FMS AIML-202-1: Generative AI Part 2

UnifabriX

Leading Memory over Fabrics™ Technology

UnifabriX

Memory over Fabrics™ for AI

Ronen Hyatt, CEO and Chief Architect, UnifabriX

August 2025

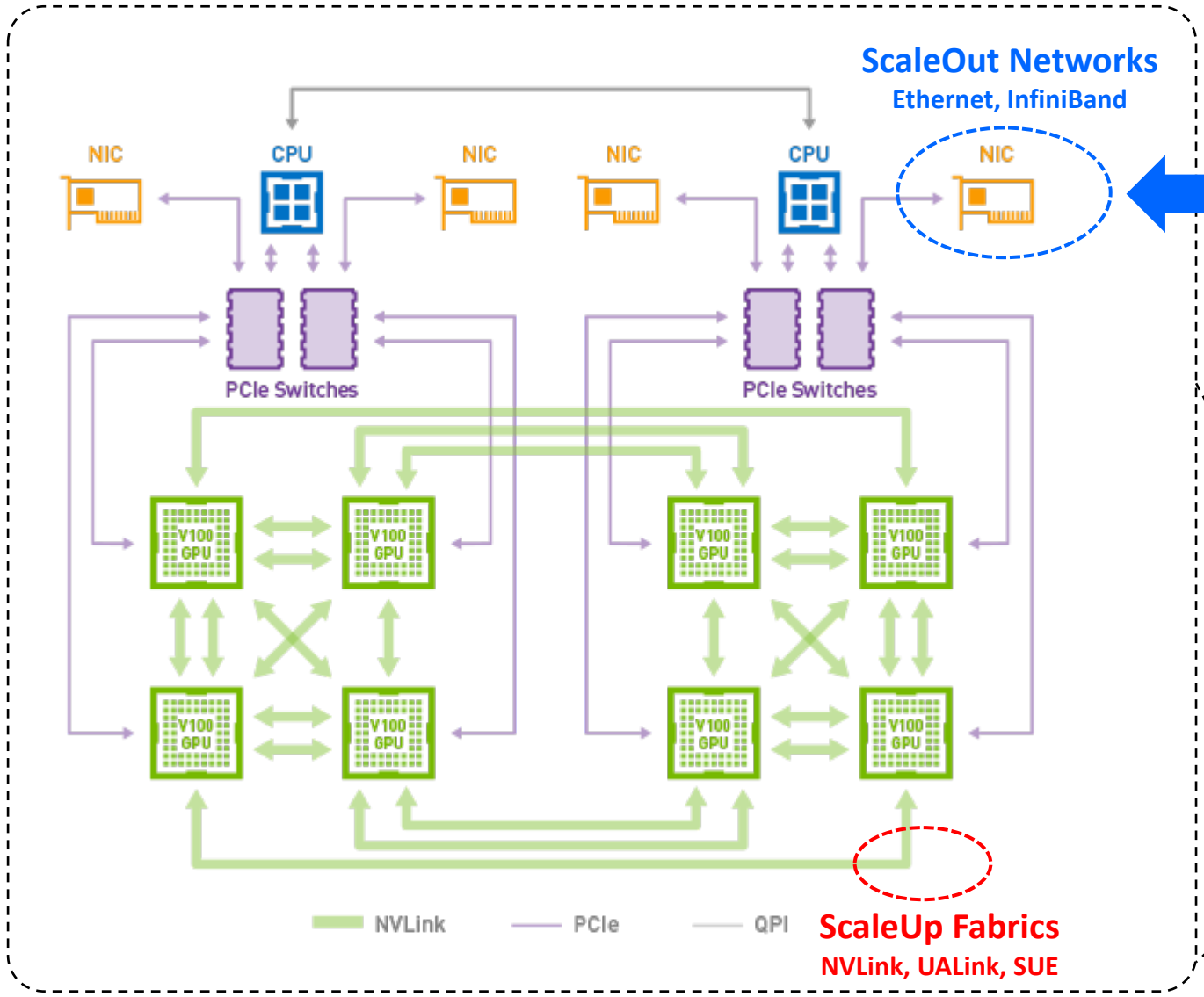
Unifabrix

/ˌjuː.niˈfæ.brɪks/

any fabric, unified memory

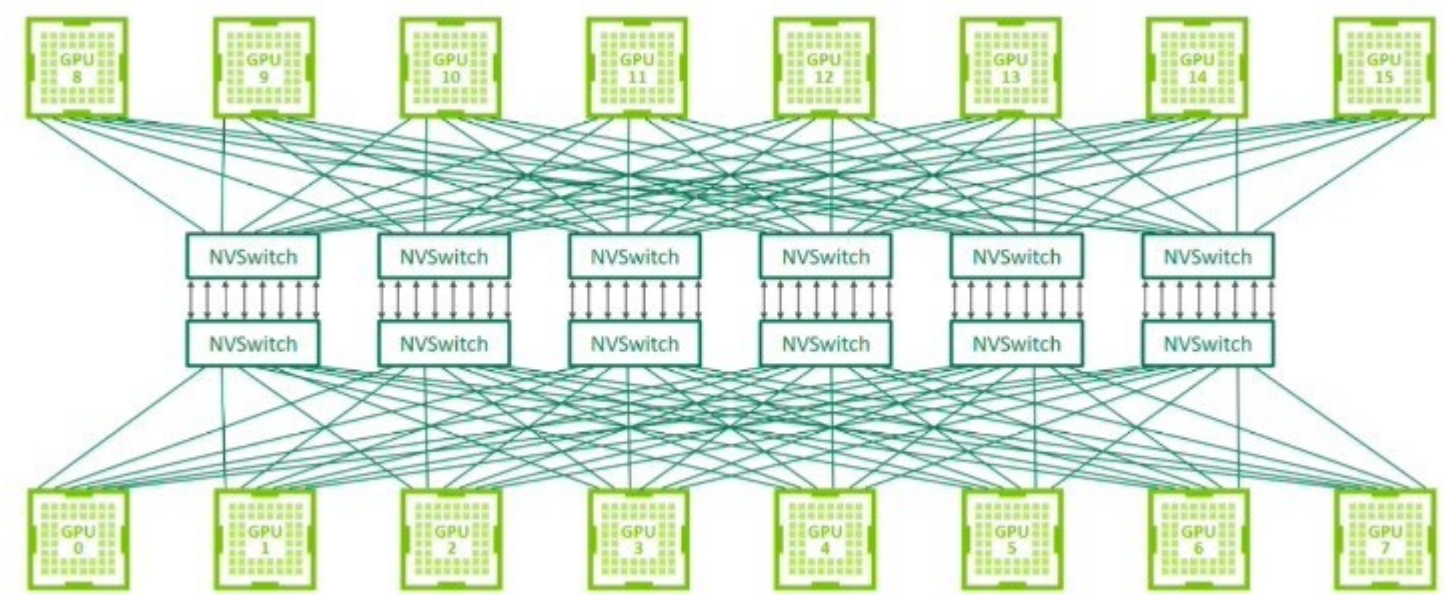
CPUs, GPUs, Accelerators

GPU-to-GPU memory transactions travel over Scaleup Fabrics



NVIDIA DGX-1 (2016)

Larger Scaleup Fabric topologies require a switch



