New Rules: Sustaining Performance Through Extreme Scale

Sudhakar Yalamanchili

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA. 30332

Sponsors: National Science Foundation, Sandia National Laboratories, SRC

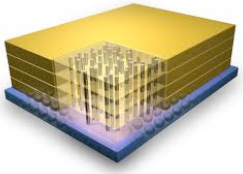
Acknowledgements - CASL Students: Haicheng Wu, William Song, He Xiao

Outline

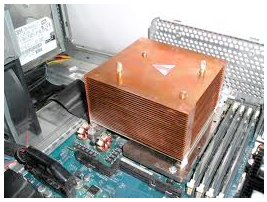
- New Rules: Consequences of Data Movement
- New Rules: Adaptation to Physical Phenomena

Scaling Computing Performance

Data Movement Costs



Thermal Limits

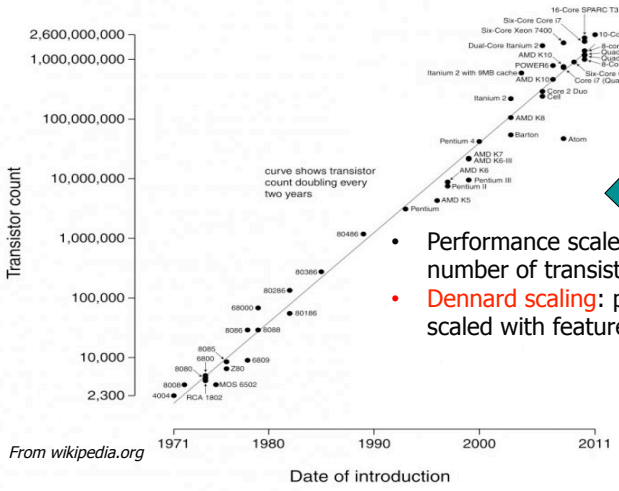


Energy Limits

Cray Titan: Heterogeneous Computing

Moore's Law

Microprocessor Transistor Counts 1971-2011 & Moore's Law



From wikipedia.org

Goal: Sustain Performance Scaling

- Performance scaled with number of transistors
- Dennard scaling: power scaled with feature size

Device or Circuit Parameter	*Scaling Factor*
Device dimension t_{ox}, L, W	$1/k$
Doping concentration N_a	k
Voltage V	$1/k$
Current I	$1/k$
Capacitance eA/t	$1/k$
Delay time per circuit VC/I	$1/k$
Power dissipation per circuit VI	$1/k^2$
Power density VI/A	1

Table I: Scaling Results for Circuit Performance (from Dennard)

From R. Dennard, et al., "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE Journal of Solid State Circuits*, vol. SC-9, no. 5, pp. 256-268, Oct. 1974.

