LA-UR-15-25938
Approved for public release; distribution is unlimited.

Title: Measurement and Characterization of Haswell Power and Energy Consumption

Author(s): Huang, Song
Lang, Michael Kenneth
Pakin, Scott D.
Fu, Song

Intended for: NMC seminar

Issued: 2016-02-19 (rev.1)

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Measurement and Characterization of Haswell Power and Energy Consumption

Song Huang, Michael Lang, Scott Pakin, and Song Fu

Los Alamos National Laboratory

University of North Texas

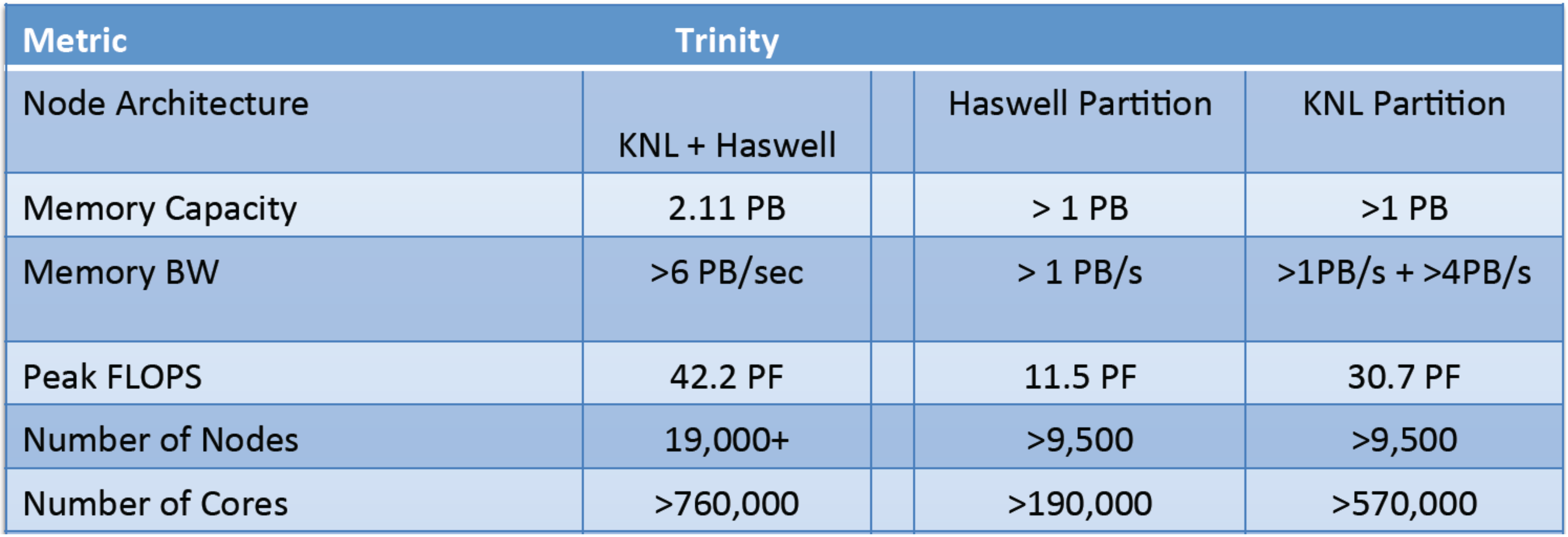
Summer 2015

Outline

- Haswell Architecture & Attractions
- Power-Performance Characterization
  - Four benchmark applications, instrumentation tools
  - Power control knobs
  - Experimental results & findings
- Summary
- Future Plans
- Acknowledgements

