3D Systems with On-Chip DRAM for Enabling Low-Power High-Performance Computing

Jie Meng, Daniel Rossell, and **Ayse K. Coskun**

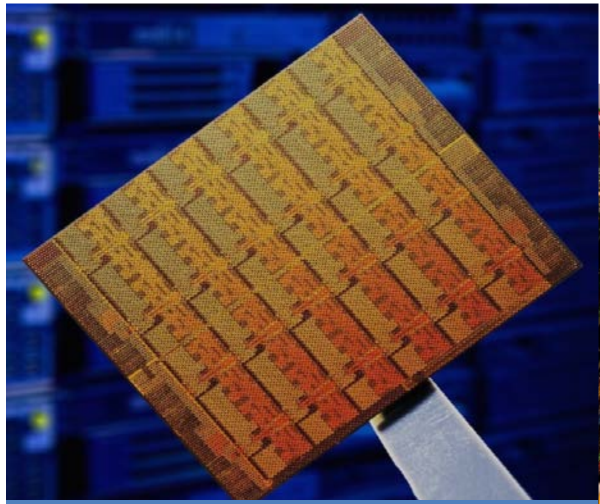
Performance and Energy Aware Computing Lab (PEAC-Lab)

Electrical and Computer Engineering Department

Boston University

HPEC'11 – September 22, 2011

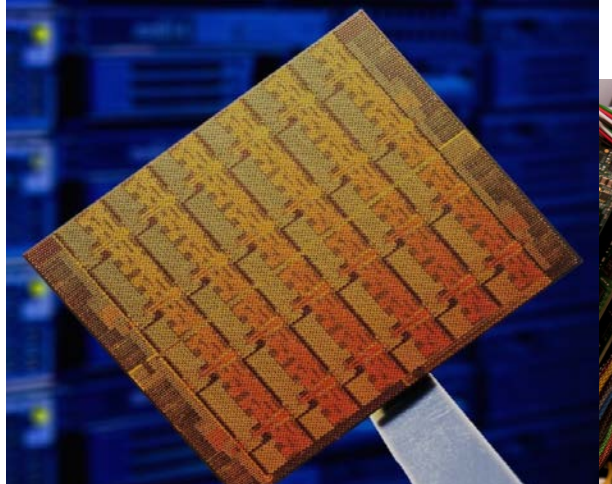
Performance and Energy Aware Computing Laboratory



Energy and thermal management of manycore systems:

-Scheduling
- Memory architecture
- Message passing / shared memory
-...

Performance and Energy Aware Computing Laboratory



Energy and thermal management of manycore systems:

-Scheduling
- Memory architecture
- Message passing / shared memory
-…

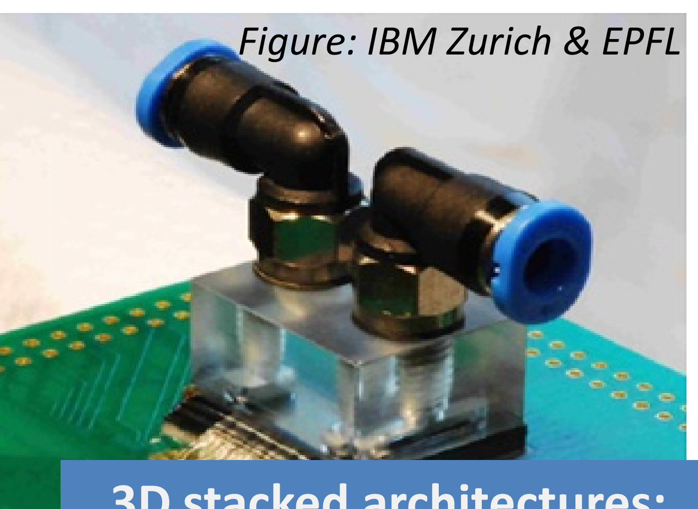
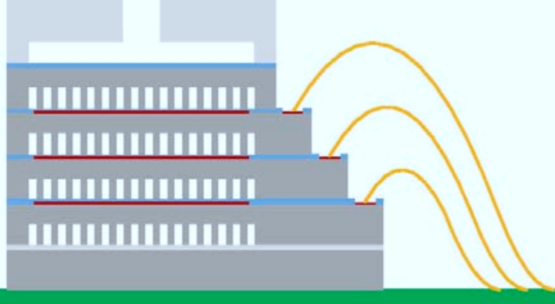


Figure: IBM Zurich & EPFL

3D stacked architectures:

Performance modeling, thermal verification, heterogeneous integration (e.g., DRAM stacking), …



Performance and Energy Aware Computing Laboratory

Green software:

Software optimization, parallel workloads, scientific & modeling applications, …

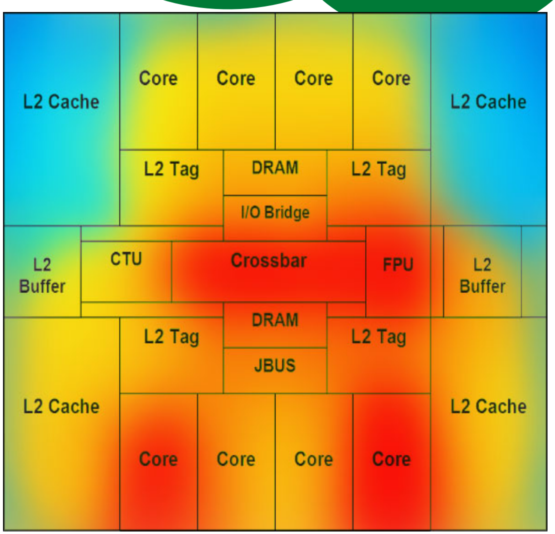
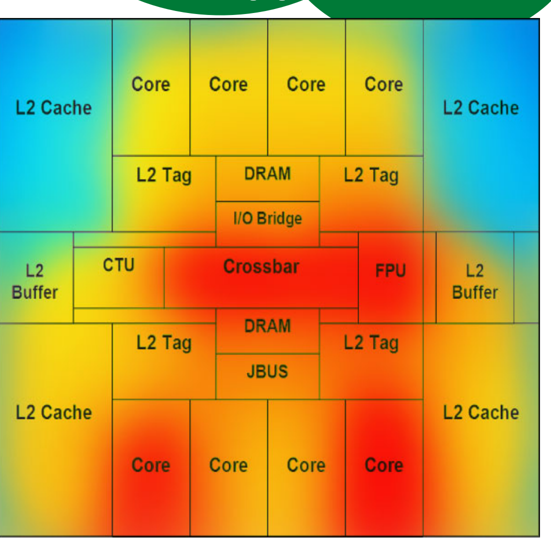


Figure: Argonne's Blue Gene/P supercomputer

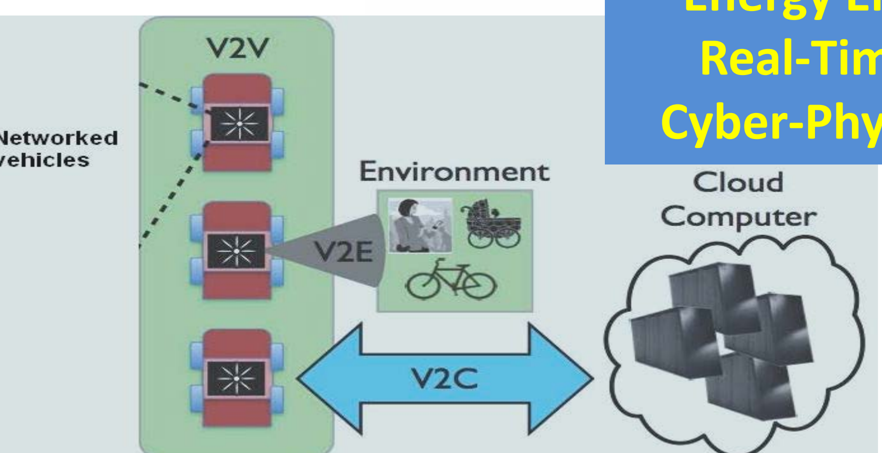
Performance and Energy Aware Computing Laboratory

Green software:

Software optimization, parallel workloads, scientific & modeling applications, …

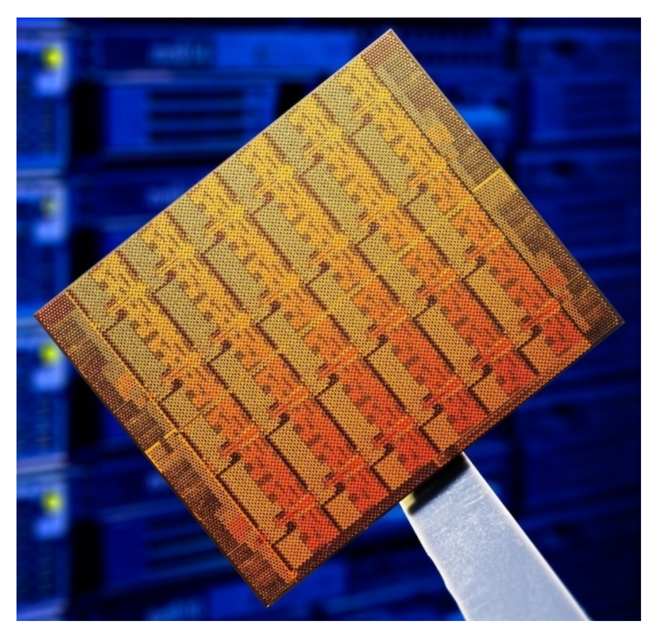


Energy Efficiency and Real-Time Design in Cyber-Physical Systems



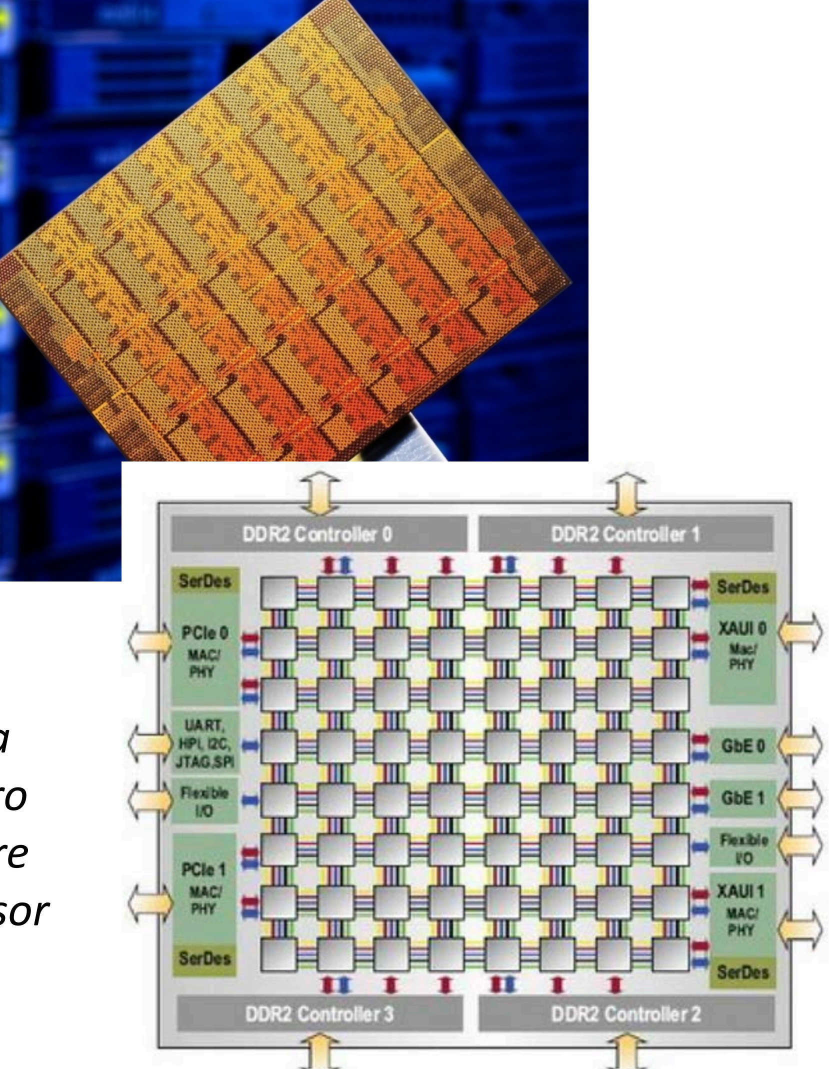
Multi-core to Many-core Architectures

Intel's 48-core Single-Chip Cloud Computer



Multi-core to Many-core Architectures

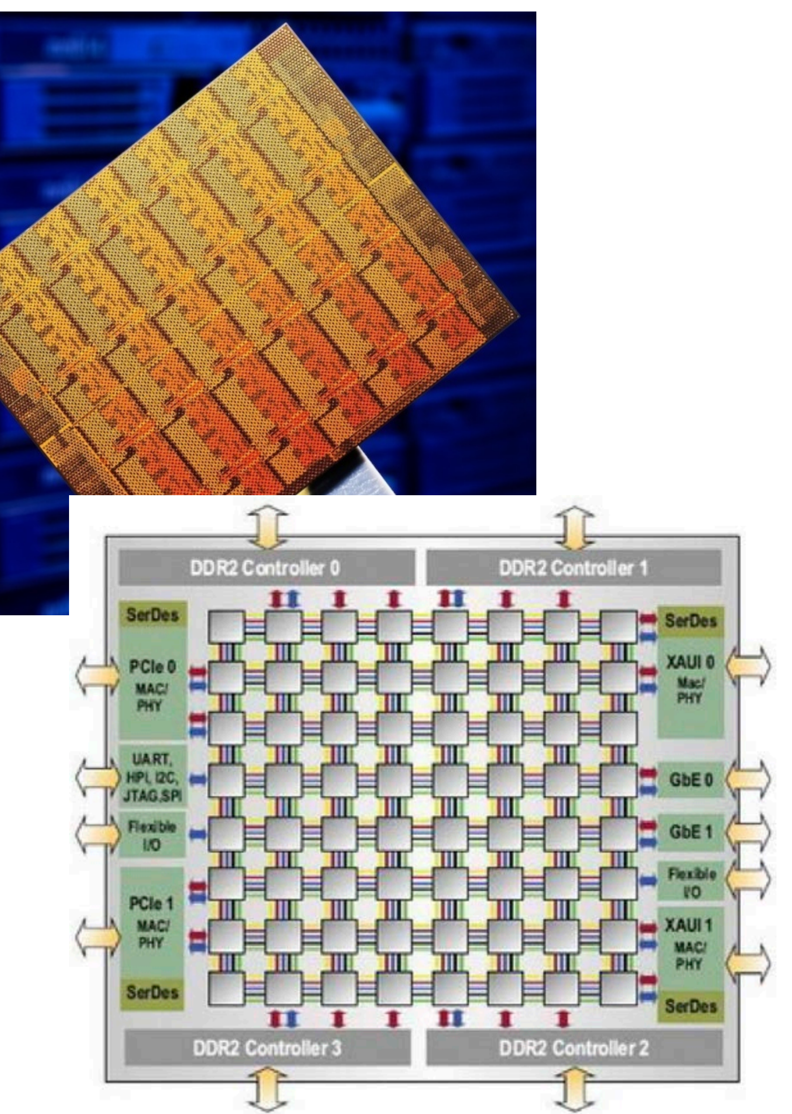
Intel's 48-core Single-Chip Cloud Computer



Tilera TILEPro 64-core Processor

Multi-core to Many-core Architectures

Intel's 48-core Single-Chip Cloud Computer



Tilera TILEPro 64-core Processor

Challenges in many-core systems

- Memory access latency
- Interconnect delay & power
- Yield
- Chip power & temperature