NVIDIA DGX-2 (2018)

NVLink
NVLink Switch
NVSwitch

UALink™
Switch

SUE
Ethernet Switch

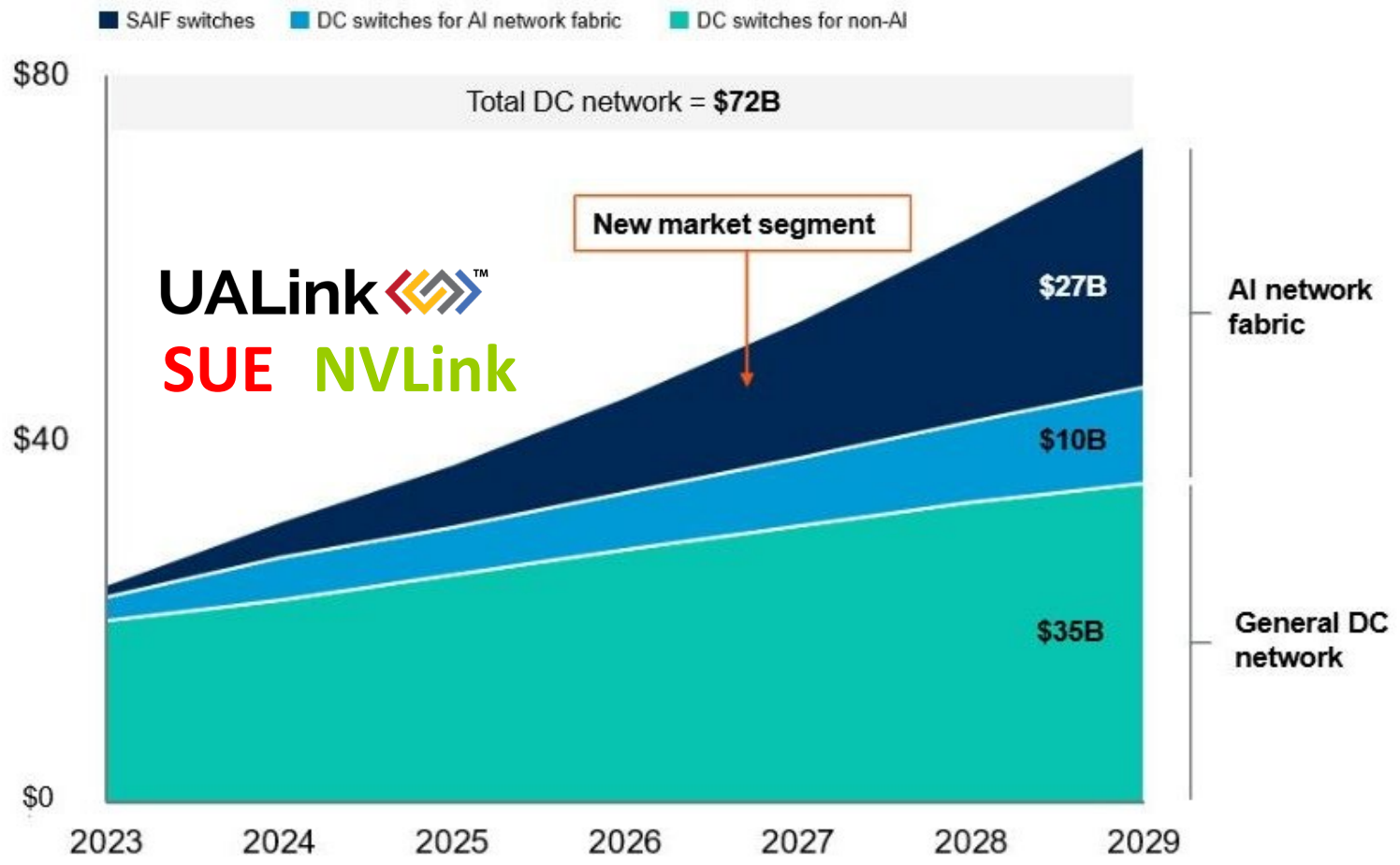
Scaleup Fabrics / SAIF (ScaleUp AI Fabrics)

Scaleup Fabrics:

- High-bandwidth
- Low-Latency
- **Memory semantics: load/store**
- **“Large GPU” software model**
- **Currently dominated by NVIDIA**

Andrew Lerner
Distinguished Analyst
3w •

Gartner July 2025



https://www.linkedin.com/posts/andrewfastlerner_scale-up-ai-fabrics-an-emerging-network-activity-7349380891481190400-e2QZ/

