Controller Architecture for Low-latency Access to Phase-Change Memory in OpenPOWER Systems

A. Prodromakis[1], N. Papandreou[2], E. Bougioukou[1], U. Egger[2], N. Toulgaridis[1], T. Antonakopoulos[1], H. Pozidis[2], E. Eleftheriou[2]

[1]University of Patras, 26504 Rio – Patras, Greece
[2]IBM Research – Zurich, 8803 Rüschlikon, Switzerland

IBM

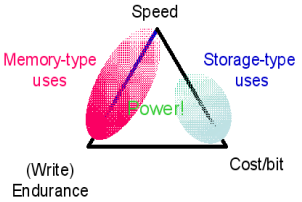
FPL 16

26th International Conference on Field-Programmable Logic and Applications
SwissTech Convention Centre, Lausanne, Switzerland, 29th August – 2nd September 2016
Session S4a: Connectivity, Communication, and Supply Chains

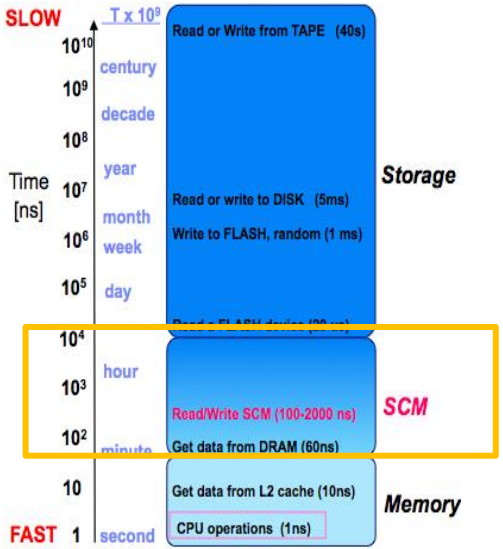
IBM

Introduction

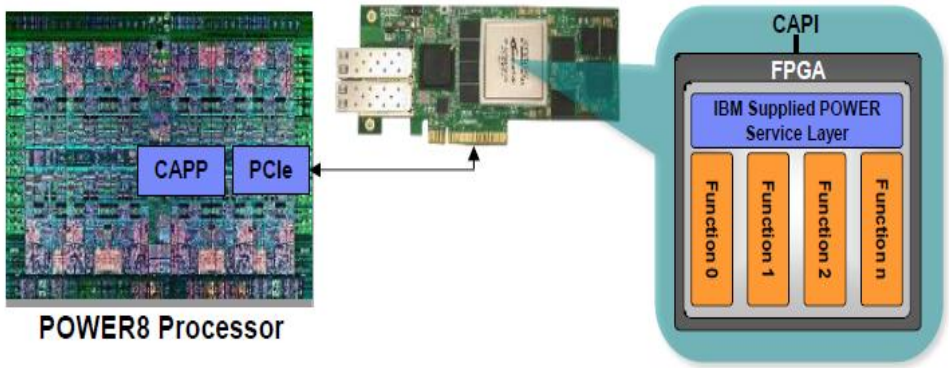
- **Phase-Change Memory** (PCM) is the top contender for realizing **Storage Class Memory**
  - read latency: faster than NAND (100s of ns vs. 100 of us)
  - write endurance: more than 10^6 cycles
  - scalable, nonvolatile, true random access
  - multi-bit capability (2016 TLC PCM demonstration by IBM)
- Exploit PCM in the system hierarchy
  - **hybrid memory**: a combination of DRAM as a 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
- This work presents the architecture, implementation and performance of an FPGA-based PCM memory controller for OpenPOWER systems
- The controller leverages the **Coherent Accelerator Processor Interface** (CAPI) of the POWER8 processor in order to offer to the CPU low-latency and small granularity access to PCM



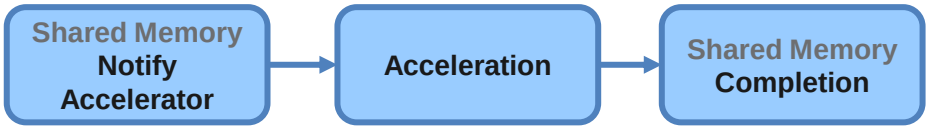
Storage Class Memory
A solid-state memory that blurs the boundaries between storage and memory by being low-cost, fast, and non-volatile.



