Processor and DRAM Integration by TSV-Based 3-D Stacking for Power-Aware SOCs

Shin-Shiun Chen, Chun-Kai Hsu, Hsiu-Chuan Shih, and Cheng-Wen Wu
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan

Jen-Chieh Yeh
Information and Communications Research Lab
Industrial Technology Research Institute
Hsinchu, Taiwan

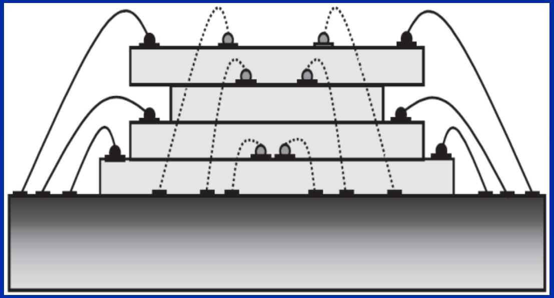
Outline

- Introduction & Motivation
- 3D-PAC Architecture
- Sans-Cache DRAM Architecture
- Experimental Result
- Conclusion

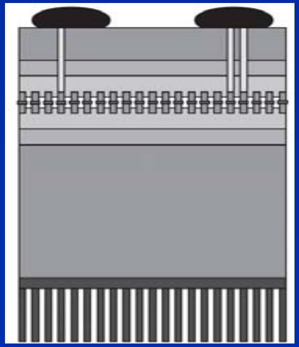
3-D Stacking Technology

- Wire-bonding technique
  - For System-In-Package (SIP)
  - Pin-count limitation
- Micro-bump technique
  - Commonly used in face-to-face style stacking
- Through Silicon Via (TSV)
  - Etch through the bulk layer

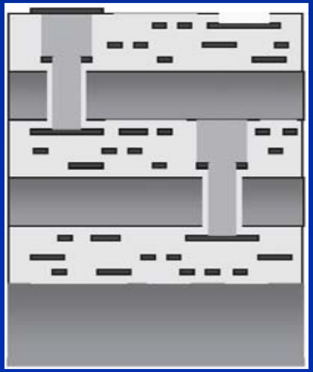
wire bonding



micro-bump



TSV



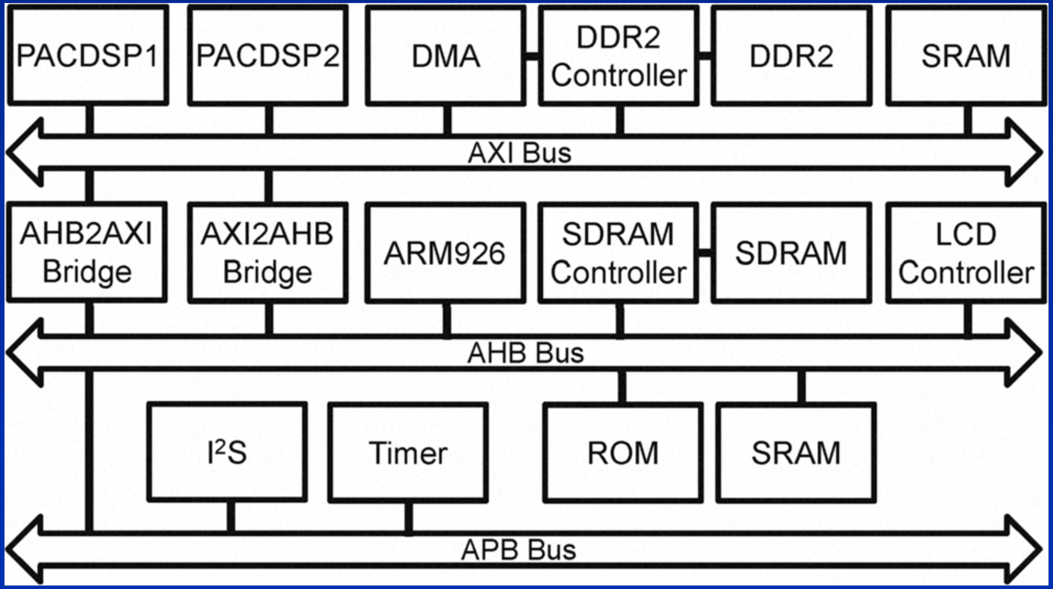
Ref: W. Rhett Davis et al. , “Demystifying 3D ICs: The Pros and Cons of Going Vertical”

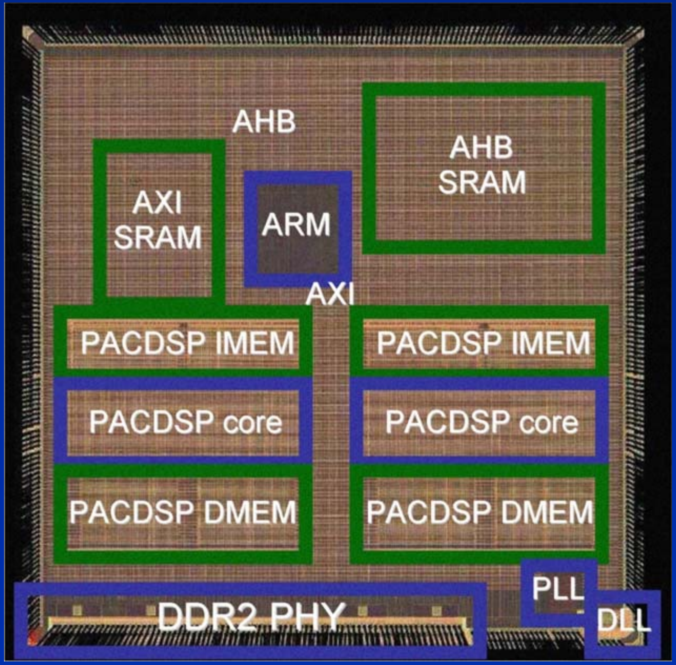
Motivation

- Memory wall problem
  - Speed mismatch between processor and memory
  - Board level signal routing
  - Additional energy for complicated memory hierarchy
- Emerging 3-D stacking technology
  - Support heterogeneous integration
  - Reduce latency, power and energy
  - Mitigate the memory wall
- A new system design
  - Greatly reduce the system power and energy
  - While maintain the system performance

PAC DSP SOC & Virtual Platform

- 32-bit DSP developed by ITRI
  - With ARM9 CPU
  - Target on low power and energy
- Cycle-accurate model
  - 99% accuracy compared to measured data

