Leakage Power Characterization and Minimization in 3D Stacked Multi-core Chips with Microfluidic Cooling

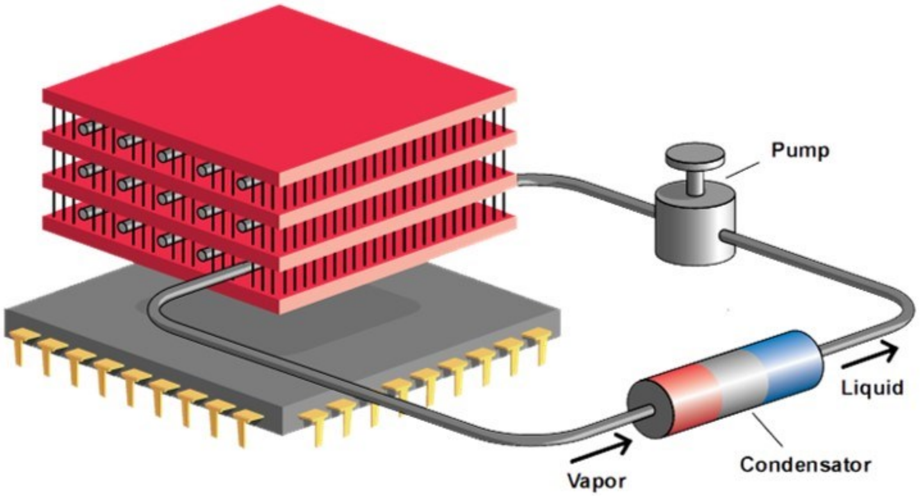
He Xiao*, Zhimin Wan+, Sudhakar Yalamanchili*, and Yogendra Joshi+

School of Electrical & Computer Engineering*
School of Mechanical Engineering+
Georgia Institute of Technology
Atlanta, GA 30332, USA

Sponsor: Sandia National Laboratories and the National Science Foundation under grant CNS-855110

Objective

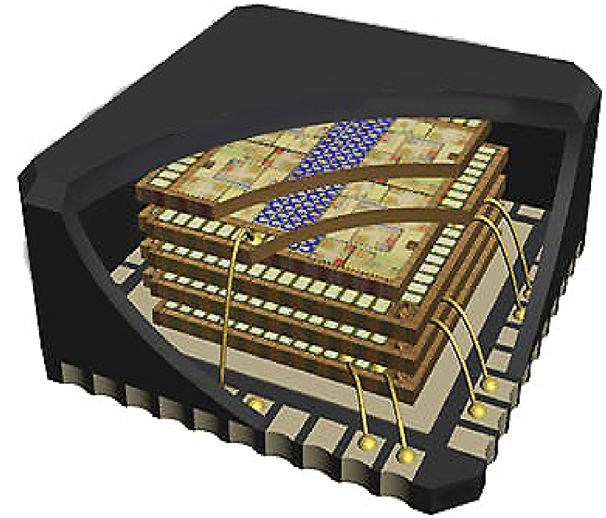
- To characterize leakage power in 3D multi-core architecture as a function of the cooling design
  - ✓ Evaluate the leakage reduction for the optimized pin fin



Infograph: Pascal Coderay, pascal@salut.ch

- To study the impact of frequency scaling in a 3D multi-core architecture with microfluidic pin fin cooling on performance speedup, power consumption and energy efficiency

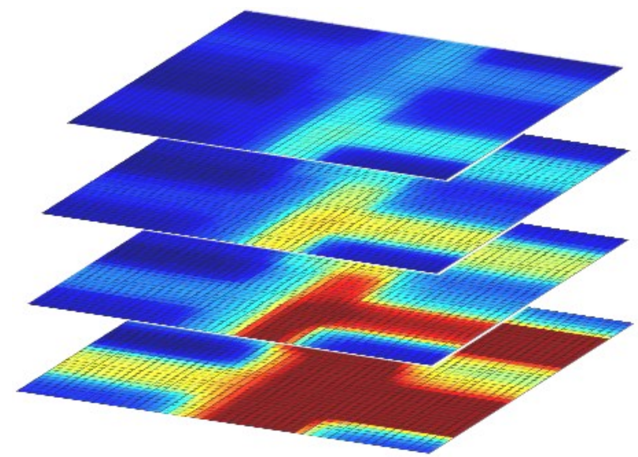
Background



3D Stacked ICs:

IC tiers are vertically integrated in a compact package.

