A Persistent CXL Memory Module with DRAM Performance

Bill Gervasi
Principal Systems Architect
bilge@nantero.com
CXL Expansion Memory is a growing trend

CXL Memory faces some unique resistance

System architecture adoption of persistent memory

Nantero NRAM® persistent main memory

How persistent memory addresses CXL concerns

AGENDA
CXL Rising, but…

I’m sure we’ll see dozens of presentations at SDC on the emergence of CXL as a revolutionary change in systems architecture.

A Bluetooth for fabrics, so to speak.

However, there is still a lot of work ahead of us to enable its potential.
Focus on Memory Expansion

ASSUMPTIONS

CXL Expansion Memory does not replace the direct attached DIMM

CXL Expansion Memory is volatile, so local SSDs for checkpointing are still needed

CXL Expansion Memory is shared by many multi-core processors (very random access)
Increasing frequency is slowing DIMM improvements

- DDR1
  1 CH, 4 DPC, 2R/D, 256Mb = 4 GB

- DDR2
  2 CH, 3 DPC, 2R/D, 1 Gb = 24 GB

- DDR3
  3 CH, 2 DPC, 2R/D, 4 Gb = 96 GB

- DDR4
  4 CH, 2 DPC, 2R/D, 16 Gb = 512 GB

- DDR5
  8 CH, 1 DPC, 2R/D, 16 Gb = 512 GB
  8 CH, 1 DPC, 2R/D, 32 Gb = 1 TB

CH = channel
DPC = DIMMs per channel
R/D = ranks per DIMM
Assumes no 3DS
The slowdown of capacity expansion from DRAM on DIMM explains the momentum for CXL memory expansion.

Memory expansion beyond DIMMs
End-user friendly pluggable modules
So, What’s Not to Love About CXL?

These concerns may slow the rate of CXL adoption.
Nantero DDR5 NRAM®

- DDR5 SDRAM speed
- Non-volatility
- Scalable beyond DRAM
- Lower power than DRAM

Let’s explore how this affects user’s concerns about CXL for memory expansion
Addressing Latency

Latency adder is real: CXL, SERDES, media

Full duplex nature of CXL helps offset penalties

Read and write buffers in a CXL DRAM controller can reduce some penalties
Addressing Performance

**CXL DRAM Controller**

Performance

Media timing is (mostly) fixed

**Non-deterministic refresh** is the performance demon

Diagram:

- **Host**
  - Read/Write
  - DRAM Refreshing?
    - YES: Take a nap
    - NO: Read/Write
  - Data/Done

- **CXL**
  - Read/Write
  - DRAM Refreshing?
    - YES: Take a nap
    - NO: Read/Write
  - DONE

## What if There Were No Refresh?

1. **111% DIMM-NRAM vs DIMM-SDRAM**

2. **96% CXL-NRAM vs DIMM-SDRAM**

3. **111% CXL-NRAM vs CXL-SDRAM**

<table>
<thead>
<tr>
<th>Ref</th>
<th>Refi</th>
<th>Avail</th>
<th>CXL</th>
<th>SDRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM, 1X Ref</td>
<td>350</td>
<td>3900</td>
<td>91%</td>
<td>111%</td>
</tr>
<tr>
<td>DRAM, 2X Ref</td>
<td>350</td>
<td>1950</td>
<td>82%</td>
<td>100%</td>
</tr>
<tr>
<td>DRAM, 4X Ref</td>
<td>350</td>
<td>975</td>
<td>64%</td>
<td>78%</td>
</tr>
<tr>
<td>NRAM</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td><strong>111%</strong></td>
</tr>
</tbody>
</table>

**NRAM’s data persistence eliminates refresh, making CXL accesses more deterministic**

Offsets some effects of latency penalty

---

* Half-duplex assumed; full duplex model to be built; likely gets better
** Assumes pipelining of CXL requests offsets 85% of penalties
Addressing Power

Compared to... what? DDR5 LRDIMM?

