The First Direct Mesh-to-Mesh Photonic Fabric

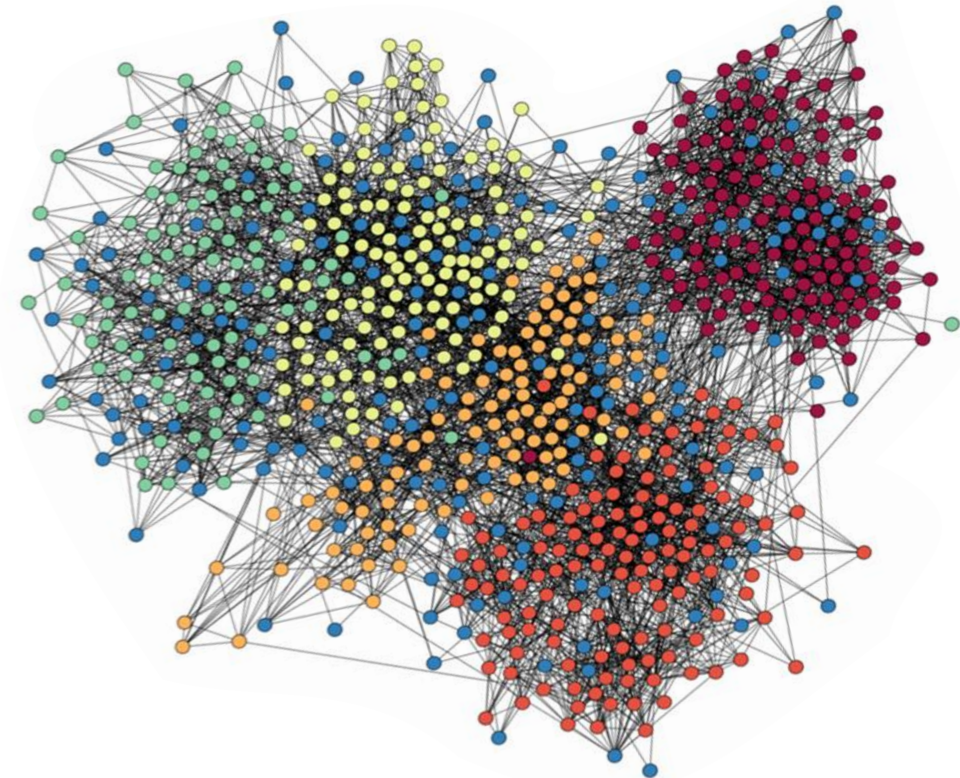
Jason Howard
Intel Principal Engineer

Outline

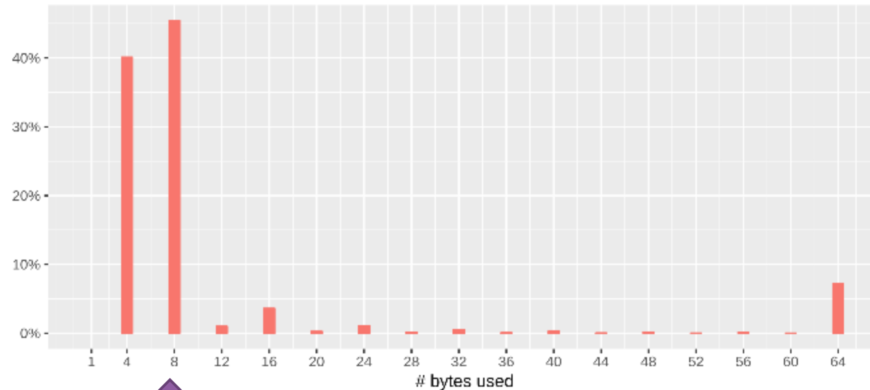
- Motivation
- Key technologies
- Architecture overview
- Testing and characterization
- Summary

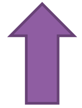
Motivation: Improved Performance in Graph Analytics

- Understand the complex relationships both within and between data sources.
- DARPA creates HIVE program for hyper-sparse data:
  - Petabyte scale graph analytics,
  - 1,000x Perf/w compared to traditional compute.

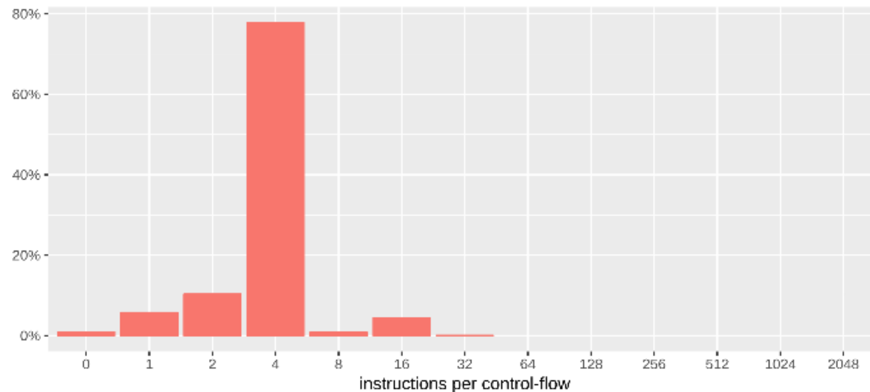


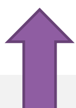
Workload Observations

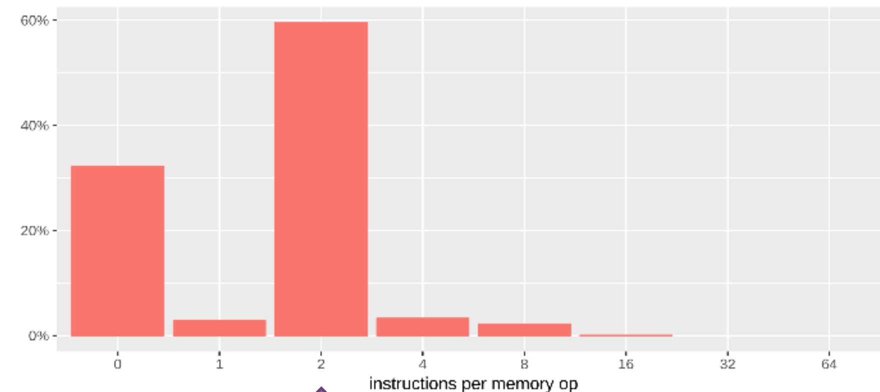




- Poor cache line utilization
  - Over 80% use 8B or less
- Extreme stress on deep pipelines, branch systems, OOO logic
  - Trending toward 2-4 ops per branch average
- Long chains of dependent loads
  - A[B[i]] += C[D[i]] * E[F[i]]
  - Constant memory pressure – every 2-3 instructions issue to memory