CAPI and OpenPOWER



I/O flow with Coherent Model



J. Stuecheli, IEEE ASAP 2014
B. Wile, IBM Enterprise2014

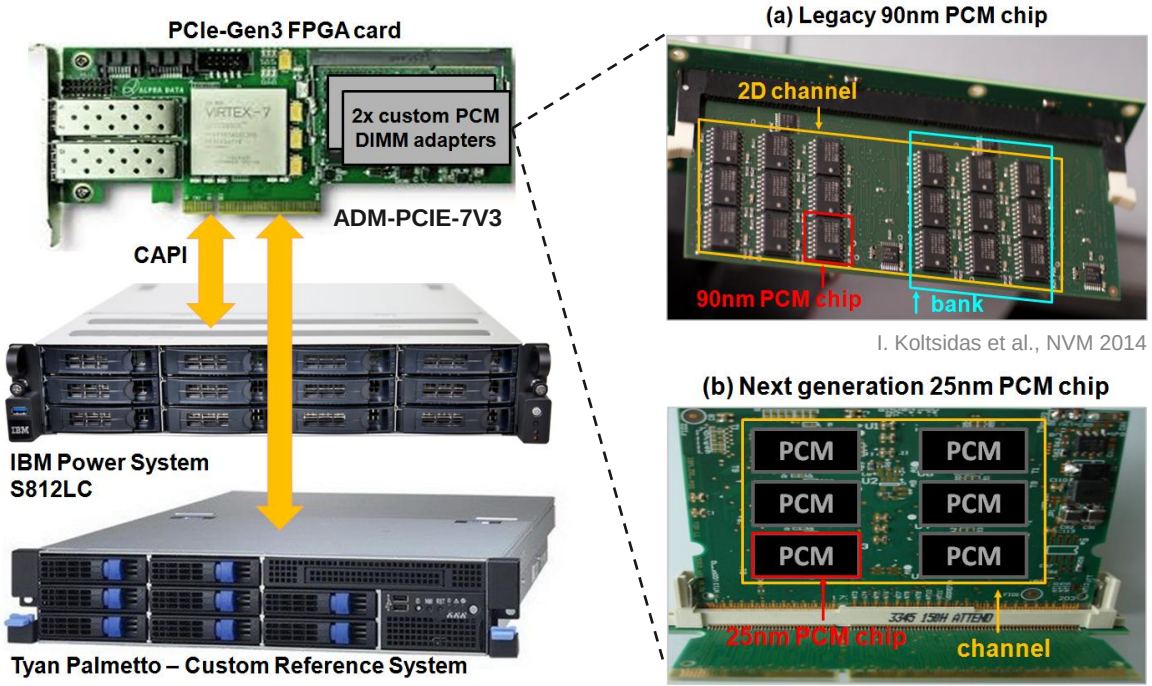
Coherent Accelerator Processor Interface (CAPI)

- CAPI connects a custom acceleration engine to the coherent fabric of the POWER8 chip
- The protocol is sent over the PCIe; Native PCIe Gen3 Support (x16); direct processor integration
- Memory coherency and address translation are handled automatically by CAPI
- CAPI removes the overhead and complexity of the I/O subsystem, allowing an accelerator to operate as an extension of an application

Advantages of CAPI over I/O attachment

- Virtual addressing and data caching (significant latency reduction)
- Easier, natural programming model (avoid application restructuring)
- Enables applications not possible on I/O (e.g. pointer chasing, shared memory semaphores)

Prototyping Platform



I. Koltsidas et al., NVM 2014

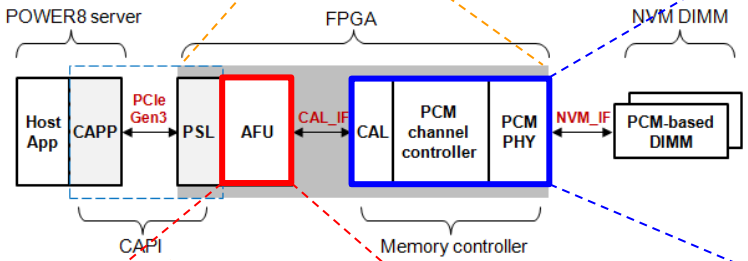
IBM Power System S812LC / Tyan Palmetto
8-core 3.32 GHz POWER8 processor
32 GB 1333MHz DDR3 DIMM memory
CAPI enabled PCIe-Gen3 slot

