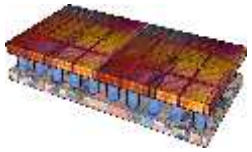
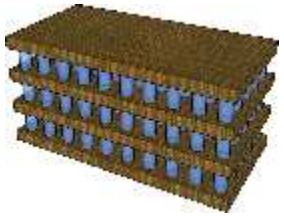
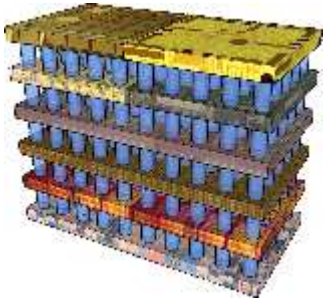
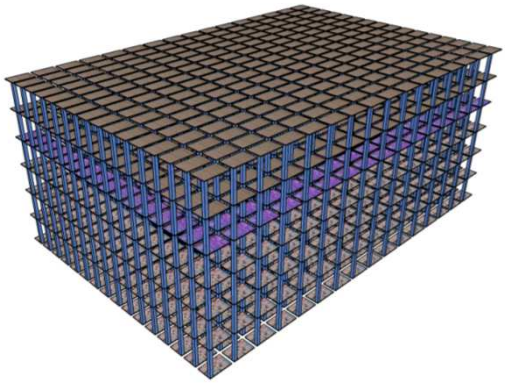
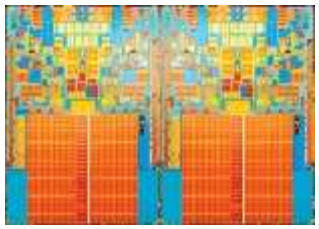
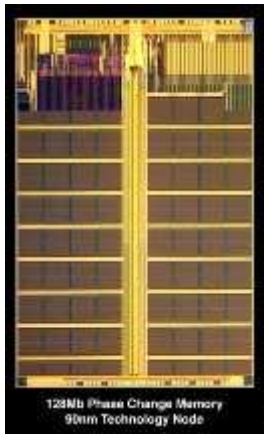
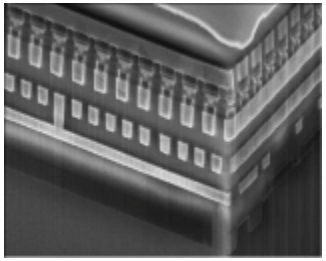







CERCS IAB Workshop, April 26, 2010



Peeling the Power Onion

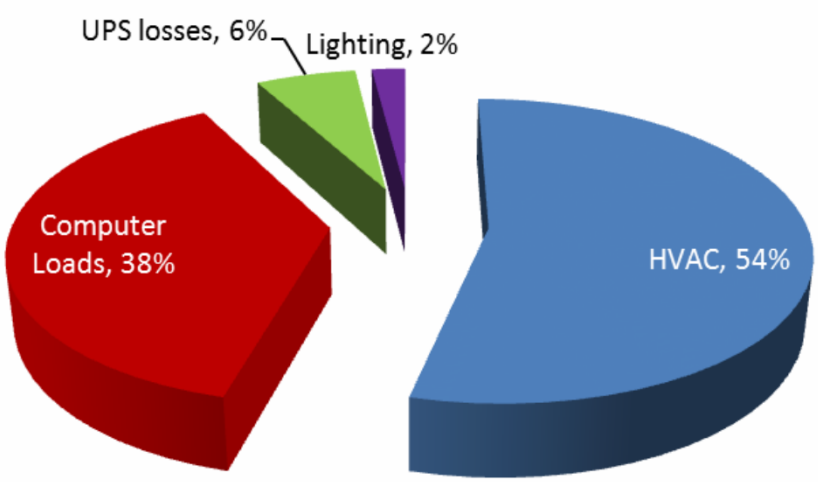
Hsien-Hsin S. Lee
Associate Professor
Electrical & Computer Engineering
Georgia Tech

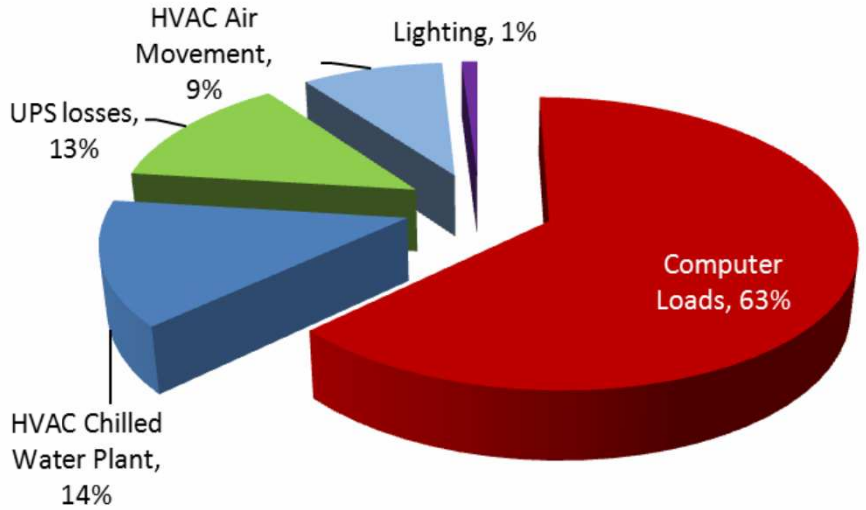
cercs

Georgia Institute of Technology

Power Allocation for Server Farm Room



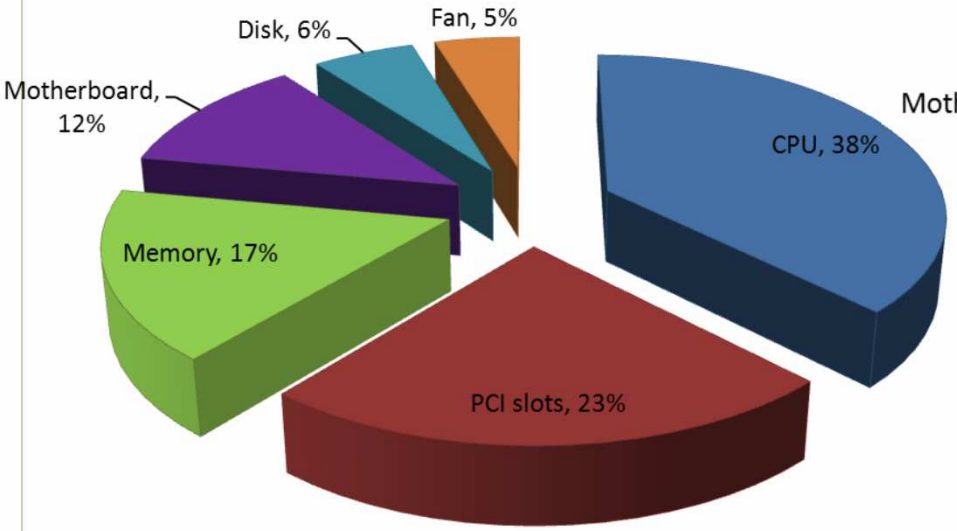
Datacenter 8.1
Total Power = 580 kW



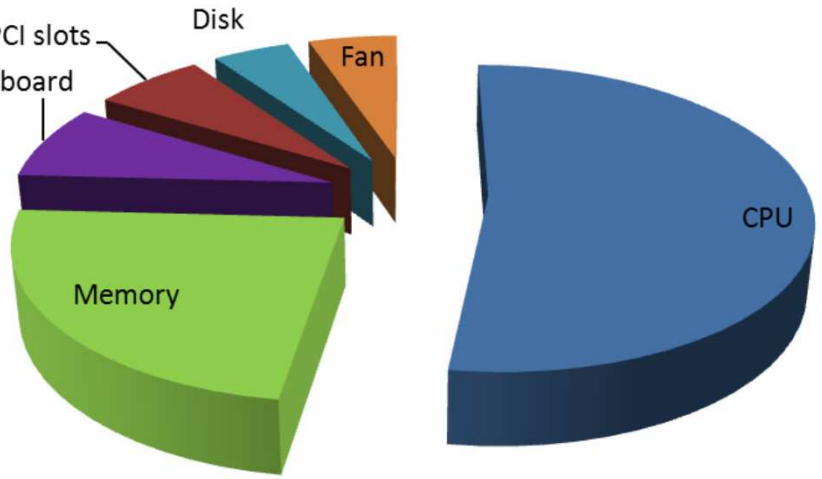
Datacenter 8.2
Total Power = 1700 kW

- Data from Lawrence Berkeley National Lab [2002]

