OpenPOWER™

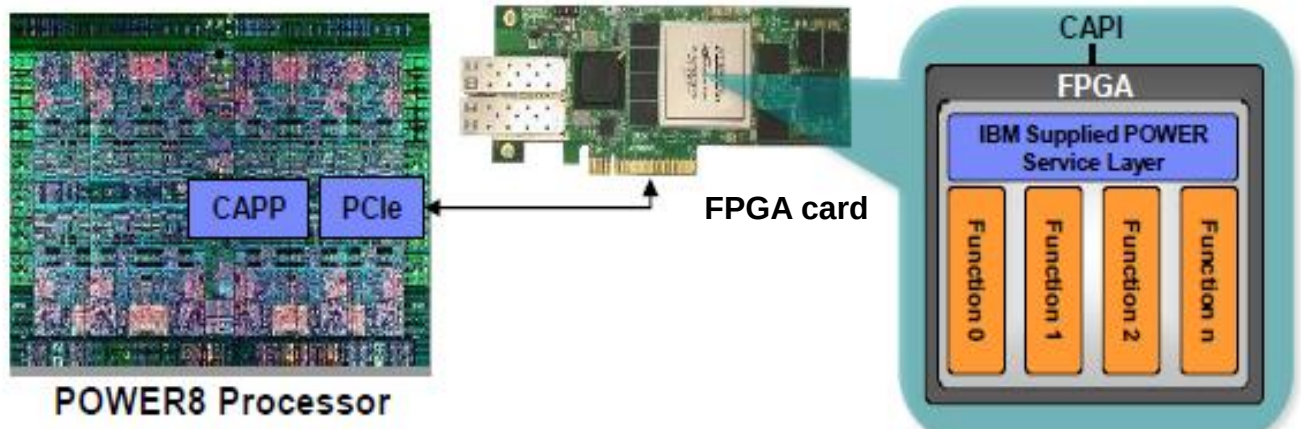
Phase Change Memory Access in OpenPOWER Systems using CAPI

E. Bougioukou, A. Prodromakis, N. Toulgaridis, T. Antonakopoulos
University of Patras, Greece

N. Papandreou, U. Egger, H. Pozidis, E. Eleftheriou
IBM Research – Zurich, Switzerland

Objective

- Demonstrate CAPI-attached Phase-Change Memory (PCM) in OpenPOWER servers
- Showcase the efficiency of CAPI protocol in data access from non-volatile memory
- Leverage the low latency and small granularity access of PCM
- Build a platform where next generation PCM chips can be attached to OpenPOWER servers via CAPI and tested on real-world workloads



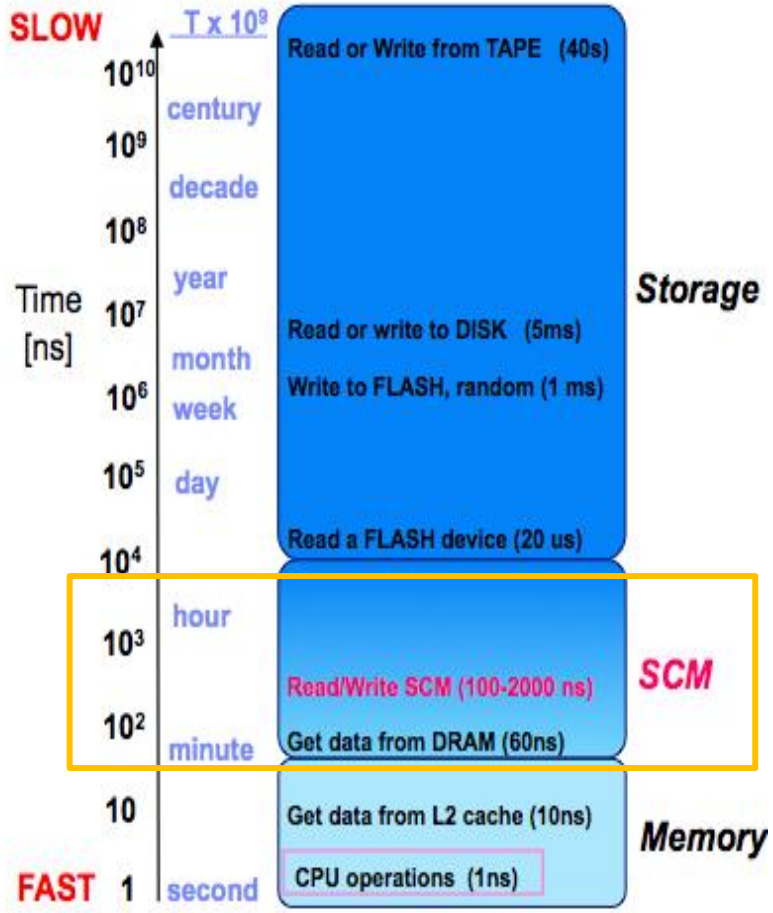
CAPP: Coherent accelerator processor proxy

