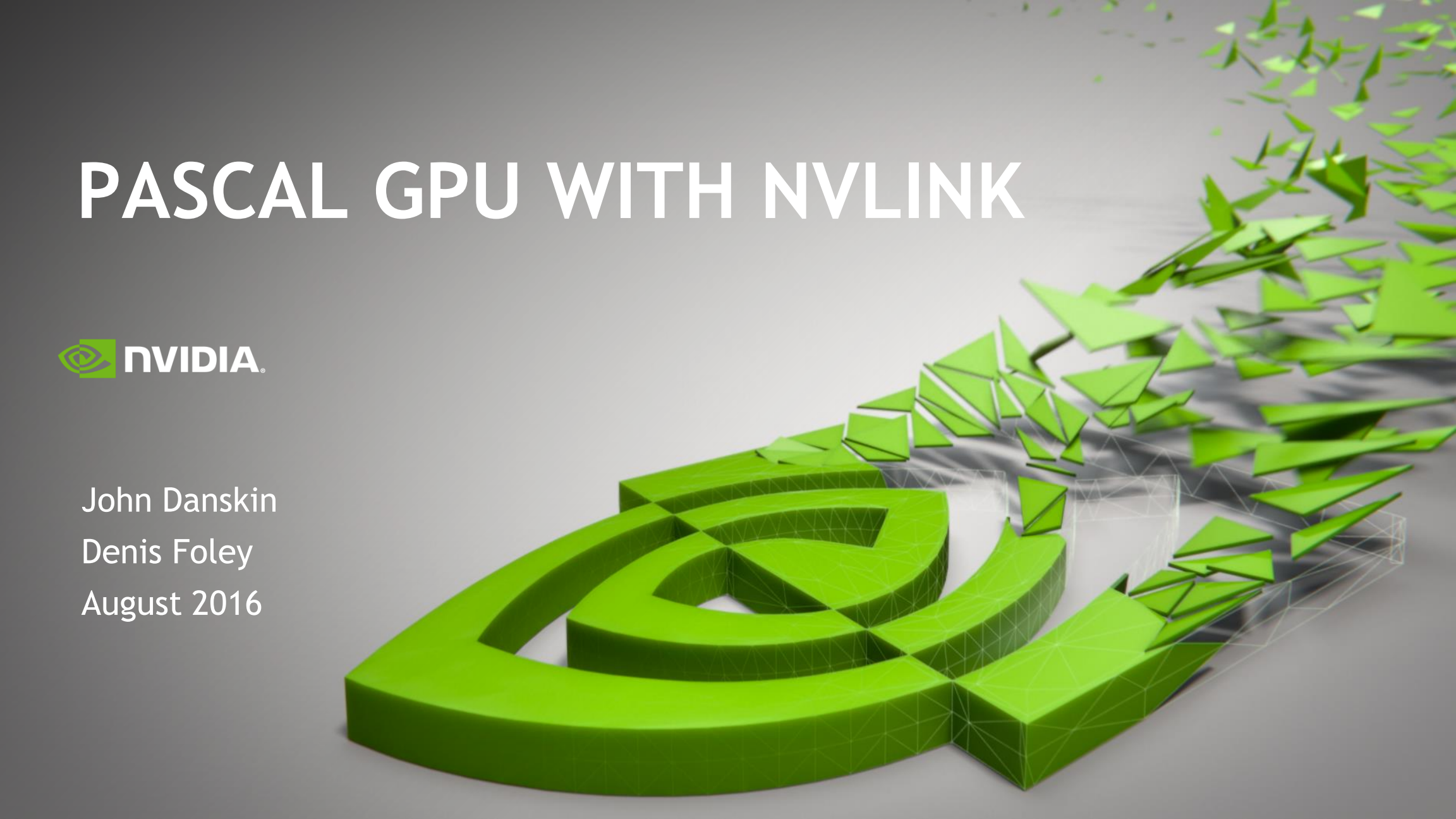
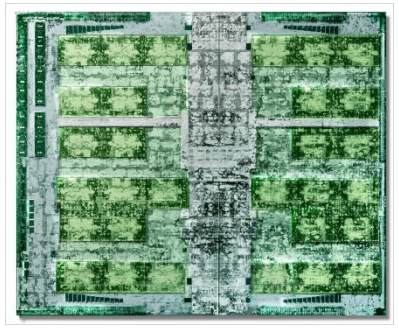
# PASCAL GPU WITH NVLINK