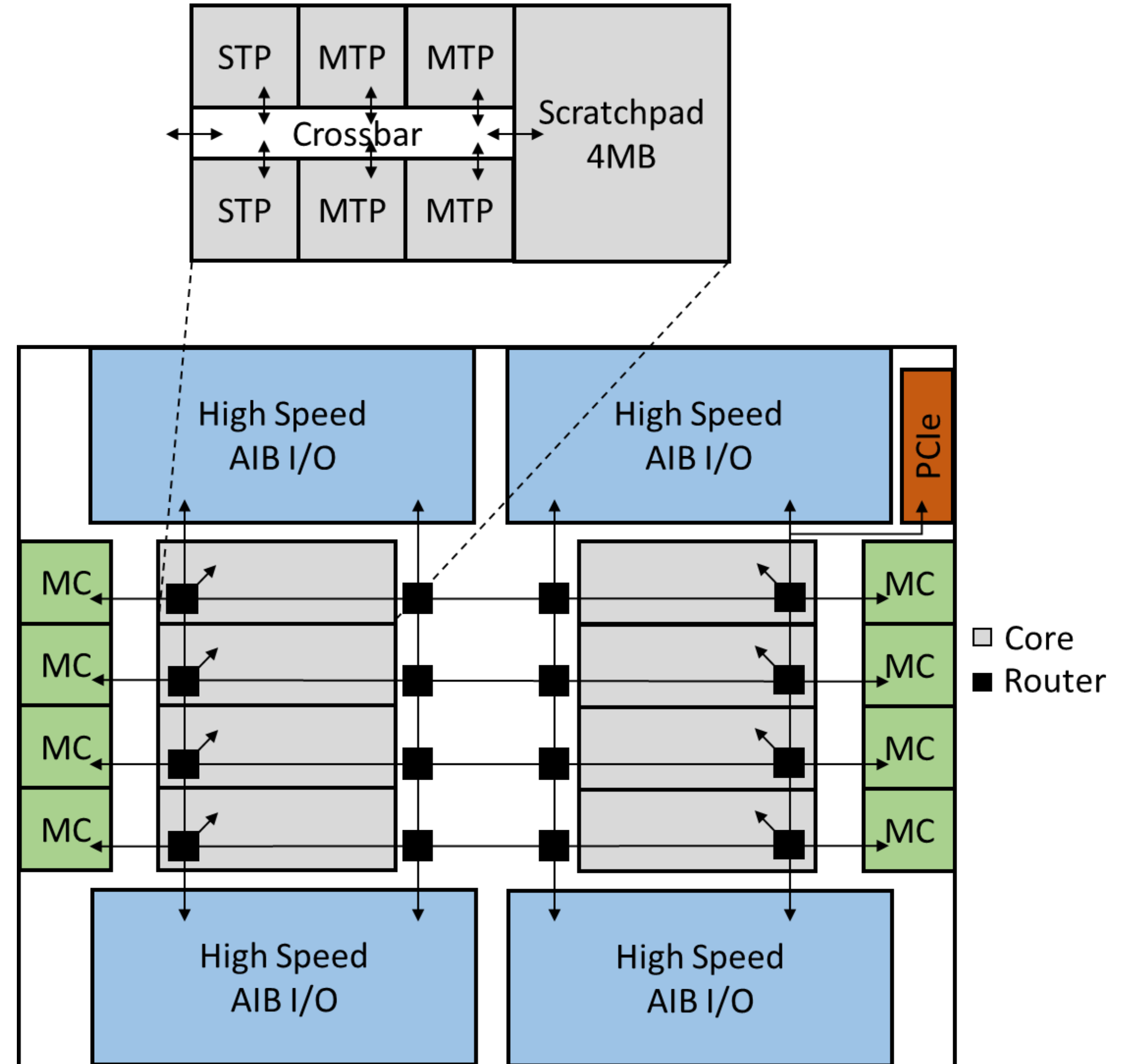


Key Technologies

- Core
  - 66 hardware compute threads per core
  - 192KB cache (i$+d$)
  - 4MB scratchpad SRAM
- Socket
  - 8 cores
  - 32 optical I/O ports at 32GB/s/dir
  - 32GB custom DDR5-4400 DRAM
- Sled
  - 16 sockets in an Open Compute Project (OCP) sled form factor,
  - 16TB/s/dir optical bandwidth
  - 0.5 TB DRAM
- Package
  - Multi-die package
  - Co-packaging of electrical to optical dies
- Network
  - HyperX network using high radix, low diameter switches
  - Optical links