3D Stacking

- Shorter interconnects → Low power and high speed

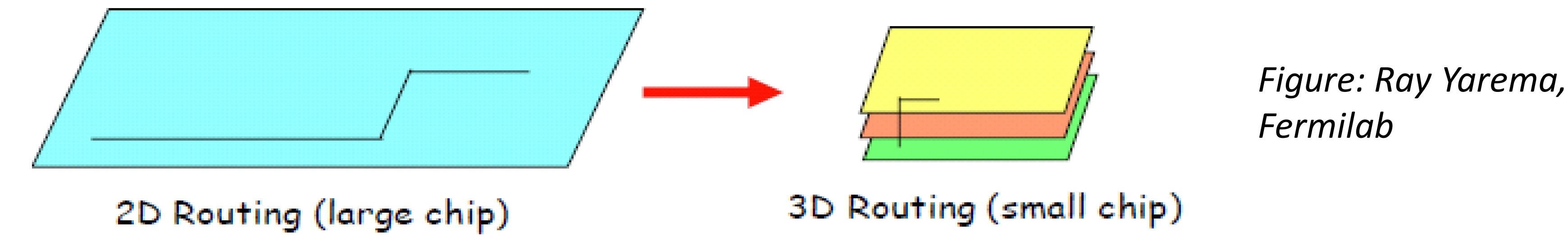


Figure: Ray Yarema, Fermilab

3D Stacking

- Shorter interconnects → Low power and high speed

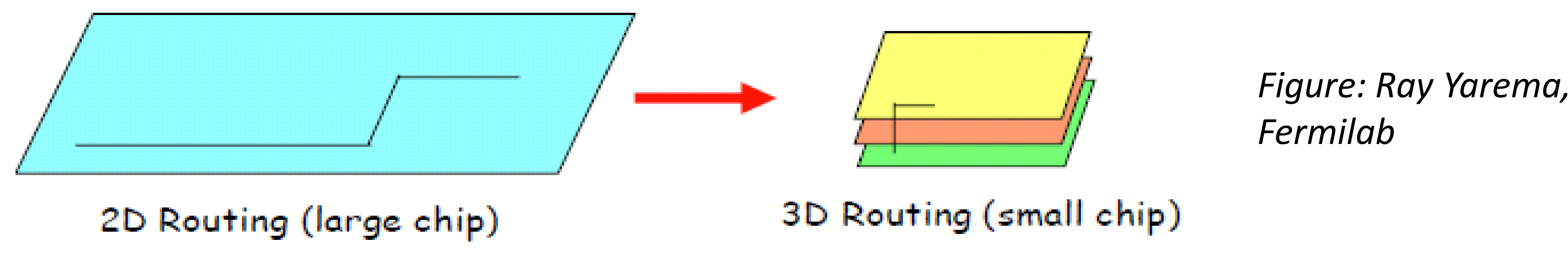


Figure: Ray Yarema, Fermilab

- Ability to integrate different technologies in a single chip

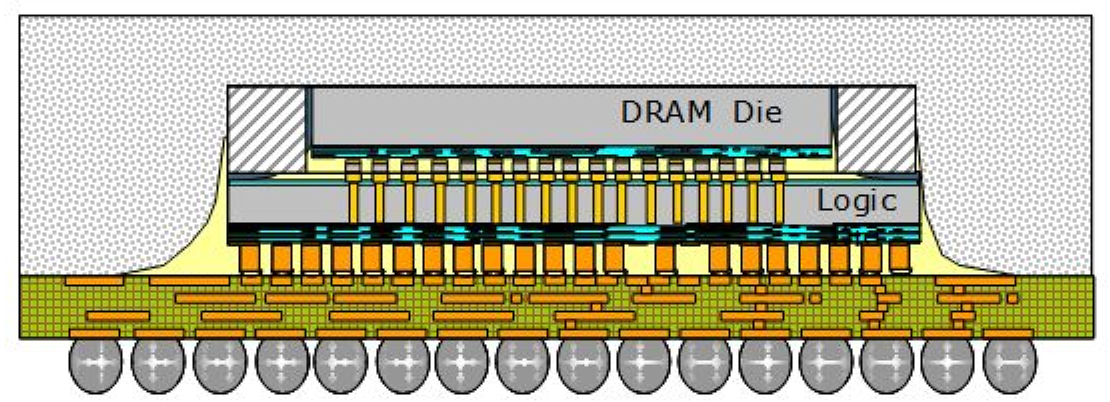


Figure: IMEC

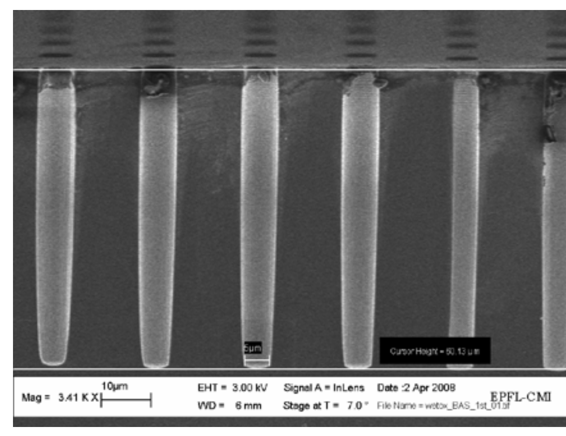


Figure: LSM, EPFL

Energy Efficiency and Temperature

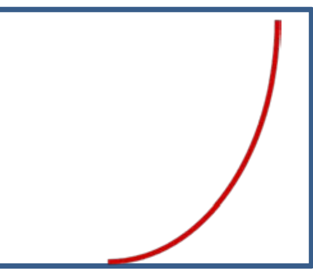
Temperature-induced challenges

Cooling Cost

Energy Efficiency and Temperature

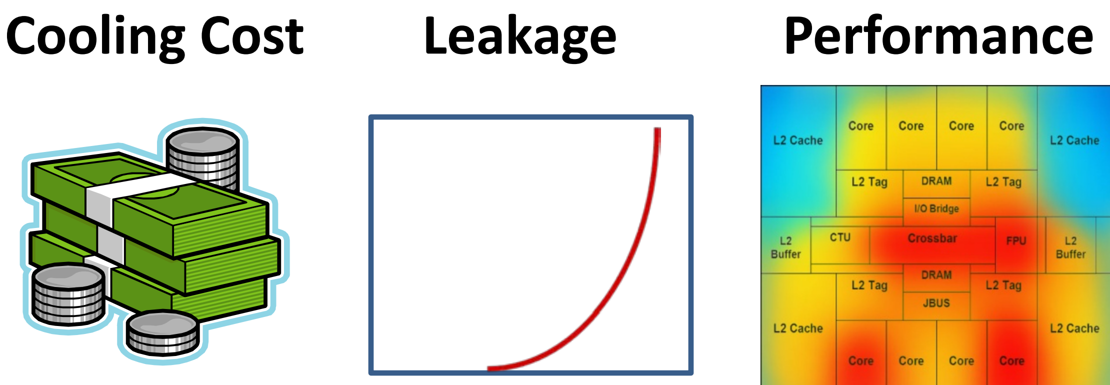
Temperature-induced challenges

Cooling Cost **Leakage**



Energy Efficiency and Temperature

Temperature-induced challenges

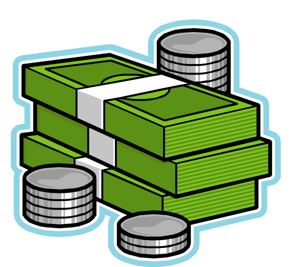


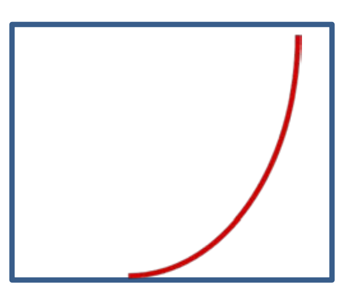
Energy Efficiency and Temperature

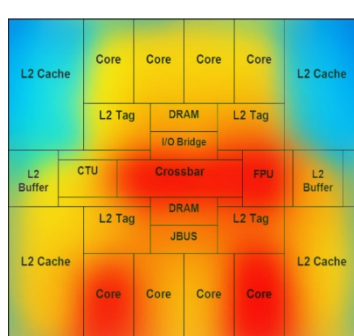
Temperature-induced challenges

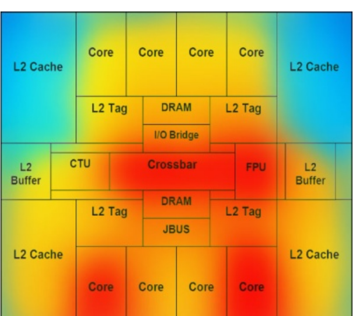
Cooling Cost

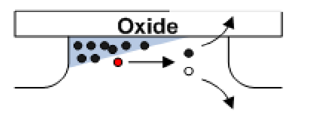
Performance



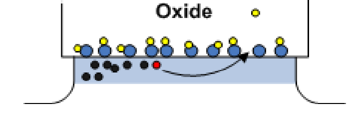




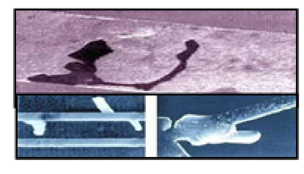




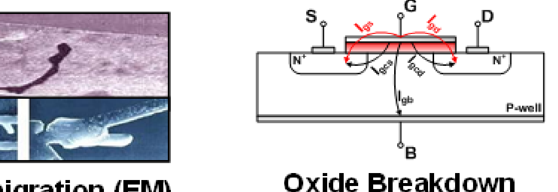
Hot carrier injection (HCI)



Negative Bias Temperature Instability (NBTI)



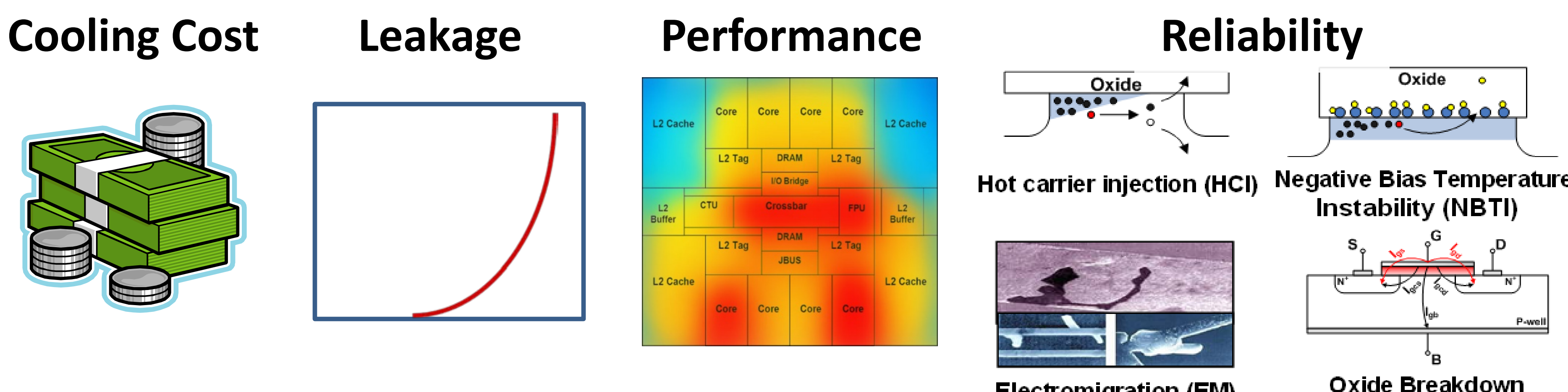
Electromigration (EM)