Post Dennard Architecture Performance Scaling

Power Delivery → Cooling

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\frac{ops}{joule}\right)$$

W. J. Dally, Keynote IITC 2012

Data_movement_cost

Three operands x 64 bits/operand

Moving 1-bit of data 1mm at 22nm[1] = ~1 pj

$$Energy = \#bits \times dist - mm \times energy - bit - mm$$

[1]HIPEAC Roadmap 2012 – 2012-9-hipeacvision.pdf

Scaling Performance: Cost of Data Movement

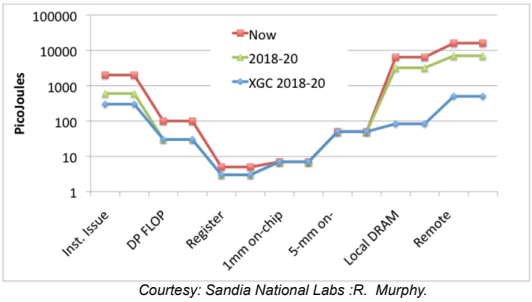
Embedded Platforms

Big Science: To Exascale

Cost of Data Movement

Goal: 1-100 GOps/w

Goal: 20MW/Exaflop

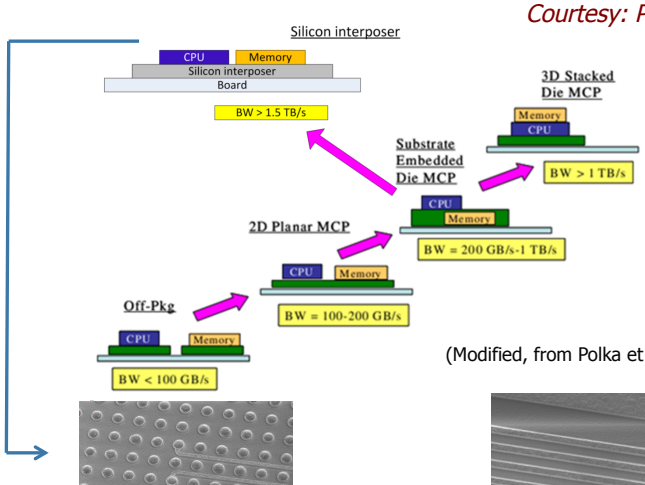


Courtesy: Sandia National Labs :R. Murphy.

- Sustain performance scaling through massive concurrency
- Data movement becomes more expensive than computation

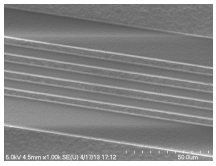
Bandwidth and Energy Tapers for Dense Interconnections

Courtesy: Professor M. Bakir (ECE) Georgia Tech



(Modified, from Polka et al, Intel Technology Journal)

Microbump array (25 μm diameter, 50 μm pitch)



Fine-pitch wires (2 μm width, 4 μm pitch)

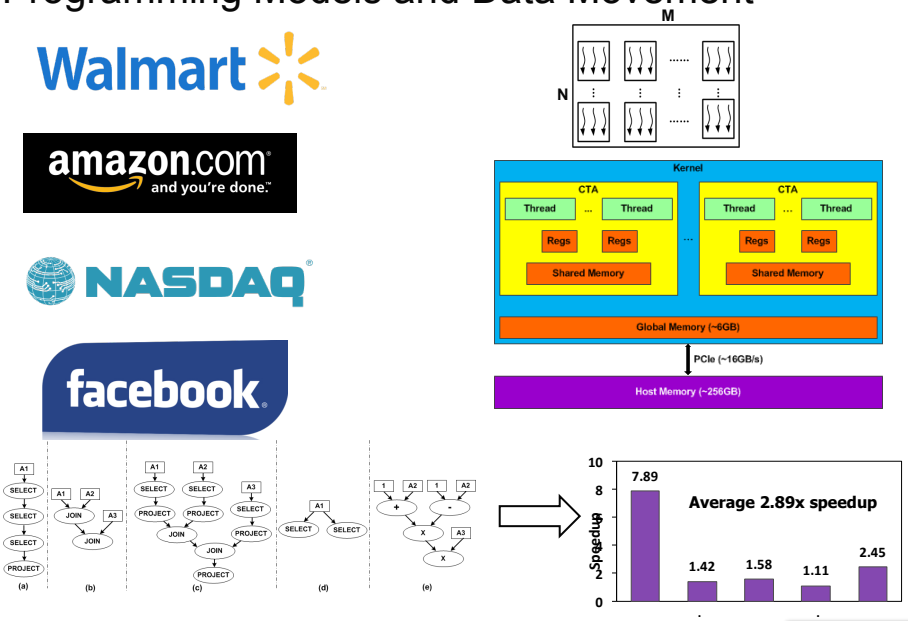
Data Movement Energy Costs

$$Energy = \underbrace{\#\,bits}_{\text{Algorithms/Applications}} \times \underbrace{dist - mm}_{\text{Architecture}} \times \underbrace{energy - bit - mm}_{\text{Technology}}$$

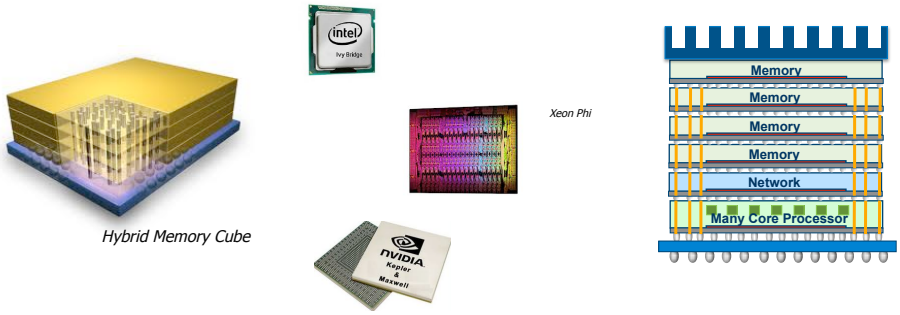
You can hide latency but you cannot hide energy!
Or can you?