Scaleup is about Memory Semantics. It means higher performance

NVLink Scaleup Fabrics outperform InfiniBand Scaleout Networks by ***x2…x6***

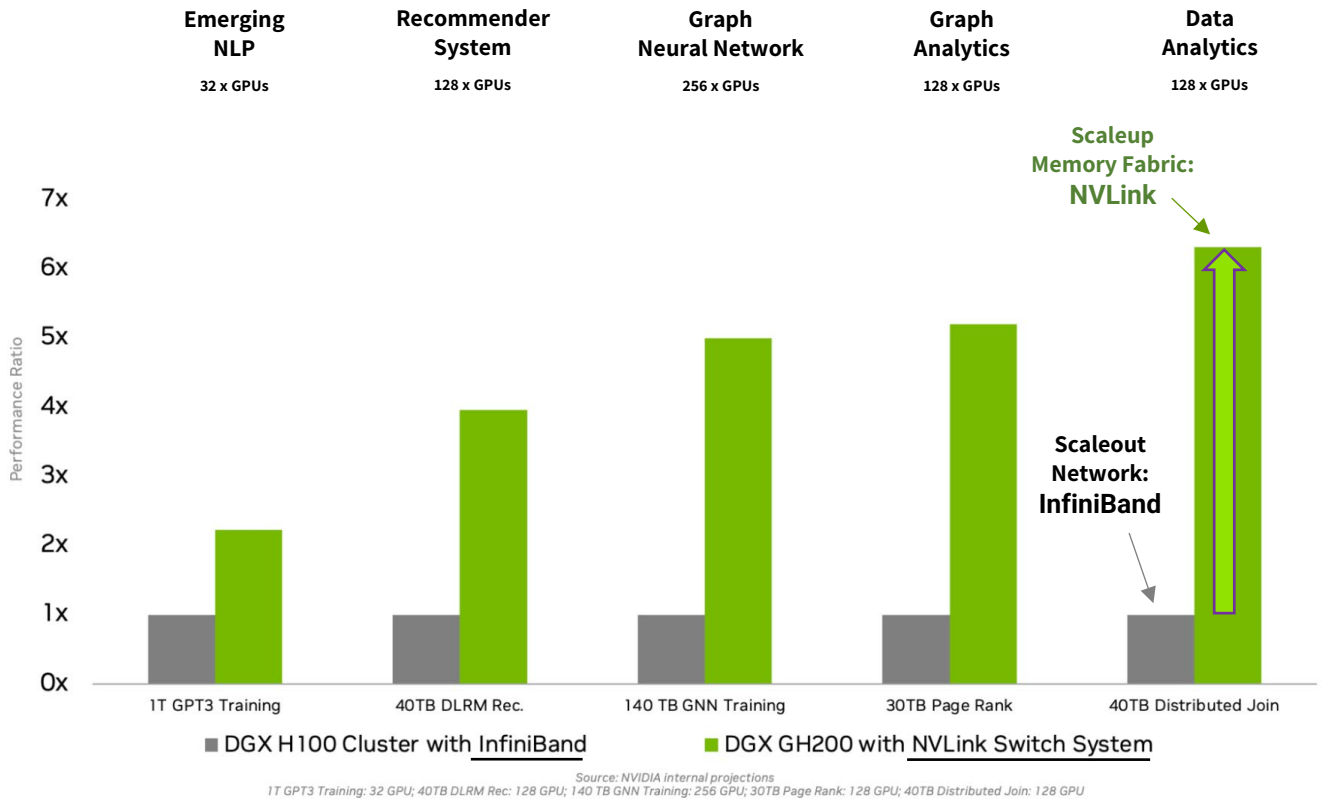
Scaleup Fabrics

Memory semantics
Native programming

Performance

Traditional Scaleout Networking

Verb/Socket semantics
Requires software refactoring

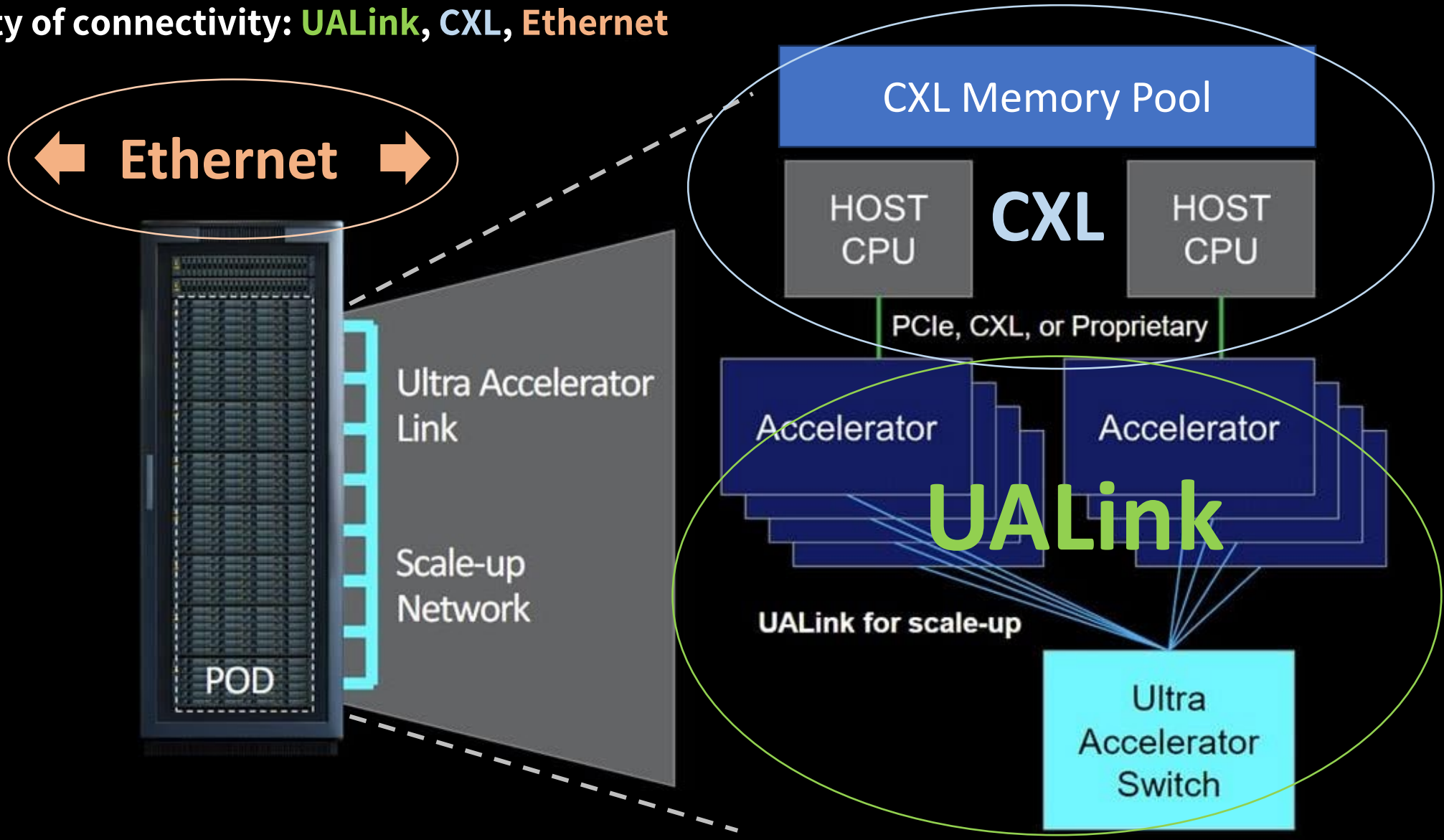


DGX GH200 Fastest for Giant Memory Models
https://developer-qa.nvidia.com/blog/announcing-nvidia-dgx-gh200-first-100-terabyte-gpu-memory-system/

Scaleup AI Fabrics are Superior to Traditional Scaleout Networking

What about CXL? CXL for AI? Definitely!