Per-System Power Breakdown



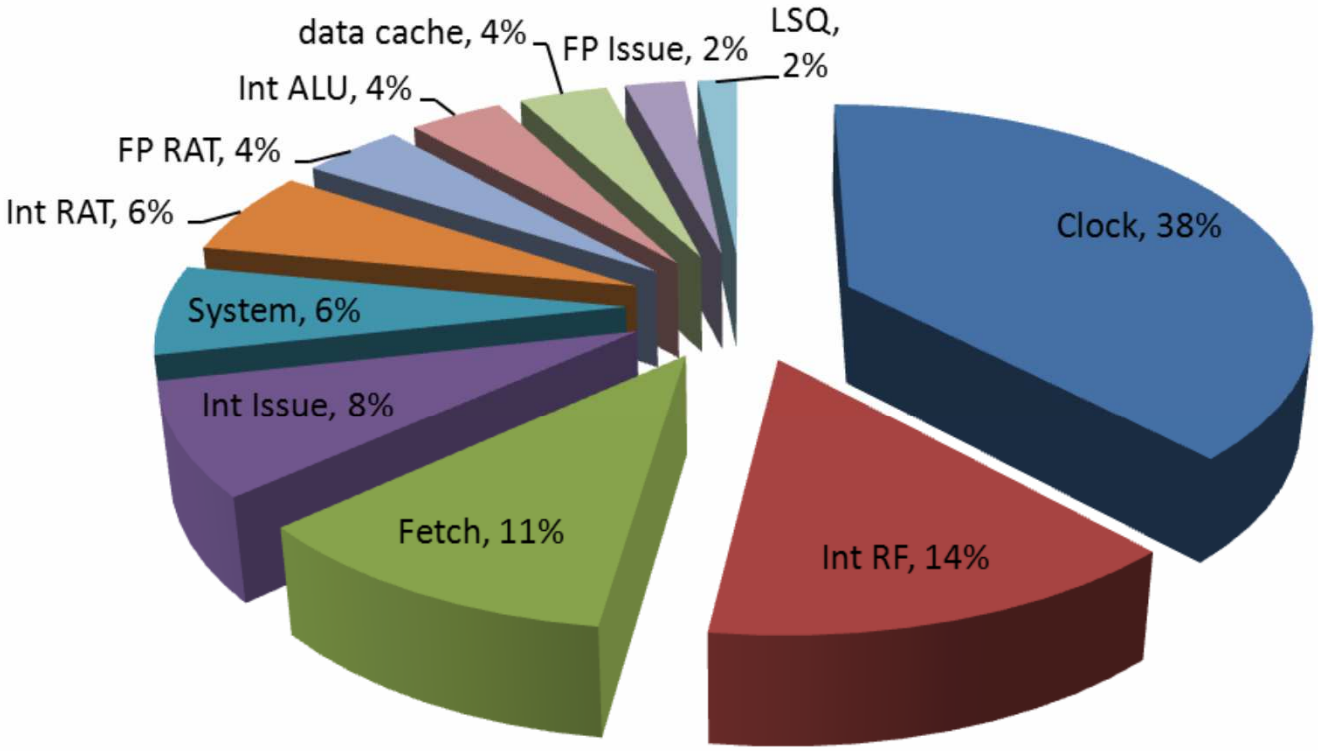
Nameplate Peak power breakdown



Likely actual power breakdown

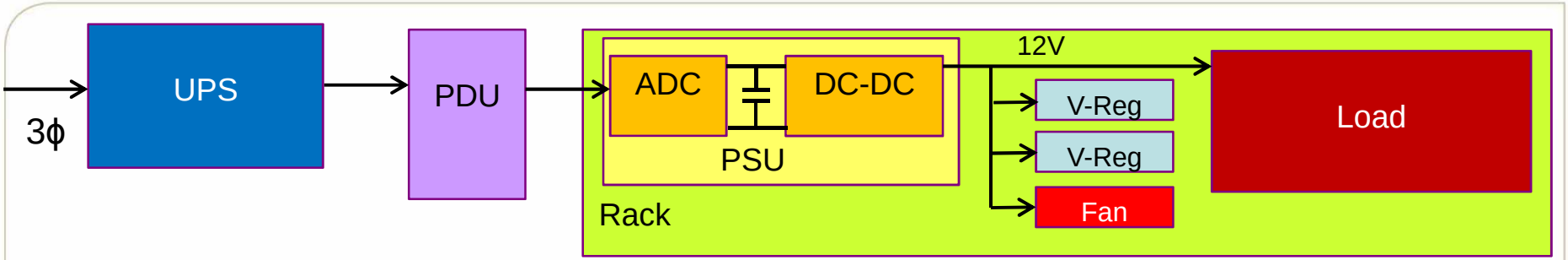
- Peak component power breakdown [ISCA-34, '07]
  - Two Intel CPUs, 2 PCI, 1 IDE disk, 4 DDR1 DIMMs
- Total = 231W, Measured = 145W
- Main variation from CPU and memory, other components correlate well with CPU activity

Per-CPU Power Breakdown



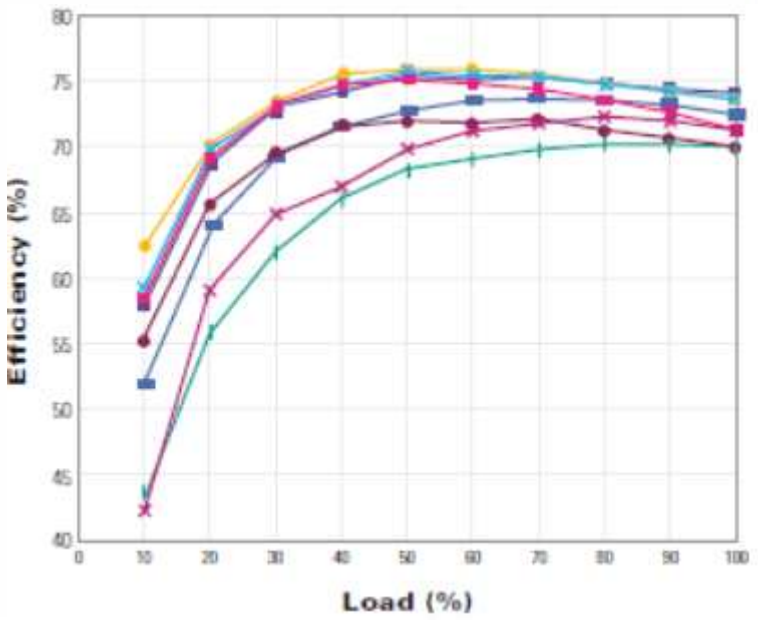
- Simulation results by UT researchers [ISLPED-03]
- Modern scenarios (requiring quantification) are
  - Multiple cores, More power aware
  - Aggressive gating, DVFS techniques
  - Larger and deeper caches

Efficiency of Power Delivery Infrastructure



Baseline 88% X 93% X 79% X 75% = **48%**

- Power conversion lost efficiency
- Supply design for fault tolerance with backup, decreasing the efficiency
- Room for improvement
  - Pay more for higher efficiency PSU (care TCO instead of the equipment price)
  - Consolidating power conversion at a higher level (reduce redundancy)



Source: datacenter journal & Rocky Mountain Institute

cercs Georgia Institute of Technology

Bottom Line

- Every **WATT** counts

- Cost of power infrastructure = $10~$20/W [Turner, uptimes'06]
- Energy cost = 10 ¢ / kW-h
- 10 years of energy bill ~ $10/W [Fan, ISC



Hierarchical Power Optimization

- Device/Circuits level
- Microarchitectural level
- Chip/Component level
- System level
- Compiler/Algorithmic level