- Modern Architectures are designed to optimize compute

Programming Models and Data Movement



Refactor Systems



- Interleaving computer, communication, and storage
- Beware the bisection bandwidth trap
- Minimize data movement → Processor in Memory?
- Programming models

Post Dennard Architecture Performance Scaling

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\underbrace{\frac{ops}{joule}}\right)$$

W. J. Dally, Keynote IITC 2012

Operator_cost + *Data_movement_cost*

Specialization → heterogeneity and asymmetry

Three operands x 64 bits/operand

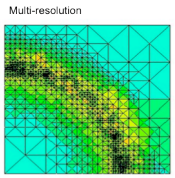
$$Energy = \#bits \times dist - mm \times energy - bit - mm$$

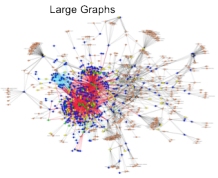
Outline

- New Rules: Consequences of Data Movement
- New Rules: Adaptation to Physical Phenomena

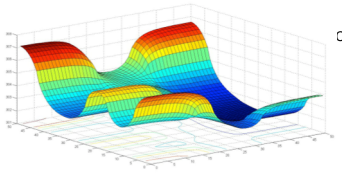
It's a Physical World: Co-Optimization is Key

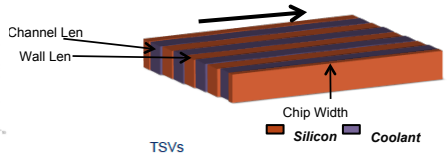
Next Generation Applications



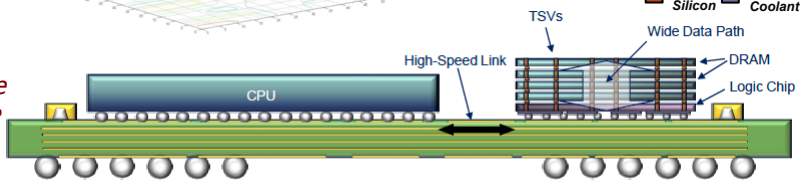


Multi-scale Physical Phenomena



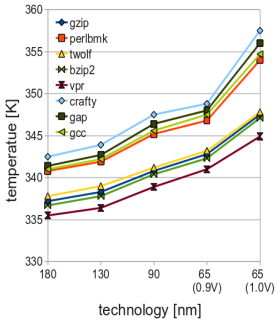


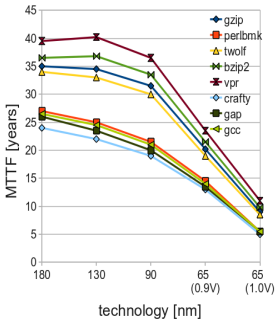
Architecture & Package



Degradation Variation

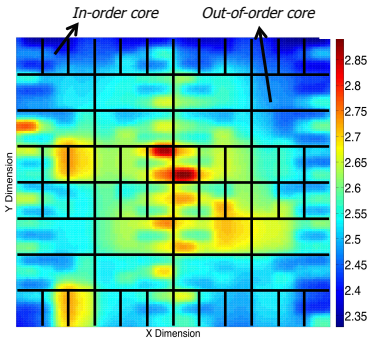
Single-core processor lifetime reliability





Peak temperature and MTTF analysis from J. Srinivasan et al., "Lifetime Reliability: Toward An Architectural Solution," Micro 2005.