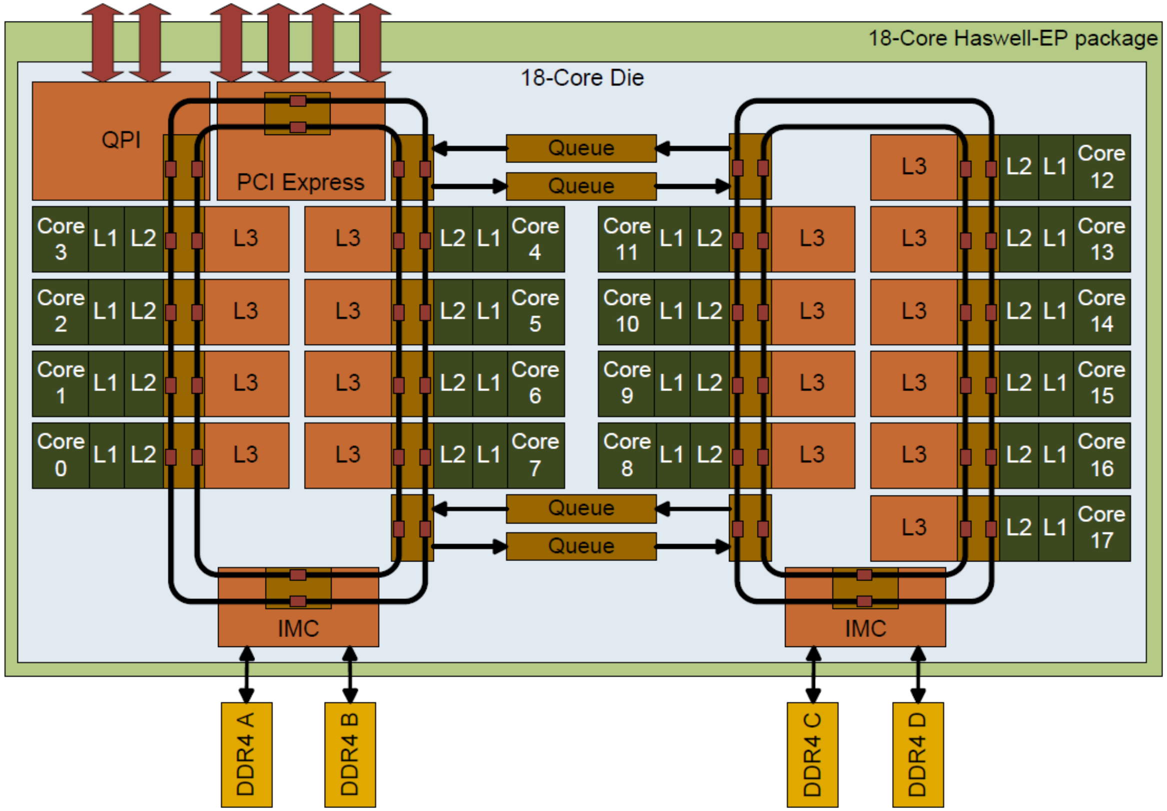
Haswell in Trinity

http://www.lanl.gov/projects/trinity/_assets/docs/trinity-overview-for-web.pdf



Metric	Trinity			
Node Architecture	KNL + Haswell		Haswell Partition	KNL Partition
Memory Capacity	2.11 PB		> 1 PB	>1 PB
Memory BW	>6 PB/sec		> 1 PB/s	>1PB/s + >4PB/s
Peak FLOPS	42.2 PF		11.5 PF	30.7 PF
Number of Nodes	19,000+		>9,500	>9,500
Number of Cores	>760,000		>190,000	>570,000