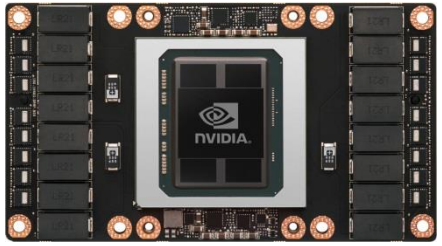
John Danskin
Denis Foley
August 2016

# OUTLINE

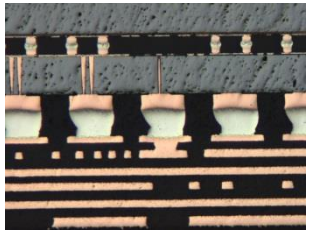
## GP100 Die



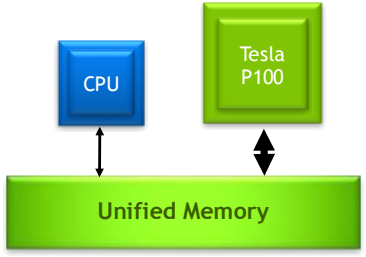
## P100 SXM2 Module



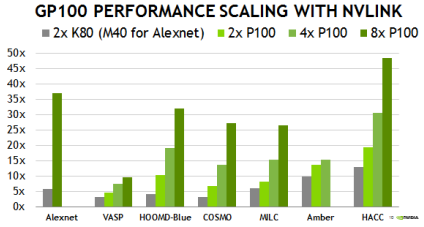

## Stacked Memory & Packaging



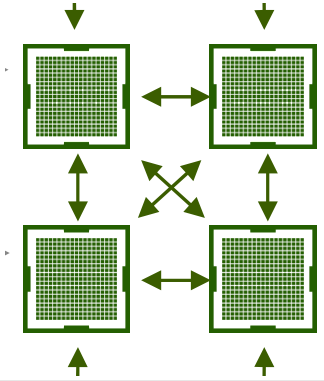
## GPU Features




## Performance




## NVLink



# GP100

610mm$^2$

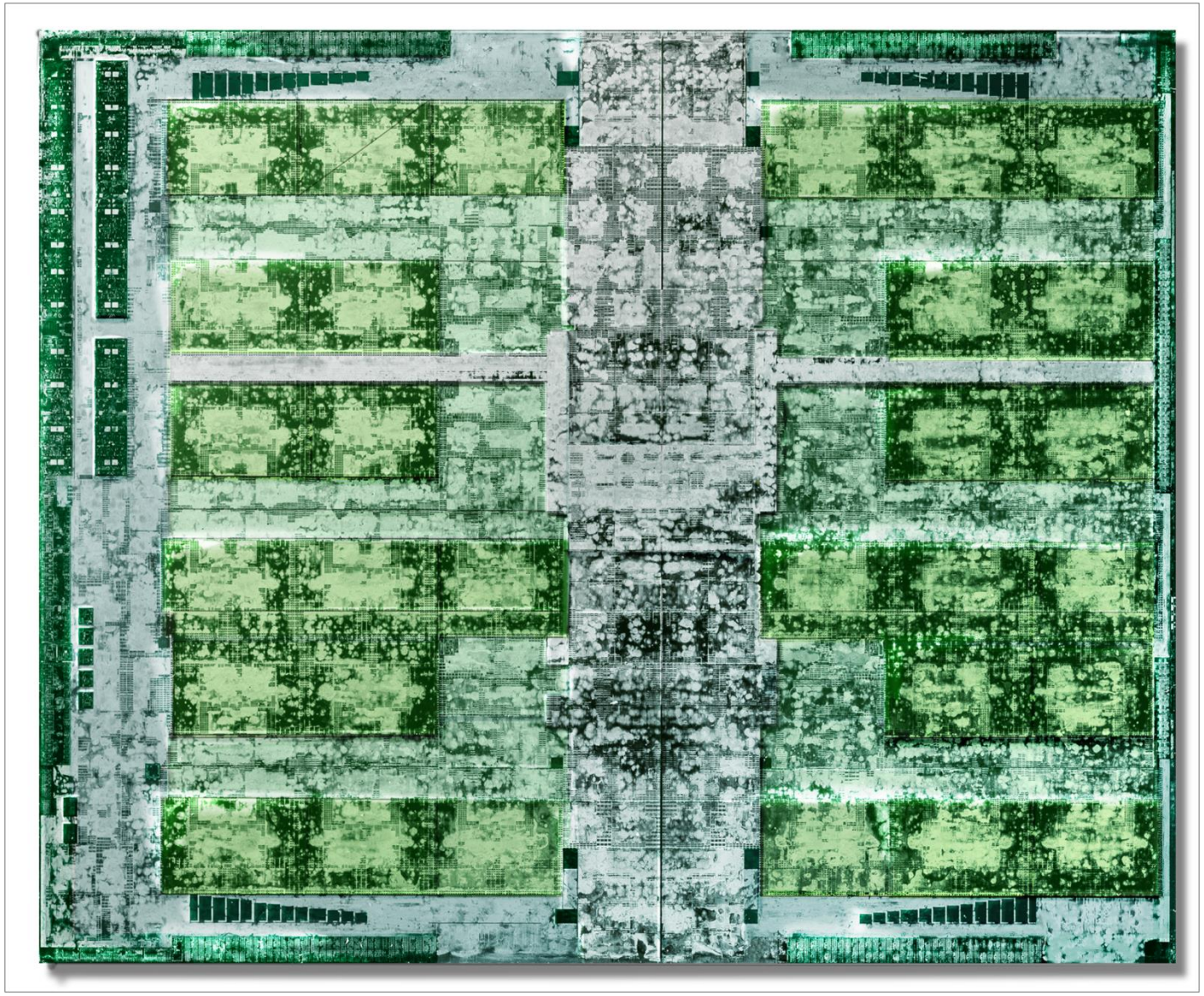
4 x HBM IO

30 SMs (28+2)

4MB L2 Cache

4 x NVLink

16x GEN3 PCIE



# SXM2 MODULE

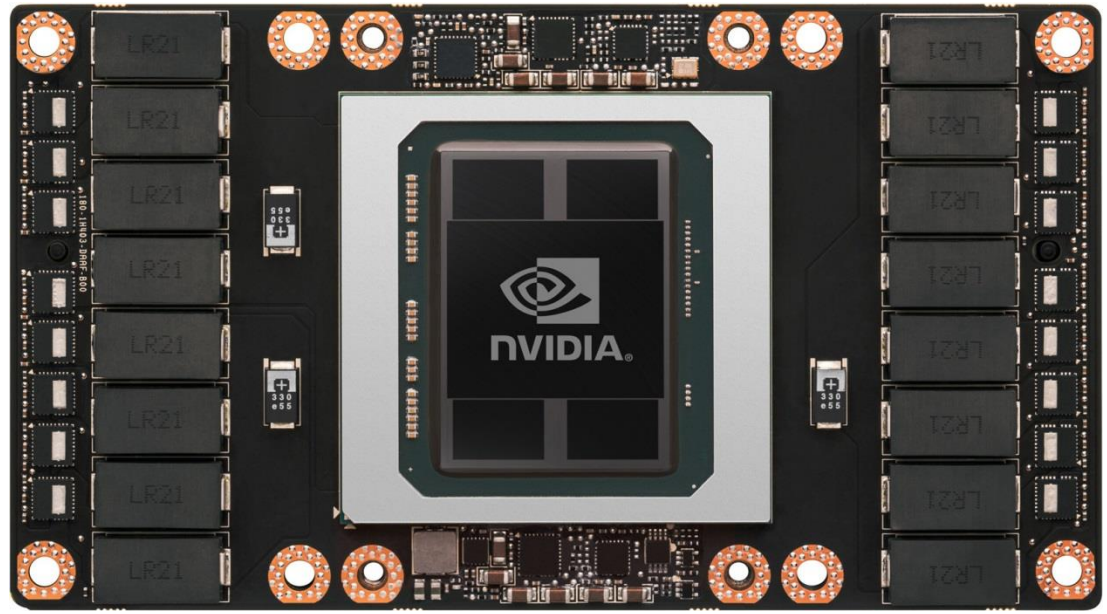
2U with heatsink

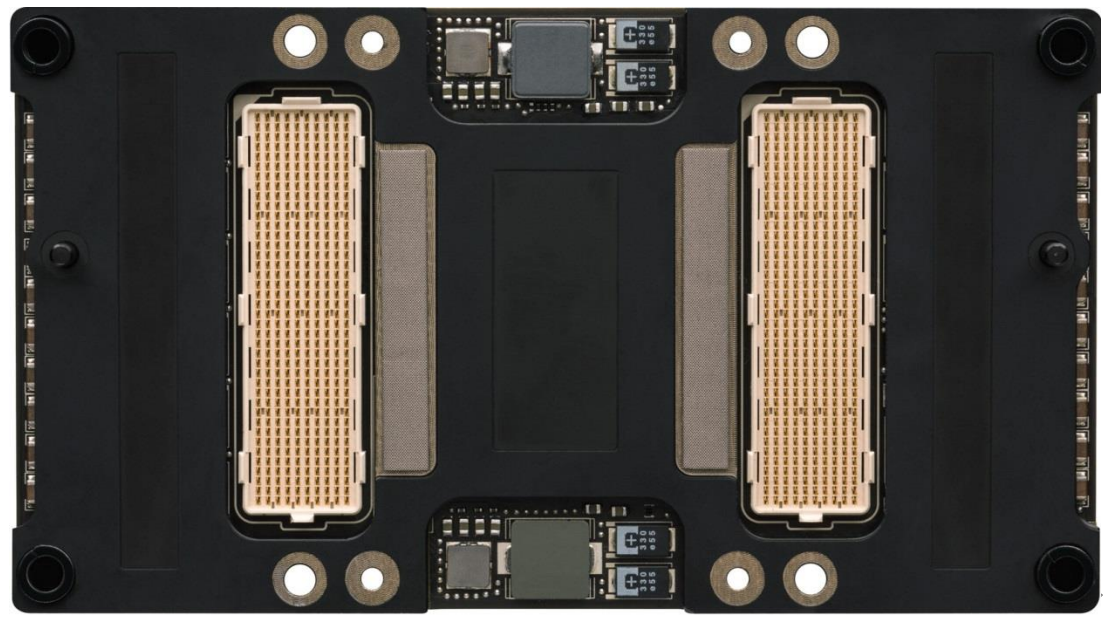
Parallel to motherboard

37% area vs K40 PCIE board

Supports NVLink

300W TDP



# DESIGNING FOR HBM

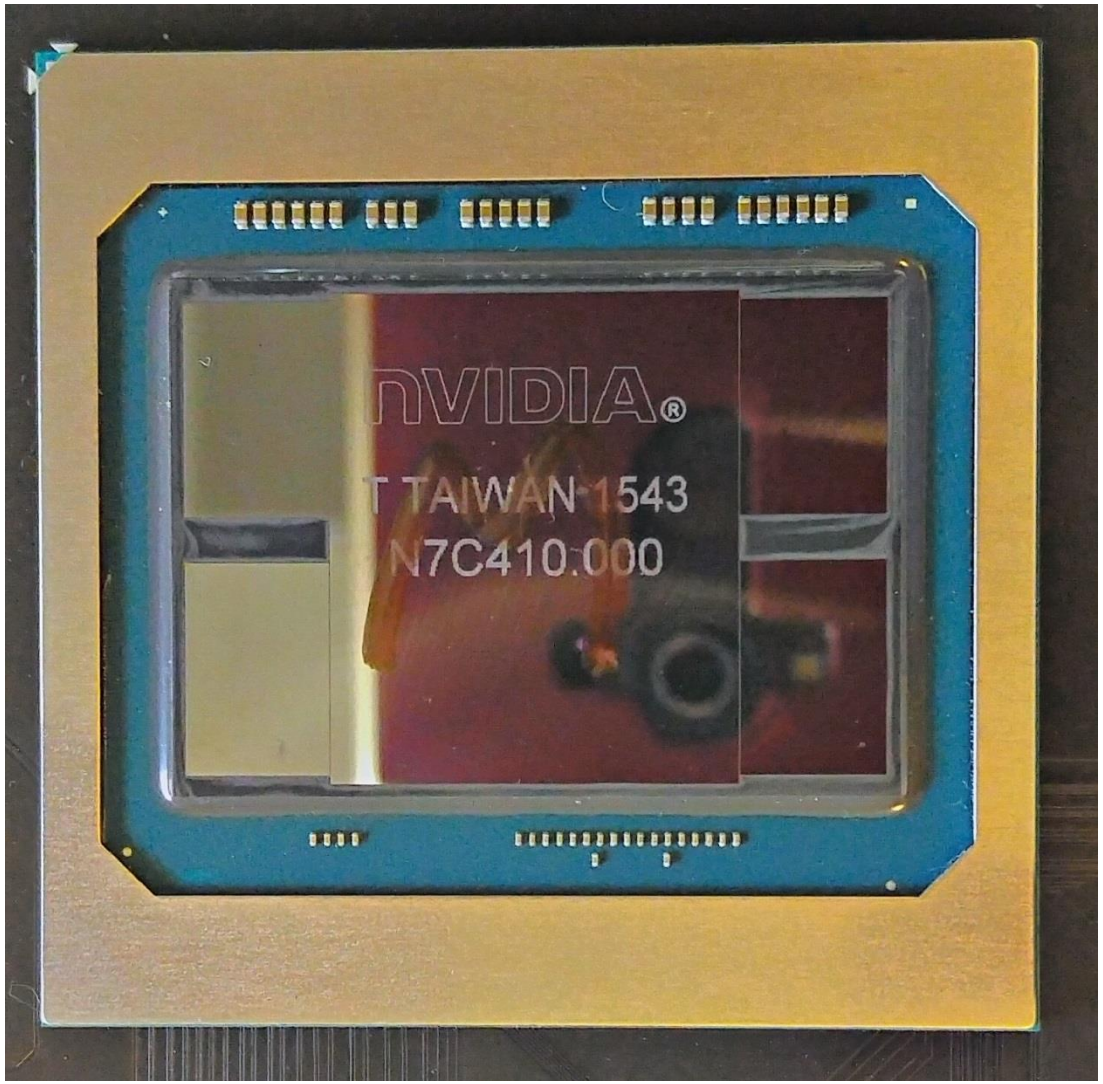
Very short wires

8 memory channels

32 L2$ slices

3TB/s internal bus

HBM bandwidth tracking GPU performance



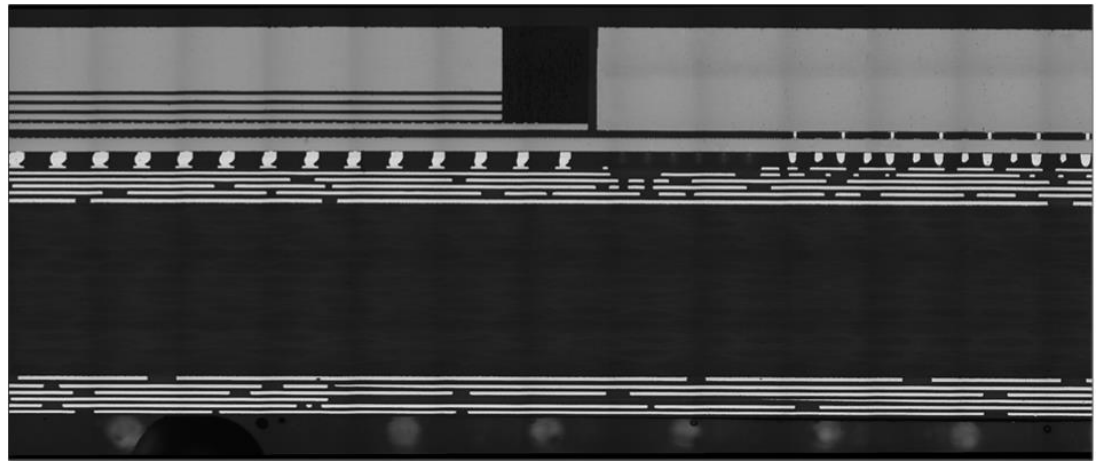
# HBM (HIGH BANDWIDTH MEMORY)