Oxide Breakdown

Energy Efficiency and Temperature

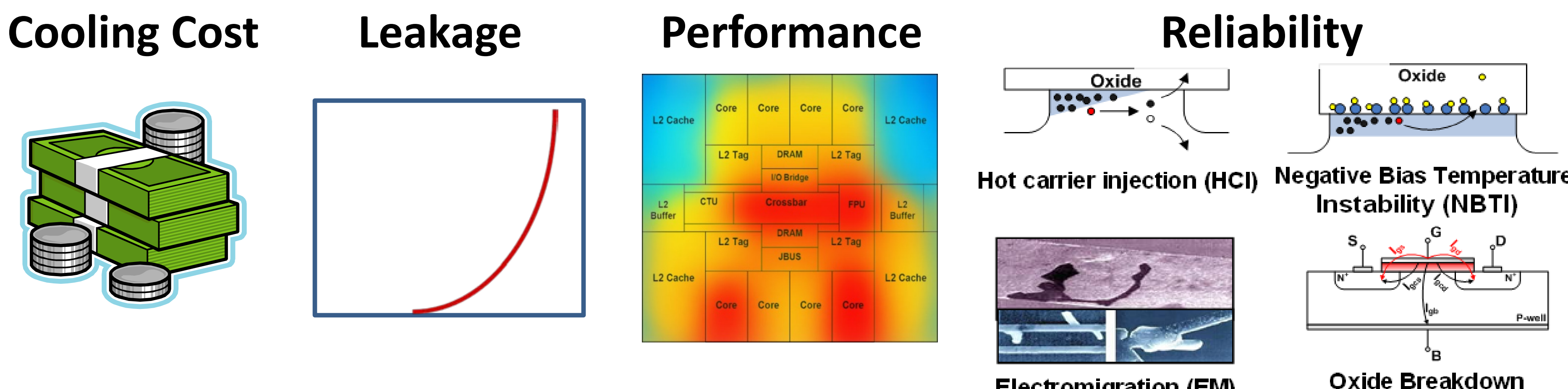
Temperature-induced challenges



Thermal challenges accelerate in high-performance systems!

Energy Efficiency and Temperature

Temperature-induced challenges



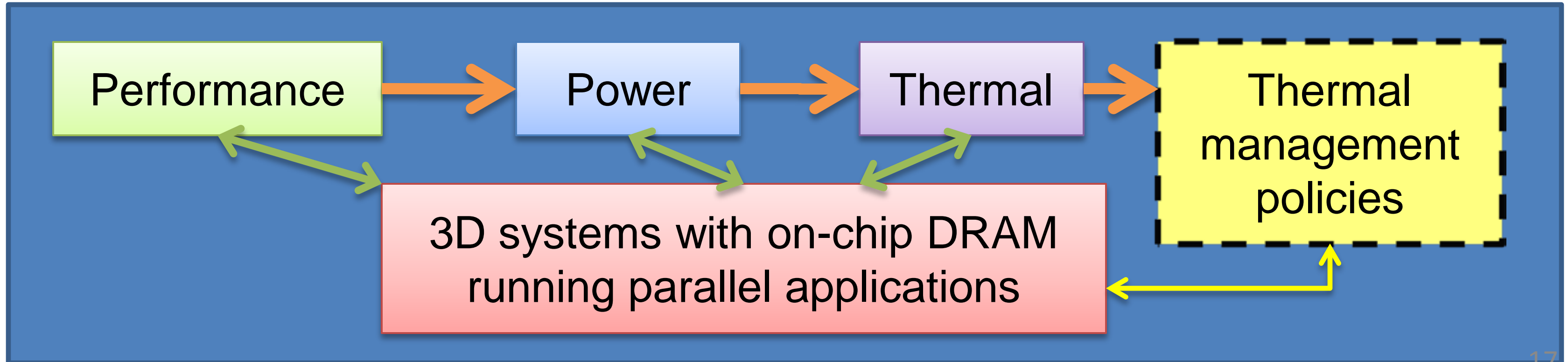
Thermal challenges accelerate in high-performance systems!

Energy problem

- High cost: a 10MW data center spends millions of dollars per year for operational and cooling costs
- Adverse effects on the environment

Contributions

- Model for estimating memory access latency in 3D systems with on-chip DRAM
- Novel methodology to jointly evaluate performance, power, and temperature of 3D systems
- Analysis of 3D multicore systems and comparisons with equivalent 2D systems demonstrating:
  - Up to **3X** improvement in throughput, resulting in up to **76%** higher power consumption per core
  - Temperature ranges area within safe margins for high-end systems. Embedded 3D systems are subject to severe thermal problems.



Outline

- System description:
  - 2D baseline *vs*. 3D target systems configuration
- Methodology:
  - Performance, power, and thermal modeling
  - Thread allocation policy
- Evaluation:
  - Exploring performance, power, and thermal behavior of 2D baseline *vs.* 3D system with DRAM stacking

Outline

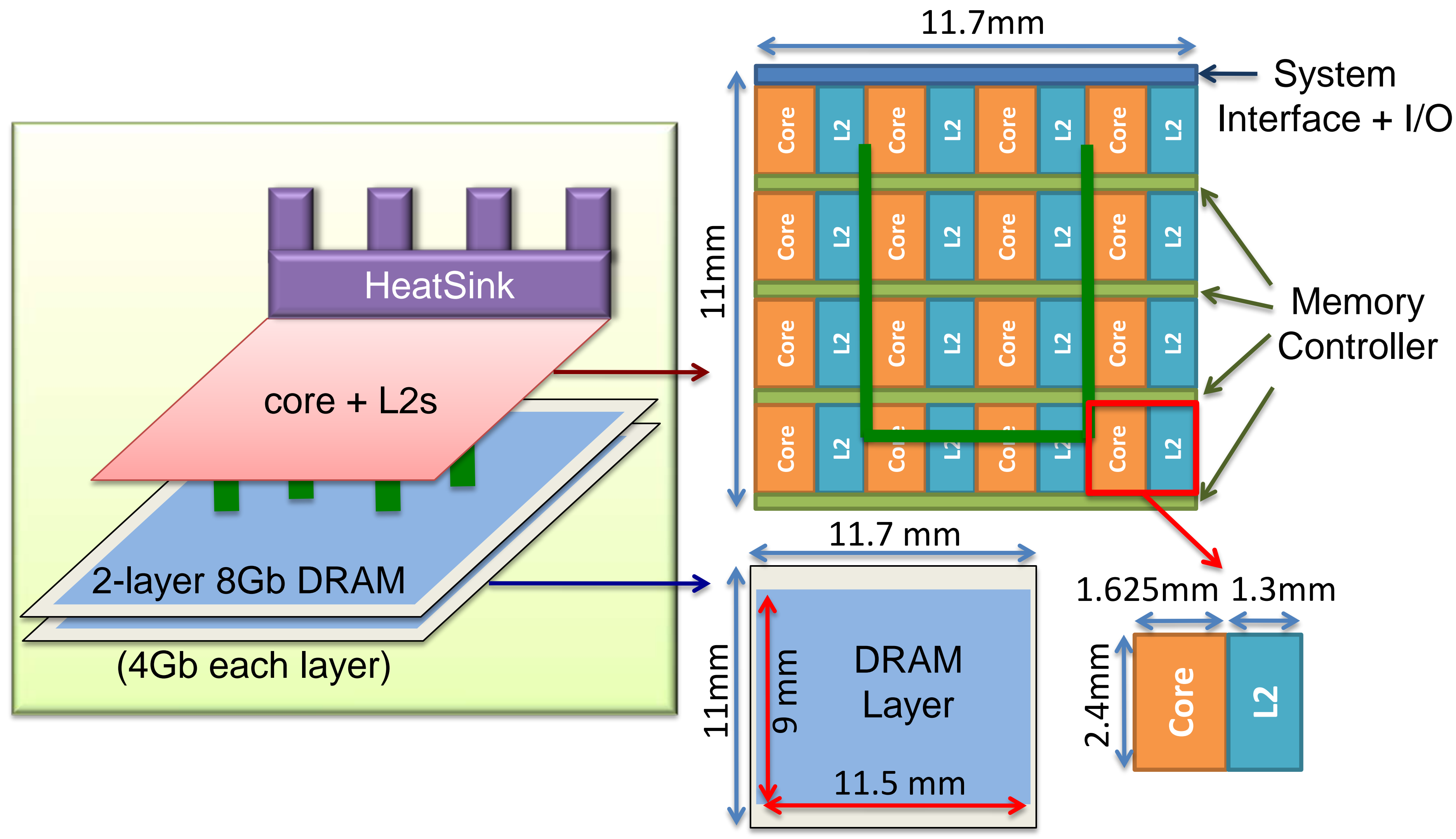
- System description:
  - 2D baseline *vs*. 3D target systems configuration
- Methodology:
  - Performance, power, and thermal modeling
  - Thread allocation policy
- Evaluation:
  - Exploring performance, power, and thermal behavior of 2D baseline *vs.* 3D system with DRAM stacking

Target System

- 16-core processor, cores based on the cores in Intel SCC [Howard, ISSCC'10]
- Manufactured at 45nm, has a die area of $128.7mm^2$