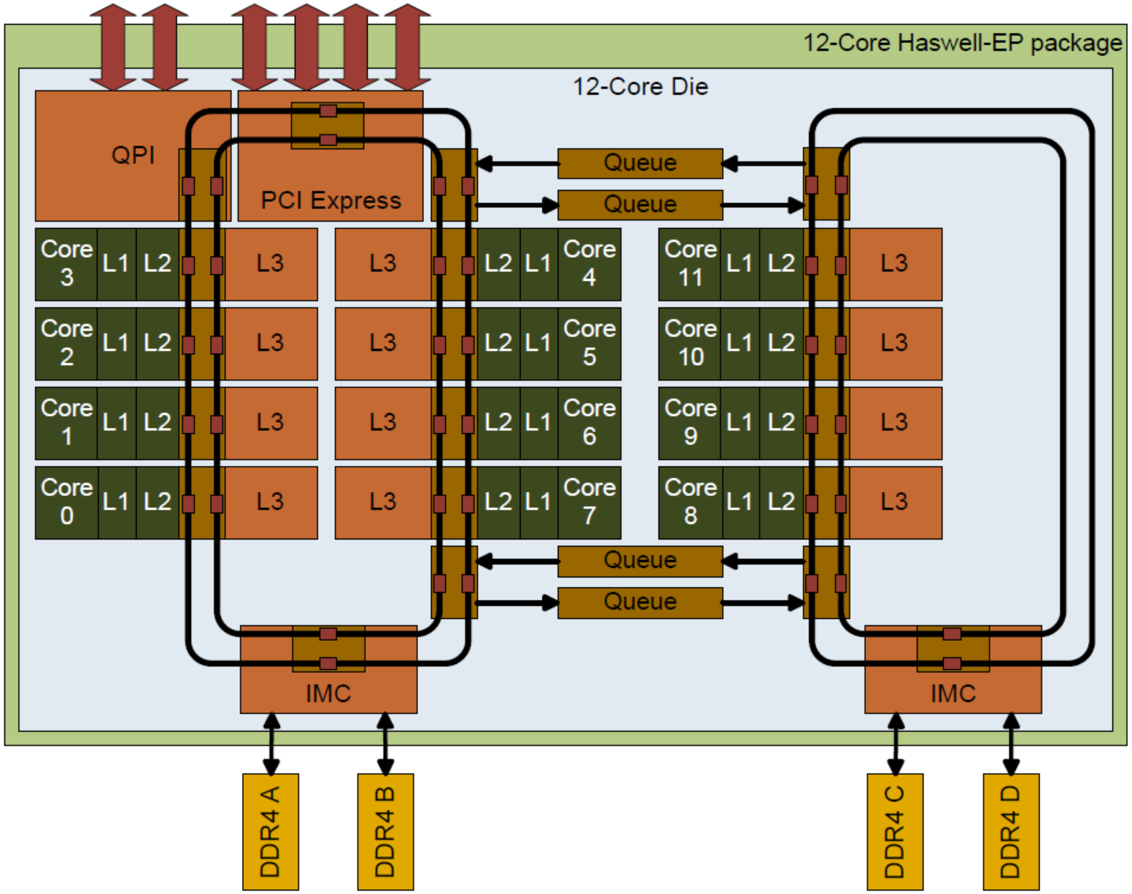
Haswell EP Microarchitecture



- HCC: 14-18 cores
- Bi-directional full rings
- Connected by queues for data transfer
- 8 + 10 cores
- 4 columns
- 2 memory controllers, 2 memory channels each, support DDR4
- SIMD ISA: AVX2
- 16 FLOPS/cycle (double)
- 9.6 GT/s QPI

https://software.intel.com/en-us/articles/intel-xeon-processor-e5-2600-v3-product-family-technical-overview

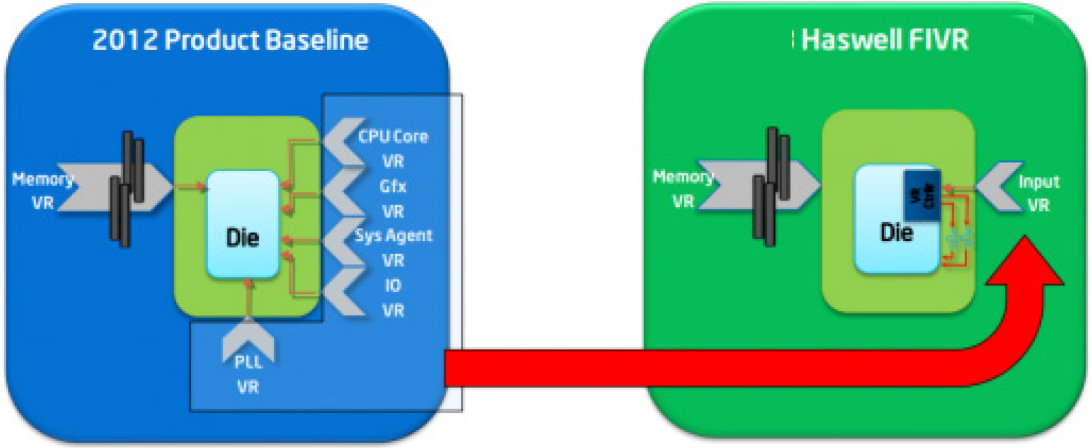
Haswell EP Microarchitecture



- MCC: 10-12 cores
- 1.5 rings
- 2 partitions
- 8 + 4 cores
- 2 memory controllers