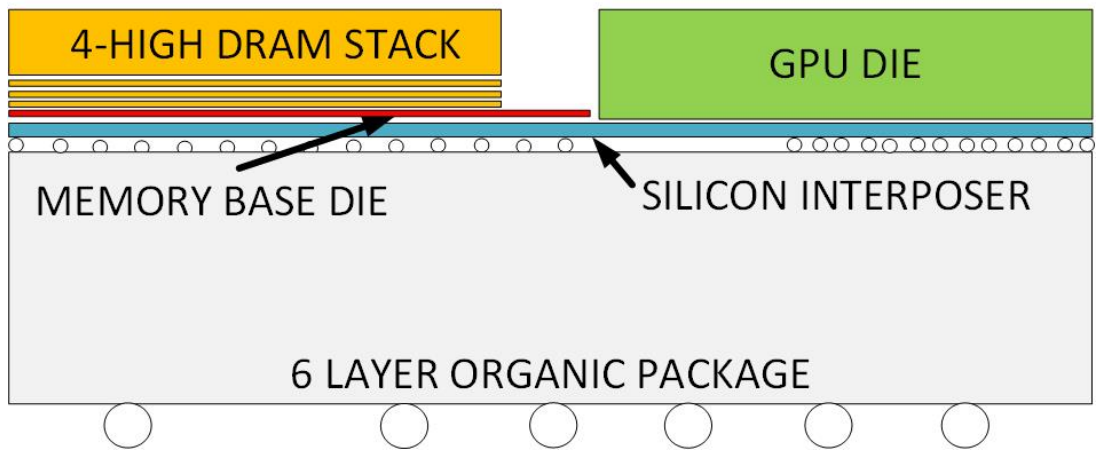
In-package HBM2 memory

4 x 4-High TSV connected DRAM stack on memory base die

DRAM and GPU die on silicon interposer

55mm x 55mm Package





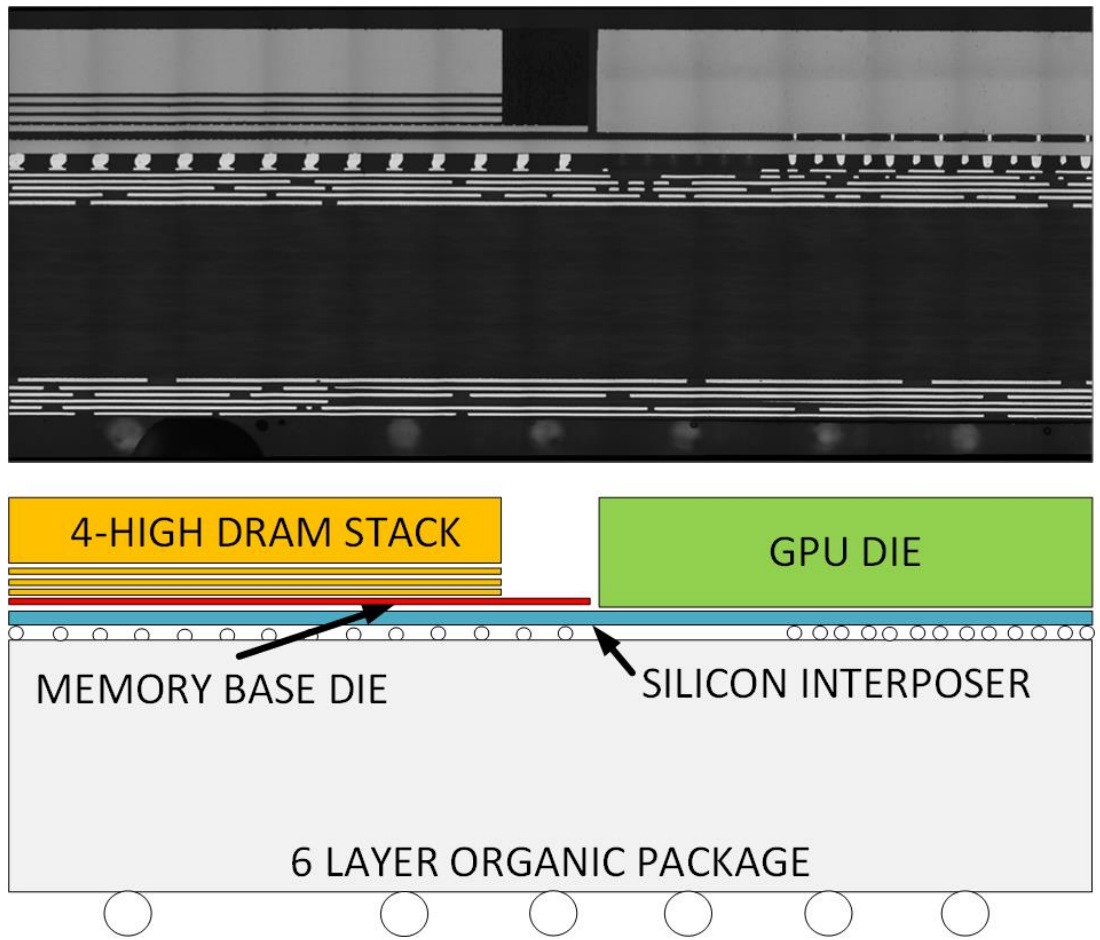
# HBM (HIGH BANDWIDTH MEMORY)