Outline

- Phase-change memory (PCM)
- CAPI-PCM prototyping system
- FPGA HW architecture
- Latency and performance measurements
- Comparison with SSD
- Summary

Phase-Change Memory (PCM)

- **Storage Class Memory**: a solid-state memory that blurs the boundaries between storage and memory by being low-cost, fast, and non-volatile
- **Phase-Change Memory (PCM)** is the top contender for realizing Storage Class Memory
  - read latency: faster than NAND (100s of ns vs. 10s of us)
  - write endurance: more than 10^6 cycles
  - scalable, multi-bit capability
  - non-volatile, true random access
- Exploit PCM in the system hierarchy
  - hybrid memory: a combination of DRAM as the small main memory and PCM as the large far memory
  - fast durable storage: PCM is used as a cache for hot data in front of a NAND flash storage pool

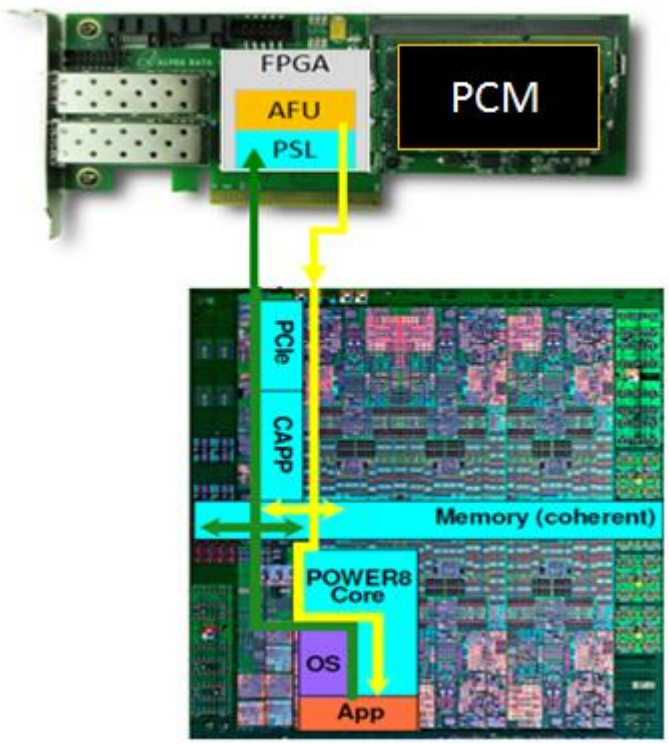


Prototyping System

IBM Power System S812LC

Tyan Palmetto - OpenPOWER CRS

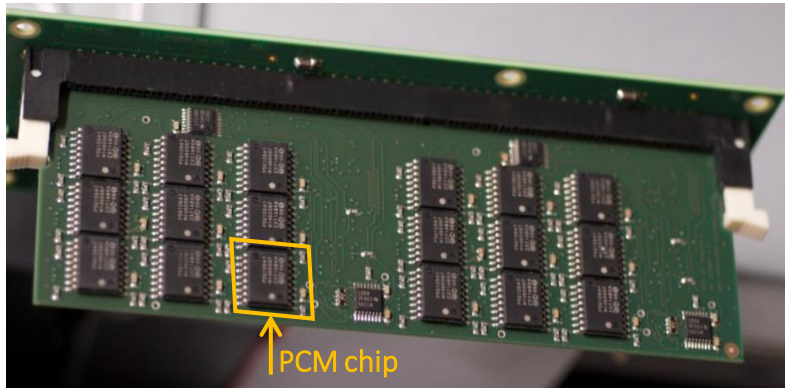
ADM-PCIE-7V3



Power 8 Processor

- IBM Power System S812LC
  - 8-core 3.32 GHz POWER8 processor
  - 32 GB 1333MHz DDR3 DIMM memory
