Rambus

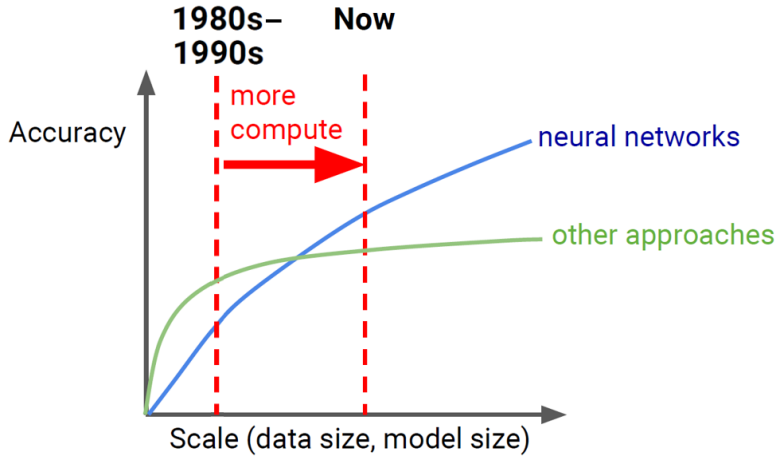
Memory Systems for AI and Leading-edge Applications

Sanjay Charagulla
Vice President, Strategy & Business Dev

Steven Woo
Fellow and Distinguished Inventor

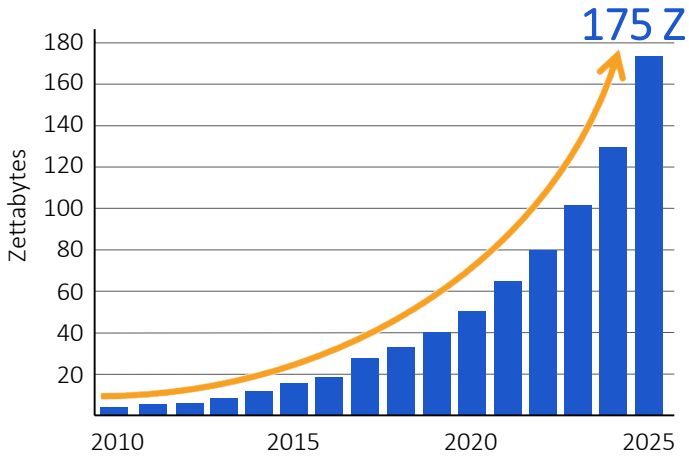
June 4, 2019

AI and the New Golden Age for Computer Architecture



Source: Adapted from Jeff Dean, "Recent Advances in Artificial Intelligence and the Implications for Computer System Design," HotChips 29 Keynote, August 2017

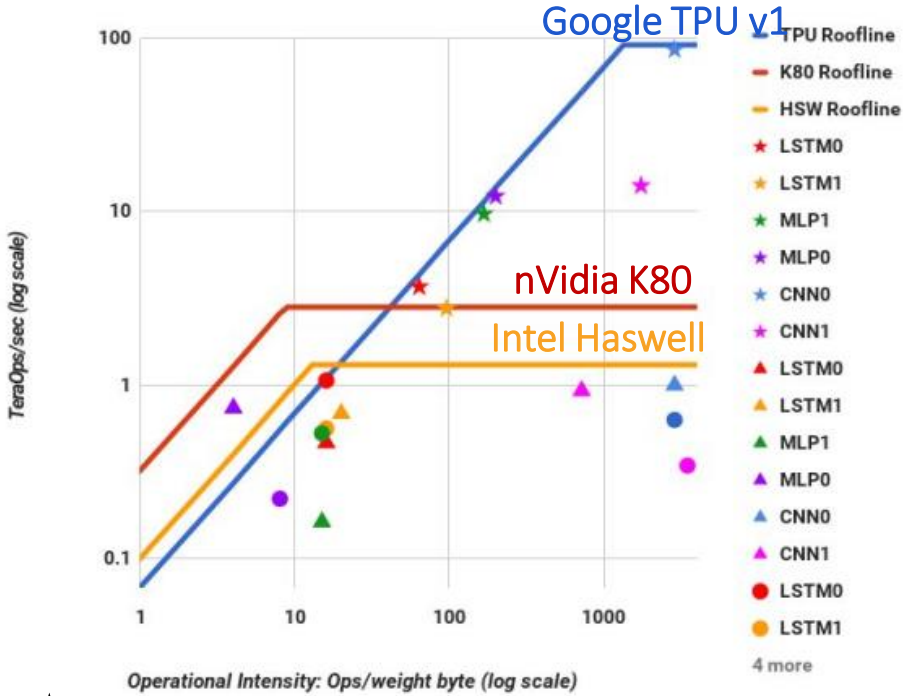
Annual Size of the Global Datasphere



Source: Adapted from Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

- Faster compute and memory + large training sets have enabled modern AI
- Much more left to do, more performance needed
- Key challenges: Moore's Law ending, energy efficiency growing in importance