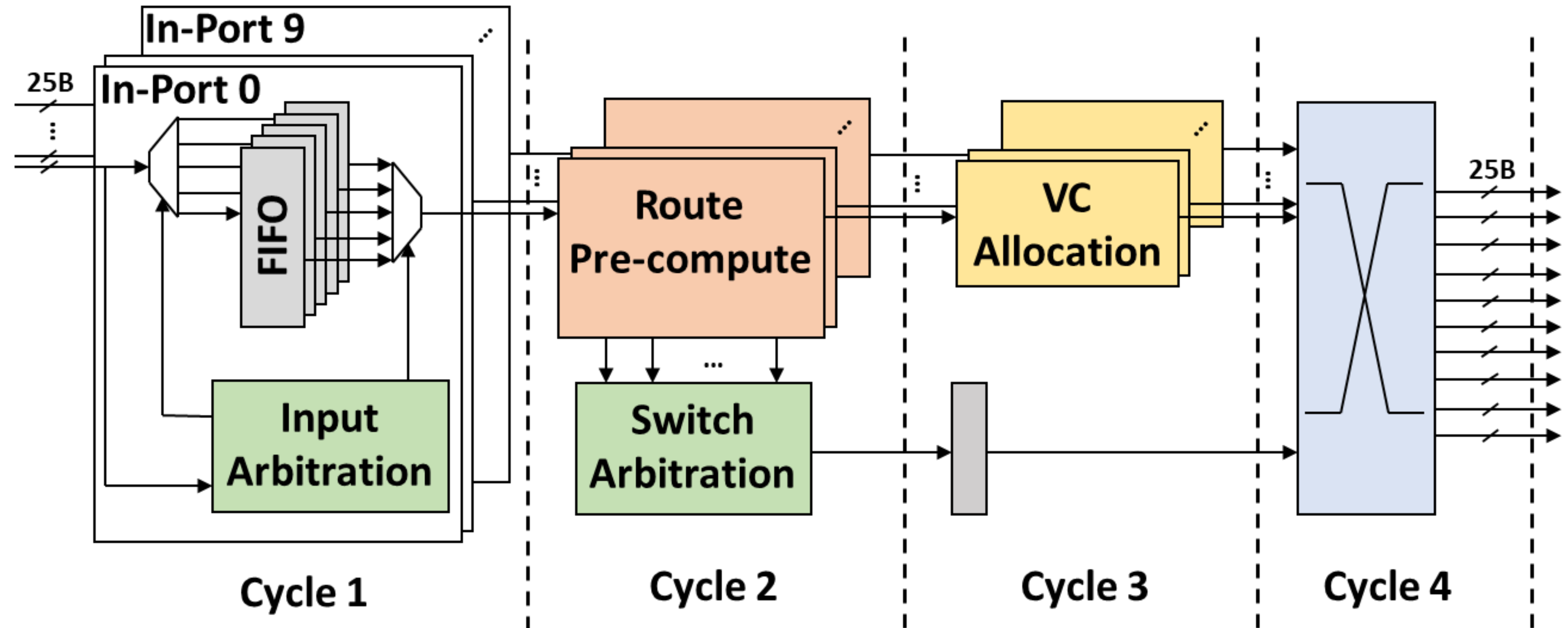
Die Architecture

- Multi-Threaded Pipelines (MTP)
  - 16 threads per pipeline
- Single-Threaded Pipelines (STP)
  - 8x higher single thread performance
- Pipeline ISA and Architecture
  - Custom RISC-based
  - Fixed length
  - 32 registers per thread
- 4MB Scratch Pad per core
  - Dual network ports
- Custom DDR5 memory controller
  - 8B access granularity
- 32 High Speed AIB ports
  - Bridges to 32 optical ports
- PCIe G4 x8



intel.

