Analyzing the Performance Benefit of Near-Memory Acceleration based on Commodity DRAM Devices

Hadi Asghari-Moghaddam and Nam Sung Kim

University of Illinois at Urbana-Champaign

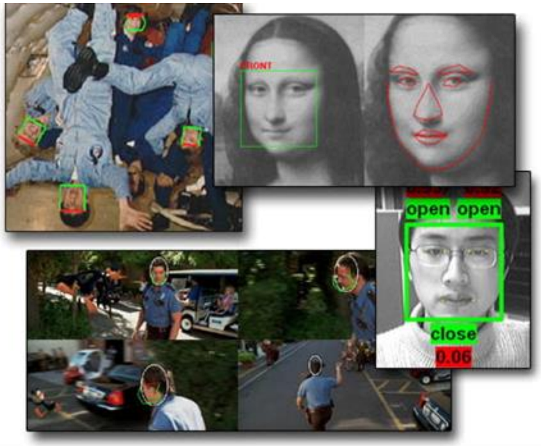
ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

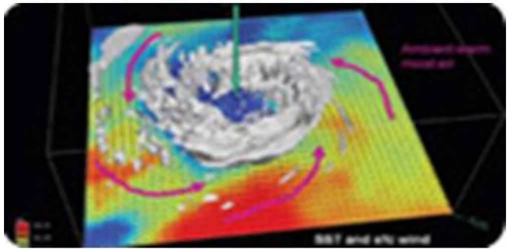
Why Near-DRAM Acceleration?

- higher bandwidth demand but stagnant increase
  - ✓ higher data rate and/or wider bus limited by signal integrity package pin constraint