The Trinity of connectivity: UALink, CXL, Ethernet



The NVLink Ecosystem vs. the Open UALink/CXL Ecosystem

CPU Fabric: Cache Coherent

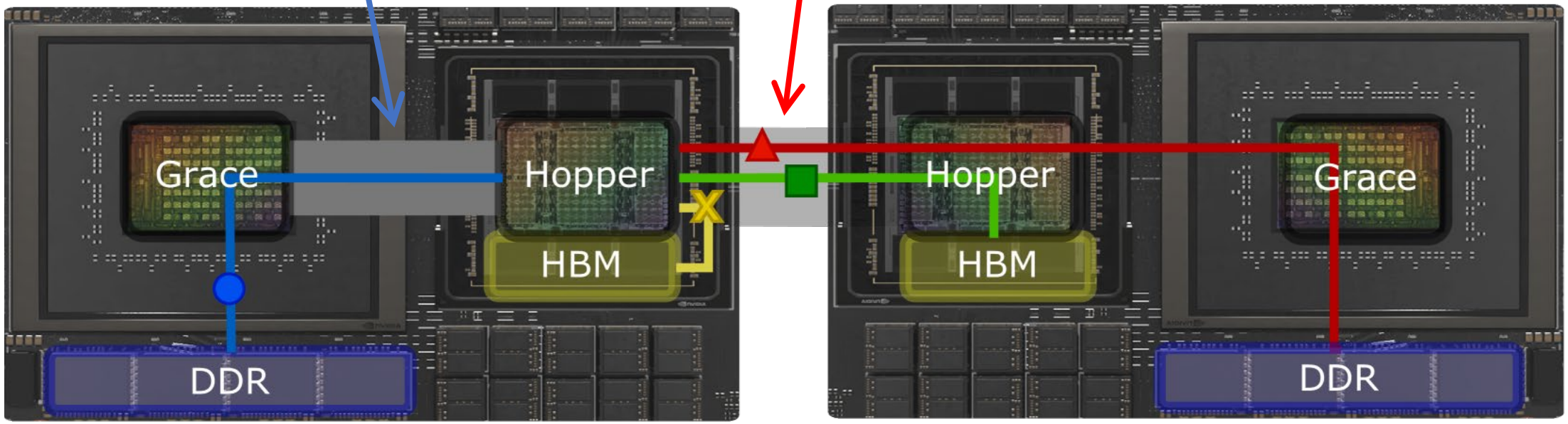
NVLink-C2C

CXL Compute Express Link

ScaleUp Fabric: I/O Coherent