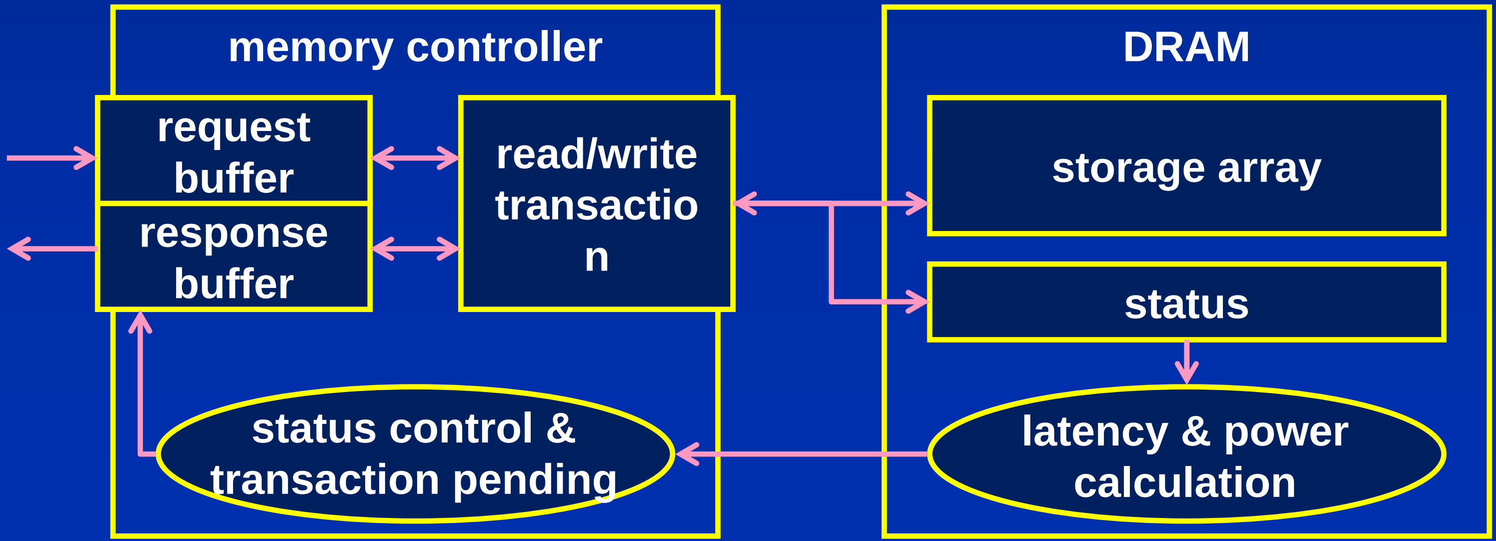




Ref: I-Yao Chung et al., “PAC Duo SOC Performance Analysis with ESL Design Methodology”

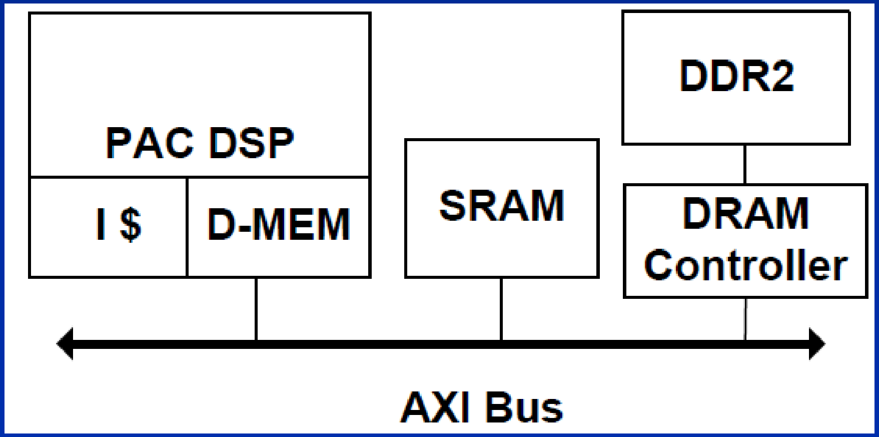
Memory System Model

- Based on DRAMSim2 library
  - Read/write behavior
    - OSCI PV (Programmer's View) level
    - DRAM status control
    - For the correctness of functionality
  - Latency and power calculation
    - According to the DRAM status
    - Add latency into response

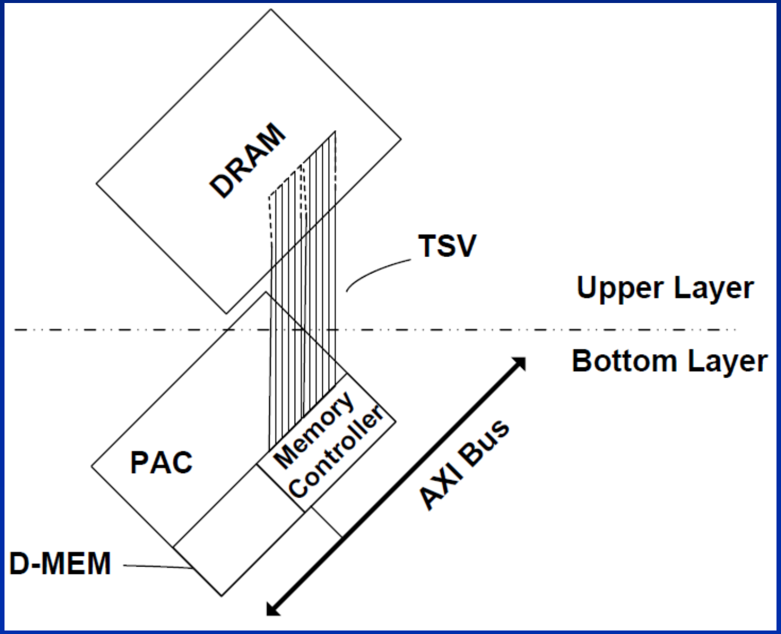


3D-PAC Architecture

- Stack DRAM on the top of PAC DSP
  - No instruction cache and AXI SRAM
  - Without bus width limitation of AXI
  - No AXI bus delay
  - High data width



