Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

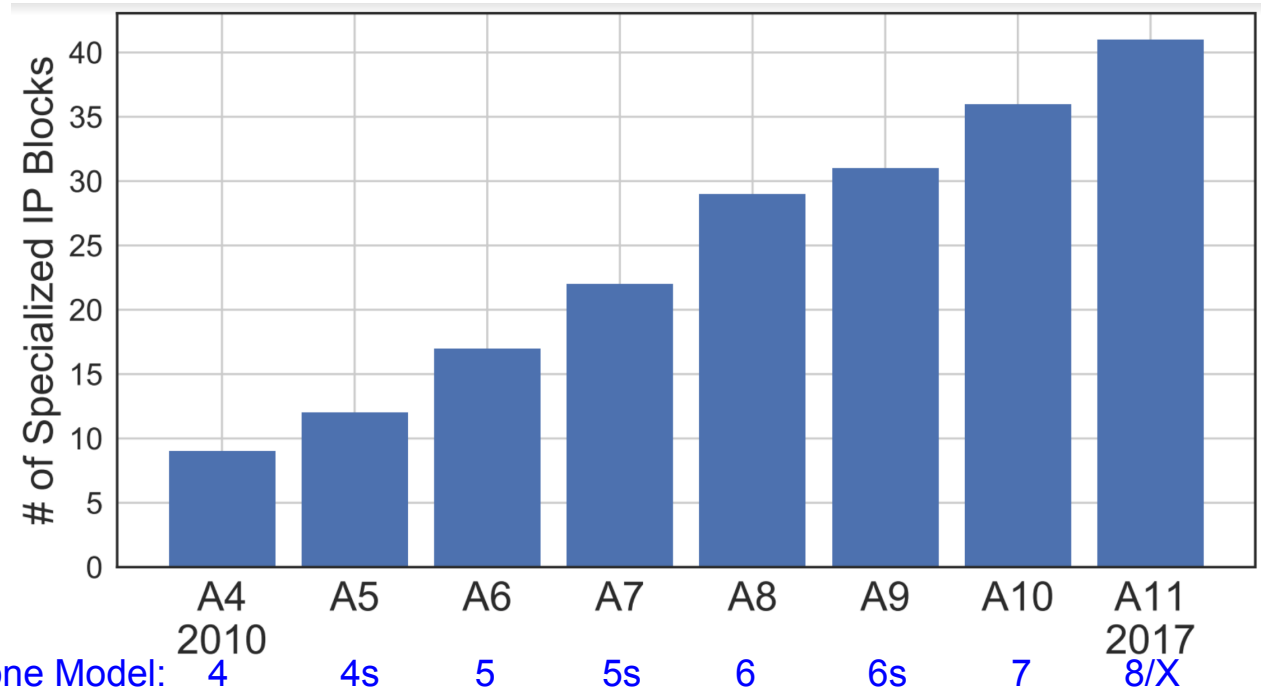
The Immediate Future is Heterogeneous but Coherent

Michael Wang
Fermilab, SCD/SSA/SSI
Computing R&D Micro-Retreat
April 20, 2018

A Path to the Post-Moore era: Heterogeneous Computing

- Our approach so far: multi-core CPUs, many-core architectures: GPU, Xeon Phi (defunct) accelerators
- Another approach: dedicated hardware for compute-intensive tasks

From: http://vlsiarch.eecs.harvard.edu/accelerators/die-photo-analysis



How the processor on the Apple iPhone has evolved in terms of the amount specialized hardware that is neither CPU nor GPU

Fermilab

A Path to the Post-Moore era: Heterogeneous Computing

FPGAs offer practical way of dedicating custom hardware for compute-intensive tasks:

- Programmable/reconfigurable
- Extremely fast
- Massively parallel: data and task parallelism, but also pipeline parallelism
- Energy efficient
- Significant expertise at FNAL with site licenses for major design tools

Difficult to program, but new developments helping to ease this:

- High-level synthesis tools: incorporated into Intel-Altera/Xilinx design tools, OpenCL, other C-based, now even Python!
- Available in the cloud – Amazon AWS. Our capable colleagues in PPD/CMS looking into this.

