Life with "dark" silicon.
Power and thermal problems in future platforms

Michael Kishinevsky
Strategic CAD Labs, Intel

Radboud University, Nijmegen, Netherlands
March 27, 2013

Outline

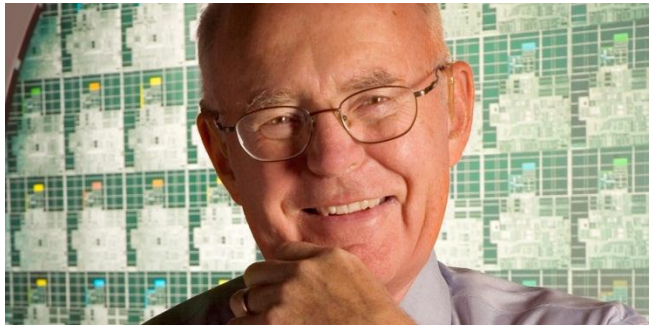
- Technology trends
- Power scaling and “dark” silicon
- Managing dark silicon

- Research challenges

Acknowledgements

- Raid Ayoub
- Umit Ogras
- Hyungjun Kim
- Steve Gunther
- Efi Rotem

Moore's Law



Gordon Moore "Cramming more components onto integrated circuits", *Electronics Magazine* 19 April 1965

"The complexity for minimum component costs has increased at a rate of roughly a factor of two per year...

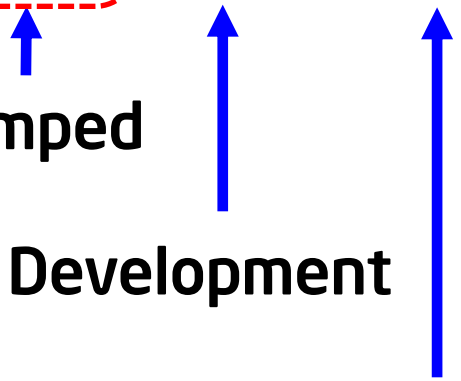
Certainly over the short term this rate can be expected to continue, if not to increase.

Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years.

That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000. I believe that such a large circuit can be built on a single wafer."

Intel Technology Roadmap

Process Name	P1266	P1268	P1270	P1272	P1274
1st Production	2007	2009	2011	2013	2015
Lithography	45 nm	32 nm	22 nm	14nm	10nm

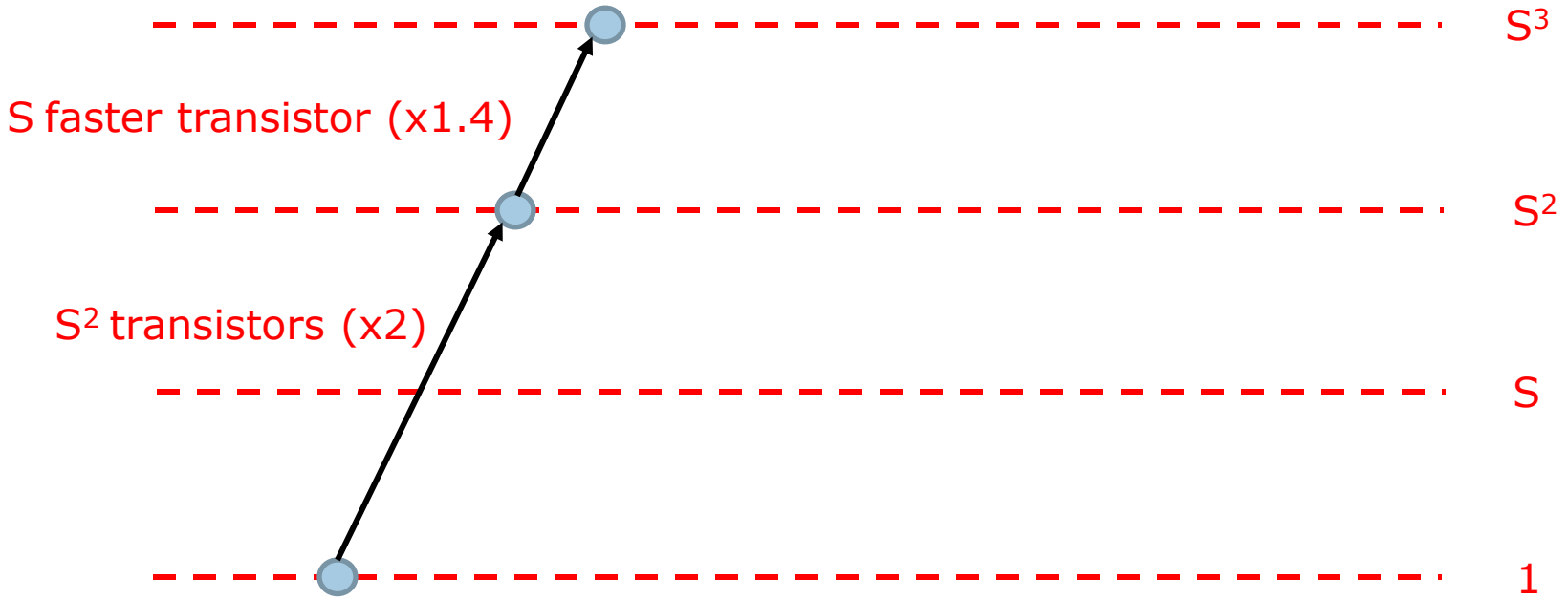


Intel continues cadence of introducing a new technology generation every two years

Technology Outlook

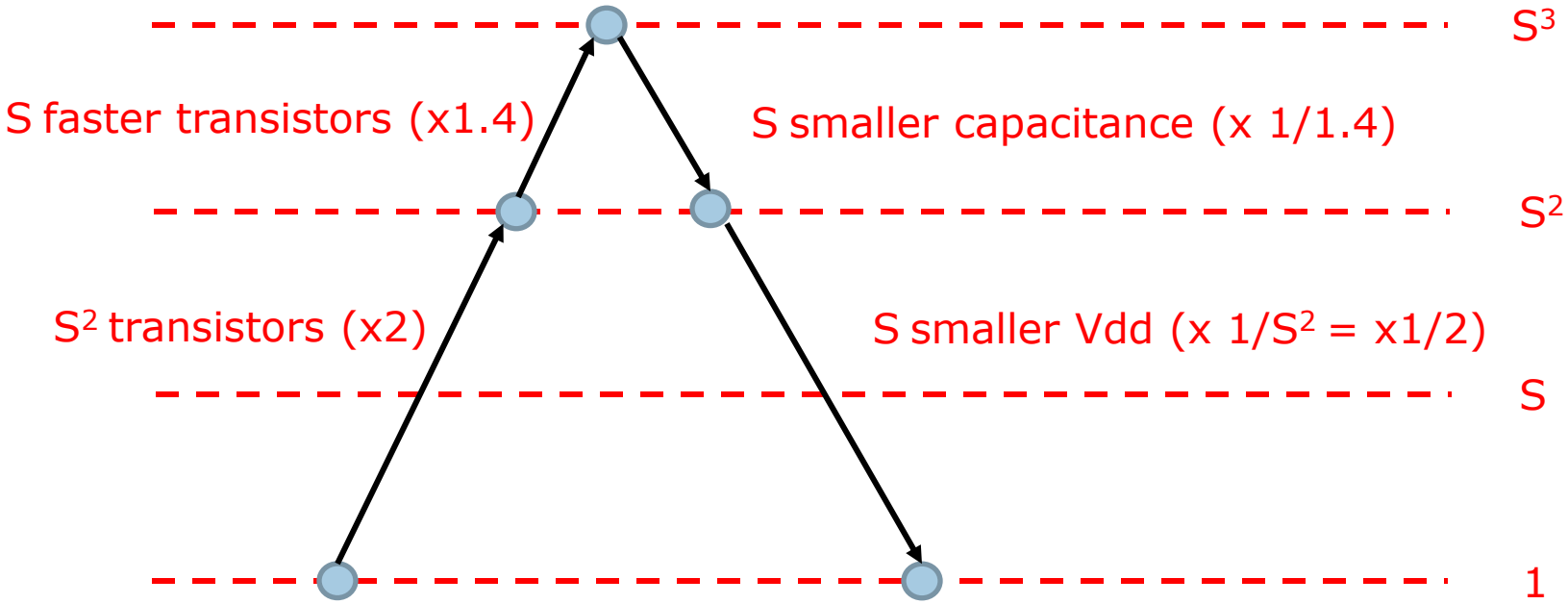
High Volume Manufacturing	2008	2010	2012	2014	2016	2018	2020	2022
Technology Node (nm)	**45**	**32**	**22**	**15**	**11**	**8**	**6**	**4**
Integration Capacity (BT)	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
Delay Scaling	>0.7			~1?				
Energy Scaling	~0.5			>0.5				
Transistors	Planar			3G, FinFET				
Variability	High			Extreme				
RC Delay	1	1	1	1	1	1	1	1
Metal Layers	8-9	0.5 to 1 Layer per generation						

Scaling Computing Capabilities



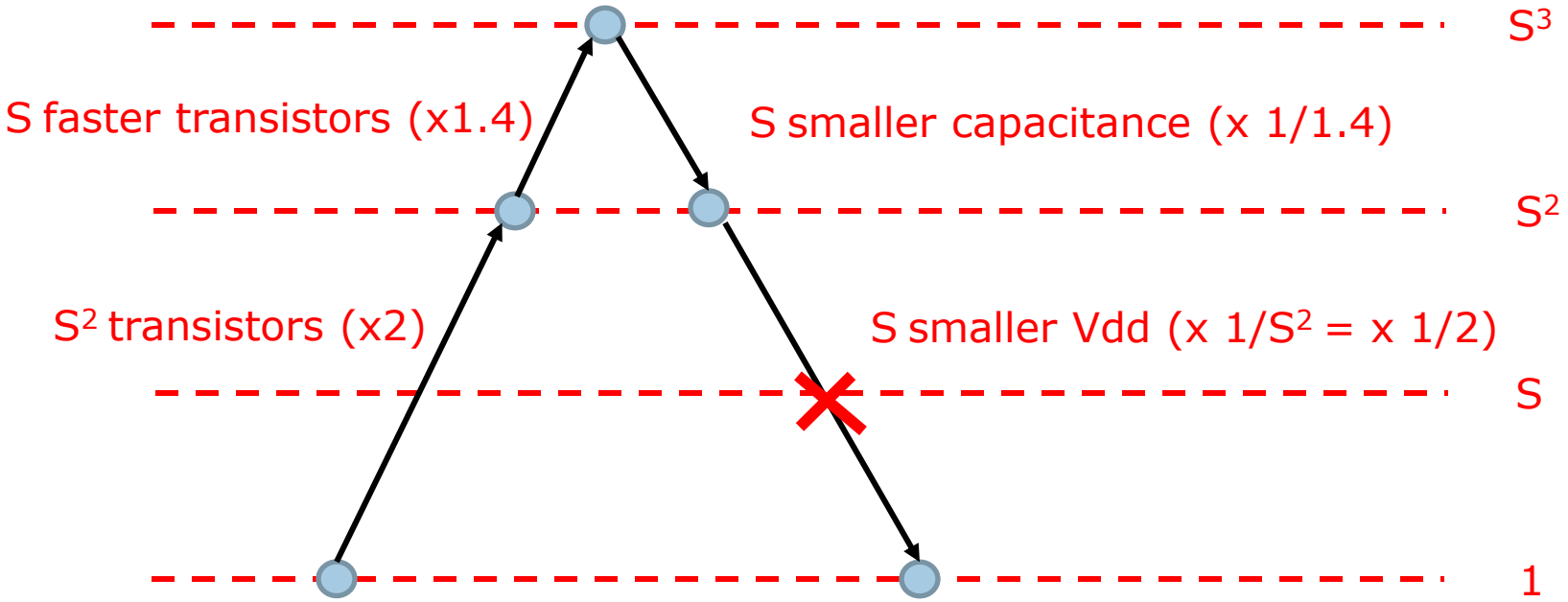
Dennard 1978: Computing capabilities scales as S^3 every generation
For S ~ 1.4 : S^3 ~ 2.8

Ideal Process Scaling. Constant Power

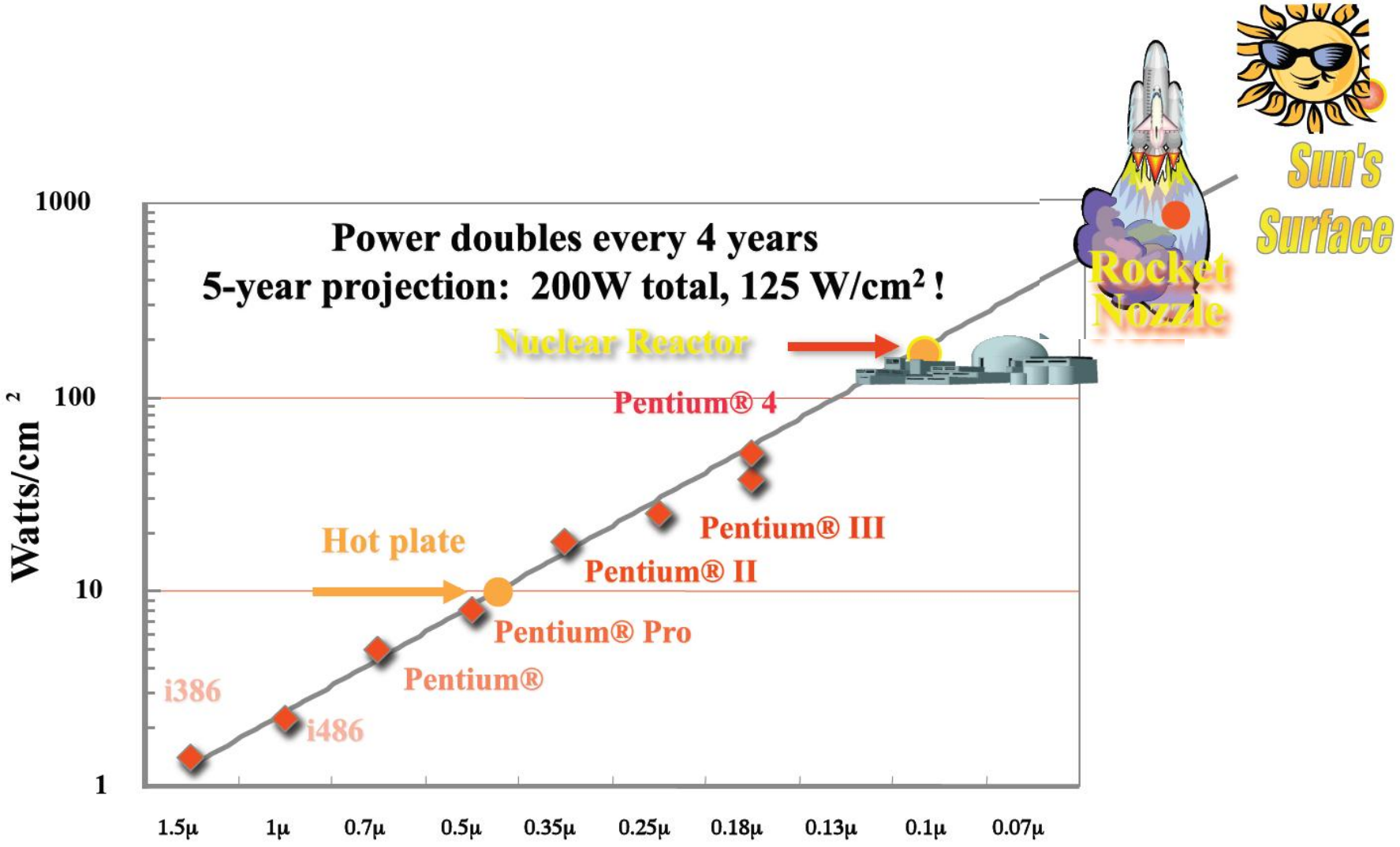


Dennard 1978: Computing capabilities scales as S^3 every generation
For S ~ 1.4 : S^3 ~ 2.8

Leakage Limited Process Scaling



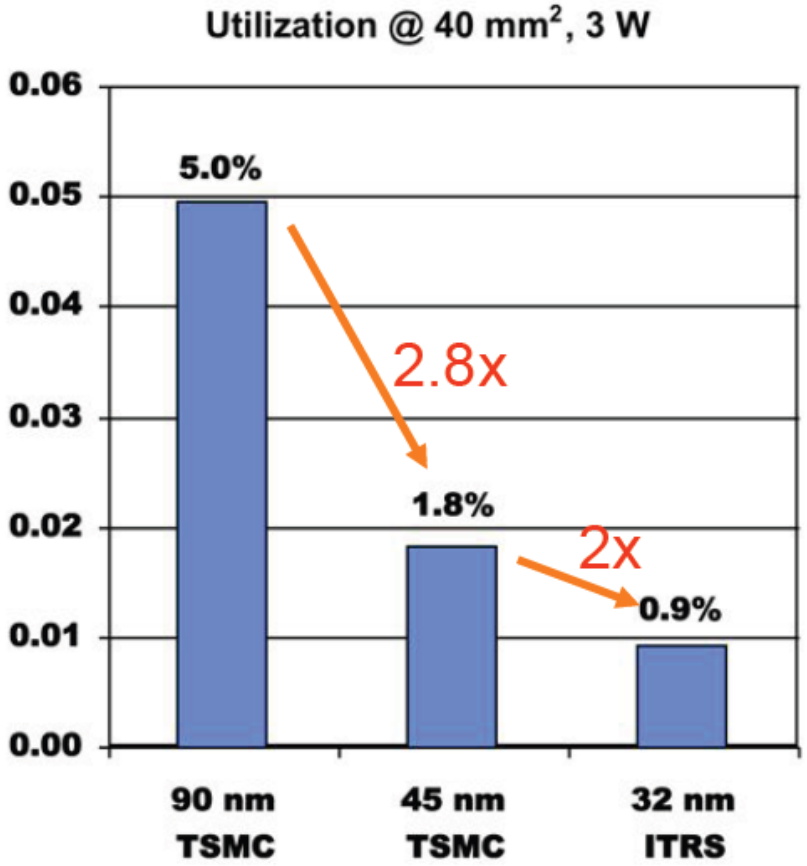
Need in Low Power Computing



Fred Pollack, Intel, 1999, keynote at Micro

Utilization Wall

Every process generation the percentage of a chip that can actively switch drops exponentially due to power constraints.
Under assumption: no radical process changes



Experiment with a small datapath

The remaining silicon that must be left unpowered is referred to as Dark Silicon
[Mike Muller, ARM, 2009]
[Michael Taylor et al.]

Ganesh Venkatesh et al., ASPLOS, 2010

Dark Silicon Apocalypse

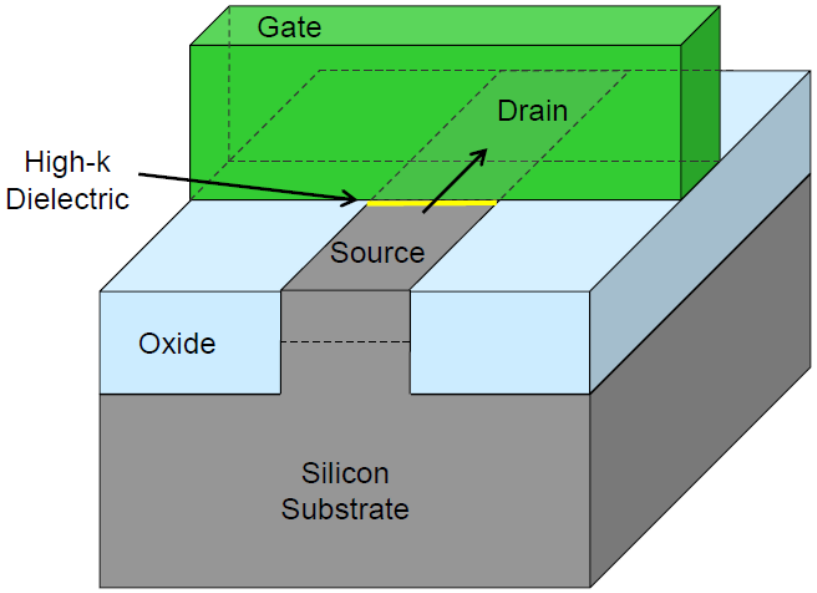
«Four Horsemen of the Apocalypse» (1887)

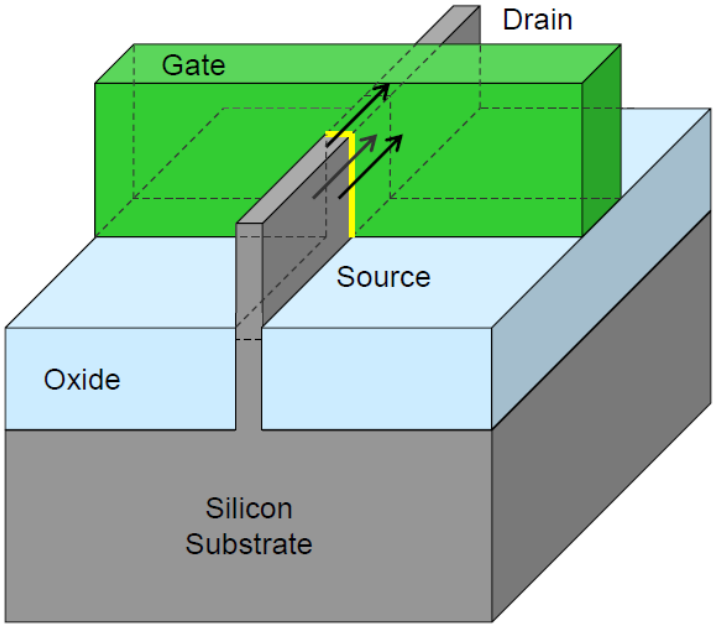
Viktor Vasnetsov

Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power management
- System-level optimization

22 nm Tri-Gate Transistor





Traditional 2-D planar transistor forms a conducting channel in the silicon region under the gate electrode when in the "on" state.