  - CAPI enabled PCIe Gen3 slot
- Tyan Palmetto Custom Reference System
  - 12-core 3.32 GHz POWER8 processor
  - 32 GB 1333MHz DDR3 DIMM memory
  - CAPI enabled PCIe Gen3 slot
- Ubuntu 15.10
- ALPHA DATA ADM-PCIE-7V3 card
  - Xilinx Virtex-7 FPGA
  - CAPI enabled
  - Custom AFU and PCM controller
- Custom prototype PCM-based NVDIMMs

Phase-Change Memory Card

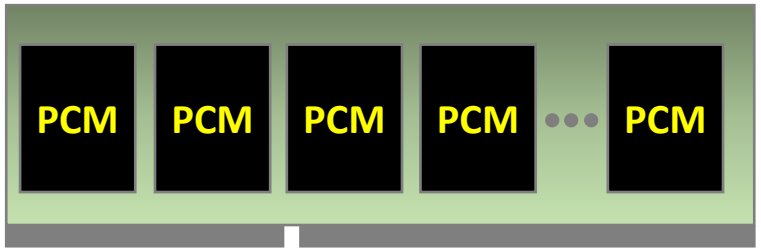


Legacy P5Q prototyping card: I. Koltsidas et al., NVMW 2014

Legacy Micron 90 nm PCM chip
128 Mb SLC PCM
SPI compatible serial interface (66 MHz)
64 bytes R/W access
READ/WRITE access time: 100ns/120usec

Case 1 (PCM parts)

- Legacy parts from Micron: 90 nm P5Q 128Mb SLC Phase-Change Memory
- Designed custom PCM-based cards and SODIMM adapters
- Connect P5Q cards directly to ADM-PCIE-7V3 FPGA card



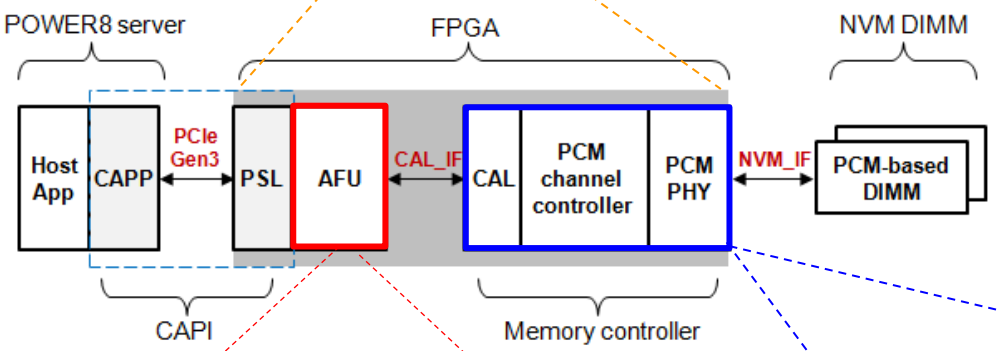
Next generation PCM technology characteristics : J. Cheon et al., IEEE CICC 2015

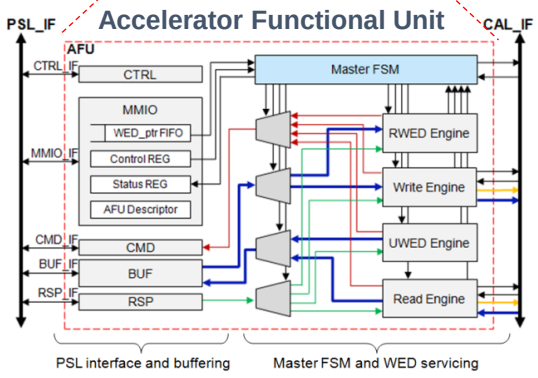
Next generation 25 nm PCM chip
16/32 Gb SLC/MLC PCM
DDR3 compatible interface
8 bytes R/W access
READ access time: 450 nsec