- Xeon E5-2660 v3
- 10 cores

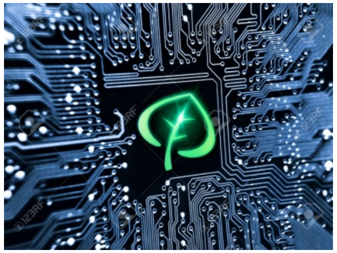
Fully Integrated Voltage Regulator



- Input voltage: sends serial voltage ID (SVID) signals to mainboard VR (MBVR)
- SVID regulates V_{ccin}
- MBVR supports 3 voltage lanes, activated by processor based on estimated power consumption
- Advantages: Simplifies power design; consolidates 5 platform VRs to one; finer-grain on-die processor delivery control.

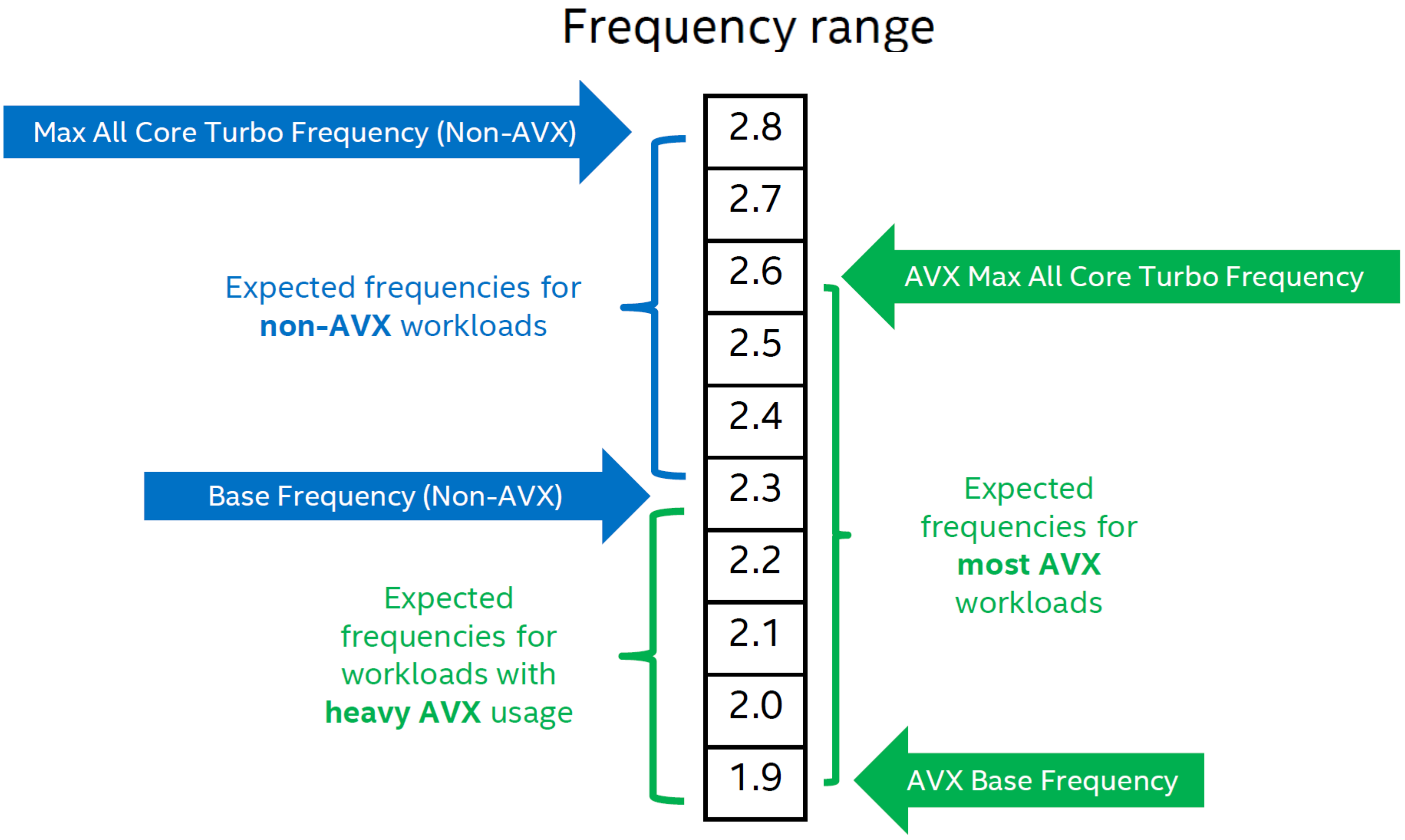
E. A. Burton, et al. “FIVR — Fully integrated voltage regulators on 4th generation Intel Core SoCs”, in *Proc. of IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2014