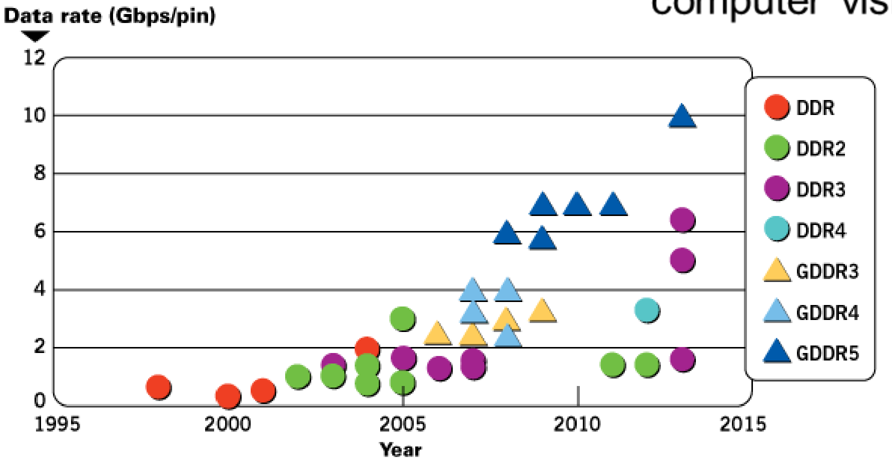
SUHD/3D-graphics



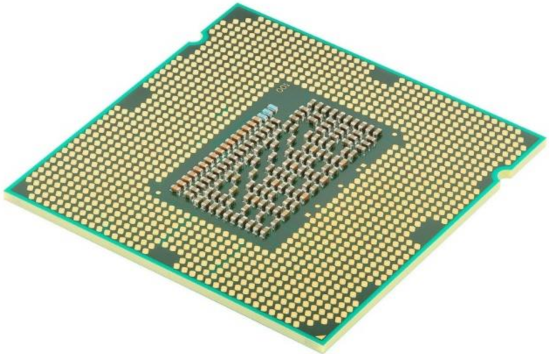
computer vision

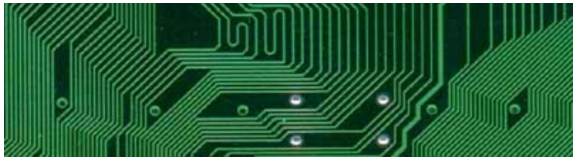


scientific/engineering



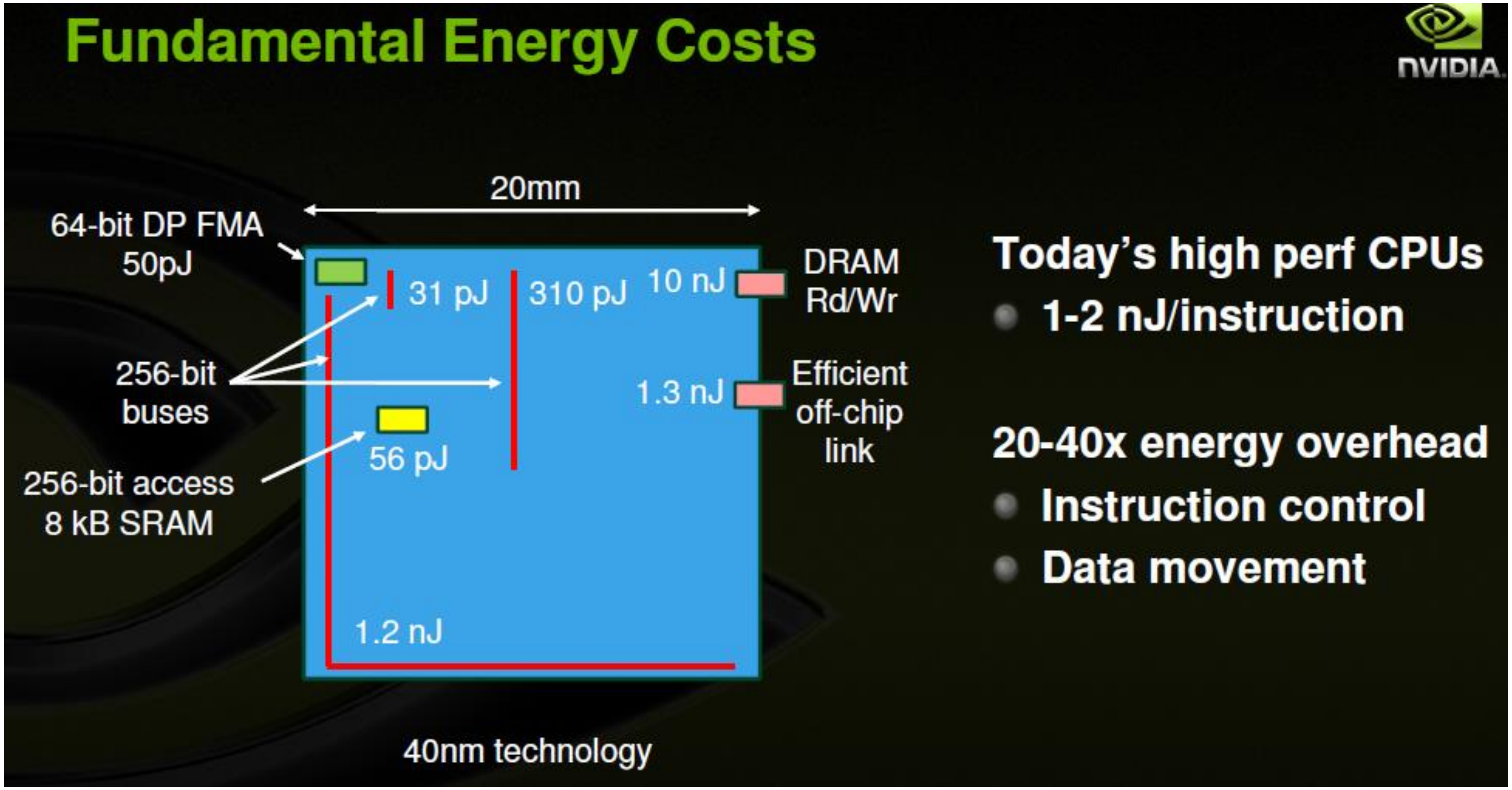
http://www.maltiel-consulting.com/ISSCC-2013-Memory-trends-FLash-NAND-DRAM.html





Why Near-DRAM Acceleration?

- data transfer energy is more expensive than computation
  - ✓ disparity b/w interconnect and transistor scaling



Fundamental Energy Costs

Today's high perf CPUs

- 1-2 nJ/instruction

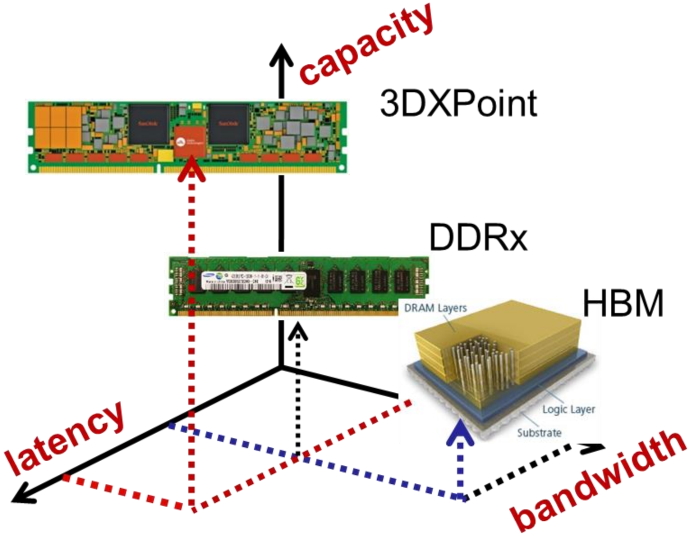
20-40x energy overhead

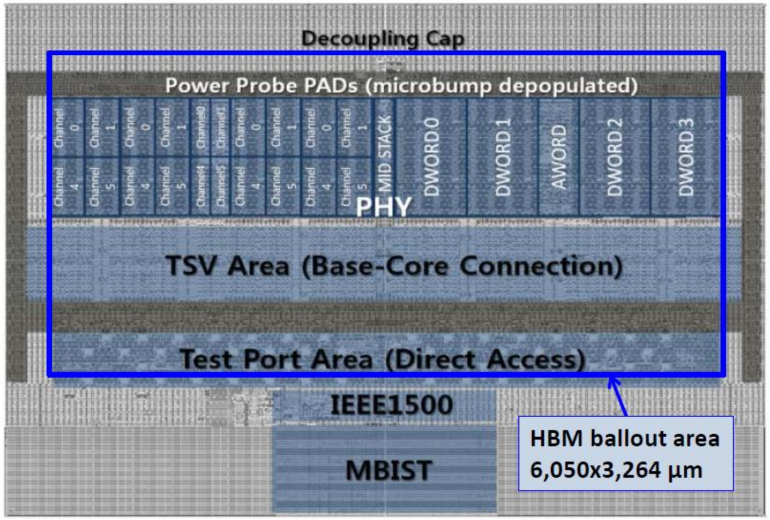
- Instruction control
- Data movement

Keckler MICRO'11 Keynote talk: "Life After Dennard and How I Learned to Love the Picojoule"