Hierarchical Power Optimization

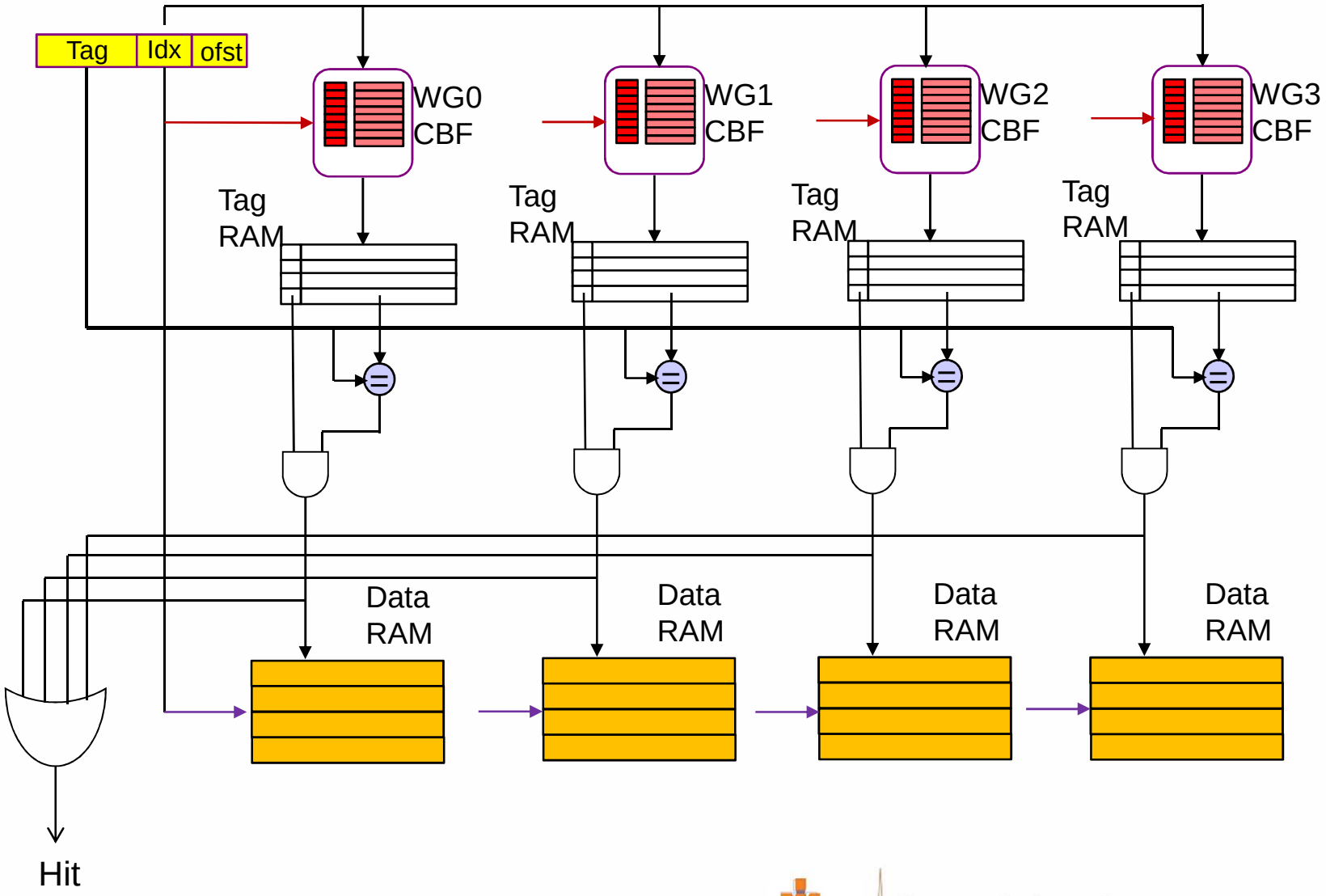
- Device/Circuits level
- Microarchitectural level
- Chip/Component level
- System level
- Compiler/Algorithmic level



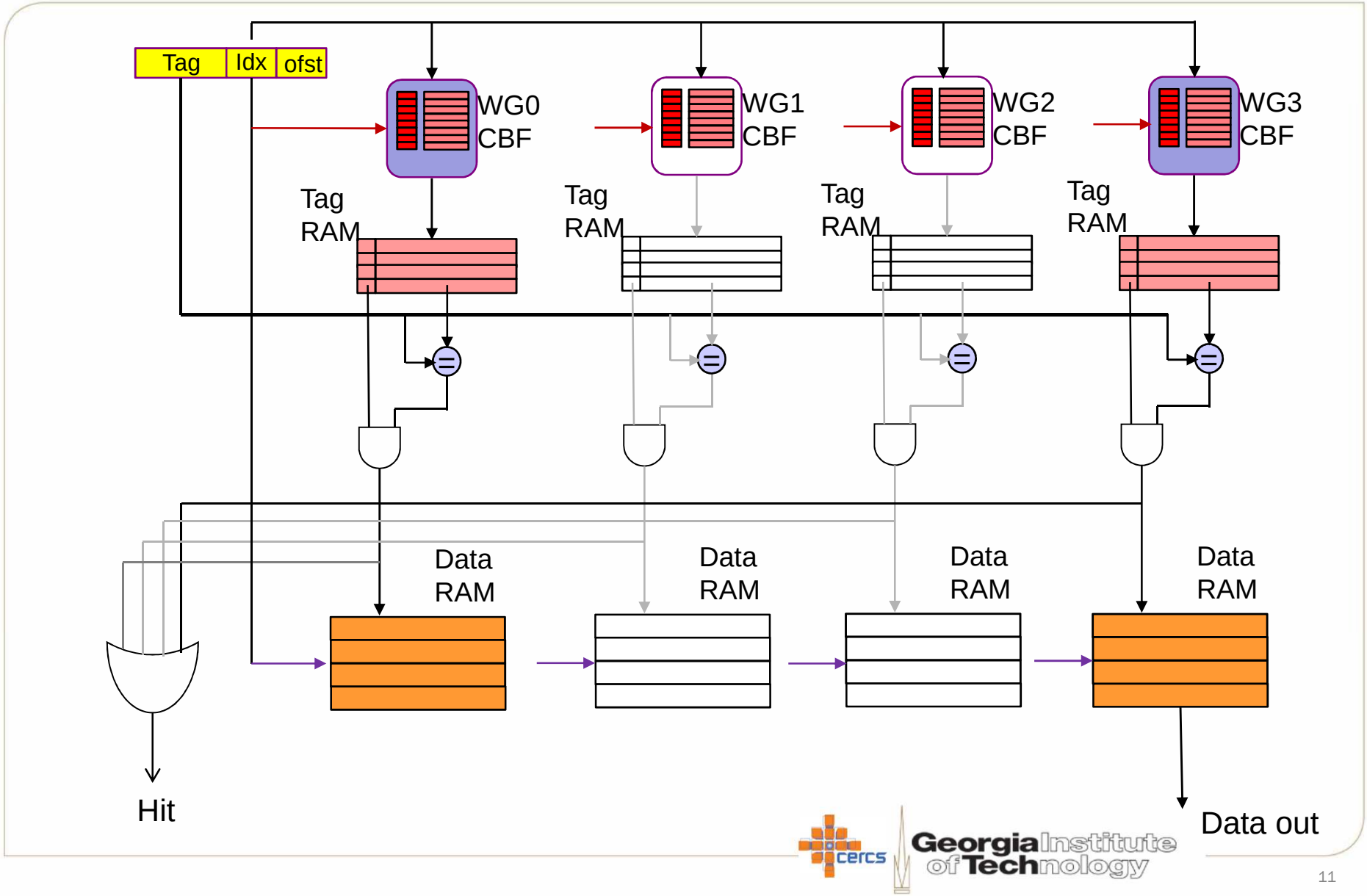
Micro-Arch. Level Power Optimization

- Utility-based Fine-grained clocking gating
- On-chip cache optimized design
  - (sub)-Banking : encapsulated computing
  - Vertical cache partition: Filter cache or Multi-level cache
  - Horizontal cache partition
    - Multi-lateral cache
    - Segmented design
    - Region-based caching
  - Lookup filtering
  - Snoop filtering

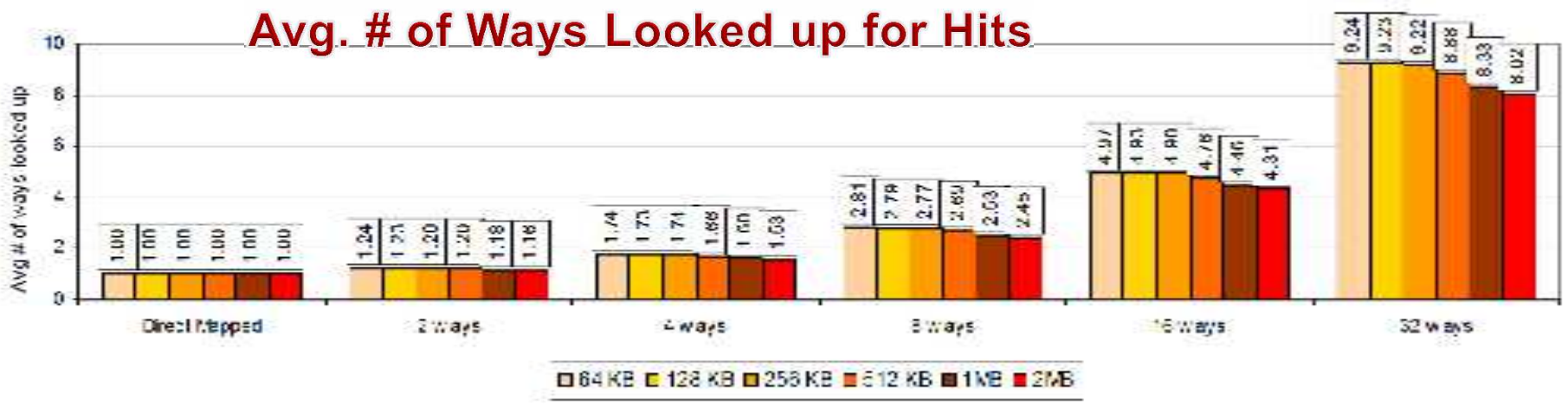
Way Guard Cache w/ Bloom Filter [Ghosh et al. ISLPED'09]