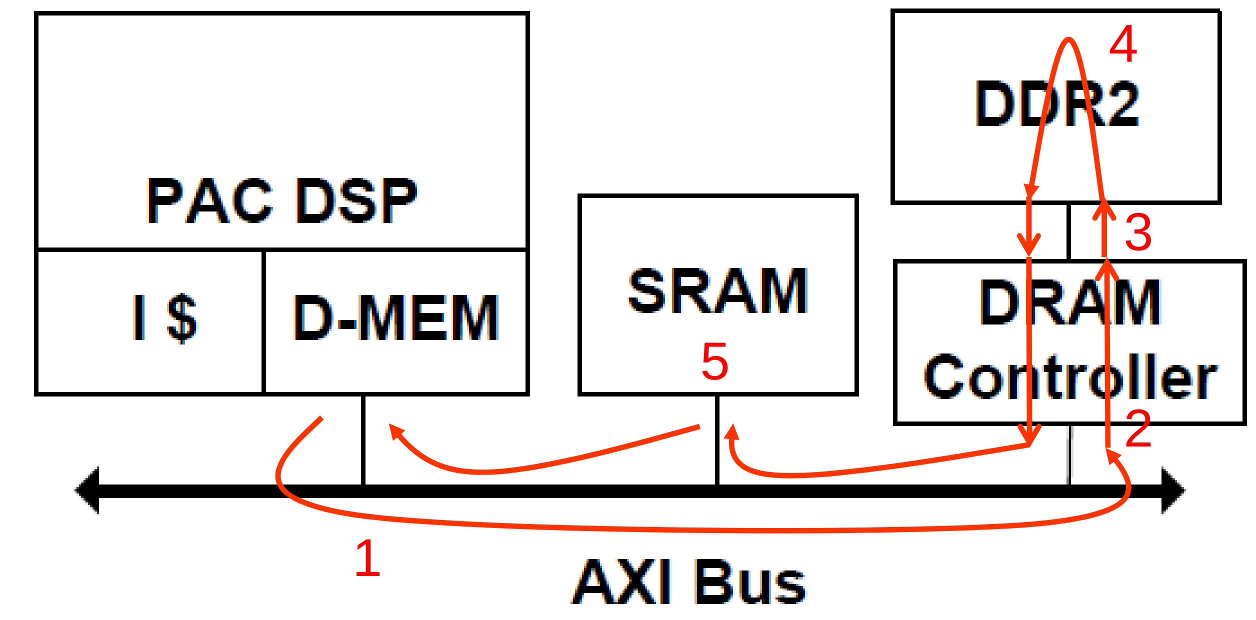
2D PAC



3D PAC

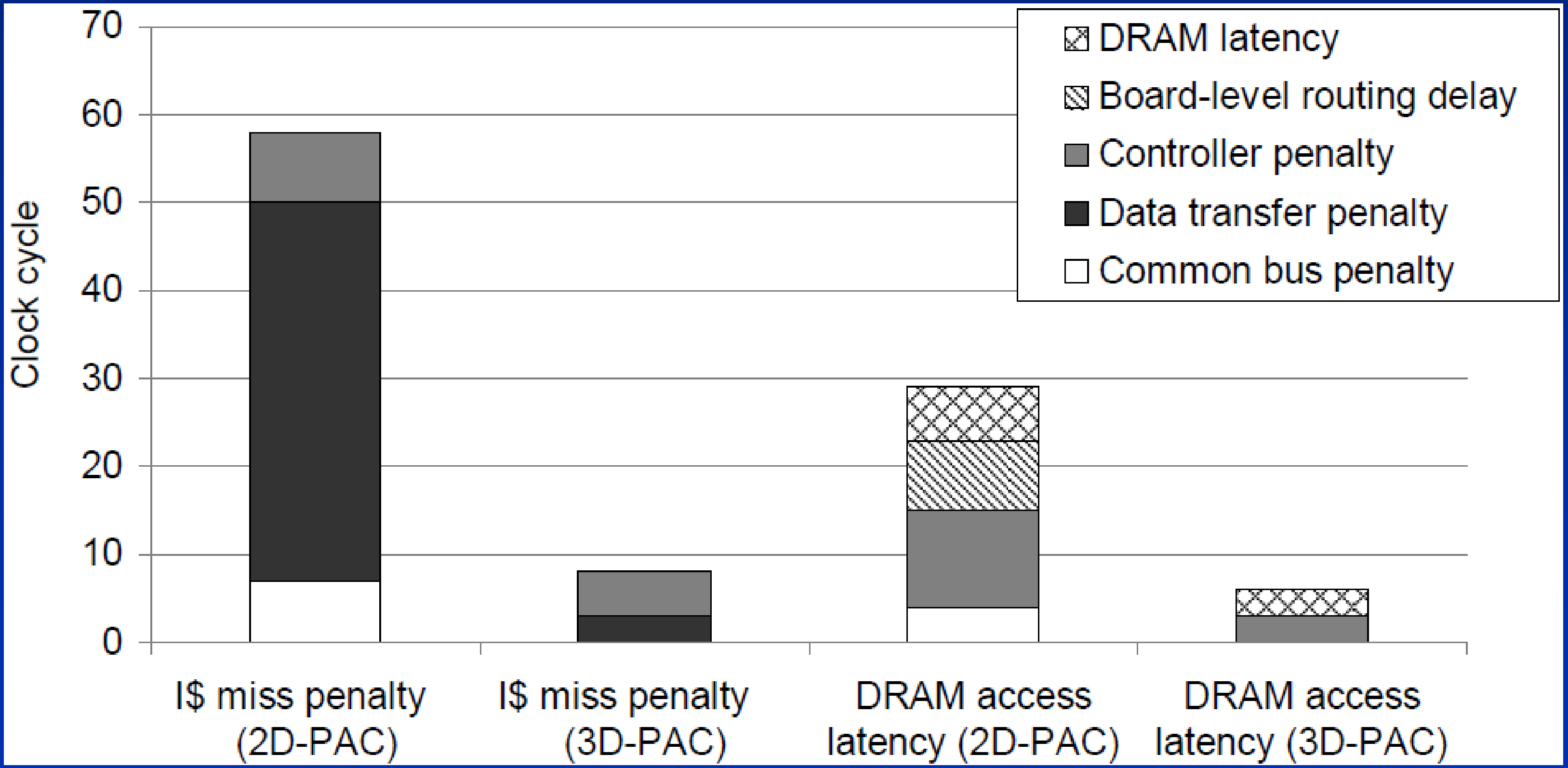
Memory System Analysis (1/2)

- Breakdown of cache miss latency
  - 1. Common bus penalty
  - 2. Controller penalty
  - 3. Board level routing delay
  - 4. DRAM latency
  - 5. Data transfer penalty



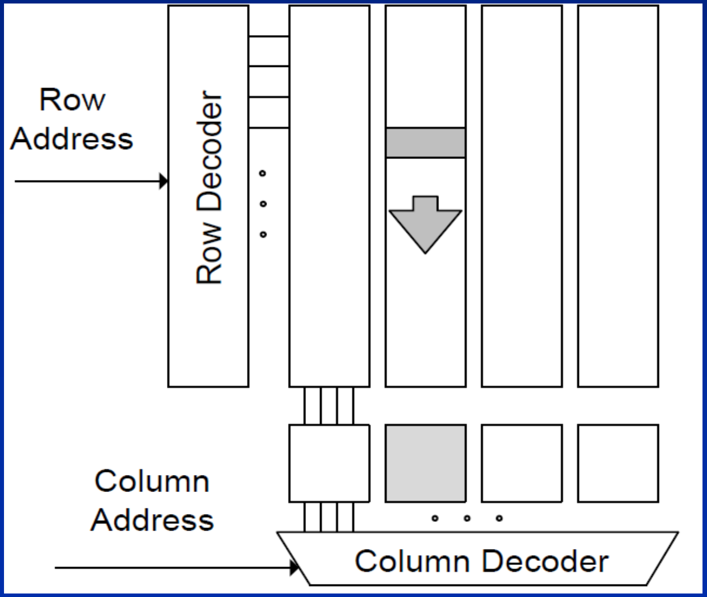
Memory System Analysis (2/2)

- Comparison between 2D-PAC and 3D-PAC
  - 3D-PAC greatly improves the latency
- Latency improvement → power/energy reduction



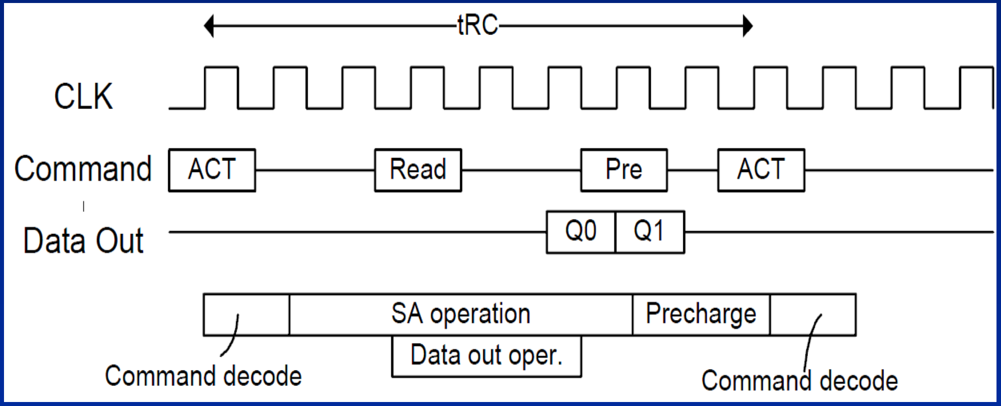
Fast-Cycle RAM (FCRAM)