3D-stacked Near-DRAM Acceleration

- conventional architectures w/ expensive 3D-stacked DRAM
  - ✓ sacrifice capcity for bandwidth (BW)
    - ○ one memory module per channel w/ point-to-point connection
  - ✓ insufficient logic die space for accelerators (ACCs)
    - ○ little space left for ACCs and/or higher BW for ACCs due to large # of TSVs and PHYs
  - ✓ not flexible after integration of ACCs w/ DRAM
    - ○ custom DRAM module tied w/ specific ACC architecture

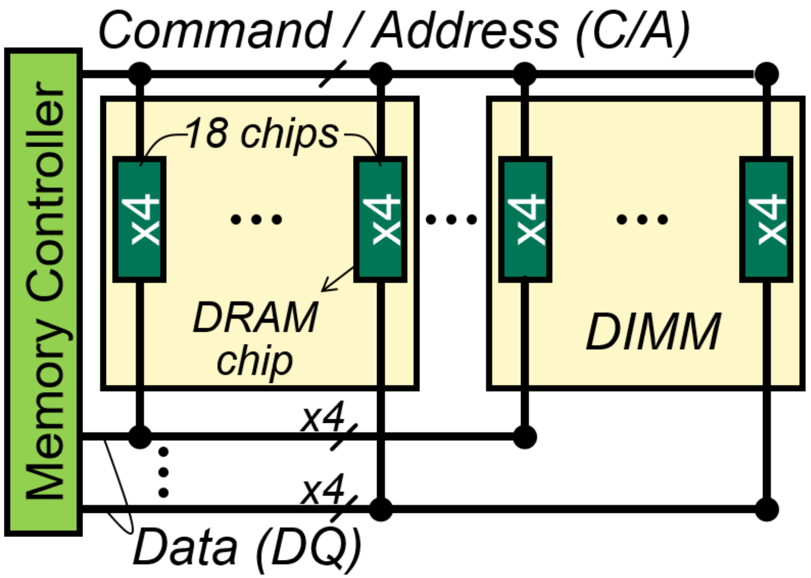



Decoupling Cap
Power Probe PADs (microbump depopulated)
MID STACK
PHY
DWORD 0
DWORD 1
AWORD
DWORD 2
DWORD 3
TSV Area (Base-Core Connection)
Test Port Area (Direct Access)
IEEE1500
MBIST

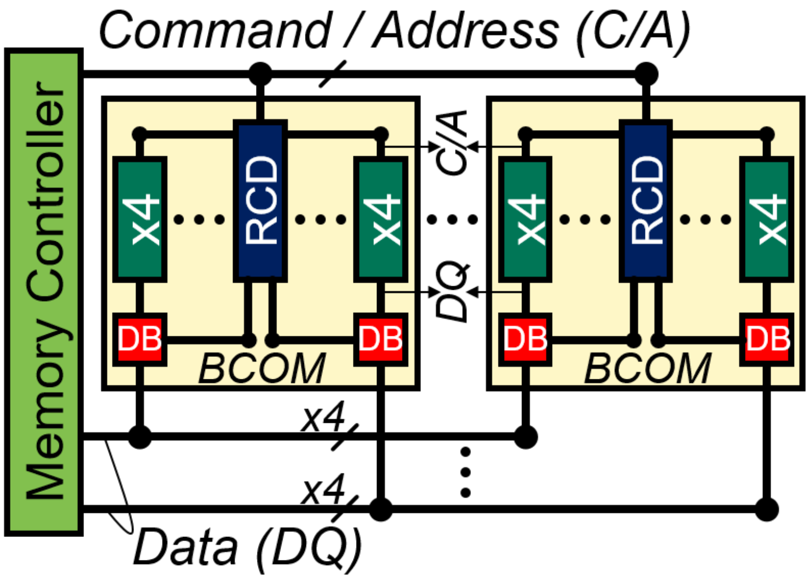
HBM ballout area 6,050x3,264 μm

Background: DDR4 LRDIMM

- higher capacity for big-data servers
  - ✓ 8 LRDIMM ranks per channel w/o degrading data rate
- repeaters for data (DQ) and command/address (C/A) signals
  - ✓ a registering clock driver (RCD) chip to repeat C/A signals
  - ✓ data buffer (DB) chip per DRAM device to repeats DQ signals