AI Application Performance: Google TPU v1 Study



☆ = Google TPU v1 △ = nVidia K80 ○ = Intel Haswell

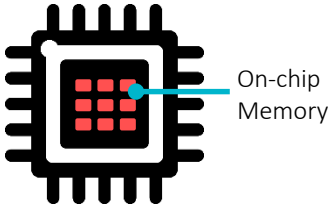
Source: N. Jouppi, et.al., "In-Datacenter Performance Analysis of a Tensor Processing Unit™," https://arxiv.org/ftp/arxiv/papers/1704/1704.04760.pdf

- Inference on older, general purpose hardware (Haswell, K80) performs well, applications can benefit from compute and memory optimizations
- Inference on AI-specific silicon (Google TPU v1) largely limited by memory bandwidth

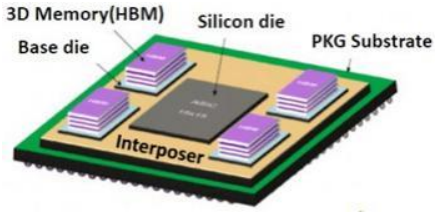
Memory Bandwidth critical for AI applications, need to make the most of what's available

AI Needs Memory Bandwidth: Common AI Memory Systems

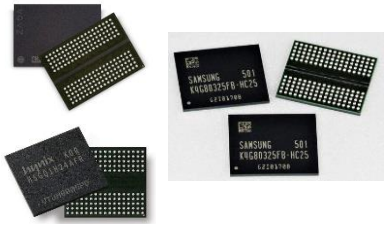
On-Chip Memory

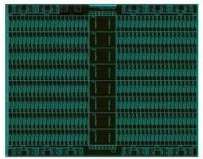


HBM

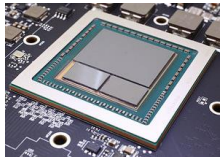


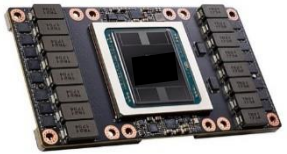
GDDR





Highest Bandwidth
and Power Efficiency





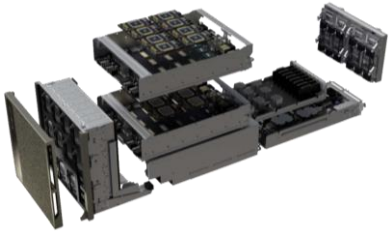
Very High Bandwidth
and Density

Good tradeoff between bandwidth,
power efficiency, cost, and reliability

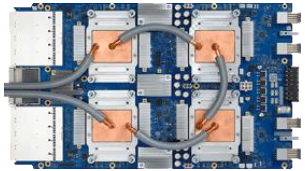
Multiple options suited to different needs