- ❖ Length of global wires reduced by 50%[1]
- ❖ Wire-limited frequency increased by 3.9X[1]
- ❖ Wire-limited area and power reduced by 84% and 51%[1]
- ❖ High communication bandwidth between tiers
- ❖ Heterogeneous tiers with multi-functionality



The Challenges:

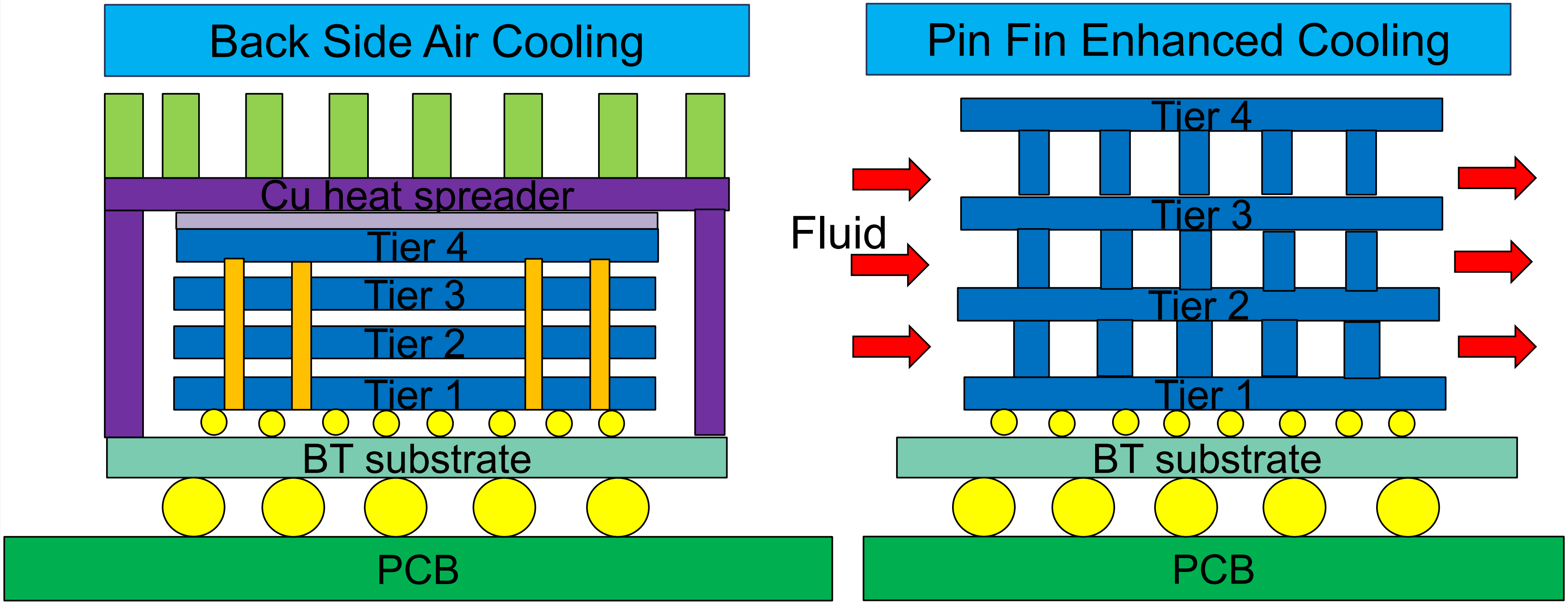
The thermal issues become pronounced, ultimately downgrading system performance.

- ❖ Highly increased heat dissipation and power density
- ❖ Non-uniformity heat flux leads to hotspot
- ❖ Strong thermal coupling between tiers

[1] N. H. Khan, S. M. Alam, and S. Hassoun, "System-level comparison of power delivery design for 2D and 3D ICs," Proc. 3DIC 2009, San Francisco, CA, USA, pp. 1-7, 2009.

Why Microfluidic Cooling?

Liquid cooling with surface enhancement such as pin fin is a viable solution to reduce the thermal stress in 3D stacked structures.



HTC: 25-250 W/m^2-K[2]

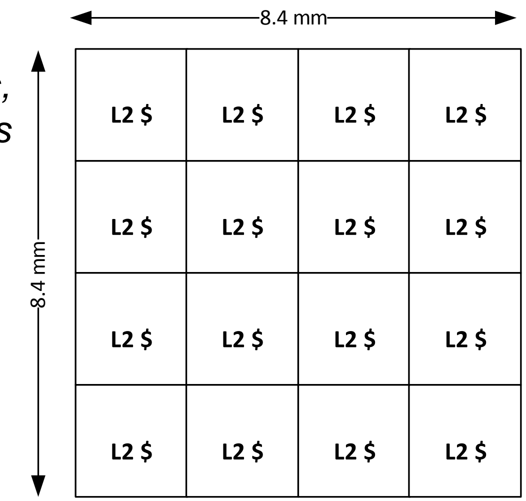
HTC: 100-20000 W/m^2-K[2]

HTC is short for Heat Transfer Coefficient

[2] T. L. Bergman, A. S. Lavine, F. P. Incropera, and D. P. DeWitt, "Fundamentals of Heat and Mass Transfer," 2011.