- A reference DRAM design
  - Small active row
  - Auto-precharge circuit
  - Pipeline scheme

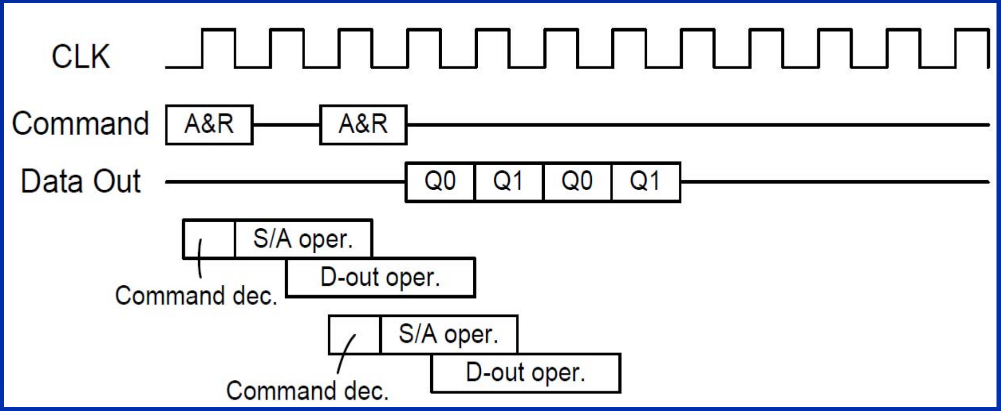


Ref: Y. *Sato et al*., “Fast Cycle RAM (FCRAM): a 20-ns random row access, pipelined operating DRAM”

Commodity DRAM



FCRAM

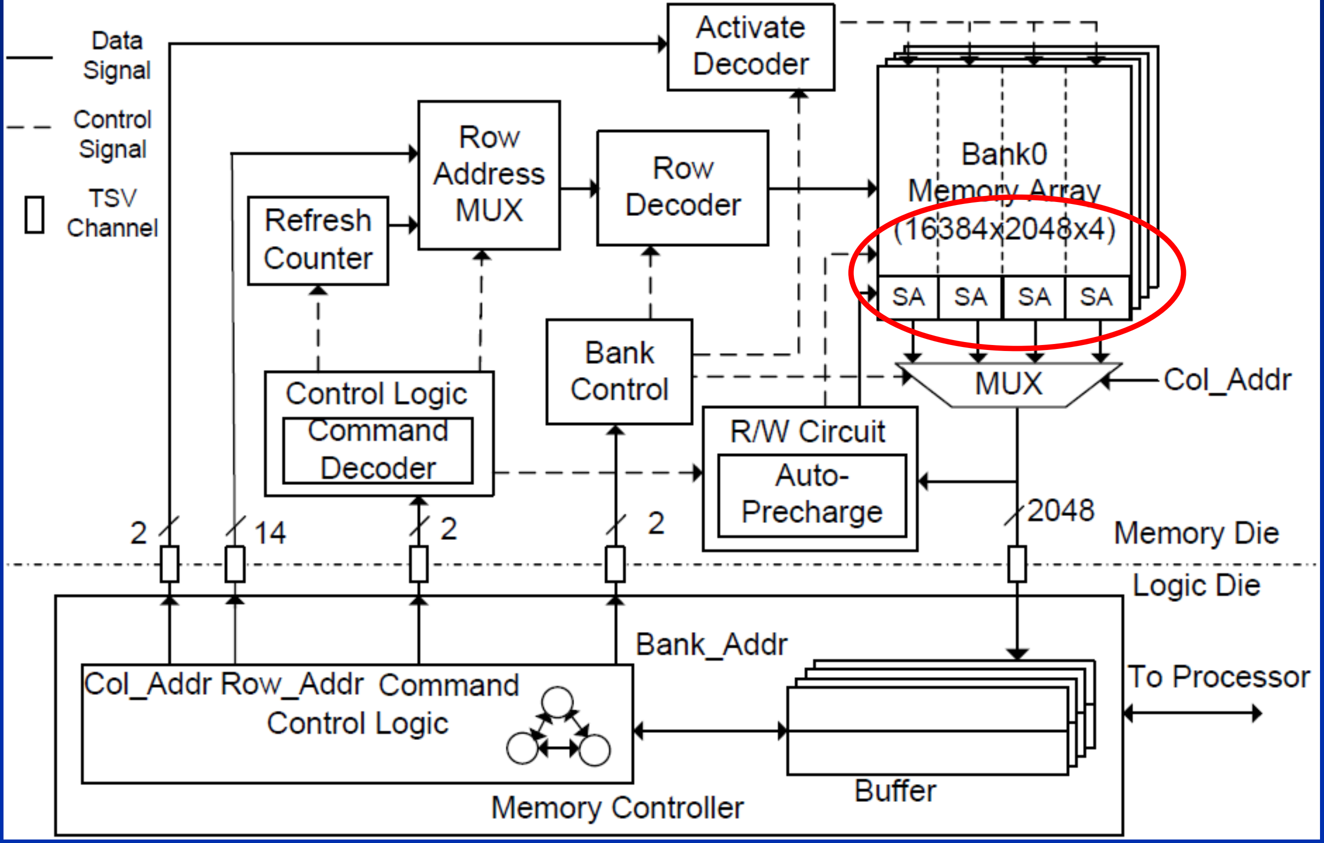


Sans-Cache DRAM (SCDRAM)

- A new DRAM interface
  - Simplify the memory hierarchy
  - Latency improvement → power/energy reduction
- Features of SCDRAM
  - Send row and column address simultaneously
  - Reduce the active row size to 2K bits
  - Auto-precharge circuit
  - Wide I/O design with ping-pong buffer
  - Without data burst and Double Data Rate (DDR)