Example AI Hardware on the Market

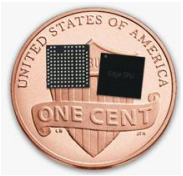
nVidia DGX-2



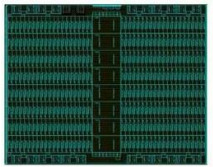
Google TPU v3



Google Edge TPU

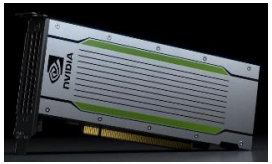


Graphcore IPU

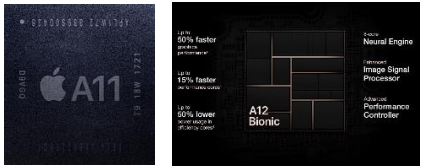


nVidia Tesla V100

nVidia Tesla T4 GPU



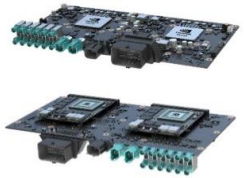
Apple A11 and A12 Bionic Processors



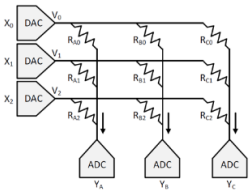
Microsoft BrainWave

nVidia GeForceRTX 2080 Ti

nVidia DRIVE PX 2

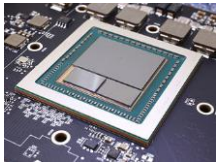


Mythic Analog MAC



Wave Computing

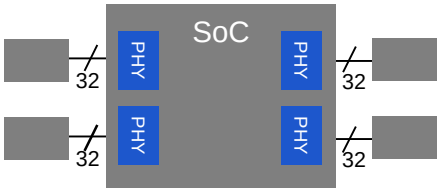
AMD Radeon Vega 56



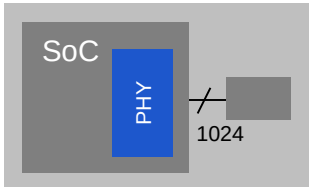
Intel Nervana NNP

Memory System Comparison: 256GB/s GDDR6 vs. HBM2

GDDR6 Memory System
Four 16Gbps x32 GDDR6 DRAMs



HBM2 Memory System
Single 2Gbps HBM2 Device



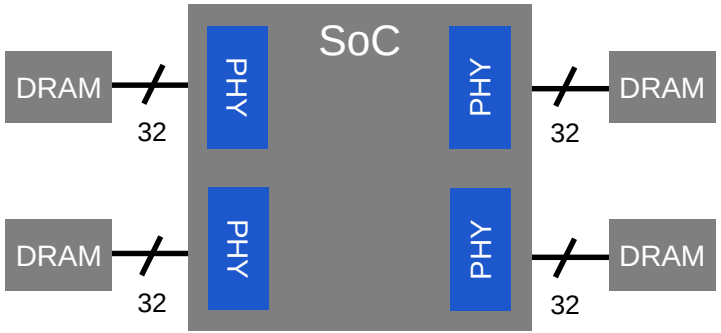
Total Bandwidth	256 GB/s	256 GB/s	
Per-pin data rate	16 Gbps	2 Gbps	
Relative Controller PHY Area[1]	1.5-1.75	1.0	Area advantage for HBM2
Relative Controller PHY Power[1]	3.5-4.5	1.0	Power advantage for HBM2
Interposer	None	Added cost[2]	Cost, complexity advantage for GDDR6
Memory	Similar to GDDR5, DDR4	Stacked, adds cost[2]	Cost advantage for GDDR6

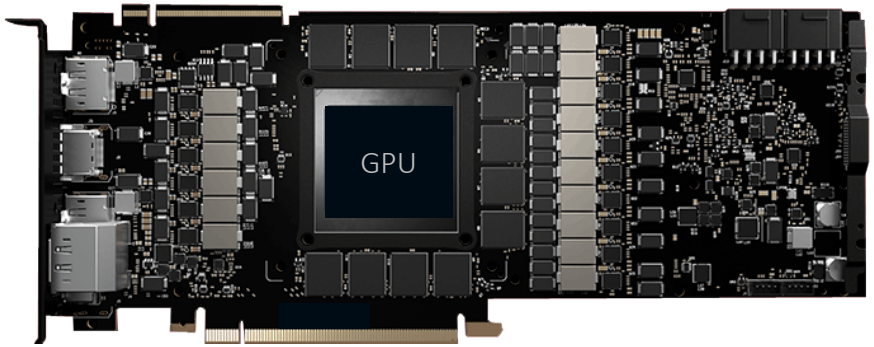
[1] Source: Rambus Inc.
[2] Source: The Cost of HBM2 vs. GDDR5 & Why AMD Had to Use It, https://www.gamersnexus.net/guides/3032-vega-56-cost-of-hbm2-and-necessity-to-use-it