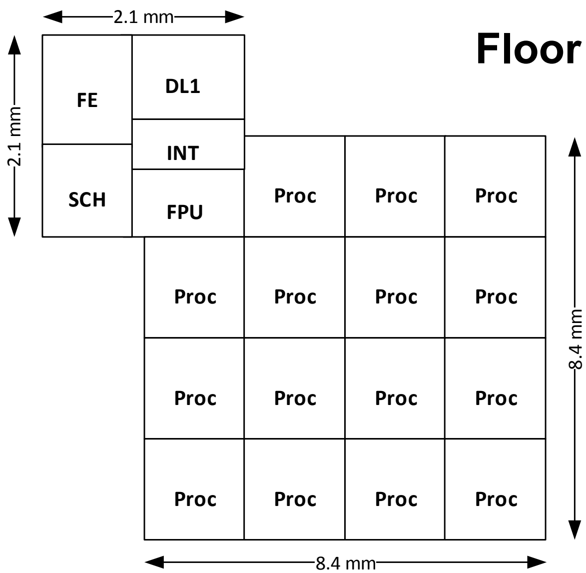
System Architecture

L2 (per core): 2MB, 4096 sets, 128B, 35 cycles



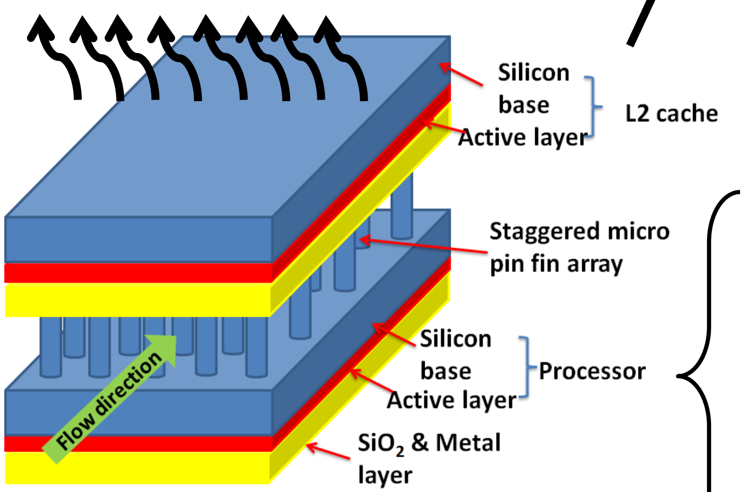
Nehalem-like, Out-of-Order cores;
3GHz, 1.0V, max temp 373K
Issue width 4, ROB size 128
DL1: 128KB, 4096 sets, 64B
IL1: 32KB, 256 sets, 32B

Floor plan is built upon 16nm technology



L2 is shared among cores

Convection air cooling:
Temperature: 300K

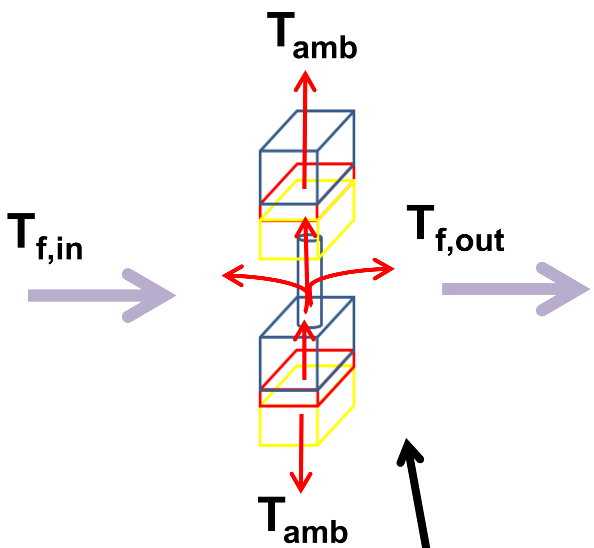


16 symmetric cores

- Linux OS
- Parsec
- SPEC
- Splash-2

Compact Thermal Model (CTM)