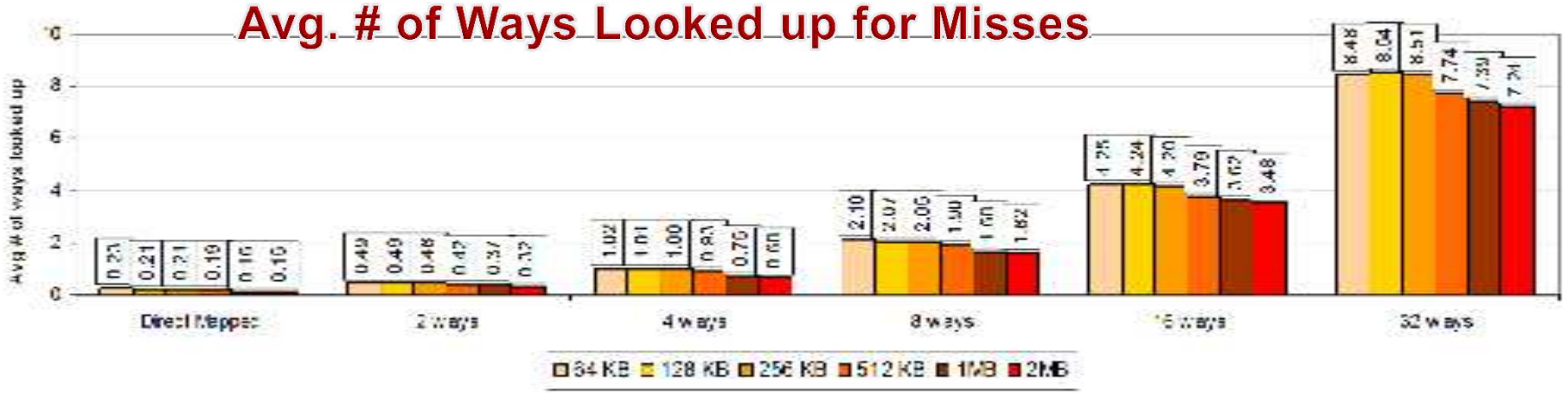


Way Guard Cache w/ Bloom Filter [Ghosh et al. ISLPED'09]



Average Number of Cache Ways Looked Up

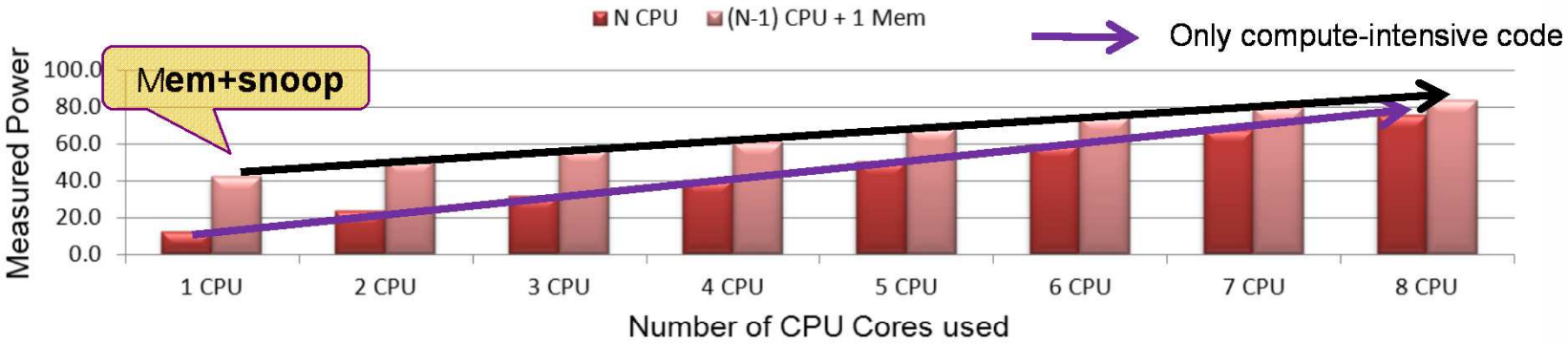




cercs Georgia Institute of Technology

Cache Snooping Power

- Snoop Power can be expensive: **~20W** seen between "Xeon" and "Core i7"
- 2-socket, 4-core Xeon Power below



- Snoop Filtering techniques for [Ballapuram, Sharif, and Lee, ASPLOS'08]
  - Self-Modifying Code: Filtered out by a Bloom Filter
  - Stack Accesses
    - Filtered out by **S**tack bit
    - Annotated **S**tack bit at decode stage based on base-register ID
  - Non-Stack Accesses
    - Track (Modified/Exclusive) state vs. (Shared) state
    - Use Bloom filter to filter out unnecessary traffic

Data: Hrishikesh Amur, Bhavani Krishnan, Eric Fontaine



Hierarchical Power Optimization

- Device/Circuits level
- Microarchitectural level
- Chip/Component level
- System level
- Compiler/Algorithmic level

Many-Core Fad

- More to be crammed on a single die, Thanks to Moore's Law
- Energy efficient or Energy inefficient?

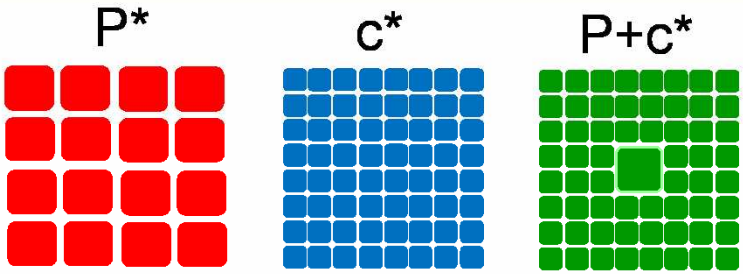
Area-Performance Scalability
- How many cores fit on a die?

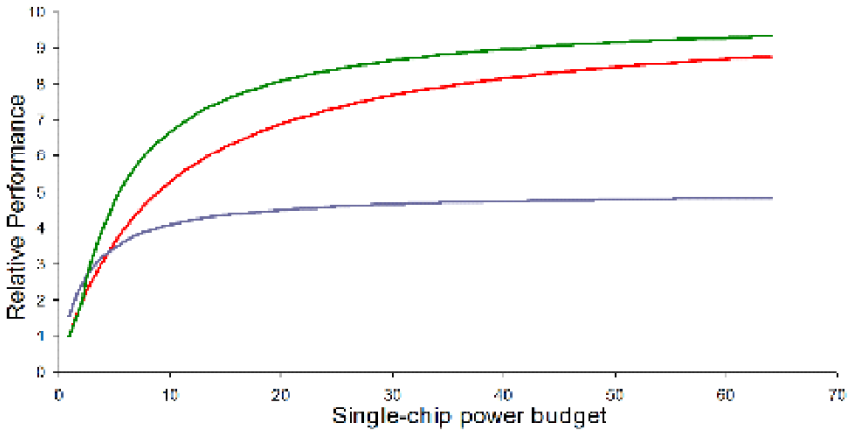
Energy Scalability
- More joules = more performance?
- Proportional energy effectiveness

Power Scalability
- Power budget ~ Cooling facility
- How much
- Cost for Cloud services

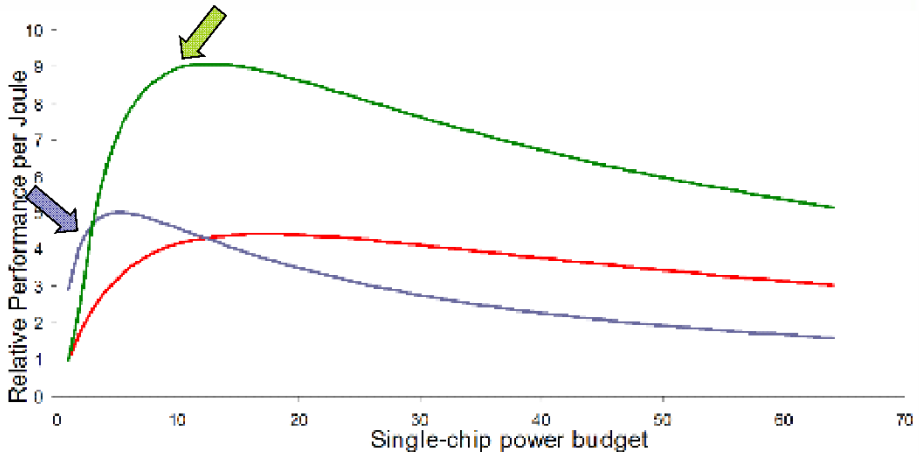
Design Space [Woo & Lee, Computer'08]

