PM: Media, Attachment, and Usage
Dave Eggleston, Intuitive Cognition Consulting
I ♥ DRAM! (But…)

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2020</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>-37%</td>
<td>-8%</td>
<td>$/year</td>
</tr>
<tr>
<td>Capacity</td>
<td>&gt;2x</td>
<td>&lt;2x</td>
<td>Gb/chip/2 years</td>
</tr>
<tr>
<td>DIMM Bandwidth</td>
<td>+31%</td>
<td>+31%</td>
<td>GB/sec/year</td>
</tr>
<tr>
<td>Channels/CPU</td>
<td>2</td>
<td>8 (going to 16)</td>
<td></td>
</tr>
<tr>
<td>DIMMs/Channel</td>
<td>4-8</td>
<td>2 (going to 0.5)</td>
<td></td>
</tr>
<tr>
<td>Channels/DIMM</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

- DRAM tech and price scaling has slowed way down
- Running out of tricks to keep DDR bandwidth increasing
- Can PM do (part of) the DRAM job? More than that?
PM Media
What, Why, and Who
PM Media: What Options

transistor-based
high-$V_T$ – low current

ferroelectric memories

MIM-based resistive switching
HRS – low current

“0“

FLASH
FeFET
FRAM
FTJ
MRAM
RRAM
PCRAM

“1“

low-$V_T$ – high current
1T up to 2T bit cell

high D current
1T-1C up to
2T-2C bit cell
“DRAM like”

LRS – high current
usually 1T-1R

Source: J. Müller, GLOBALFOUNDRIES, “Ferroelectric Memories & Beyond”, IEDM 2019
PM Media: Something New!

Process and Packaging Innovations for Moore’s Law Continuation and Beyond

Robert Chau
Components Research, Technology Development, Intel Corporation, Hillsboro, OR 97124, USA
Contact E-mail: robert.chau@intel.com

Figure 18 shows the measured $I_{DS}-V_{G}$ of a FeFET single transistor memory after applying programming voltage of +1.7V and erase voltage of -1.5V at $V_{DS}=0.1V$, demonstrating a memory window of 1V.

1T-1C Ferro HfOx Bitcell*
1E4 sec data retention
1E9 endurance cycles

* Verbal mention in talk

PM Media: Why we Love/Hate each

DENSE & CHEAP

FLASH

FAST & CHEAP

FeFET

FAST & CHEAP

FRAM

FAST & FOUNDARY

FTJ

CHEAP & ???

MRAM

JUST RIGHT!

RRAM

PCRAM

WAY TOO SLOW!

WAY TOO NEW!

WAY TOO $$$!

REL TOO ???!

HOW TO USE?

© 2020 SNIA Persistent Memory Summit. All Rights Reserved.
PM Media: Who is Doing What

Source: Industry announcements & scuttlebutt
PM Attach
Options, DIMMs, Fabrics, Pooling, and Who
PM Attach: Options aplenty

DATA ACCESS HIERARCHY IN 2019 – GETTING WORSE!

Source: P. Faraboschi, HPE, “The Data Access Continuum”, SC’19 MCHPC
PM Attach: DIMMs getting complicated

Possible DDR5 Server Subsystem

- RDIMMs
- LRDIMMs
- UDIMMs
- DDIMMs
- Optane DIMMs
- NVDIMM-N
- NVDIMM-P
- Etc...

Different types of DIMMs will be common in DDR5

Source: D. Rhoden, Montage, “Server Memory Evolution”, JEDEC Training Class Oct 2019
PM Attach: Think Beyond the DIMM

- Today memory is direct-attached to the CPU
- New emerging interfaces will add high-speed differential CPU-attach options
- Systems will be aware of what type of memory or storage is available and how it is connected
- **Lots of new types of memory, persistent memory and storage products are possible!**

Source: A. Sainio, SMART, “PM Benefits AI/ML”, IMC Summit 2019
PM Attach: Think Pooling and Fabrics

DATA ACCESS INTERCONNECTS

30-100ns
200-500 GB/s
Memory semantics (ld/st)
Invisible to software, CC
Point-to-point
Dedicated resources

OPTANE®

OPTANE®

JEDEC.