Control volume around one pin

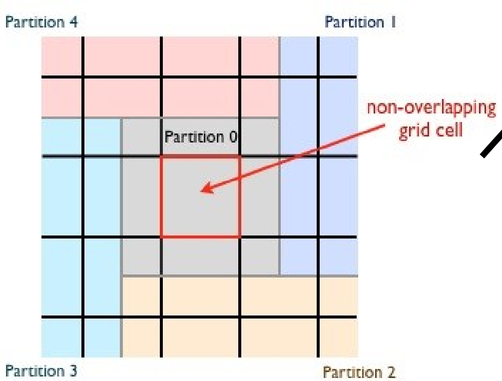


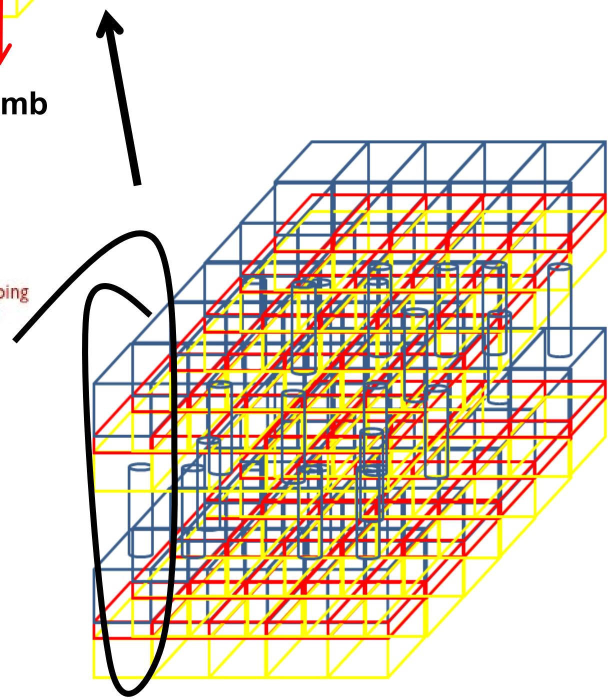
Energy balance for Solid:

$$\dot{q}_{gen} + \dot{q}_{cond} + \dot{q}_{conv} = 0$$

Energy balance for Fluid:

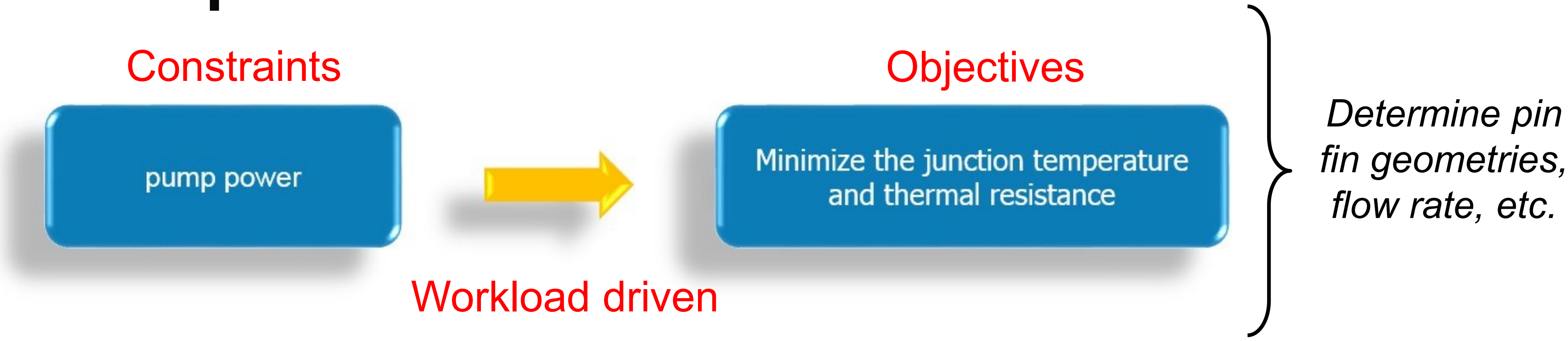
$$\dot{m}C_p\left(T_{f,in} - T_{f,out}\right) + \dot{q}_{conv} = 0$$





