Evaluating On-Node GPU Interconnects for Deep Learning Workloads

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

NATHAN TALLENT, NITIN GAWANDE, CHARLES SIEGEL
ABHINAV VISHNU, ADOLFY HOISIE

Pacific Northwest National Lab

PMBS 2017 (@ SC)

November 13, 2017

Acknowledgements

Center for Advanced Technology Evaluation (ASCR)
Performance Prediction & Diagnosis for Extreme Scientific Workflows (ASCR)

Scaling 'Deep Learning' Increasingly Important

PNNL

- Scaling some workloads requires a high-performance interconnect
- Motivating Example: KNL/Omni-path vs. DGX-1 (NVLink 1.0)

What is scaling behavior given workload and interconnect?

Single-KNL/GPU performance very similar, despite GPU's higher peak!

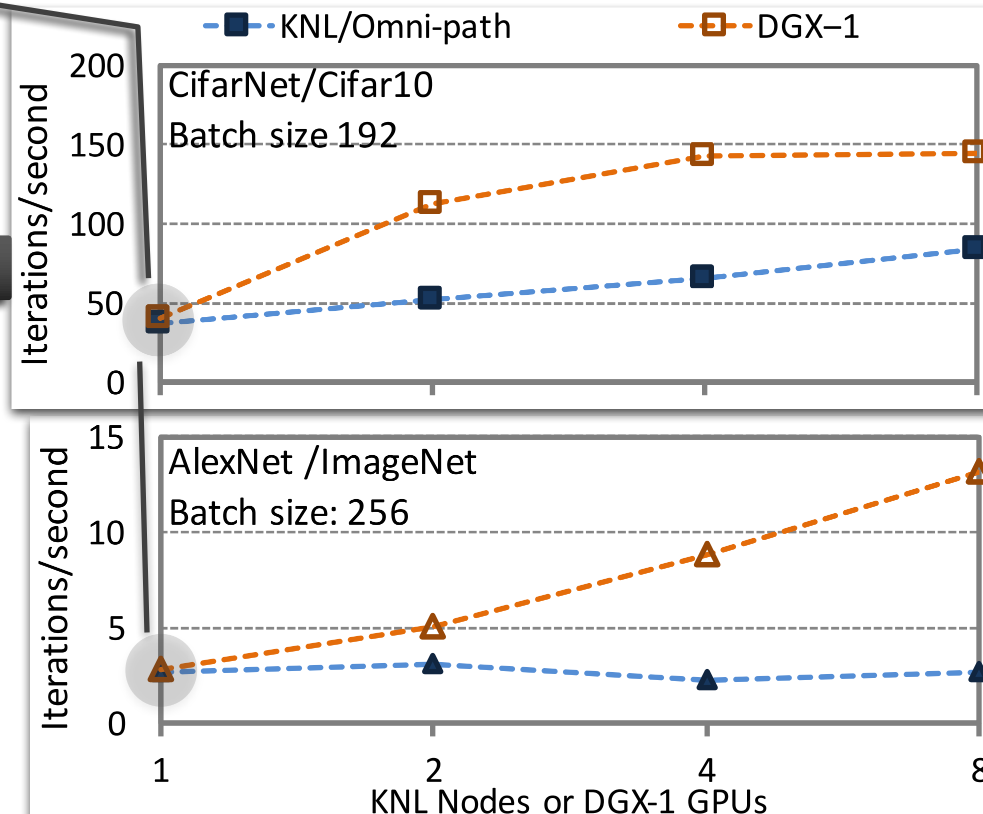
DGX-1: better absolute performance...

...but scaling behavior is quite different

With Omni-Path, CifarNet scales better than AlexNet

With NVLink, AlexNet scales better than CifarNet

AlexNet's much larger all-to-all reduction operations stress interconnect bandwidth



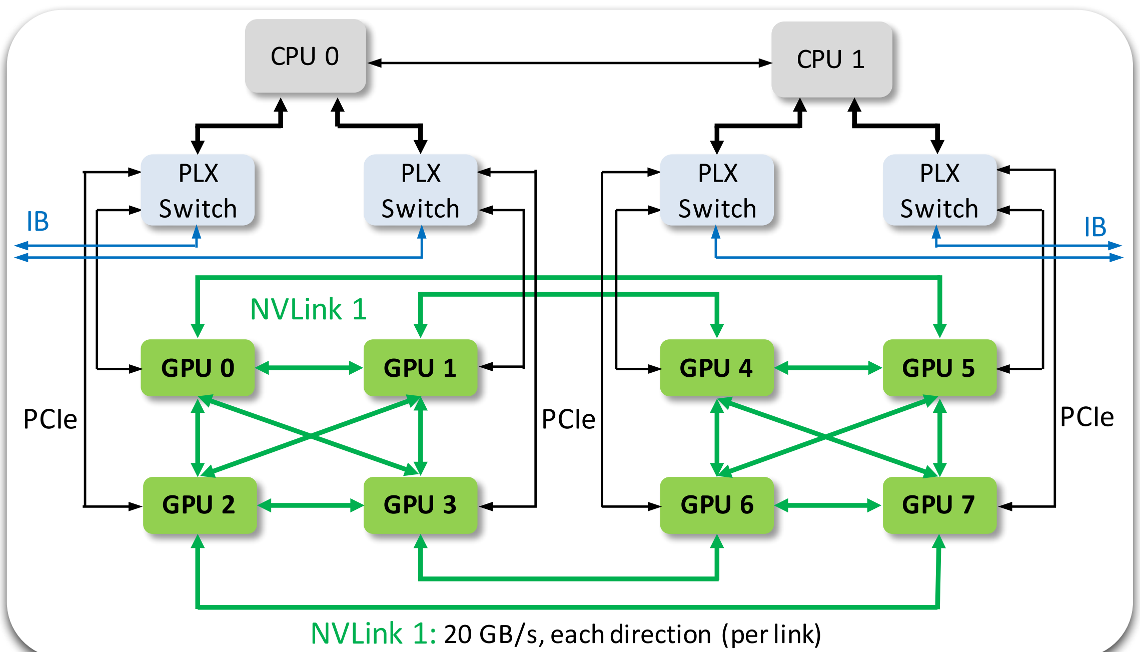
Which On-Node GPU Interconnect is Best For Me?

