LIGHTELLIGENCE

Harnessing light to power new possibilities

Flash Memory Summit

Advantages of optical CXL for disaggregated compute architectures

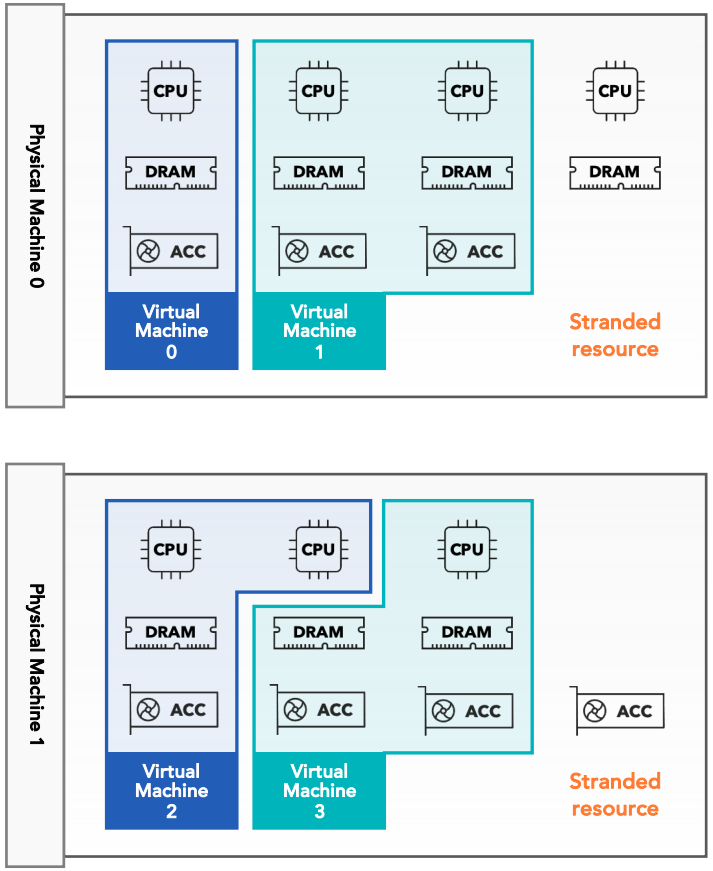
Ron Swartzentruber
Director of Engineering

Agenda

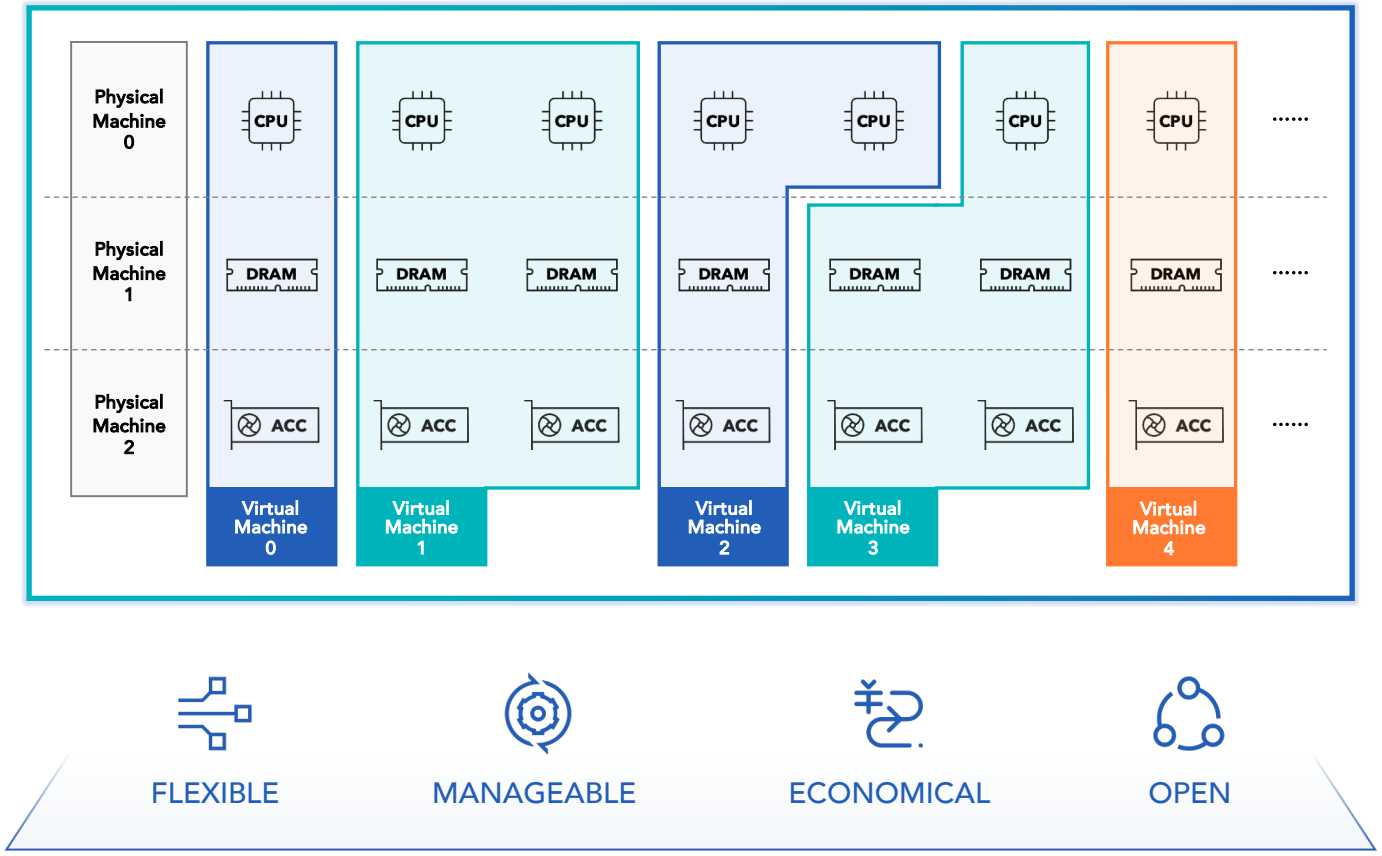
- Memory centric shift in the data center
- AI Large Language Model growth
- Need for optical CXL technology
- Case study: OPT inference benefits using optical CXL