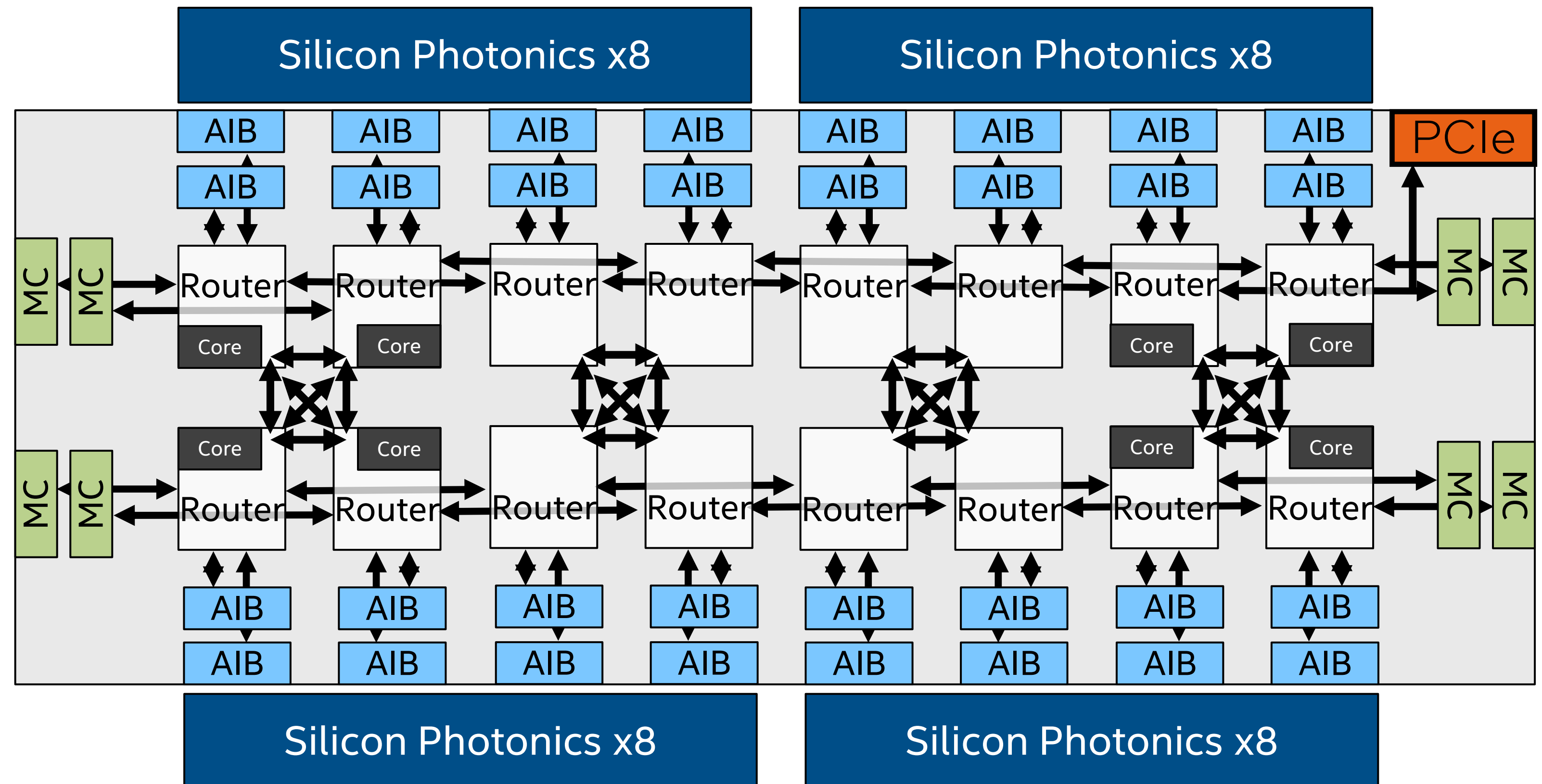
Router Architecture



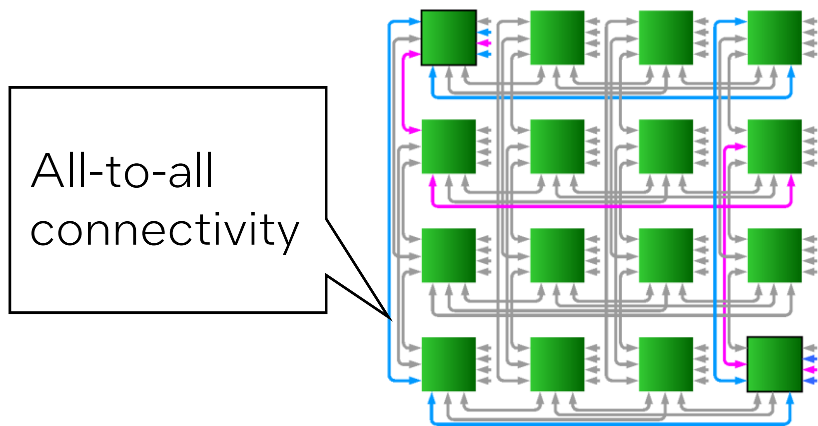
Frequency	1GHz @ 0.75V
Latency	4 cycles
Link Width	25 Bytes
Bandwidth	64GB/s per link
Architecture	10 VCs over 4 MCs

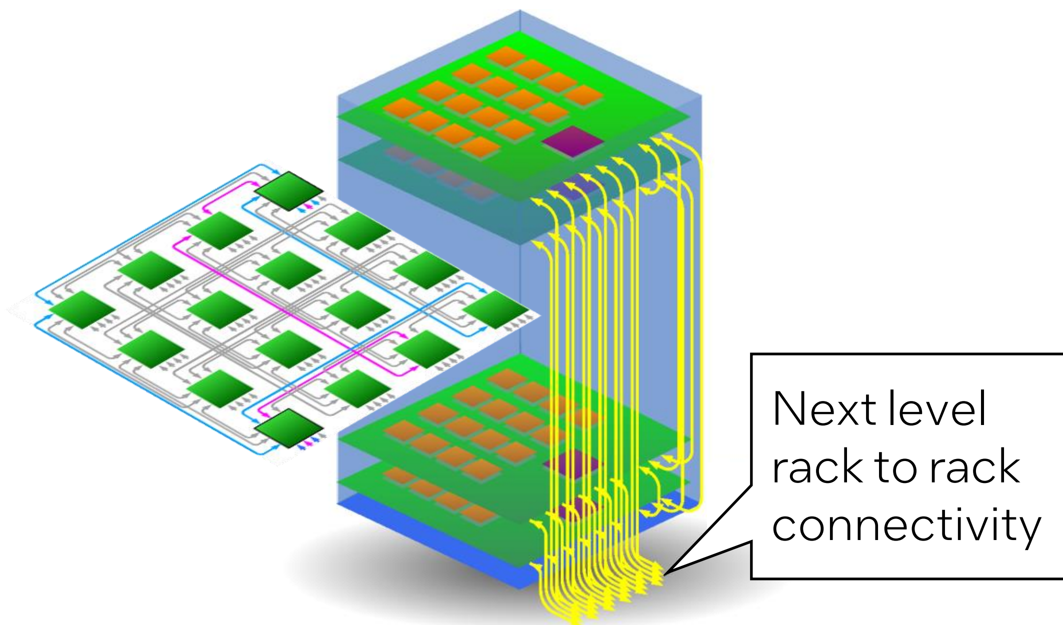
On Die Network

- 2D on-die mesh
- Uses 16 routers
  - Required for high off-die I/O bandwidth
- Silicon photonics extend the on-die network between sockets
- Silicon photonics links are only used as a Phy (physical) layer