4096 Wide interface GPU⇔DRAM

720 GBps DRAM bandwidth

3x GDDR5 BW at equal power

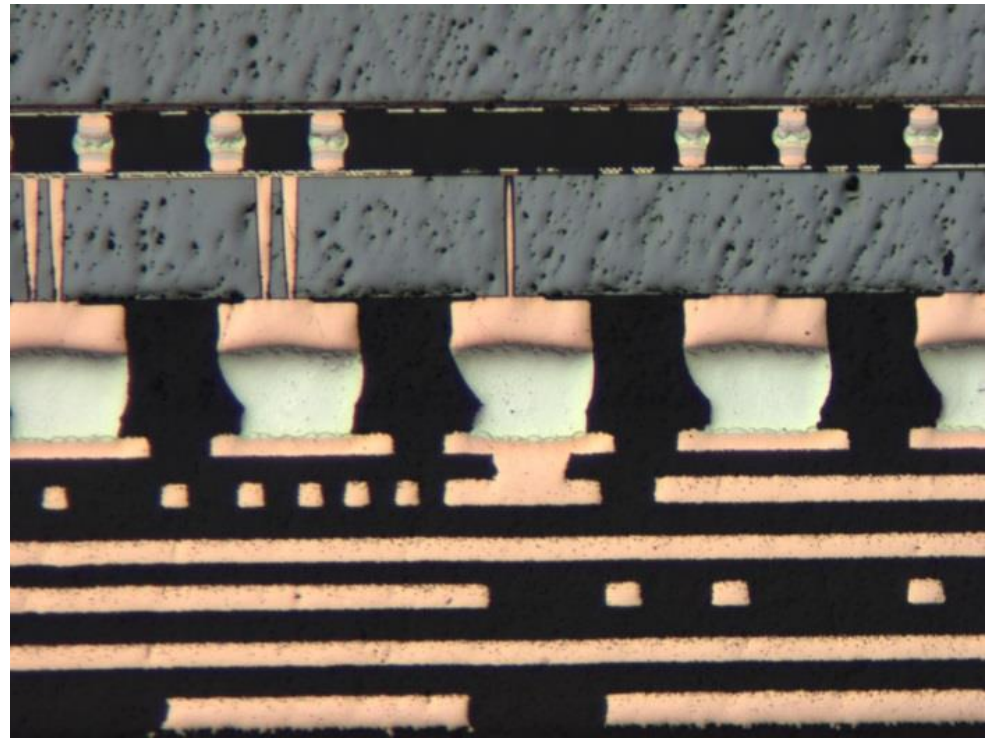
4-8x GDDR5 physical density



# MANUFACTURING
## Yield

Die & 4 stacks known good before assembly

Silicon interposer visually Inspected, not tested



# MANUFACTURING

## Warpage and Thermal

### Chip on Wafer on Substrate mitigates thin interposer warpage

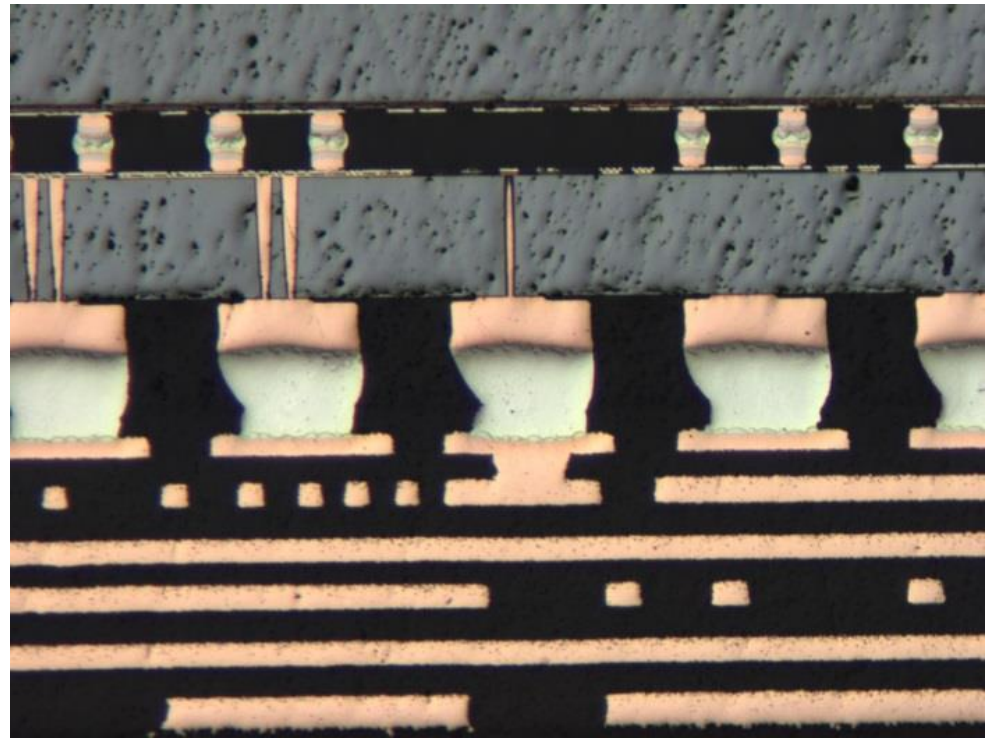
5 devices attached to thick interposer

Interposer thinned after attach

### 55x55 Package

5-2-5 buildup with a thick 1200um low CTE Core

Stiffener ring rather than a lid for better thermals



# NEW GP100 FEATURES

| FEATURE | JUSTIFICATION |
|---|---|
| **FP16 at 2x FP32 Rate** | Deep Learning |
| **FP16 Atomics** | Multi-Thread Math |
| **FP64 Atomics** | Multi-Thread Math |
| **1.5x DRAM BW / Flop** | Enabled by HBM |
| **Instruction Preemption** | Interrupt and Restart (Debug) |
| **Page Fault Stall** | Unified Memory (Next Slide) |
| **49 Bit Virtual Address** | Unified Memory |

# PASCAL UNIFIED MEMORY

## Automatic Page Migration

### CUDA 6 Code with Unified Memory

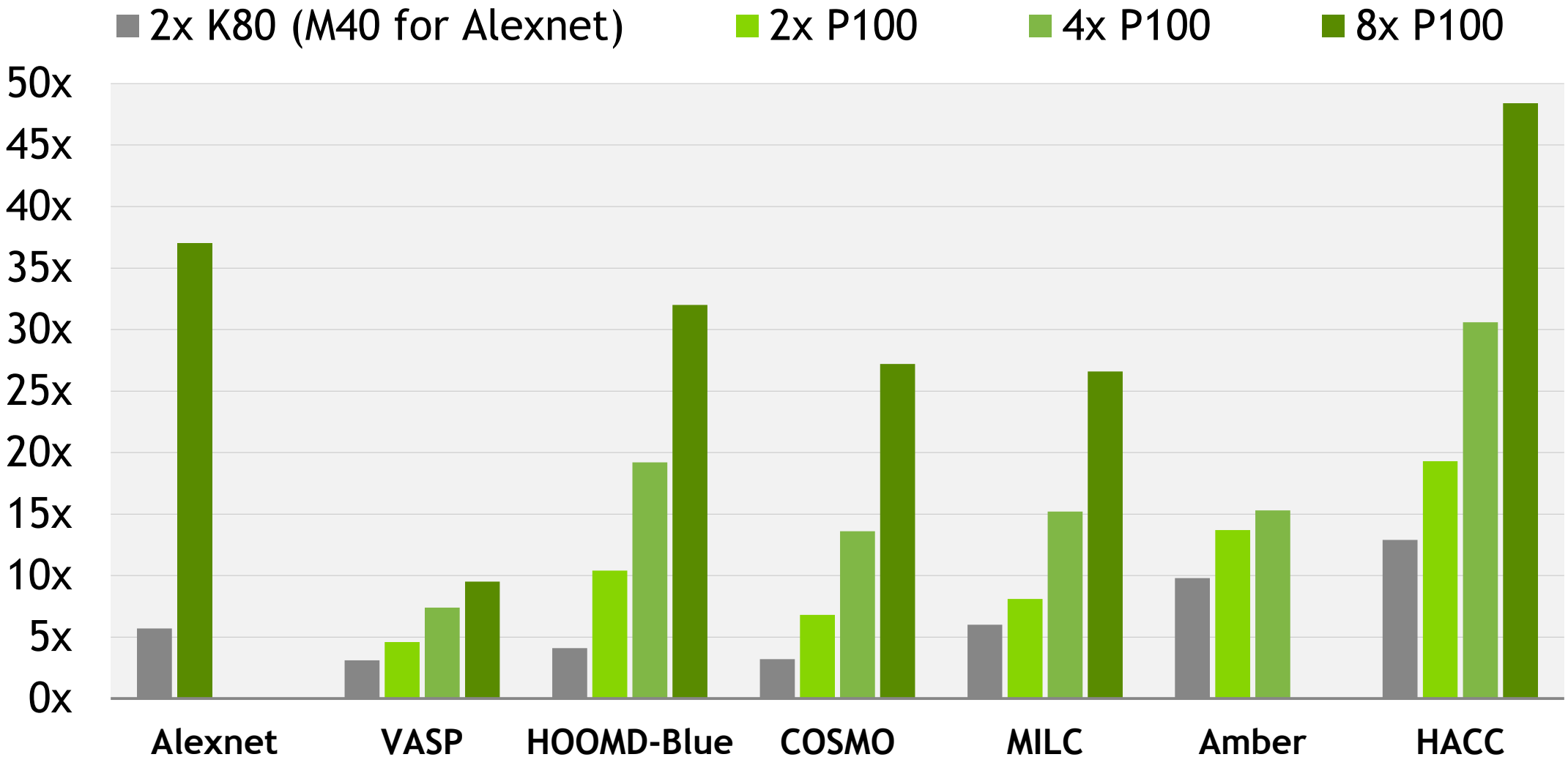
```
void sortfile(FILE *fp, int N) {
  char *data;
  cudaMallocManaged(&data, N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  cudaFree(data);
}
```