Haswell Power Management



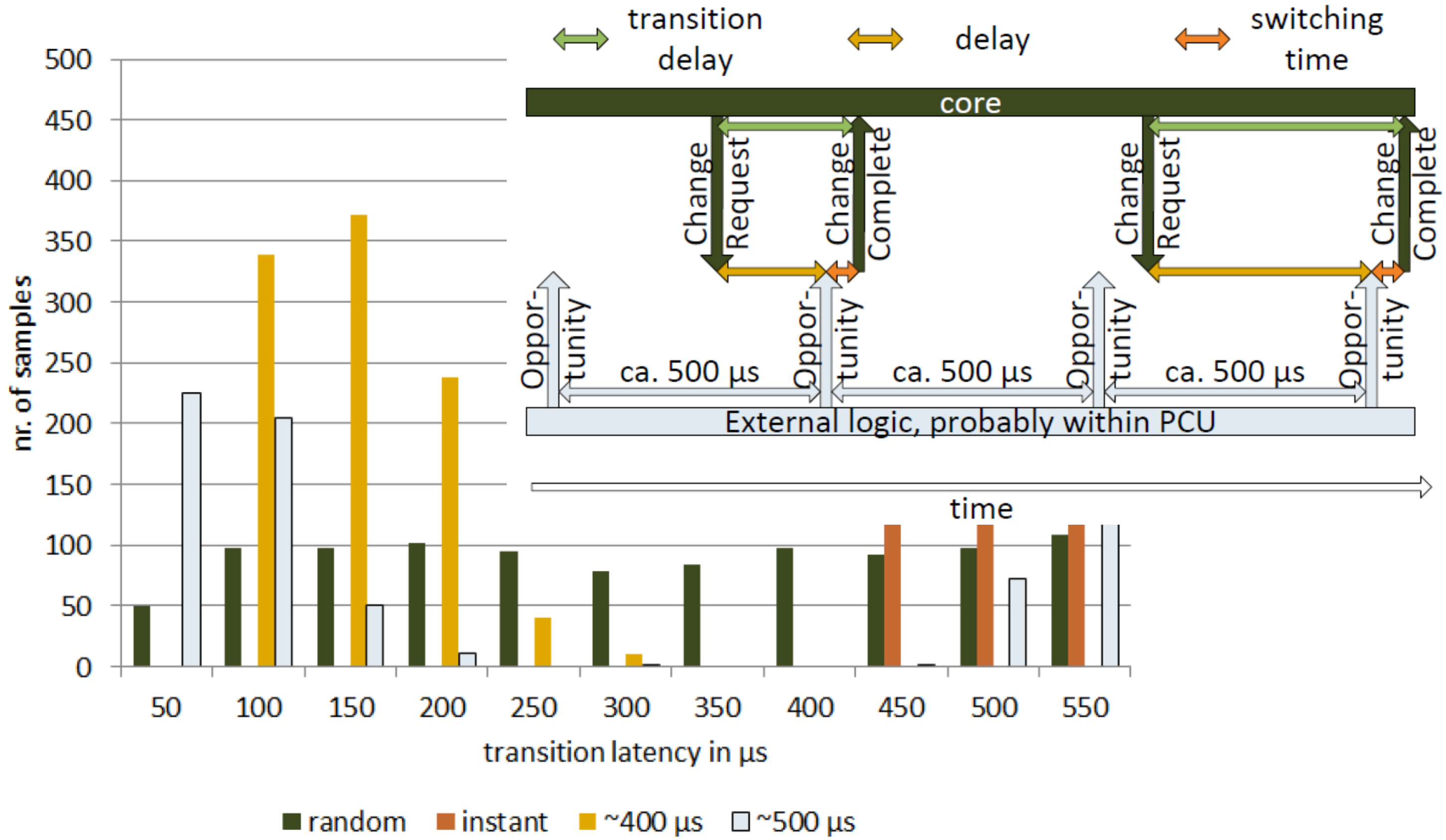
- Per-Core Power States (PCPS)
  - individual cores to have their own voltage and therefore frequency domains, can change their freq independently
- Uncore Frequency Scaling (UFS)
  - components outside of the cores to scale their frequency up and down independently of the cores
- Energy Efficient Turbo (EET)
  - reduce usage of turbo frequencies that do not significantly increase the performance
- Turbo State Limiting
  - puts a cap on the turbo states on the cores to cut down on the variability of clock frequencies