PNNL

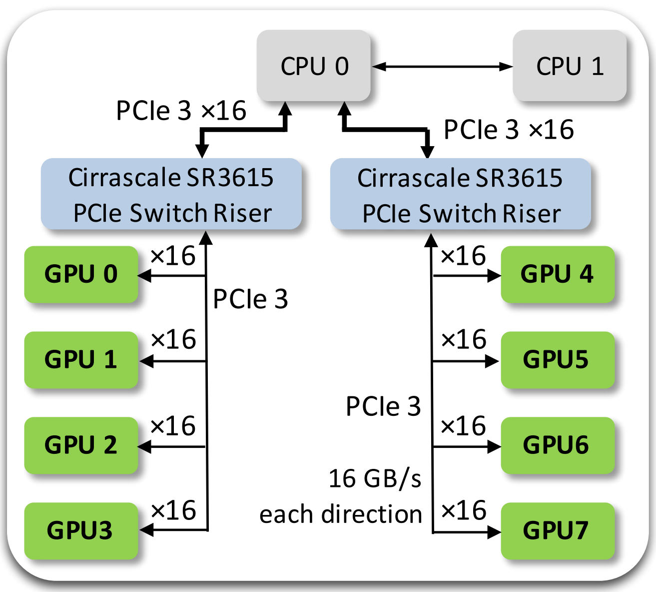
- Our focus: Scaling Deep Learning across *on-node* GPUs:
  - Is a high-performance interconnect required (e.g., NVIDIA NVLink)
  - Are PCIe-based interconnects adequate?
  - How dependent is the answer on my workload?

Answers not obvious!

NVIDIA DGX-1 (NVLink 1.0)