Used in Intel's 32 nm technology

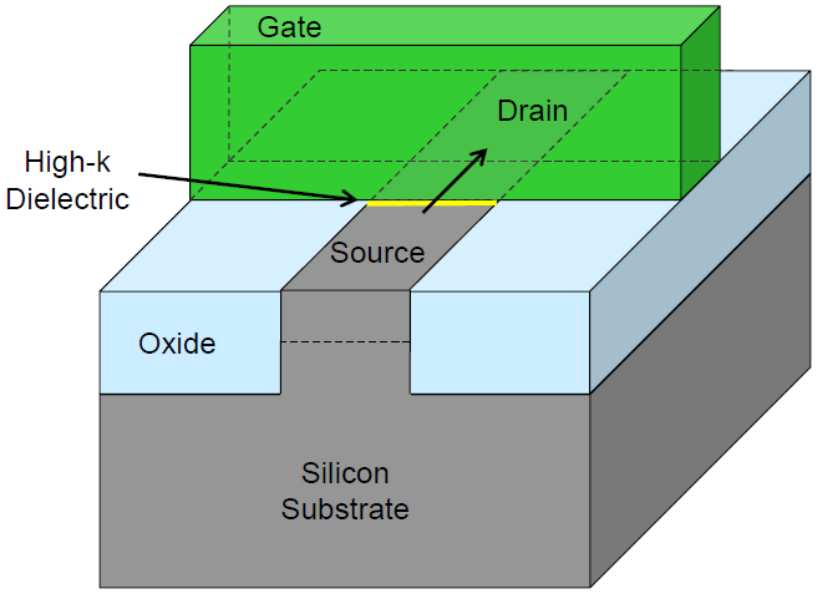
3-D Tri-Gate transistor forms conducting channels on three sides of a vertical fin structure, providing "fully depleted" operation.

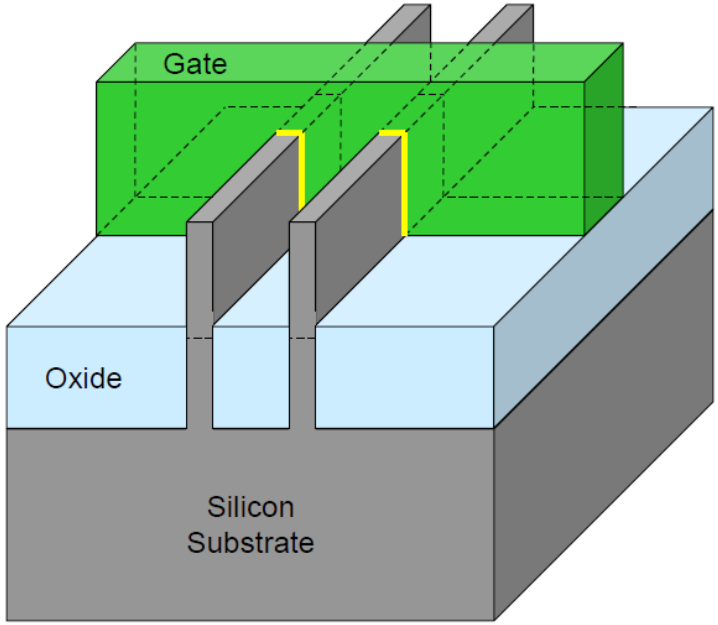
Only 2-3% cost adder.

Used in Intel's 22 nm technology

Intel, 2011 http://newsroom.intel.com/docs/DOC-2032

22 nm Tri-Gate Transistor

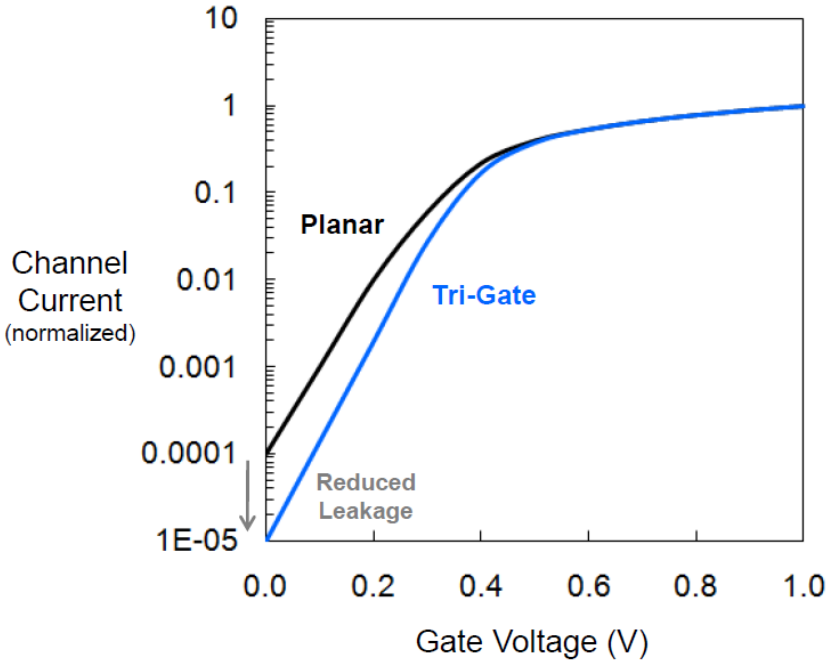


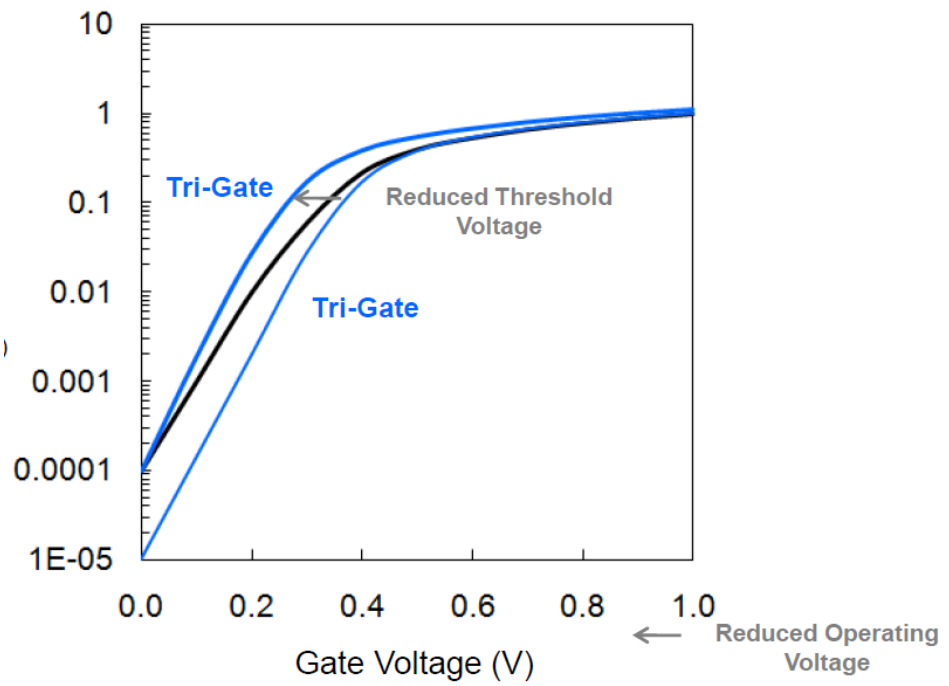


3-D Tri-Gate transistor can have multiple fins connected together to increase total drive strength for higher performance.

Intel, 2011 http://newsroom.intel.com/docs/DOC-2032

22 nm Tri-Gate Transistor



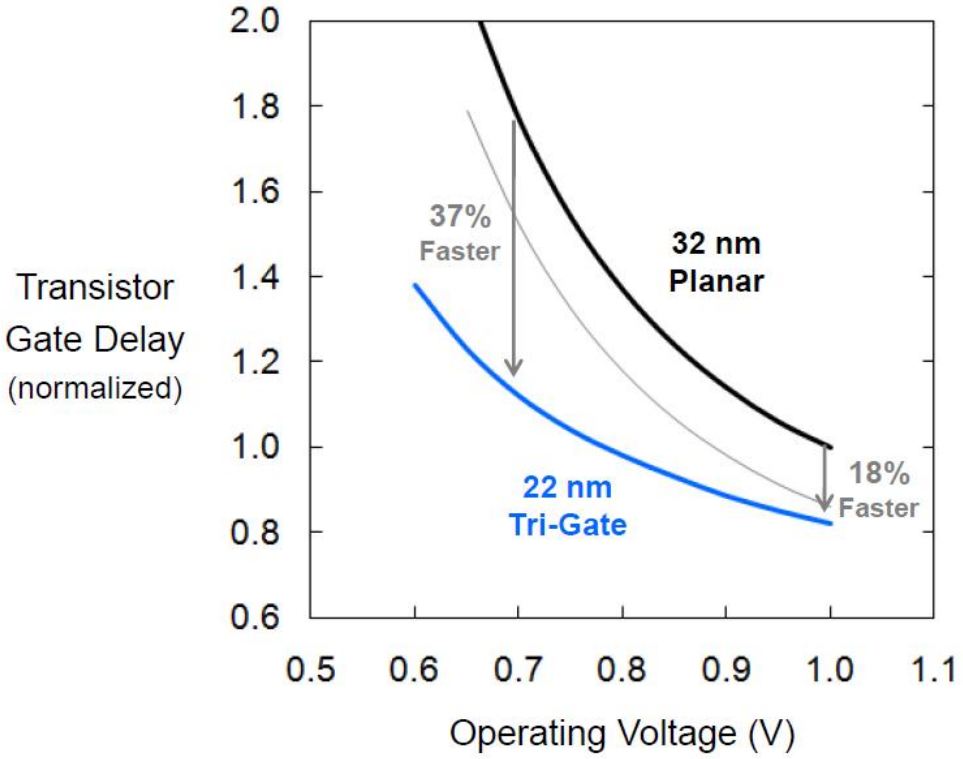


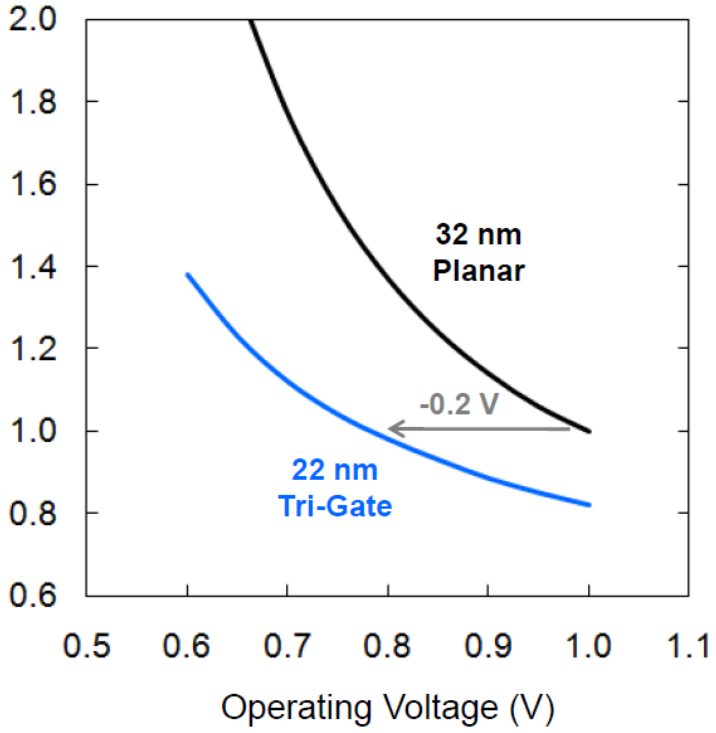
The fully depleted characteristics of Tri-Gate transistors provide a steeper sub-threshold slope that significantly reduces leakage current

Steeper sub-threshold slope can be used to target lower threshold voltage, allowing transistor to operate at lower operating voltage

Intel, 2011 http://newsroom.intel.com/docs/DOC-2032

22 nm Tri-Gate Transistor





Improved performance at high voltage and unprecedented performance gain at low voltage

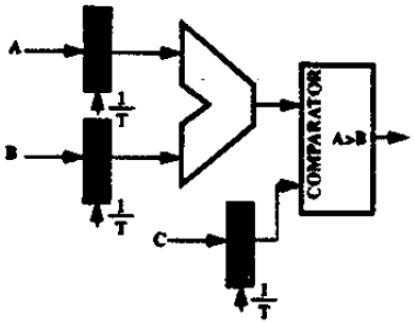
Can operate at lower voltage with same performance.
Active power reduced by > 50%.

Intel, 2011 http://newsroom.intel.com/docs/DOC-2032

Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization

Parallelism for power efficiency



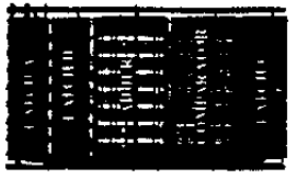
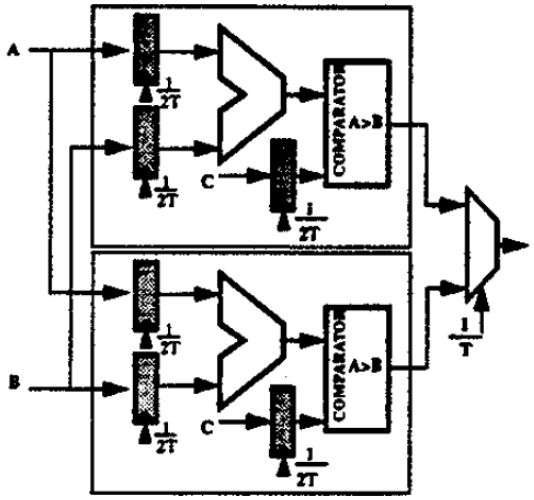


Fig. 7. A simple data path with corresponding layout.

$$P_{par} = C_{par} V_{par}^2 f_{par}$$

$$= (2.15C_{ref})(0.58V_{ref})^2 \left(\frac{f_{ref}}{2}\right) \approx 0.36\, P_{ref}.$$



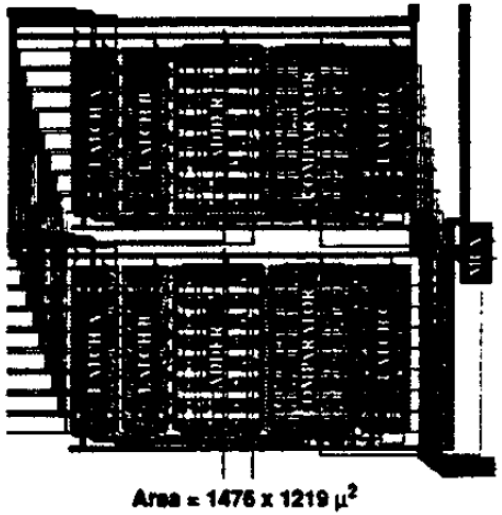


Fig. 8. Parallel implementation of the simple data path.

Ideal voltage scaling

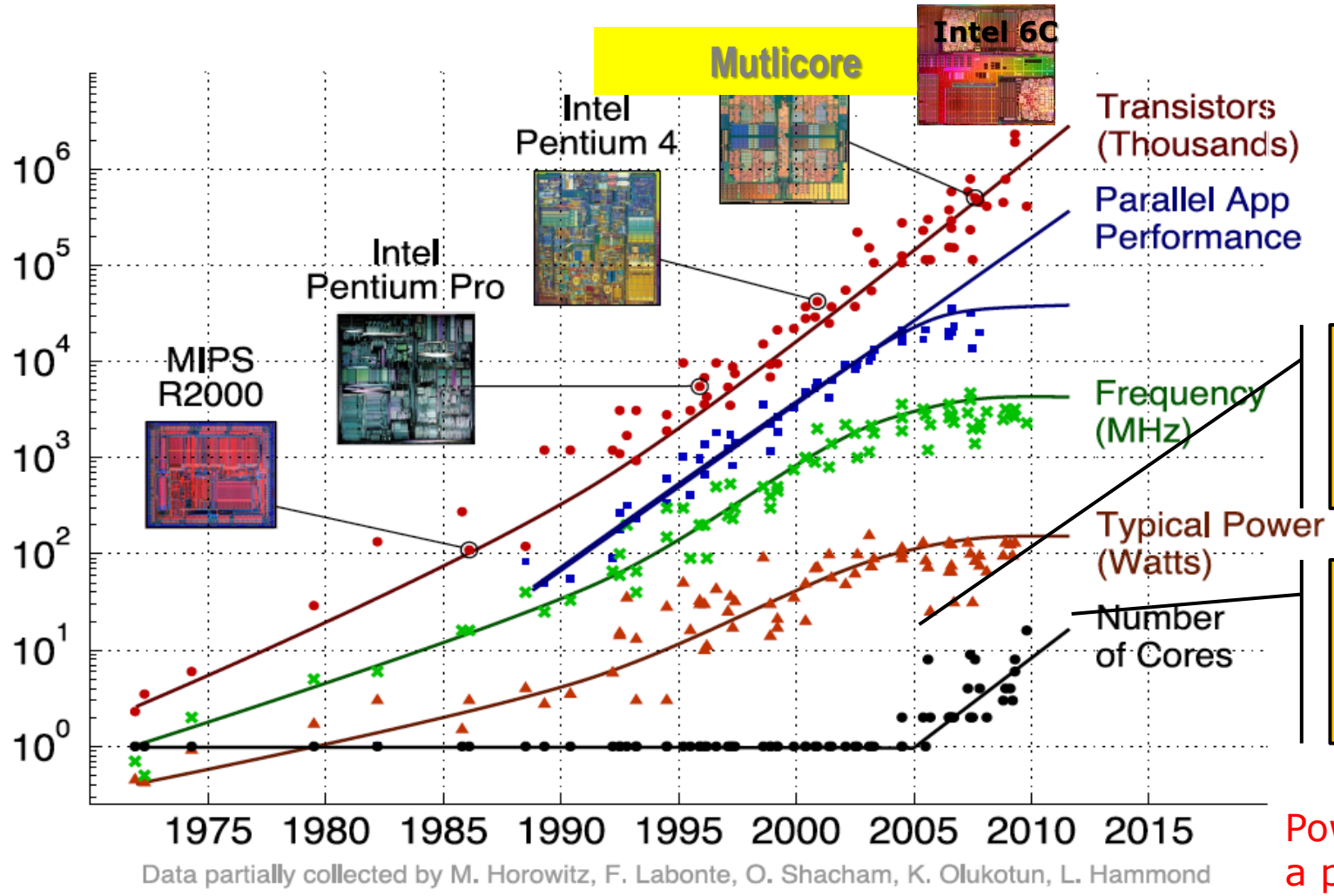
Ideal parallelization

IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 27, NO. 4, APRIL 1992

Low-Power CMOS Digital Design

Anantha P. Chandrakasan, Samuel Sheng, and Robert W. Brodersen, *Fellow, IEEE*

Power and Area Constrained Performance



Power constrained performance caused transition to multi-cores

Area constraints limiting the number of cores.