Disaggregation is the Future for Datacenter

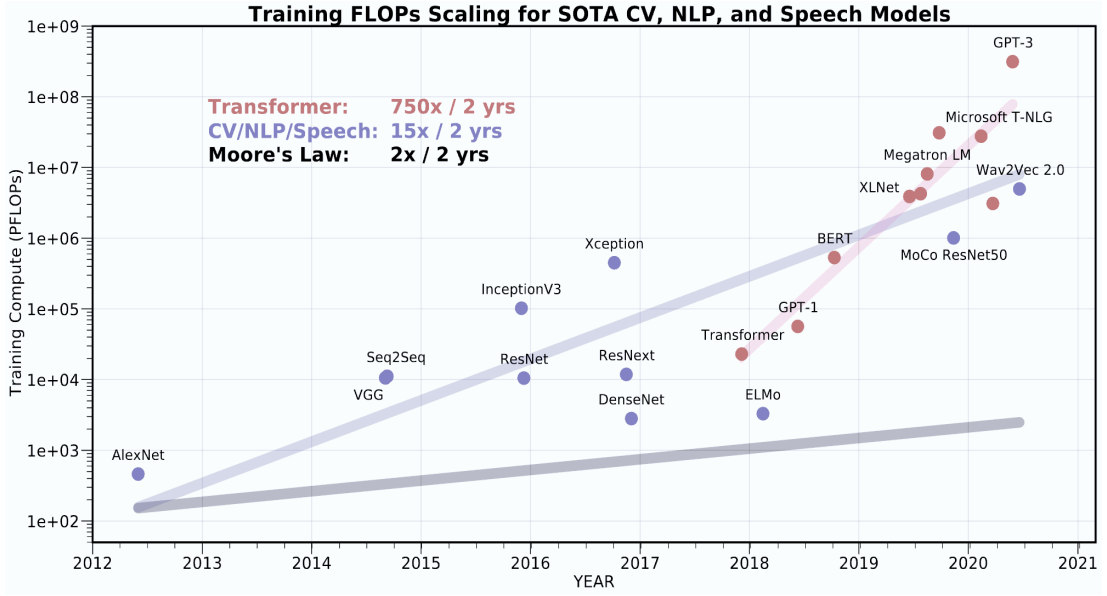
Traditional Datacenter



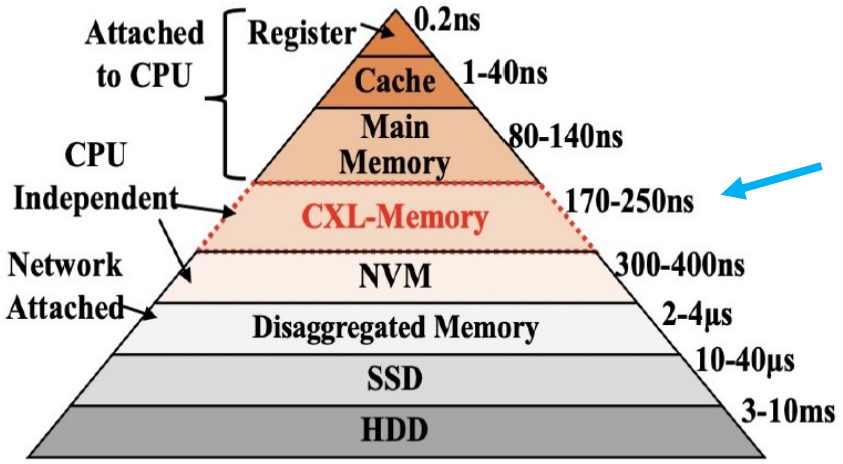
Disaggregated Datacenter



AI trends



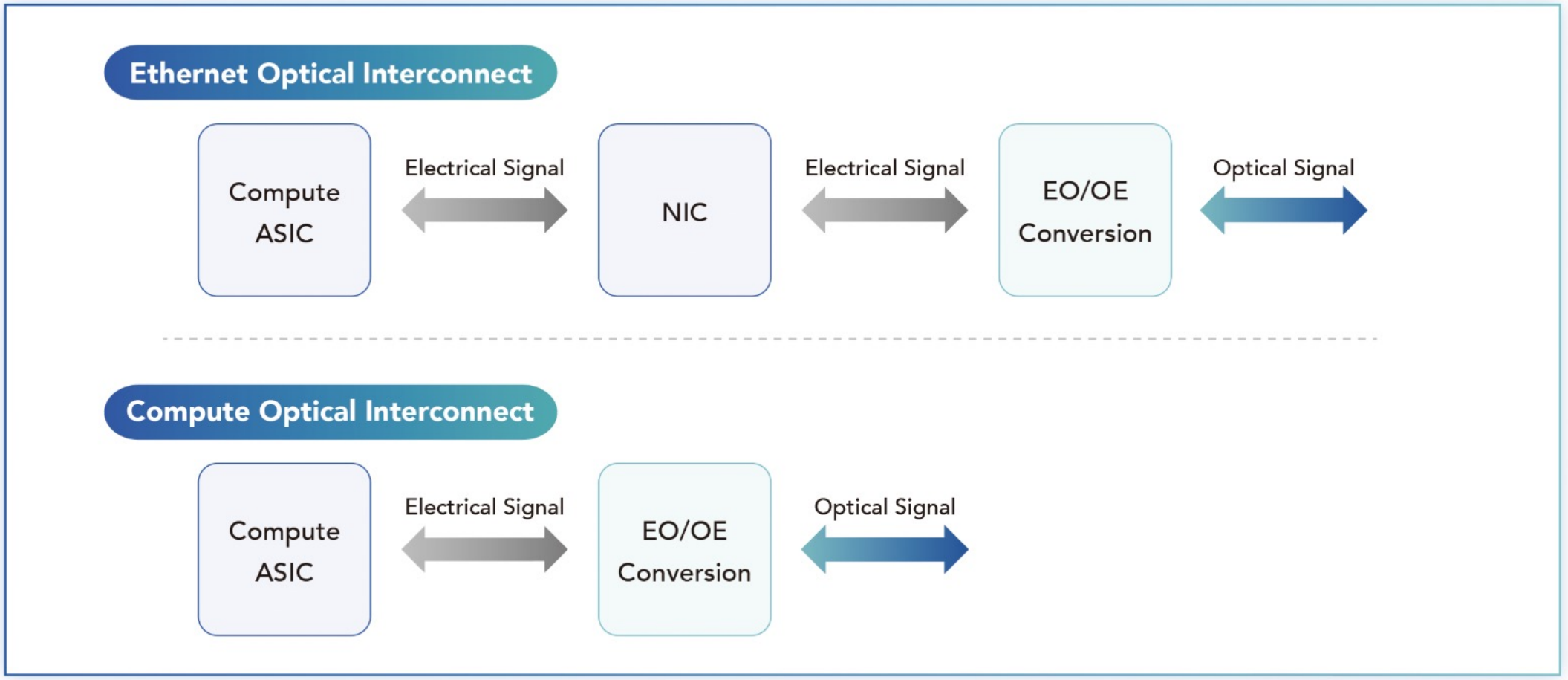
