Custom Memory Cube

F3

Yu Zou, Vinayak Deshpande, Herman Lam, Alan George

CHREC NSF Center for High-Performance Reconfigurable Computing

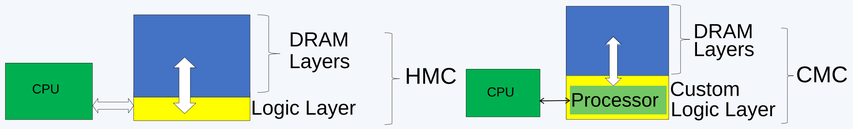
UF UNIVERSITY of FLORIDA

MOTIVATION & GOAL

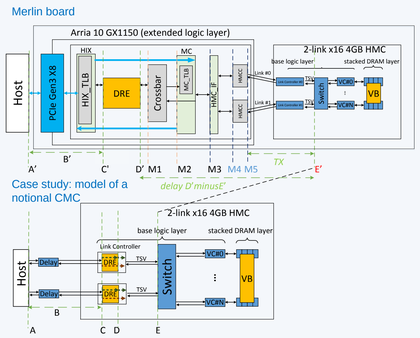
Motivation

- Memory bottleneck – critical for memory-intensive Big Data apps
- Promise of CMC for computational ram & processor-in-memory processing

Goal: Create flexible research platform for design space exploration of CMC apps & arch. before existence of CMC



APPROACH



Phase 1: Prototype Platform

- Create CMC emulation platform using Convey's FPGA-HMC platform (Merlin board)

Phase 2: Case Study

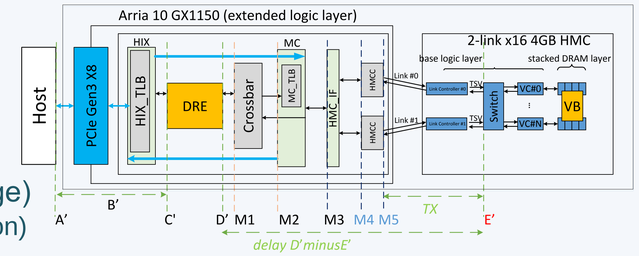
- Select model of CMC
- Develop mapping from measured parameters to model parameters
- Perform design space exploration of CMC arch & CMC operations/apps

PHASE 1: Prototype Platform

- ❑ Observability
  - Instrument Merlin infrastructure with hardware performance monitors (C', D', M1, M2, M3, M4, M5, and E')
- ❑ Usability & Flexibility
  - Implemented & measured kernels and apps using HT (Convey's tool & language)
    - Vector Addition (emulates DRE* fill operation)
    - DRE app: SpMV

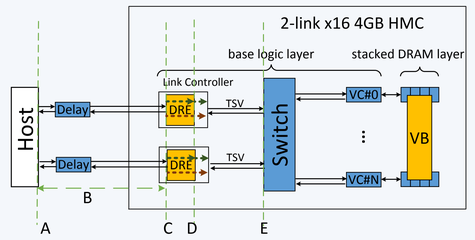
*DRE: Data Reordering/Rearrangement Engine from LLNL

CMC platform on Merlin board

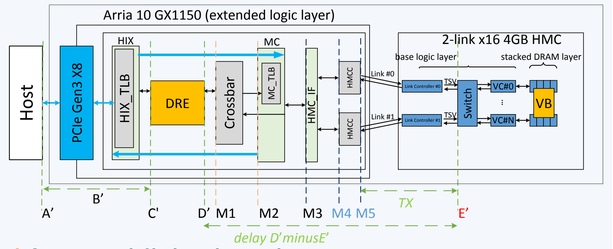


PHASE 2: Case Study

Model of notional CMC



CMC platform on Merlin board



- ❑ Selected notional CMC model from published work
- ❑ Identified hardware parameters required by performance model
- ❑ Mapped measurement points to model parameters as required

PHASE1: Initial Results

Measured performance: simple read/write (µs)
(Averaged over 100,000 iterations)

Oper.	A' (µs)	C'	D'	M1	M2	M3	M4	M5	E'
Read	1095.6	1.03	1.006	0.923	0.725	0.519	-	-	-
Write	1104.0	1.071	1.011	0.991	0.793	0.577	-	-	-

Measured performance: Vector Addition in ms
(100,000 iterations of 2 reads & 1 write)

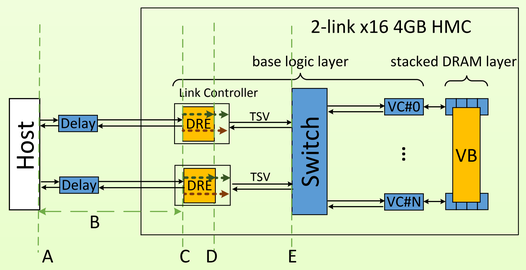
A' (ms)	C'	D'	M1	M2	M3	M4	M5	E'
282.55	281.72	278.72	258.96	199.68	151.31	-	-	-

Measured performance: SpMV – DRE app in seconds
(2^{24} X 2^{24} matrix)

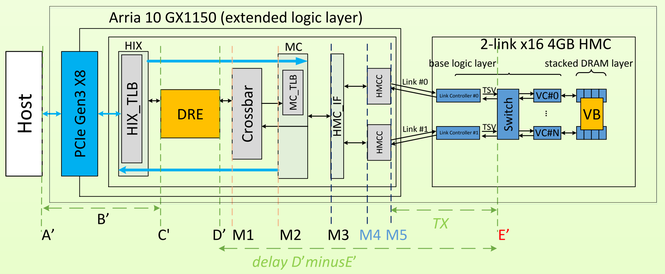
A' (s)	C'	D'	M1	M2	M3	M4	M5	E'
1658.67	1580.14	1559.65	1460.57	1122.42	839.84	-	-	-

PHASE2: Initial Results

Model of notional CMC



CMC platform on Merlin board



$Latency\ A = latency\ C + transfer\ time\ B$

$Transfer\ time\ B$: $time\ to\ transfer\ result\ data\ to\ host\ (e.g.\ VB\ to\ host\ in\ DRE)$

$Latency\ C = latency\ C' - delay\ D'minusE' + delay\ TSV\ *$

$Delay\ D'minusE' = latency\ D' - latency\ E'$

$Latency\ E' = latency\ M_5 - delay\ TX\ **$

$* delay\ TSV = time\ to\ transfer\ request + time\ to\ transfer\ response$

$** delay\ TX = time\ to\ transfer\ request\ to\ HMC\ switch\ logic + time\ to\ transfer\ response$

PLANS FOR 2017

Task 1: Platform development

- Develop library for customization of notional CMC architecture under study
- Explore CMC apps using HT

Task 2: Case studies to explore

- Notional CMC architectures
- CMC apps: DRE, Sorting, Bloom filter
- Characteristics of CMC-amenable apps

Members e.g.,

LPS The Laboratory for Physical Sciences

CHREC NSF Center for High-Performance Reconfigurable Computing

picocomputing is now Micron

CONVEY COMPUTER is now Micron

Micron