### Pascal Unified Memory*

```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  free(data);
}
```

*with operating system support

# GP100 PERFORMANCE SCALING WITH NVLINK



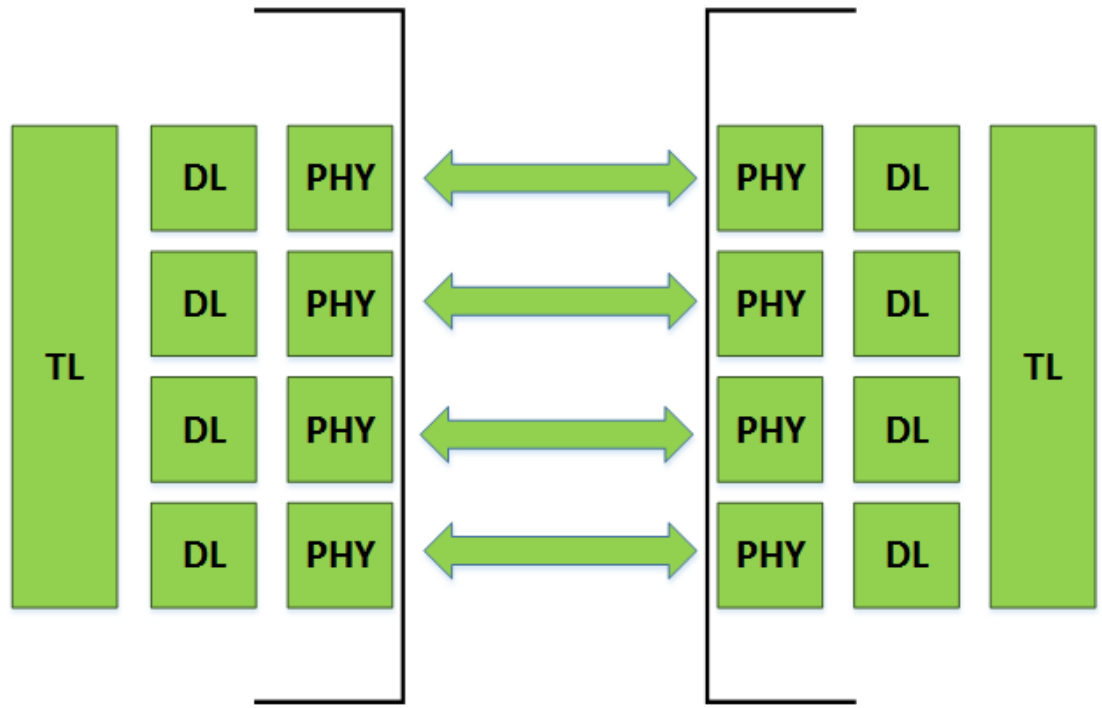
# NVLINK1.0 — PHYSICAL

Differential signaling at 20 Gbps

Building block is x8 with 16 wires in each direction

40 GBps bi-directional bandwidth

Pascal supports 4 NVLinks



# NVLINK1.0 — PHYSICAL

Embedded clock

85 Ohm terminated

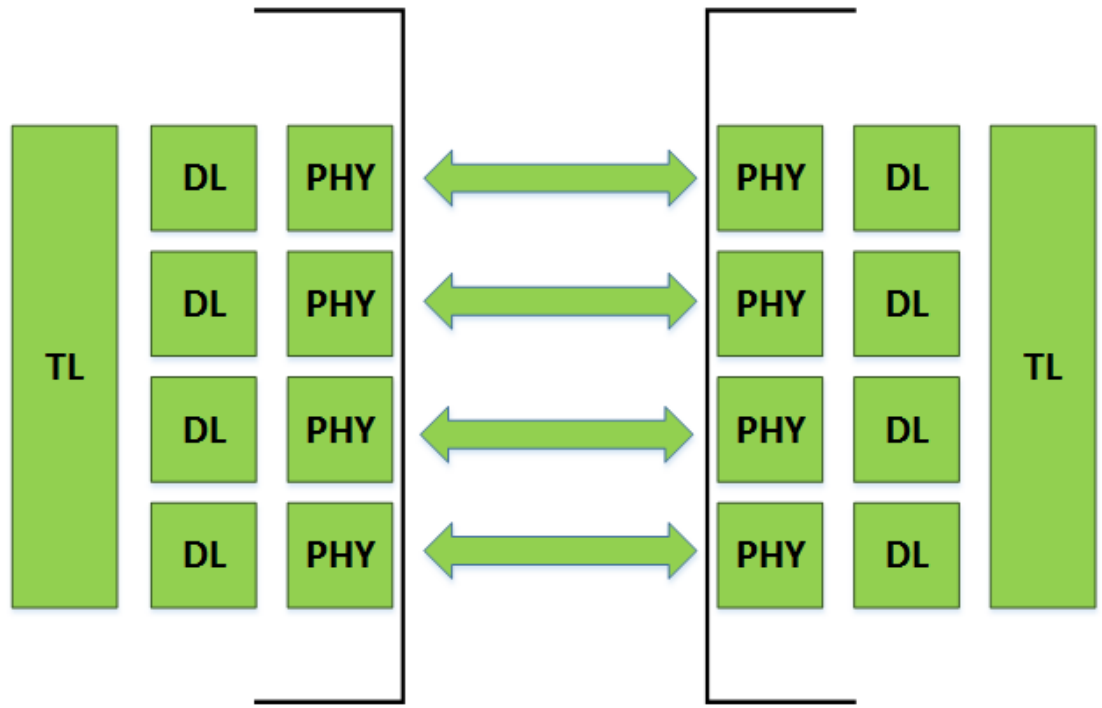
DC coupled

Bit Error Rate 1e-12

-22dB insertion loss (~15”)