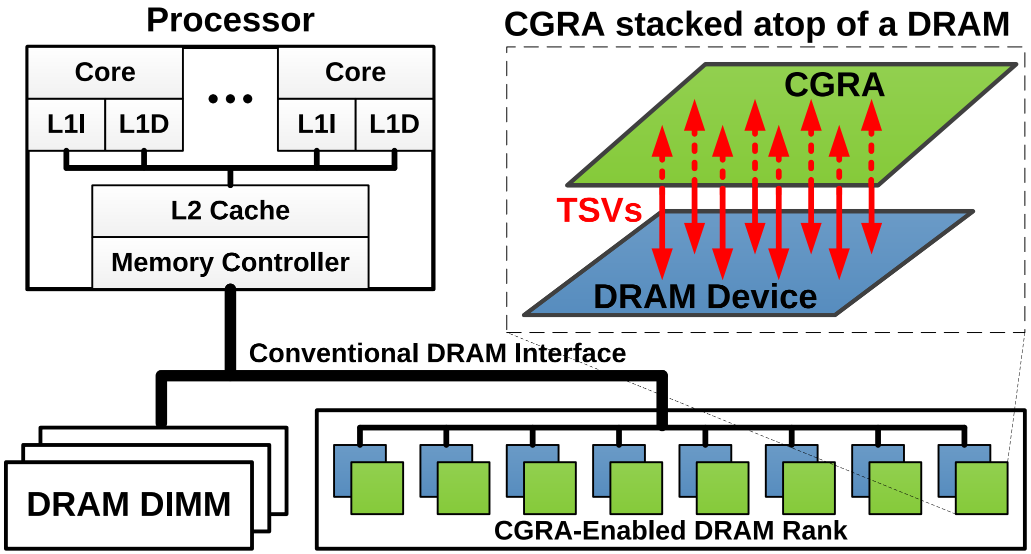
UDIMM



DDR4-LRDIMM

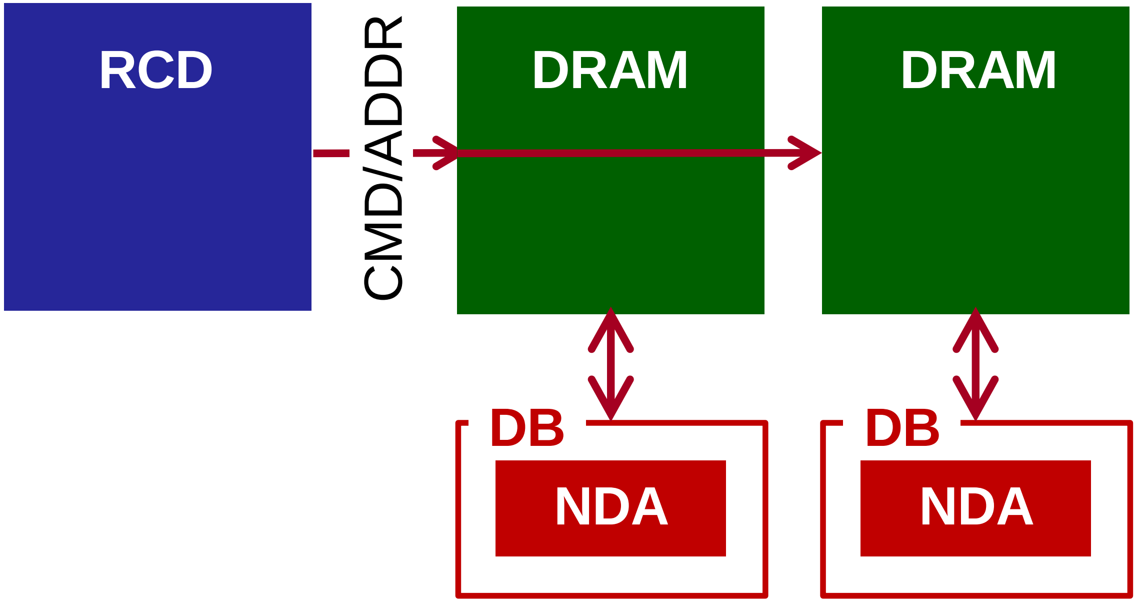
Proposal: In-Buffer Processing[1]

- built upon our previous near-DRAM acceleration architecture
  - ✓ accelerators (e.g., coarse-grain reconfigurable accelerator (CGRA)) 3D-stacked atop commodity DRAM devices
    - ○ Farmahini-Farahani, et al. NDA: Near-DRAM acceleration architecture leveraging commodity DRAM devices and standard memory modules, HPCA 2015
- processor offloads compute- and data-intensive operations (application kernels) onto CGRAs
  - ✓ CGRAs process data locally in their corresponding DRAM