Core architecture	
CPU clock	1.0GHz
Branch Predictor	Tournament predictor
Issue Width	2-way out-of-order
Functional Units	2 IntAlu, 1 IntMult, 1 FPALU, 1 FPMultDiv
Physical Registers	128 Int, 128 FP
Instruction Queue	64 entries
L1 ICache / DCache	16 KB @2 ns (2 cyc)
L2 Cache(s)	16 private L2 Caches Each L2: 4-way set-associative, 64B blocks 512 KB @5 ns (5 cyc)

3D System with On-chip DRAM



Memory Access Latency: 2D *vs.* 3D

Memory access latency		
	2D-baseline design	**3D system with on-chip DRAM**
memory controller (MC)	4 cycles controller-to-core delay, 116 cycles queuing delay, 5 cycles MC processing time	4 cycles controller-to-core delay, 50 cycles queuing delay, 5 cycles MC processing time
main memory	off-chip 1GB SDRAM, t_{RAS} = 40ns, t_{RP} = 15ns, 10ns chipset request/return	on-chip 1GB SDRAM, t_{RAS} = 30ns, t_{RP} = 15ns, no chipset request/return
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

Memory Access Latency: 2D *vs.* 3D

Memory access latency		
	2D-baseline design	**3D system with on-chip DRAM**
memory controller (MC)	4 cycles controller-to-core delay, 116 cycles queuing delay, 5 cycles MC processing time	4 cycles controller-to-core delay, 50 cycles queuing delay, 5 cycles MC processing time
main memory	off-chip 1GB SDRAM, t_{RAS} = 40ns, t_{RP} = 15ns, 10ns chipset request/return	on-chip 1GB SDRAM, t_{RAS} = 30ns, t_{RP} = 15ns, no chipset request/return
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

Memory Access Latency: 2D *vs.* 3D

Memory access latency		
	2D-baseline design	**3D system with on-chip DRAM**
memory controller (MC)	4 cycles controller-to-core delay, 116 cycles queuing delay, 5 cycles MC processing time	4 cycles controller-to-core delay, 50 cycles queuing delay, 5 cycles MC processing time
main memory	off-chip 1GB SDRAM, t_{RAS} = 40ns, t_{RP} = 15ns, 10ns chipset request/return	on-chip 1GB SDRAM, t_{RAS} = 30ns, t_{RP} = 15ns, no chipset request/return
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

Memory Access Latency: 2D *vs.* 3D

Memory access latency		
	2D-baseline design	**3D system with on-chip DRAM**
memory controller (MC)	4 cycles controller-to-core delay, 116 cycles queuing delay, 5 cycles MC processing time	4 cycles controller-to-core delay, 50 cycles queuing delay, 5 cycles MC processing time
main memory	off-chip 1GB SDRAM, t_{RAS} = 40ns, t_{RP} = 15ns, 10ns chipset request/return	on-chip 1GB SDRAM, t_{RAS} = 30ns, t_{RP} = 15ns, no chipset request/return
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

Memory Access Latency: 2D *vs.* 3D

Memory access latency		
	2D-baseline design	**3D system with on-chip DRAM**
memory controller (MC)	4 cycles controller-to-core delay, 116 cycles queuing delay, 5 cycles MC processing time	4 cycles controller-to-core delay, 50 cycles queuing delay, 5 cycles MC processing time
main memory	off-chip 1GB SDRAM, t_{RAS} = 40ns, t_{RP} = 15ns, 10ns chipset request/return	on-chip 1GB SDRAM, t_{RAS} = 30ns, t_{RP} = 15ns, no chipset request/return
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

Memory Access Latency: 2D *vs.* 3D

Memory access latency		
	2D-baseline design	**3D system with on-chip DRAM**
memory controller (MC)	4 cycles controller-to-core delay, 116 cycles queuing delay, 5 cycles MC processing time	4 cycles controller-to-core delay, 50 cycles queuing delay, 5 cycles MC processing time
main memory	off-chip 1GB SDRAM, t_{RAS} = 40ns, t_{RP} = 15ns, 10ns chipset request/return	on-chip 1GB SDRAM, t_{RAS} = 30ns, t_{RP} = 15ns, no chipset request/return
memory bus	off-chip memory bus, 200MHz, 8Byte bus width	on-chip memory bus, 2GHz, 128Byte bus width

Outline

- System description:
  - 2D baseline *vs*. 3D target systems configuration

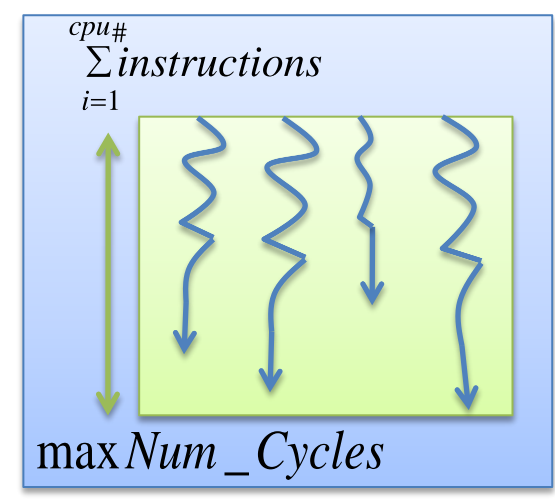
- **Methodology:**
  - **Performance, power, and thermal modeling**
  - **Thread allocation policy**

- Evaluation:
  - Exploring performance, power, and thermal behavior of 2D baseline *vs.* 3D system with DRAM stacking

Performance Model

- Performance metric: *Application IPC*

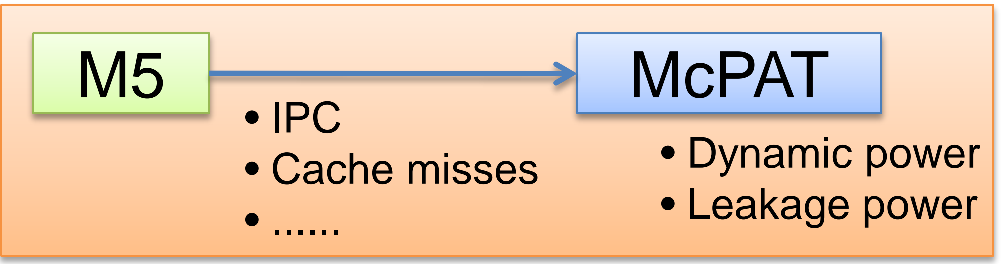
$$IPC_{app} = \frac{\sum_{i=1}^{cpu\#} Committed_Instrctions_{cpu[i]}}{\max_{1 \le i \le cpu\#} Num_Cycles_{cpu[i]}}$$



- Full-system simulator:
  - M5 (gem5) simulator [Binkert, IEEE Micro'06]
  - Thread-binding in an unmodified Linux 2.6 operating system
- Parallel benchmarks:
  - PARSEC parallel benchmark suite [Bienia, Princeton 2011]
  - Sim-large input sets in region of interest (ROI)

Power Model

- Processor power:
  - McPAT simulator [Li, MICRO' 06]
  - Calibration step to match the average power values of the Intel SCC cores



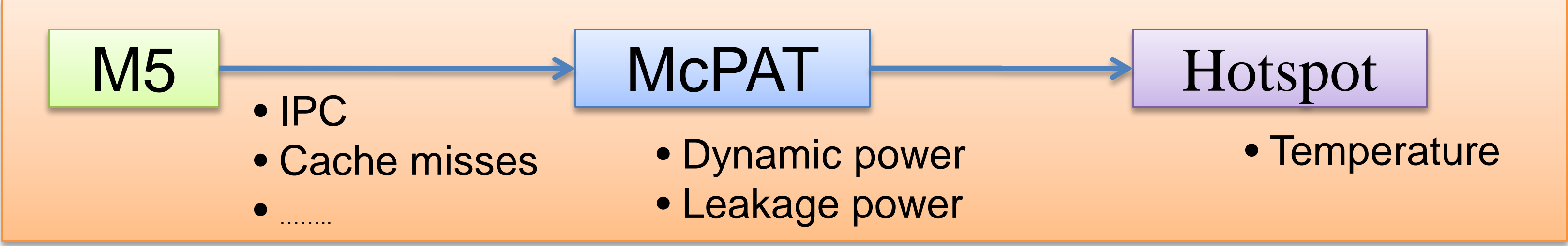
- L2 cache power:
  - CACTI 5.3 [HPLabs 2008]
  - Dynamic power computed using L2 cache access rate
- 3D DRAM power:
  - MICRON's DRAM power calculator [www.micron.com]
  - Takes the memory read and write access rates as inputs

Thermal Model

- Hotspot 5.0 [Skadron, ISCA' 03]
- Includes basic 3D features

Thermal simulation parameters	
Chip thickness	0.1mm
Silicon thermal conductivity	100 W/mK
Silicon specific heat	1750 kJ/m^3K
Sampling interval	0.01s
Spreader thickness	1mm
Spreader thermal conductivity	400 W/mK
DRAM thickness	0.02mm
DRAM thermal conductivity	100 W/mK
Interface material thickness	0.02mm
Interface material conductivity	4 W/mK

Thermal Model (Cont'd)



- We consider two additional packages representing smaller size and lower cost embedded packages.