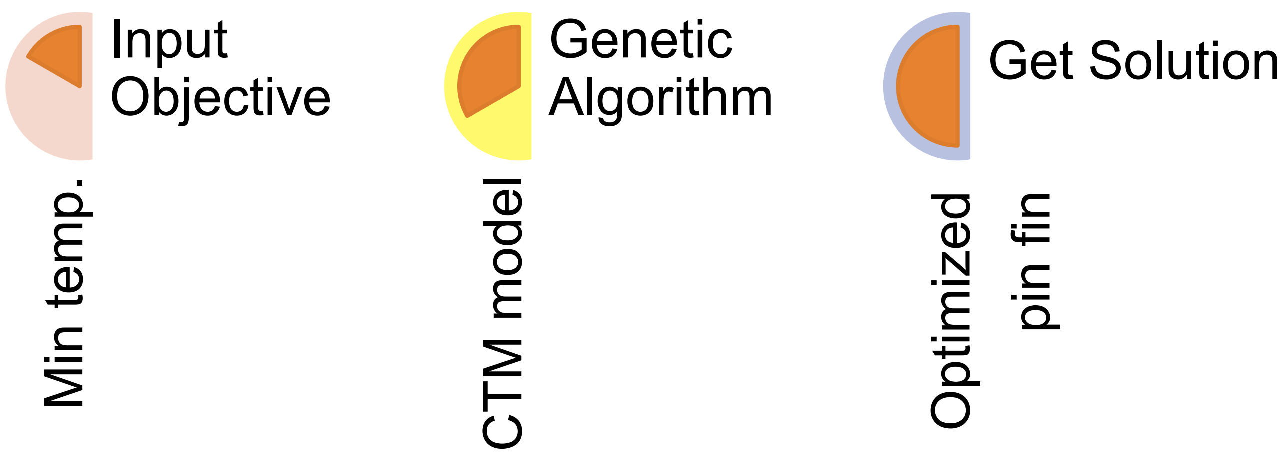
	Thickness	Thermal Cond.
SiO2&Metal	10 μm	1.4 W/m-K
Silicon	100 μm	149 W/m-K

Pin Fin Optimization Process

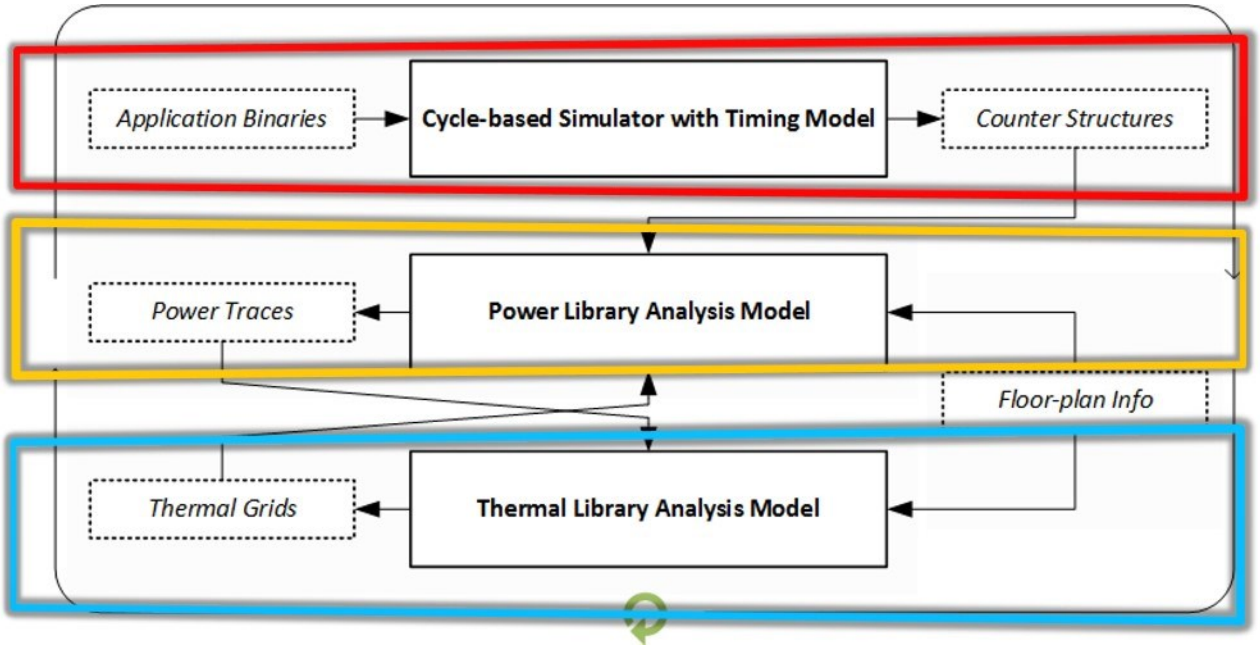


Optimization Algorithm:

Genetic algorithm is used to do the optimization. The compact model is embedded into a genetic algorithm as a function. After optimization, the optimized pin fin dimensions are obtained.



Simulation Framework



Phase 1 (timing simulation[3]): The benchmark is executed on a cyclebased simulator, which collects the info of pipeline execution and cache reference.

Phase 2 (power analysis[4]): The floor plan and architecture info are used to generate the power traces of different components

Phase 3 (thermal analysis[5]): The 3D floor plan and power traces are used as the input of the thermal library to compute the thermal fields, which update the leakage power for the next iteration.

[3] J. Wang, J. Beu, R. Behda, T. Conte, Z. Dong, C. Kersey, M. Rasquinha, G. Riley, W. Song, He Xiao, P. Xu, and S. Yalamanchili, "Manifold: A Parallel Simulation Framework for Multicore Systems," ISPASS, 2014.

