Accelerating Time to Insight
Performance optimized emerging system architectures

Balint Fleischer
Senior Director
Advanced Computing Solutions

September 25th, 2019
Rapidly growing Dark Data

Rapidly growing Data sets from many sources
+ Increasing complex algorithms
+ Need for faster Time to Insight
+ Affordability challenges

Dark Data

Estimated data
1 Zettabyte = 10^21 (one billion, trillion) bytes
Perfect storm:

Moore’s law is slowing, Dennard scaling is ending and von Neumann architecture became a bottleneck.

B. Meisner, The Bump in the road to ExaFlops and Rethinking LINPACK, HPC User Forum, June 2014
Large, industry wide approach but without a master plan to insure continued scaling.
New, Non Von Neumann Architectures Scale well (for a while) with process technology improvements
General purpose CPU performance CAGR is declining
Machine Learning + Deep Learning emerging as the key application for data analytics

A rich target for Domain specific computing
On a given technology node, Domain Specific Architectures deliver greater performance on targeted applications.

Pattern for domain specific architectures
Simpler, lower performance, more efficient processing elements
High degree of parallelism (1000s of processing elements)
Highly optimized on die data movement
New, application specific ISA
Custom compiler
An example of Domain Specific Architecture delivering greater performance

Optimized AI processors deliver even better gain
There are over 300 AI processor designs worldwide innovating from Low End to High End.
“Cost” of IO is critical to performance scaling and energy consumption

The goal is improving Bandwidth, reducing Latency & reducing Data movement
Improving Data Proximity is a vehicle to address IO cost

For simplicity single socket CPU Hosts are shown.
NAND die capacity continues to grow enabling more dense SSDs and storage solutions.
SSD IO Bandwidth is lagging SSD Capacity growth

Query of very large data sets will take an increasingly long time
NAND stacks and SSDs are designed for capacity scaling

Device level aggregate BW is \(~100\times\) vs. SSD

External IO Bandwidth

Die level Bandwidth is \(~1000\times\)
Moving processing into SSD can benefit from the high Embedded Bandwidth. Time to Insight on very large shardable data will also benefit from Massive parallelism.

Note: In both cases there are 24 drives per JBOF, identical PCIe G5 NVMe interface.

* For simplicity single socket CPU Hosts are shown.
Near Memory Computing can Deliver 5-10x better Bandwidth to applications in a 2.5D packaging technology
Tighter integration of processing and memory can lead to even greater Performance at lower Power

Near Memory Computing example

Q1. Neural computing layer should meet thermal and area constraint in 3D stacked DRAM
Q2. NeuroCube should be programmable to cover different types of neural network

[Kim et al., NeuroCube, ISCA 2016]
On Die integration of compute and memory can take advantage of ~160x on Die bandwidth on smaller Data sets.
Large investment into improving Platform IO
Increasing integrated IO lane count to grow connectivity

Faster PCIe will enable higher speed devices

Silicon photonics to extend reach

* For simplicity single socket CPU Hosts are shown
Emerging Media technologies to improve capacity, latency and access methods.
Emerging Memory research is intensifying

Difficult to beat DRAM Performance & Energy

But

DRAM density growth is slowing
Emerging Memories can fill data access latency gaps and enable new storage models.
Attaching EM requires server architecture changes

CXL is the emerging Standard for EM attach

* For simplicity sake single CPU Hosts are shown
CXL enables innovations ranging from different memory types to heterogeneous computing.

Some possible memory examples:

- CXL Block EMM
- CXL LD/ST EMM
- CXL LD/ST DRAM
- CXL LD/ST DRAM
- CXL Near Memory DRAM
- CXL AI Engine NAND
Scaling Emerging Memory capacity is critical to address use case requirements.

* For simplicity, single CPU Hosts are shown.
The collection of these new technologies and architectures will impact how we build future Data Centers and Systems.
Co-locating Compute and Data requires change in data placement, provisioning and load balancing strategies.

* For simplicity single socket CPU Hosts are shown.
Heterogeneous Resource Pools Example

* For simplicity single socket CPU Hosts are shown
Exciting times.
The Data Growth and the need for Faster Insight drives transitioning from decades old architectures to a new, emerging model utilizing breakthrough technologies

Buckle your seatbelt!
Thank you!

bfleischer@micron.com