Heat sink parameters for three different packages		
Package	**Thickness**	**Resistance**
High Performance	6.9 mm	0.1 K/W
No Heatsink (Embedded A)	10 µm	0.1 K/W
Medium Cost (Embedded B)	6.9 mm	1.0 K/W

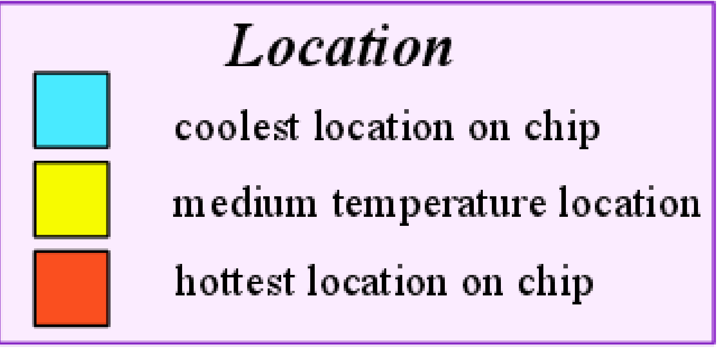
Outline

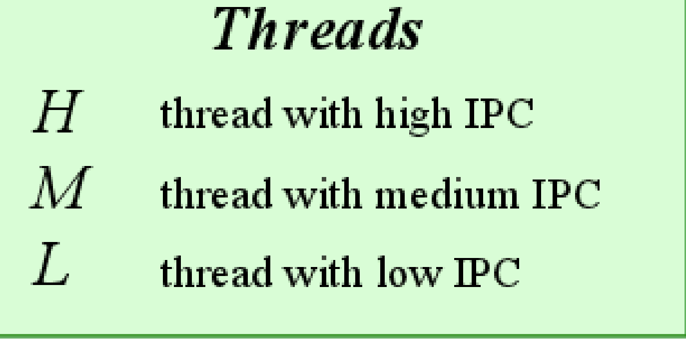
- System description:
  - 2D baseline *vs*. 3D target systems configuration
- Methodology:
  - Performance, power, and thermal modeling
  - Thread allocation policy
- **Evaluation:**
  - **Exploring performance, power, and thermal behavior of 2D baseline *vs.* 3D system with DRAM stacking**

Thread Allocation Policy

- Based on the *balance_location* policy [Coskun, *SIGMETRICS '09*]
- Assigns threads with the highest *IPC*s to the cores at the coolest locations on the die

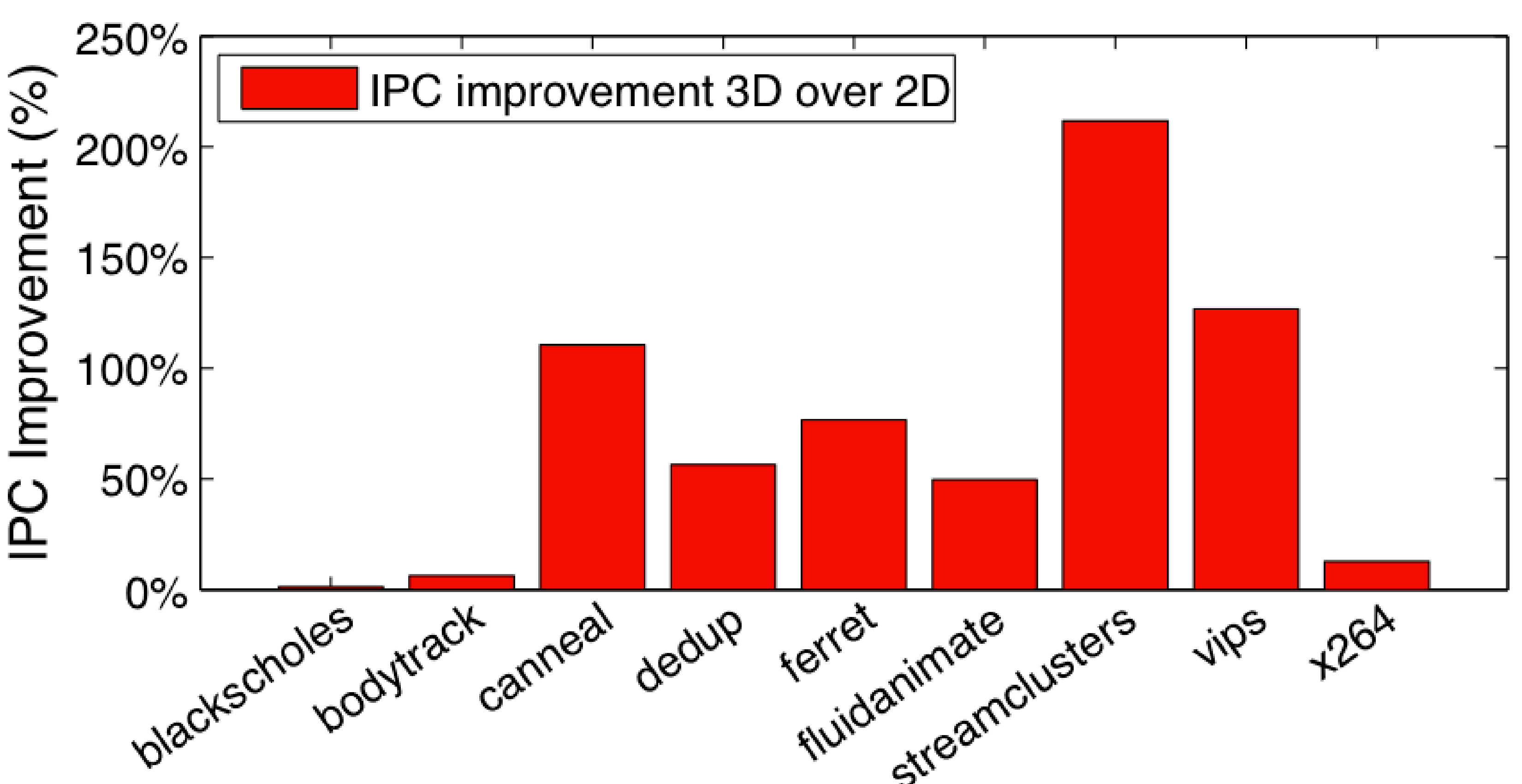
H	*M*	*M*	*H*
M	*L*	*L*	*M*
M	*L*	*L*	*M*
H	*M*	*M*	*H*





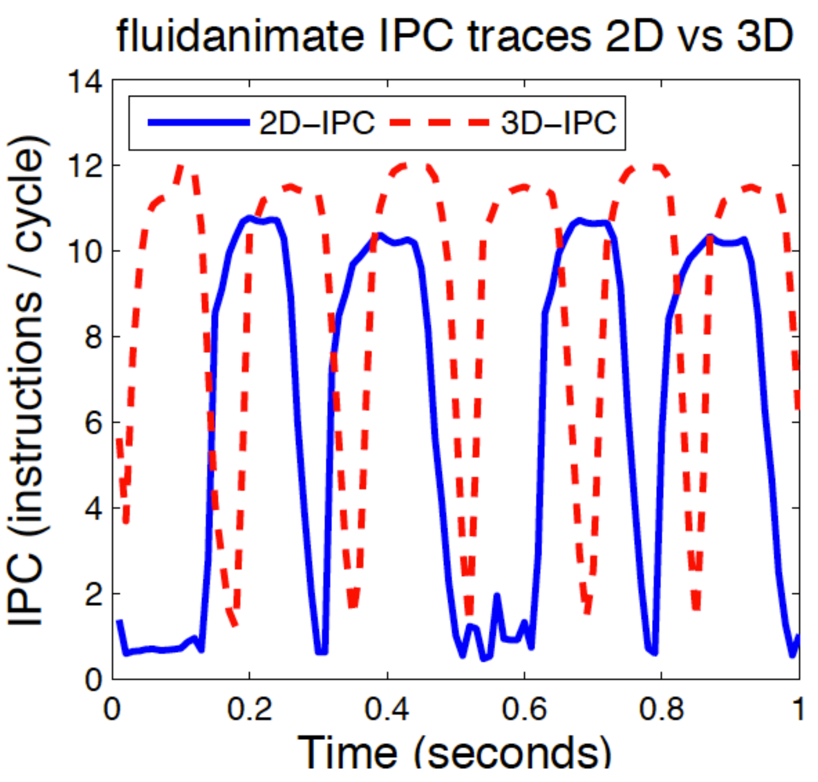
Performance Evaluation

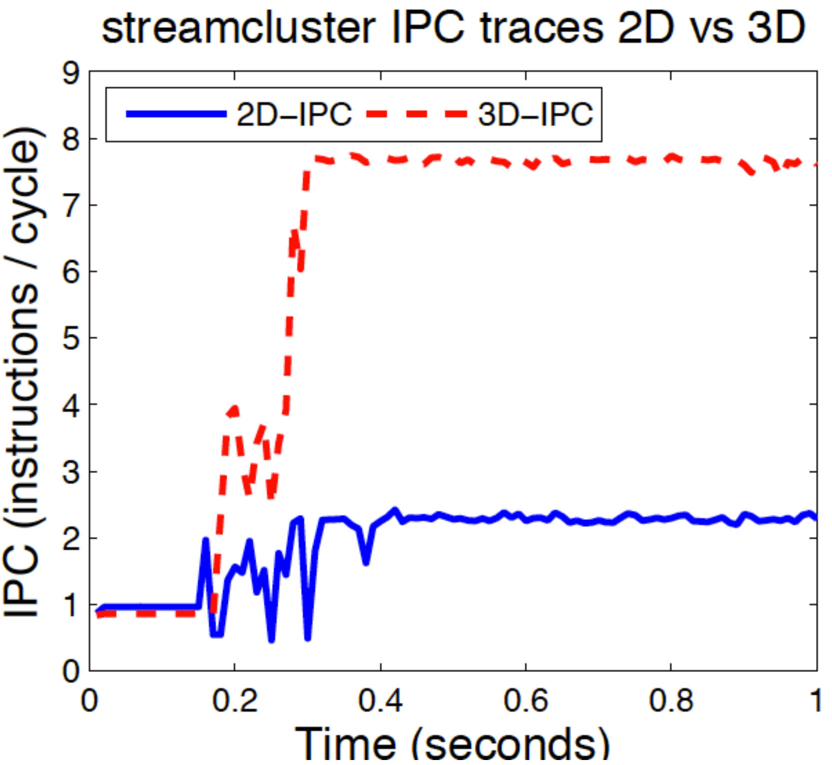
- 3D DRAM stacking achieves an average IPC improvement of 72.55% compared to 2D.