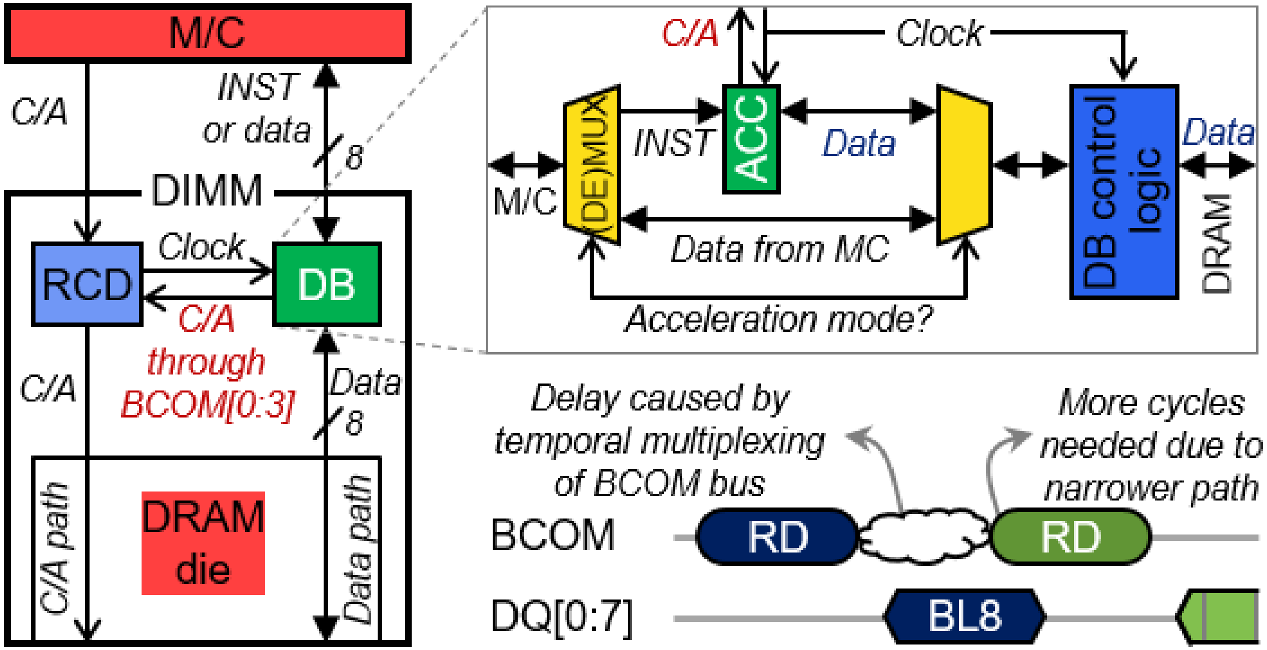
Proposal: In-Buffer Processing^2

- place near-DRAM accelerators (NDA) in buffer chips
  - ✓ require no change to
    - ○ processor
    - ○ processor-DRAM interface
    - ○ DRAM core circuit and architecture
  - ✓ propose three Chameleon microarchitectures
    - ○ Chameleon-*d*, *t* and -*s*



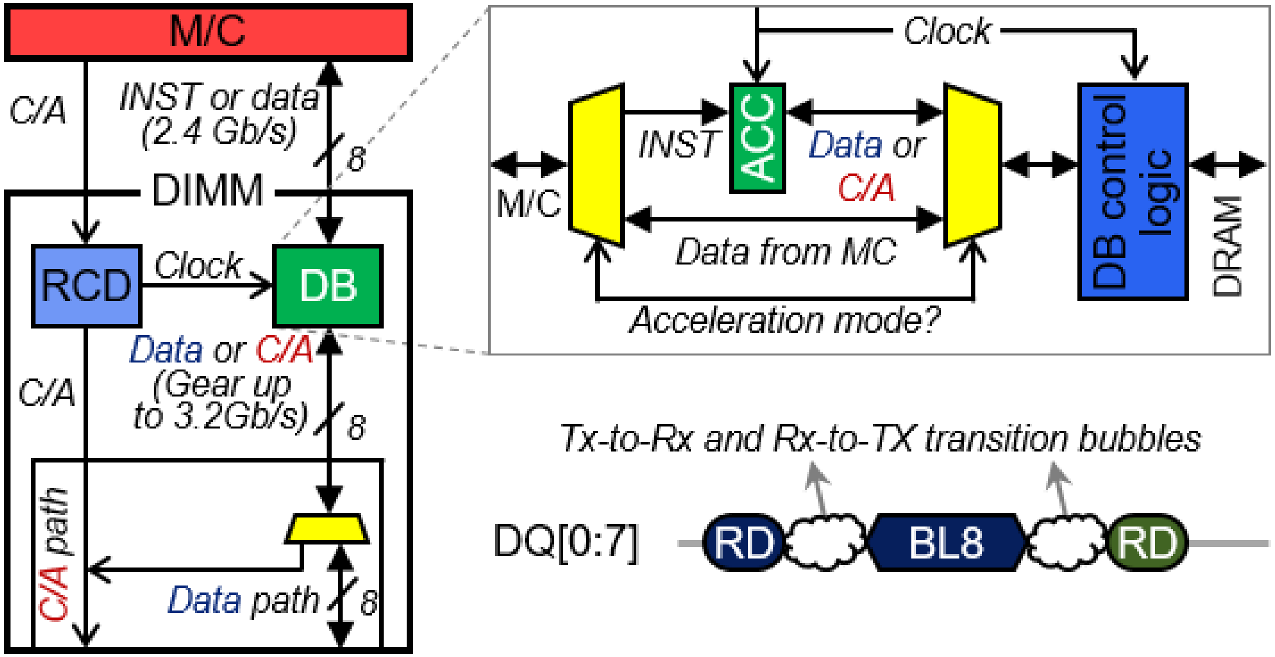
ACC-DRAM Connection: Chameleon-d

- allocate full DQ bus bandwidth to data transfer b/w ACC and DRAM modules vertically aligned in a LRDIMM
  - ✓ 8-bit data bus b/w ACC and DRAM
- connect C/A pins to the RCD through BCOM bus (400MHz)
  - ✓ RCD arbitrates among C/A requests of all ACCs
  - ✓ limited bandwidth of the RCD becomes the bottleneck