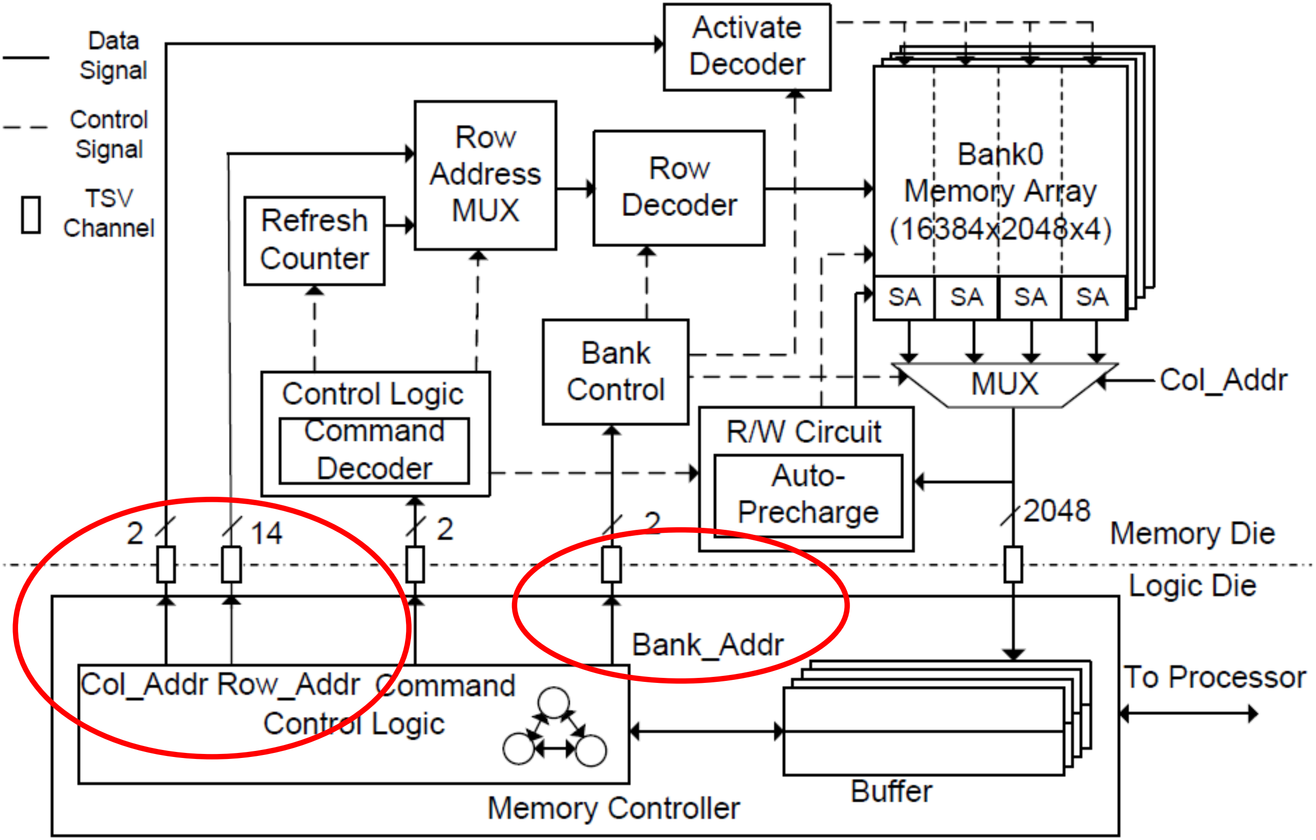
Small Page Size

- Reduce the active row size to 2K bits
  - Reduce power and access latency
  - Improve energy efficiency