Temporal Performance Behavior

- *streamcluster* and *fluidanimate* improve their *average IPC* by 211.8% and 49.8%, respectively.

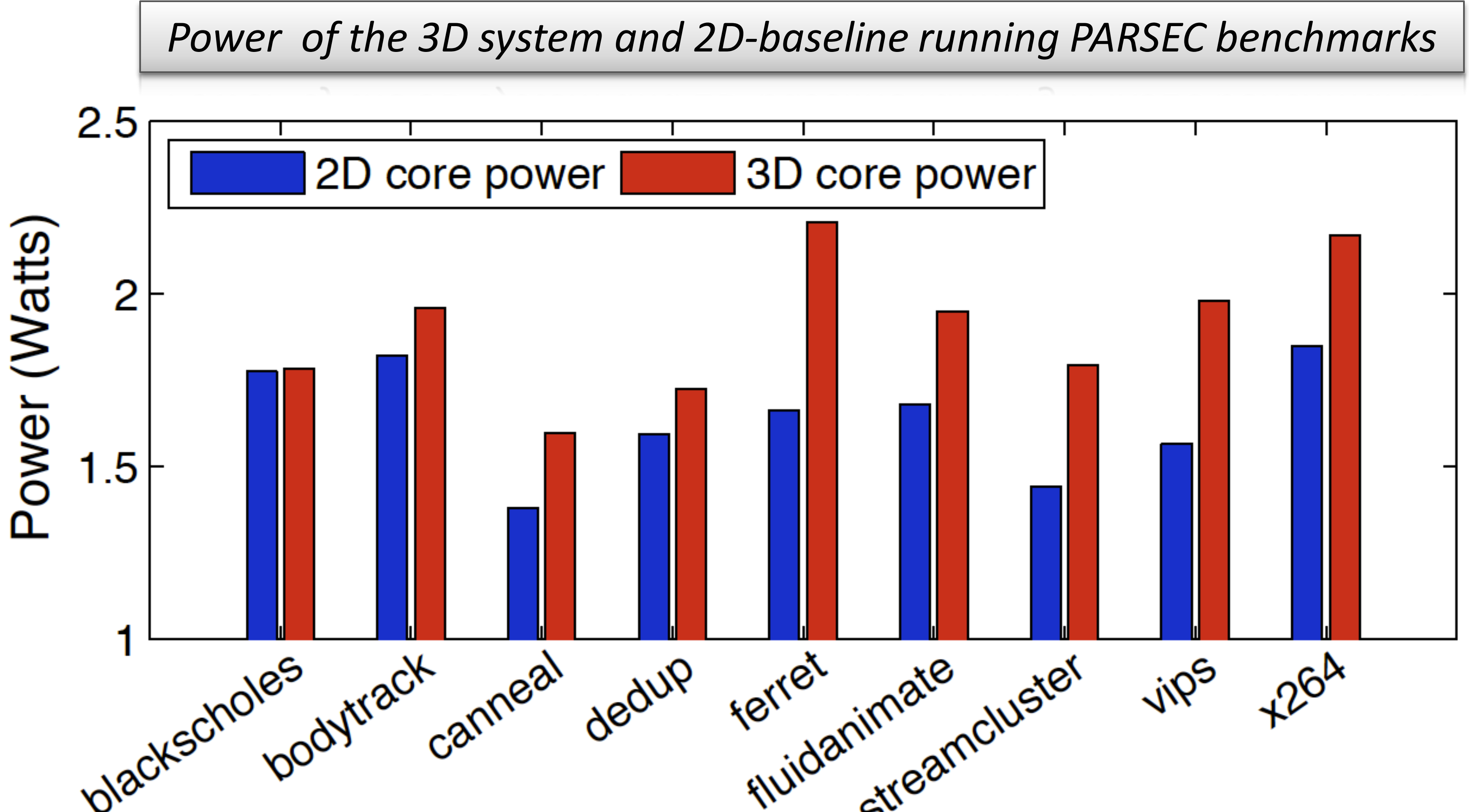




Power Evaluation

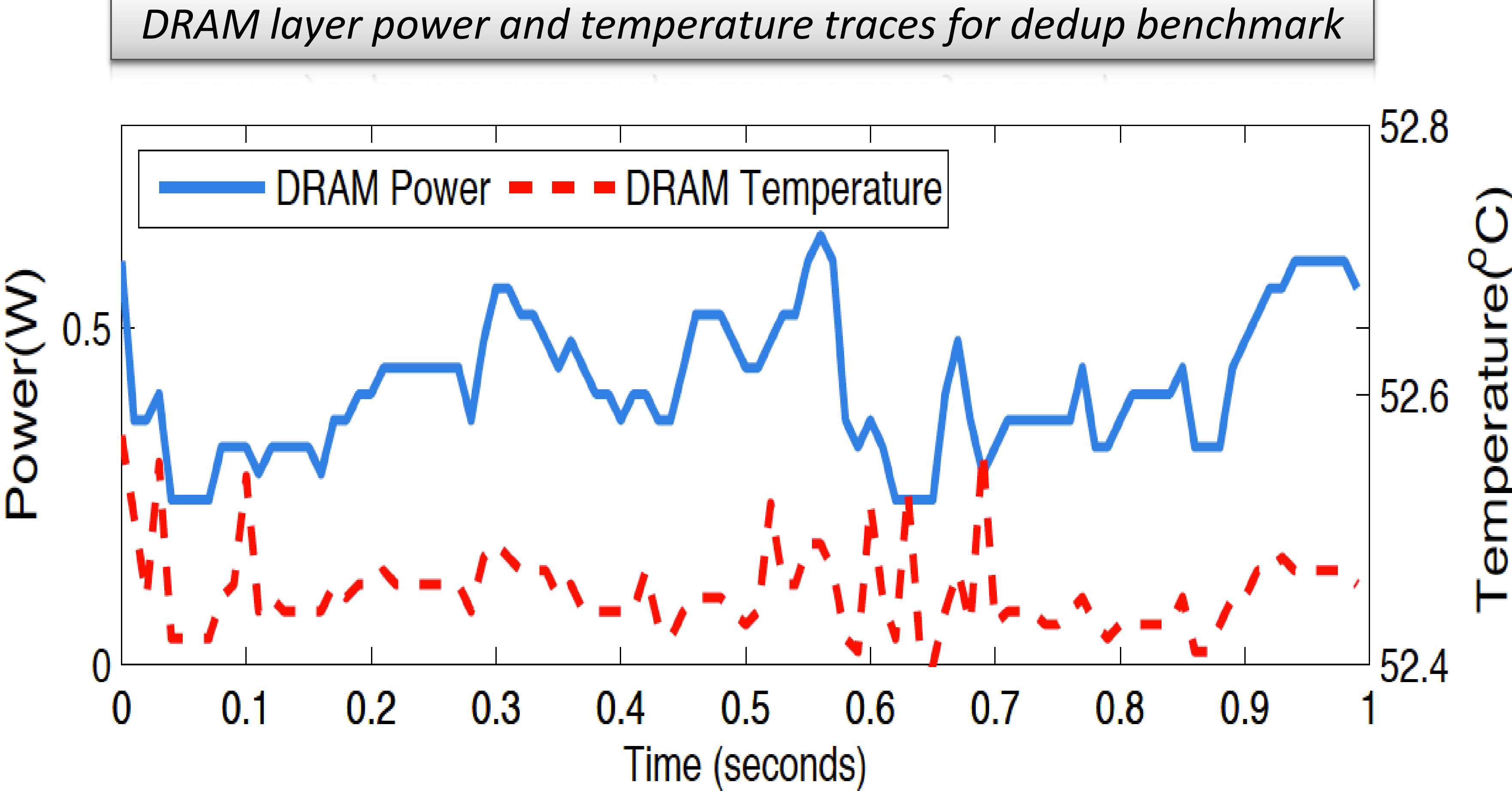
- Per-core power increases by 16.6% on average for the 3D system.