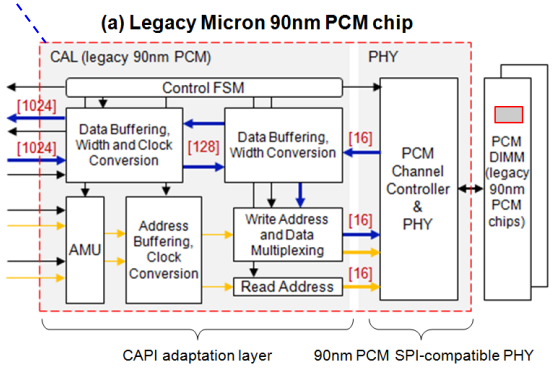
Case 2 (test scenario)

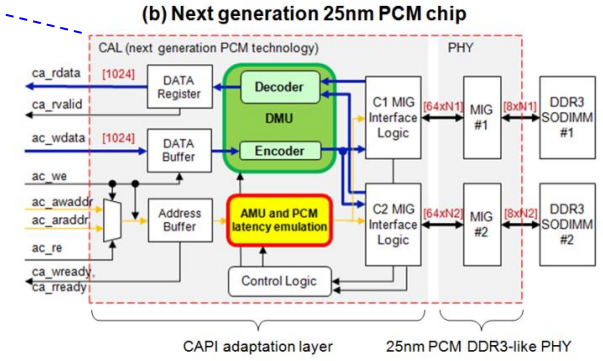
- Next generation high capacity Phase-Change Memory
- Used DRAM modules and special HW to emulate PCM channel R/W latency
- Architecture enables evaluation of different NVM technologies

FPGA HW architecture







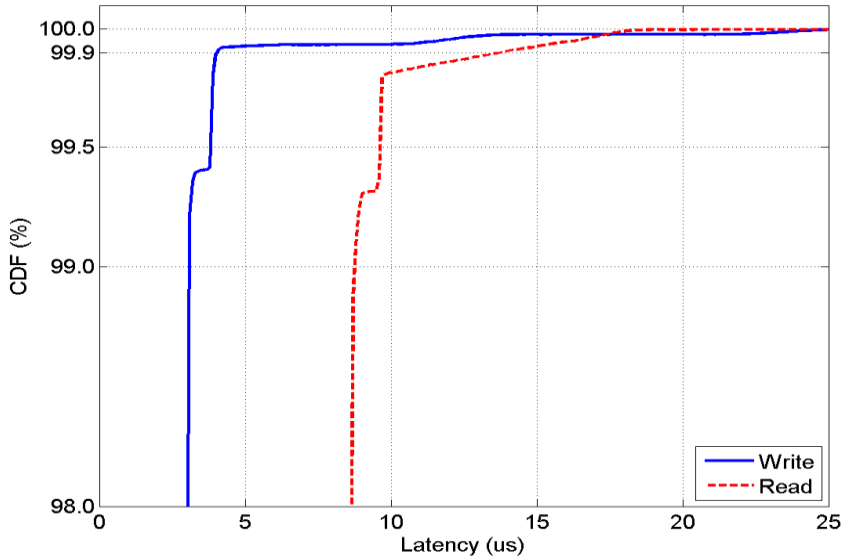


- AFU implements PSL interface and WED management
  - WEDs support multiple R/W commands
  - Multiple threads from the Host can form a single WED
- PCM channel controller for legacy PCM parts
  - Memory channel consists of 2x3x3 P5Q chips
  - Controller supports 8 channels in total
- PCM channel controller for next generation PCM technology
  - User-defined channel configuration
  - Special HW exposes host to PCM chip R/W latency