Multi-core / Many-core Scaling Alternatives





Power-Equivalent "Performance"

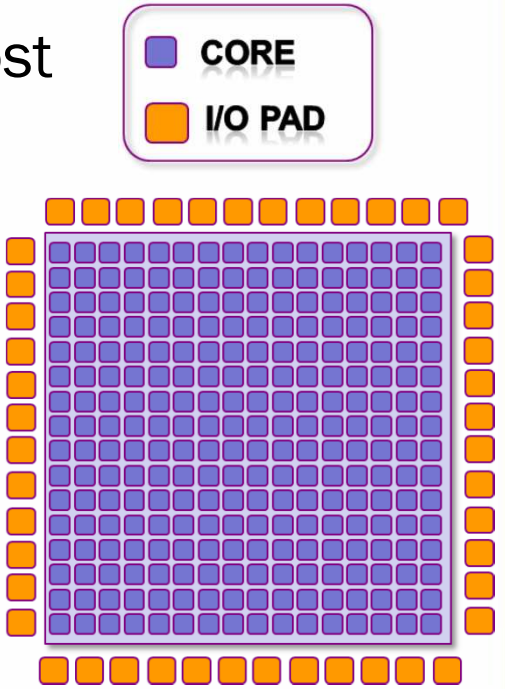


Power-Equivalent "Performance Per Joule"

f = 0.9 (parallelizable fraction)
Perf(P) = 2 x Perf(c) ; Power(P) = 4 x Power(c)

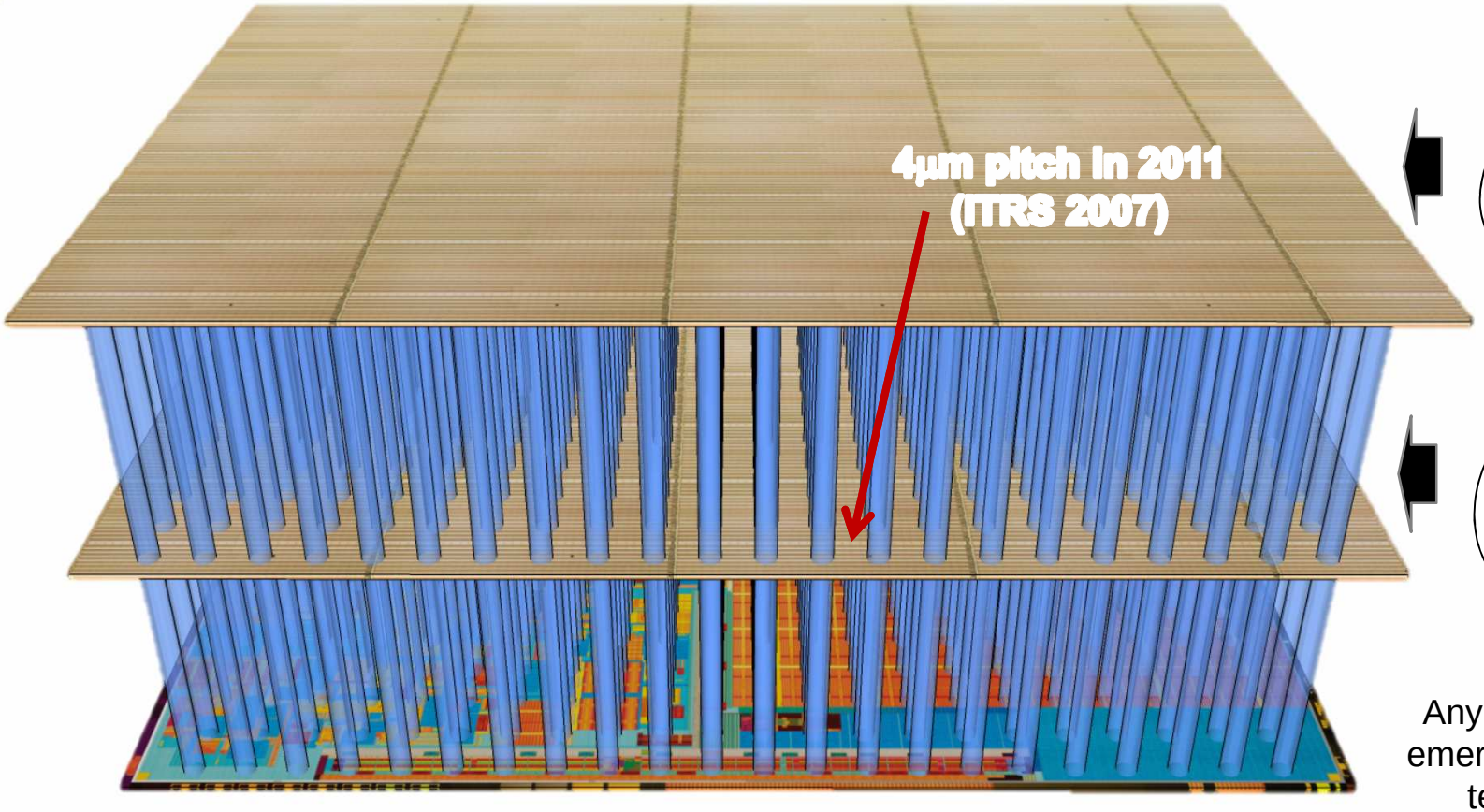
Chip Level Power Challenge for Many-Core

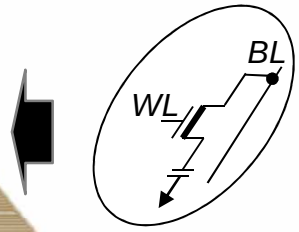
- I/O power — ITRS predicts slow growth in pin count
  - 2/3 for **Power and ground**, 1/3 for **Signal I/O**
  - Limited by physical metal properties and cost
    - DDR3 ~ 40mW per pin
    - 1024 data pins ~ **40W**
    - 4096 data pins ~ **160W**



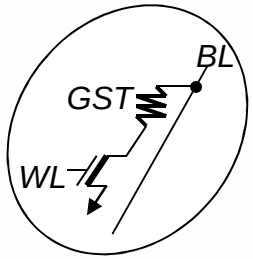
- Techniques to avoiding I/O pad power
  - Memory Locality optimization
  - Very tight Integration
    - Intel: Cores+MCH+ICH → Cores+PCH
    - 3D integrated circuits

3D Memory-Stacked Processor [Woo et al. HPCA-16]