Multicore processor lifetime reliability

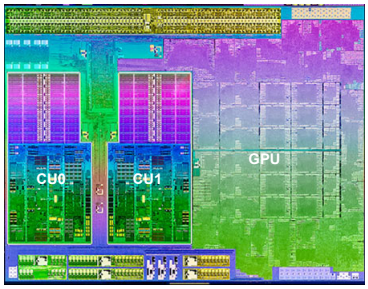


64-core asymmetric chip multiprocessor layout and failure probability distribution

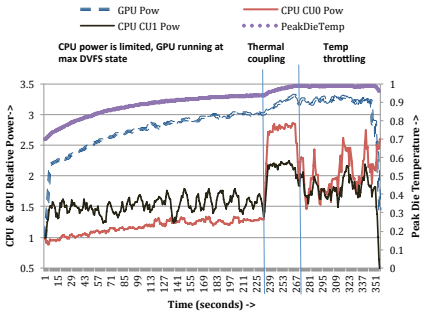
25% peak-to-peak difference of failure distribution across the processor die; induced by architectural asymmetry, thermal coupling, power management, and workload characteristics

Thermal Coupling

Paul, Manne, Bircher, Arora, Yalamanchili (ISCA 2013)

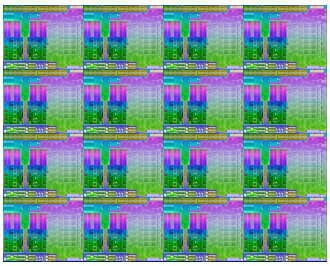


AMD Trinity APU

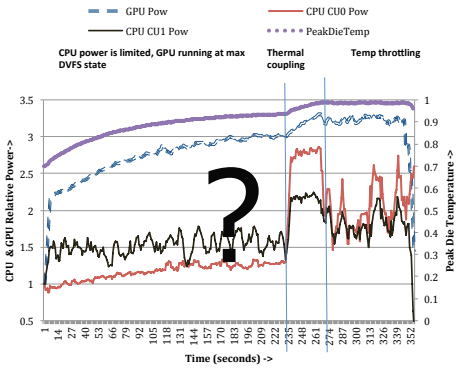


- Thermal coupling between CPU and GPU accelerates temperature rise
- Performance coupling between CPU and GPU impacts power management! ⇨ Interconnect and Memory!

Thermal Coupling



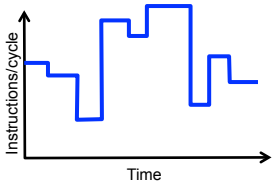
64 CPU cores, 6144 CPU cores



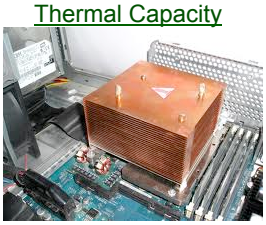
- Thermal impact?
- Power management?
- Asymmetric network bandwidth demand?

Adaptation to the Physics

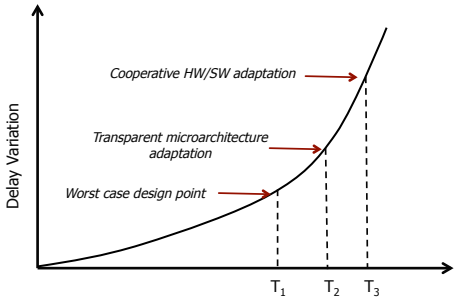
Time Varying Workload



Thermal Capacity

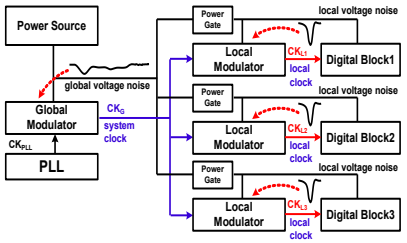


Adaptive Microarchitecture



Adaptive Clocking

Courtesy S. Mukhopadhyay

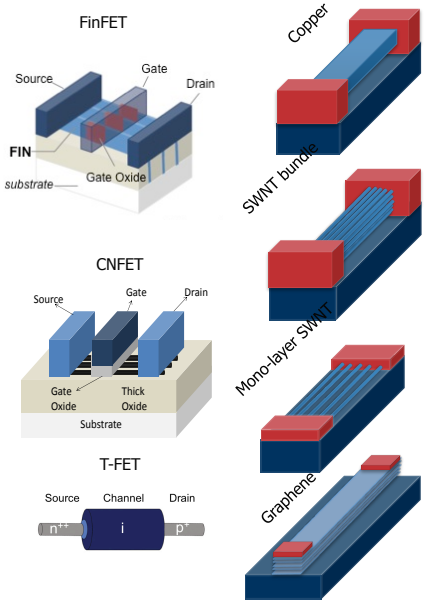


Modulate the arrival time of clock at the flip-flops to track the variation in the critical path delay of a logic block