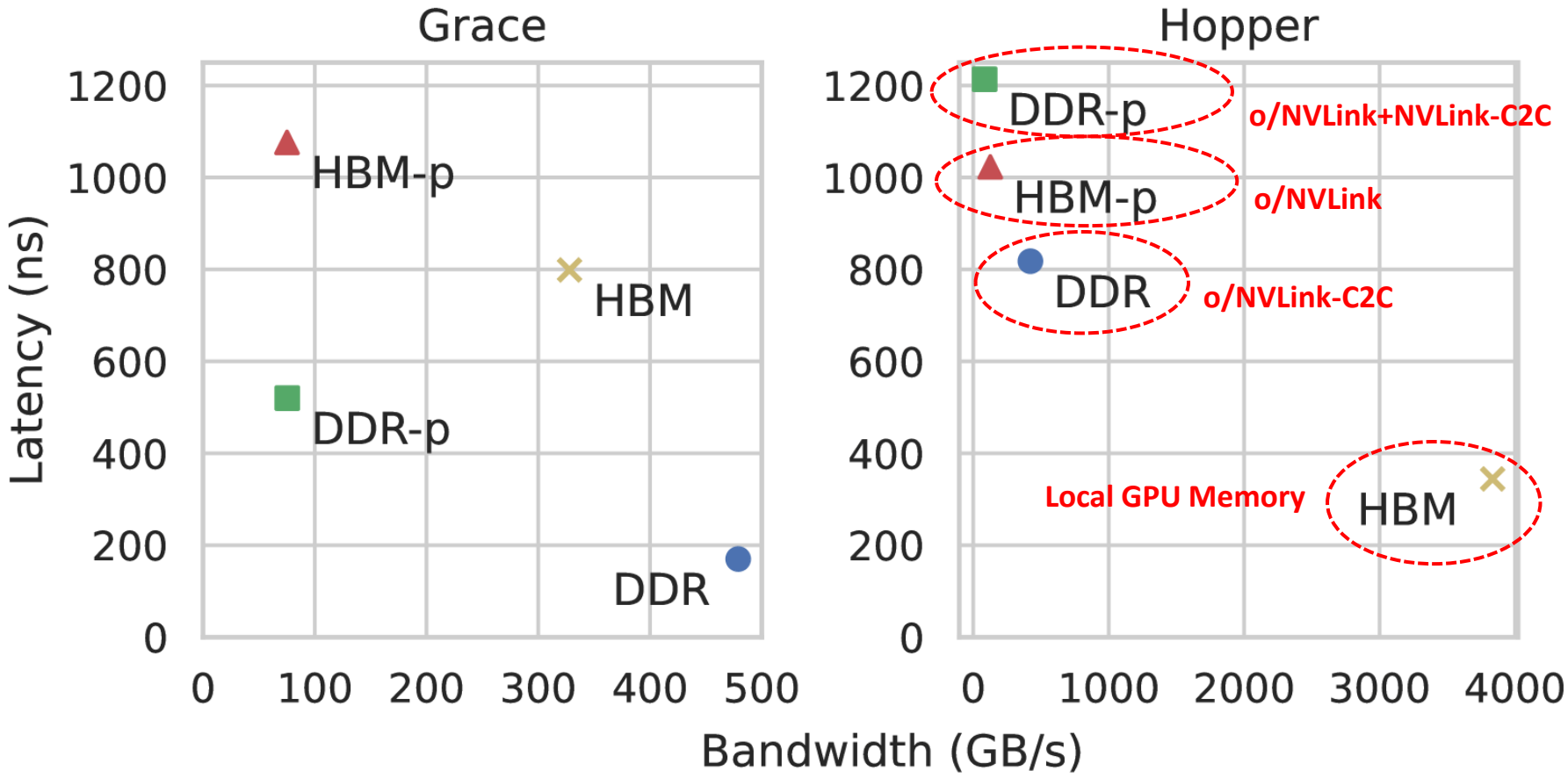
NVLink

UALink™



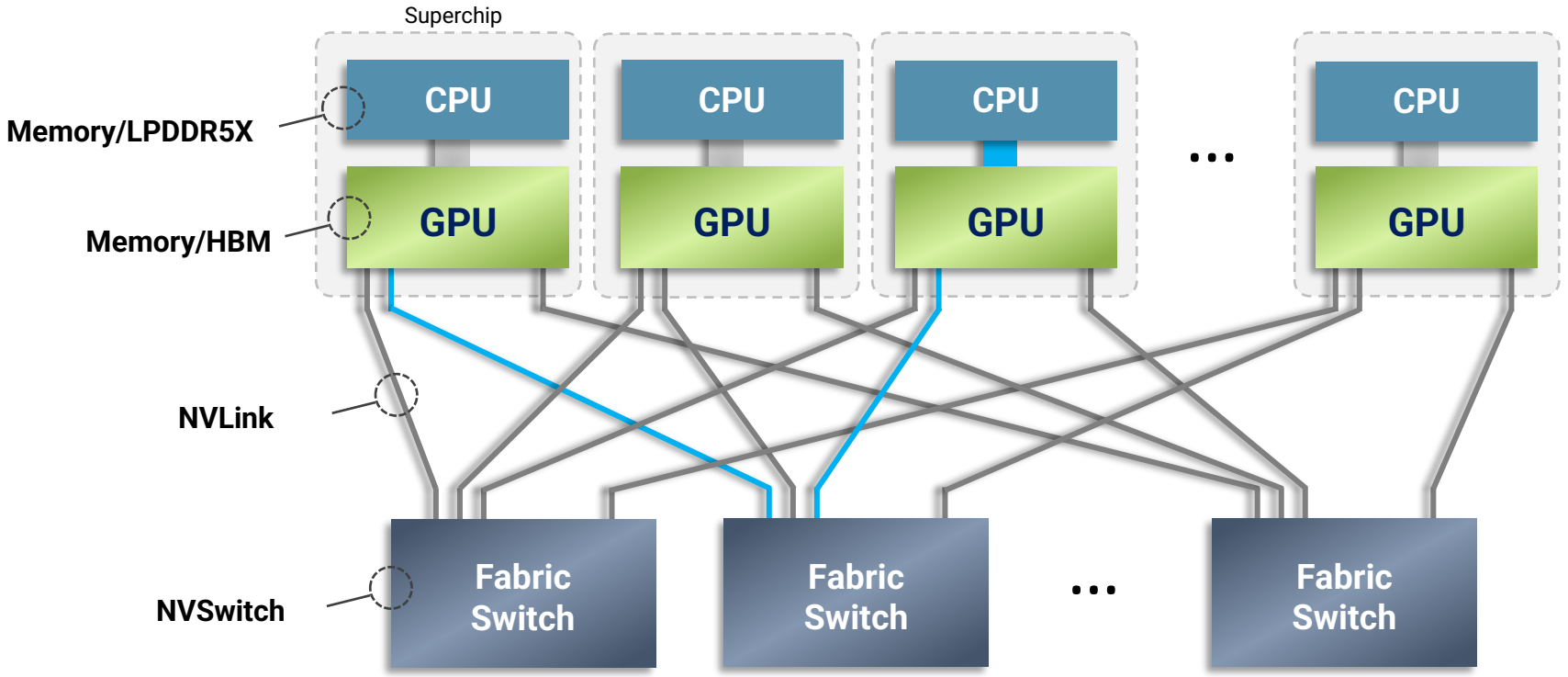
Understanding Data Movement in Tightly Coupled Heterogeneous Systems: A Case Study with the Grace Hopper Superchip
https://arxiv.org/html/2408.11556v2
https://arxiv.org/pdf/2408.11556v2

