashBlade

40GbE

100GbE

40GbE

(4) NVIDIA DGX-1

NetApp A800

100GbE

100GbE

(5) NVIDIA DGX-1

EMC/Isilon F800

40GbE

100GbE

(9) NVIDIA DGX-1 Servers
Future Network Performance

Ethernet

InfiniBand

TO TERABIT SPEEDS

**Ethernet**

- 200 Gb/s (2019)
- 400 Gb/s (2019)
- 800 Gb/s (>2020)
- 10/40/100 Gb/s (2010-2017)
- 25/50 Gb/s (2017)

**InfiniBand**

- 100 Gb/s (2013)
- 168 Gb/s (2014)
- 200 Gb/s (2015)
- 300 Gb/s (2016)
- 600 Gb/s (2017)
- 1.2 Tb/s (2018)

**Bandwidth per Direction (Gb/s)**

- 1x
- 4x
- 12x
- 14G
- 56G
- 25G
- 100G
- 200G
- 400G
- 600G
- 1.2T
- 1.6T
- 2T
- 4T

**Link Speed (b/s)**

- 10G
- 25G
- 50G
- 100G
- 200G
- 400G
- 800G
- 1 Tb/s
- 1.6 Tb/s
- 2 Tb/s
- 4 Tb/s
- 6.4 Tb/s
- 10 Tb/s
- 800 Gb/s (>2020)
- 10 Tb/s (>2020)
- 1.6 Tb/s (>2020)
- 6.4 Tb/s (>2020)
- 10 Tb/s (>2020)

**Highly Parallel Speeds**

**Quad Speeds**

**Serial Speeds**

**Mellanox Technologies**

© 2019 Mellanox Technologies | Confidential
Future Storage Advancements

- Ethernet SSDs
- Remote Persistent Memory (RPM)
- Computational Storage

![Diagram of Future Storage Advancements](image.png)

Facebook Example – from OCP 2019

Architectural Options

- MySQL/ZippyDB
- RocksDB
- FS/Kernel
- Kernel
- Drivers

Network PM

© 2019 Mellanox Technologies | Confidential
Thanks You!

Michael Kagan

Chris Lamb
Thanks You!

Chris Lamb
VP Compute Software at NVIDIA
Michael Kagan
CTO at Mellanox