Source: https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8



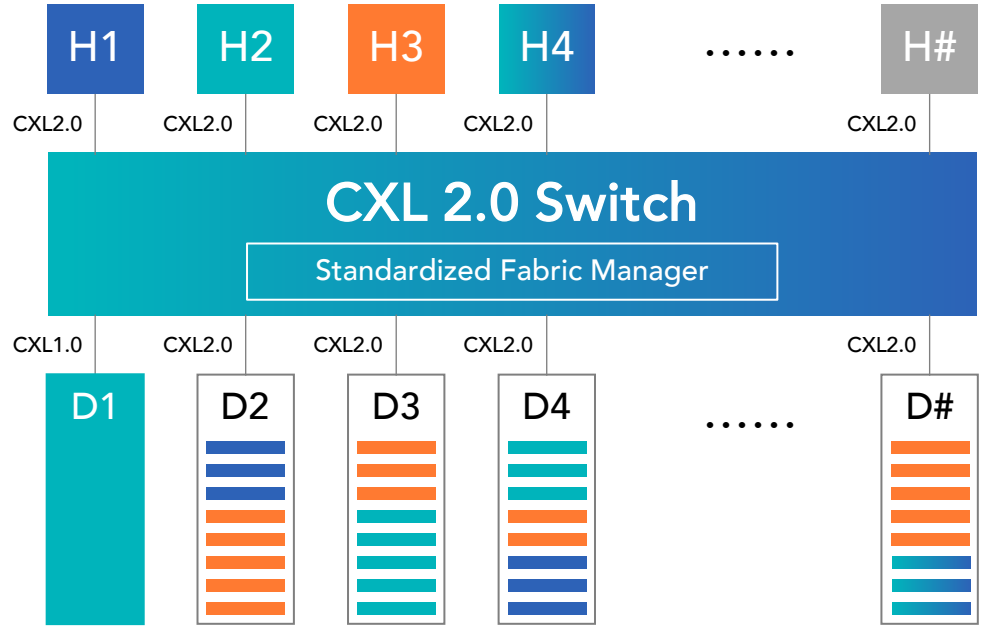
Source: https://hc34.hotchips.org/assets/program/tutorials/CXL/Hot%20Chips%202022%20CXL%20MemoryChallenges.pdf