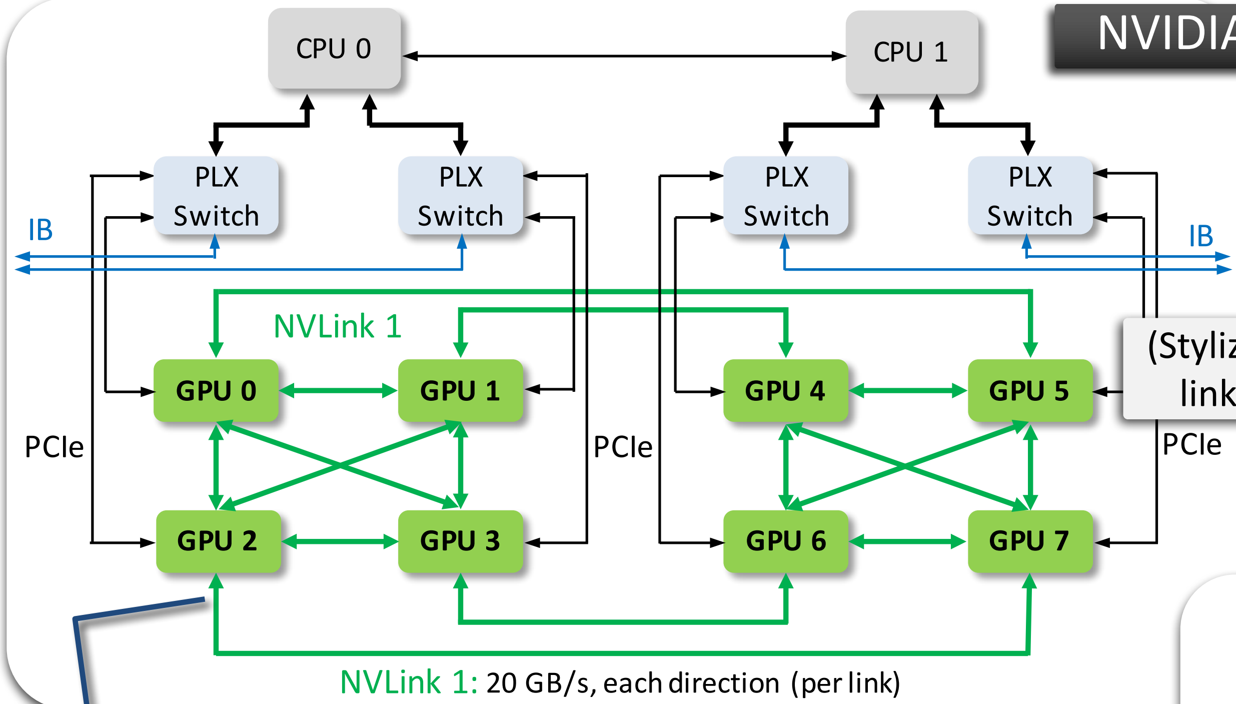
Cirrascale GX8



On-Node GPU Networks: DGX-1 vs. GX8

PNNL

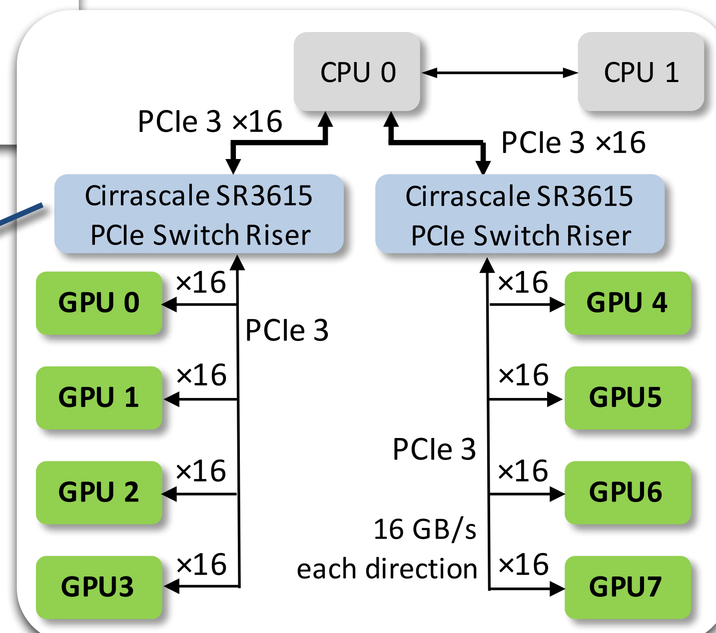
NVIDIA DGX-1



DGX-1 appears to offer much higher performance…

(Stylized to avoid crossing links: GPU0 ↔ GPU4)

Cirrascale GX8



Hybrid cube mesh:

- Two (fully connected) 4-GPU meshes
- Each GPU: 4 links = 80 GB/s (uni)

Two-level tree (PCIe):

- Two (fully connected) 4-GPU clusters
- Each GPU: 16 GB/s (uni) PCIe ×16
- Switch upstream: 16 GB/s

Outline of Deep Learning Workload

- Outline of deep learning training algorithm
  - Replicate neural network architecture on each GPU
  - For each batch in image data set:
    - Distribute images among GPUs (data parallel)
    - Process images → activations → parameters (per-GPU)
      - activation: floating point operations
    - Synchronize parameters: all-to-all reduction (allreduce)
- Use NCCL for GPU collectives:
  - NCCL: NVIDIA Collective Communications Library
  - topology-aware rings, optimized for throughput (pipelined)
  - interconnect-aware
- Train on ImageNet Dataset:
  - ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
  - Well known benchmark for object classification and detection

Workloads:

- AlexNet (high comm)
- GoogLeNet (high compute)
- ResNet/*x*: everything in-between & more

Parameterize ResNet: Control Compute Intensity

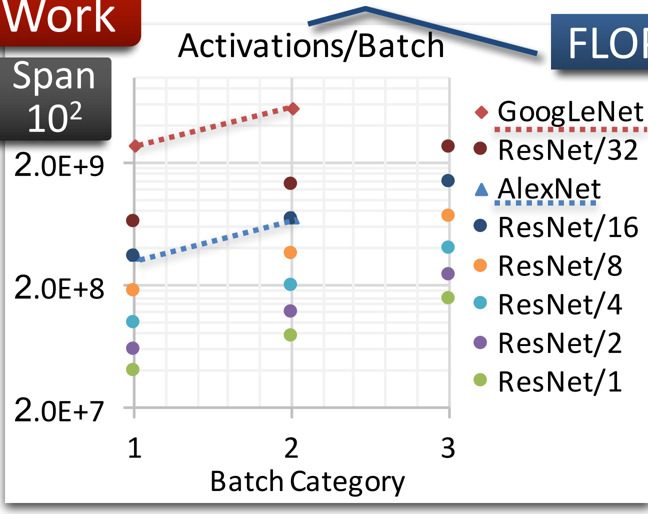
PNNL

Parameterized workload: systematically represent range of neural network depths & batch sizes