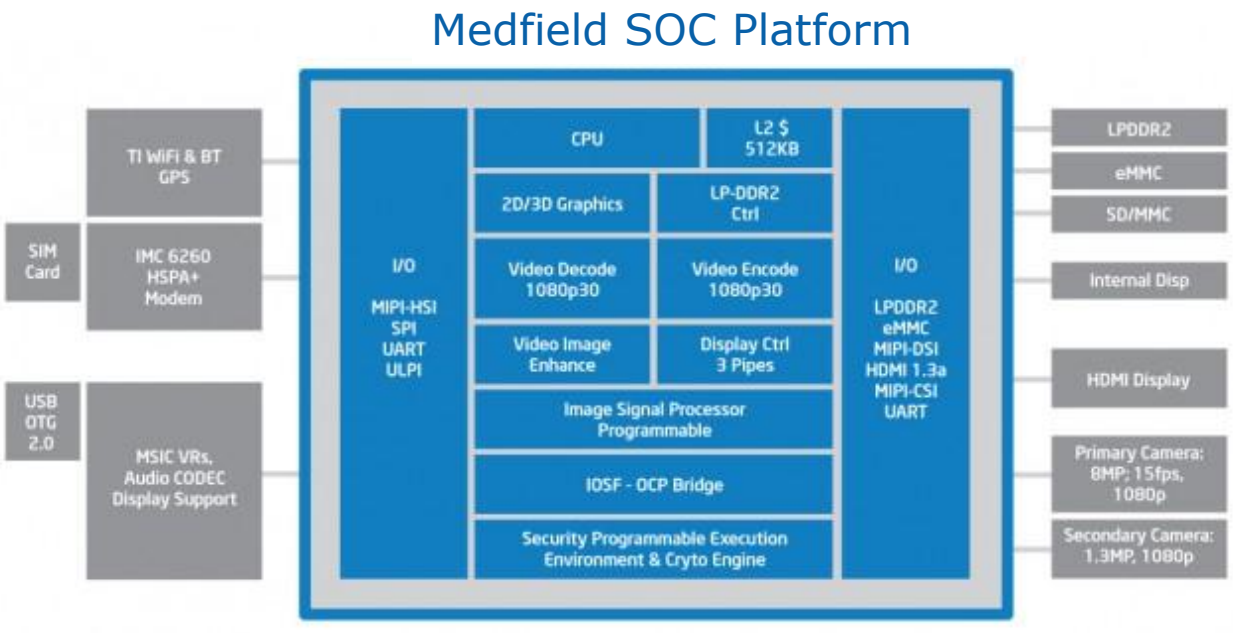
Power constraint is a primary limiter in the future

Life with Dark Silicon

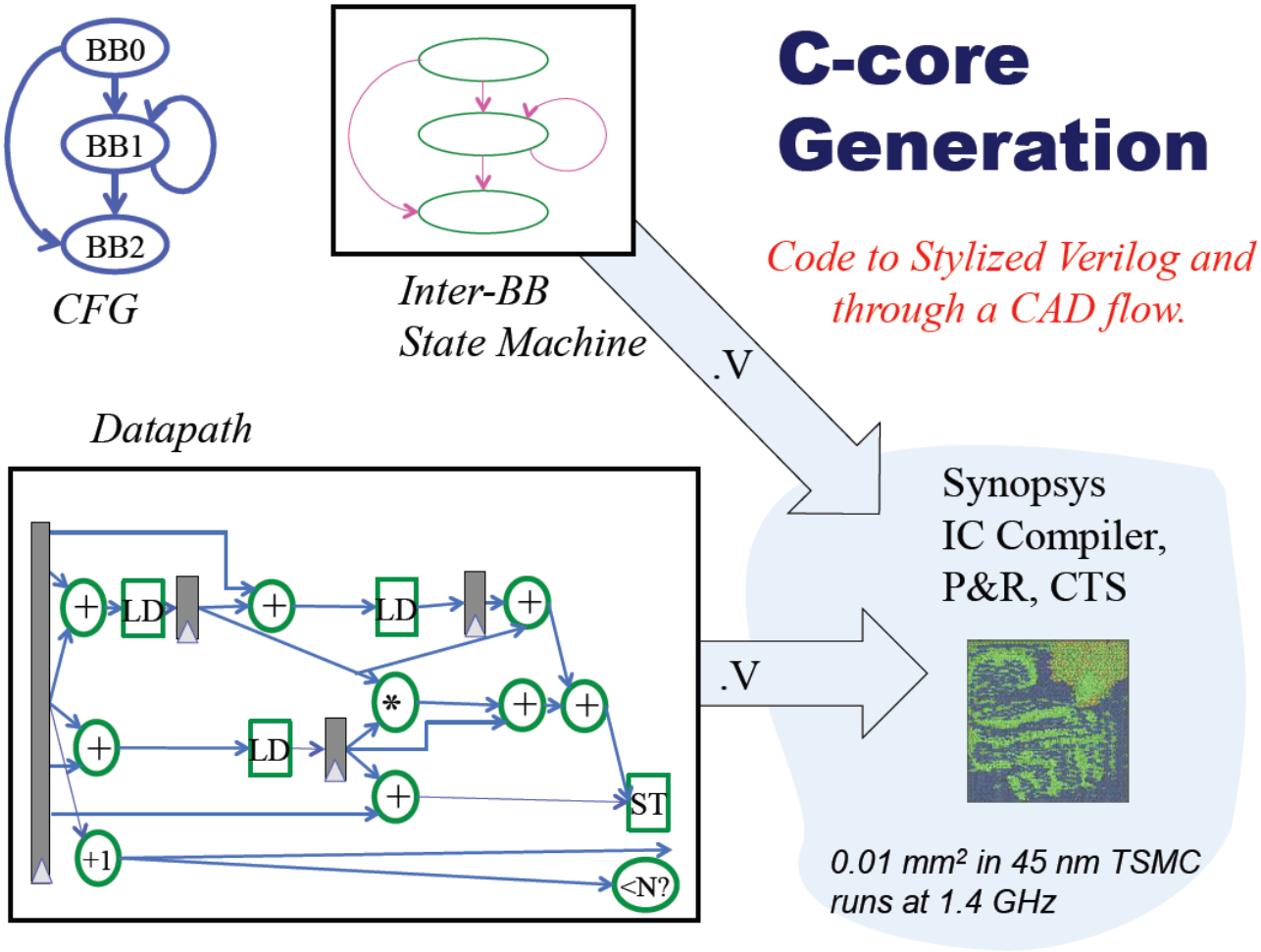
- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization

SoC. Collection of specialized block

- ASIC specialized block is an order more energy-efficient per unit of performance than universal CPU
  - Specialization: only good at one task
  - Specialized logic and shorter wires to deliver computation results



Research on quick generation of accelerators (an example)



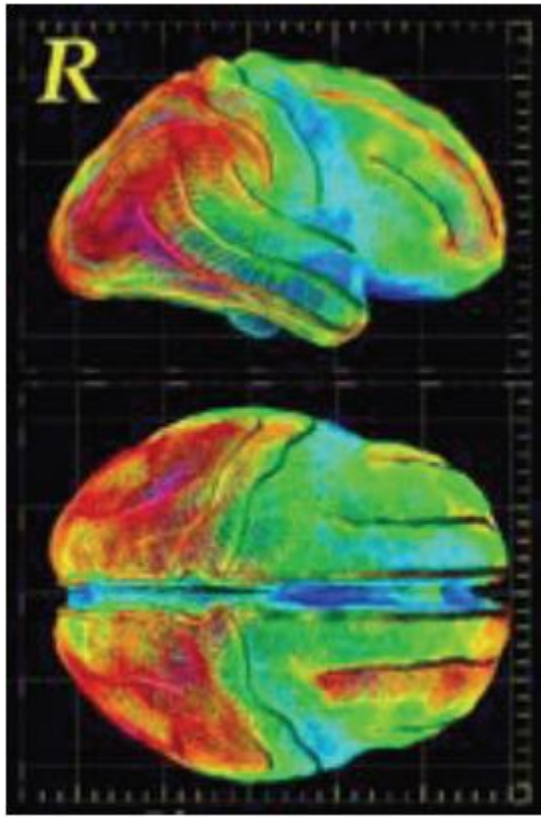
Ganesh Venkatesh et al., ASPLOS, 2010

Life with Dark Silicon

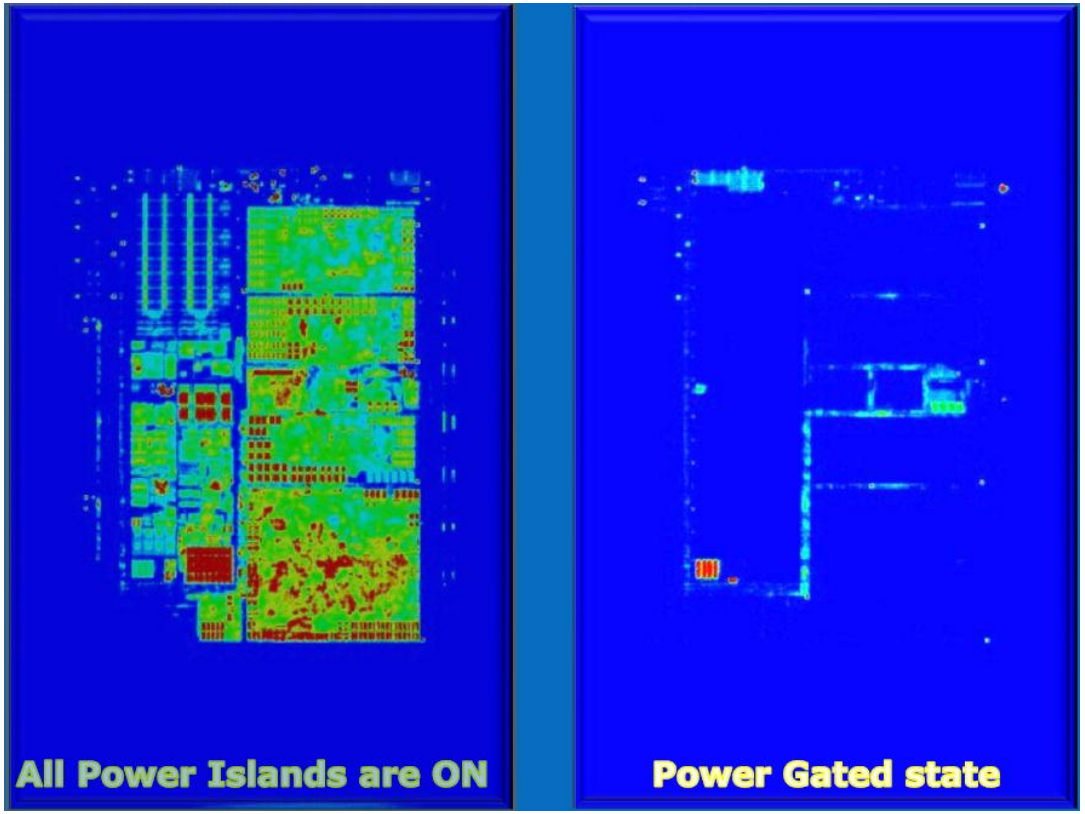
- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization

Dimming

- Human Brain – very “dark” circuit
- 100 trillion synapses at 20W



- Lincroft SoC IREM
- 50x idle power reduction via power gating

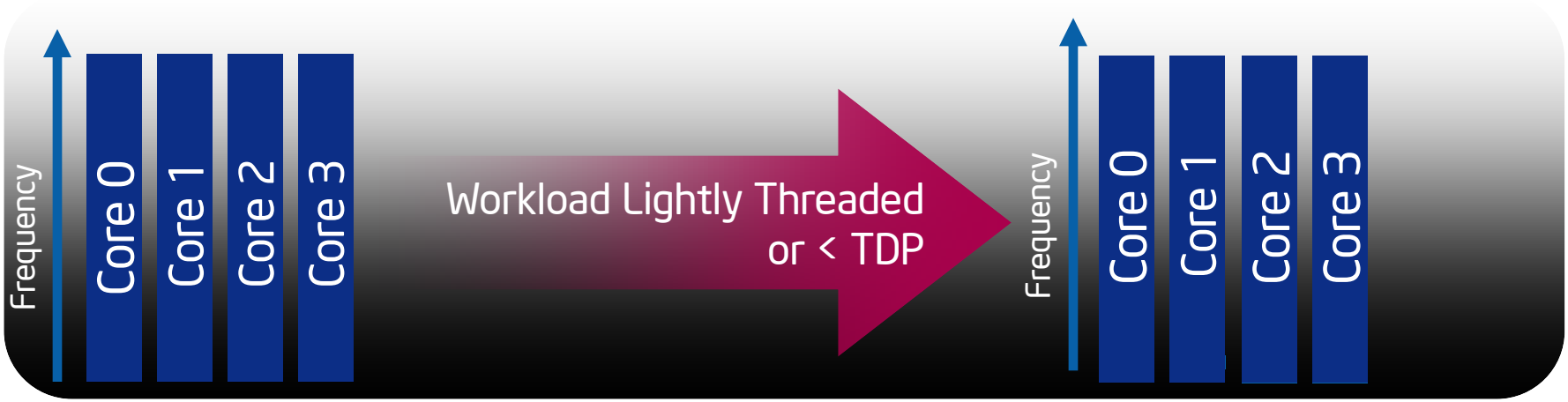


Power management complexity.
Intel® Turbo Boost™ Technology (Nehalem)

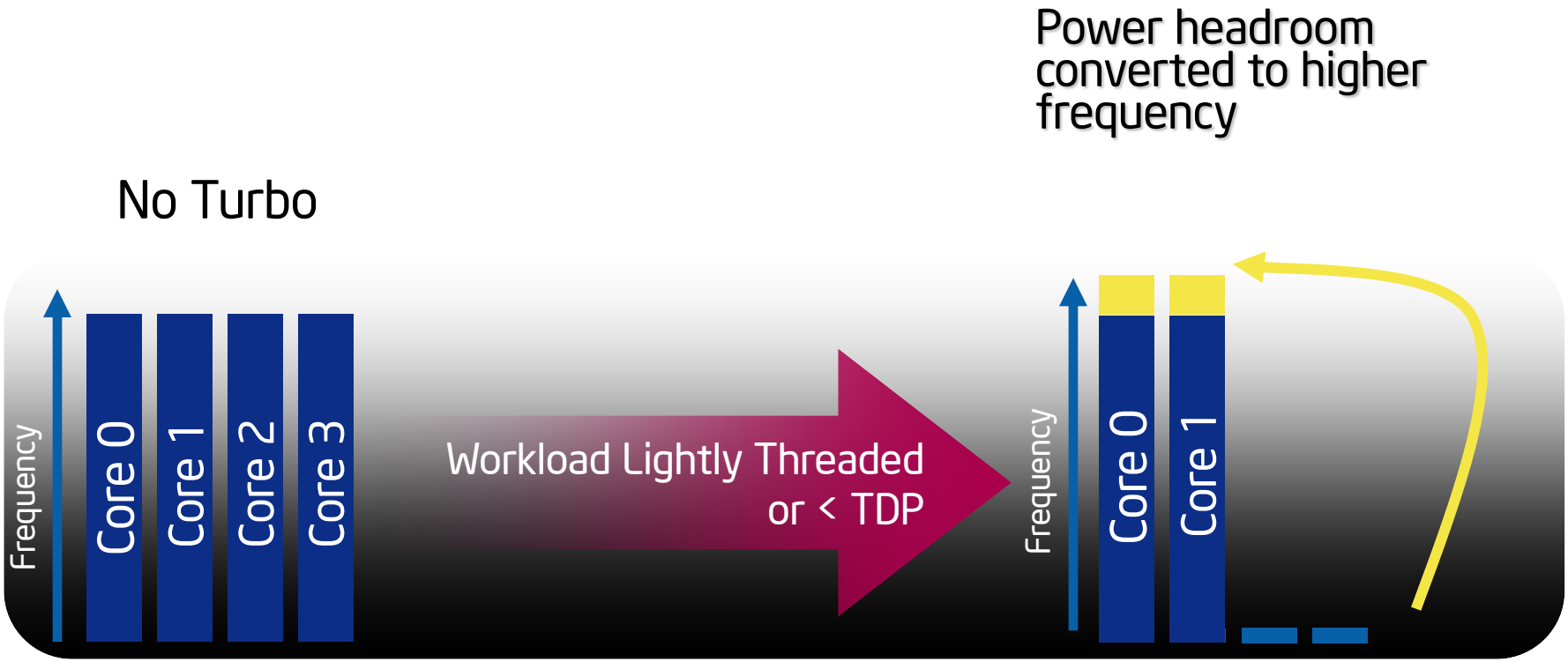
C6

Zero power for inactive cores

No Turbo



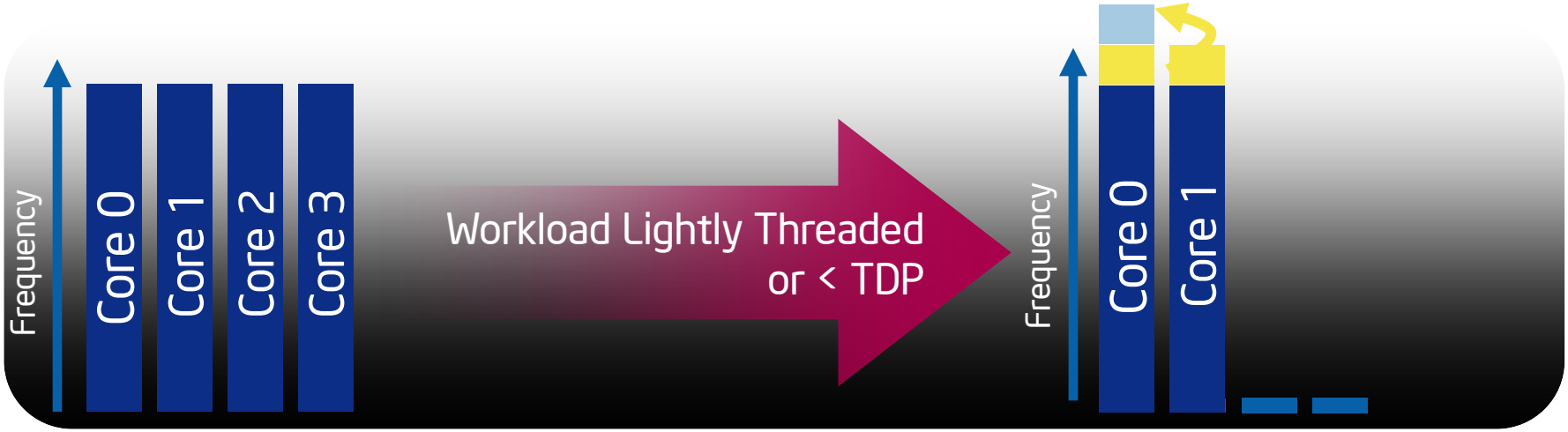
Intel® Turbo Boost™ Technology (Nehalem)



Intel® Turbo Boost™ Technology (Nehalem)

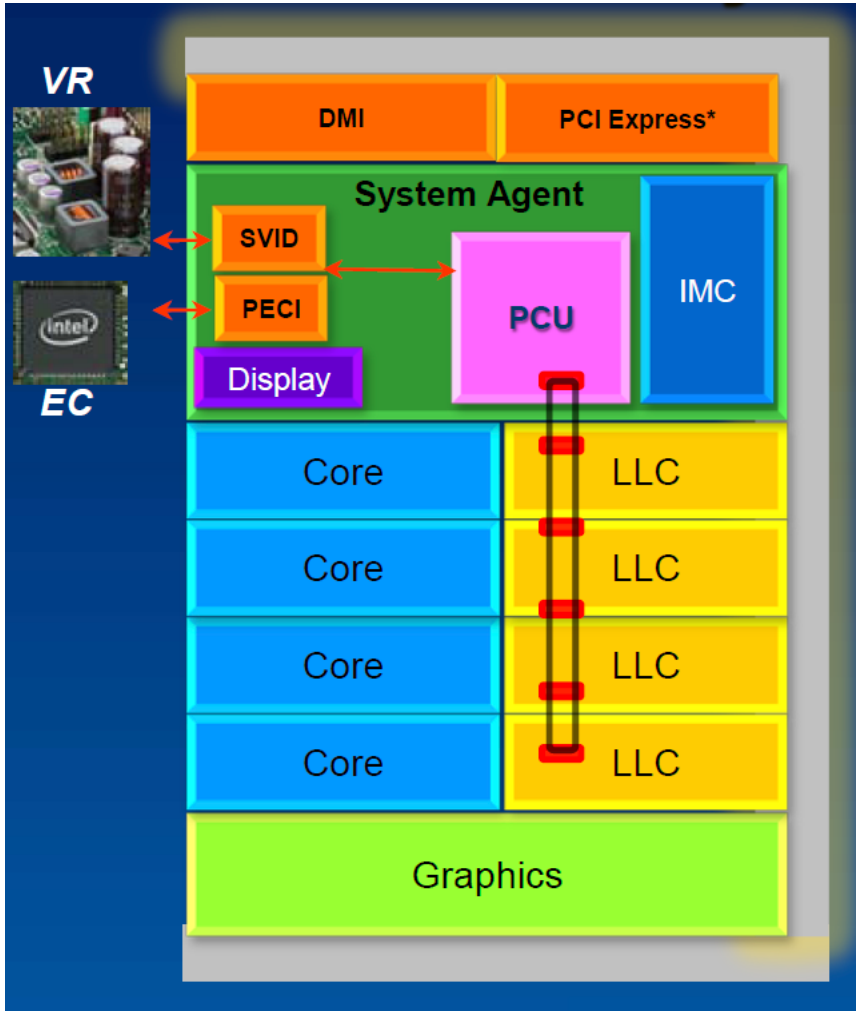
Power headroom converted to higher frequency

No Turbo



Steve Gunther, Intel, 2010

Sandy Bridge Power Management

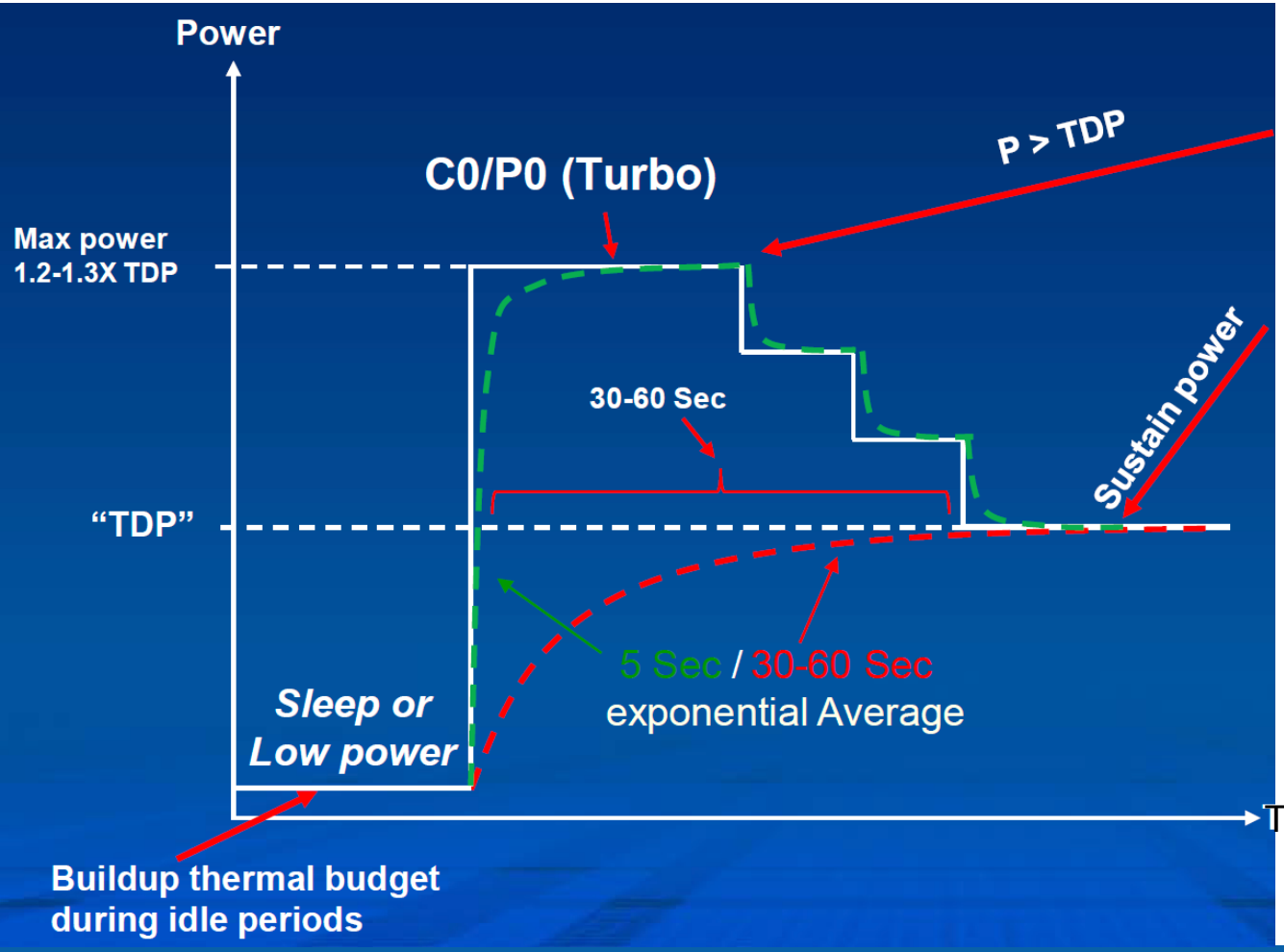


- 1-4 CPU cores + Gfx
- Integrated system agent
  - Package control unit
  - Memory controller
- Sliced LLC shared by cores/Gfx
- Ring interconnect + power management link
- PCU – control system
  - Logic and embedded controller running power and thermal management firmware
  - Monitors physical conditions (V, T, P)
  - Controls power states of CPU, Gfx (V,f)
  - Controls voltage regulators, DDR and system
- Excepts external inputs
  - System power management requests and limits
  - Power and temperature reading