- AI and Large Language Models will continue to grow and consume more compute
- Disaggregated memory architectures are required in order to continue to scale
- Optical interconnect is required to extend reach

Optical Interconnect Latency



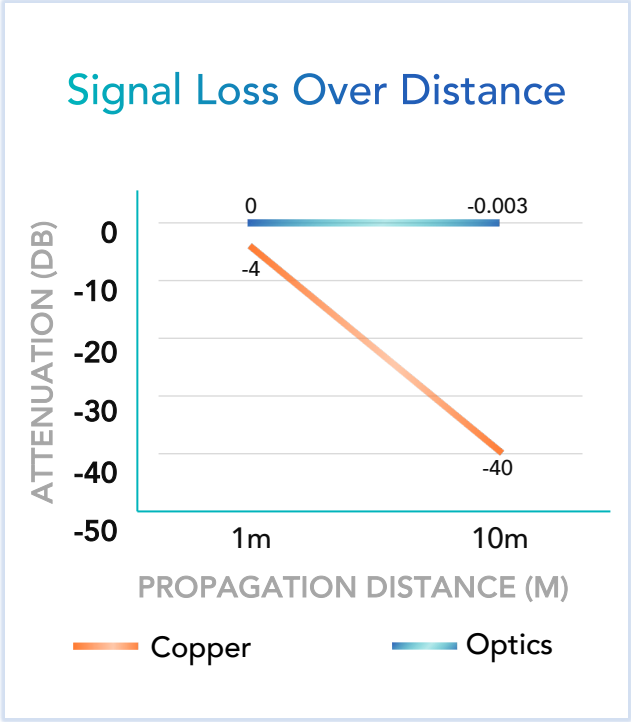
CXL is the Predominant Standard for Disaggregation



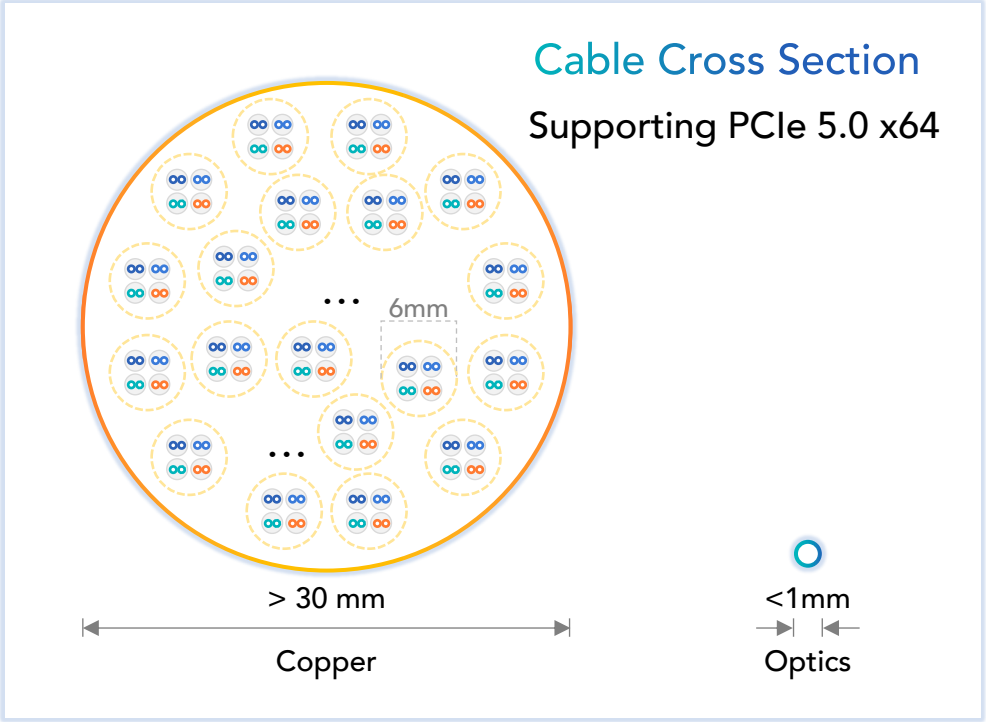
	Cache-coherence	Latency	Memory decouple
CXL	Yes	~100ns	Supported
RDMA (ethernet)	No	~3μs	Not supported

Wide adoption from most major industry players

Optical CXL is Required for Scaling



Assuming AWG26 wire, PCIe 5.0 signal



32 cables with diameter > 6mm (CAT8)