Or



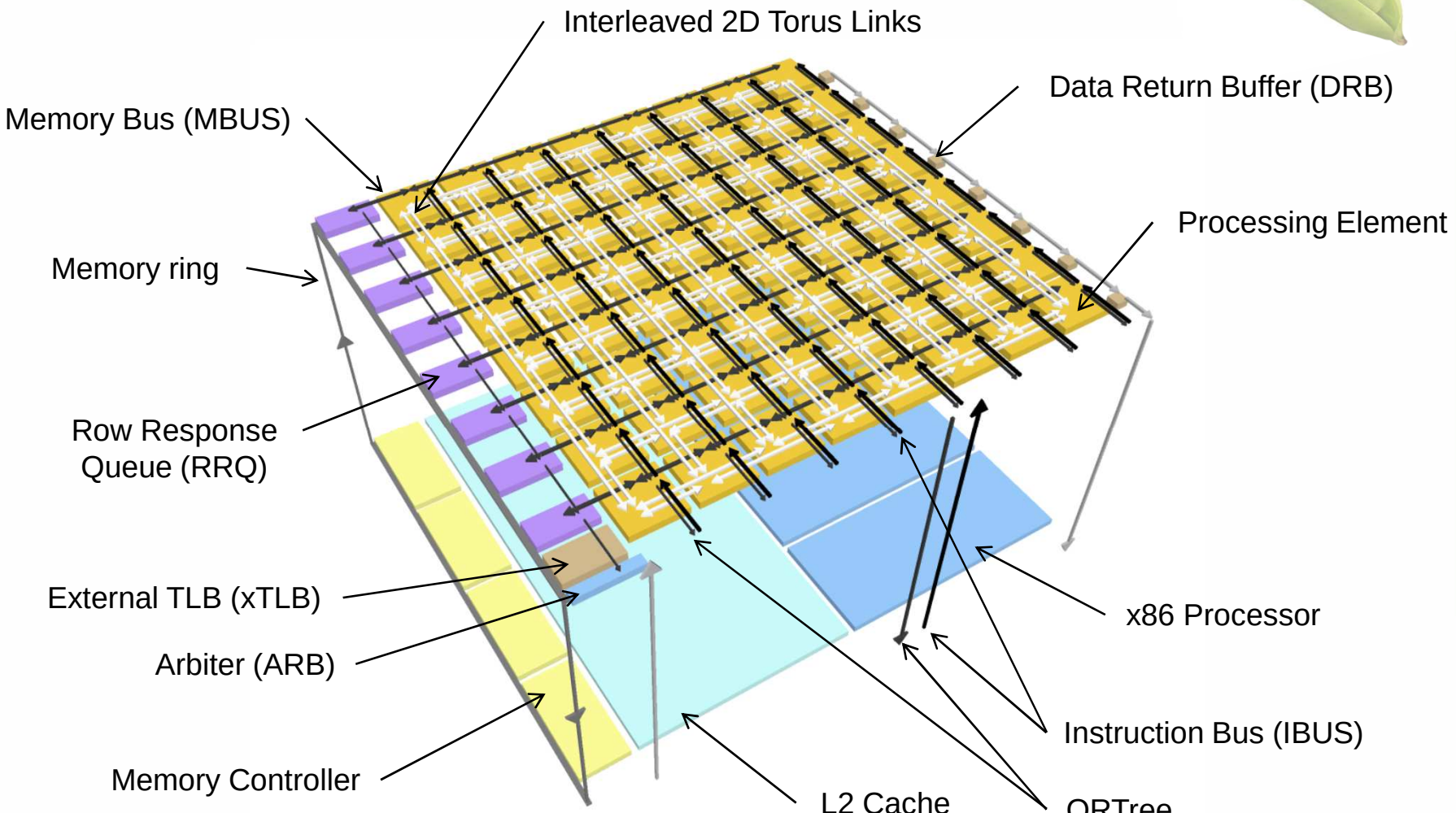
Or

Any high-density emerging memory technology

- Can reduce power in I/O pad
- Can reduce power in Interconnects on package
- Can reduce power in clock routing

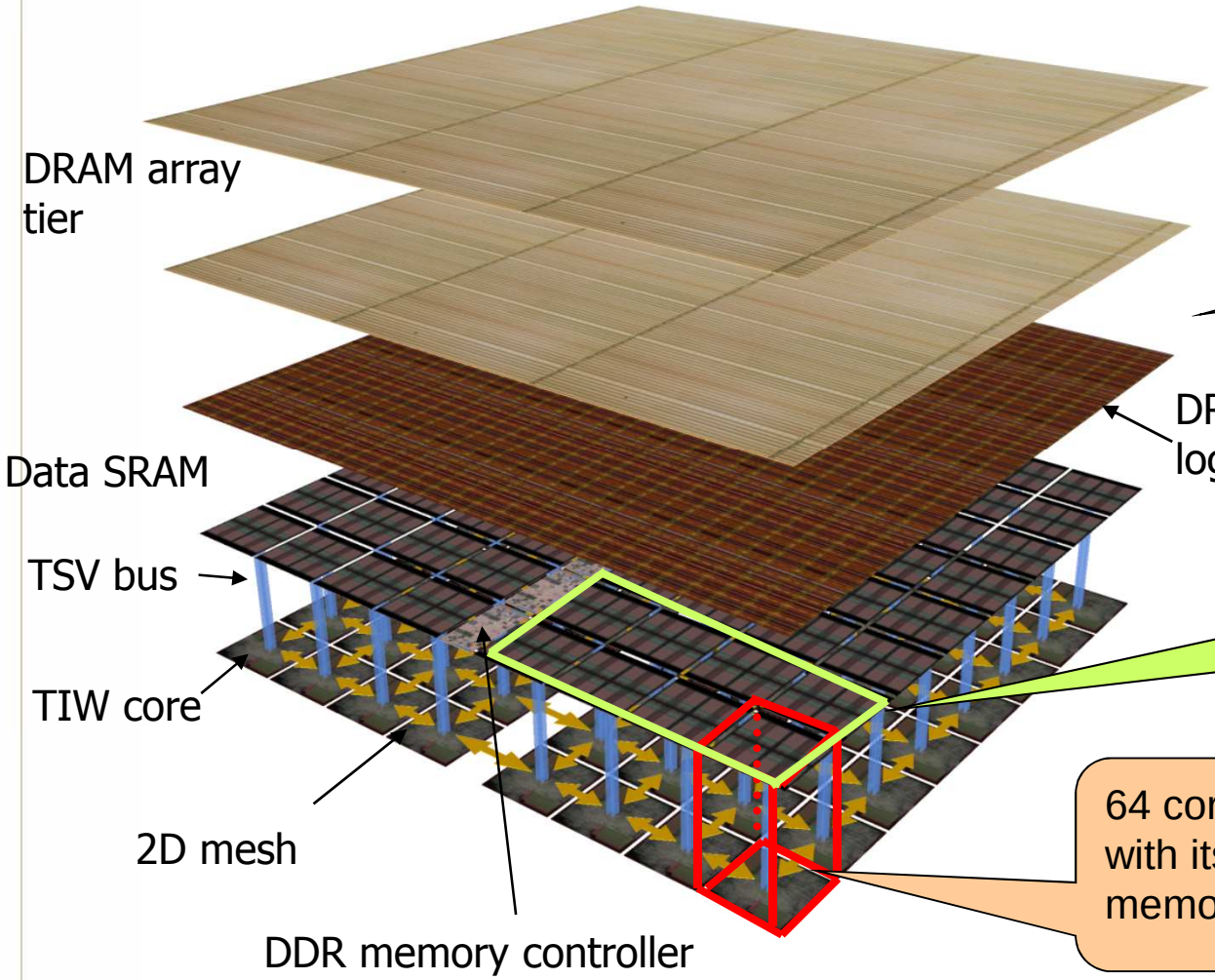
POD: 3D-Integrated Broad-Purpose Co-Processor

[Woo et al. IEEE MICRO '08]



Collaborated with Intel (Allan Knies, Josh Fryman)

3D-MAPS Many-Core Processor @GA Tech



Optimized 3D DRAM design by moving peripheral logic to one separate die tier.

8 SRAM tiles share one memory controller.

64 cores, each communicates with its private scratchpad memory tile thru TSV



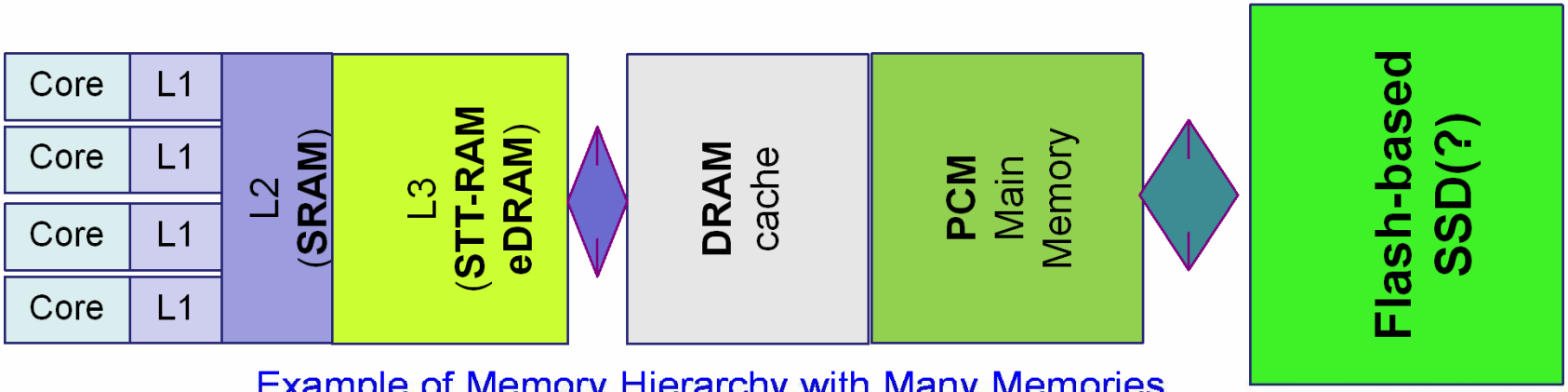
Hierarchical Power Optimization

- Device/Circuits level
- Microarchitectural level
- Chip/Component level
- System level
- Compiler/Algorithmic level

cercs Georgia Institute of Technology

Memory Subsystem Level Power Strategy

Memory Type	DRAM	6T SRAM	NOR FLASH	PCM	MRAM	STT-RAM	FeRAM	Memristor
Volatility	Volatile	Volatile	NVM	NVM	NVM	NVM	NVM	NVM
Cell Size (F^2)	6 - 12	50 -80	7 - 11	5 - 8	16-40	6-20	Large	scalable
Read	Destructive	Partial Destructive	Non-Destructive	Non-Destructive	Non-Destructive	Non-Destructive	Destructive	Non-Destructive
Erase Granularity	Direct	Direct	Block	Direct	Direct	Direct	Direct	Direct
Write/Erase/Read Time	50ns/50ns/50ns	8ns/8ns/8ns	1μs/1-100ms/60ns	10ns/50ns/20ns	30ns/30ns/30ns	20ns/20ns/20ns	80ns/80ns/80ns	??
Programming Energy	Medium	Medium	High	Medium	Medium	Low	Medium	Low?
Write/Read Endurance	~∞/~∞	~∞/~∞	10^6/~∞	10^8~10^{12}/~∞	10^{12}/10^{12}	10^{15}/10^{15}(?)	10^{12}/10^{12}	10^7/?
Multi-Level Cell	No	No	Yes	Yes	stacking	stacking	No	stacking
Cost per bit	Low	High	Medium	Low	??	??	High	??
Supply Voltage	3V	<1V	6-8V	1.5-3V	3V	<1.5V	2-3V	<1.5V?