Efi Rotem et al., Intel, 2011, HotChips

http://www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.19.9-Desktop-CPUs/HC23.19.921.SandyBridge_Power_10-Rotem-Intel.pdf

Intel® Turbo Boost™ Technology 2.0



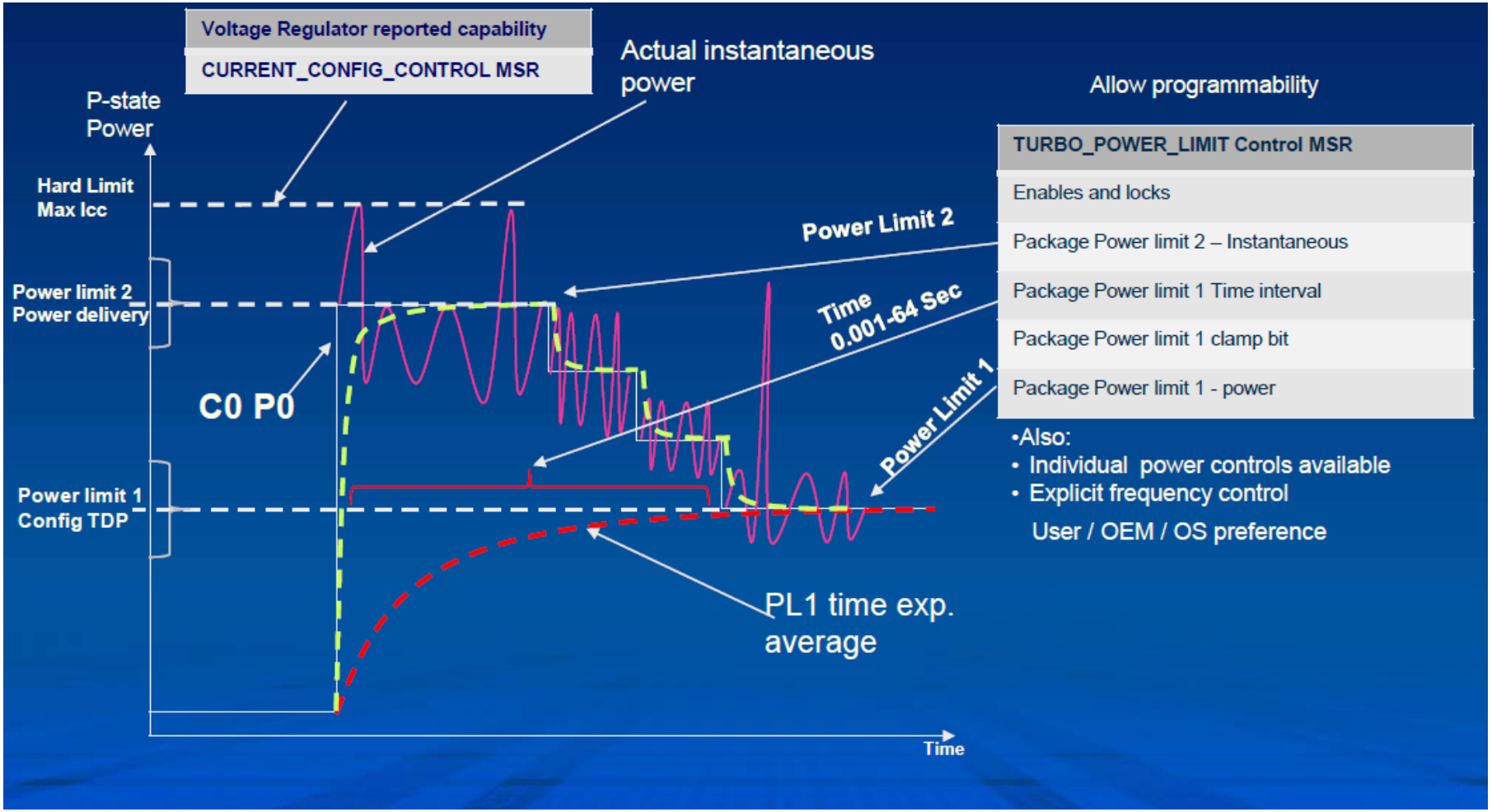
After idle periods, the system accumulates energy budget and can deliver higher power/performance for up to a minute. Used to enhance user experience.

In steady state power stabilizers on TDP

Efi Rotem et al., Intel, 2011, HotChips

http://www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.19.9-Desktop-CPUs/HC23.19.921.SandyBridge_Power_10-Rotem-Intel.pdf

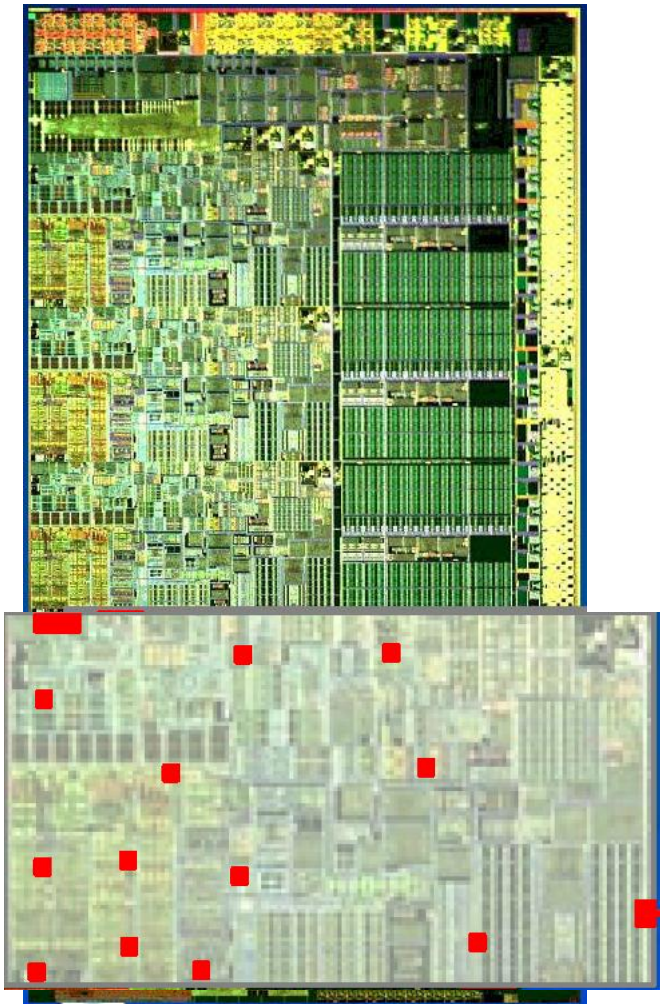
Turbo Control Dynamics



Efi Rotem et al., Intel, 2011, HotChips

http://www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.19.9-Desktop-CPUs/HC23.19.921.SandyBridge_Power_10-Rotem-Intel.pdf

Sandy Bridge Package Thermal Management



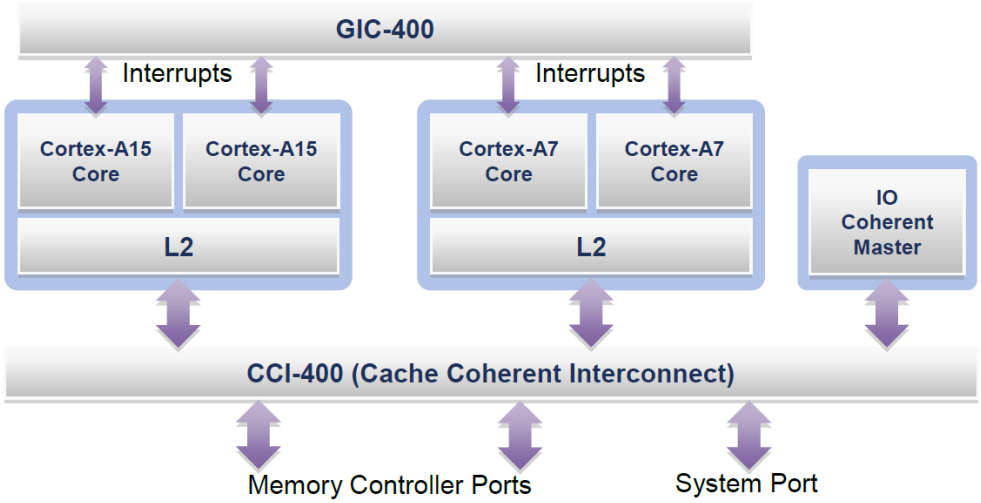
- On-die thermal sensors: 12 on each CPU + Gfx + ring + system agent. Operating range 50-100C
- Report max T per block and total chip
- Management: critical thermal protection (notification, throttle, shutdown)
- In addition used in power management for
  - Leakage calculation of power meter
  - PM optimization algorithms
  - External controls (e.g. fans)

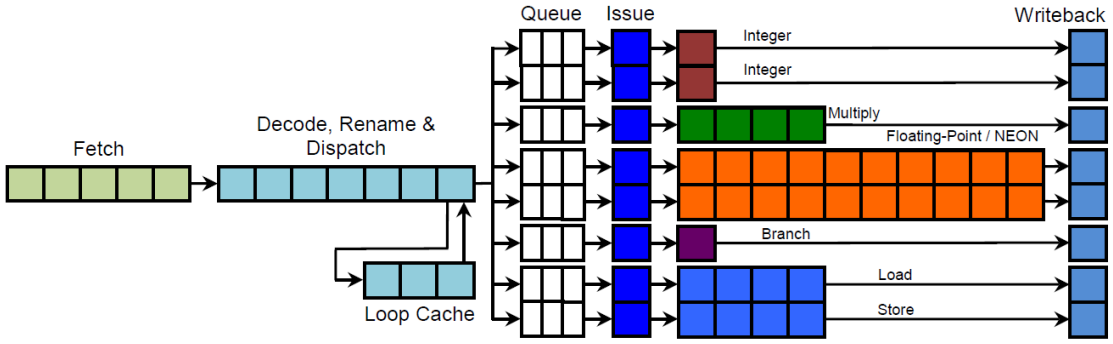
Efi Rotem et al., Intel, 2011, HotChips

http://www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.19.9-Desktop-CPUs/HC23.19.921.SandyBridge_Power_10-Rotem-Intel.pdf

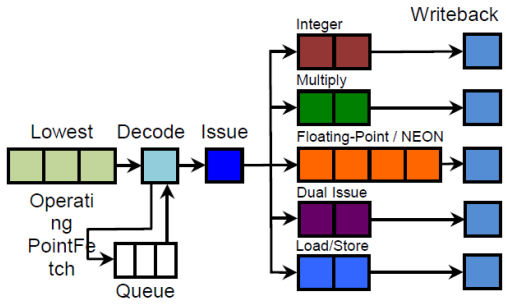
Hetero – extending DVFS. ARM's Big.LITTLE

System





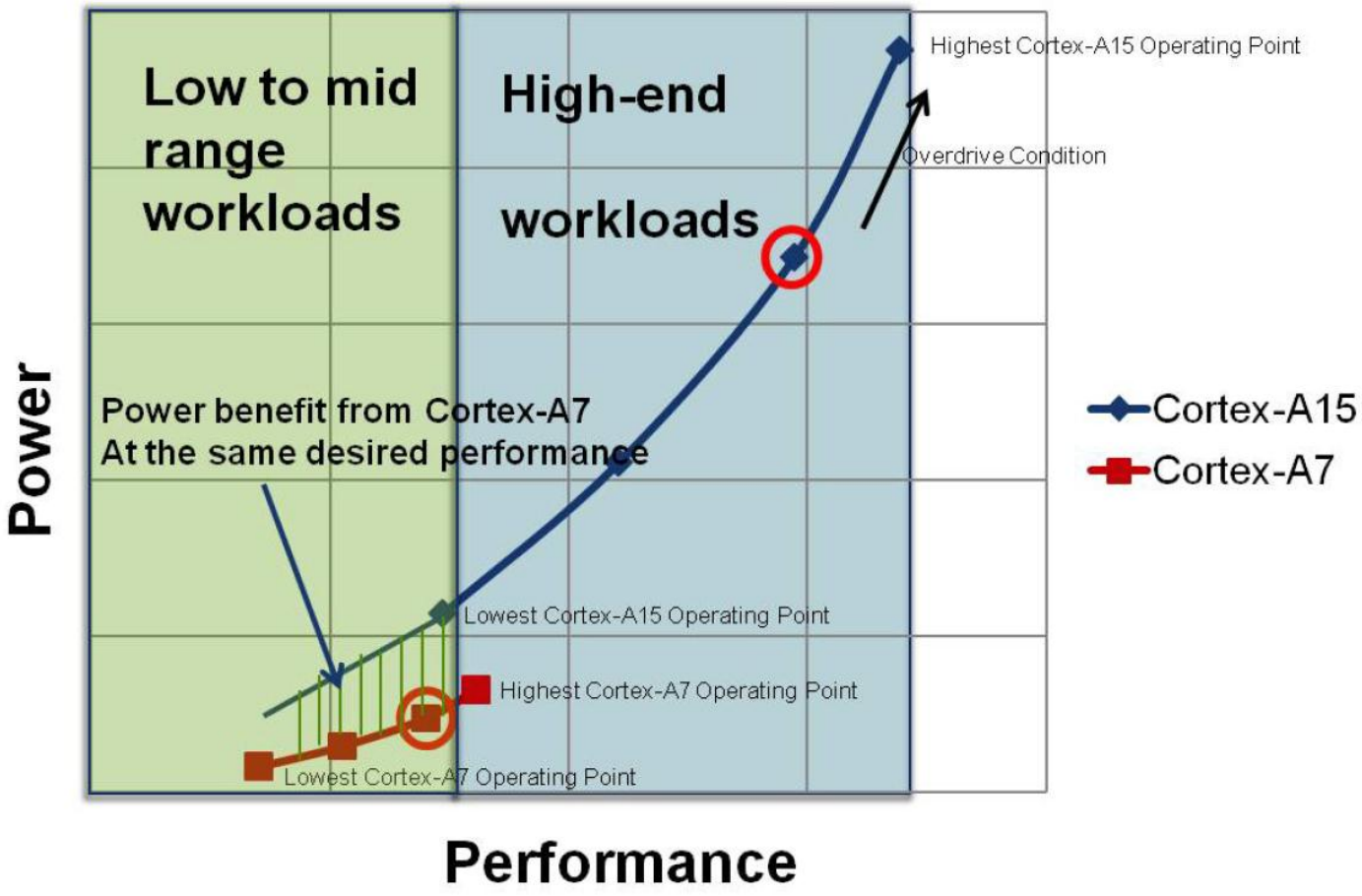
Big: 3-issue, out-of-order



LITTLE: 2-issue, in-order

Peter Greenhalgh, ARM, 2011
http://www.arm.com/files/downloads/big.LITTLE_Final.pdf

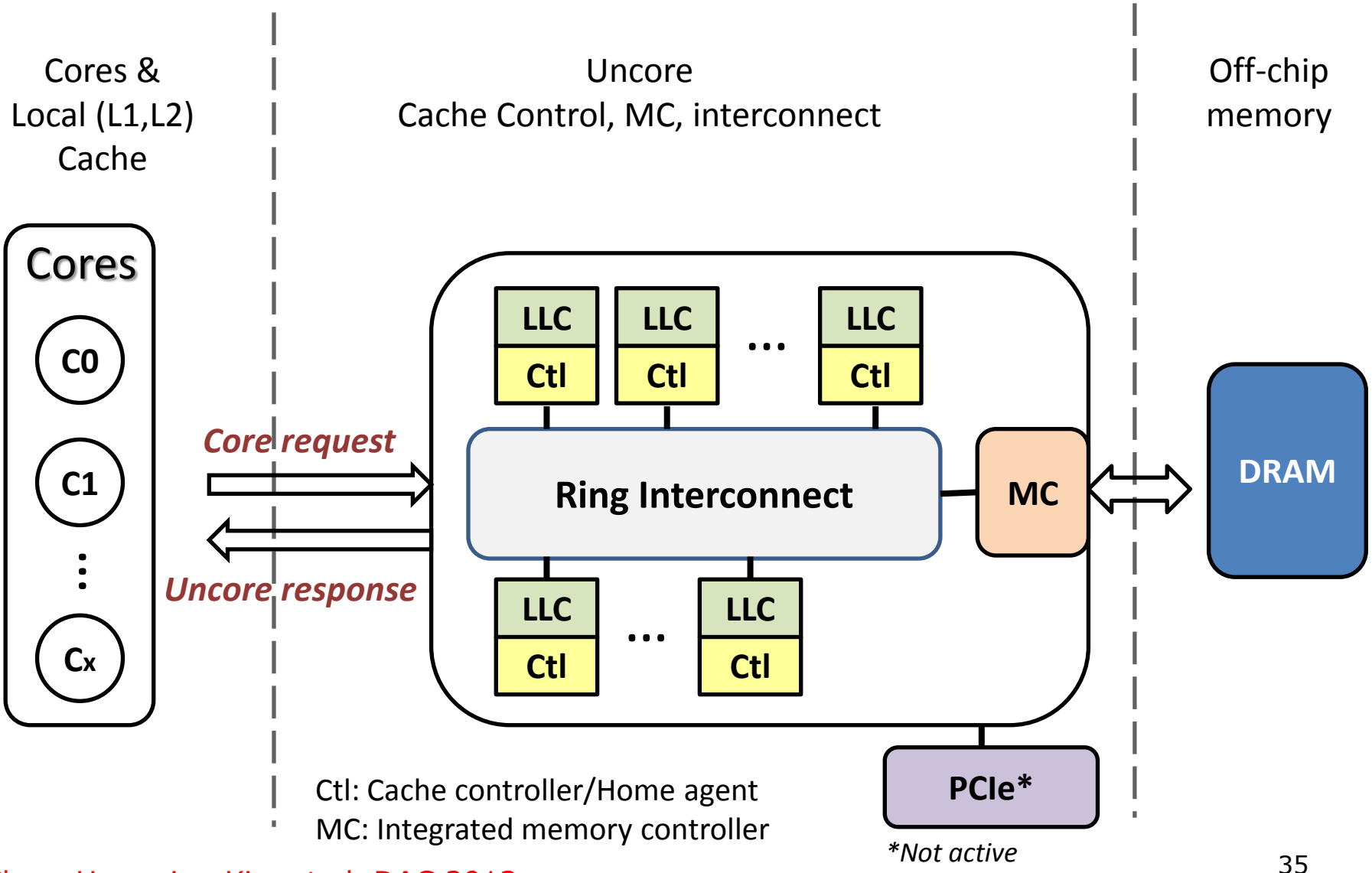
Big.LITTLE Power-performance Trade-off



Challenge: Migration of the application

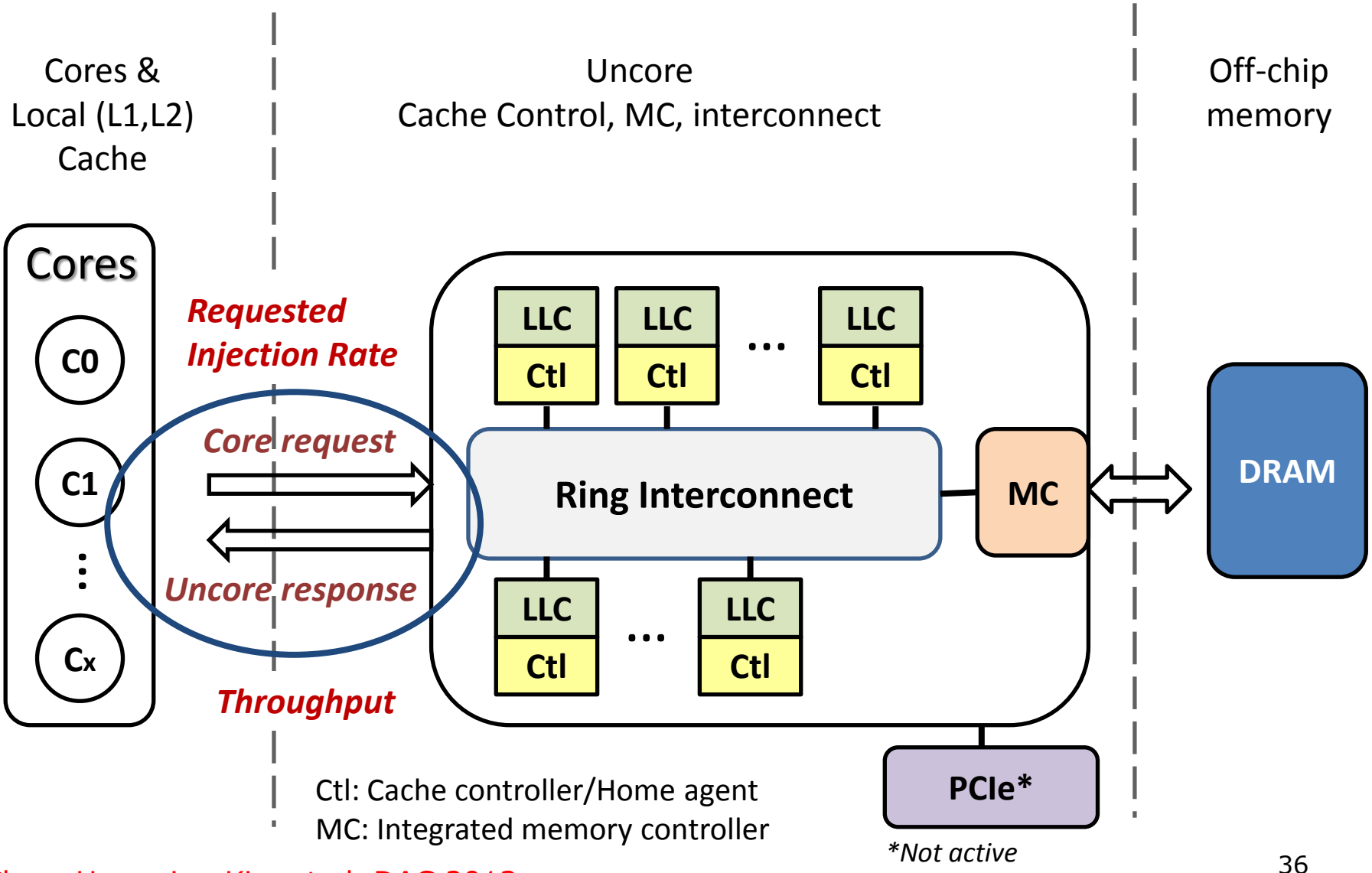
Peter Greenhalgh, ARM, 2011
http://www.arm.com/files/downloads/big.LITTLE_Final.pdf