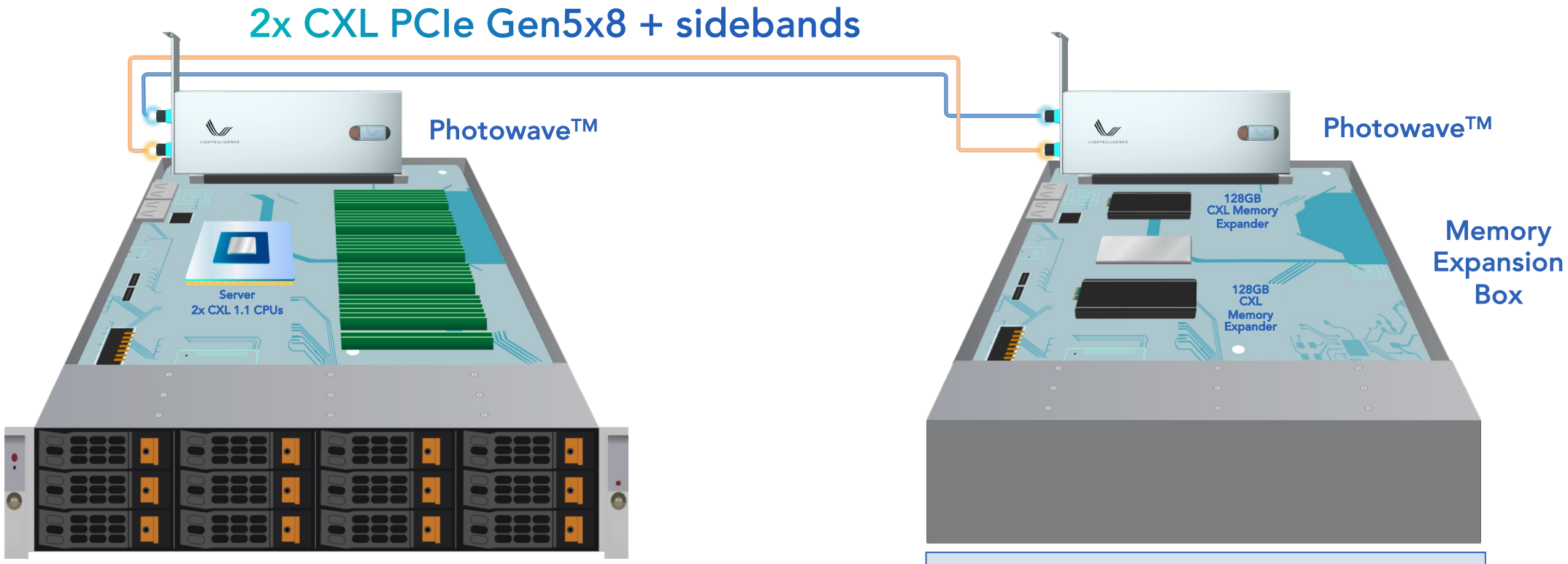
16 fibers with diameter of 0.125mm

Copper cable struggles to support CXL scaling beyond a few servers.

Optical CXL in the Datacenter

Photowave™ CXL over optics
Low Latency, High Bandwidth, Data Center Reach

Case study: LLM Inference



- 2U Supermicro server
- 2x AMD Genoa CXL 1.1 CPUs
- MemVerge Memory Tiering and Pooling Software
- Nvidia GPU running LLM inference
- All VMs access to CXL memory
- Secure application, encrypted data

- 2x Samsung 128GB Memory Expanders each with CXL/PCIe Gen5x8 link
- Upgrade to Memory Pooling in Q4

Demo @ FMS booth #915

LIGHTELLIGENCE

LLM Model List

Flash Memory Summit

Model	Weight Memory(float16)	KV-Cache per sample(float16)	Activation per sample(float16)	Context length
OPT-1.3B	2.4 GB	0.095 GB	0.002 GB	512
OPT-13B	23.921 GB	0.397 GB	0.005 GB	512
OPT-30B	55.803 GB	0.667 GB	0.007 GB	512
OPT-66B	**122.375 GB**	**1.143 GB**	**0.009 GB**	**512**
OPT-175B	325GB	2.285GB	0.012GB	512