Data Movement Energy Costs

$$Energy = \underbrace{\#\,bits}_{\text{Algorithms/Applications}} \times \underbrace{dist - mm}_{\text{Architecture}} \times \underbrace{energy - bit - mm}_{\text{Technology}}$$

Emerging Electrical Switches & Interconnects

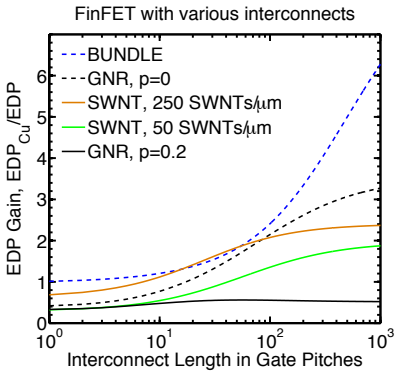


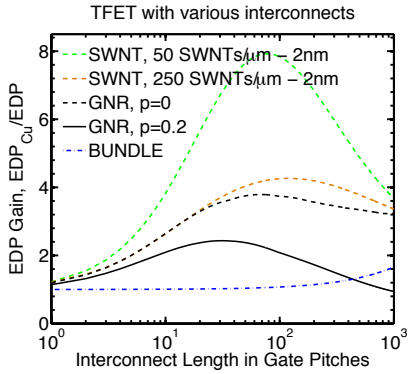
Prof. A Naemi, ECE, GT
http://users.ece.gatech.edu/~azad/

- A host of novel devices and interconnects are being pursued.
- Various devices offer vastly different characteristics in terms of output current, input capacitance, subthreshold swing, etc.
- Emerging interconnects such as CNT and graphene may offer lower interconnect capacitances but are mostly more resistive.
- The speed, energy and EDP advantages offered by novel interconnects will highly depend on the transistors used.

Interconnect EDP Comparisons (16nm ITRS Node)

Prof. A Naemi, ECE, GT
http://users.ece.gatech.edu/~azad/

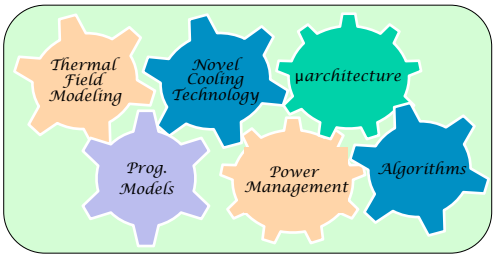




For low power devices such as TFETs, interconnect capacitance is the most important parameter since interconnect resistance is dominated by transistors.

For high performance devices, resistance and capacitance are equally important.

Research Needs: Co-Design

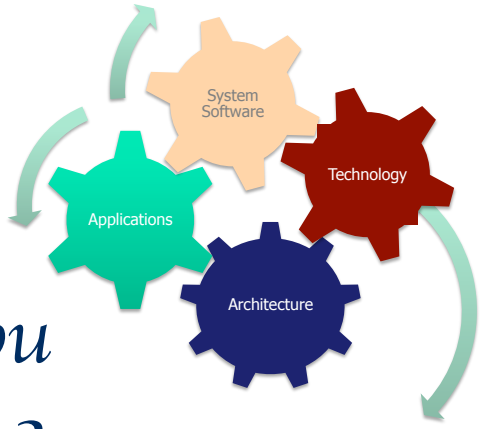


$$Energy = \#bits \times dist - mm \times energy - bit - mm$$

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\frac{ops}{joule}\right)$$

- Look for system level multipliers of communication

Scaling Rules ⇨



Thank You

Questions?

Acknowledgements – CASL Students: Haicheng Wu, William Song,