P-States for Workloads


Frequency range
Max All Core Turbo Frequency (Non-AVX)
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
Expected frequencies for **non-AVX** workloads
AVX Max All Core Turbo Frequency
Base Frequency (Non-AVX)
Expected frequencies for **most AVX** workloads
Expected frequencies for workloads with **heavy AVX** usage
AVX Base Frequency

P-State Transition Latency

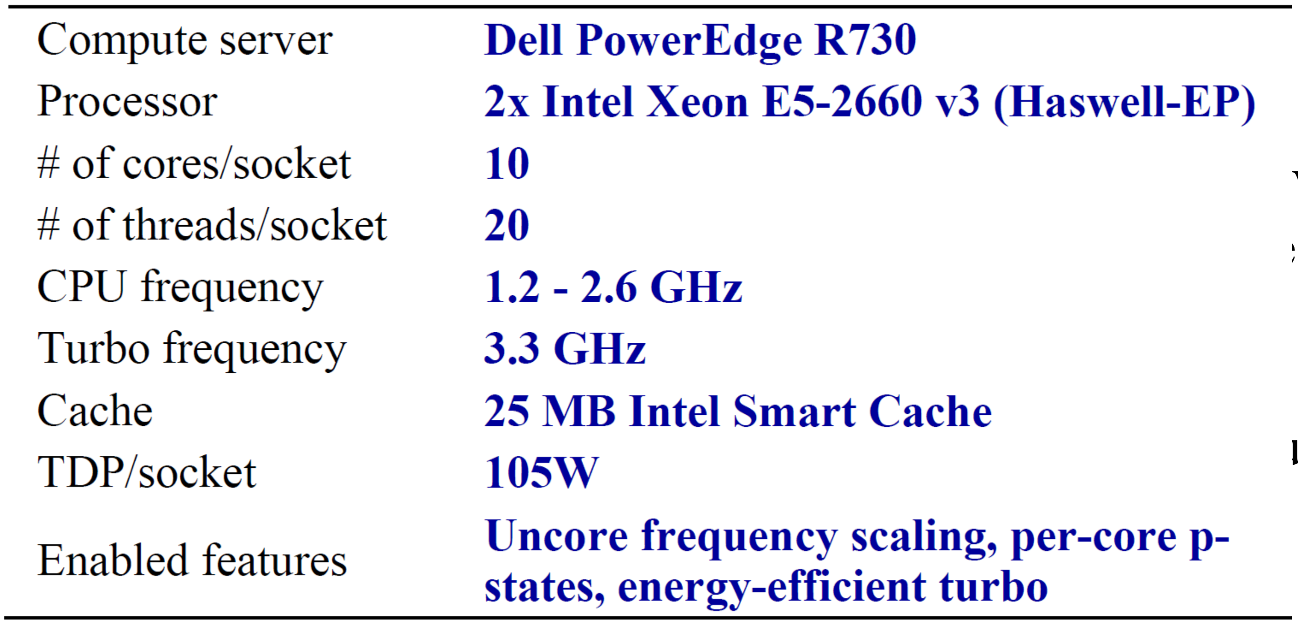
- Frequency changes occur in regular intervals of about 500 µs.