Successfully applied to mainstream applications in other fields: finance, bioinformatics, etc.

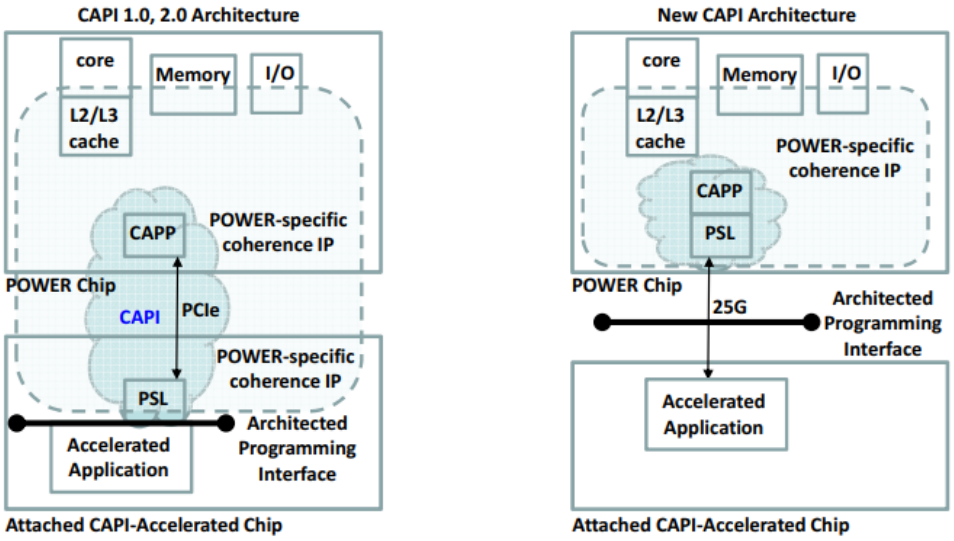
- So far mostly used in DAQ/Trigger applications in HEP:
- Start to look into possible offline HEP applications: Reco, DL, simulations, multivariate analyses, etc.

The Missing Link to a Heterogeneous Future: Coherent Interconnects

- Current accelerators provide some respite in the Post-Moore computing era, but for how long?
- Despite the raw computational power, not tapping full potential because used in a sub-optimal way -- as loosely-coupled IO attached devices that operate in separate address spaces. Significant cycles wasted from device driver overheads, copying memory back and forth, set up time for pinning/unpinning memory, etc.
- Fortunately, industry figured out a practical way to address this – **Coherent Bus/ Interconnect**:
  - allows the accelerator to share the same virtual address space as the host– all virtual-real address translations done in hardware. Accelerator standing elevated to a "peer" of the host processor
  - exposes the internal cache bus of the host to the accelerator, maintaining cache-coherency throughout.
- All of this is transparent to the user and greatly simplifies the programming model.
- One could imagine implementing a library of hardware accelerated functions on an FPGA that can be invoked as straightforward function calls. (Think LAr"Hard")

The missing link to a Heterogeneous Future: Coherent Interconnects

- Available standards (backed by major players in industry):
  - OpenCAPI (Open Coherent Accelerator Processor Interface)
  - CCIX (Cache Coherent Interconnect for Accelerators)
  - GEN-Z
- OpenCAPI:
  - available now on IBM Power8 and 9 families.
  - OS support available as of RHEL7.2.
  - Power9 servers with OpenCAPI available from IBM & 3rd party suppliers
  - NVLink used in Summit (ORNL) and Sierra (LLNL) Supercomputers on Power9 platforms are proprietary form of coherent interconnect designed specifically for NVidia GPUs.



Fermilab

The missing link to a Heterogeneous Future: Coherent Interconnects

- GEN-Z goes further than OpenCAPI and CCIX by bringing the coherent bus outside the rack and supports disaggregated memory and simple memory operations for accessing storage/persistent memories. Enables in-memory processing for all the processors/accelerators on the coherent bus. Imagine the possibilities in implementing an entire computing cluster or a HEP Online Trigger or DAQ system on such a bus. What would seem like a random collection of distributed, disparate, and heterogeneous devices and components can now operate cohesively and coherently as a single unified and organic whole.