Addressing the Latency Gap with Composable Architectures

Session : SOFT-201-1: Composable Infrastructure and Software Defined Storage
Larrie Carr
Abstract

In the past few years, the latency gap between the traditional load/store memory infrastructures and block-focused fabric and storage networks has seen the introduction of new composable architectures, CPU interfaces, processing accelerators, memory technologies and software stacks into the industry.

This presentation explores why addressing this latency gap has become so important and discusses the composable architectures enabled by technologies like OpenCAPI, CCIX, Gen-Z and CXL.
Latency Landscape

- Data needs to move from the right to the left in order to be “processed”
- Data movement and storage at the 1us “gap” is difficult for any technology
• Bit density increasing
• Cost per bit decreasing
• Connectivity rates doubling
• Latency and access times decreasing

• Lots of design flexibility and choice

• Life is pretty good…. 

Block Latency Landscape
Load Store Latency Landscape

- Moore’s law is dead - costs increasing
- IPC efficiency decreasing
- Latency increasing
- Connectivity rates evolving incrementally
- Closed proprietary interfaces
- Architecture flexibility limited to time of installation
- DRAM is a unique player…
Processor Memory Bottleneck

- Current processor DRAM buses provide limited quantity & performance scaling
- Multi-core processor devices experiencing *increased* latency per core

https://blog.dellmc.com/en-us/memory-centric-architecture-vision/
Reducing Stranded Resources

- Only 16% of Edison jobs would NOT run in Hopper 32 GB nodes
- 71% of the Edison jobs only need 16 GB of DRAM and fit into Cori’s HBM memory
- By sizing each node’s DRAM capacity to service the largest of jobs, overall DDR efficiency decreases and much of the system DRAM capacity is stranded

<table>
<thead>
<tr>
<th>NERSC HPC Datacenters</th>
<th>Hopper</th>
<th>Edison</th>
<th>Cori</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2011</td>
<td>2013</td>
<td>2016</td>
</tr>
<tr>
<td>Datacenter Nodes</td>
<td>6,384</td>
<td>5,576</td>
<td>9,300</td>
</tr>
<tr>
<td>DDR / Node</td>
<td>32 GB @ 54 GB/s</td>
<td>64 GB @ 102 GB/s</td>
<td>96 GB @ 90 GB/s</td>
</tr>
<tr>
<td>HBM / Node</td>
<td>N/A</td>
<td>N/A</td>
<td>16 GB @ 400 GB/s</td>
</tr>
</tbody>
</table>
DRAM Power Consumption

- DRAM consumes 18% of Google’s WSC total power consumption
- Memory costs upwards of 50% of total server costs
- Up to 25% of the memory bandwidth consumed by “overhead”
  - Memmove, allocation, compression, ….
Compressing Memory

- Google: Reduce the memory TCO by improving the density of data store in “colder” memory
  - Data gets cold after 2 minutes
  - Applications access 15% of their total cold memory every minute
  - Compress the cold memory to reduce footprint

- Results:
  - 20% of memory was compressible on average
  - 4% to 5% TCO savings on DRAM

- Microsoft: Open sourcing their Zipline memory compression RTL at OCP 2019
Composable architectures are about choice
- Sharing pools of resources (CPU, memory, FPGA) to reduce resource stranding
- Scaling to the server resources to match the workload
- Connecting to pools of resources

Forcing emerging memory and compute technologies behind legacy interfaces (PCIe/SATA) is inefficient

New open load/store standards provide the low-latency connectivity required to build flexibility **inside and outside the server...**
Serial Memory Controllers

- Dramatic bandwidth expansion with significant reduction in pin count
- Media independence - enable multiple memory types
- Lower cost SoC packaging by using serial memory interfaces
The Smart Memory Controller

8x25G Open Memory Interface (OMI) Serial DDR4 Smart Memory Controller

- **INCREASED MEMORY BANDWIDTH**: Enables 4x memory channels vs. x72 DDR4
- **MEDIA INDEPENDENCE**: Single OMI interface provides for multiple media types
- **LOWER SOLUTION COSTS**: Reduced silicon, IP and package costs for CPUs and SoCs
Smart Memory Controller PM8596

Open Memory Interface
- OIF-28G-MR
- Dynamic low-power modes

DDR Interface
- x72 bit DDR4-3200
- Up to 4 ranks, 3D stack
- 16Gbit DRAM support

Persistent Memory
- NVDIMM-N module support

On-Chip Processor
- Initialization, monitoring and diagnostics
- Open source firmware

Security
- Hardware root-of trust
- SECDEC with memory scrub

Applications
- JEDEC DDIMM applications
- Chip-down with conventional DIMMs
- FPGA accelerator memory fanout
- ML memory bandwidth expansion (versus HBM architectures)
Summary

- New low-latency load/store memory interfaces allow new compute and memory architectures

- New low-latency non-persistent load/store memory technology will play a role with these new architectures
  - Less expensive and power-hungry than DRAM

- Come see our Serial Memory Controller demo at our booth…
Thank You