Off Die Optical Network - HyperX

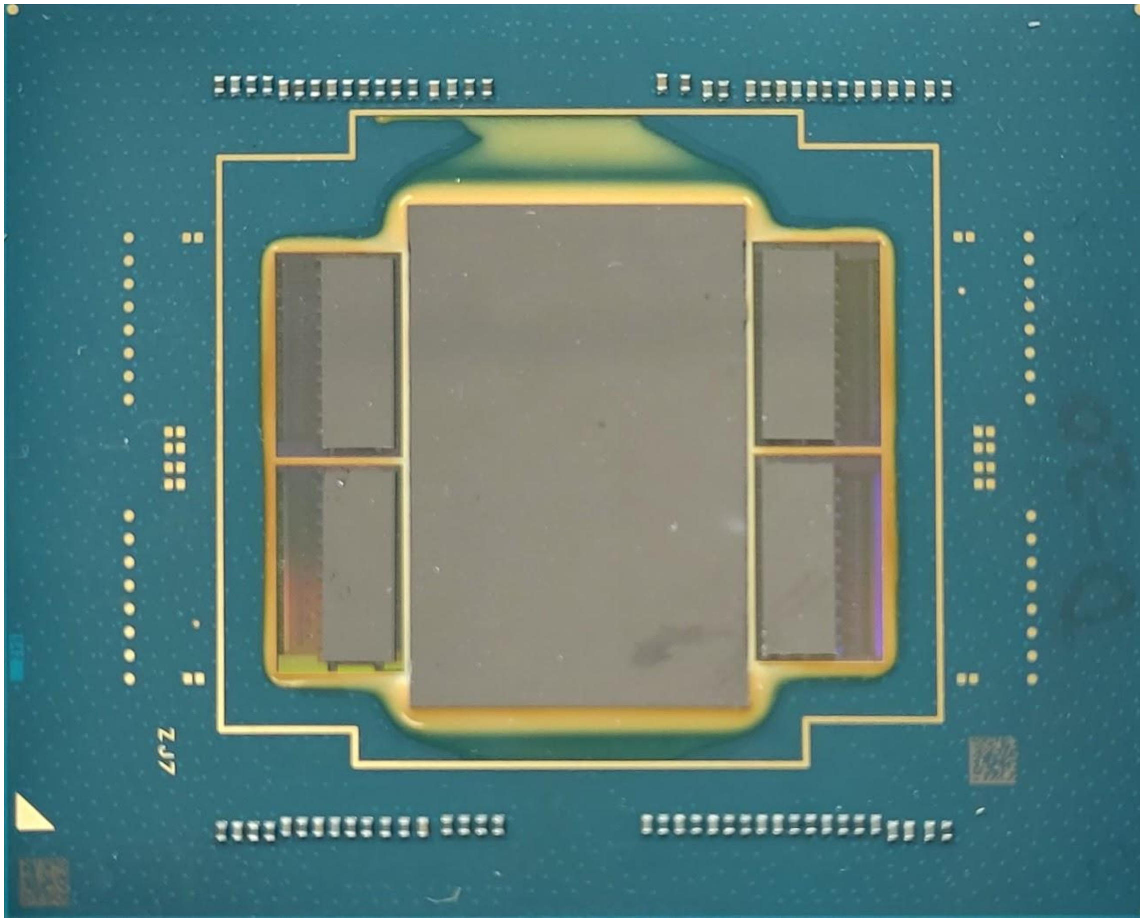




$$\sum_{i=0}^{L-1} n_i m_i = M; \quad M = 256 \text{ ports}$$

Number of Sleds	HyperX Levels	Configuration			Port count per sled pair			Uni-directional bisection bandwidth (TB/s)
		n_0	n_1	n_2	m_0	m_1	m_2	
2	1	2			128			1
4	1	4			64			2
8	1	8			32			4
16	1	16			16			8
32	1	32			8			16
64	1	64			4			32
128	1	128			2			64
256	1	256			1			128
512	2	32	16		4	8		128
1,024	2	32	32		4	4		256
2,048	2	64	32		2	4		512
4,096	2	64	64		2	2		1,024
8,192	2	128	64		1	2		2,048
16,384	2	128	128		1	1		4,096
32,768	3	32	32	32	2	2	2	4,096
65,536	3	64	32	32	2	2	2	8,192
131,072	3	64	64	32	2	1	2	16,384