Research on Uncore Power Management



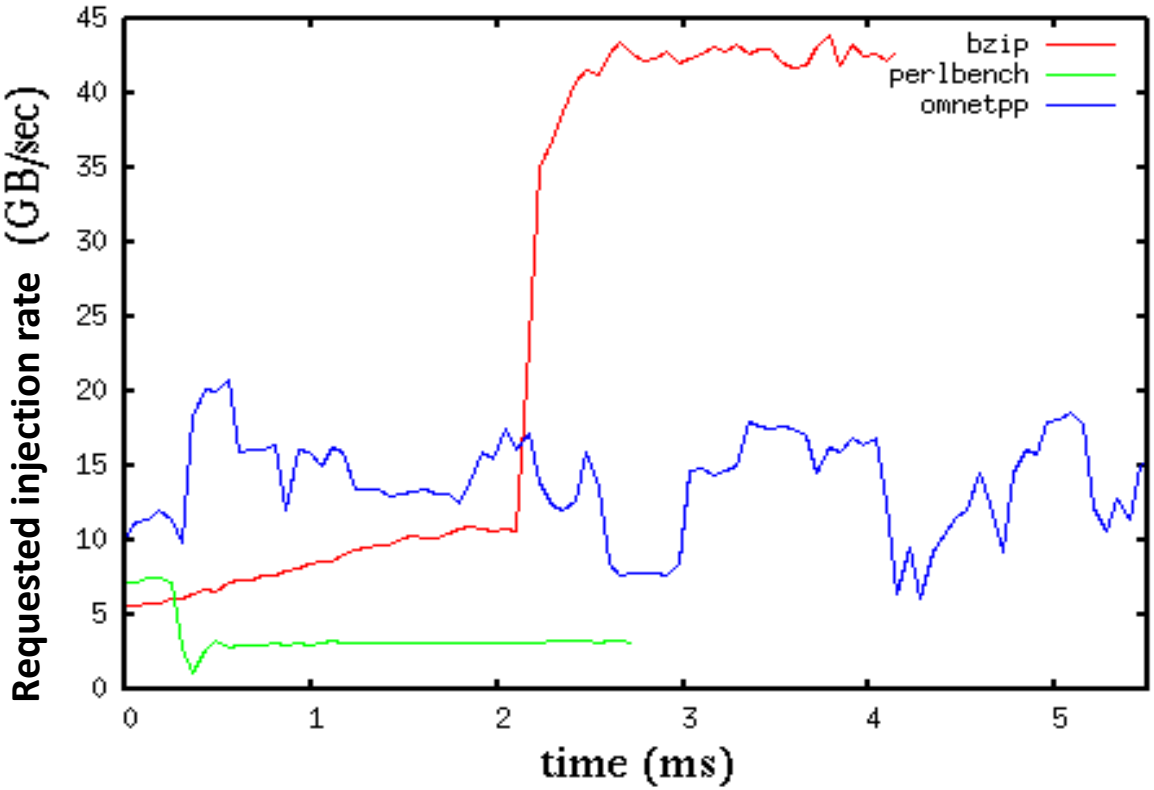
Xi Chen, Hyungjun Kim et al. DAC 2013

Research on Uncore Power Management



Xi Chen, Hyungjun Kim et al. DAC 2013

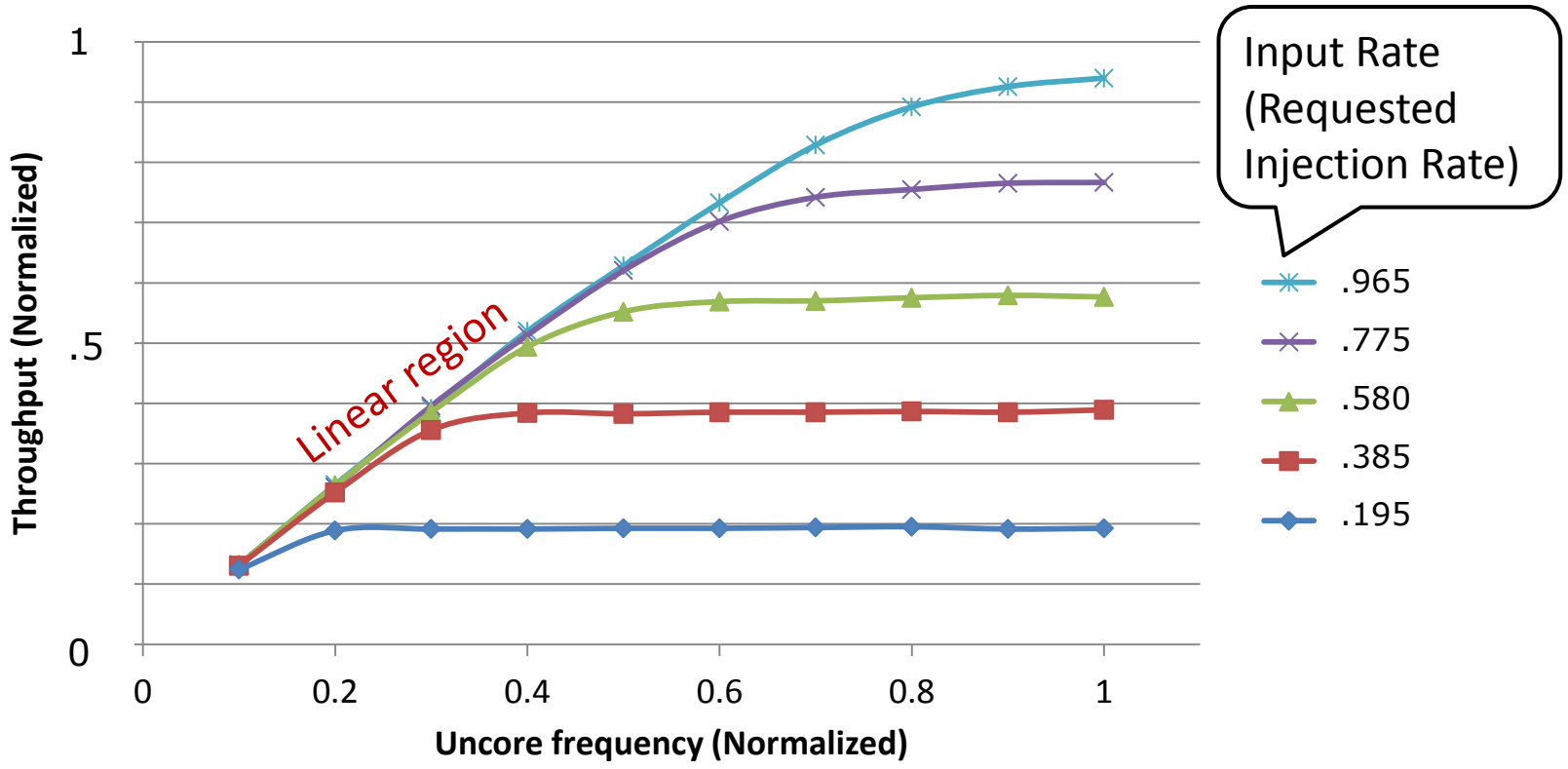
Workload Variation



- Injection rate changes significantly both
  - During the run-time of applications and
  - Across different applications
- Designing the uncore for the worst-case may cause the uncore operate faster than necessary for most of the time

Xi Chen, Hyungjun Kim et al. DAC 2013

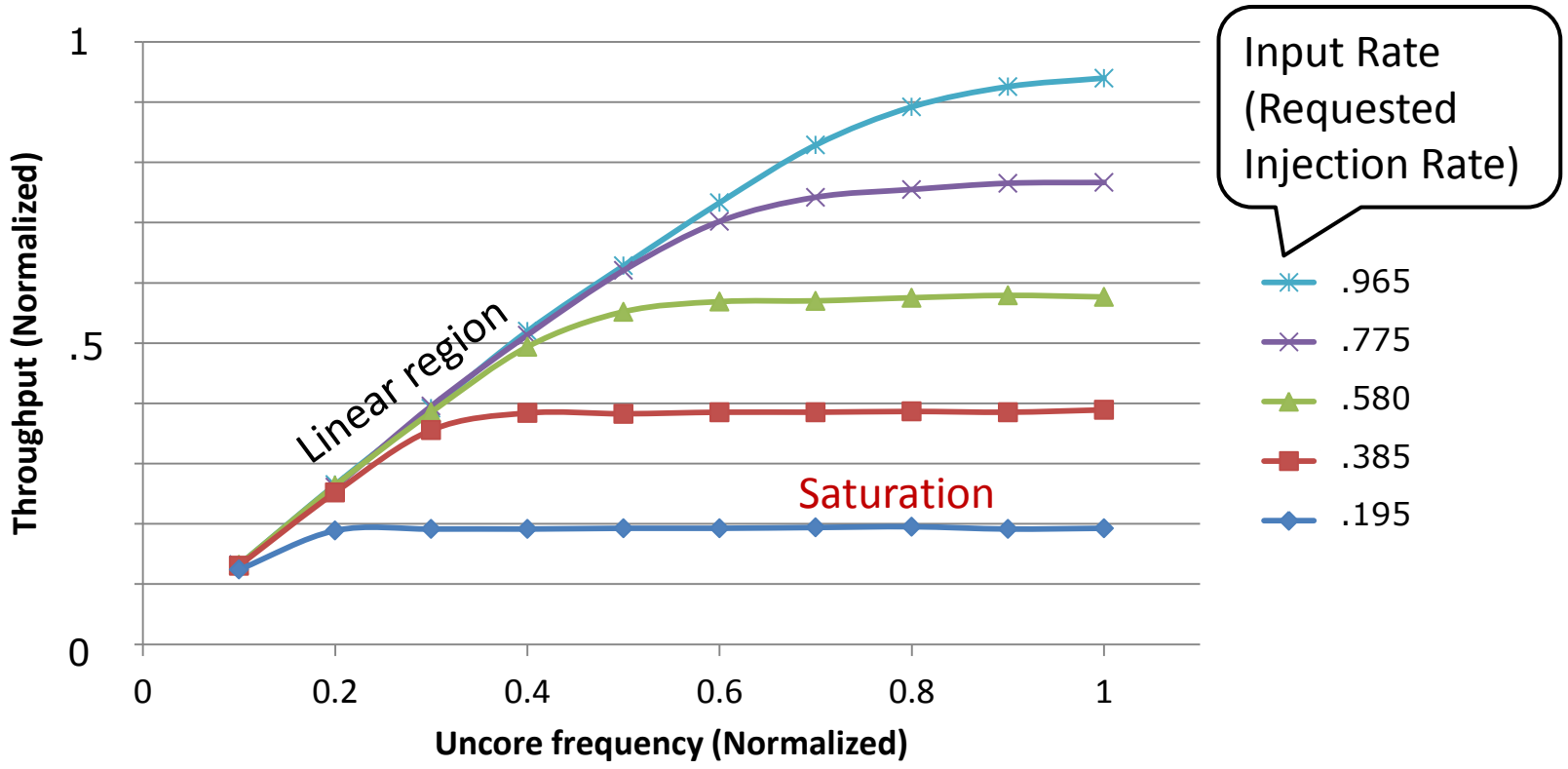
Core Bandwidth-Throughput Characteristics



- ***Linear region:*** *Almost proportional increase in throughput with increasing frequency*

Xi Chen, Hyungjun Kim et al. DAC 2013

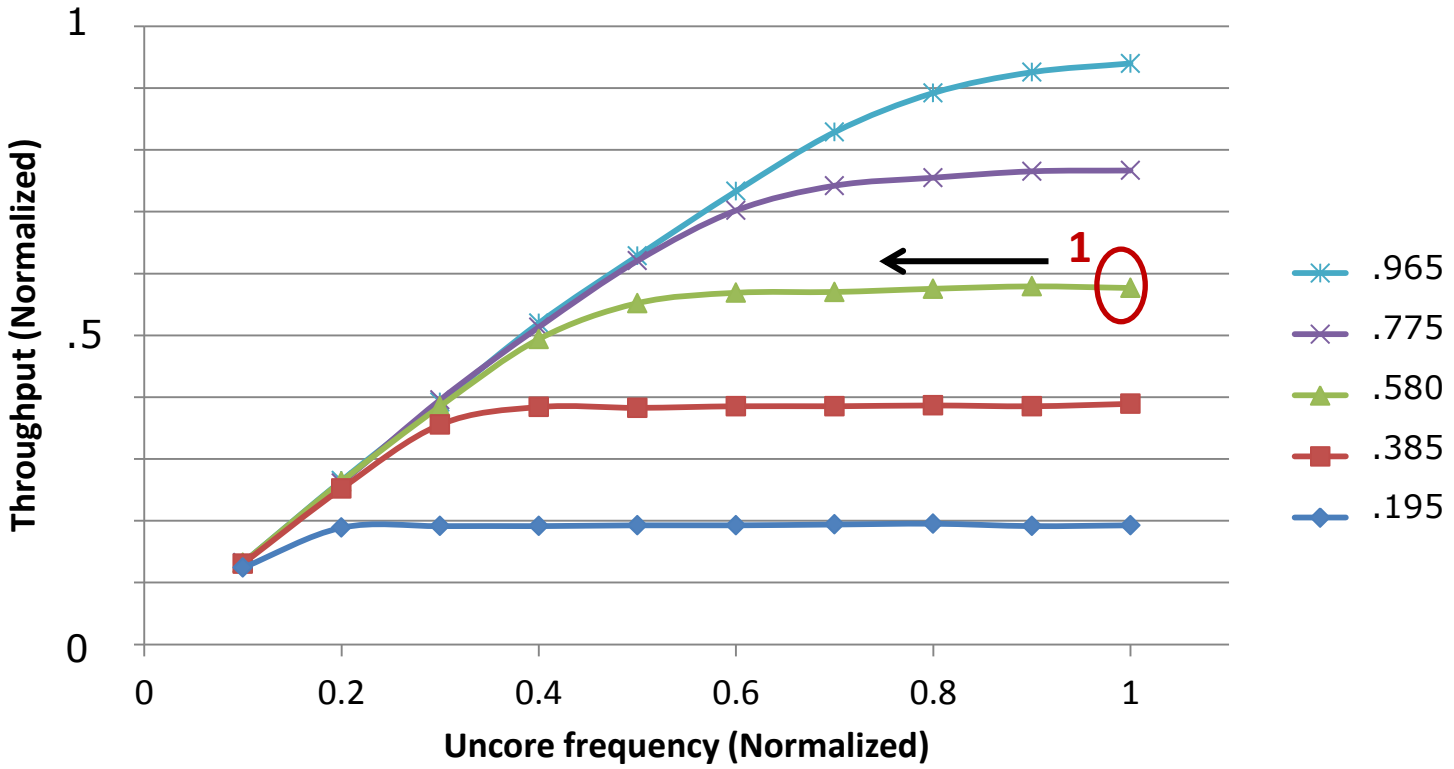
Core Bandwidth-Throughput Characteristics



- **Linear region:** Almost proportional increase in throughput with increasing frequency
- ***Saturation due to requested injection rate***