Work

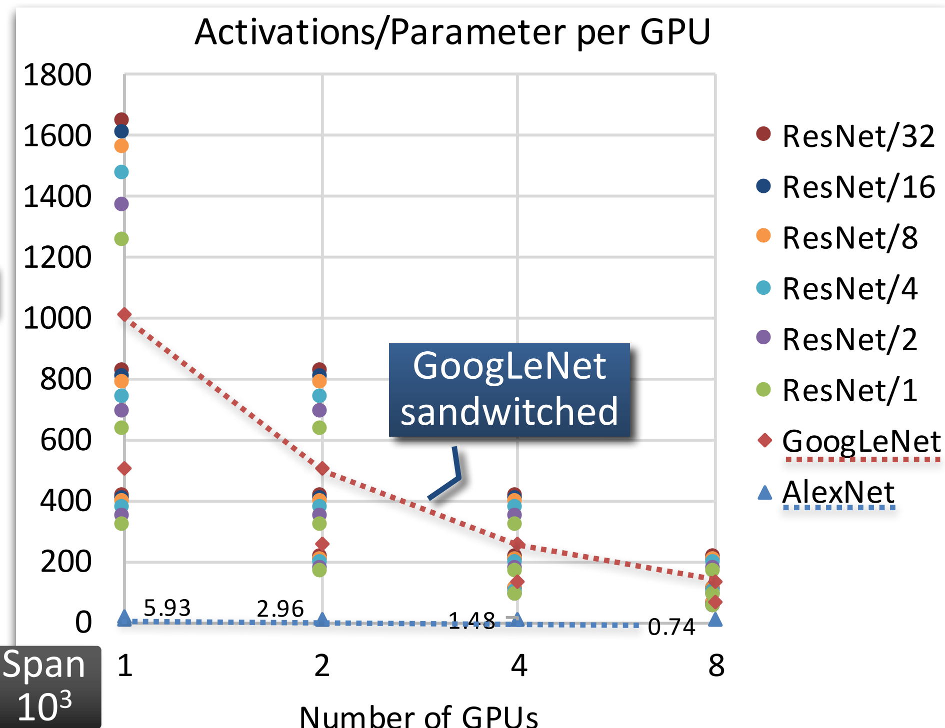
Span 10^2

FLOPS



ResNet/x Replicate a ResNet 'block' x times where is x is {1, 2, 4, 8, 16, 32}

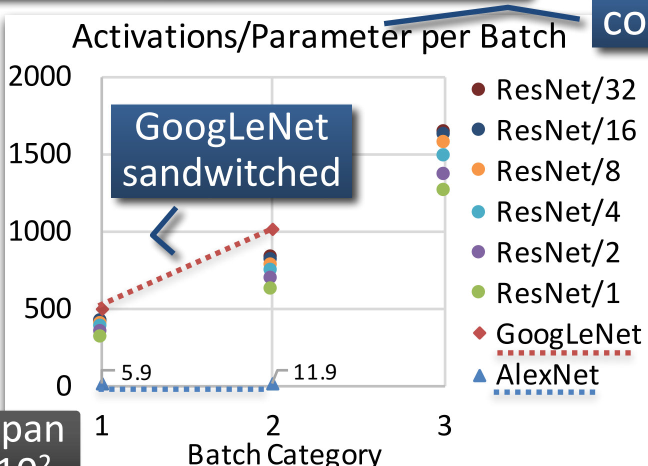
Intensity When Strong Scaled



Span 10^3

Intensity (Work/Comm)

communication (allreduce)



Span 10^2

GPU-to-GPU Memory Copy: Bandwidth

MGBench: unidirectional; GPU-GPU; pipelined using CUDA’s async memcopy

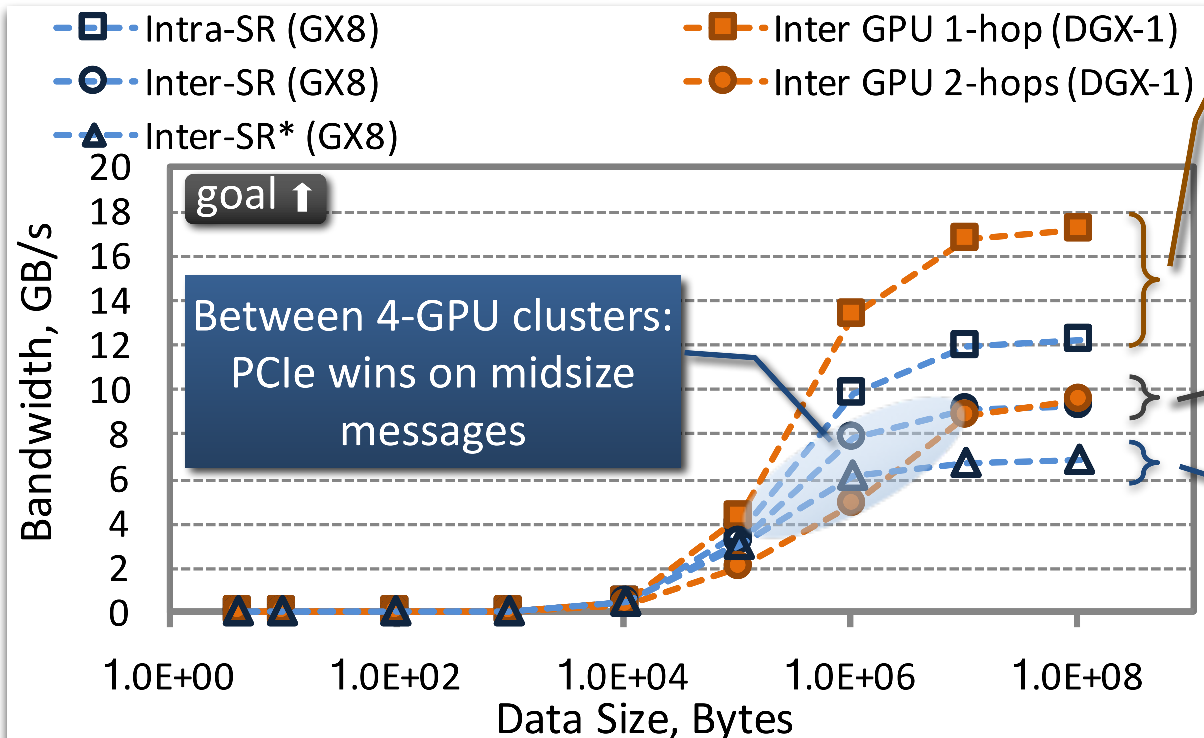
Group results by value clusters

GX8 has *three* groups:
- Intra-SR: within switch
- Inter-SR: between switches
- Inter-SR*: anomaly

DGX-1 has two groups (expected)

Within 4-GPU clusters (1-hop; intra-switch): NVink wins (85% of 1 link)

(Uses only 1 NVLink; software has to manage routing, etc.)