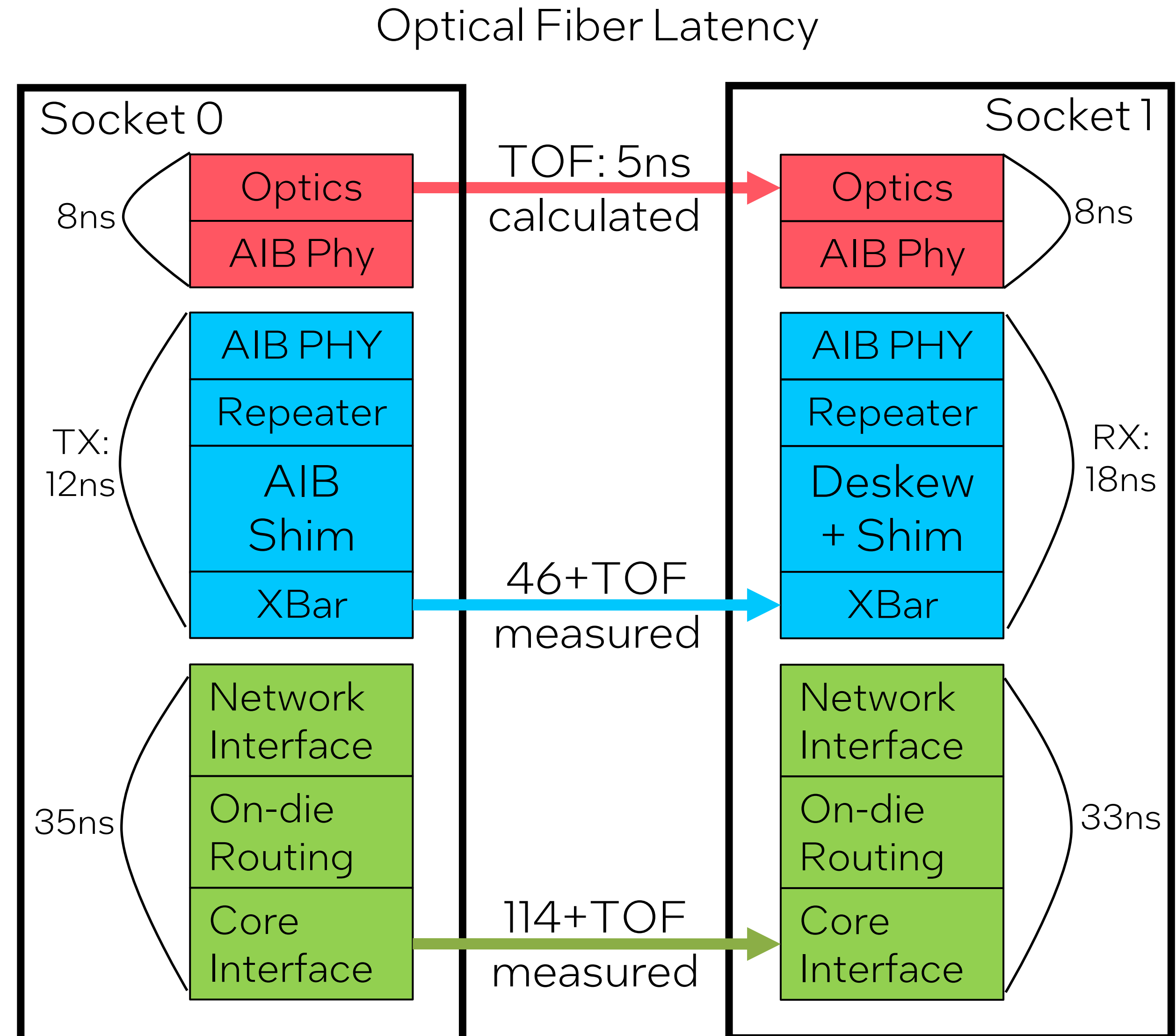
Co-packaged Optical Silicon Photonics



- Multi-chip package design
- Electrical to optical chips
  - AIB over EMIB technology between chips
  - V-groove for optical alignment during attachment
- Challenges
  - Fiber attach stability
  - Reflow and thermal issues resulted in yield loss of optical attachment
  - New materials developed for high volume yields