Xi Chen, Hyungjun Kim et al. DAC 2013

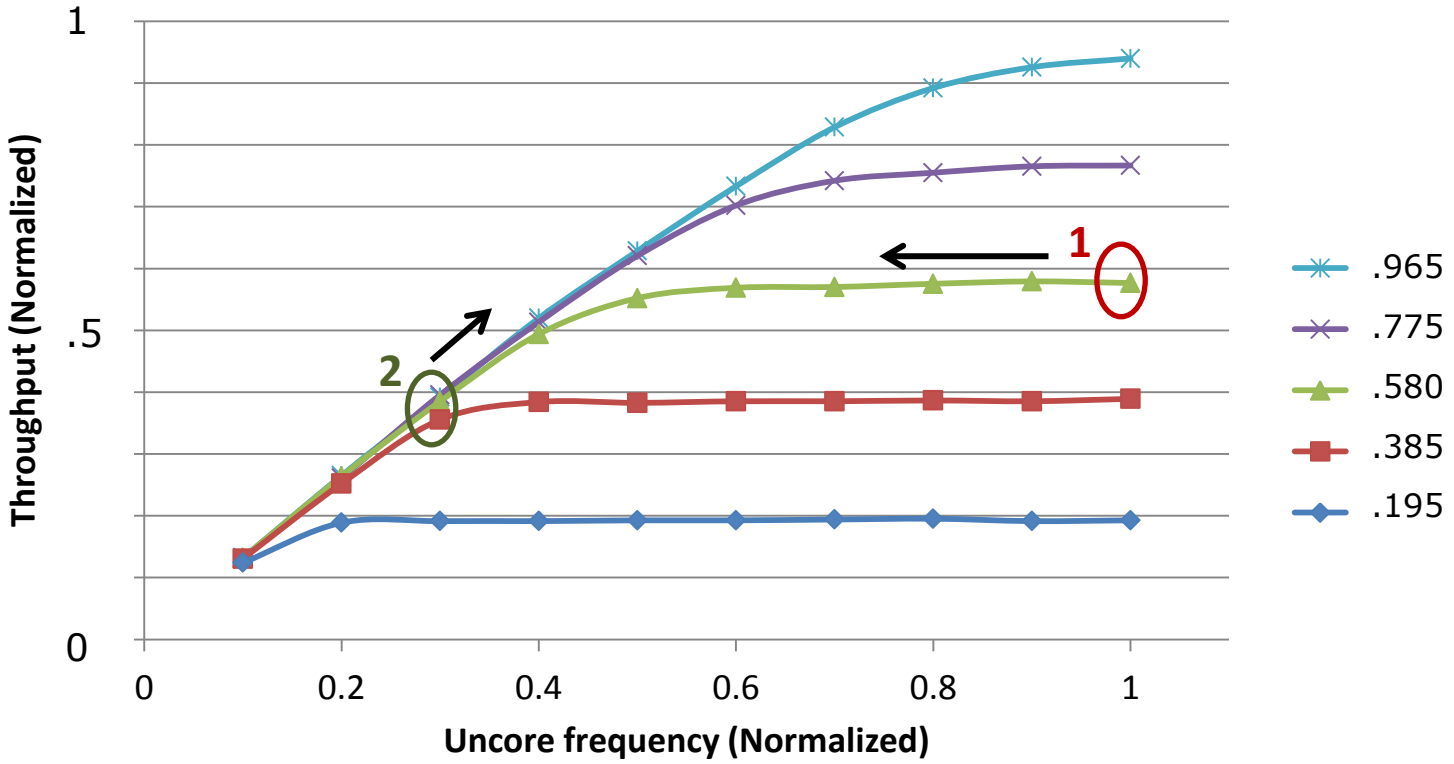
Power/Performance Implications



- **Point 1:** Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput

Xi Chen, Hyungjun Kim et al. DAC 2013

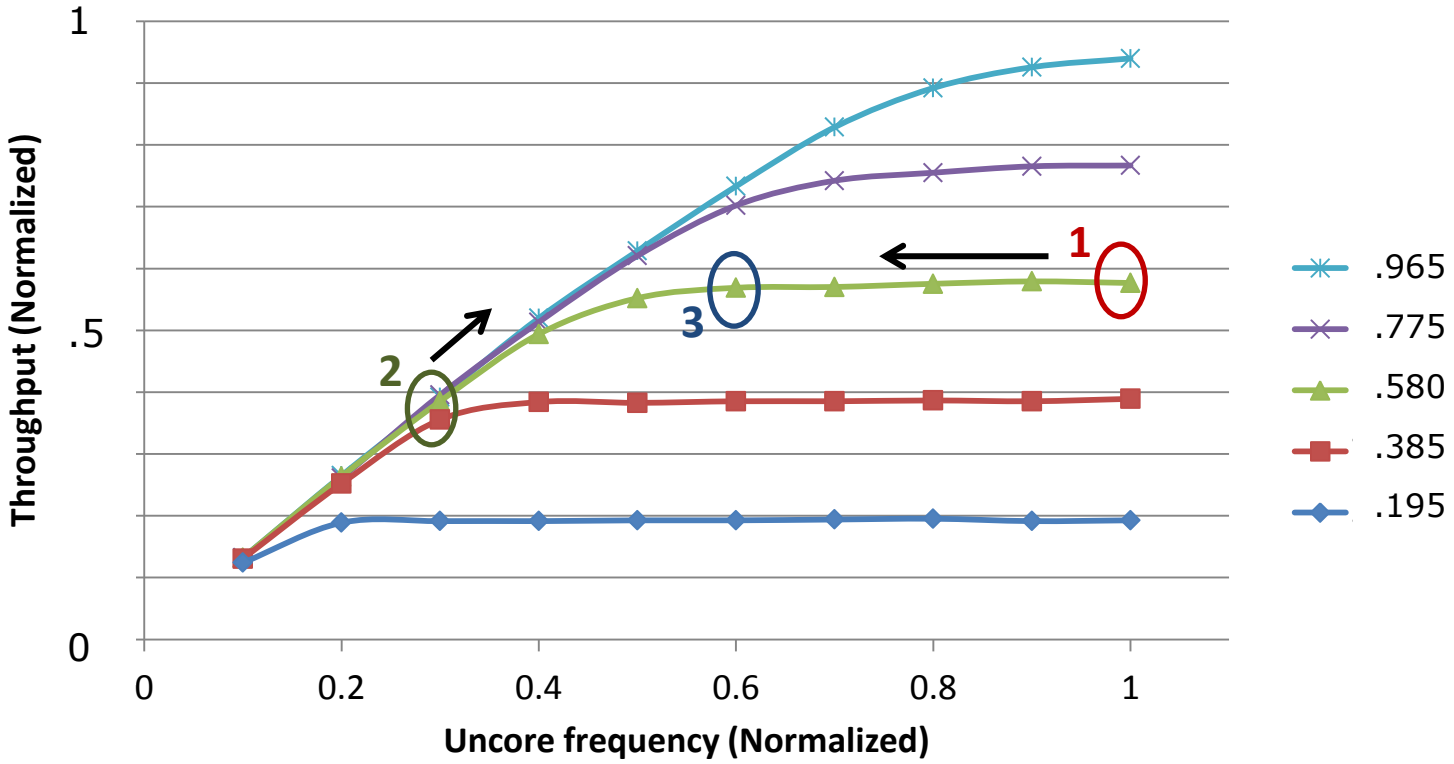
Power/Performance Implications



- **Point 1:** Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput
- **Point 2:** Performance requirement is *not* met. Frequency *should* be increased

Xi Chen, Hyungjun Kim et al. DAC 2013

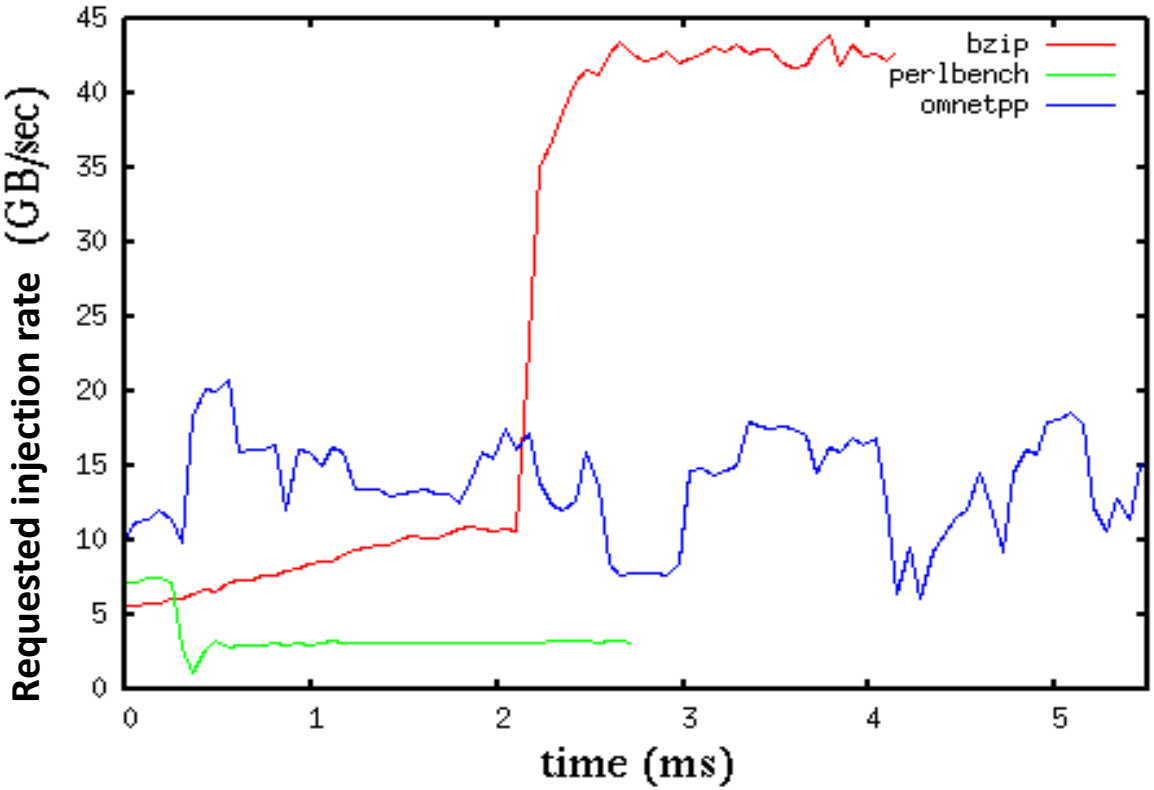
Power/Performance Implications



- **Point 1:** Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput
- **Point 2:** Performance requirement is *not* met. Frequency *should* be increased
- **Point 3:** "Optimum" for this load.

Xi Chen, Hyungjun Kim et al. DAC 2013

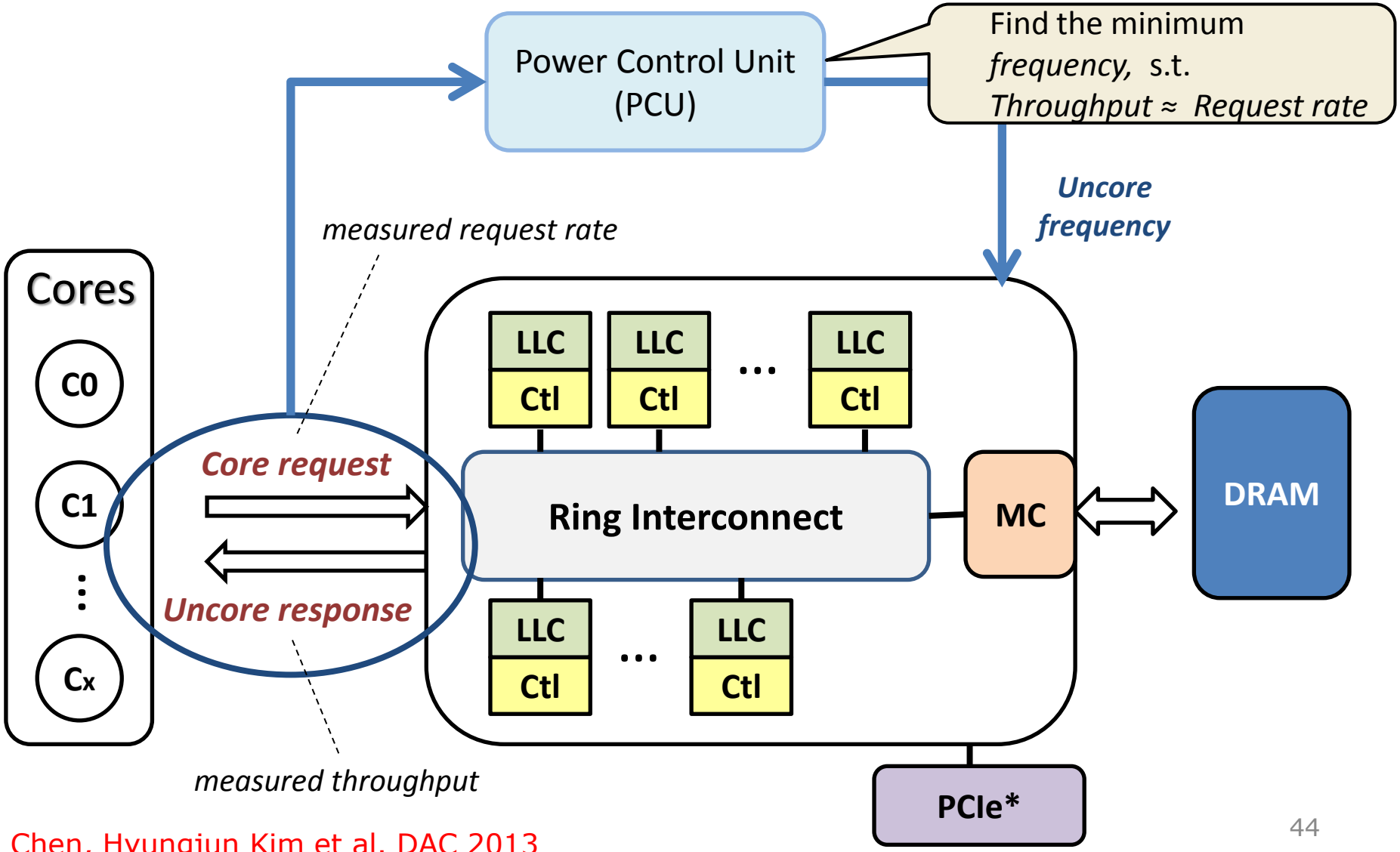
Power/Performance Implications



- **Point 1:** Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput
- **Point 2:** Performance requirement is *not* met. Frequency *should* be increased
- **Point 3:** "Optimum" for this load. Optimum point changes dynamically!

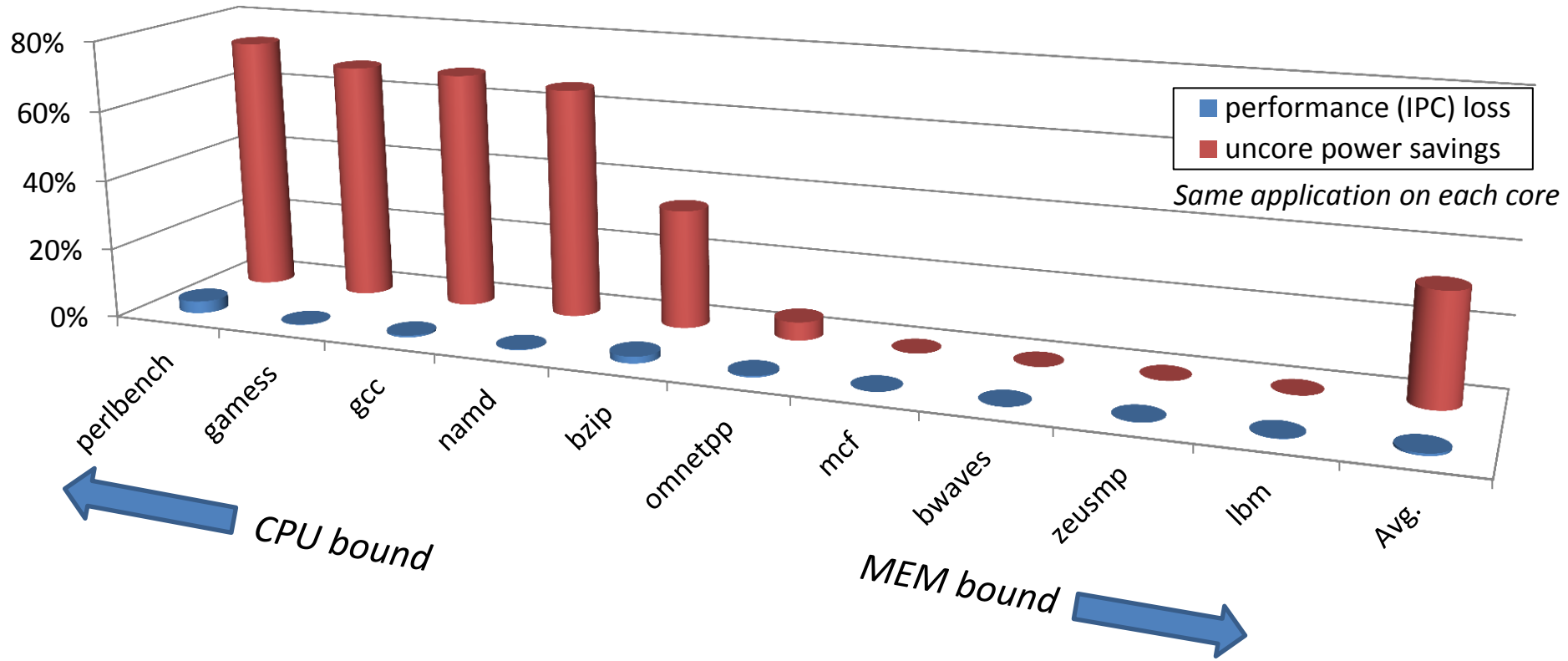
Xi Chen, Hyungjun Kim et al. DAC 2013

Objective



Xi Chen, Hyungjun Kim et al. DAC 2013

Improving Power Efficiency



- **31%** average uncore power savings at minimal perf. cost of **0.6%**
- Up to 73% power savings at 3.5% perf. cost
- Almost no effect on memory bound apps

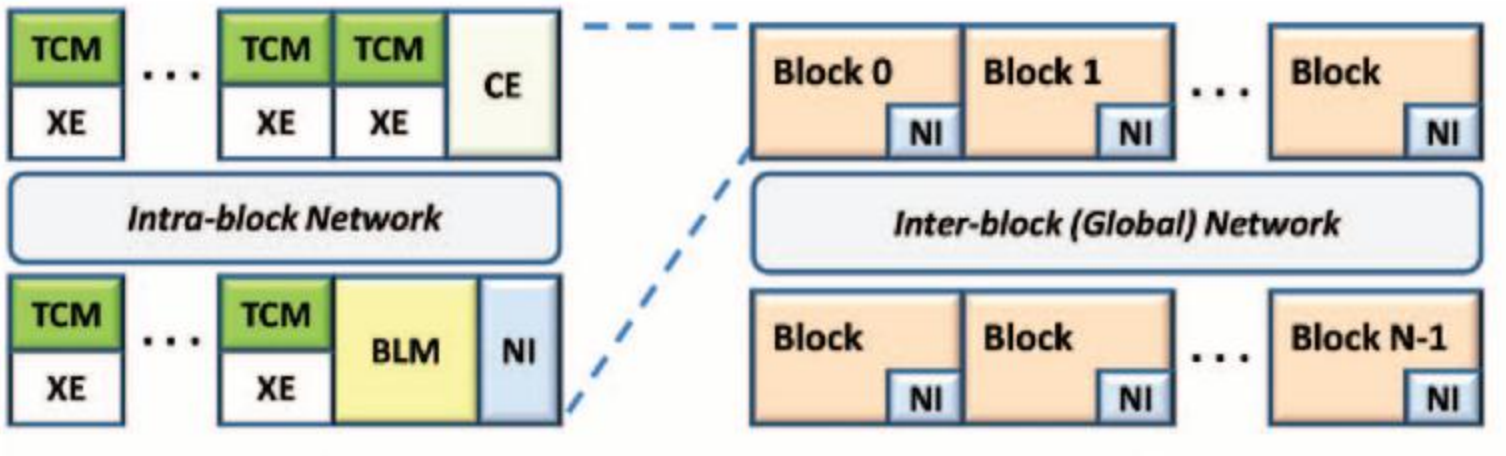
Xi Chen, Hyungjun Kim et al. DAC 2013

Exascale computing

Goals:
Exascale performance: 10^{18} flops
Low energy: 20 pJ/flop (including moving data) -> 20MW per system

John Shalf et al. "Exascale Computing Technology Challenges" VECPAR 2010, LNCS 6449

UHPC experiment



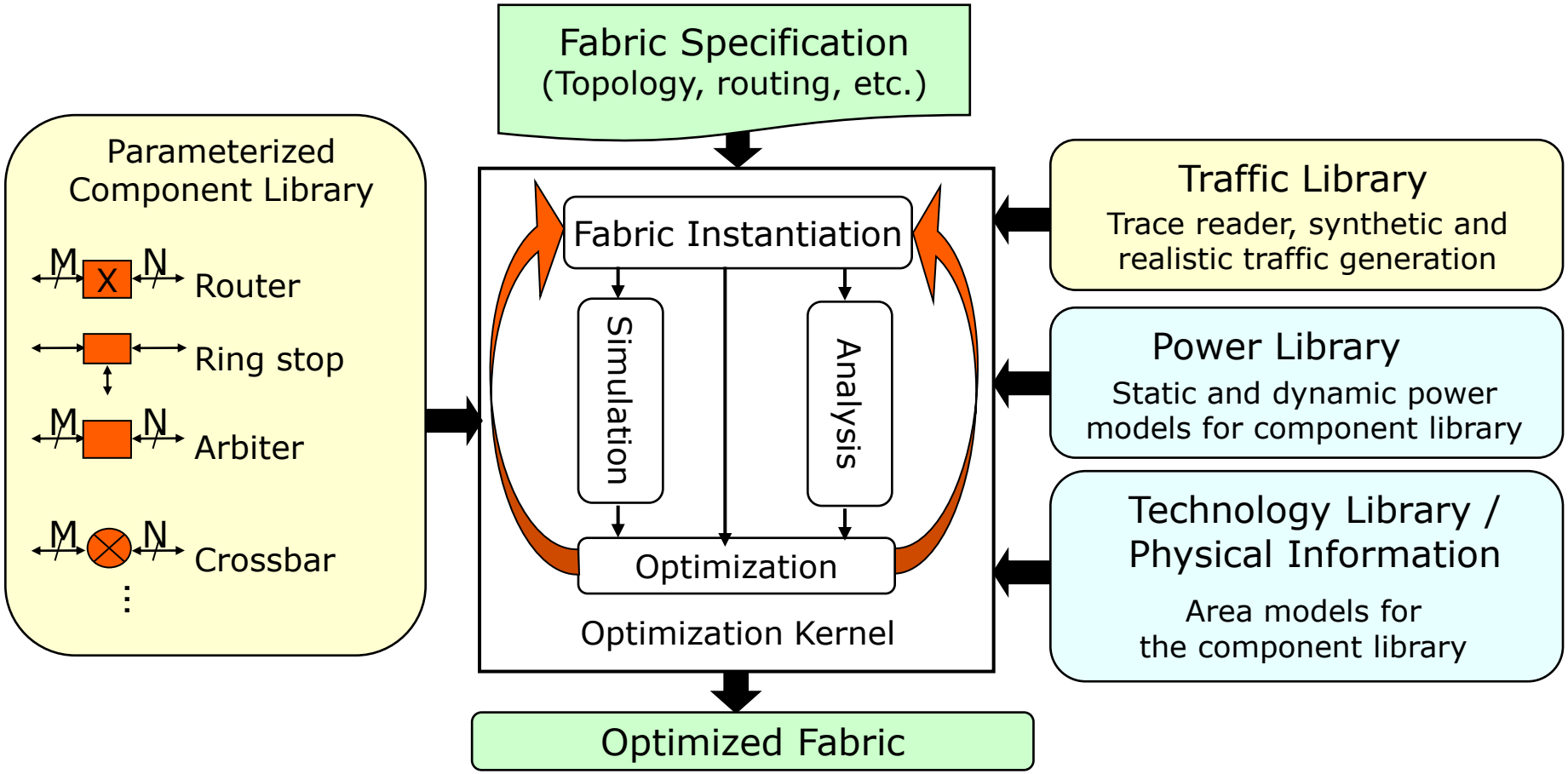
cluster

die

Baseline: 720 nodes on die

Ogras et al. "Energy-Guided Exploration of On-Chip Network Design for Exa-Scale Computing" SLIP 2012