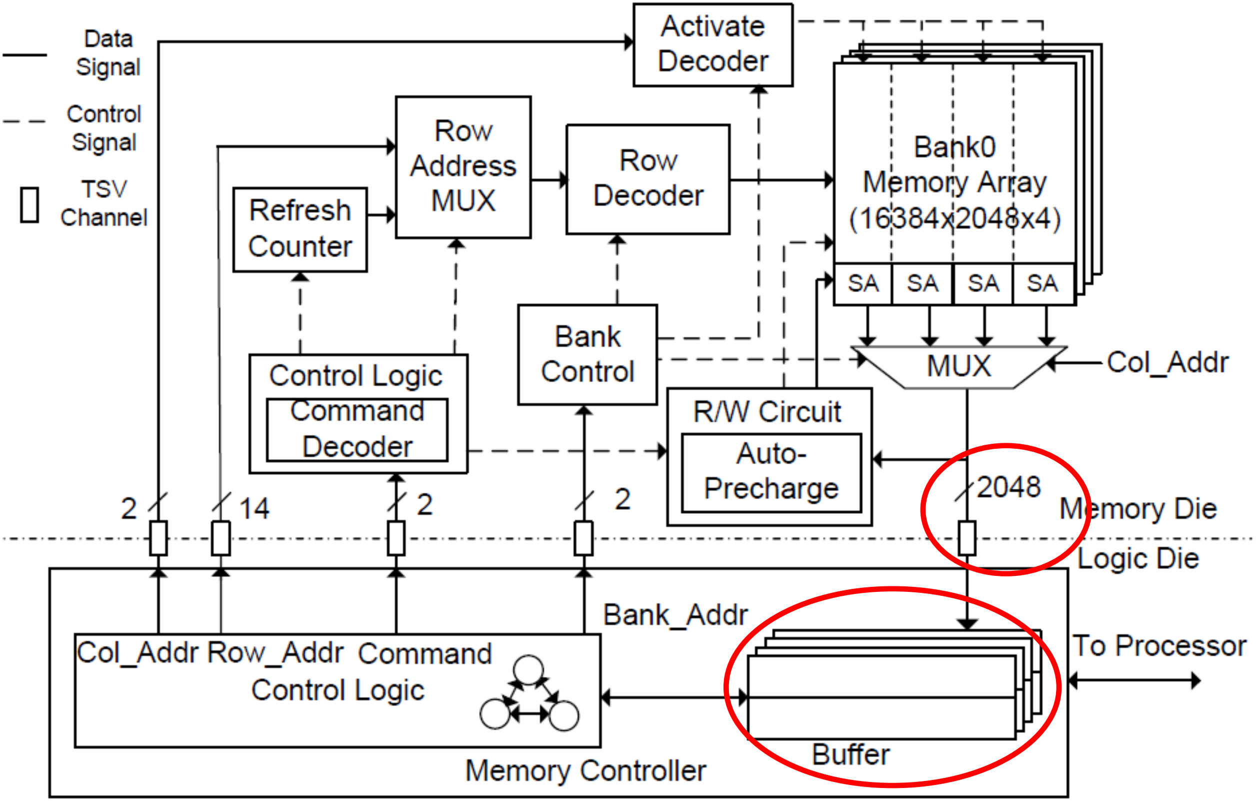
Address Interface

- Eliminate address multiplexing
  - Send row and column address simultaneously



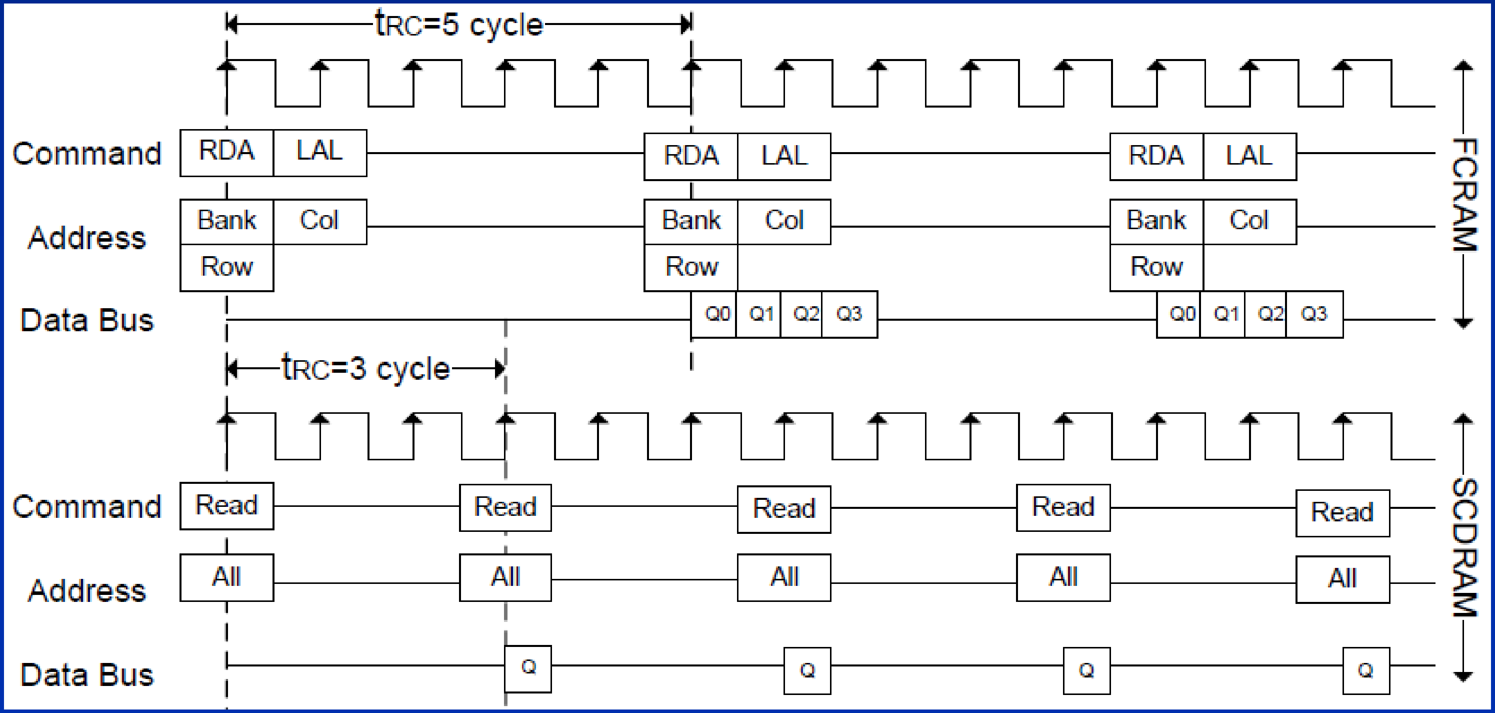
Wide I/O with Ping-Pong Buffer

- Simplify cache to a ping-pong buffer
  - Each bank has 512B size of buffer
- I/O bus width is the same to the row size
  - Remove data burst operation



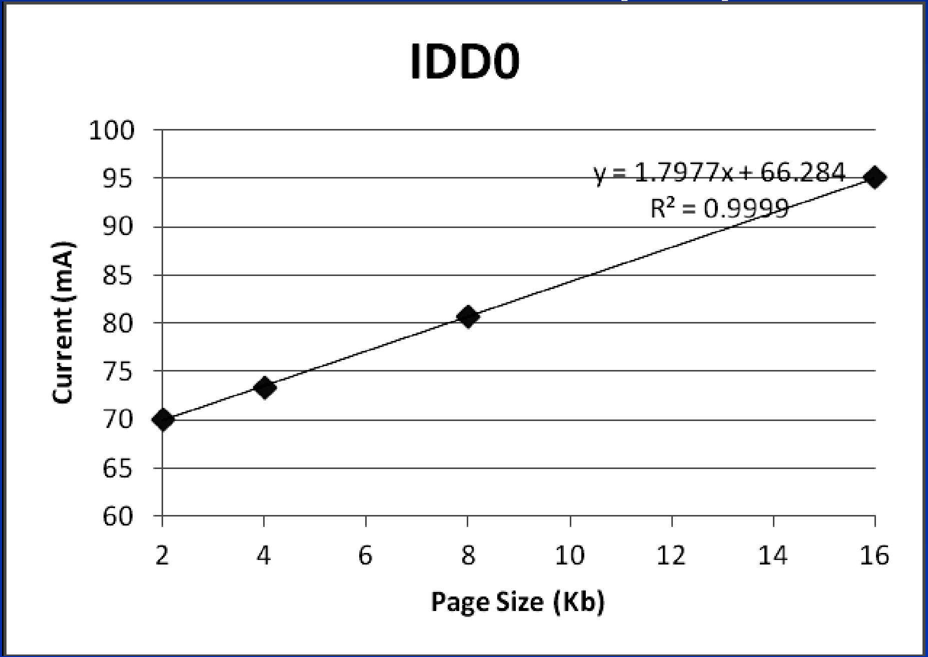
Timing Model of SCDRAM

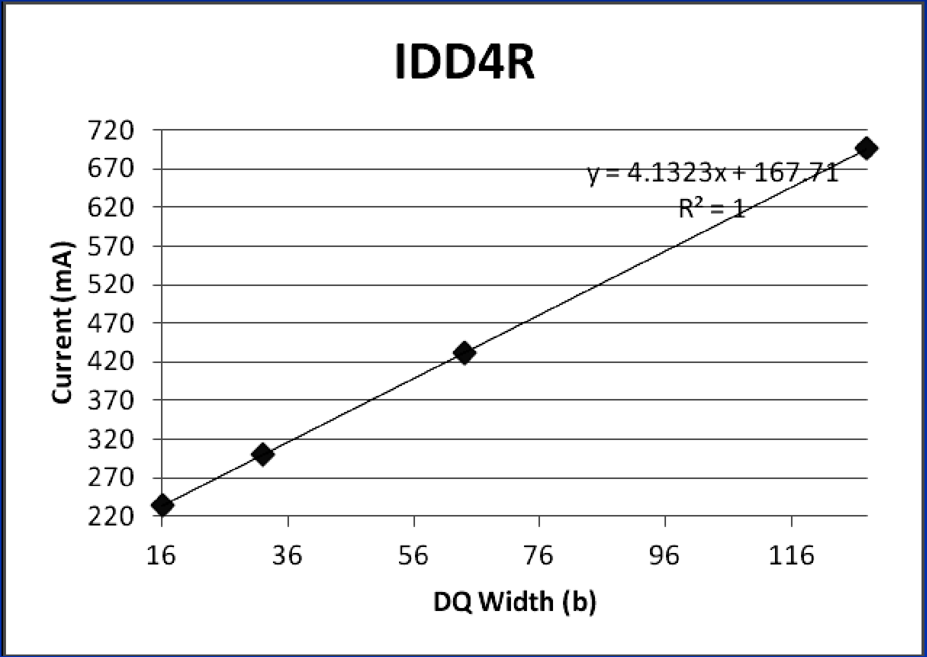
- Predicted from FCRAM
  - – Eliminate address multiplexing (1 cycle)
  - – Reduce the column decoder level (1 cycle)
  - 1 cycle to enter/exit power-down mode
- The same refresh period as FCRAM
  - – Parallel refresh (power constraint)



Observation of DRAM Power

- Important current parameters in standard
  - IDD0: the activation and precharge current
  - IDD4R/W: the burst read/write current
- Based on RAMBus power model
  - IDD0 is proportional to the page size
  - IDD4R/W is proportional to the I/O width