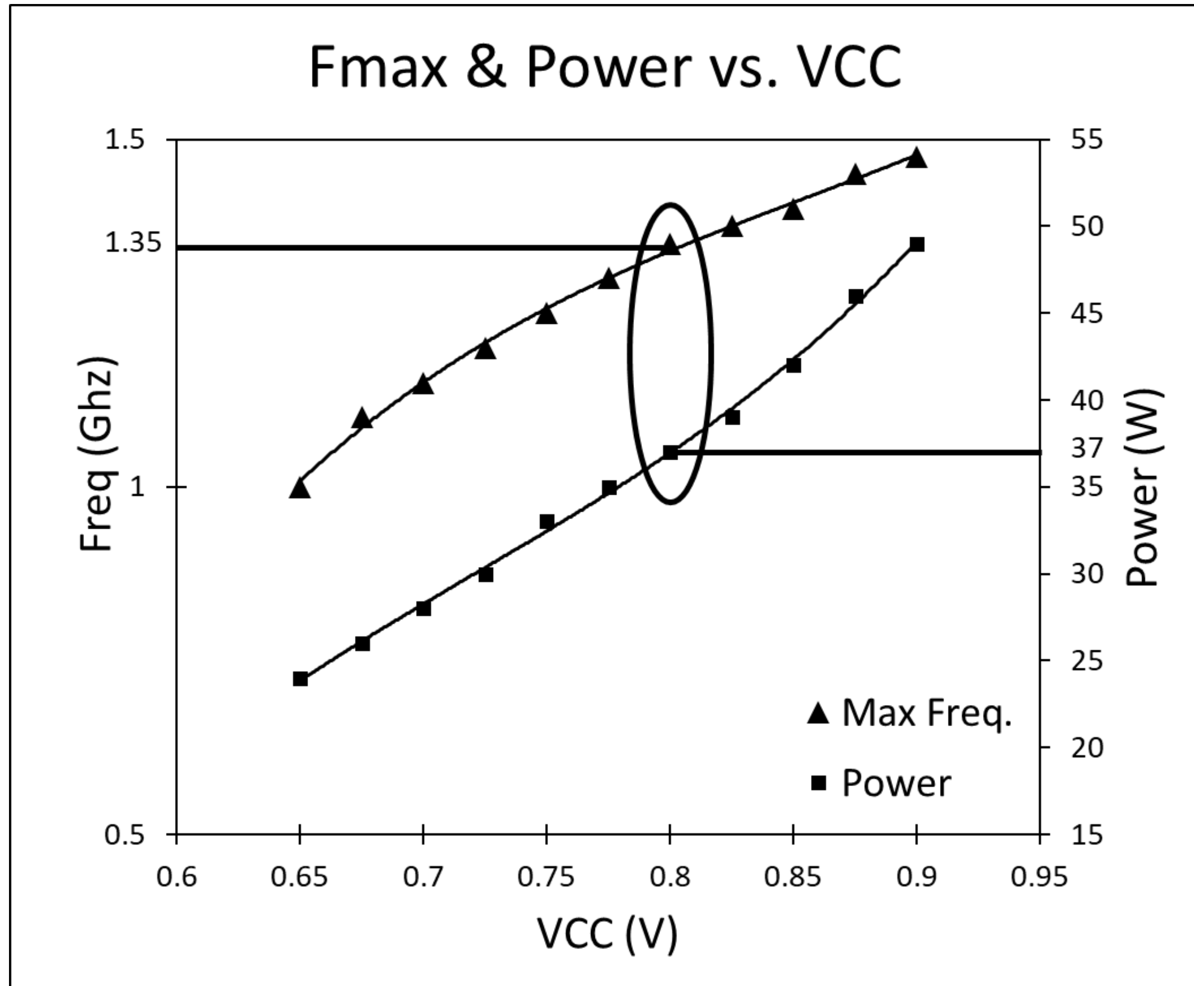
Measured Optical Performance

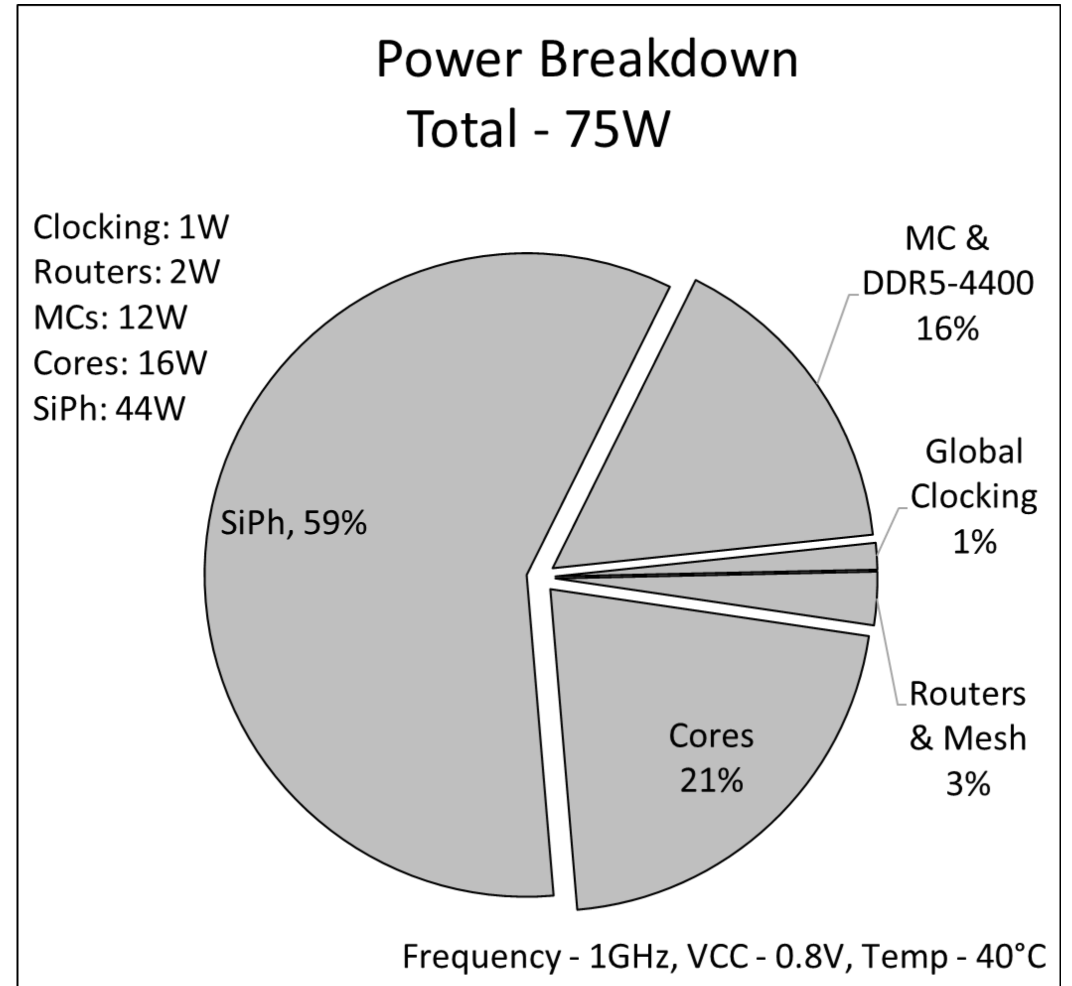
- Latency calculated using on-die measurement points and known on-die cycle accurate values
  - Core to AIB routing: 35ns
  - AIB TX: 12ns
  - AIB RX : 18ns
  - AIB to core routing: 33ns
- Optical Latency
  - 5ns
- Single fiber optical bandwidth
  - 32GB/s/dir (design point)
  - 16GB/s/dir (achieved)



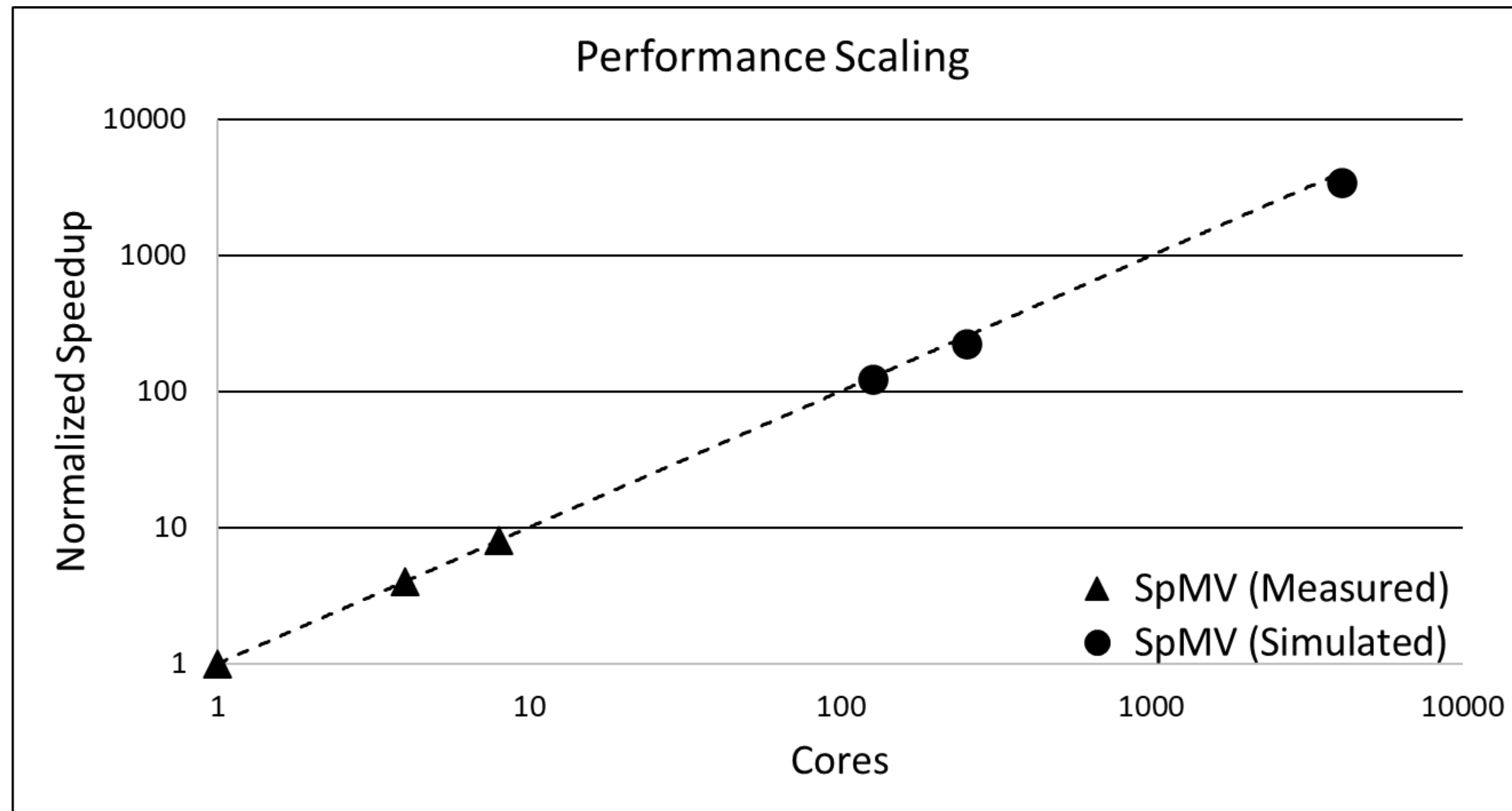
intel.