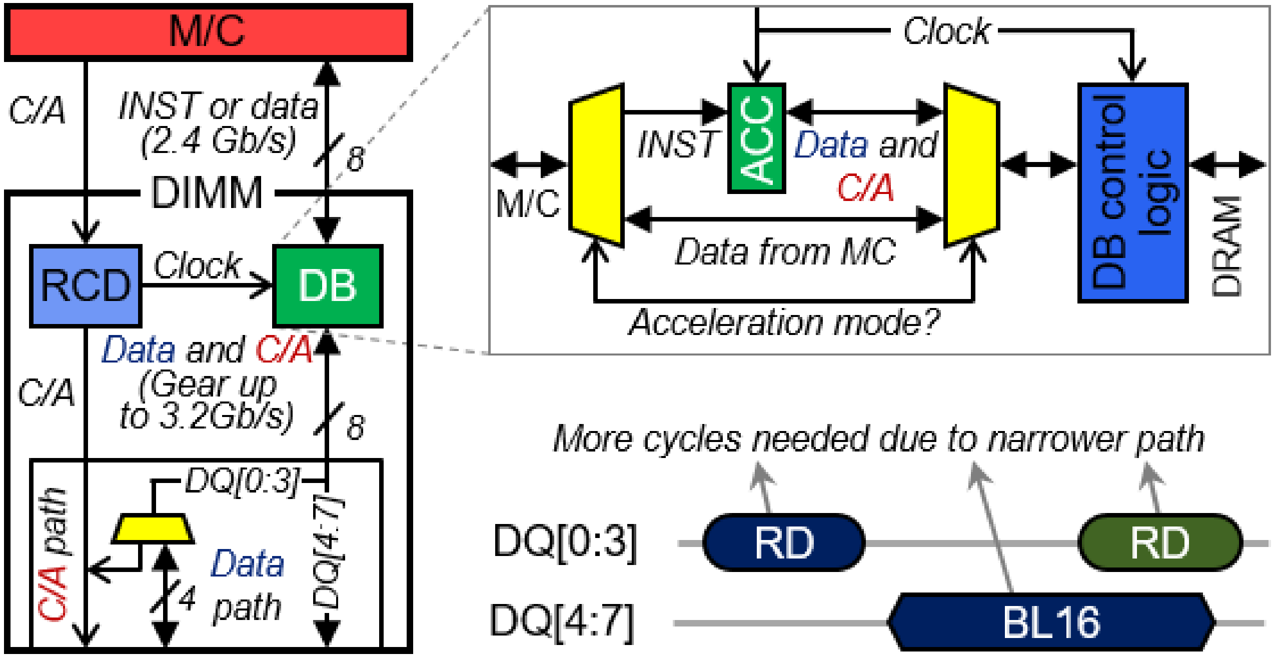
ACC-DRAM Connection: Chameleon-t

- DQ pins are temporally multiplexed b/w DQ and C/A signals
  - ✓ previous DRAM shared I/O pins for C/A and DQ signals
    - ○ e.g., FBDIMM
  - ✓ 1tCK, 1tCK, 2tCK for activate, pre-charge, and read/write commands, respectively
  - ✓ cons: a bubble cycle required for every read operation



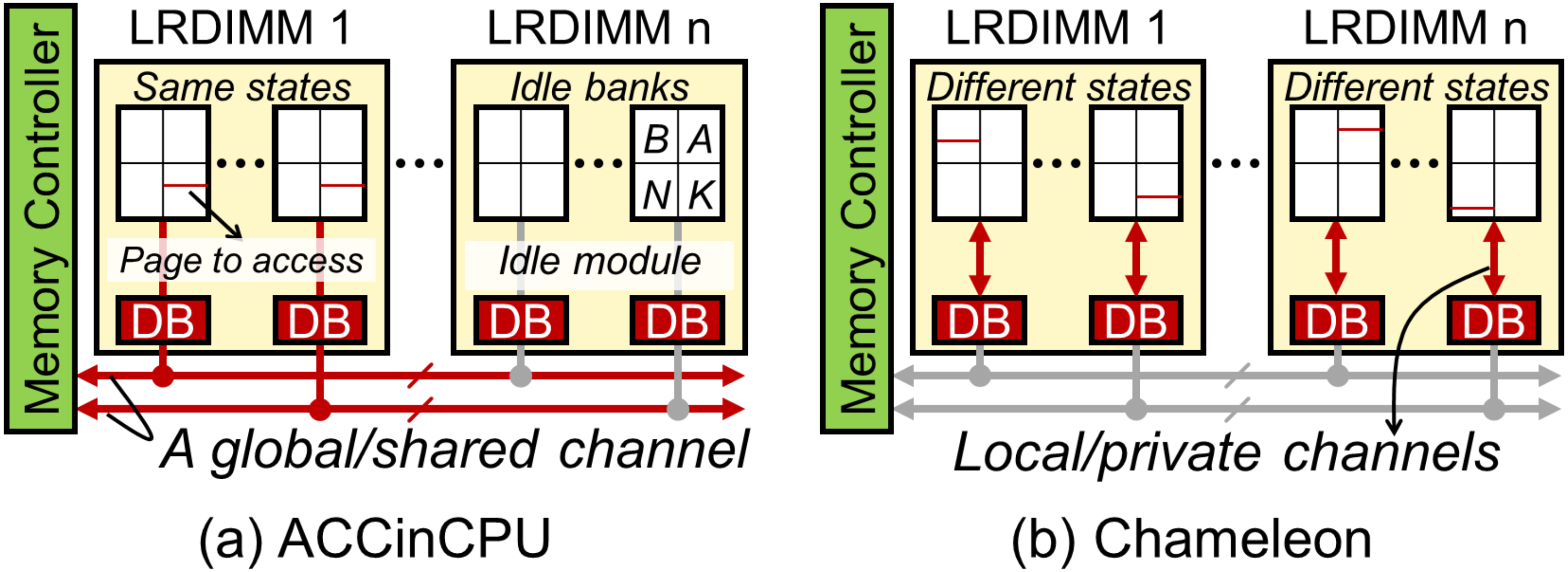
ACC-DRAM Connection: Chameleon-s

- DQ pins are spatially multiplexed b/w DQ and C/A signals
  - ✓ pros: avoids bubble for bus direction changes for every read trans.
  - ✓ cons: burst length increased from 8 to 16 if 4 out of 8 lines are used for data transfer