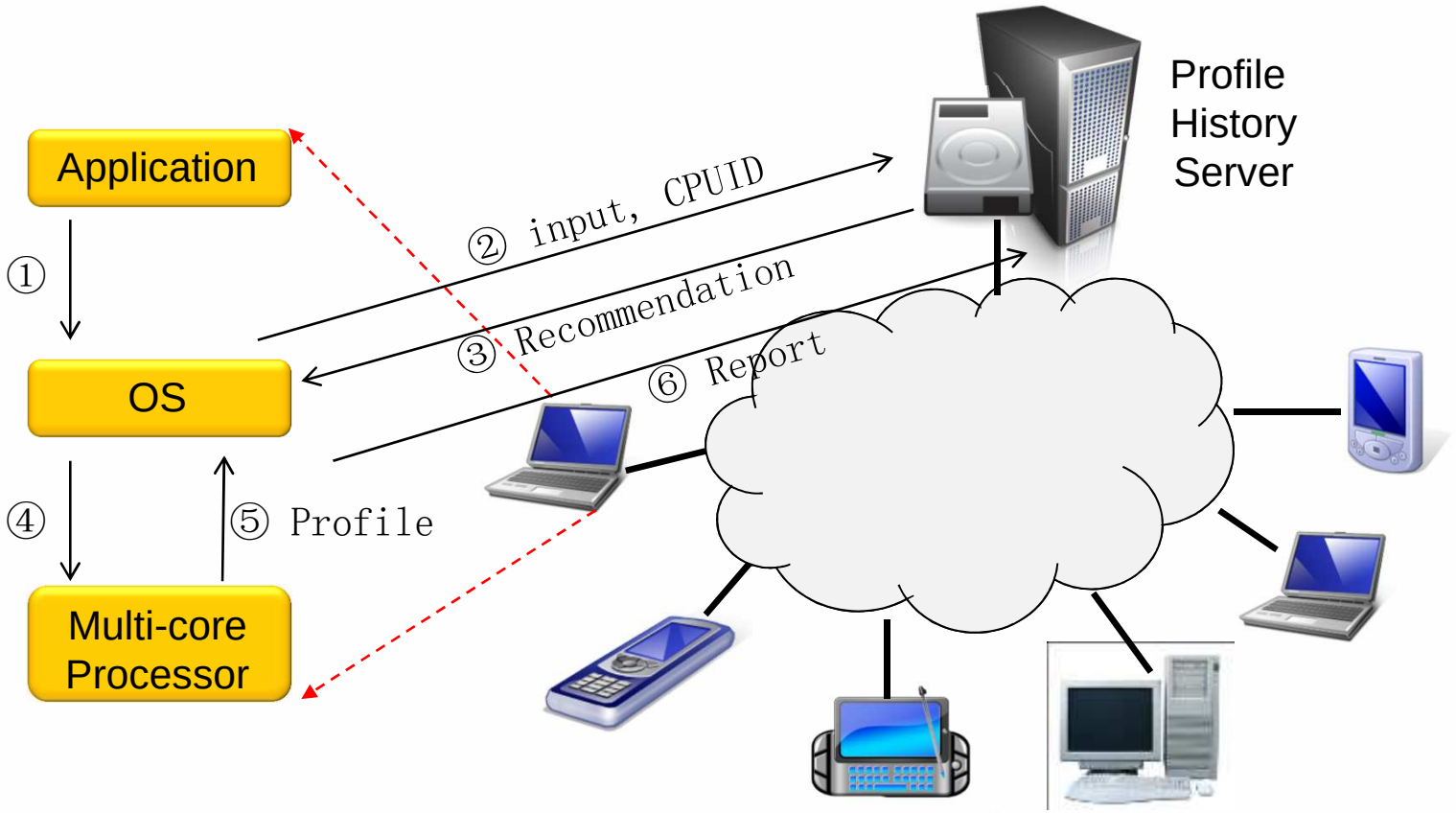
Example of Memory Hierarchy with Many Memories

- Many Memories, Many Hierarchies
- Co-optimize: Density, Power, Speed, and Reliability
- No "one-size-fits-all"

System Level Power Provisioning (I)

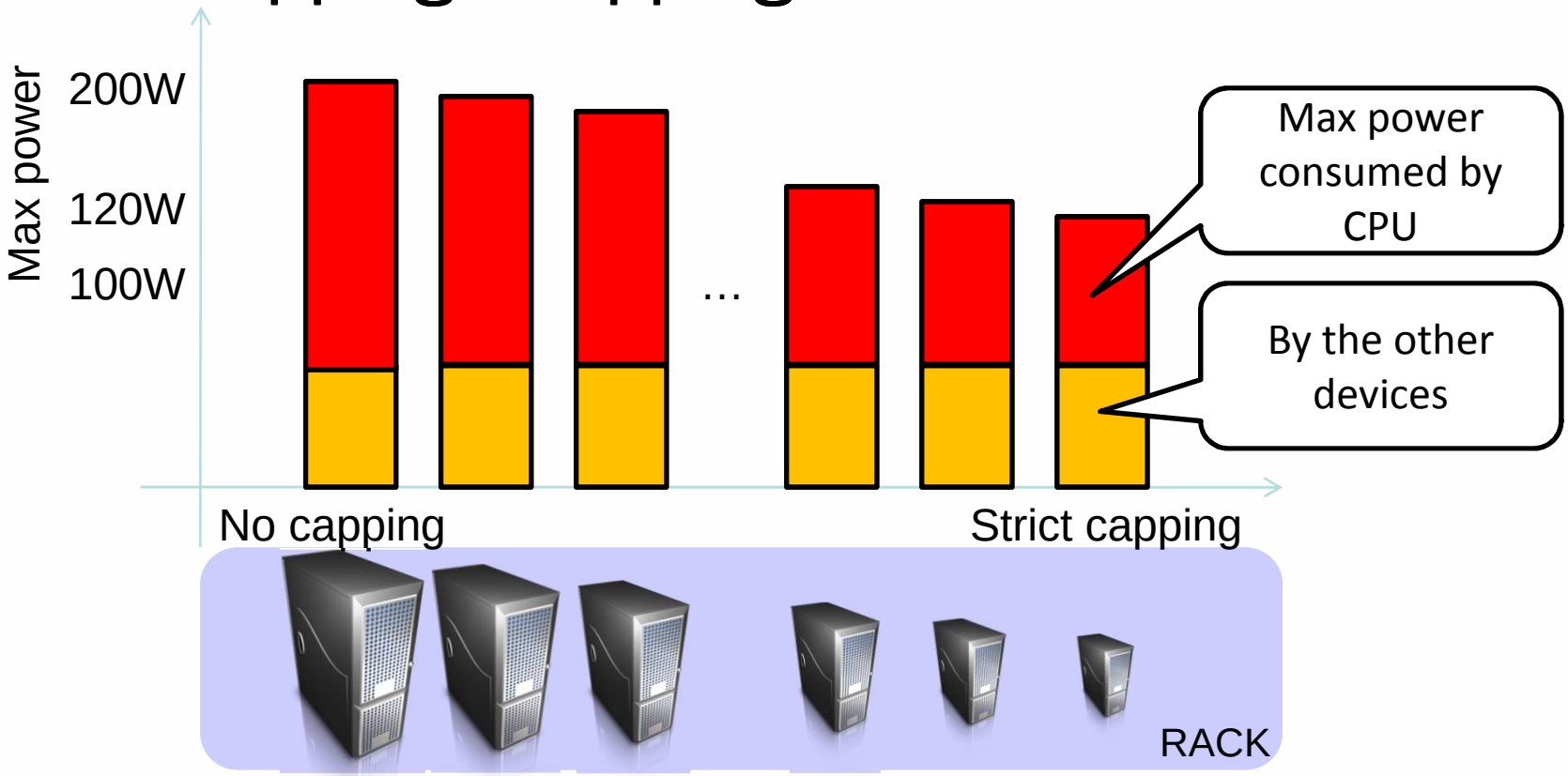
[Woo & Lee, ACM OS Review, 2009]

- PROPHET: PROvisioning Processing in a Heterogeneous EnvironmenT
- Use history to guide best-power-efficiency execution for a given application
  - Netflix, youtube, Hulu.com, etc.



