Memory Processing Units
Jaikrishnan Menon, Lorenzo De Carli, Vijayraghavan Thiruvengadam
Karthikeyan Sankaralingam and Cristian Estan*
UW-Madison and *Google

Today’s Processors
- Cores are extraordinarily inefficient
  - High freq., deep pipeline, speculation, caching
    - 50 pj in ALUs, 400pj in overhead
- Memory is far away
  - 100s of cycles
  - 500 pj to fetch 32-bit word
- Optimized for single-thread “arbitrary” programs

Opportunity 1: Applications
- Modern applications are highly concurrent
- Offloadable
  - Have many parts that don’t need high single-thread performance
- Caching doesn’t do well anyway!

Opportunity 2: 3D stacked memory
HMC (Micron & others): high bandwidth, low latency, and low power to move bits

Opportunity 3: “Economics”
- Moore’s law dead, Dennard scaling dead
- Conventional approaches or “traditional” technology scaling provides no improvement
- Economics of new chip design can be tolerated

Memory Processing Units
Jaikrishnan Menon, Lorenzo De Carli, Vijayraghavan Thiruvengadam
Karthikeyan Sankaralingam and Cristian Estan*
UW-Madison and *Google

Today’s Processors
- Cores are extraordinarily inefficient
  - High freq., deep pipeline, speculation, caching
    - 50 pj in ALUs, 400pj in overhead
- Memory is far away
  - 100s of cycles
  - 500 pj to fetch 32-bit word
- Optimized for single-thread “arbitrary” programs

Opportunity 1: Applications
- Modern applications are highly concurrent
- Offloadable
  - Have many parts that don’t need high single-thread performance
- Caching doesn’t do well anyway!

Opportunity 2: 3D stacked memory
HMC (Micron & others): high bandwidth, low latency, and low power to move bits

Opportunity 3: “Economics”
- Moore’s law dead, Dennard scaling dead
- Conventional approaches or “traditional” technology scaling provides no improvement
- Economics of new chip design can be tolerated

Ongoing work
- Database workloads: TPC-H Engine
- End-to-end study
- Multi-HMC system and how to get to 1TB
- FPGA prototyping & Custom cores
- Your suggestion here

MPU hardware
- Compute fabric (90nm)
  - ARM Cortex M3
    - 250 MHz, 1-issue
    - 16KB stack, code
  - 8 cores / vault
  - 128 cores = 1.1W
- MMU can ensure sequential semantics

Programming MPUs
- Shard data among DRAM chips
- CPU-to-MPU memory procedure calls start MPU threads
- MPU-to-MPU continuations

Performance & Energy
- Low IPD (11-29)
- Medium IPD (79-256)
- High IPD (648-4400)

CPU Thread
- MPU Threads (MPCs)

(a) Execution flow for MPU applications
(b) MPU-based Key-Value store
(c) Key-Value store application kernels