Polarity inversion

Lane reversal



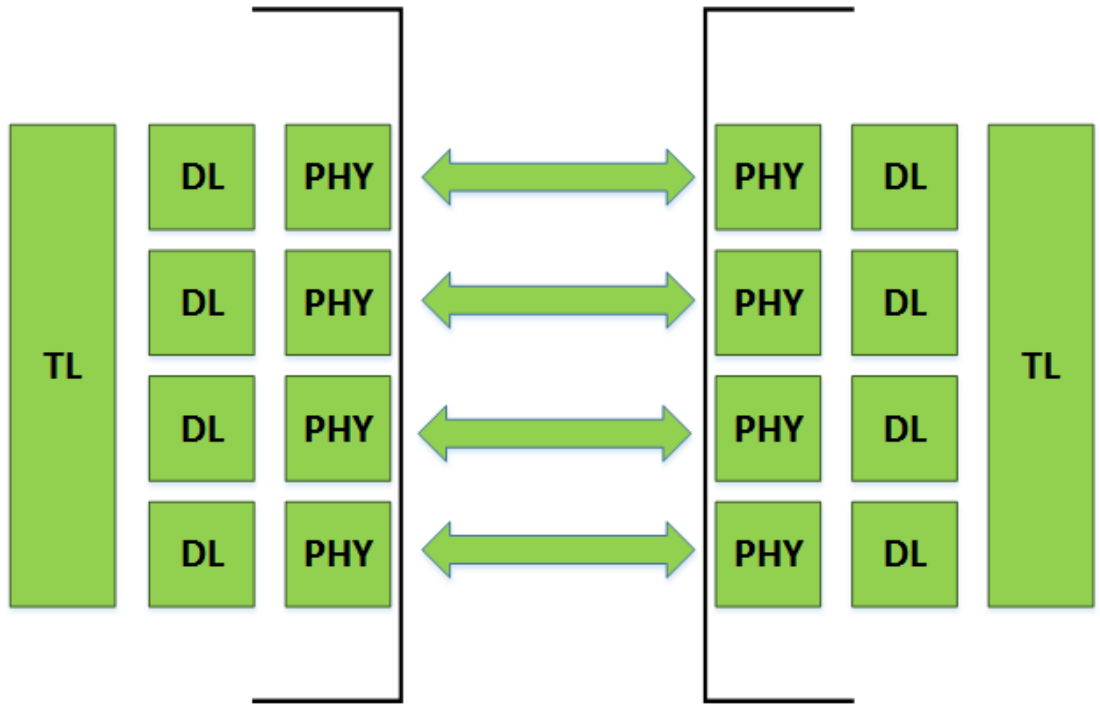
# NVLINK1.0 — PROTOCOL

Packetized protocol with variable length packet

CRC protected

Supports up to 256B transfers

94% efficiency with 256B transfers



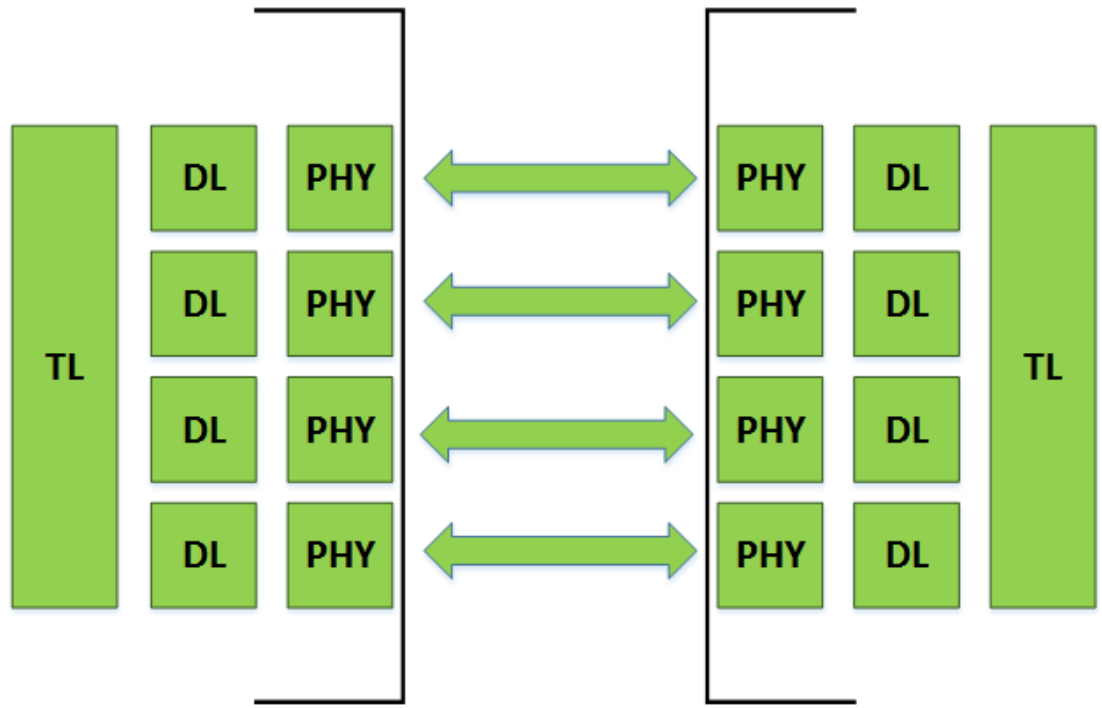
# NVLINK1.0 — PROTOCOL

Ganged links for higher bandwidth

Data sprayed across ganged links

Supports read/writes/atomics to peer GPU

Supports read/write access to NVLink enabled CPU



# NVLINK PACKET

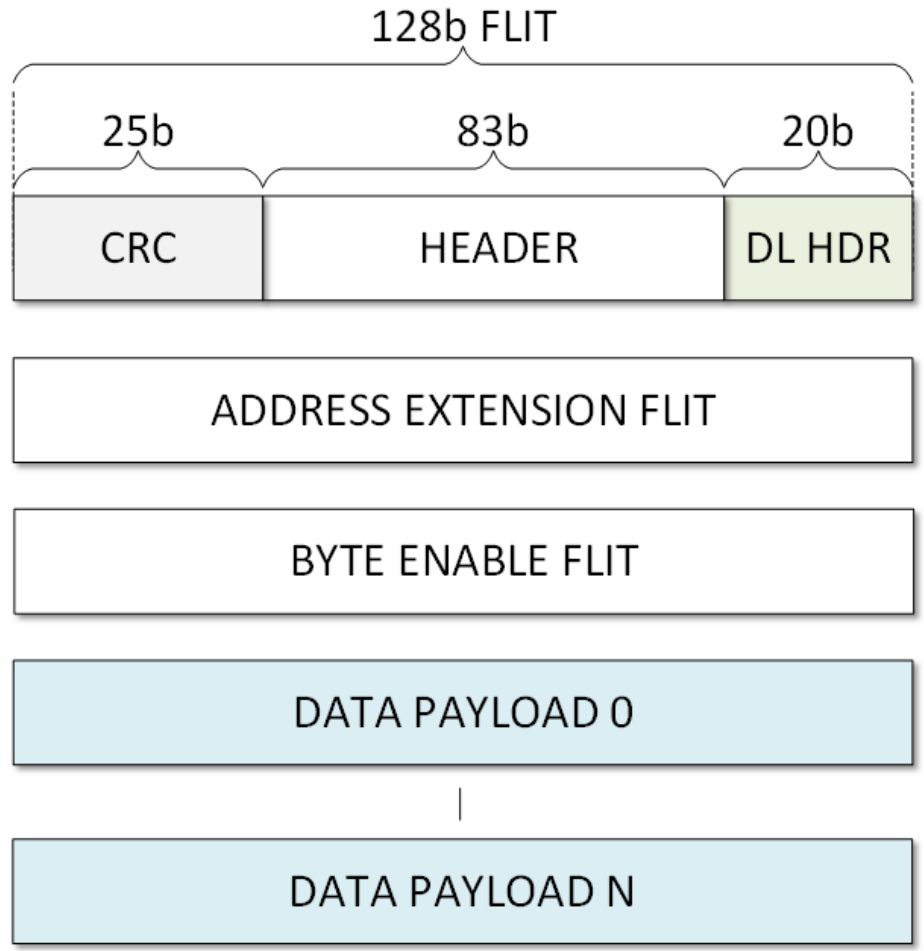
Variable length packet

Header flit contains CRC, DL header and TL header information

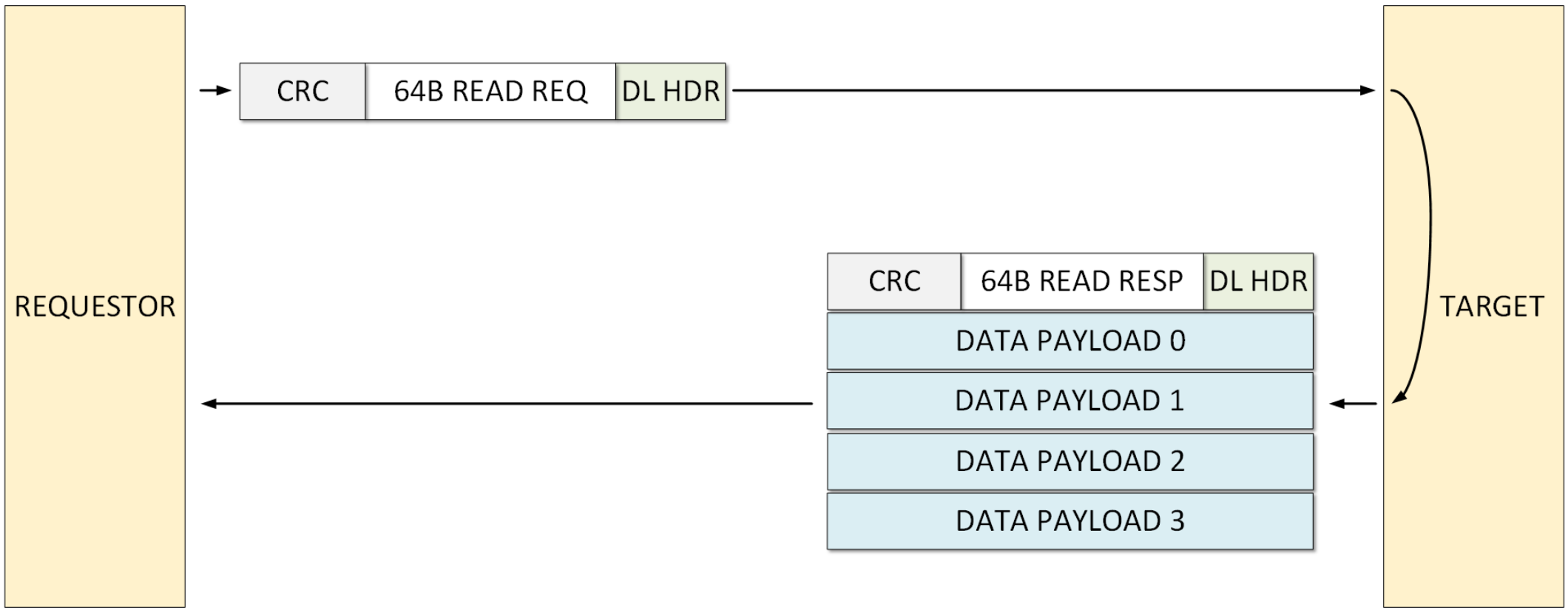
AE flit contains information likely to be shared by multiple commands (e.g. upper address bits).

Command specific BE flit used as required.

0–16 data payload flits



# SIMPLE READ



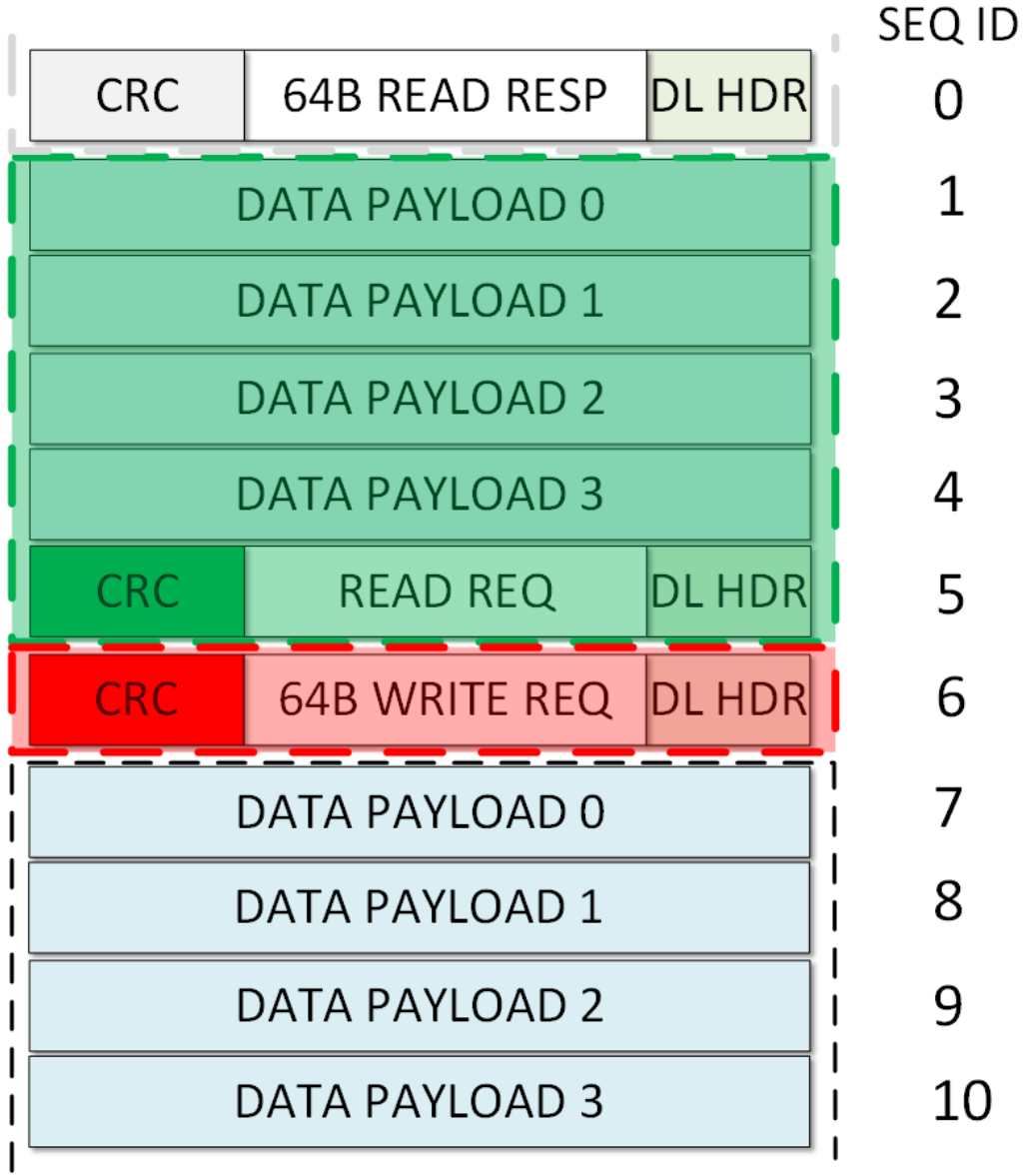
# CRC AND REPLAY

25b CRC on each packet

5 random bits in error in max packet or a burst of up to 25b on a single lane

Packet header contains packet length

CRC calculated over current header and previous payload to free up packet length info asap