System Level Power Provisioning (II)

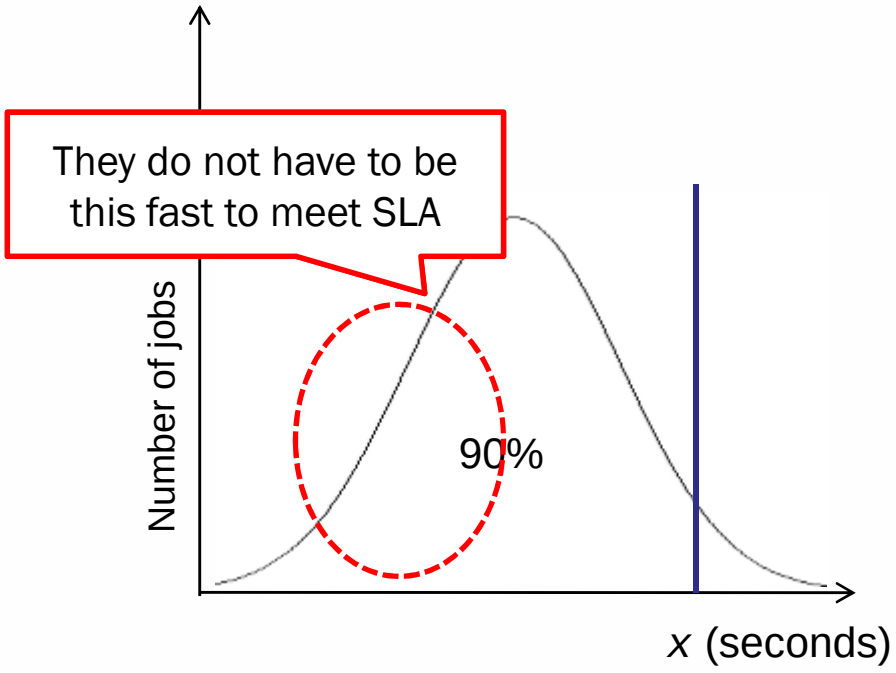
Power-Capping Stepping



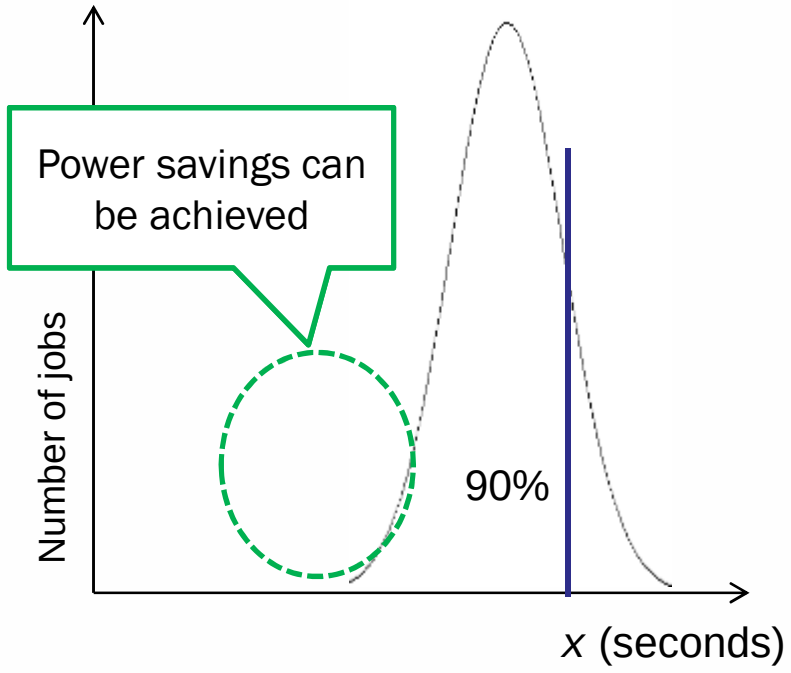
- Current work: to provision jobs to different capped servers based on power-use prediction
- Lower the overall infrastructure cost

System Level Power Provisioning (II)

- Schedule jobs
  - to a specific power-capped machine
  - to just meet performance requirement of SLA

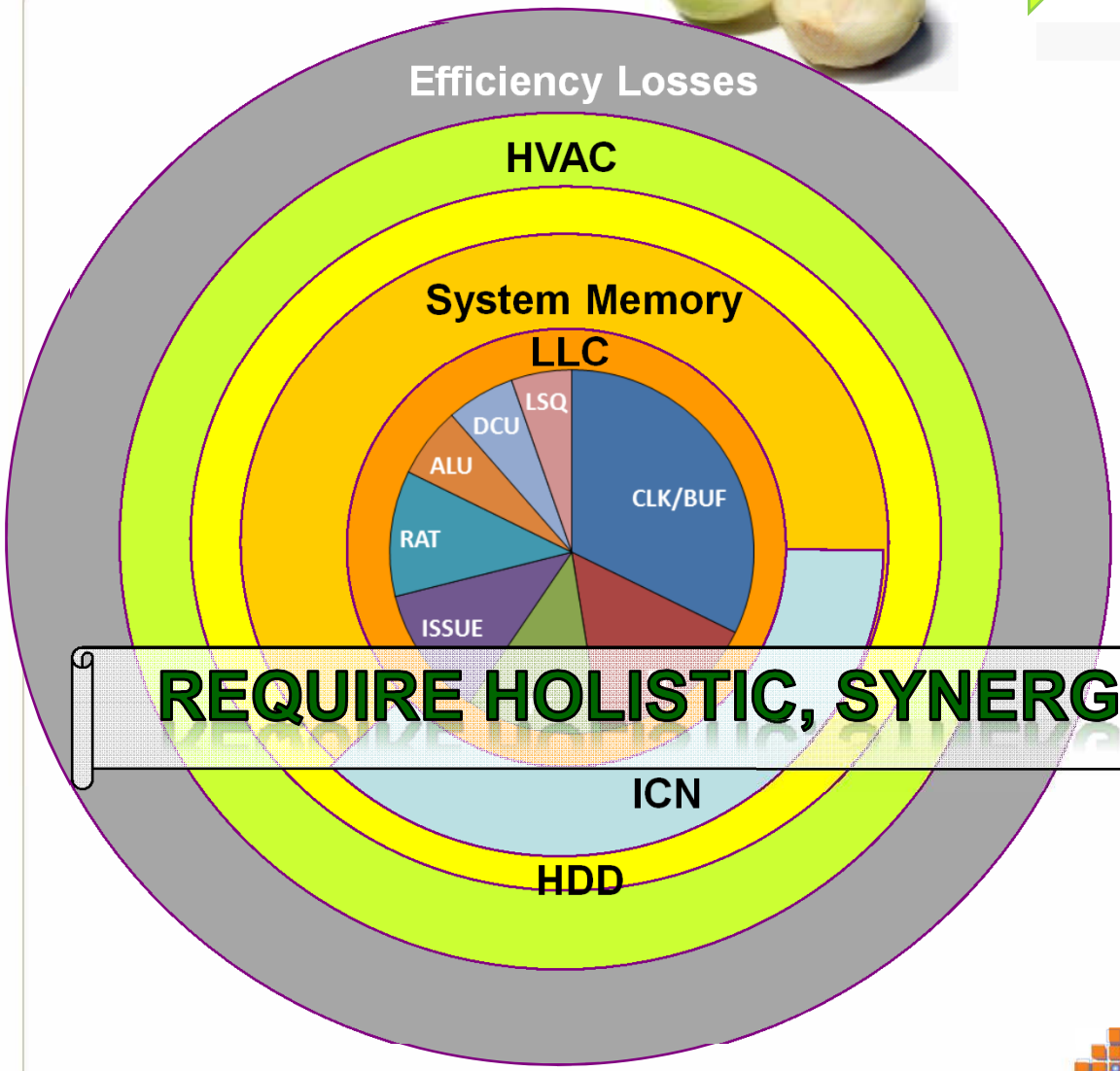


Round robin job placement



Intelligent job placement

Power Onion and Optimization Stack



Infrastructure Level
(Power conversion efficiency, Env. cooling, e.g., datacenters in Artic, ocean?)

Compiler/Algorithmic Level
(code gen impv., memory optimization, lookup tables, etc.)

System Level
(Consolidation, Locality mgmt, Power capping, DVFS, Smart provisioning, etc.)

Chip Level
(Emerging NVM, SSD, SoC Integration, 3D integration)

REQUIRE HOLISTIC, SYNERGISTIC APPROACHES

On-chip memory: Banking, Deeper hierarchy, Snoop filtering, etc.

Circuits Level
(Device scaling, ASV, Razor)

Thank You!

Georgia Tech
ECE MARS Lab
http://arch.ece.gatech.edu