Entire OPT-66B model fits within one 128GB CXL memory expander

KV-cache Size: data_type * dimension* num_layers* batch_size * Context_len * 2

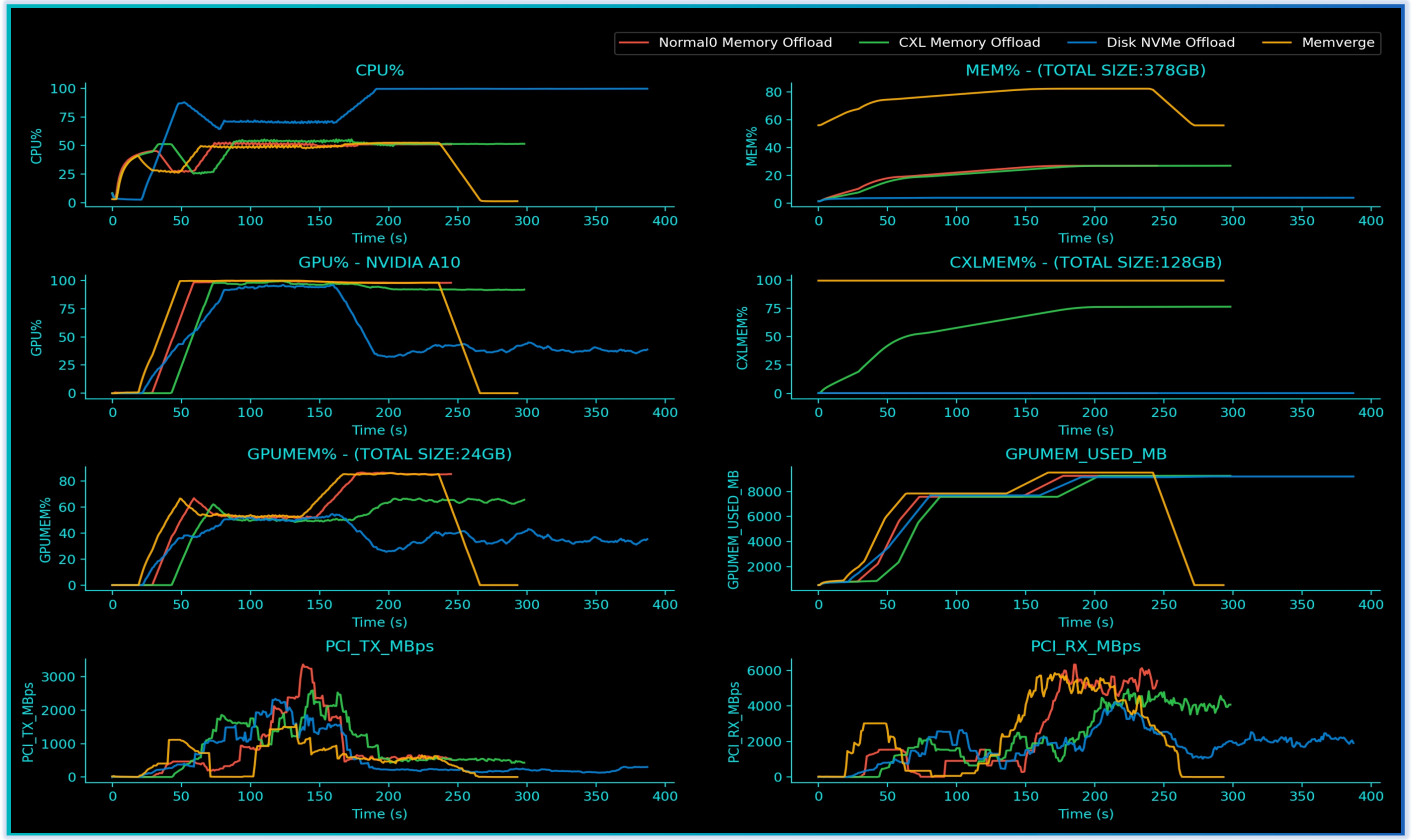
e.g., for opt-1.3B, FP16 -> 2Bytes * 2048 * 24 * 1 * 512 * 2 = 100,663,296 Bytes