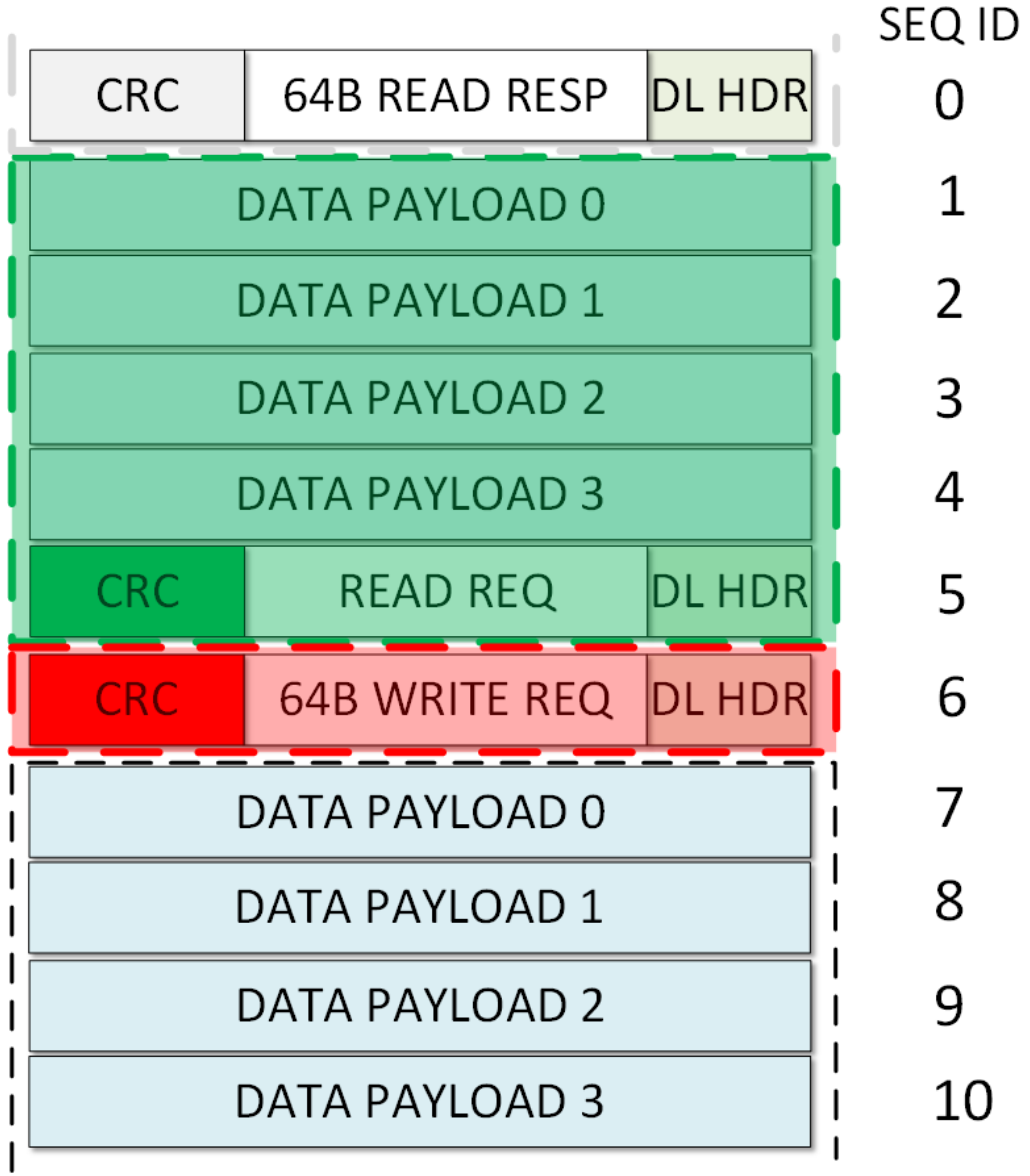


# CRC AND REPLAY

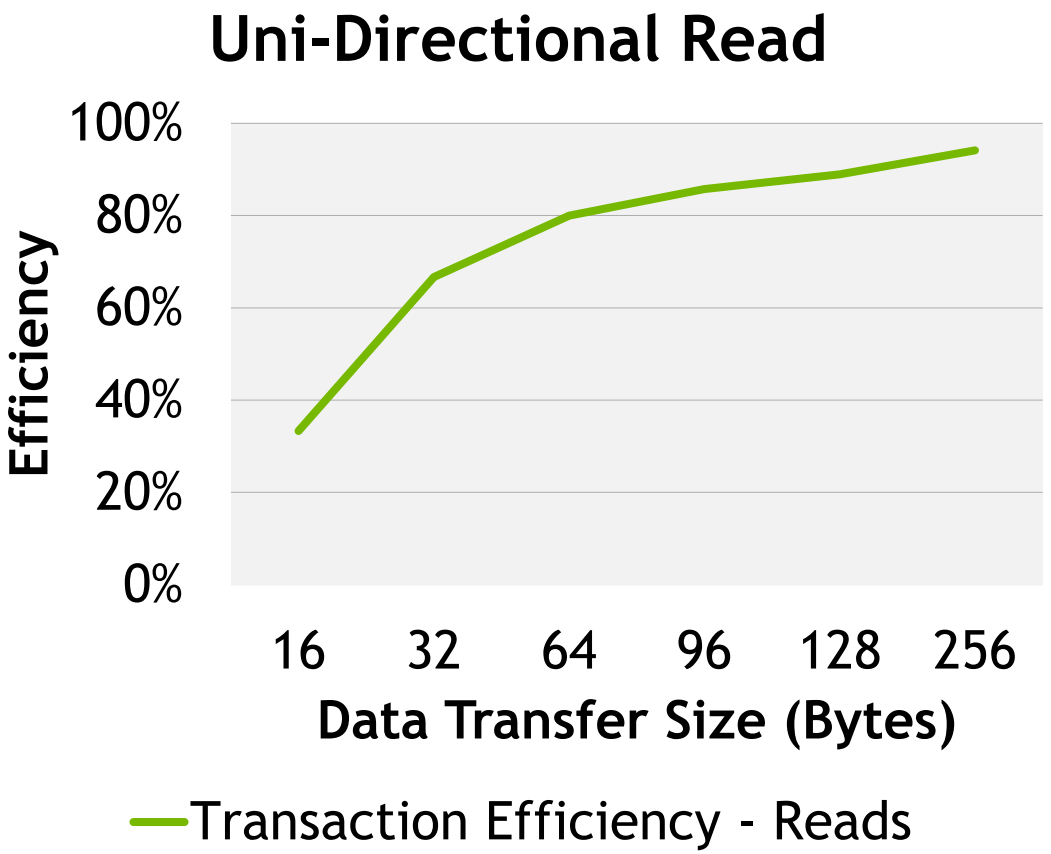
Positive acknowledgement of good CRC

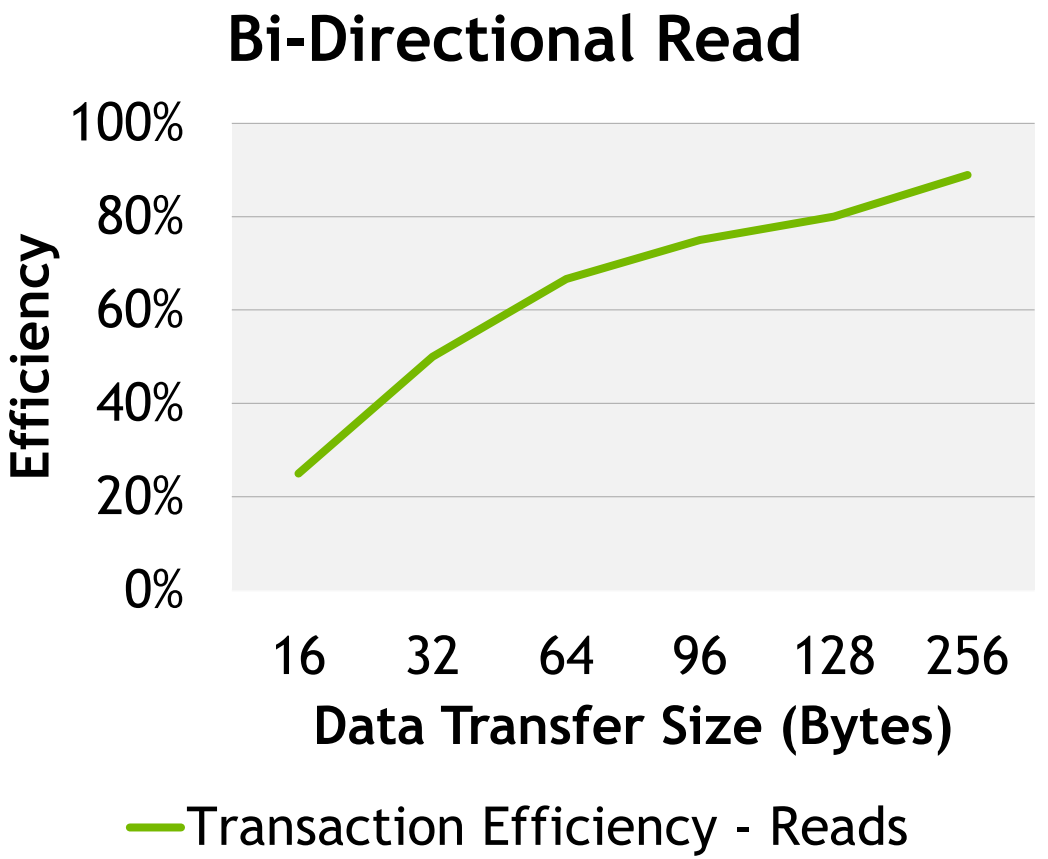
Transmitted data stored in replay buffer

In the event of an error replay sequence is started from last ACKed packet



# EFFICIENCY


Uni-Directional Read
100%
80%
60%
40%
20%
0%
Efficiency
16
32
64
96
128
256
Data Transfer Size (Bytes)
Transaction Efficiency - Reads


Bi-Directional Read
100%
80%
60%
40%
20%
0%
Efficiency
16
32
64
96
128
256
Data Transfer Size (Bytes)
Transaction Efficiency - Reads