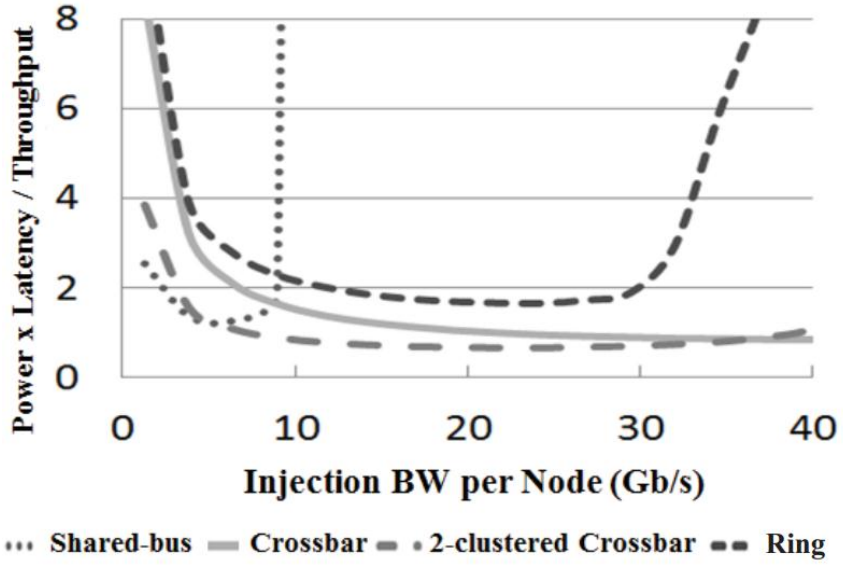
Exploration tool



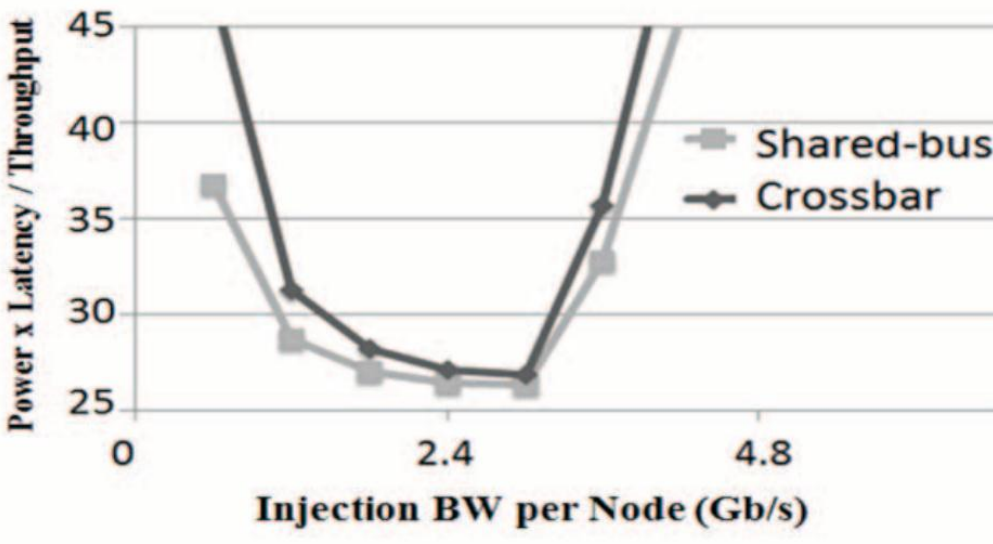
Metric: power x latency / throughput = energy x delay/bit for networks (Jxsec/bit)

Ogras et al. "Energy-Guided Exploration of On-Chip Network Design for Exa-Scale Computing" SLIP 2012

Balancing hierarchical interconnect

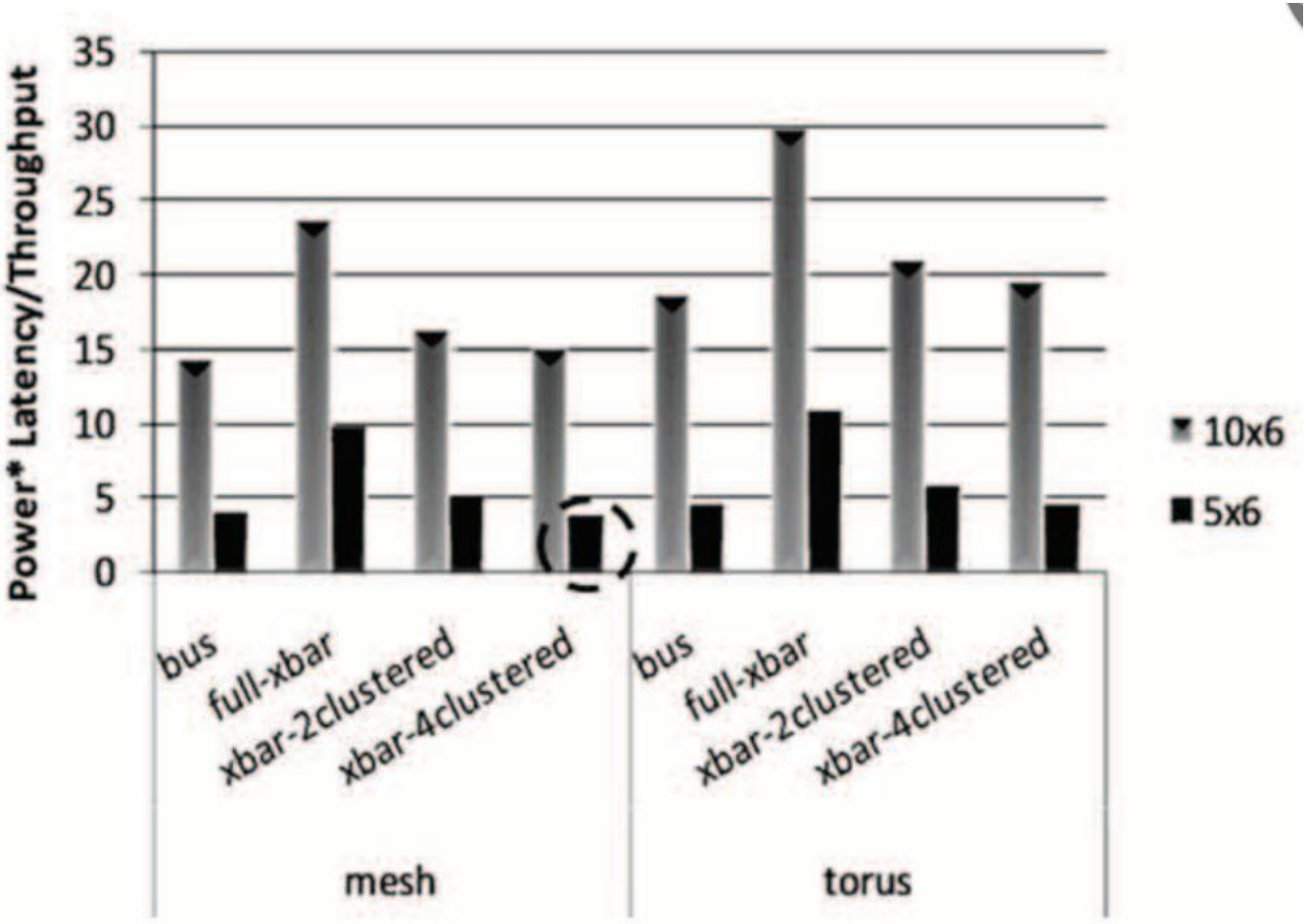


Intra-block network



Total system network with 10x6 2D mesh for inter-block network (12 nodes cluster)
Uniform traffic

Evaluation on SAP from SPEC2000



24-node clusters more energy-delay efficient than 12-node

Life with Dark Silicon

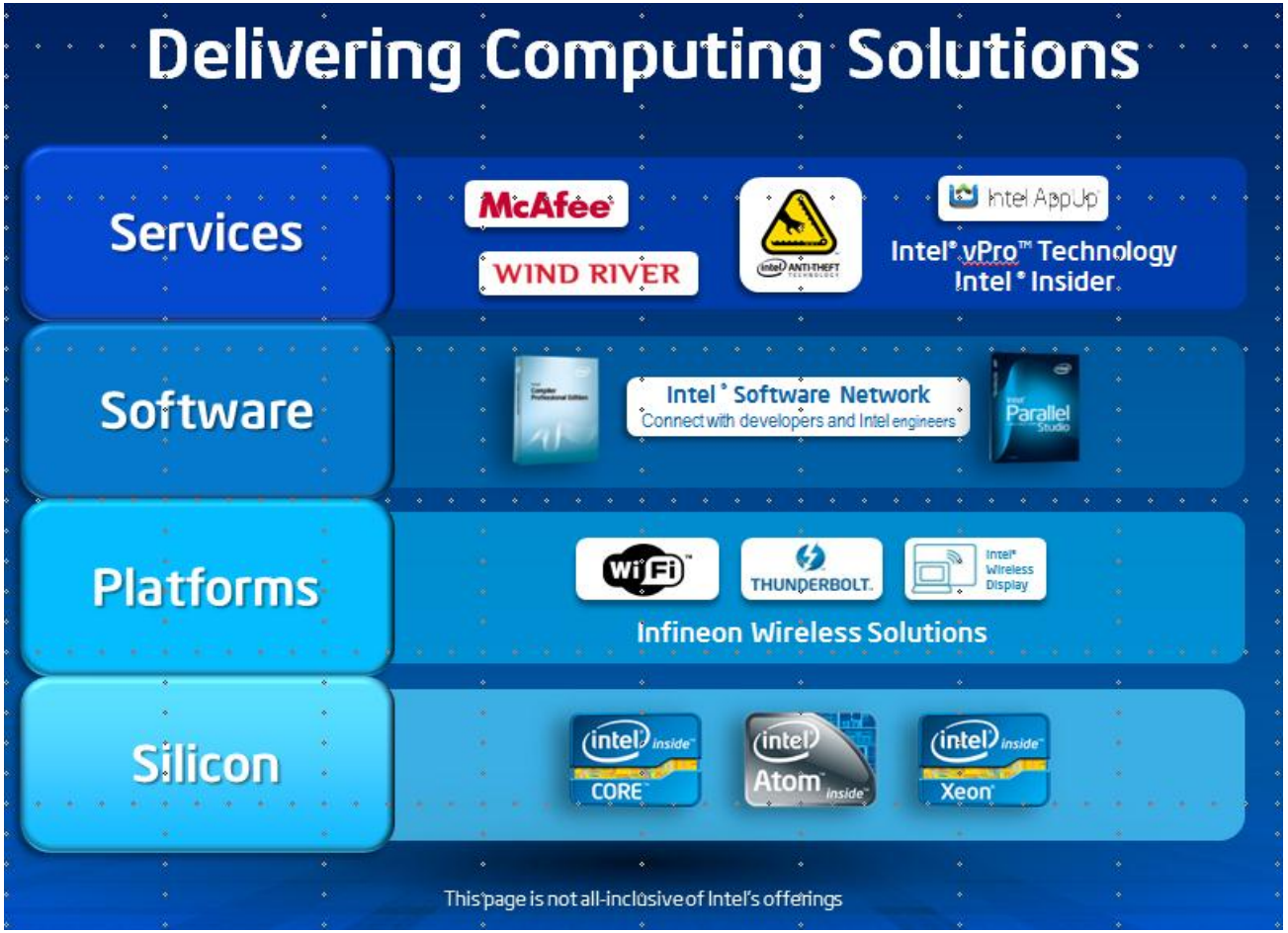
- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization

Why system-level?

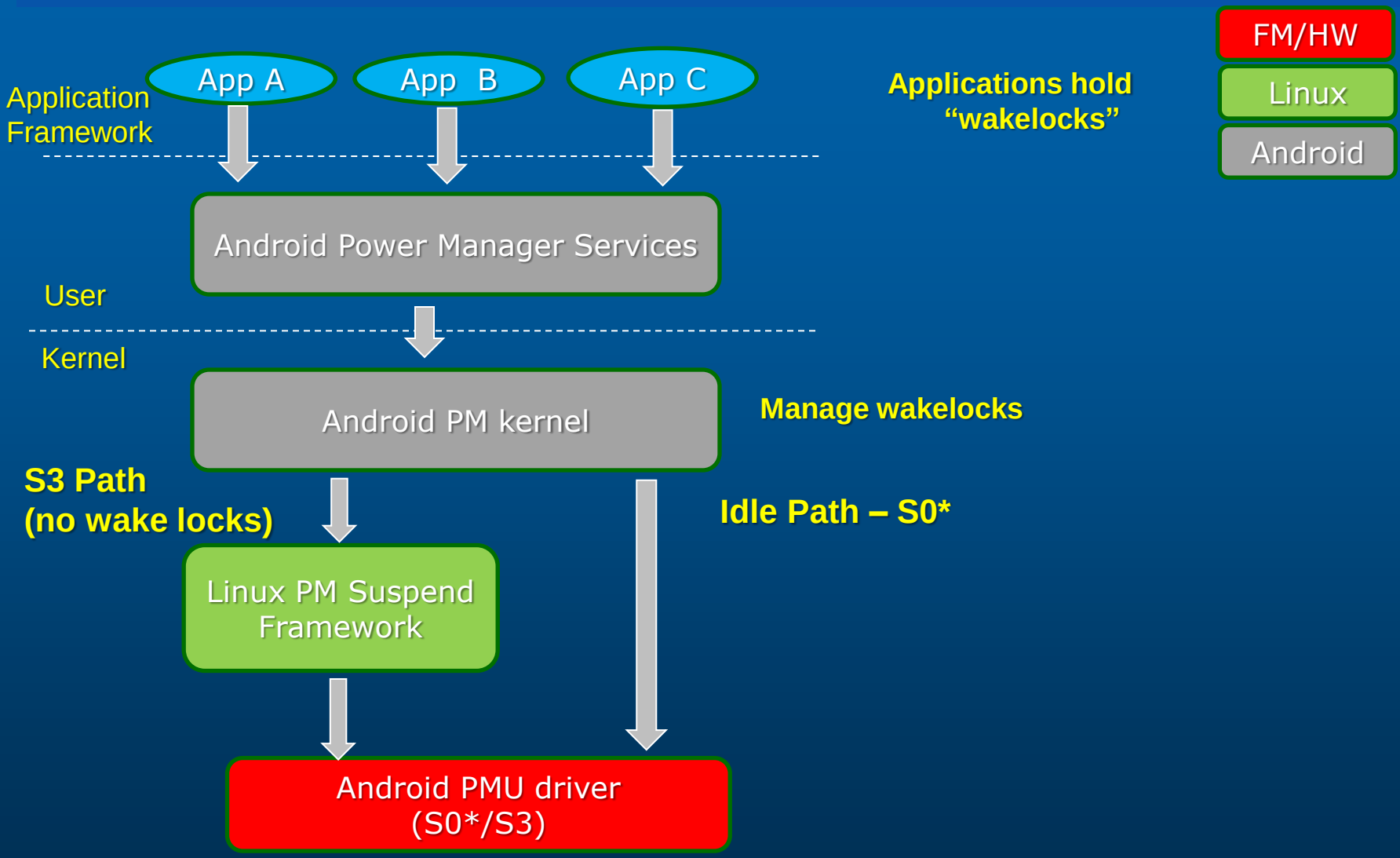
What matters is the final product

Need a system view: SW stack + Platform + Silicon components

March 2012



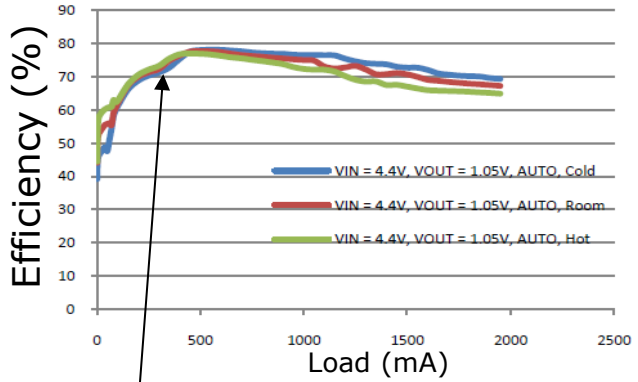
Android Power Management Stack

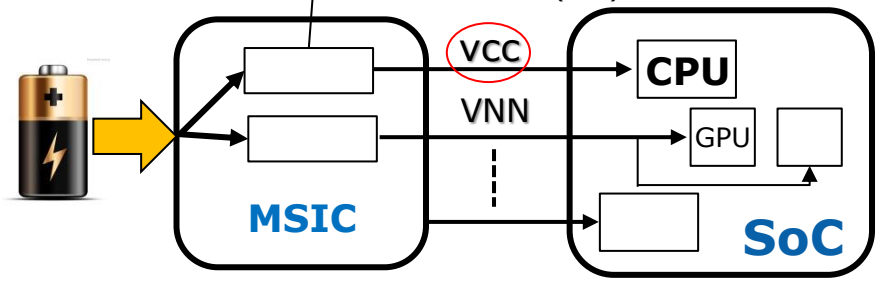


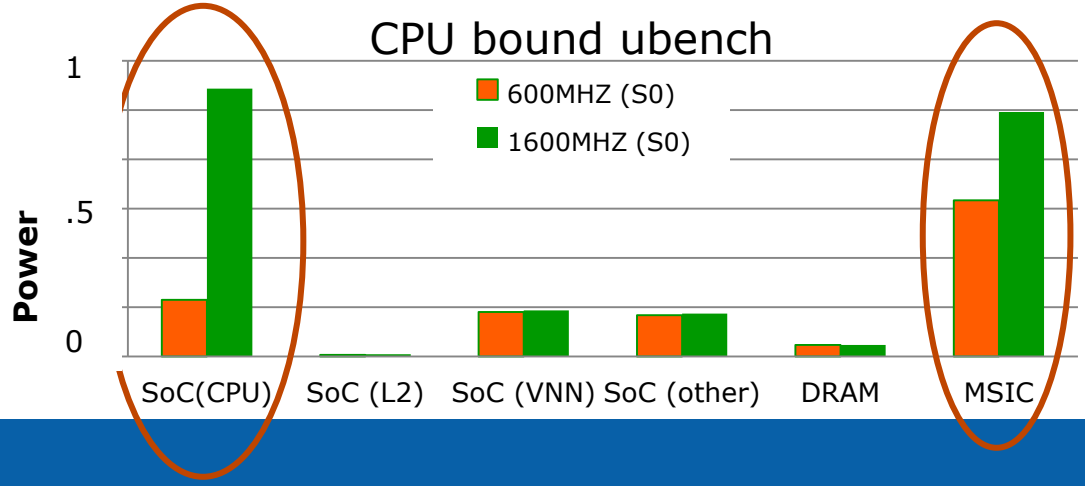
System level power delivery

- Power delivery uses significant power
- *Case study*: Increasing CPU frequency-voltage leads to:
  - Higher **SoC** power
  - Higher **MSIC** power
- Ignoring inefficiency in power delivery leads to inaccurate power estimations

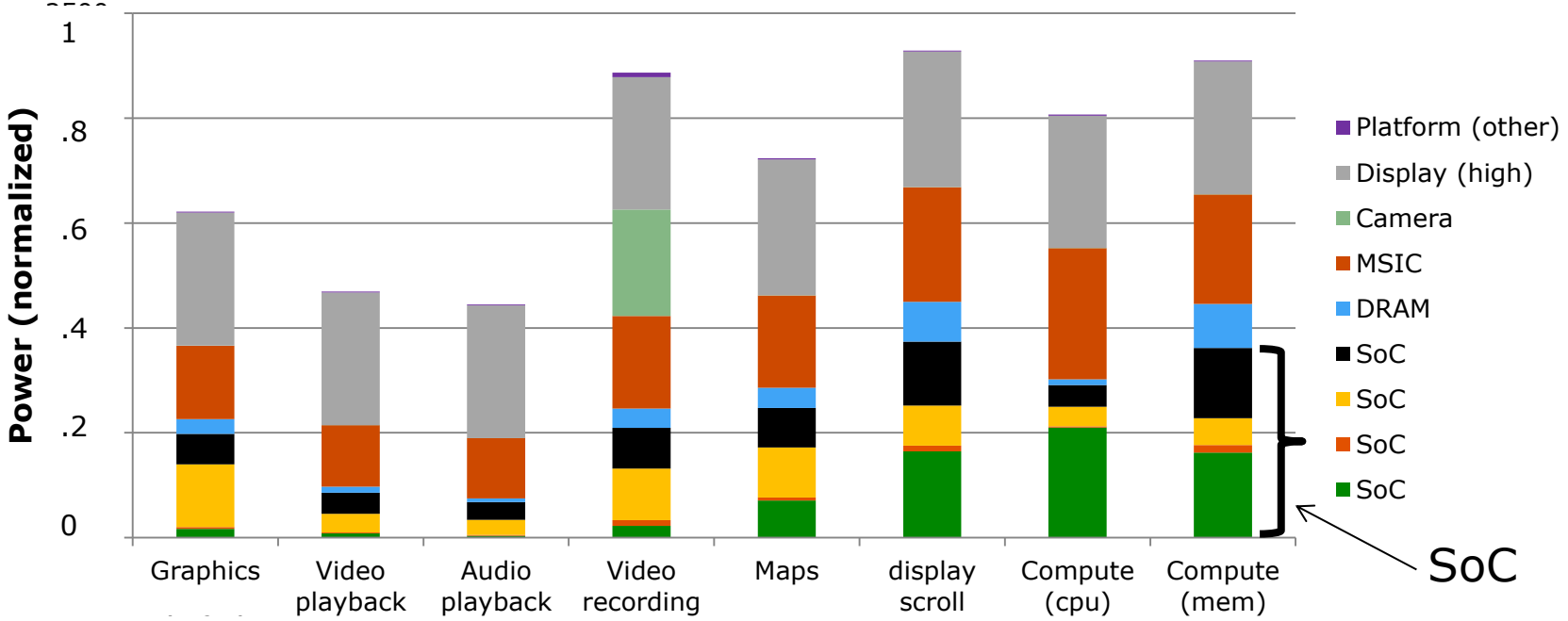
Power optimizations must be applied at the (*system level*)