DDR5 \(\rightarrow\) 6400 Mbps speed @ 25% increase in power
Register, data buffers consume at least 2W per DIMM
40 DRAMs per DIMM: \(11 + 2 = 13\)W total DIMM power

DDR5 \(\rightarrow\) Same power issues as DIMM
CXL controller consumes at least 6W
Comes down to media capacity

E1.S: 20 chips
GB/W: \(\frac{1}{2}\) DIMM capacity @ 11.5W
\(~ 56\%\) efficient

E3.S: 40 chips
Better GB/W: \(1X\) @ 17W
\(~ 77\%\) efficient

E3.S: 80 chips
Even better GB/W: \(2X\) @ 28W
\(~ 92\%\) efficient
(Non-)Destructive Activation

Precharge operation required even if data is only read
8192+ECC bits always rewritten

Activate in place,
No precharge operation required,
64+ECC bits directly read/written
NRAM Non-Destructive Activation = Lower Power

Even with the volume cranked to 11*

(Data transfers when a DRAM would be refreshing)**

NRAM saves 21% power over DRAM while delivering 11% more data at the same clock rate

34% better throughput per watt than DRAM

<table>
<thead>
<tr>
<th>CMD</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>ACT</th>
<th>REF</th>
<th>REF</th>
<th>NOP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>RAP</td>
<td>WAP</td>
<td>WAP</td>
<td>WAP</td>
<td>WAP</td>
<td>WAP</td>
<td>WAP</td>
<td>WAP</td>
<td>RAP</td>
<td>WAP</td>
<td></td>
</tr>
<tr>
<td>DRAM</td>
<td>65.9</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>205.6</td>
<td>139.7</td>
<td>115.5</td>
<td>115.5</td>
<td>82.55</td>
</tr>
<tr>
<td>NRAM</td>
<td>65.9</td>
<td>106.75</td>
<td>172.65</td>
<td>106.75</td>
<td>172.65</td>
<td>106.75</td>
<td>172.65</td>
<td>106.75</td>
<td>172.65</td>
<td>82.55</td>
<td>82.55</td>
<td>82.55</td>
<td>106.75</td>
<td>82.55</td>
<td>53.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The number 11 is a registered trademark of Spinal Tap
NRAM CXL lower power than two DDR5 DIMMs or DDR5 on CXL

Additional power saving if SSDs can be eliminated
Addressing Volatility

Let your data die on power fail
Checkpoint to SSD for recovery

Add supercapacitors
to NVDIMMs, CXL

Use non-volatile NRAM
Addressing Capacity

**DDR5 NRAM®**

- 16 Gb
- 32 Gb

DDR5 SDRAM compatible

DDR5 NVRAM extensions
Massive Memory Expansion!

64 GB to 2 TB per module

CXL E3.S

* PCIe 6.0 unlikely to be in first generation CXL systems
CXL E1.S

Commodity Memory Expansion

32 GB to 256 GB per module

DDR5-6400 NVRAM
1 channels x
2 subchannels x
1 ranks
@ 25.6 GB/s

CXL 3.0
PCIe 5.0
8 lanes
@ 32 GB/s
Full duplex
So Where Do We Go From Here?

But the next step in exploiting persistent main memory requires

- Boot from CXL
- Power fail restart from CXL

Similar to BAEBI

Software support via DAX assists in moving...

- from mounted drives...
  - ...to RAM drive...
  - ...to direct access mode
The old paradigm
“checkpointing”

DDR SDRAM
DDR SDRAM
DDR SDRAM
DDR SDRAM

Run
checkpoint
Run
Run
Run

The new paradigm
“leave it in place”

CXL

Run

21 | ©2022 Storage Networking Industry Association. All Rights Reserved.
The Holy Grail of computing systems is when Instant On is realized.

Non-volatile memory media with DRAM performance makes Instant On achievable.
SUMMARY

- CXL for memory expansion has some valid concerns
- Persistent memory on CXL addresses these concerns
- The Holy Grail is end to end data persistence for Instant On
- Nantero NRAM is a DDR5 class persistent memory
- The DAX model helps existing systems move gracefully to PMEM
Thank You

Please take a moment to rate this session.

Your feedback is important to us.

Bill Gervasi
bilge@nantero.com