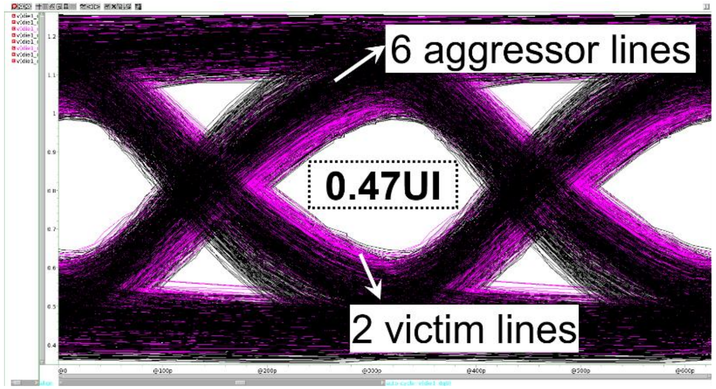
Transcending Limitation of DIMMs

- no change to standard DRAM devices and DIMMs
  - ✓ no BW benefit w/ the same bandwidth as traditional DIMMs?
- in NDA mode
  - ✓ DRAM devices coupled w/ accelerators can be electrically disconnected from global/shared memory channel
    - ○ short point-to-point local/private connections b/w DRAM and DB devices

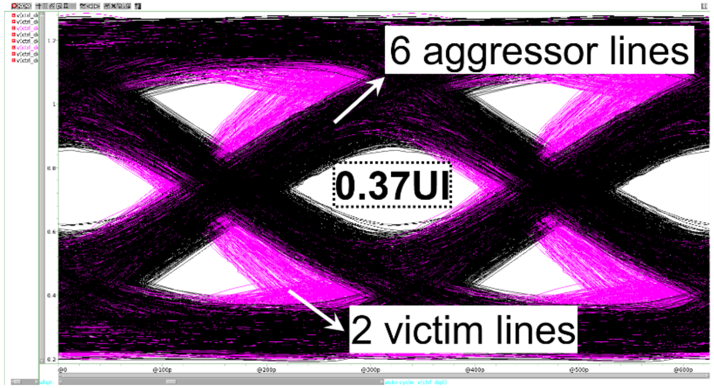


Gear-up Mode

- short-distance point-to-point local/private connections allows
  - ✓ higher I/O data rate w/ better channel quality b/w DB and DRAM device (from 2.4GT/s to 3.2GT/s)
    - ○ DRAM device clock is remains intact



DB to DRAM (Tx) at 3.2GHz



DRAM to DB (Rx) at 3.2GHz

  - ✓ scaling aggregate bandwidth w/ more DIMMs
    - ○ ACCs concurrently accessing coupled DRAM devices across multiple DIMMs

compensating the bandwidth and timing penalty incurred by Chameleon-s and Chameleon-t

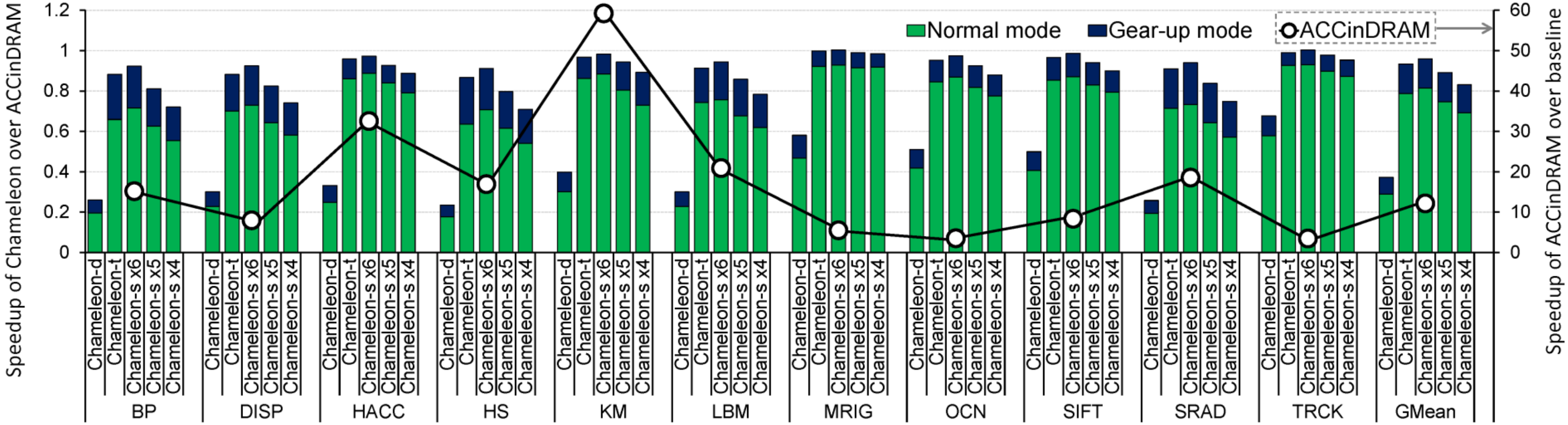
Evaluated Architectures

Architecture	# of ACCs	Description
Baseline	-	4-way OoO processor at 2GHz
ACCinCPU	32	32 on-chip CGRAs co-located with the processor
ACCinDRAM	32	4 CGRAs stacked atop each DRAM [HPCA'2015]
Chameleon	32	4 CGRAs in each DB device