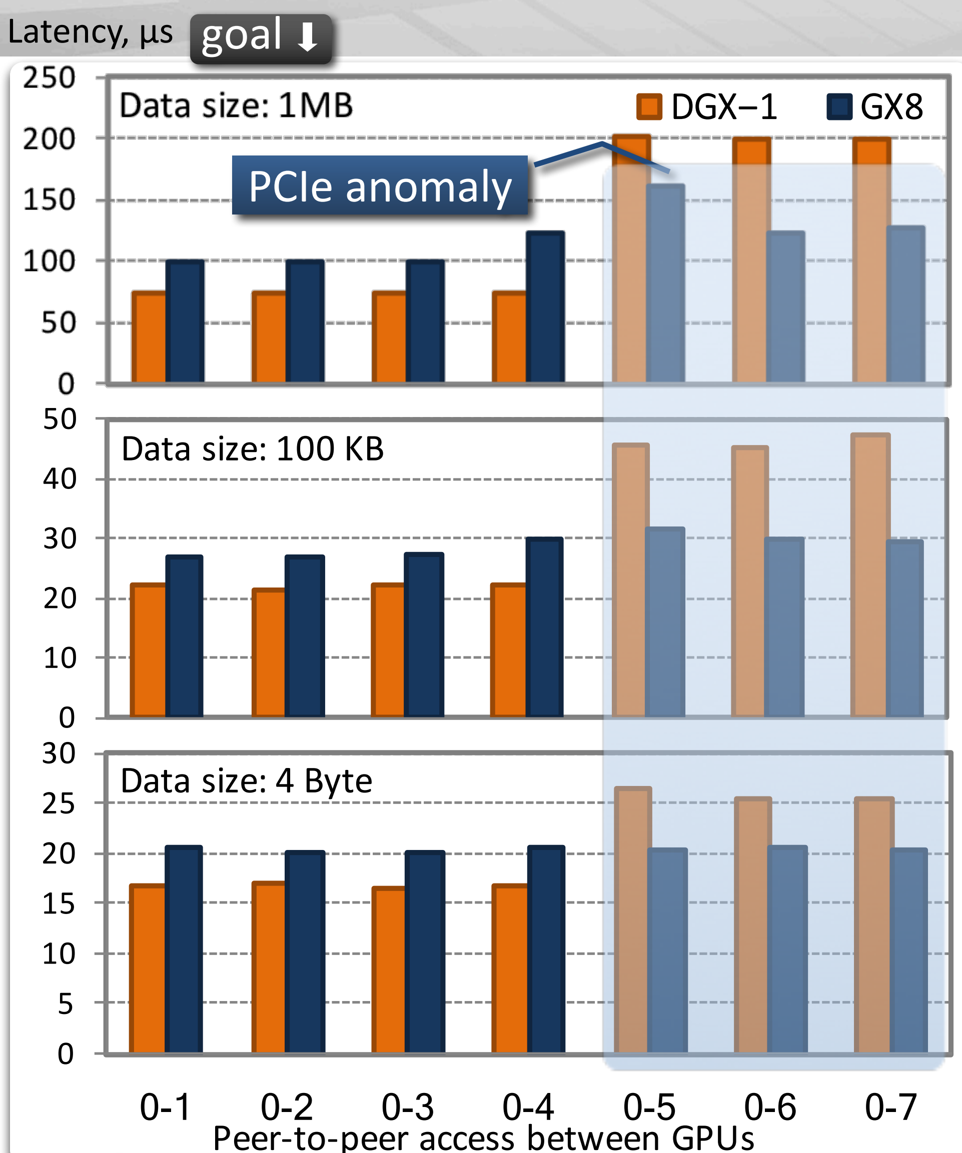


Between 4-GPU cluster (2-hop; inter-switch): depends on payload size

PCIe anomaly (see latency plots)