Power of the 3D system and 2D-baseline running PARSEC benchmarks



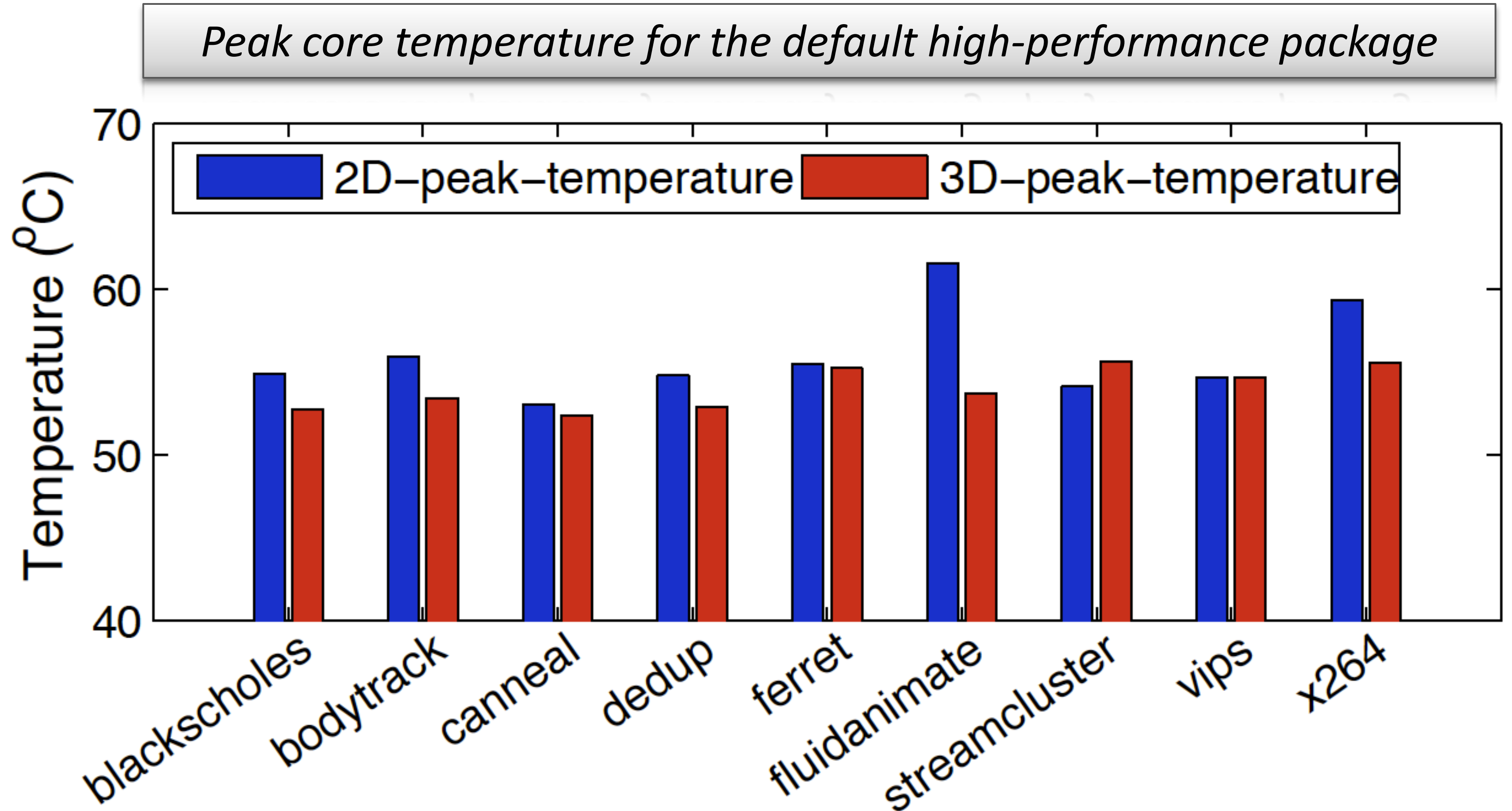
DRAM Power and Temperature

- DRAM power changes following the variations in memory access rate.



Temperature Analysis

- Temperature decreases because the lower power DRAM layer shares the heat of the hotter cores.

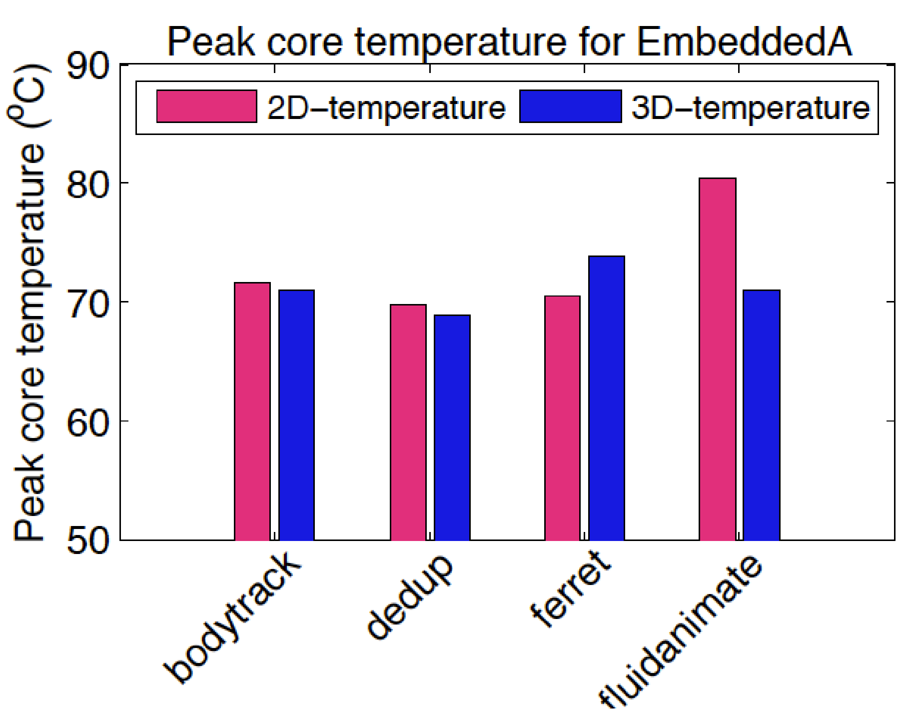


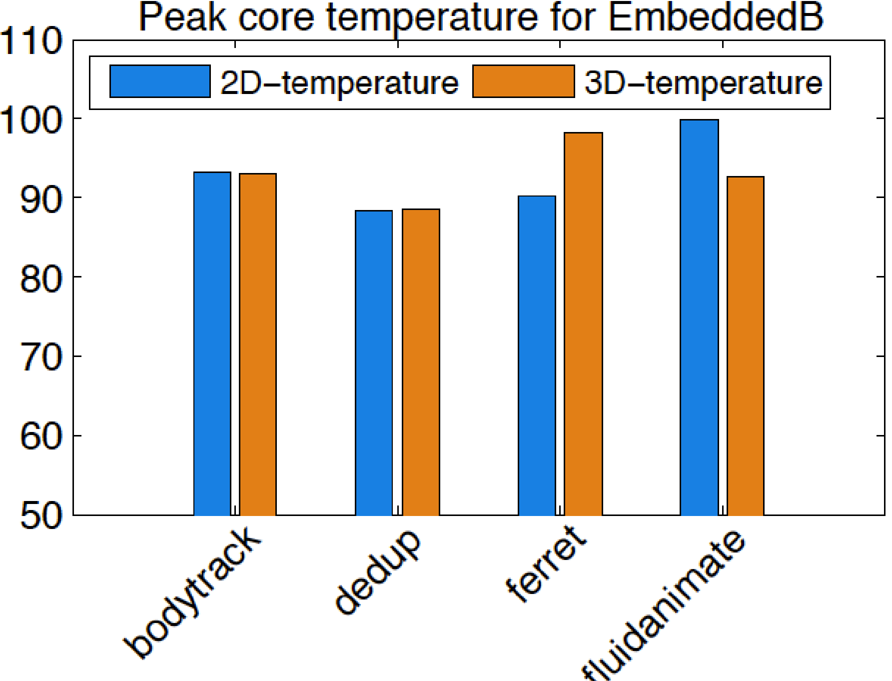
Temperature Analysis (Cont'd)

- Temperatures increase more noticeably in 3D systems with small-size and low-cost embedded packages

small-size embedded package

low-cost embedded package





Conclusion

- We provide a comprehensive simulation framework for 3D systems with on-chip DRAM.
- We explore the performance, power, and temperature characteristics of a 3D multi-core system running parallel applications.
- Average IPC increase is 72.6% and average core power increase is 16.6% compared to the equivalent 2D system.
- We demonstrate limited temperature changes in the 3D systems with DRAM stacking with respect to the 2D baseline.
- Future work: Detailed DRAM power models, higher bandwidth memory access, new 3D system architectures, new thermal/energy management policies.

Performance and Energy Aware Computing Laboratory

- Collaborators
  - EPFL, Switzerland
  - IBM
  - Oracle
  - Intel
  - Brown University
  - University of Bologna, Italy
- Funding:
  - DAC, Richard Newton Award
  - Dean's Catalyst Award, BU
  - Oracle
  - VMware

Contact:
http://www.bu.edu/peaclab
http://people.bu.edu/acoskun
acoskun@bu.edu