- accelerator
  - ✓ coarse-grain reconfigurable accelerator (CGRA) w/ 64 FUs
- LRDIMM w/ DDR4-2400 ×8 DRAM devices
- area of CGRA w/ local memory controller
  - ✓ ~0.832 mm2 for 64-FU CGRA + ~0.21 mm2 for MC, fitting in a DB device
- benchmarks
  - ✓ the same ones used in ``NDA'' in HPCA'2015

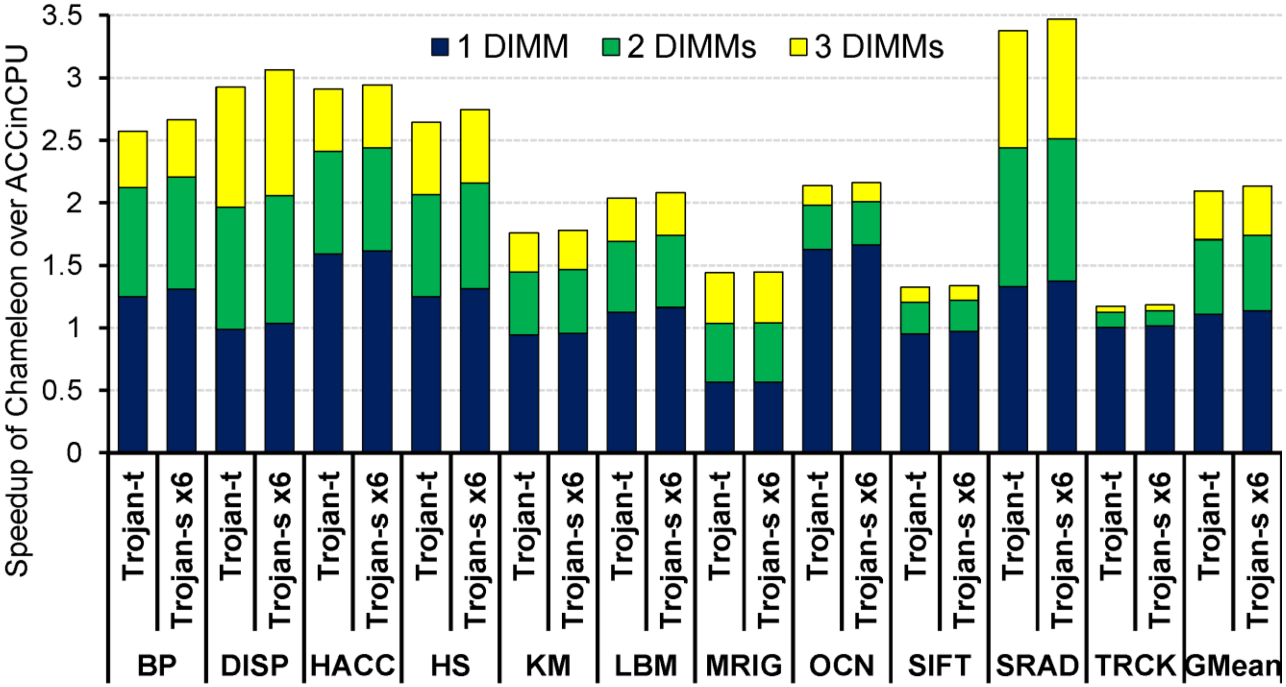
Speedup

- Chameleon-*s* & -*t* offer competitive performance compared to ACCinDRAM relying on 3D-stacking ACCs atop DRAM
  - ✓ Chameleon-s x6 (6 and 2 pins for data and command/address)
    - ○ 96% performance of ACCinDRAM w/ gear-up mode
    - ○ 3% better than Chameleon-*t* w/ no bubble for every read
    - ○ 9%/17% higher performance than Chameleon-s x5/x4



Speedup

- Chameleon architectures scale w/ # of LRDIMMs
  - ✓ ACCinCPU performance marginally varies w/ # of ACCs
  - ✓ each Chameleon LRDIMM operates independently
    - ○ for 1, 2, and 3 LRDIMMs , Chameleon-*s* x6 performs 14%, 74%, and 113% better than ACCinCPU, respectively



Conclusions

- Chameleon: practical, versatile near-DRAM acceleration architecture
  - ✓ propose in-buffer-processing architecture, placing accelerators in DB devices coupled w/ commodity DRAM devices
  - ✓ require no change to processor, processor-DRAM interface, and DRAM core circuit and architecture
  - ✓ achieve 96% performance of (expensive 3D-stacking-based) NDA architecture [HPCA'2015]
  - ✓ improve performance by 14%, 74%, and 113% for 1, 2, and 3 LRDIMMs compared w/ ACCinCPU
  - ✓ reduce energy by 30% compared w/ ACCinCPU