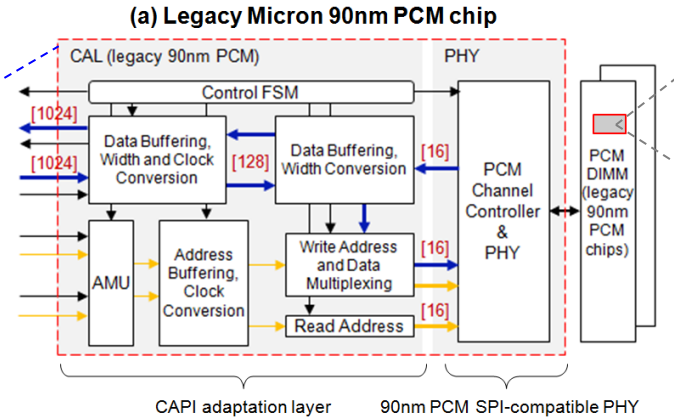
Legacy Micron 90nm PCM chip
128 Mb SLC PCM
SPI compatible serial interface (66 MHz)
64 bytes R/W access
WRITE access time: 120 usec
READ access time: 100 nsec

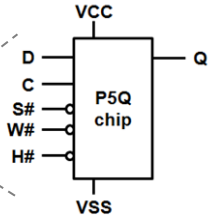
Next generation 25nm PCM chip
16/32 Gb SLC/MLC PCM
DDR like interface
READ access time: 450 nsec

- OpenPOWER servers running Ubuntu 15.10 (IBM Power System S812LC, Tyan Palmetto CRS)
- CAPI-enabled FPGA cards (Alpha Data ADM-PCIE-7V3 – Xilinx Virtex 7)
- Custom made PCM DIMMs and adapter cards (legacy 90nm Micron PCM, next generation 25nm PCM)

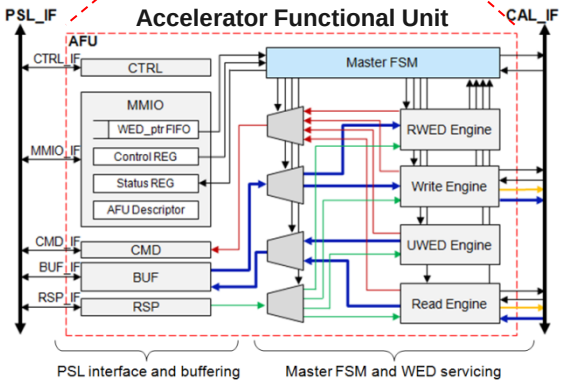
FPGA Architecture of CAPI-based PCM controller



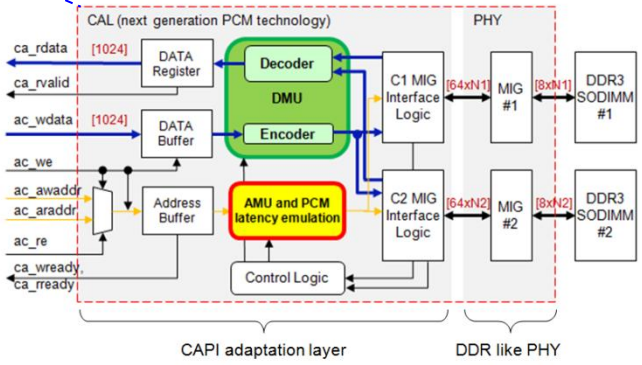




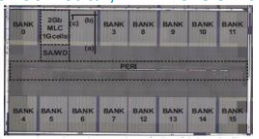
- PCM channel consists of 2x3x3 PQ5 chips
- Controller supports 8 channels in total
- Data width & clock conversion due to slow serial interface



- AFU implements PSL I/F along with WED management and control
- 4 special HW engines prepare the data and service the R/W requests
- WED supports multiple R/W commands; multiple threads from the Host can form a single WED