Measured Fmax and Power



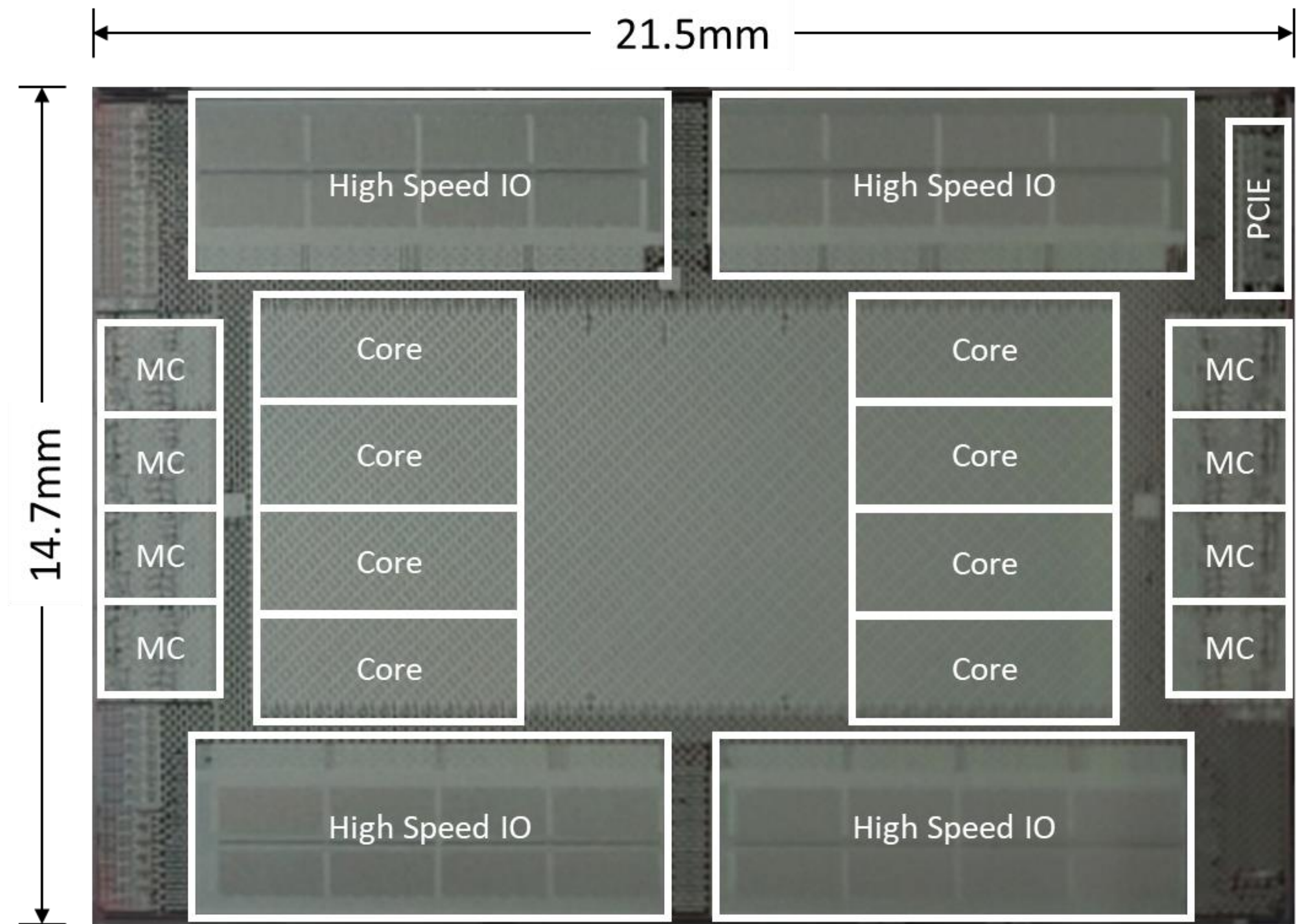


Workload Profile



Die Photograph and Characteristics

Technology	TSMC 7nm FinFET
Interconnect	15 Metal Layers
Die Transistors	27.6B
Die Area	316mm^2
Core Transistors	1.2B
Core Area	9.3mm^2
Signals	705
Package	3275 pin BGA Package



Package and Test Board

Summary

- 8 core processor in 7nm FinFET CMOS
  - Over 500 concurrent threads
  - Fused network and compute die
  - 1TB/s of optical I/O
- Socket-to-socket photonic links
  - New package technology allowing 32 fibers per socket
- HyperX topology allows scaling to over 100k sleds and over 1M sockets
- A socket dissipates 75W at nominal voltage and workloads
  - A 16-socket sled dissipates 1.2kW

Acknowledgements

- Program Direction & Funding

  DARPA, US Government
  Prime Agreement No.: HR0011-17-3-0004
  Performer Name: Intel Federal LLC

- Experimental Chiplets and Co-assembly Support
  - Ayar Labs

- Package design and assembly

  Intel Assembly Technology Development
  (ATTD, Chandler, AZ)