Software Defined Memory: A Meta perspective

Chris Petersen, Hardware Systems Technologist, Meta
Increasing Memory Cost and Power

Price per Gb (Log Scale)

- Memory an increasing % of system power and cost
  - Memory price (cost/bit) flat due to scaling challenges
  - Memory power scaling with speed
Increasing Core Counts Drives Growth

- Increased Bandwidth
- Increased Capacity

Increasing core counts driving memory demand

<table>
<thead>
<tr>
<th>Year</th>
<th>CPU core count</th>
<th>Memory Channel BW per core</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2014</td>
<td>1.5</td>
<td>0.8</td>
</tr>
<tr>
<td>2016</td>
<td>2.0</td>
<td>0.7</td>
</tr>
<tr>
<td>2018</td>
<td>2.5</td>
<td>0.6</td>
</tr>
<tr>
<td>2020</td>
<td>3.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Normalized growth rate
Machine Learning Growth

ML Model Capacity Growth

ML Models are growing rapidly
- ~50X growth in ~5 years
- Existing memory hierarchy can’t keep pace
Compute Express Link (CXL) Introduction

- **Processor Interconnect:**
  - Open industry standard
  - High-bandwidth, low-latency
  - Coherent interface
  - Leverages PCI Express®
  - Widths: x4, x8, x16
CXL Memory Tiers

Requirements:
- Memory, not storage
- Cache-line reads/writes
- Scalable
- Heterogenous
- Standard interfaces

CXL-attached
Bandwidth Memory Tier

- **Use Cases:** Warm Pages, Page Migration
- **BW:** BW per GB close to that of DDR4 memory
- **Latency:** NUMA-like
- **Power:** ~90% of DDR5 at ISO capacity
- **Capacity:** Scales with standard RDIMMs
- **Form factor:** Initial solutions focused on “chip down + DIMMs”
Capacity Memory Tier

- Use Cases: Caching and ML Models
- BW: BW per GB 5-10% of DDR5 memory
- Latency: Hundreds of ns
- Power: ~50% of DDR5 at ISO capacity
- Capacity: 256GB - 1TB
- Form factors: Use hot-pluggable form factors (like E1 or E3)
CXL Memory Evolution

Direct-attach

CPU

CXL

CXL Controller

Media

CXL Memory

DIMMs

DDR*

Small Pools

CXL Memory Expander

Host 1

Host 2

CXL

CXL

Host 3

Host 4

CXL

CXL

Host 3

Host 4

CXL

CXL

Host 3

Host 4

Rack-scale Pools

Host 1

Host 2

CXL Memory Expander

Host 3

Host 4

CXL Memory Expander

Host 1

Host 2

Host 3

Host 4
Parting thoughts

• Lots of work ahead of us! Industry collaboration is critical.

• Think at the system level including SW integration, and also in phases

• Multiple CXL memory tiers are needed for multiple use cases. One size does not fit all!