OPTANE®

CXL

OpenCAPI™

Inside the server

300ns-1,000ns
25-100 GB/s
Software-exposed (libfabrics)
Messaging semantics
(send/receive or put/get)
Switched (high throughput)

Compute
High Bandwidth Memory

Capacity Memory/NVM

Switch

Compute
High Bandwidth Memory

Capacity Memory/NVM

Switch

Pooled Composable Memory/NVM

100-300ns
100-400 GB/s
Invisible to software, not CC
Exposed to HV/container
Memory semantics (ld/st)
Switched (low latency)
Shared resources

Cluster

Source: P. Faraboschi, HPE, “The Data Access Continuum”, SC’19 MCHPC
PM Attach: Who is Doing What

OPTANE™

JeDEC®

CXL

OpenCAPI™

GenZ

DDR-T

NVDIMM-N

NVDIMM-P

Intel

Micron

Samsung

SMART Modular Technologies

AGIGA Tech

SK hynix

Source: Industry announcements & scuttlebutt
PM Usage
Value Prop, Tiering, Pooling, Computational Memory
1. **PM = More Memory**
   - Pro: Existing applications want it
   - Con: Compete directly vs. DRAM on price

2. **PM = Persistent & Coherent**
   - Pro: Take full advantage of features
   - Con: Application needs to know

PM Usage: Data Tiering

With DRAM:
- OLAP (In-Memory Processing)
- Hot Tier: DRAM 8-256GB
- Warm Tier: NVMe SSD 1TB
- Cold Tier: SAS/SATA SSD 10TB

With Persistent Memory:
- Collapse Hot & Warm Tier to process or store critical data
- Persistent Memory: 16GB – >1TB
- OLAP – Online Analytical Processing
- OLTP – Online Transaction Processing

Source: A. Sainio, SMART, “PM Benefits AI/ML”, IMC Summit 2019
PM Usage: Move SSD tasks into PM

PROOF POINT: CASSANDRA 4.0* DATABASE

IOPS performance vs. Comparable Server System with DRAM and NAND SSD

- 9x the Read IOPS!
- Persistent
- Coherent

## PM Usage: Move SSD tasks into PM

### FORSA Enhances Performance and Usability of Persistent Memory:

<table>
<thead>
<tr>
<th>MySQL Transactions Per Second</th>
<th>Intel NVMe SSD P4510</th>
<th>Intel Optane DCPMM Linux Native (Block)</th>
<th>Intel Optane DCPMM FORSA LEM (Block)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,824</td>
<td>9,211</td>
<td>10,299</td>
<td></td>
</tr>
<tr>
<td>84.5</td>
<td>21.5</td>
<td>17.6</td>
<td></td>
</tr>
</tbody>
</table>

- **MySQL 99th Pctl. Latency (ms)**
  - **Intel Cascade Lake - NVMe SSD**: 2x 8160M CPU 24C, 2.1GHz, 24x DDR4 2666 MT/s (384GB)
  - **Intel Cascade Lake - DCPMM**: 12x DDR4 2666 MT/s (384GB)
  - **n/a**: 12x Intel Optane DCPMM (1.5TB)
  - **FORSA 3.0**: FORSA 3.0
  - **System Memory: 384GB**: System Memory: 384GB
  - **MySQL TPS: 3,824**: MySQL TPS: 10,299
  - **MySQL 99 Pctl. Latency: 85ms**: MySQL 99 Pctl. Latency: 18ms
  - **System Price: $22,879.80**: System Price: $29,614.95
  - **$USD / TPS: $5.98**: $USD / TPS: $2.88

### Source: J. Xie, Formulus Black, “PM Delivers Superior ROI”, Webinar 2020

**2.4x the transactions...**

**@ 50% lower $ per transaction!!!**

[SNIA Persistent Memory Summit](https://www.snia.org)
PM Future: “Pooled Memory Appliance”

All Top Supercomputers Are Heterogeneous
Both within their Processors/Accelerators and their Memories