PCIe can win on ‘long’ midsize transfers

GPU-to-GPU Memory Copy: Latency

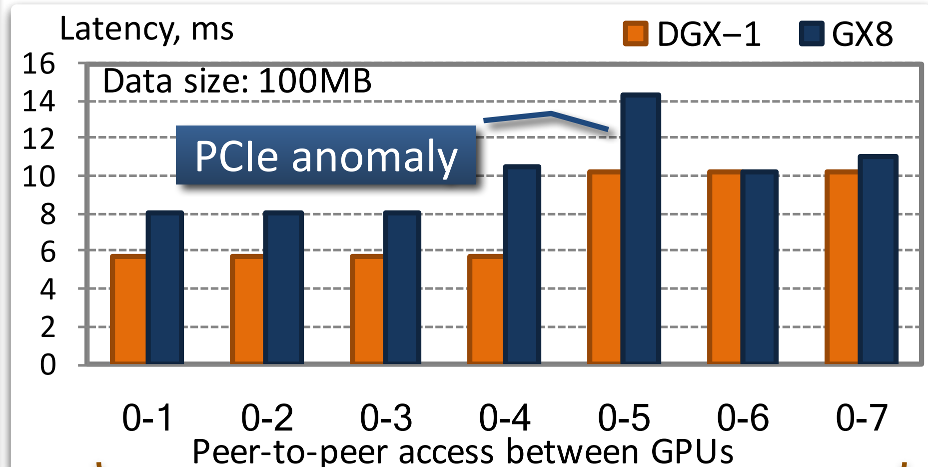


NVink wins

PCIe wins: bandwidth saturates more quickly w.r.t payload

Details at four different data sizes

x-*y* means GPU *x* sent data to GPU *y*



NVink wins for large payloads

NVLink: 2 groups, independent of data size

PCIe: 1—3 groups, dependent on data size

PCIe Anomaly Cirrascale SR, 2nd slot (GPU5) has longer signal paths; delays

NCCL: NVIDIA Collective Communications Library

- NCCL uses topology-aware & interconnect-aware rings

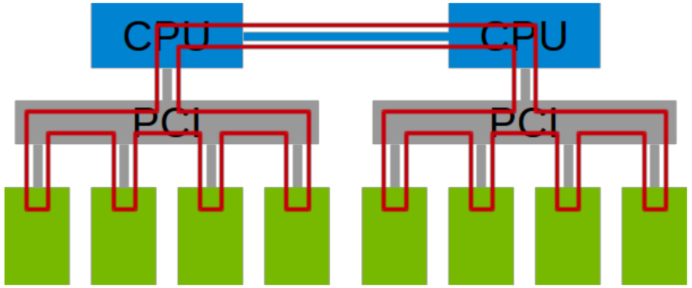
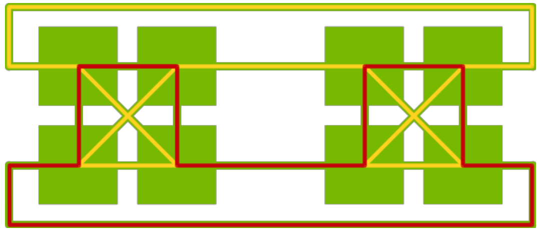


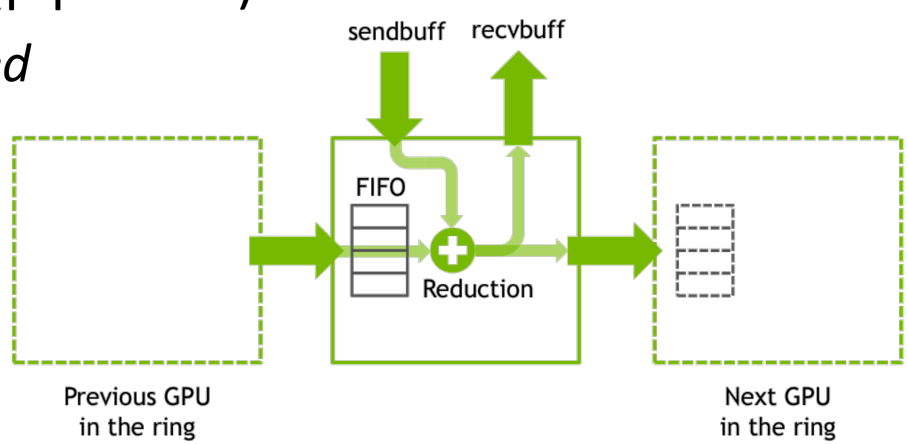
Image: Sylvain Jeaugey



PCIe / QPI : 1 unidirectional ring

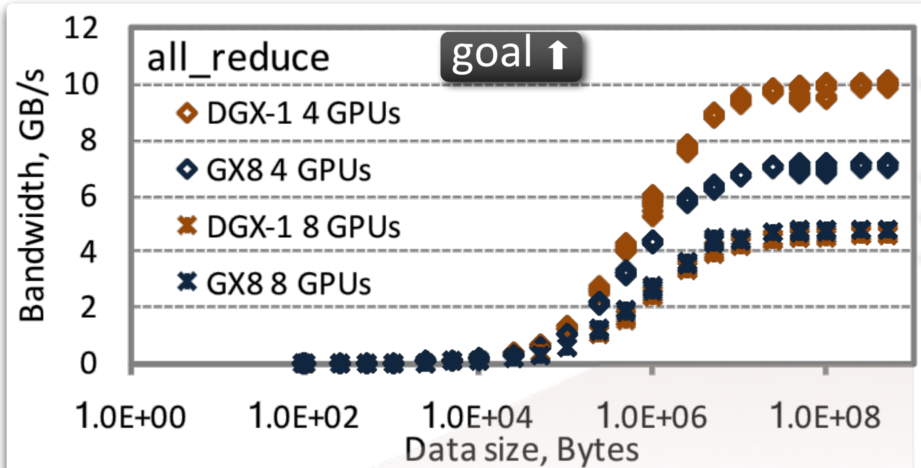
DGX-1 : 4 unidirectional rings

- NCCL is optimized for throughput (pipelined)
  - Small payload: ring latency *exposed*
    - time = hops × link latency
  - Large payload: ring latency *hidden*
    - time = payload / bandwidth

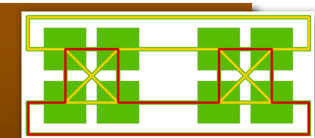


NCCL Allreduce: Effective Bandwidth

Effective BW: bandwidth relative to a *single* GPU's payload. Max is BW of 'memcopy.'