Power Model of SCDRAM

- Based on the equation from Micron spreadsheet

$$I_{act} = IDD0 - \frac{IDD3N * tRAS + IDD2N * (tRC - tRAS)}{tRC}$$

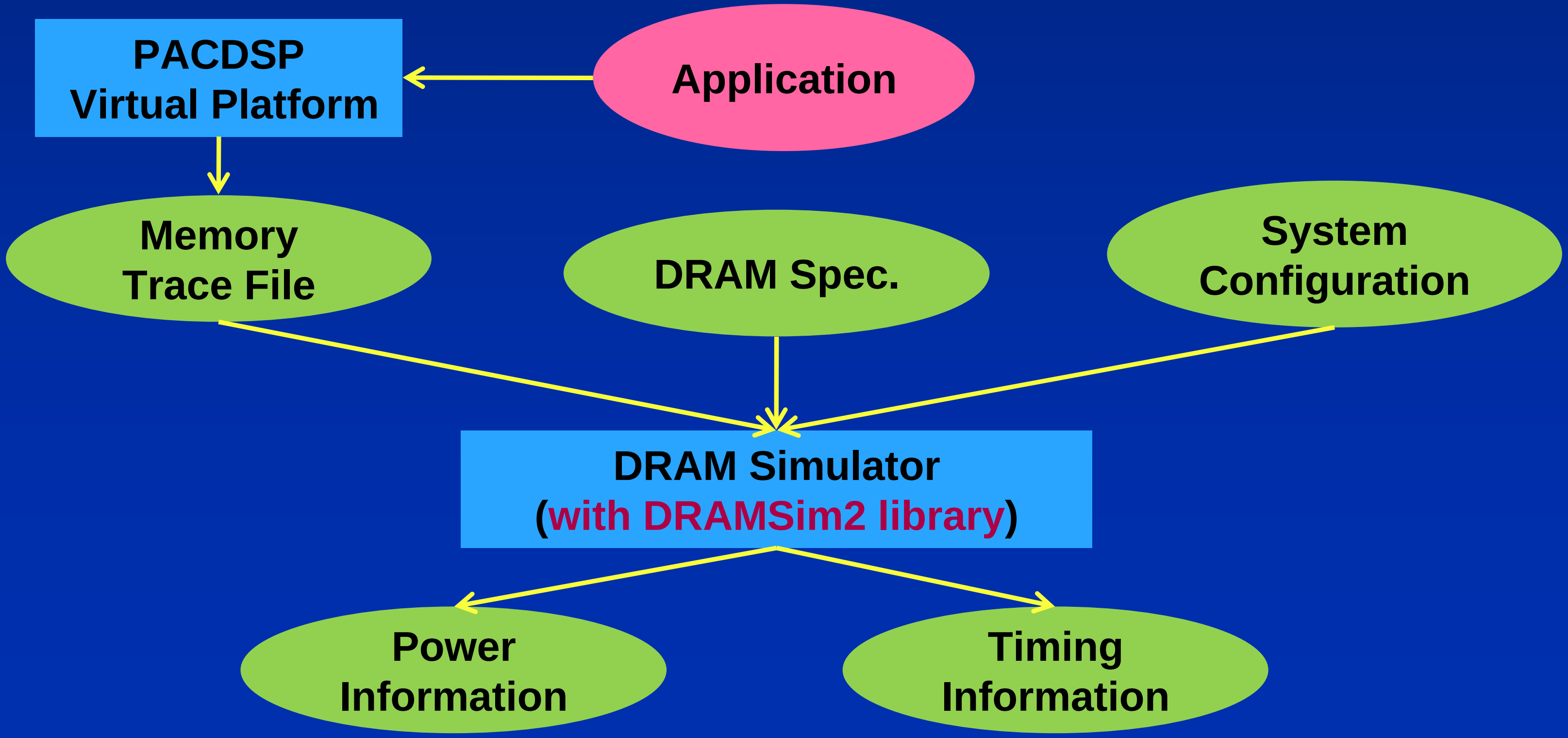
$$I_{read} = IDD4R - IDD3N$$

$$I_{write} = IDD4W - IDD3N$$

- Timing model (tRAS, tRC)
- Assume the same static current (IDD2N, IDD3N)
- Extrapolate the IDD0 and IDD4R/W
  - From the Micron DDR2 datasheet
  - IDD0 is proportional to the page size
  - IDD4R/W is proportional to the I/O width

Experimental Environment

- Power of cache and SRAM from CACTI6.5
- SCDRAM model
  - Developed with DRAMSim2 library
- The 3D-PAC virtual platform

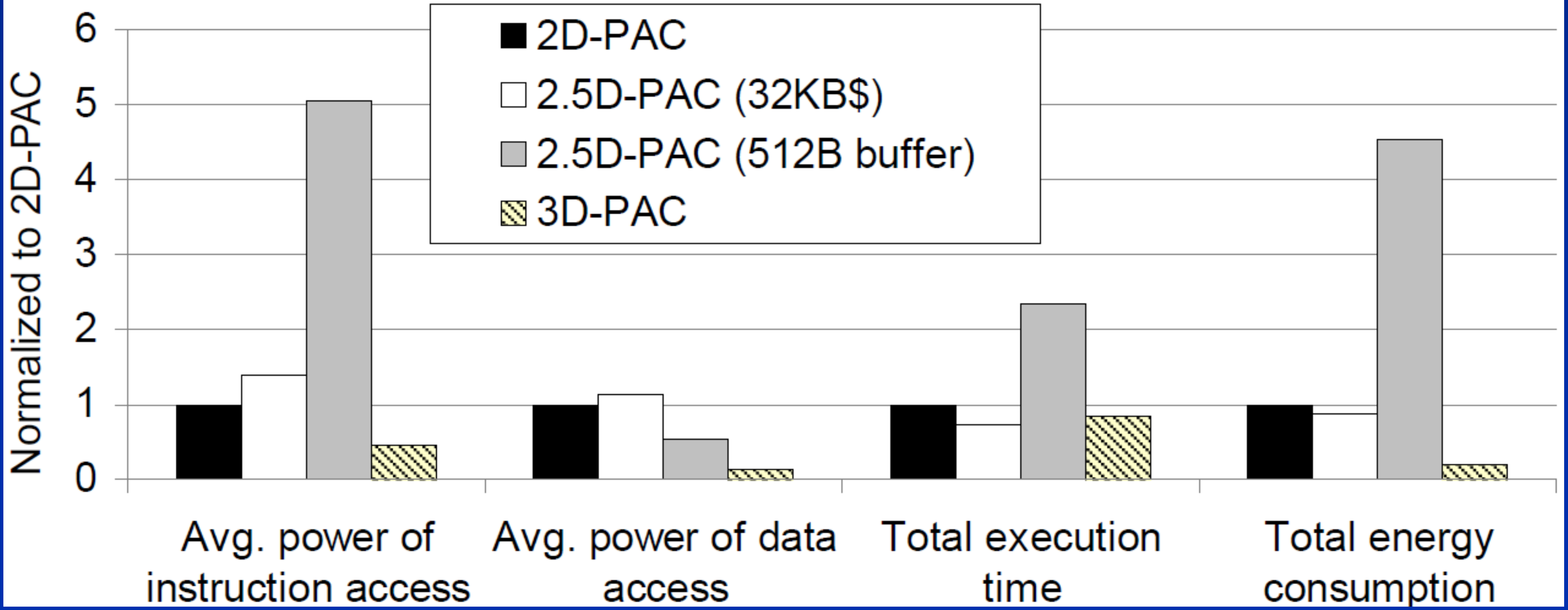


Architecture Comparison (1/2)