Details of HW architecture: A. Prodromakis et al., IEEE FPL 2016

Latency measurements

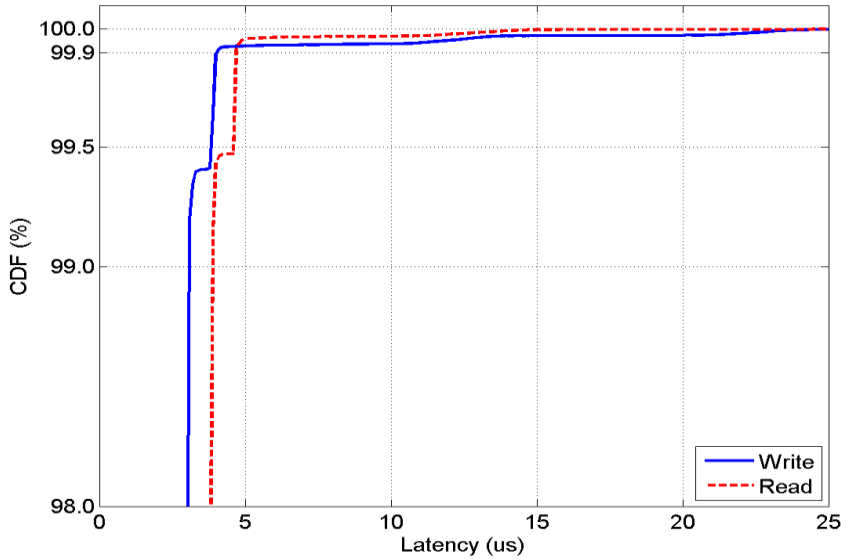
Legacy PCM parts



Workload	50%	99%	99.9%
128B Write	2.9 us	3.1 us	4.1 us
128B Read	8.6 us	8.8 us	13.8 us

↑ (~ 4.5 us due to chip serial command/data interface)

Next generation PCM technology



Workload	50%	99%	99.9%
128B Write	2.9 us	3.1 us	4.1 us
128B Read	3.7 us	3.9 us	4.7 us

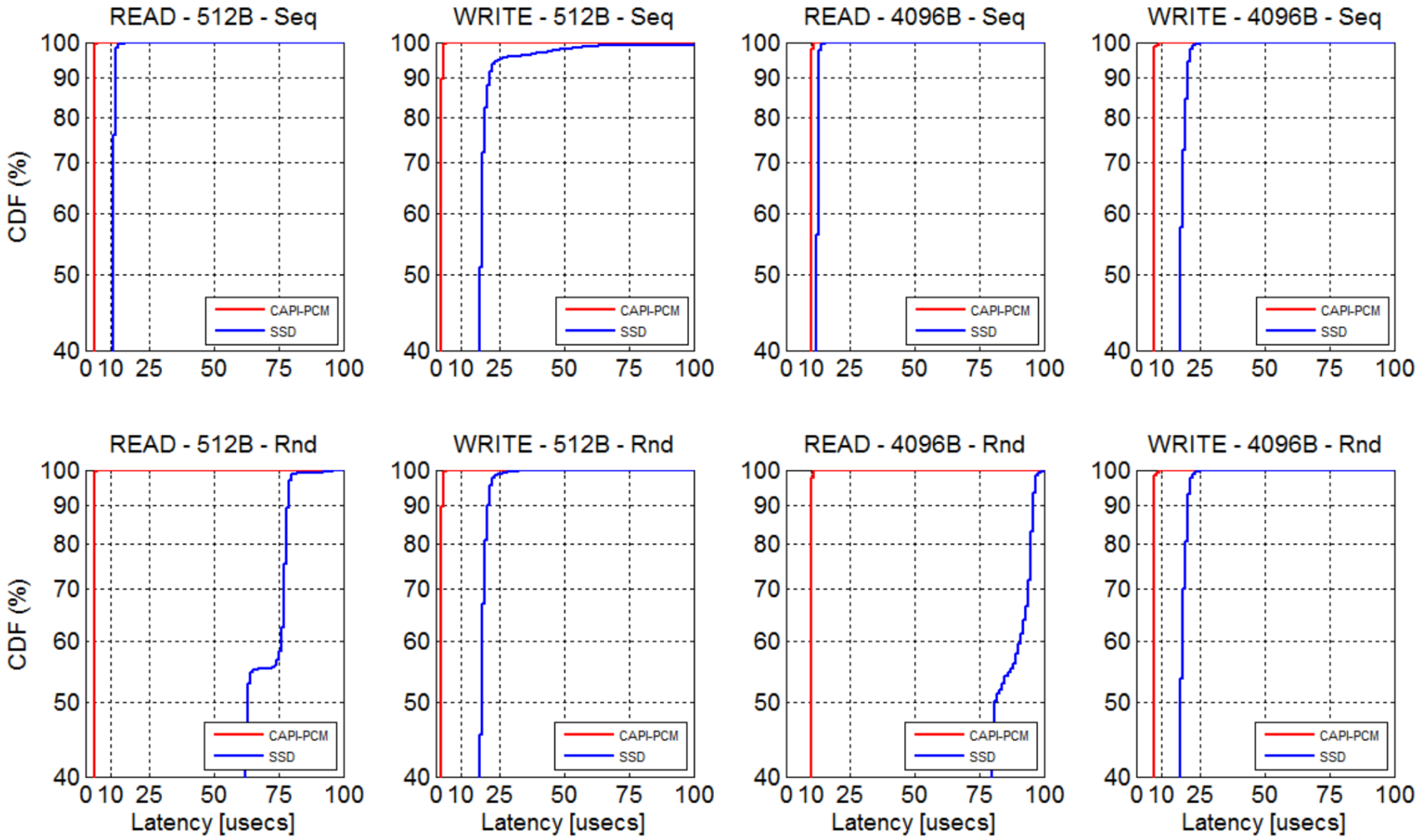
- Latency measured on 128B R/W access (CAPI cache line size)
- 99% of reads complete within **8.8 us** for **legacy PCM parts** and **3.9 us** for **next generation PCM** technology