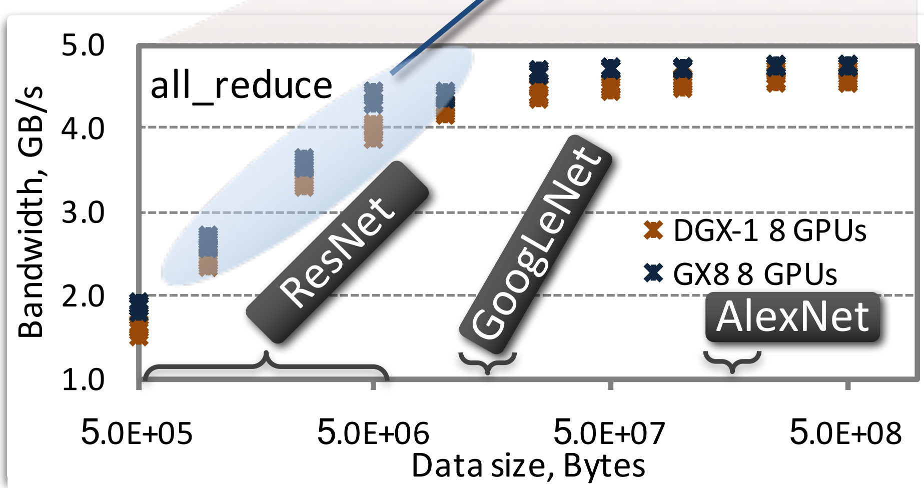


4-GPUs (*within* cluster); ideal allreduce is 1 step. NVink wins by 40% (60% of max)



8-GPUs (*between* clusters); ideal allreduce is 2 steps: PCIe wins by 3%!

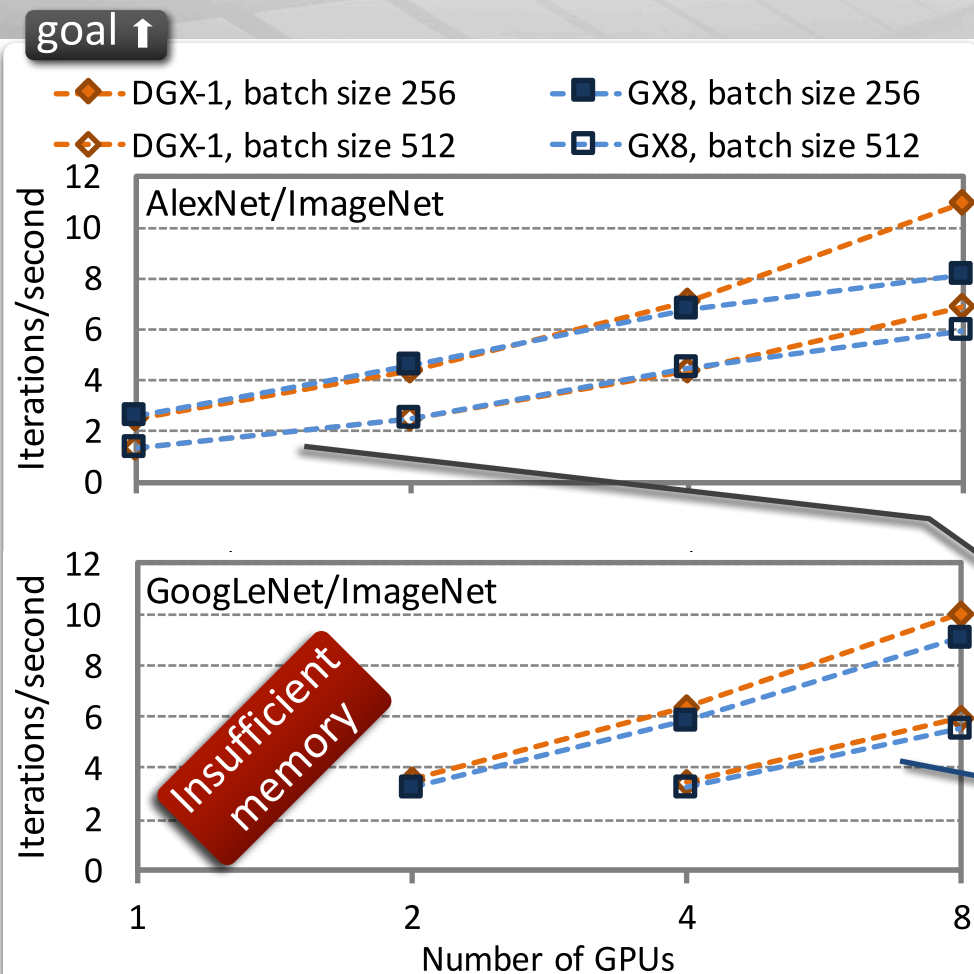
8-GPUs: PCIe wins by 10% on midsize messages



PCIe Bandwidth saturates more quickly with respect to payload size. More hardware for switching and flow control?

Broadcast Performance differs with collective. On 8-GPU broadcast, NVLink has slight advantage: single-root has less synchronization vs. all-to-all.

Strong-scaling (ImageNet): AlexNet & GoogLeNet



NVLink important for AlexNet (NVlink has 36% advantage)

Unexpected! Although AlexNet is communication intensive, GX8 has slightly higher 8-GPU allreduce performance!