- Four cases of architecture
  - 2D-PAC
    - Board level interconnect with DDR2
  - 2.5D-PAC
    - Stack DDR2 by interposer
  - 2.5D-PAC with 512B ping-pong buffer
    - Stack DDR2 by interposer
    - Simplify instruction cache to ping-pong buffer
  - 3D-PAC
    - Stack SCDRAM by TSV
    - Simplify instruction cache to ping-pong buffer
- Benchmark
  - H.264 decoder with QVGA bitstream

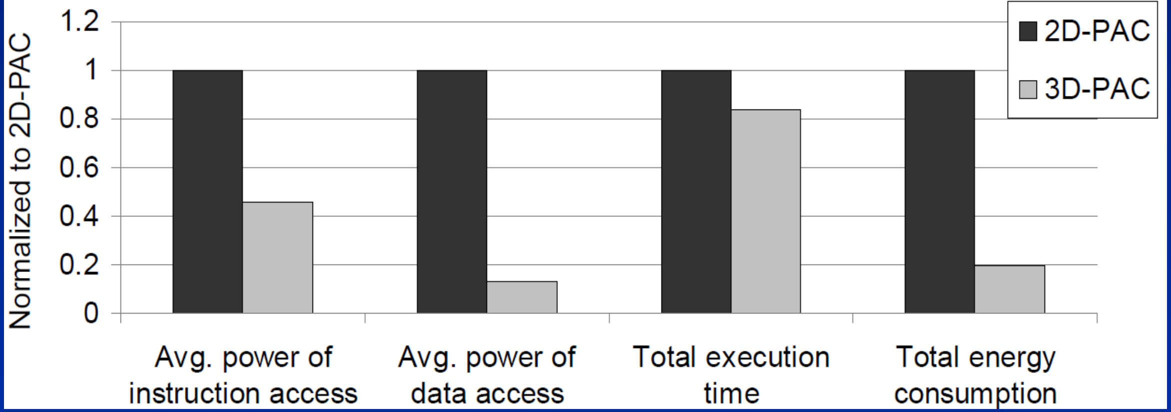
Architecture Comparison (2/2)

- The architecture effect
  - Change board level to Interposer
    - Slightly reduce the execution time and energy
  - Replace cache by ping-pong buffer
    - Significant degradation due to rise of miss rate
  - Stack SCDRAM with TSV
    - Greatly improve the power and energy

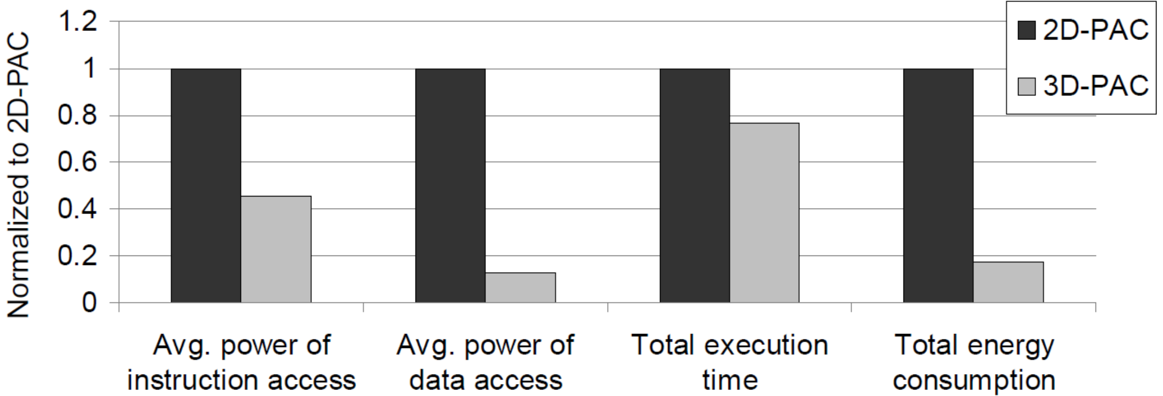


Bitstream Comparison

- High-motion-rate vs. low-motion-rate
  - 80% energy reduction in both cases
  - 16% performance improvement on high-motion
  - 23.5% performance improvement on low-motion



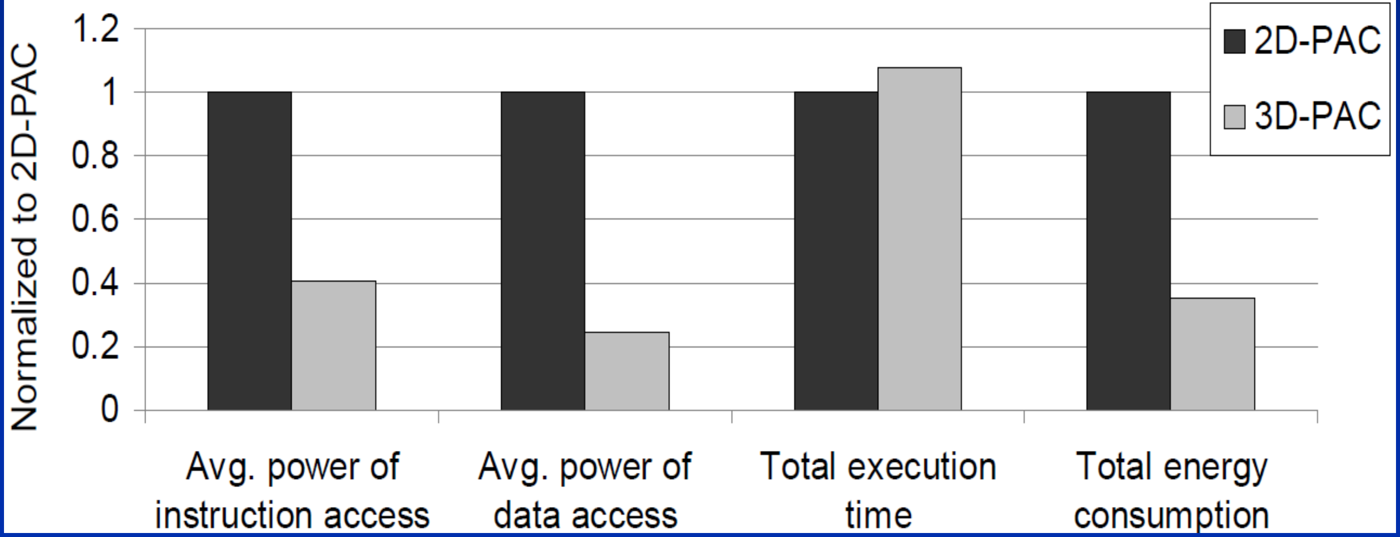
high-motion-rate



low-motion-rate

H.264 Encoder

- Also has 65% energy reduction
- Performance degradation
  - Computing intensive & high spatial data locality
  - The limitation of ping-pong buffer
    - Solved by increasing the buffer size



Conclusion

- A new memory hierarchy for PAC DSP SOC
  - DSP and DRAM integration by TSV stacking
  - Simplify cache to a ping-pong buffer
  - Proposed SCDRAM interface
- Greatly reduce the system power and energy, While maintain the system performance
  - For memory-intensive applications
    - 80% of energy reduction
    - With 23% of performance improvement
  - For computing-intensive applications
    - 65% of energy reduction
    - Slight performance degradation