Very low R/W latency with very low variance

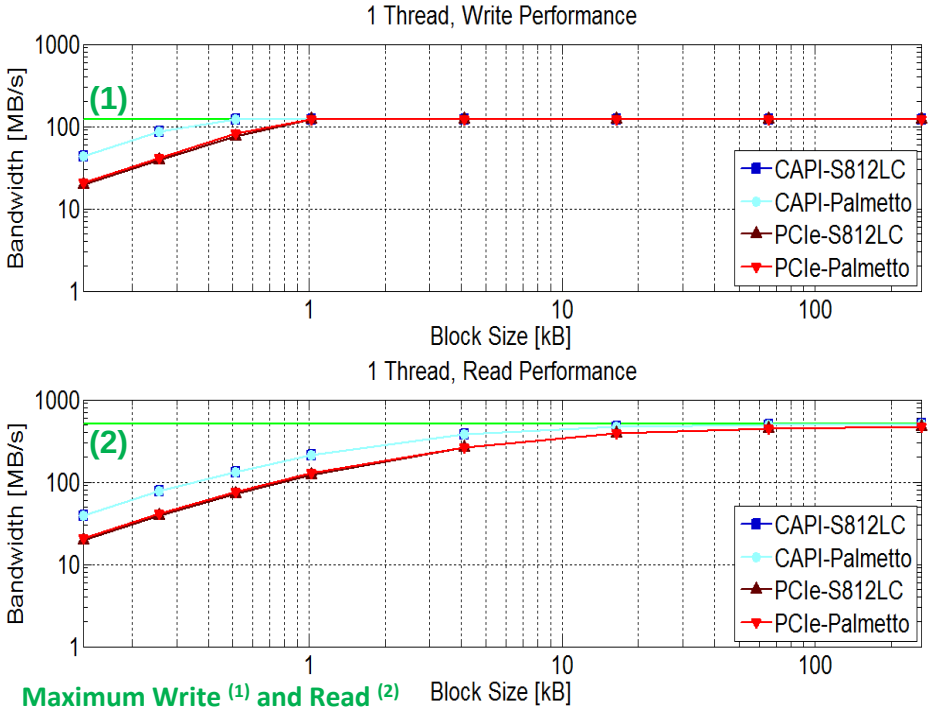
CAPI-PCM vs SSD: Latency

- Various data sizes:
  - 512 B (sector)
  - 4096 B (OS page)
- Access modes
  - Sequential
  - Random
- SSD specs:
  - PCIe Gen. 3, 4 lanes
  - NVM Express
  - 1.2 TB