- World’s Top: Summit 200 PetaFlops at 14MW
- 4,608 nodes
  - 9,216 IBM POWER9 22-core CPUs and 27,648 Nvidia Tesla GPUs
  - > 95% of Flops comes from GPUs
- Over 600 GiB of coherent memory per node
  - 6×16 = 96 GiB HBM2
  - 2×8×32 = 512 GiB DDR4 SDRAM
  - Addressable by all CPUs and GPUs
- 800 GB of non-volatile RAM that can be used as a burst buffer or as extended memory
- Challenging to program system with heterogeneity
  - Not all software can be ported

Opportunity of Memory Heterogeneity

- Why have systems become heterogeneous?
  - Slowdown in frequency scaling, lithography scaling, interconnect scaling
  - Increase in specialization, problems that are embarrassingly parallel, higher intensity compute
  - Mixed compute requirements - variable access patterns, some very regular, others very irregular such as pointer chasing, indirection
  - High energy and latency of data movement - finer grain accelerators, data transformations
  - Advent of emerging memory technologies
  - Advances in packaging
- Challenges
  - Compute and Software - offload and communications overhead, variable performance, determinism, programming complexity, mapping increasingly varied workloads
  - Emerging memory explosion
    - Recent introduction of STT-MRAM, ReRAM into foundries, introduction of 3DXpoint by Intel/Micron, 5 classes of NAND Flash – fast SLC, SLC, MLC, TLC, QLC, with varied maturity and performance
  - Attaching the memory into hardware
  - Additional memory tiers presents challenge to app’s software and operating system
  - Data movement between memory types
  - Widened variety of data access patterns

PM Future: “Pooled Memory Appliance”

All Top Supercomputers Are Heterogeneous
Both within their Processors/Accelerators and their Memories

- World’s Top: Summit, 200 TFlops, 140,000 CPUs
- 4,608 nodes
  - 9,216 IBM POWER
  - > 95% of Flops comes from CPUs
- Over 600 GiB of cache
  - 6x16 = 96 GiB HBM
  - 2x8x32 = 512 GiB DRAM
- Addressable by all devices and GPUs
- 800 GB of non-volatile memory that can be used as a buffer or as extended main memory

Challenging to process with heterogeneity
- Not all software can run on all devices

Source: P. Faraboschi, HPE, “The Data Access Continuum”, SC’19 MCHPC
PM Future: “Computational Memory”

Use Case: CXL-Based NVDIMM + Accelerator

- Controller chip includes compute functions (e.g. AI, search, graph database)
- Controller chip can be programmed via PCIe driver (e.g. NVMe).
- NVM can still be exposed to host and accessed via CXL.mem (volatile or persistent)

Source: S. Bates, Eideticom, “Successfully Deploying PM and Acceleration”, PIRL 2019
Conclusions

4 Things I think I think
4 Things I Think I Think

1. I ♥ DRAM! (But...)  
   • Watch: Tech and Price scaling stall gets worse

2. 3DXP/PCM still the top PM Media for the job  
   • Watch: Ferro HfOx 1T-1C research advances

3. DIMM slots too precious for PM  
   • Watch: CXL becomes primary PM attachment; big bandwidth increase

4. Best PM usage values Persistence and Coherence  
   • Watch: Pooled Memory Appliances and “Computational Memory”
Thank you

Please visit www.snia.org/pmsummit for presentations

Dave Eggleston
Intuitive Cognition Consulting
Technology & Business Strategy

Email: dave@in-cog.com
Twitter: @NVM_DaveE
LinkedIn: linkedin.com/in/deggleston/
Abstract

The commercialization of Persistent Memory is well underway, fundamentally changing the way processors load and store information. The characteristics of the non-volatile memory media (3DXP, PCM, MRAM, RRAM, others) determine the basic performance, reliability, and cost that Persistent Memory can deliver to the system. Brand new methods of processor attachment (DDR5, DDR-T, NVDIMM, CXL, CCIX, Gen-Z, OpenCAPI) for Persistent Memory promise to broaden its usage far beyond “DRAM replacement”, opening up vast resources enabling memory and power hungry applications.

In this talk, the speaker will:

• Discuss the basic characteristics of Persistent Memory Media
• Articulate how Persistent Memory enhances server architectures
• Identify, compare, contrast the new methods for Persistent Memory attachment
• Propose innovative ways applications will utilize Persistent Memory to overcome memory and power limitations
• Discuss who is doing what within the Persistent Memory ecosystem