Example of system level power breakdown

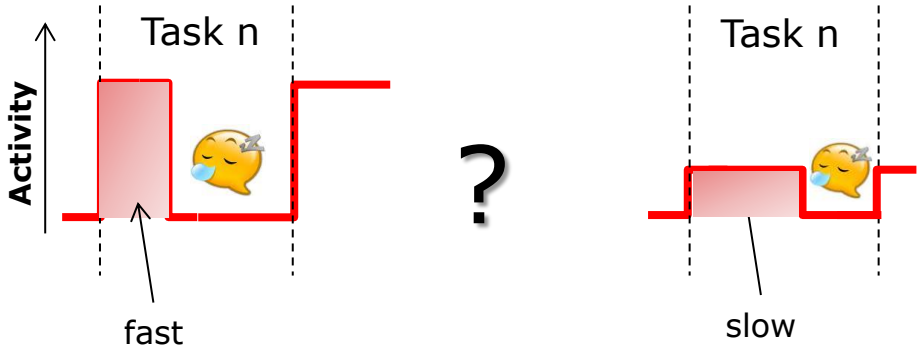


- CPU does not dominate the SoC
- SOC is not the biggest contributor of power
- MSIC dissipates significant portion of platform power

Power savings while using the phone

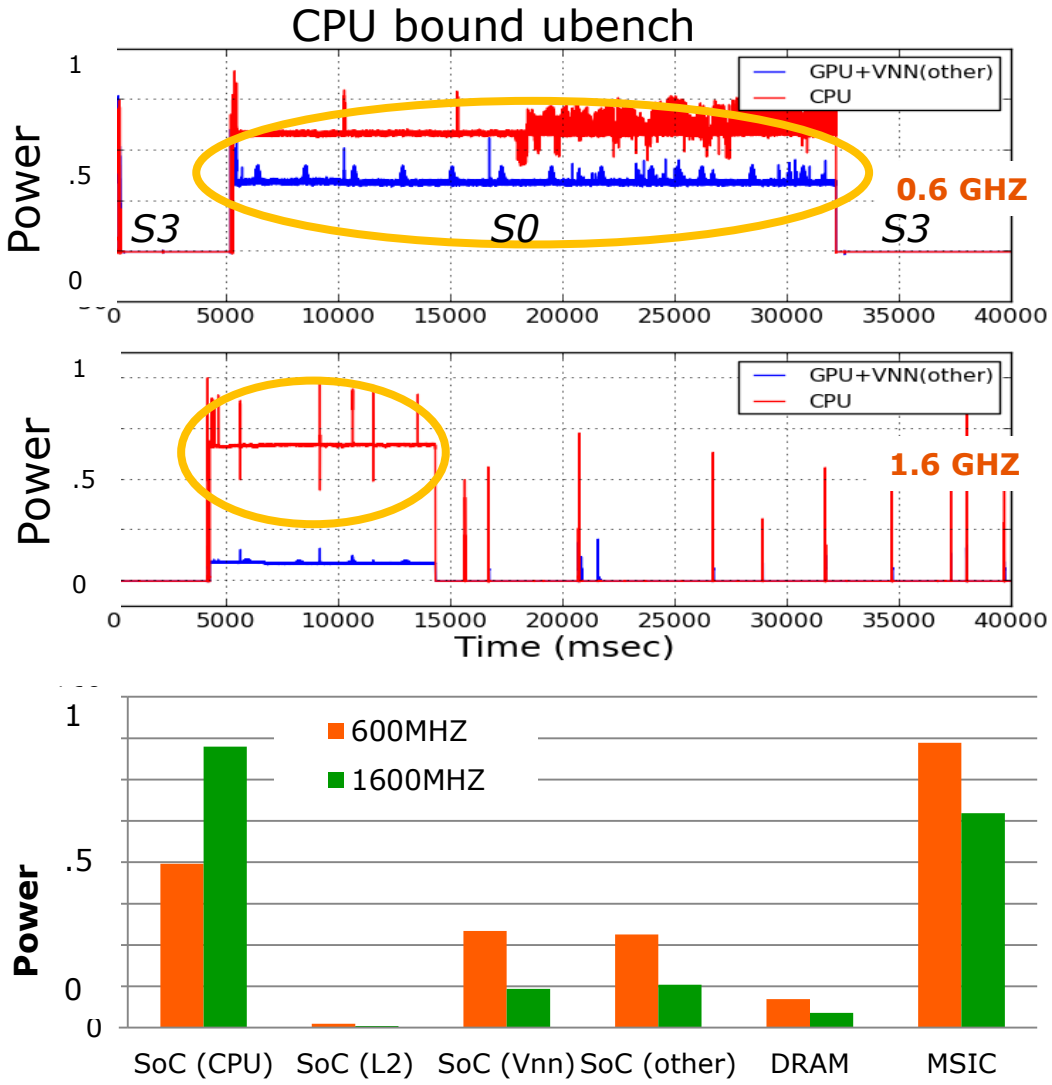
- Assume current task has a slack
  - **Utilize slack to optimize for execution speed**

    Which scenario is better for energy savings?
    - Execute fast and entertain a longer sleep
    - Run slow to reduce dynamic power at the cost of a shorter sleep



Power savings: run CPU faster vs. slower

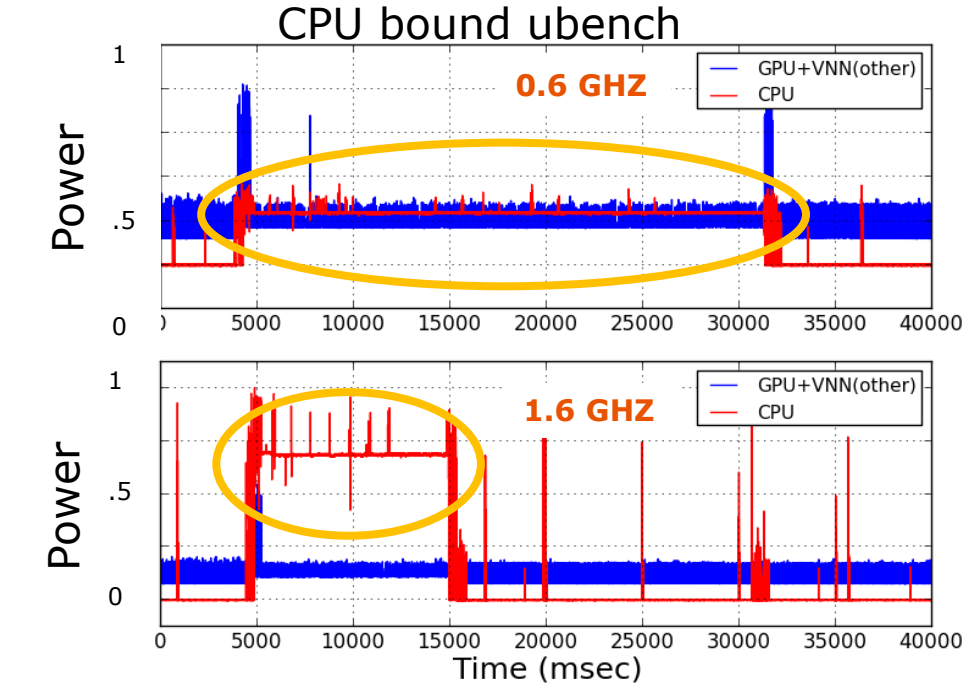
- **Case 1**: Running a task at two frequencies, enter *S3* when no execution (using *wait* function in Java app)
- Running the CPU faster increases its average power
- Still save at the platform level when CPU runs faster
  - Savings come from putting the platform in a deep system state for a longer time

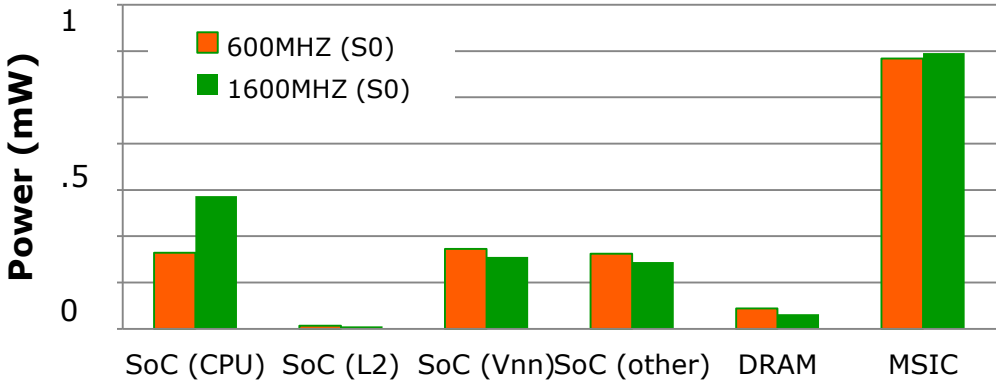


Power savings: run CPU faster vs. slower

- **Case 2**: Running a task at two frequencies, <u>always run in *S0*</u> (long setting on display ON prevents entering S3)
- Running the CPU slower reduces its average power
- Running slower saves power at the platform level
  - Savings in this case come from utilizing DVFS to reduce dynamic power

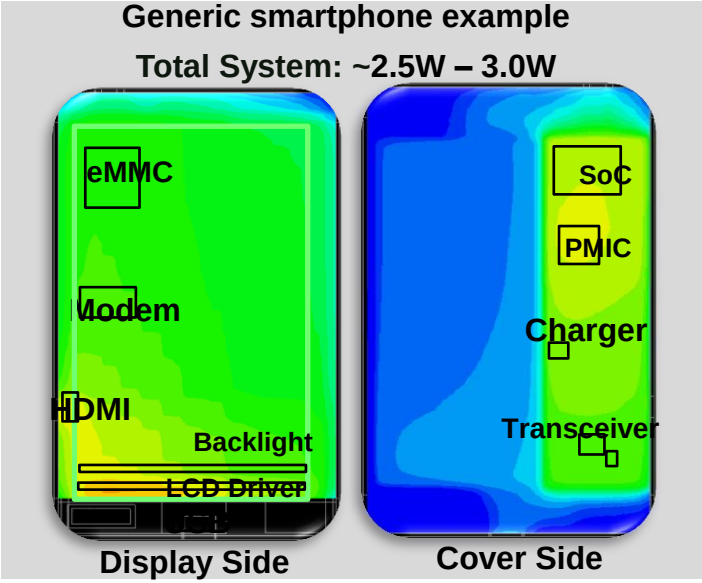
Finding optimal speed is challenging, requires system level optimizations



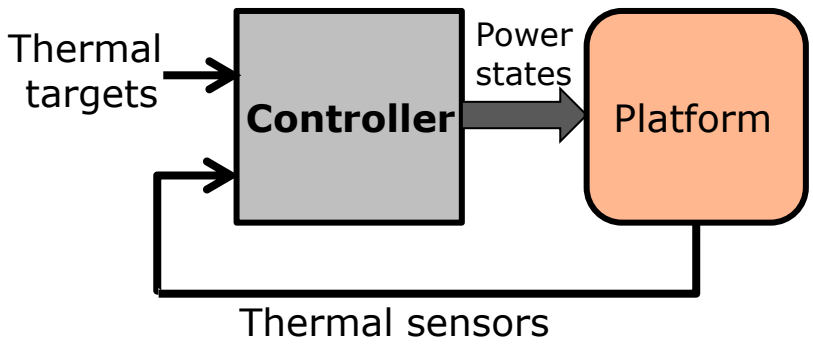


Challenge: Temperature estimation and management

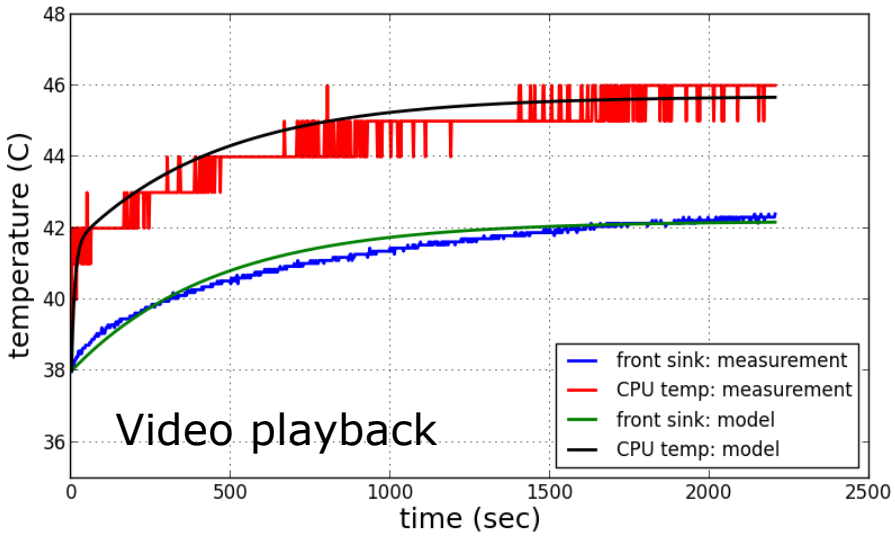
- **Current skin estimation models are off**
  - Challenging problem due to the complexity of thermal distribution in space and time and limitation in number of thermal sensors
  - How to place thermal sensors to maximize thermal observability?
  - Can we create accurate thermal models that are simple enough for run time estimations?

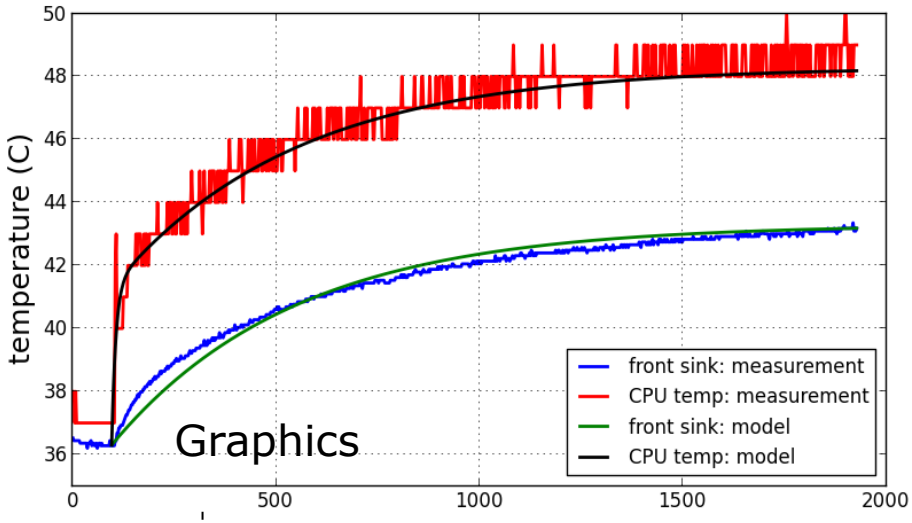


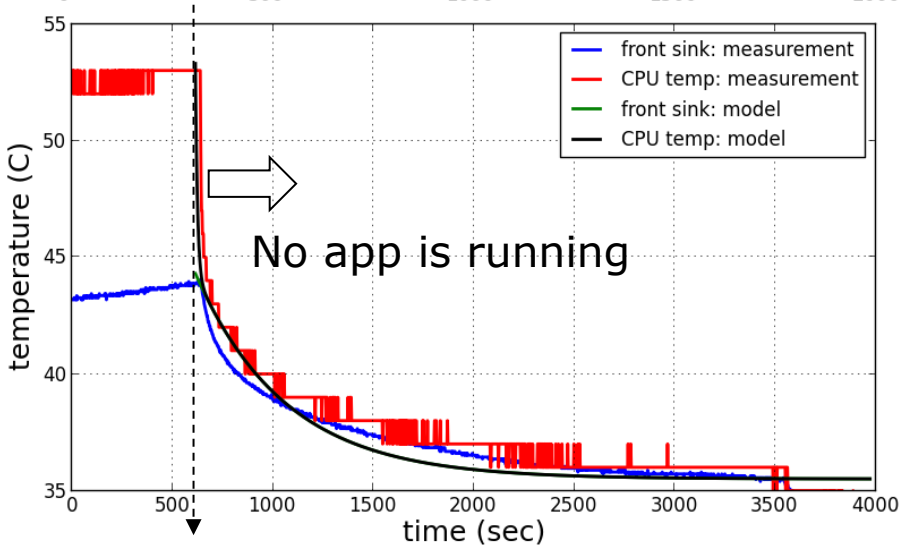
- **Temperature management**
  - Current thermal management techniques are reactive and heuristic
  - Better methods (e.g. MIMO control) are needed to ensure optimality



Model validation (simple phone)







- The model is able to track the temperature of skin and SoC's die
- Skin time constant is much larger than die

Temperature prediction

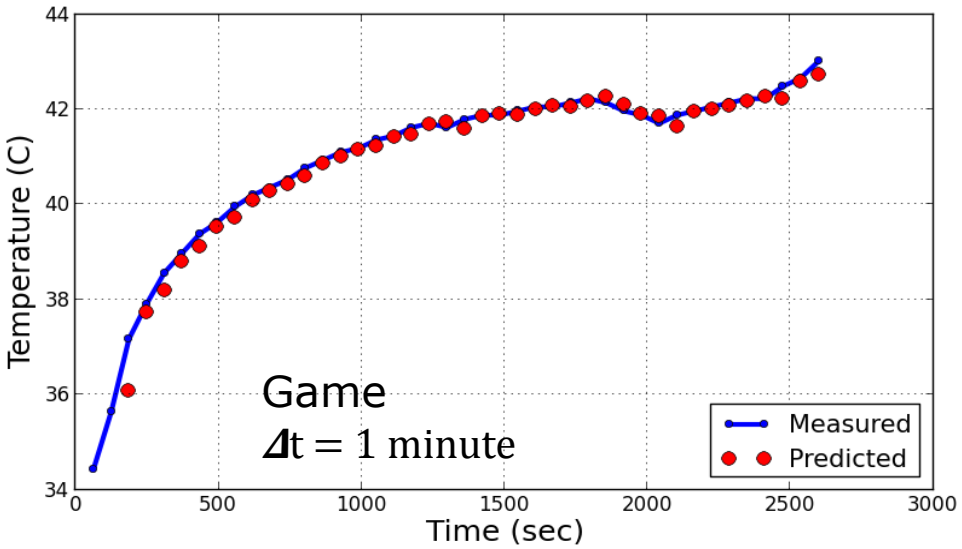
- Proactive thermal managements provide better thermal decisions than reactive*
- A temperature predictor* can be formulated using our thermal model

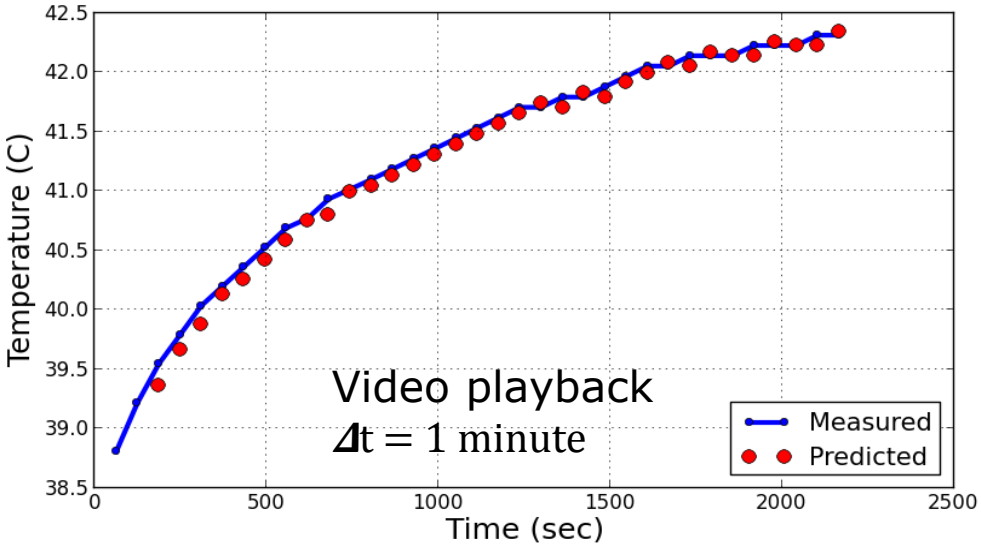
$$T_{skin}(\mathrm{K+1}) = T_{skin}(\mathrm{k}) + (T_{skin}(\mathrm{k}) - T_{skin}(\mathrm{k-1}))\mathrm{e}^{\frac{-\Delta t}{\tau}}$$

$$\tau = R_{skin}C_{skin}$$

Δt : prediction distance

- Good prediction can be achieved

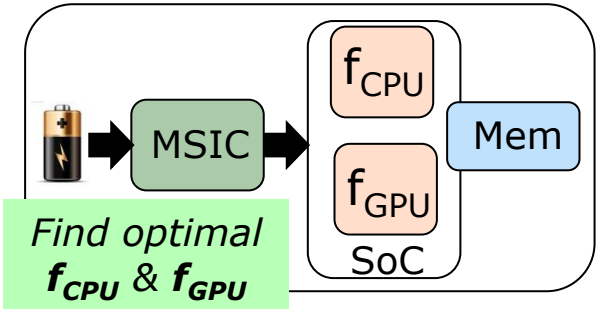




*R. Ayoub, et al "Energy Efficient Proactive Thermal Management in Memory Subsystem," ISLPED 10.

Challenge: System level power/energy optimizations under given constraints

- How to distribute power between components to *maximize performance while meeting power/thermal constraints*
  - Runtime optimizations at the system level
    - Low overhead
    - Power, performance and thermal models
    - Better observability of runtime characterization



- Dual problem: *optimizing power/thermal under given performance constraints*

Summary

- Modern silicon systems are "dark silicon"
  - Cannot operate all parts of the system at full speed simultaneously for a long time
  - This trend increases in the future systems
- Process technology continue delivering breakthrough improvements. However it is not sufficient alone
- Dimming technologies (thermal and power management) getting more elaborate
  - Applying control theory instead of heuristics seem promising
  - Hetero helps if can solve SW thread migration
- Specialization in the form of accelerators is used in all modern SoC
  - Higher energy efficiency
  - Automatic support for quick generation of accelerators and IP blocks is needed
- Need optimization at system-level (software, platform) not just SoC
- Shrinking die size without increase in functionality is an option, but not economically sustainable