[4] S. Li, J. Ahn, R. Strong, J. Brockman, D. tullsen, and N. Jouppi, "Mcpat: An integrated power, area, and timing modeling framework for multi-core and manycore architectures", Micro-42, pp. 469–480, 2009.

[5] M. Sridhar, T. B., A. V., D. A., and M. R., 3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling, IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, pp. 463-470, 2010.

Simulation Framework Cont'

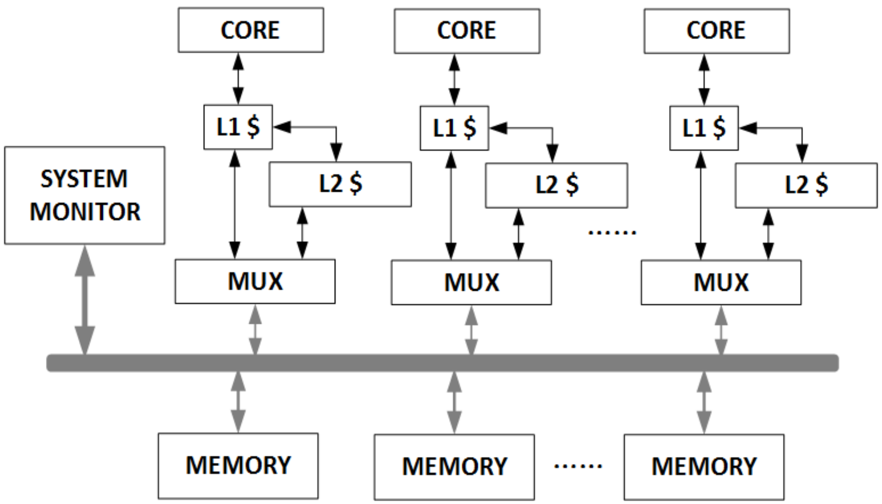
Convection Heat Sink	
Heat transfer coefficient	1.2e-11 W/μm^2K
Ambient temperature	300 K

Micro Pin Fin	
Pin material	Silicon
Pin distribution	Staggered
Coolant vol. heat cap.	4.17e-12 J/μm^3K
Coolant incoming temp.	300 K

	DP (um)	**PS** (um)	**HP** (um)
baseline	100	200	200
optimized	180	320	400

DP: diameter, PS: pitch spacing, HP: height;
Pumping power 0.03W

The structure of 16-core simulator



The system monitor is built inside the simulator to coordinate the execution of timing model and the invocation of the physical models across sampling windows.

Power Profile over Splash-2 Benchmark

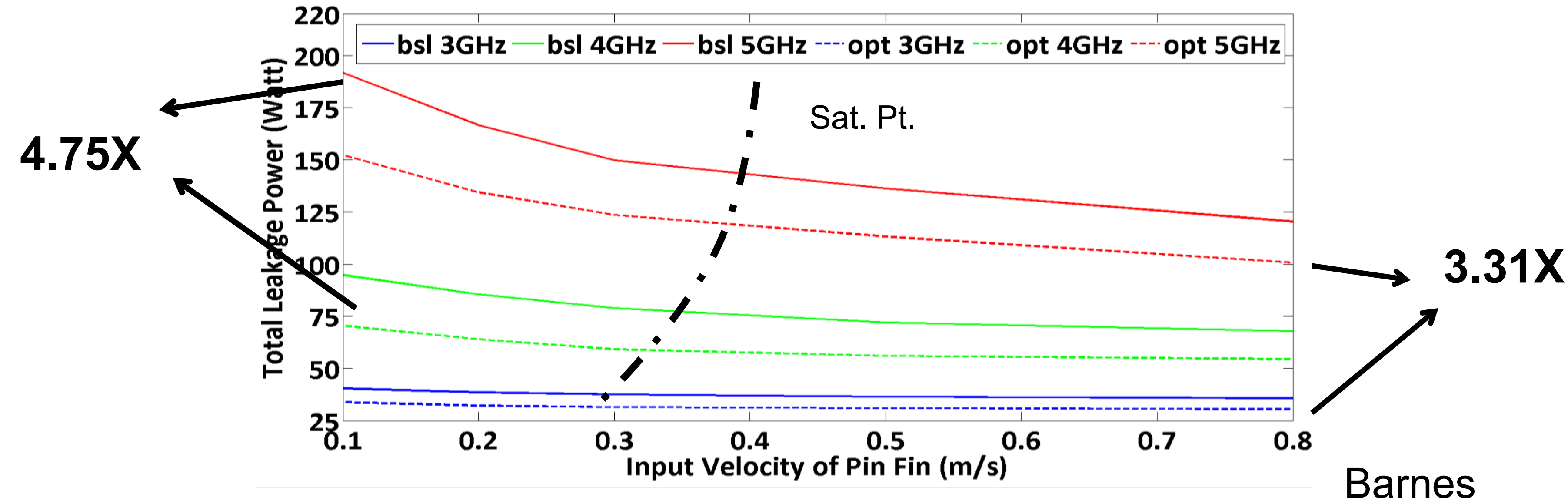
The pin fin is fixed to baseline configuration with input velocity 0.3m/s