# NVLINK — GPU CLUSTER — DGX-1

Two fully connected quads

Quads connected at corners

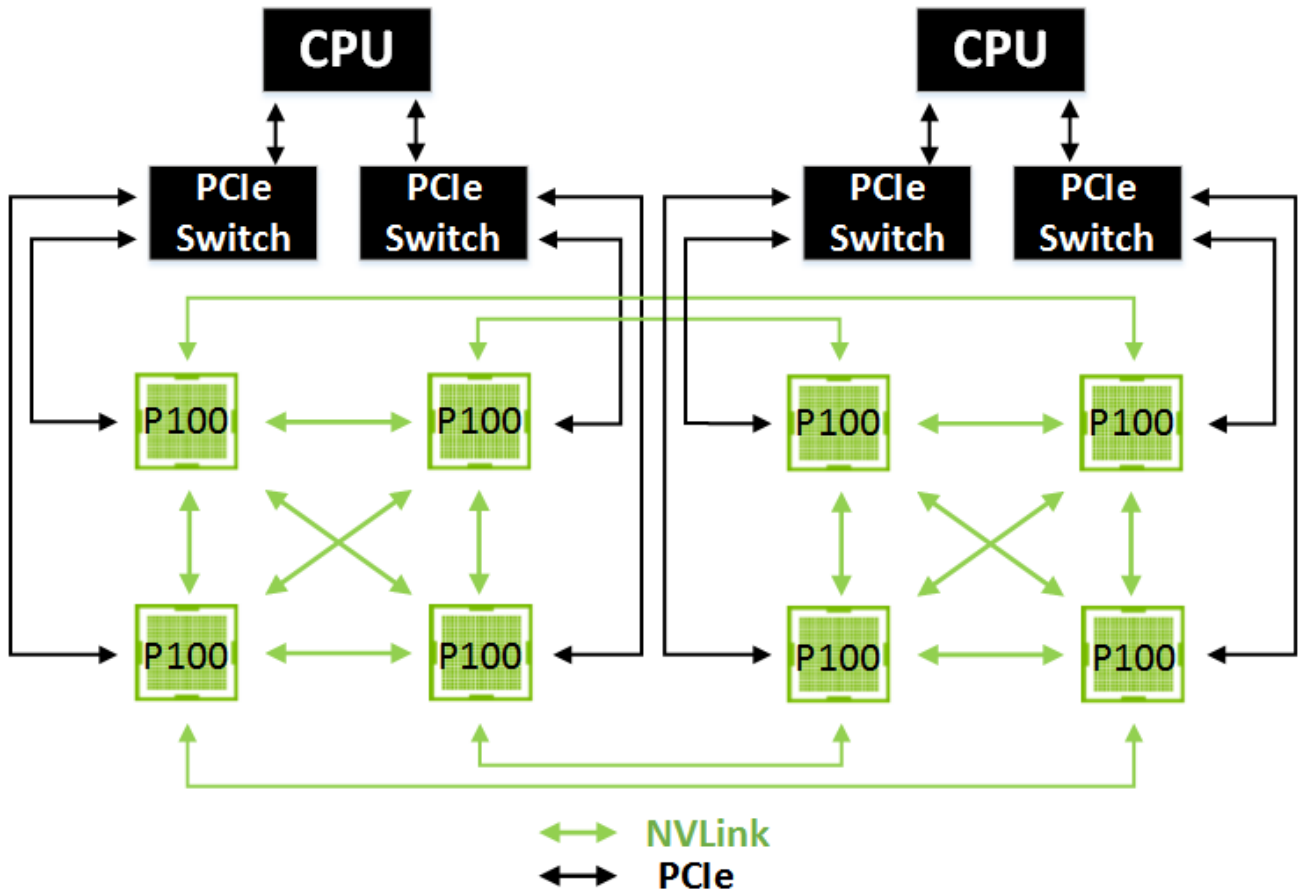
640 GBps of NVLink bidirectional bandwidth

Load/store access to peer memory

Full atomics to peer GPUs

High speed copy engines for bulk data copy

PCIe to/from CPU



# NVLINK TO CPU

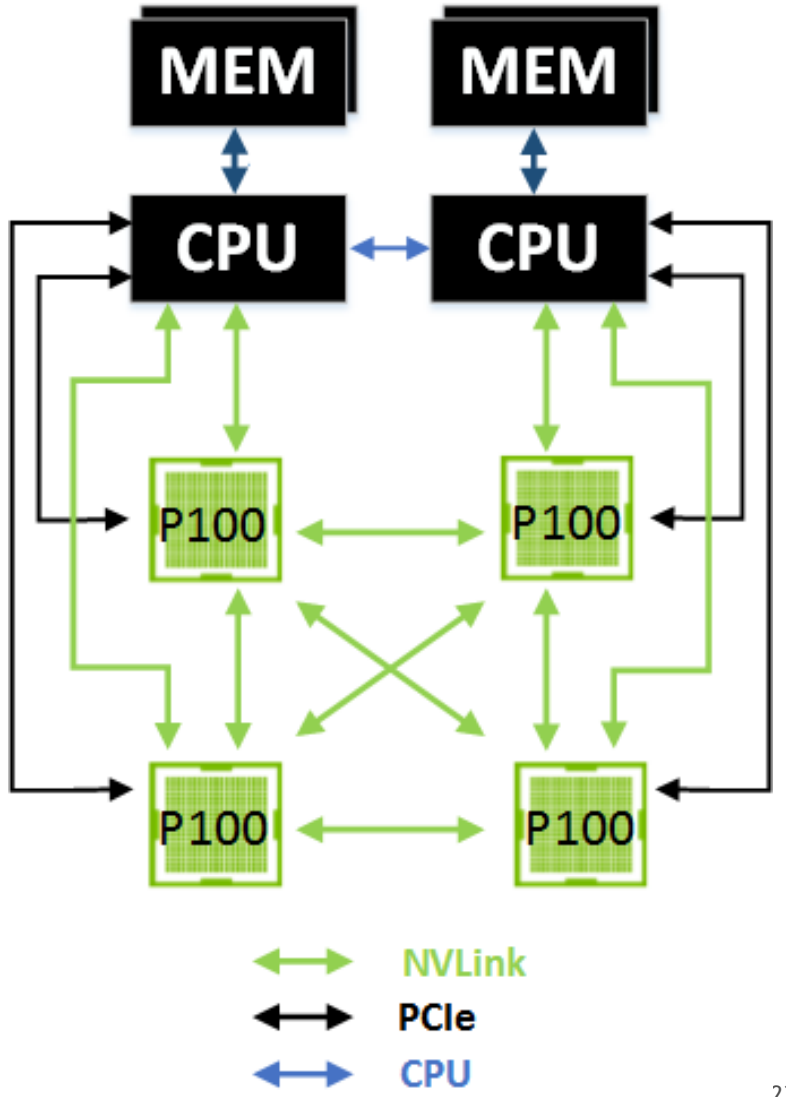
Fully connected quad

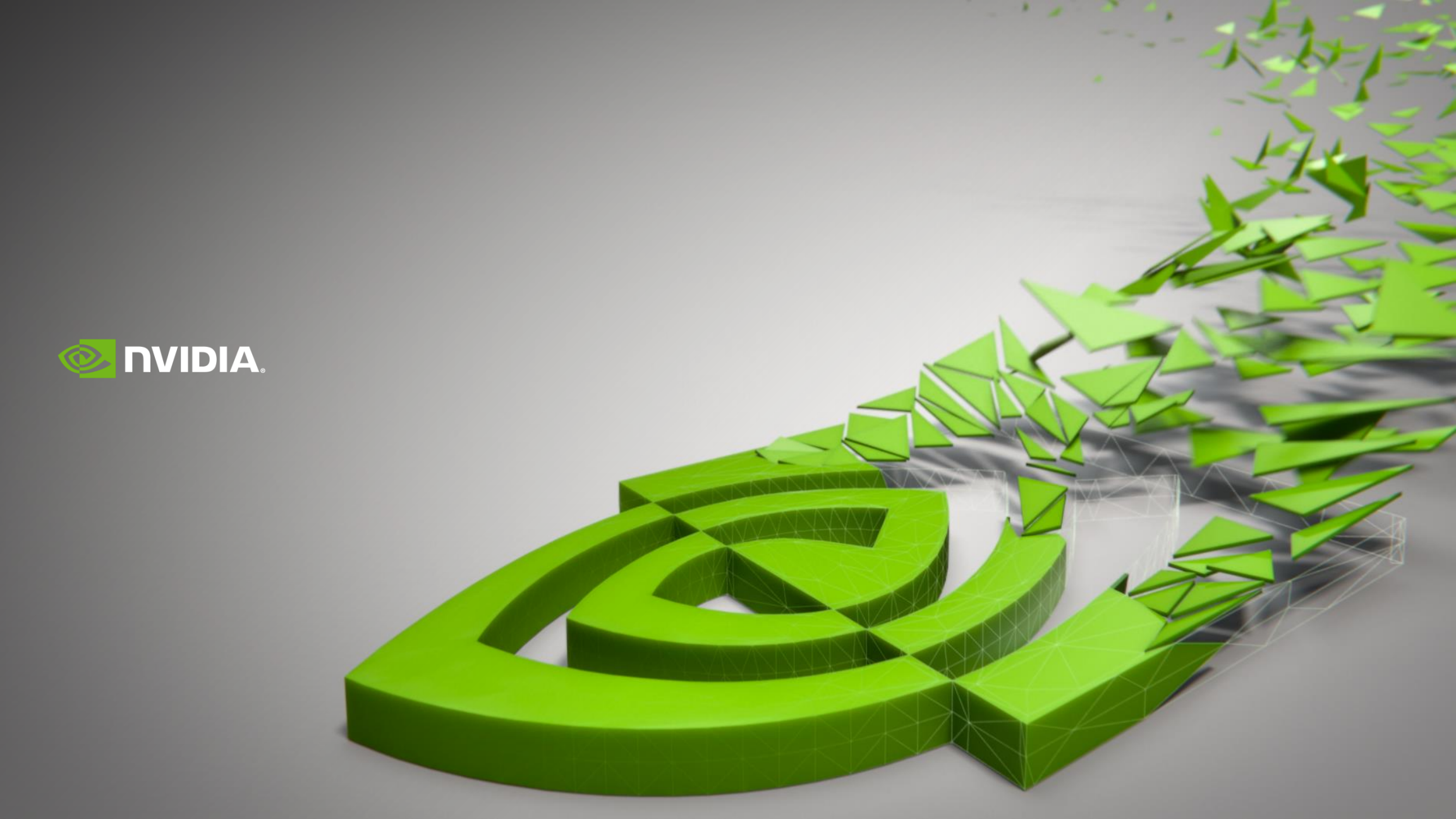
120 GBps per GPU bidirectional BW to peers

40 GBps per GPU bidirectional BW to CPU

Direct Load/Store access to CPU memory

High Speed Copy Engines for bulk data movement




NVIDIA