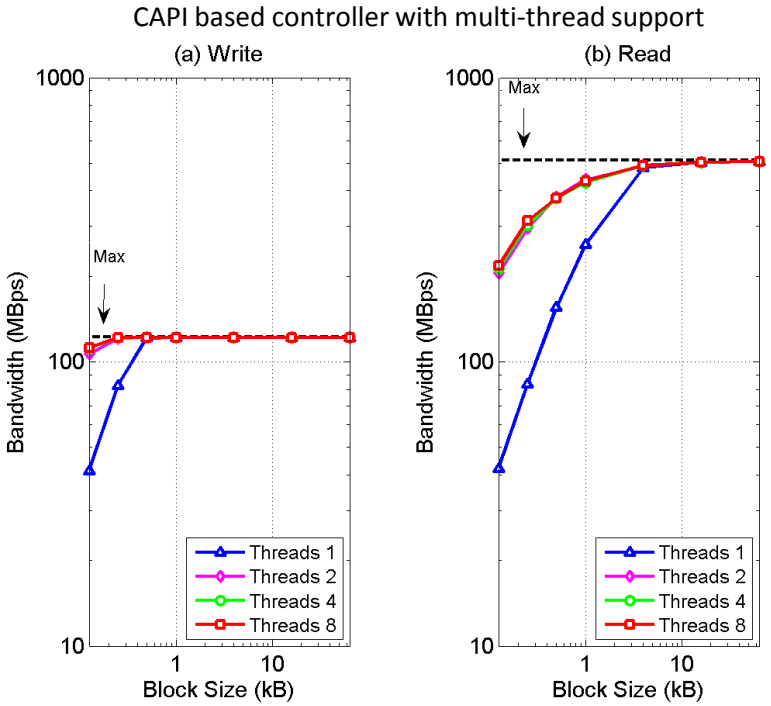
The CAPI-PCM demonstrates faster and more consistent response times.

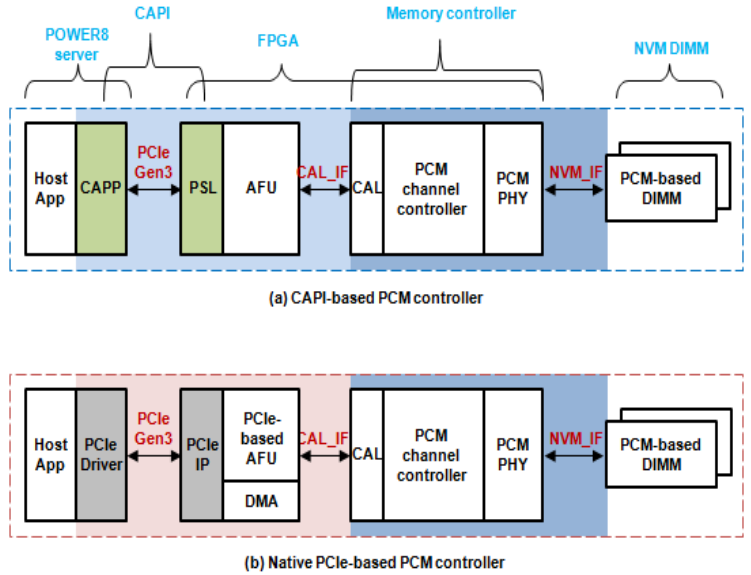


Performance Measurements



Maximum Write [1] and Read [2] PHY bandwidth of next generation PCM channel scenario





- CAPI shows better performance than native PCIe Gen3 design
- System approaches maximum memory channel BW with multi-threads support
- Optimization of descriptors and data management at AFU can further improve the R/W performance (same is true for PCIe design)

Summary

- Demonstrated Phase-Change Memory access in OpenPOWER servers using CAPI
  - Implemented custom AFU on Xilinx FPGA
  - Designed custom CAPI-based PCM controllers for 2 different technologies (old and new PCM)
- CAPI-PCM demonstrates low read & write latency with very low variance
  - 99% of reads of 128B complete within 3.9 usec for the new PCM technology test-case
- CAPI-PCM demonstrates faster and more consistent read & write latency than SSD
- Ongoing work:
  - Get more performance out of the protocol by optimization
  - Extend design to support multiple PCM channels for higher throughput

Acknowledgements

- **Nonvolatile Memory Systems Group, IBM Research – Zurich**
  N. Papandreou, U. Egger, T. Mittelholzer, M. Sifalakis, H. Pozidis, E. Eleftheriou

- **Cognitive Computing Machines and Embedded Systems Group, University of Patras, Greece**
  E. Bougioukou, A. Prodromakis, N. Toulgaridis, M. Varsamou, T. Antonakopoulos

- **Accelerator Technologies Group, IBM Research – Zurich**
  R. Polig, H. Giefers, C. Hagleitner