How Ultra Ethernet and UALink Accelerate Token-to-Token Performance

Rajesh Arsid
Principal Engineer,
Synopsys Inc.

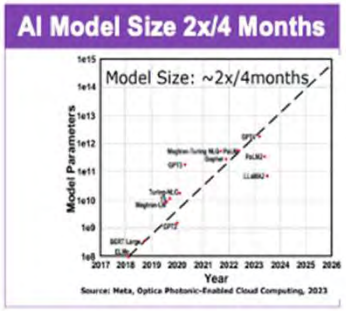
AGENDA

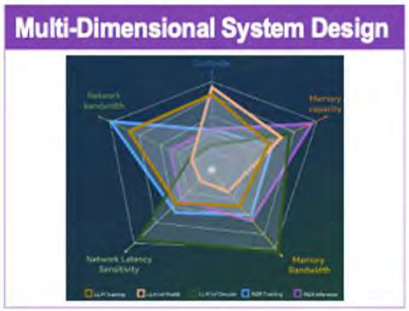
- AI infrastructure bottleneck
- Need for enhanced interconnect technology
- UA link
- Ultra Ethernet
- Conclusion

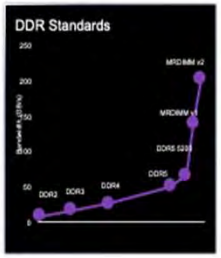
Source https://www.tomshardware.com/tech-industry/ualink-consortium-officially-incorporates-nvlink-competitor-headed-by-amd-and-intel-opens-doors-to-contributor-members

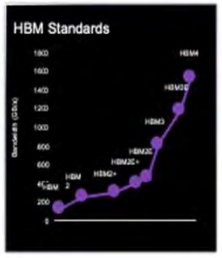
AI INFRASTRUCTURE BOTTLENECK

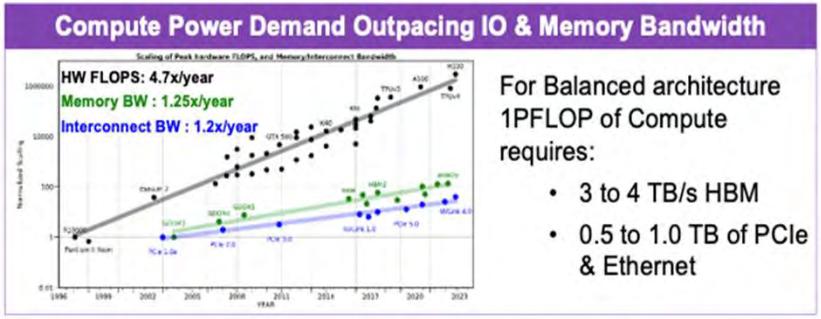
1. Exponential growth in Compute demand
2. Despite parallelization training time of models have raise from weeks to months
3. Model parameters doubling ~3-4 months
4. System design specifications are exhausting
5. Design complexities are increasing

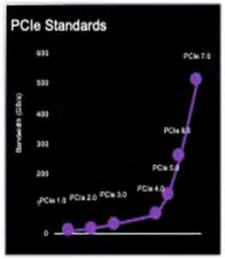












AI infrastructure needs enhance Interconnect technology to meet current and future demands

NEED FOR ENHANCED INTERCONNECT TECHNOLOGY

1. Deep learning models continuously learn from feedback loops by fine-tuning the model
2. Split data across multiple GPUs and Multithreading for CPU operations and sometimes multiple machines at larger scale

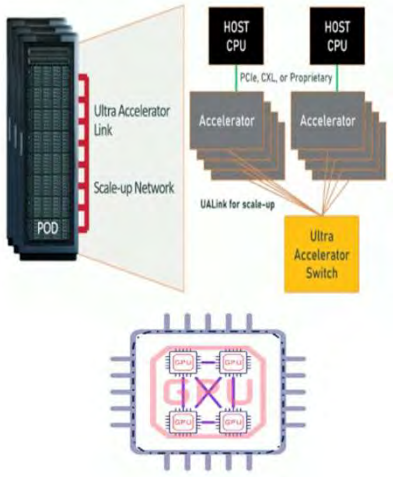
AI/ML life cycle

Source: Tech Field Day: High Performance Ethernet NIC for AI/ML by Hemal Shah

GPU performance influences the timeline of deep learning

UA LINK – (SCALE-UP)

1. Scale-Up enables the ability to make several XPUs/GPUs act like one giant XPU/GPU to complete the task
2. Enables memory sharing and synchronization b/w accelerators
3. Direct load, store and atomic operations enabled b/w accelerators
4. Low Latency, high bandwidth fabric for 100's of accelerators in a POD

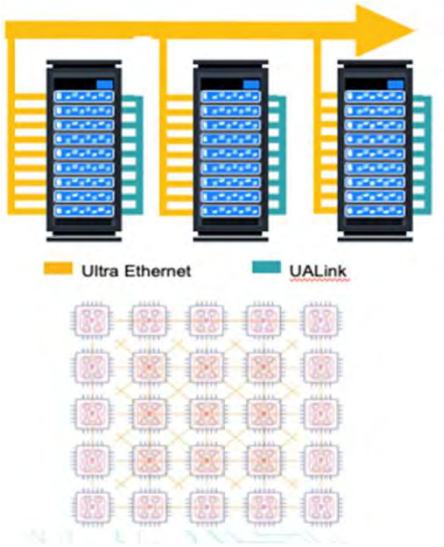


Accelerator-to-accelerator links

Open-source interconnect technology developed to scaleup accelerators for AI work loads

ULTRA ETHERNET – (SCALE-OUT)

1. XPU to XPU communication is critical and requires special consideration at large scale
2. High BW, Multi-Path, Open Standard, Highly Configurable interface
3. Delivers unparalleled speeds essential for advanced AI clusters
4. Implements intelligent congestion control for managing intense burst traffic
5. Lightweight and Low Latency
6. Focused on AI workload resource sharing & synchronization between 1M endpoints



cluster-to-cluster links

Open-source, high-performance networking technology developed by the Ultra Ethernet Consortium to offload AI workloads and HPC.

CONCLUSION

Token-to-Token Performance

UALink

1. **Direct Peer-to-Peer -** token exchange directly b/w accelerators
2. **Efficient synchronization-** memory pooling
3. **Optimized for AI work loads -**Rapid token passing

Ultra Ethernet

1. **Low Latency -** microseconds latency
2. **High Throughput -** Rapid token exchange
3. **Scalability -** Bandwidth aggregation across all connections

THANK YOU