NVLink: Memory paths available for the GPU: Bandwidth and Latency



Understanding Data Movement in Tightly Coupled Heterogeneous Systems: A Case Study with the Grace Hopper Superchip
https://arxiv.org/html/2408.11556v2
https://arxiv.org/pdf/2408.11556v2

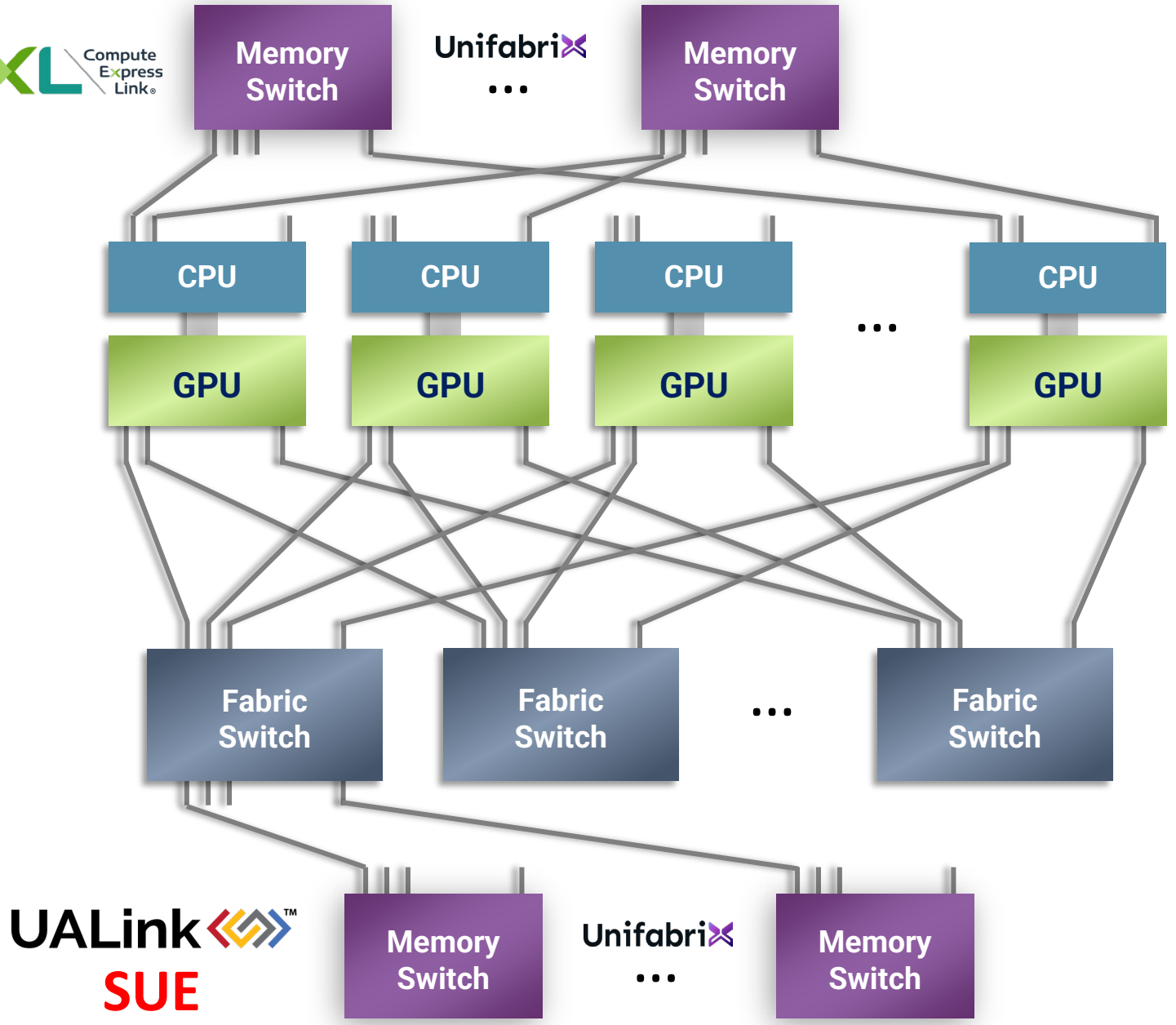
Scaleup Fabrics with Larger LLMs: Independent Scaling of Memory and Compute?



- GPUs utilize HBMs, and can borrow memory from directly-attached CPUs (e.g., Grace-Hoper superchip o/NVLink-C2C) or remote GPUs/CPUs
- Memory resides only on the edges of the fabric, in GPUs and in CPUs
- Challenges:
  - (1) **Difficult scaling**: Memory capacity and memory bandwidth cannot increase without adding more GPUs or superchip modules
  - (2) **Fair solution for training** but **expensive for Inference** which is relied upon as the primary source of revenue for AI applications

Scaleup Fabrics: Memory Pooling

Memory is fungible
and can scale
Independently of Compute
to support even larger models



UnifabriX Memory over Fabrics™ – MAX Memory Switch

Memory Pooling, Sharing, and Switching over CXL, Ethernet and UALink, powered by UnifabriX Silicon

- Fabric Independent
- Memory Pooling
- Adaptive Memory Sharing
- Transactional Switch
- Large Address Space
- Server-Grade RAS
- Link Redundancy
- PFA
- Mirroring and Striping
- Resource Interleaving
- Processing Near Memory
- Security
- Memory Encryption
- FlexMemory