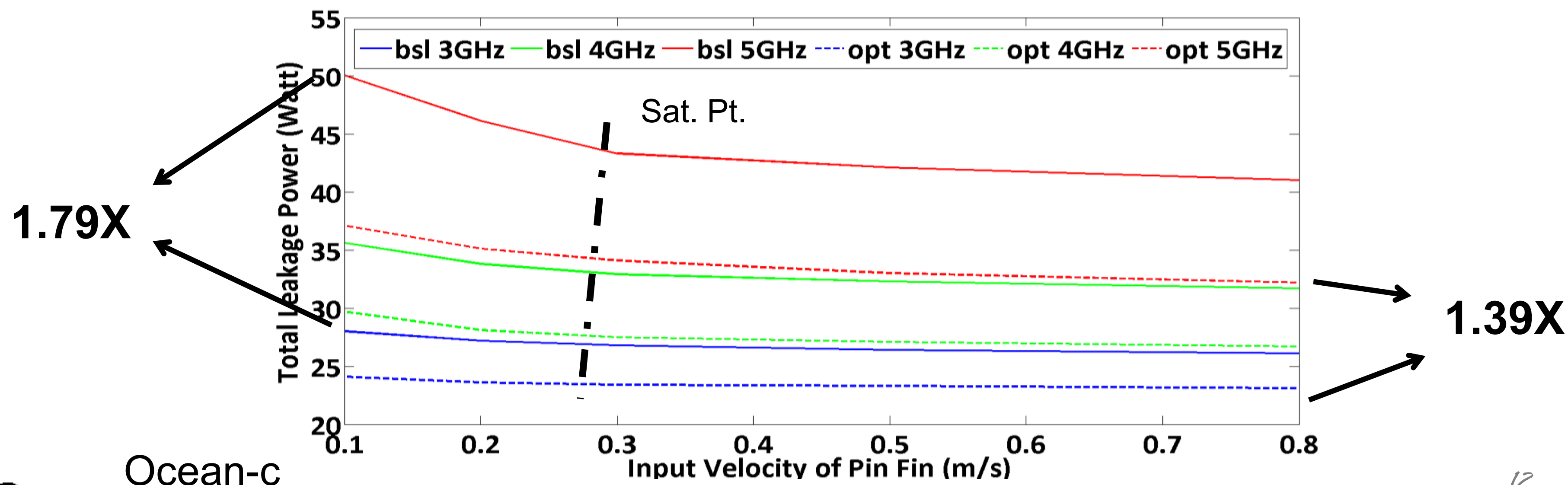
Case Study

- ✓ **Barnes:**(computational-bounded) simulates the interaction of a system of bodies in three dimensions using Barnes-Hug hierarchical N-body method.
- ✓ **Ocean-c:** (Memory-bounded) studies large-scale ocean movements based on eddy and boundary currents, using a red-black Gauss-Seidel multigrid equation solver.

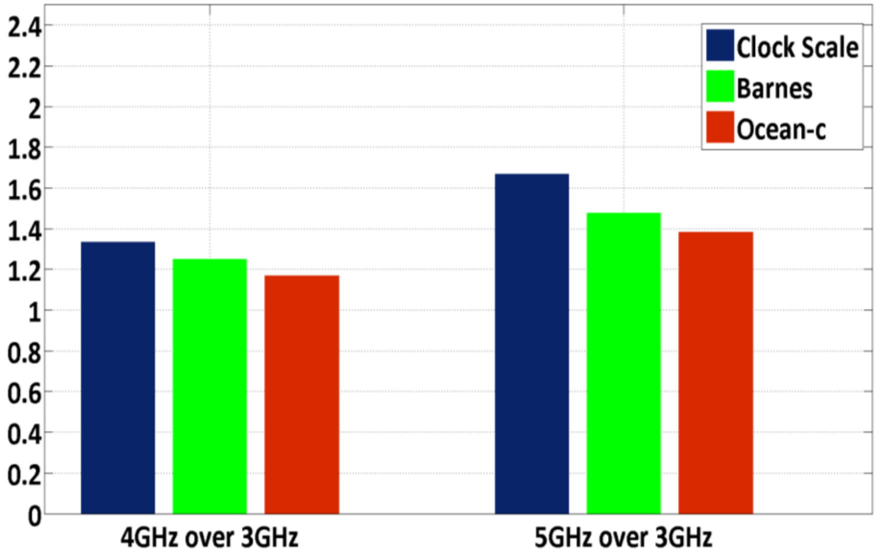
	Time	**Space**
Barnes	NlogN	N
Ocean-c	N^3	N^2