Activation Size: data_type * dimension * batch_size * Context_len

LIGHTELLIGENCE

Results

Flash Memory Summit



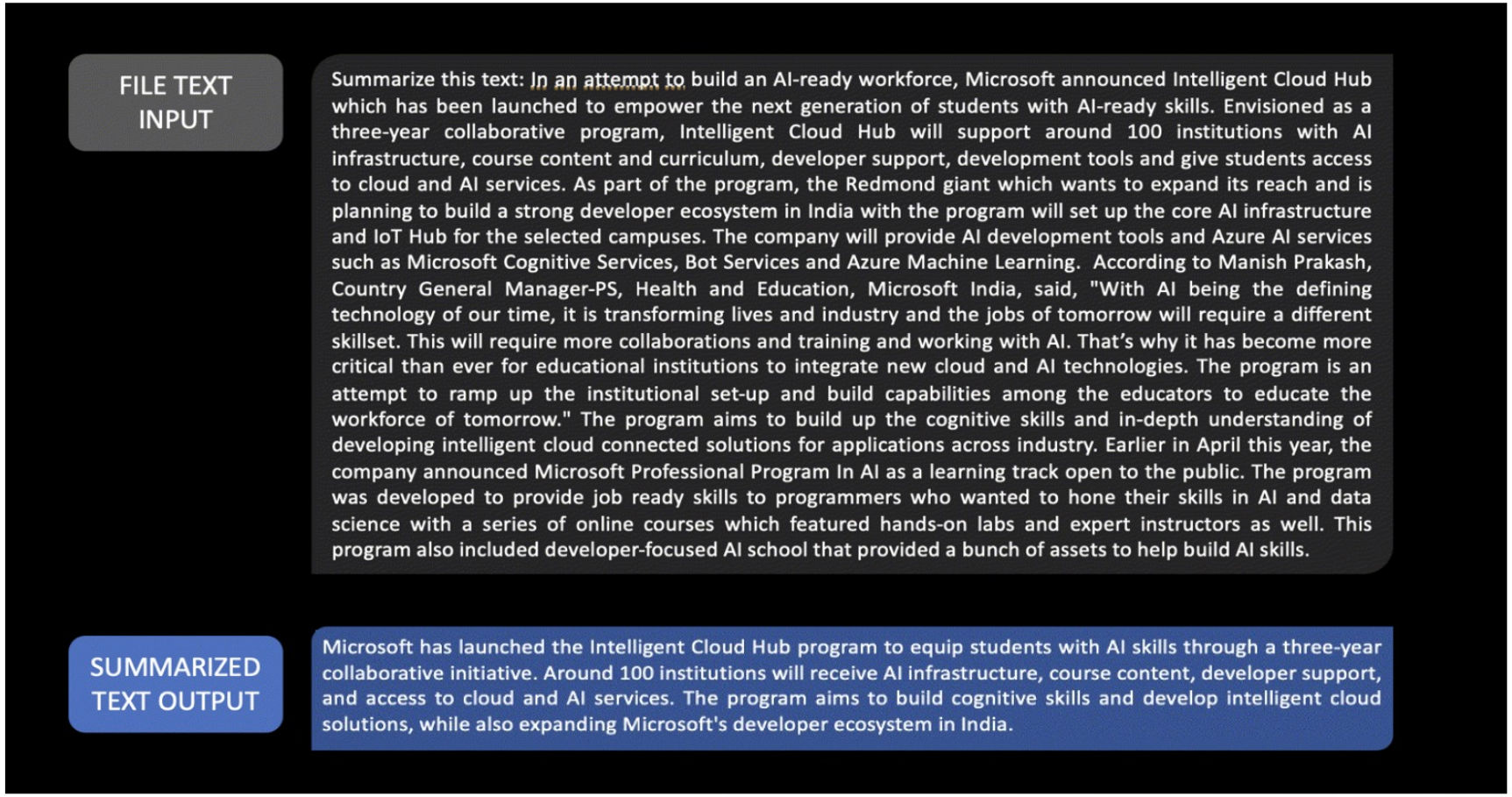
OPT-66B model results	Disk (NVMe)	CXL Memory	System Memory	MemVerge 60:40 Policy
Decode Throughput (Tokens/s)	1.984	4.859	6.216	6.237
Decode Latency (s)	338.7	138.2	108.1	107.7

CXL Mem. Achieves ~2.4x Higher Throughput than Disk

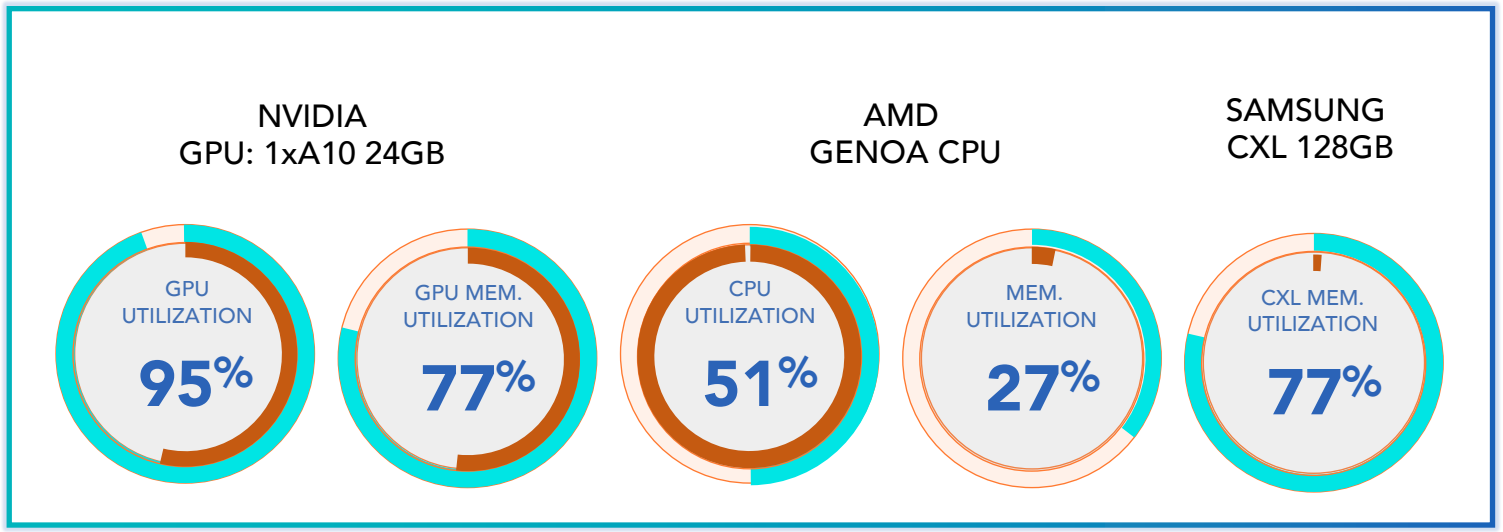
- **Average GPU-PCIe-Transmit-Traffic:**
  CXL: 882MB/s, System Memory 857MB/s, Disk: 582MB/s, MemVerge: 493MB/s
- **Average GPU-PCIe-Receive-Traffic:**
  CXL: 2365MB/s, System Memory: 2609MB/s, Disk: 1887MB/s, MemVerge: 2173MB/s