Standard 2U FF

4-32 TB Memory

UALink CXL SUE

GPUs CPUs xPUs

OSFP Cabling

Scaleout:
Network Ports

Scaleup:
Fabric Ports

UnifabriX: Fabric Manager Dashboard

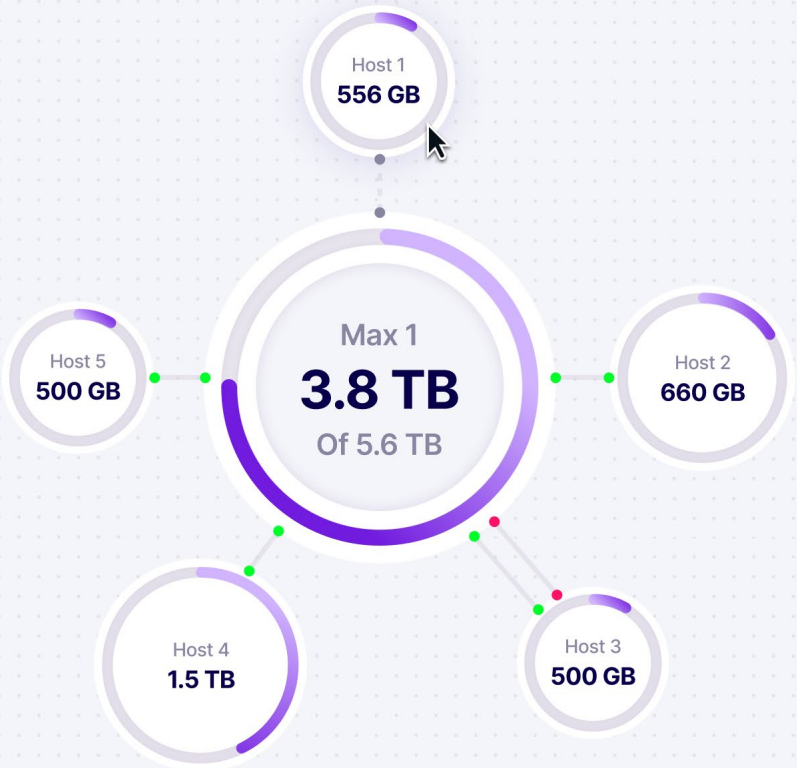
3,892 GB
(Max.01) Total Memory Provisioned

48%
(Max.01) Memory Utilization

5
(Max.01) CXL Ports

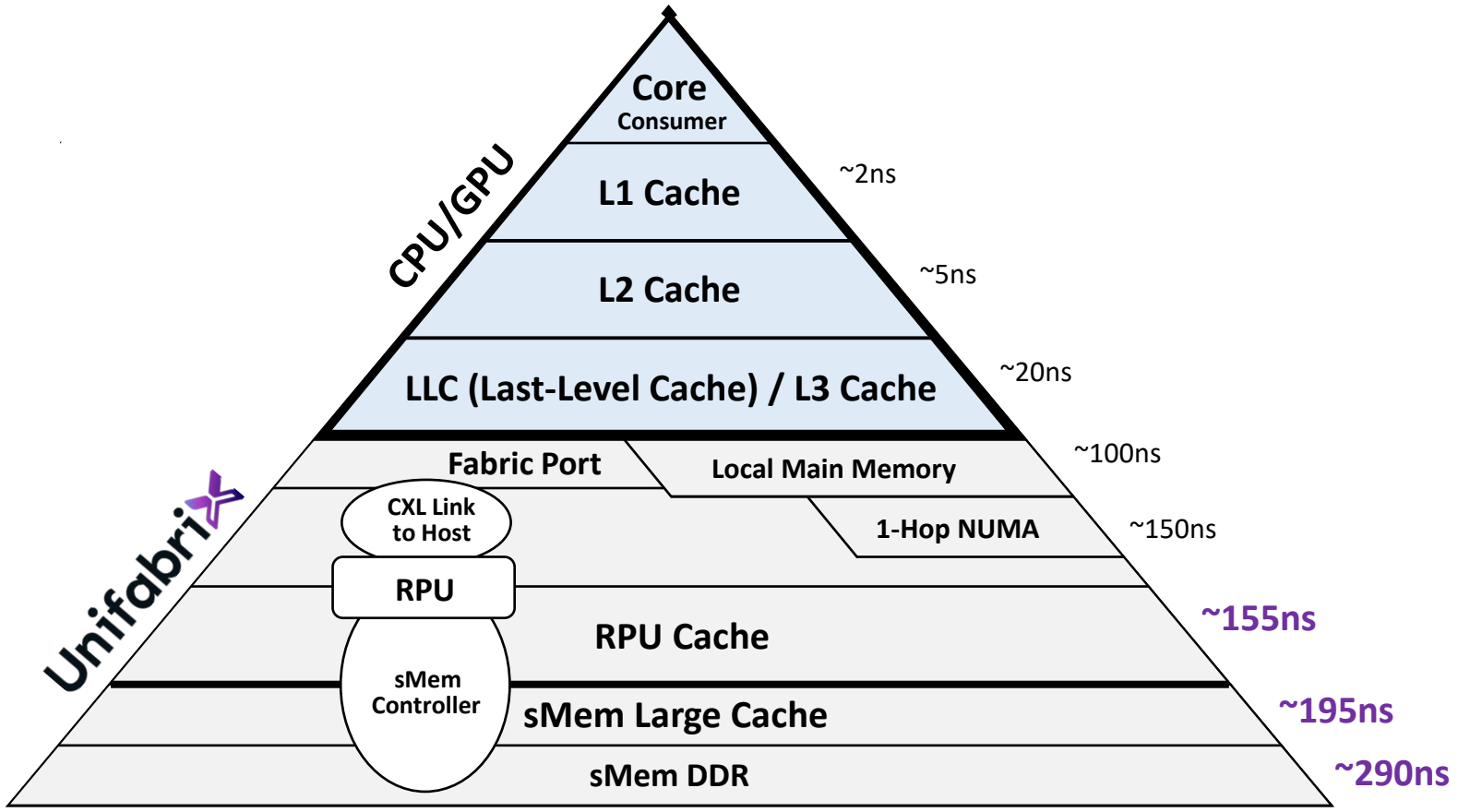
5
(Max.01) Active Hosts

Search..