Leakage Characterization

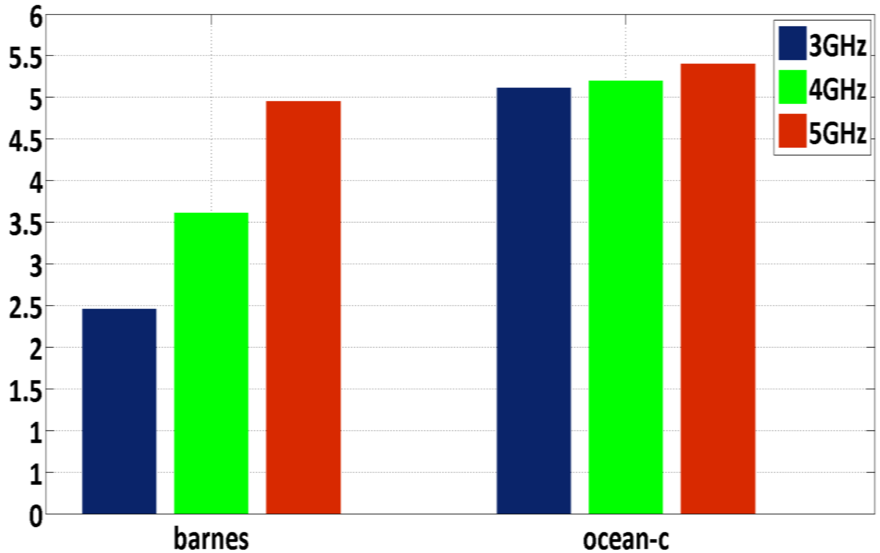




Performance Speedup and Energy Efficiency



Normalized system throughput in terms of core frequency



Energy per Instruction (nJ/inst) in terms of core frequency under optimized pin fin with input velocity 0.8m/s

Clock scaling improves overall 20% and 40% performance in 4GHz and 5GHz respectively.

- ✓ Barnes benefits more than ocean-c because of lower miss rate and fewer memory interaction

EPI degradation is detected due to the exponential relationship between leakage, temperature and supply voltage.

- ✓ Barnes suffers from 2X degradation in EPI
- ✓ Ocean has a roughly constant EPI
- ✓ The results overestimate the power as we still use the 16nm model from ITRS 2007.

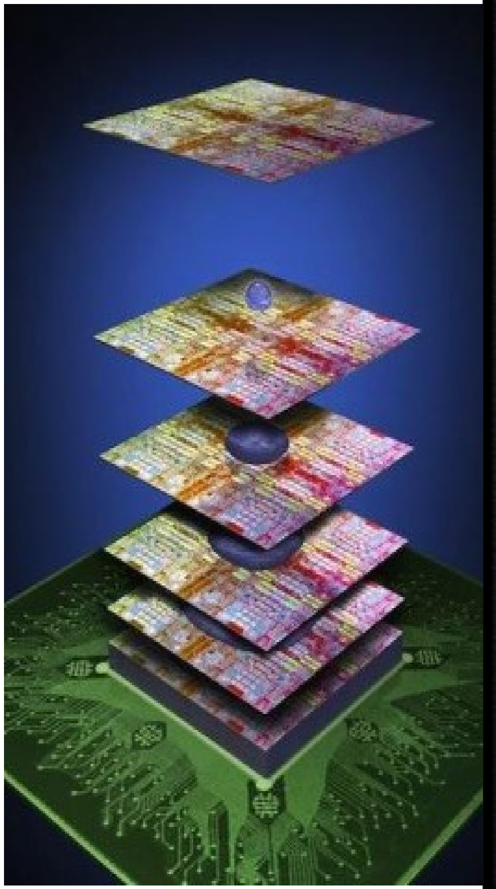
Conclusion

- The saturation point of coolant velocity is determined by the pin fin geometry, system frequency and runtime application. A typical saturation value for system within 5GHz is 0.4 m/s.

- Barnes saves 37.2% and 33.9% leakage power respectively for the baseline and optimized system running at 5GHz with input velocity from 0.1 m/s to 0.8m/s, compared to ocean-c 18.3% and 13.2%.

- The optimized pin fin can save up to 20% leakage dissipation with a same pumping power.

- The energy efficiency tends to increase with frequency for applications running at low temperatures (i.e. below 335K), as the performance speedup will compensate the increase of power.



Thank You

Questions?