PHOTOWAVE™ OPTICAL CXL MEMORY EXPANDER

NEWS TEXT SUMMARIZATION



PHOTOWAVE™ OPTICAL CXL MEMORY EXPANDER



PARAMETERS

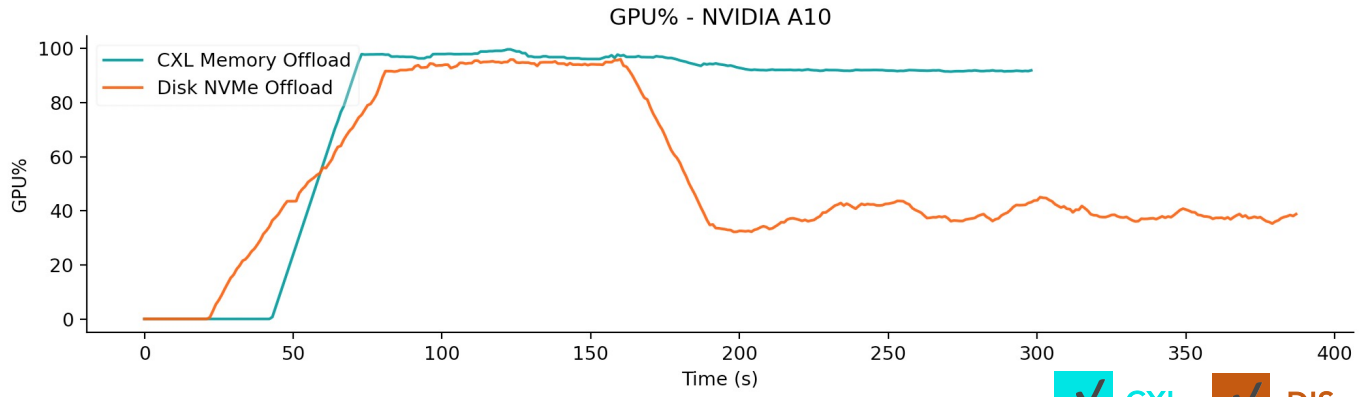
RUN MODE: **CXL**

INFERENCE ENGINE: FLEXGEN

WEIGHTS: 122.375GB

KV CACHE: 109.688GB

OPT-66B MODEL PROGRESS 99%



CXL NVMe

CXL DISK

Summary of Results

CXL memory offloading is efficient and beneficial

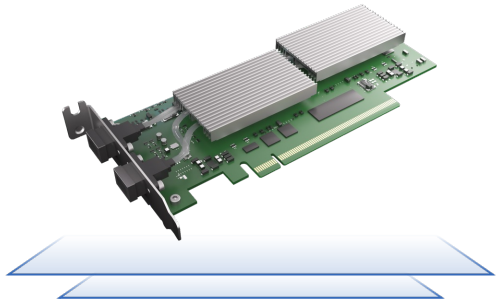
- LLM inference case study
- Allows use of lower cost memory

Similar performance compared to pure system memory

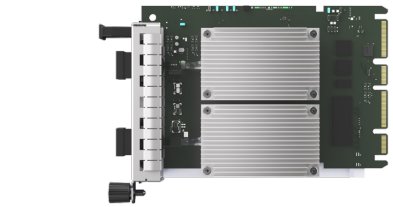
2.4x performance advantage compared to SSD/NVMe disk offloading

>1.92x TCO improvement using inexpensive GPUs at similar throughput

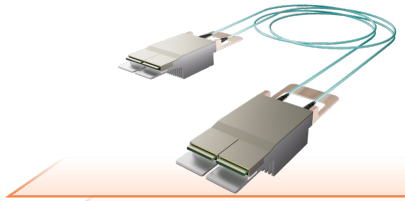
Photowave™ Form Factors



Low Profile PCIe Card



OCP 3.0 SFF Card



Active Optical Cables

Product Suite Features

- CXL 2.0/PCIe Gen5 x16
- Jitter reduction, SI cleanup
- Sideband signals over optics
- x8, x4 or x2 bifurcation
- End-to-end latency:
  - Card: <20ns + TOF
  - AOC: <1ns + TOF

Endnotes

Algorithm & Software

- LLM: OPT-66B
- Batch size = 24
- Context length = 512
- Output length = 8
- FlexGen

Hardware configuration

Super Micro Server
- AMD EPYC 9124 16-Core CPU
- Samsung DDR5 4800 MT/s
- MEM0 size: 256GB
- MEM1 size: 256GB
- Bandwidth: 307GB/s

Nvidia GPU
- Gen4x16, DMEM size: 24GB
- Bandwidth: 32GB/s

Samsung NVME
- Gen4x4, MEM size: 1.92TB
- Bandwidth: 8GB/s

Samsung CXL Memory
- Gen5x8, MEM size: 128GB
- Bandwidth: 32GB/s