GDDR6 and HBM2 offer different system design tradeoffs

256GB/s GDDR6 Memory System: GDDR6

GDDR6 Memory System
Four 16Gbps x32 GDDR6 DRAMs

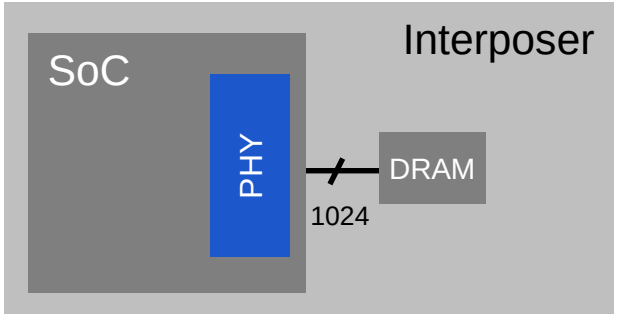


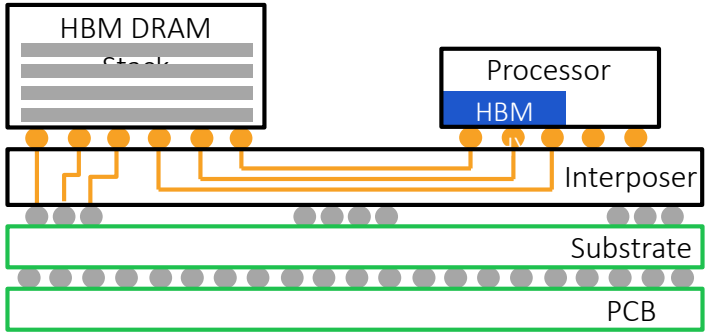


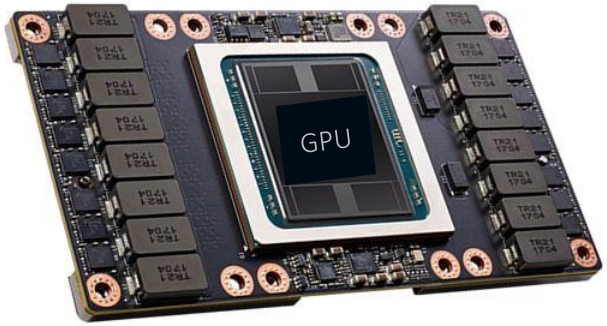
- Good tradeoff between bandwidth, capacity, power-efficiency, cost, reliability, design complexity
- Easier integration / system engineering compared to HBM
- Well-understood assembly and manufacturing (similar to DDR)

256GB/s HBM2 Memory System

HBM2 Memory System
Single 2Gbps HBM2 Device







- Additional interposer/substrate, new manufacturing and assembly methods (not like DDR or GDDR)
- Very high bandwidth (256GB/s per HBM2 DRAM)
- High power-efficiency (short interconnects, wide and slow interface, 1024b@2Gbps)

Summary/Conclusion

- AI driving the development of new silicon and new system architectures
- Memory bandwidth a critical resource for AI applications, memory systems are once again a hot topic in the semiconductor industry
- Multiple memory options to suit different AI application needs
  - On-chip memory: Highest bandwidth and power efficiency, lowest latency, but storage capacity limited
  - HBM: Extremely high bandwidth and power efficiency, but higher cost and more challenging integration and design complexity
  - GDDR: Good tradeoff between bandwidth, capacity, power efficiency, cost, reliability, design complexity

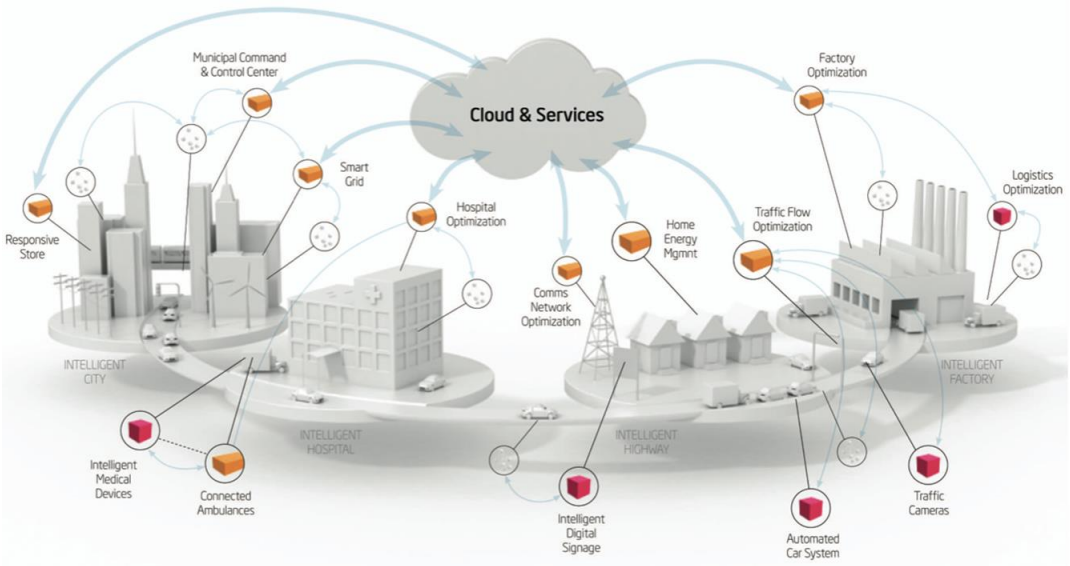
Thank you

Rambus

Data · Faster · Safer

Pervasive Connectivity and Data Computing

1. Memory, link, and storage performance must continue to increase
2. Data and insights are increasingly more valuable, security a growing concern
3. Compute and I/O power efficiency must also continue to improve



In our increasingly connected world, architectures are evolving to more efficiently capture, secure, move, and process the growing volume of digital data