J. Cheon et al., IEEE CICC 2014



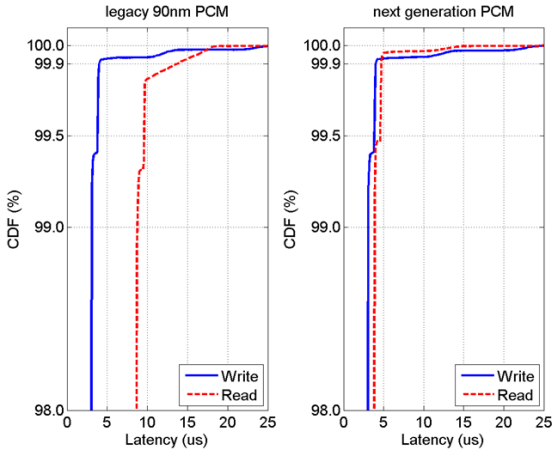
- Special HW for PCM chip R/W latency emulation
- BCH encoder/decoder
- Supports user-defined channel configuration: number of PCM chips per DIMM

Performance results

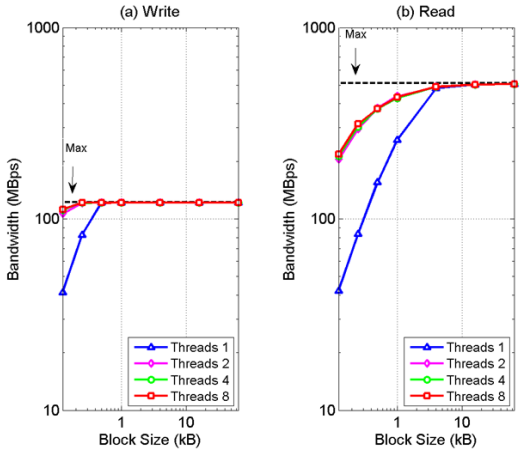
Latency of 128 Byte READ/WRITE access

Legacy 90nm PCM chip	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 IF			
Next generation PCM chip	**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 Measurements



Performance Measurements



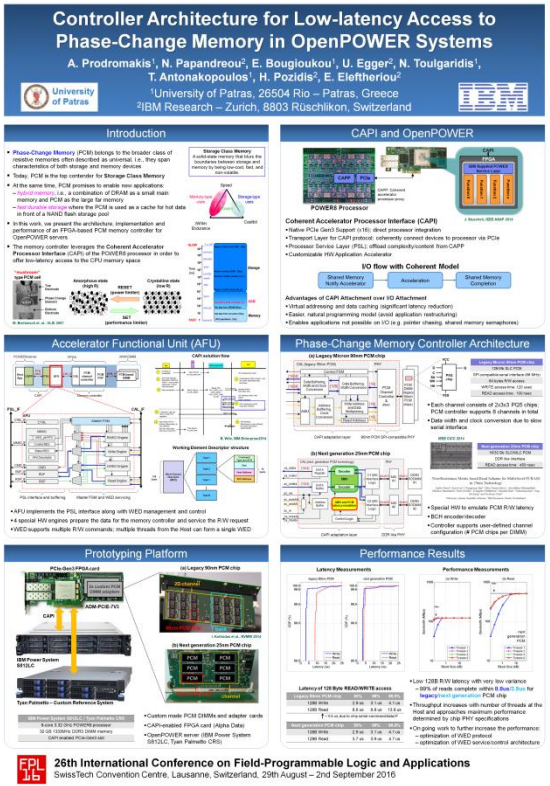
Next generation PCM technology

- 128B R/W access: low latency with very low variance
  - **99%** of reads complete within **8.8us**/**3.9us** for **legacy**/**next generation** PCM chip
- Throughput increases with number of threads at the Host and approaches maximum determined by PCM chip PHY
- On going work to further increase the performance:
  - optimization of WED protocol
  - optimization of WED service/control architecture

Poster Session

Wednesday
31st August

09:00	**K2** Gustavo Alonso (ETHZ, CH)
10:00	Coffee Break
10:30	**S4a** Connectivity, Communication, and Supply Chains / **S4b** Low Level Architecture (2), Pipelining, and Timing
12:35	Lunch
14:15	**K3** Christoph Hagleitner (IBM, CH)
15:15	Coffee Break and Posters
16:00	**S5a** Data Analysis and Databases / **S5b** Compilation
17:30	PhD Forum Elevator Pitch Session
18:00	



For more details and fruitful discussions

visit us at the

Poster Session

Wednesday 31st August
3:15pm – 4:00pm