Optimal · No connection · No link · Degraded

Fabric-Attached Memory: Round-Trip Latency



Fabric-Attached Memory: Optimal Link Utilization: CXL o/PCIe Gen5

Table 4. Realizable Bandwidth in GB/s with CXL.mem for Different Traffic Mixes Across CXL 2.0 68 B Flits and CXL 3.0 256 B and 128 B Flits for a x16 link at 32 GT/s

Data B/W (GB/s) x16 @ 32G or x8 @ 64 G (Raw B/W: 64 GB/s/dir)		Type-3		Type-2	
		M2S	S2M	M2S	S2M
1R/0W	68B Flit	0	53.5	0	48.1
	256B Flit	0	54.0	0	50.3
	128B LO	0	51.9	0	49.1

```
pilmeni:~ # numactl -m3 test/mlc --peak_injection_bandwidth -k56-72 -b500m
Intel(R) Memory Latency Checker - v3.11a
Command line parameters: --peak_injection_bandwidth -k56-72 -b500m

Using buffer size of 500.000MiB/thread for reads and an additional 500.000MiB/thread for writes

Measuring Peak Injection Memory Bandwidths for the system
Bandwidths are in MB/sec (1 MB/sec = 1,000,000 Bytes/sec)
Using all the threads from each core if Hyper-threading is enabled
Using traffic with the following read-write ratios
ALL Reads        :      53174.1
3:1 Reads-Writes :      66903.3
2:1 Reads-Writes :      70883.3
1:1 Reads-Writes :      61888.6
Stream-triad like:      69637.6
```

GNR: 99.4%
of Link Bandwidth
53,174.1 / 53,500 = 0.9939

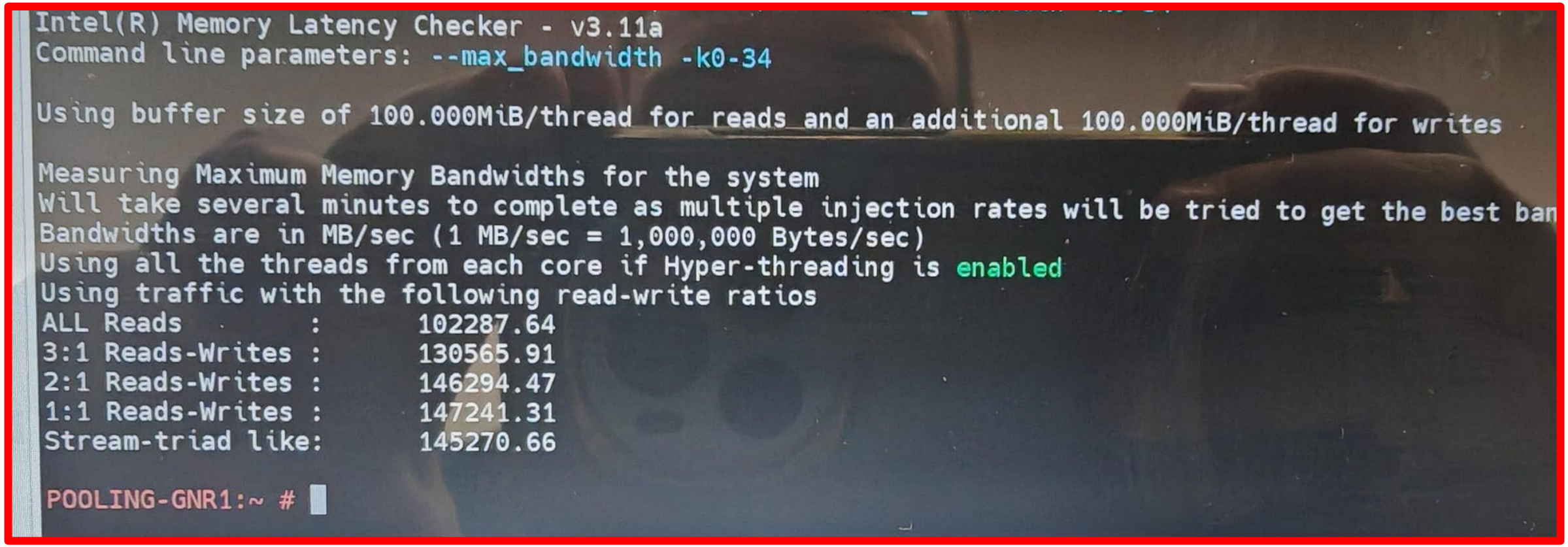
```
  2:  24  14  10
kugel:~ # numactl -m 2 test/mlc --max_bandwidth -b500m -t5 -k16-31
Intel(R) Memory Latency Checker - v3.11a
Command line parameters: --max_bandwidth -b500m -t5 -k16-31

Using buffer size of 500.000MiB/thread for reads and an additional 500.000MiB/thread for writes

Measuring Maximum Memory Bandwidths for the system
Will take several minutes to complete as multiple injection rates will be tried to get the best bandwidth
Bandwidths are in MB/sec (1 MB/sec = 1,000,000 Bytes/sec)
Using all the threads from each core if Hyper-threading is enabled
Using traffic with the following read-write ratios
ALL Reads        :      46034.66
3:1 Reads-Writes :      54899.42
2:1 Reads-Writes :      61287.77
1:1 Reads-Writes :      62123.36
Stream-triad like:      53427.39
```

EMR: 95.7%
of Link Bandwidth
46,034.66 / 48,100 = 0.9570

An Introduction to the Compute Express Link (CXL) Interconnect
Debendra Das Sharma, Robert Blankenship, Daniel Berger
https://dl.acm.org/doi/pdf/10.1145/3669900

Fabric-Attached Memory: Link Aggregation: >147GB/s

Read:Write Bandwidth



Visit our Booth

#940

Thankyou

Unifabrix

/ˌjuː.niˈfæ.brɪks/

any fabric, unified memory