Same single-GPU performance. Power cap GPUs to equalize the slightly different SM frequencies

PCIe is close to NVLink for GoogLeNet

Expected GoogLeNet is more compute intensive than AlexNet by 100× (activations/parameter/batch)
AlexNet: 5.9 and 11.9
GoogLeNet: 500 and 1004

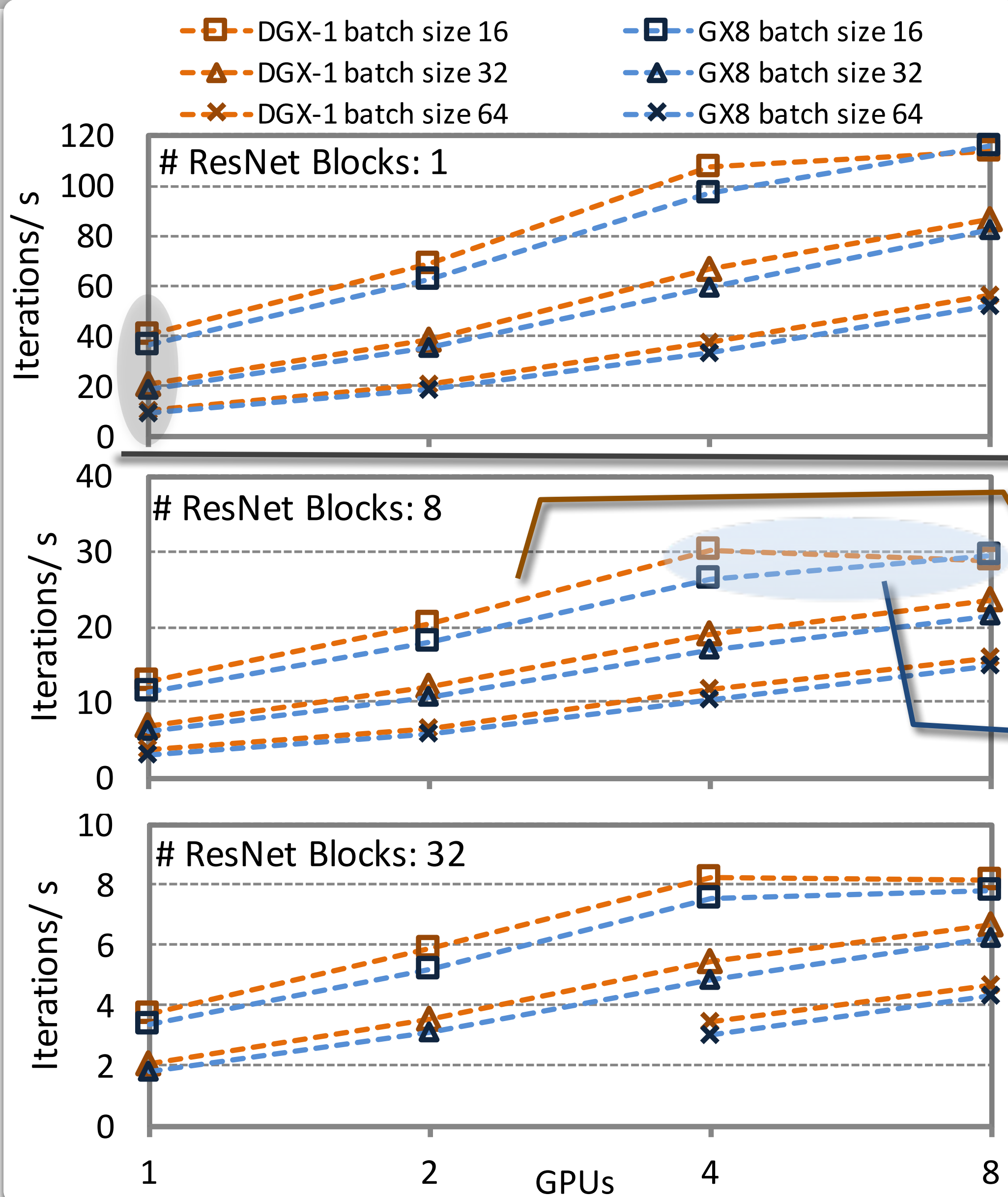
Expected NVLink becomes less important as batch size increases (more computation).

Gripe: GPUs have very poor performance tools

Strong-scaling (ImageNet): ResNet/x

2, 4, 16 in paper

PNNL



Performance expectation

- Identical GPU work
- NVLink/PCIe win/loss: fraction of allreduce × allreduce win/loss

Single-GPU performance slightly different! Converges as batch size increases. But why? CPU-based overheads on smaller batch sizes?

Expect DGX-1 win for 2 and 4 GPUs. Holds.

Expect GX8 win for 8 GPUs. Explains ‘knee’ on batch size 16. Why no more ‘knees’?

GX8 is competitive for ResNet-style workloads.

Smaller batch sizes (vs. AlexNet, G-Net). Comports with ResNet’s deeper network & fewer parameters; highlight interconnect.

Conclusions

PNNL

- Scaling ML across multiple on-node GPUs is increasingly important
- 'Workload Intensity' helps explain scaling performance
  - Parameterized ResNet captures large space of workload intensities
    - Systematically characterize & specify neural network workloads
  - Workload intensity: reflects computation/communication
- DGX-1 typically has superior performance
  - More links than GX8's PCIe bus; and higher bandwidth/link
- GX8 is very competitive for all ResNet-style workloads
  - On 8 GPUs, the GX8 can slightly outperform Unexpected
    - GX8's PCIe bandwidth saturates more quickly w.r.t. to payload size
    - For medium-sized messages, GX8 has better memory copy latency and an average of 10% better allreduceop performance
  - ResNet currently more popular than AlexNet (large allreduce)
- GX8 may be especially attractive if cost is considered

Hiring!