D. Hackenberg, et al. “An Energy Efficiency Feature Survey of the Intel Haswell Processor”, in *Proc. of IEEE International Parallel and Distributed Processing Symposium Workshop (IPDPSW)*, 2015

Experiment Setup

- Hardware
  - Dell PowerEdge R730 rack servers
  - 2 Xeon E5-2660 v3 processors
  - 128 GB RAM, 200 GB SSD
  -



Compute server	**Dell PowerEdge R730**
Processor	**2x Intel Xeon E5-2660 v3 (Haswell-EP)**
# of cores/socket	**10**
# of threads/socket	**20**
CPU frequency	**1.2 - 2.6 GHz**
Turbo frequency	**3.3 GHz**
Cache	**25 MB Intel Smart Cache**
TDP/socket	**105W**
Enabled features	**Uncore frequency scaling, per-core p-states, energy-efficient turbo**

ver issue)

cific

ıncores

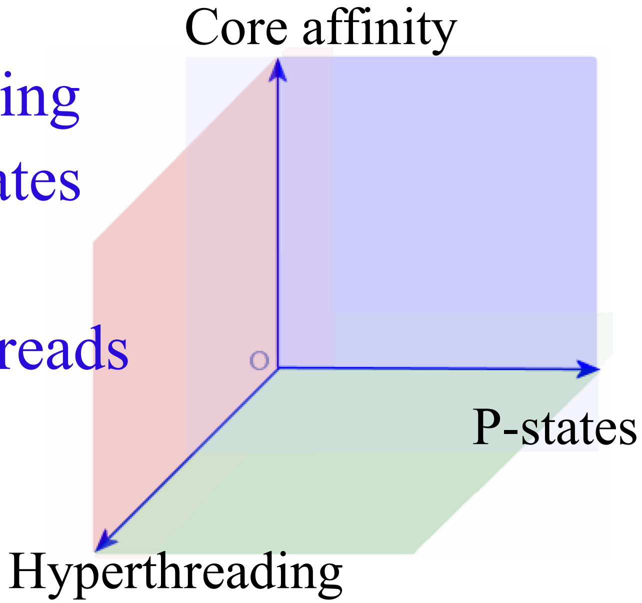
Supported o
Haswell-EI

Benchmark Code & Applications

- Software
    - Benchmarks
        - Compute bound: HPL and FIRESTARTER
        - Memory bound: STREAM
        - Mixed: CLAMR (OpenMP version)
    - Tools
        - Performance API (PAPI): platform-independent library, PAPI-C provides in-line power & energy measurement
        - FTaLaT to control p-states
        - A tool developed by Schone to control c-states
        - LIKWID to measure the uncore freq

Experiment Design

- Issues of interest
  - Overhead of onboard power monitoring
  - Power-performance at different P-states
  - Hyperthreading influence
  - Core Affinity – floating or pinned threads

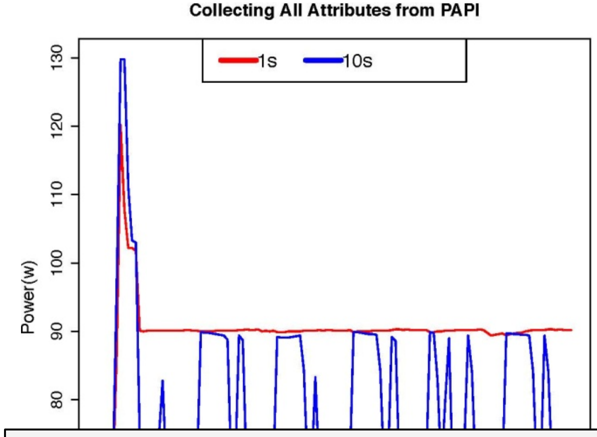


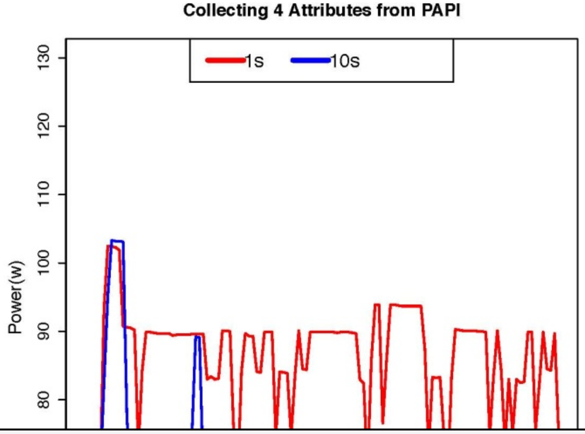
- Metrics
  - Energy consumption: E = P * T
  - Performance per watt: FLOPS/W = Throughput / P

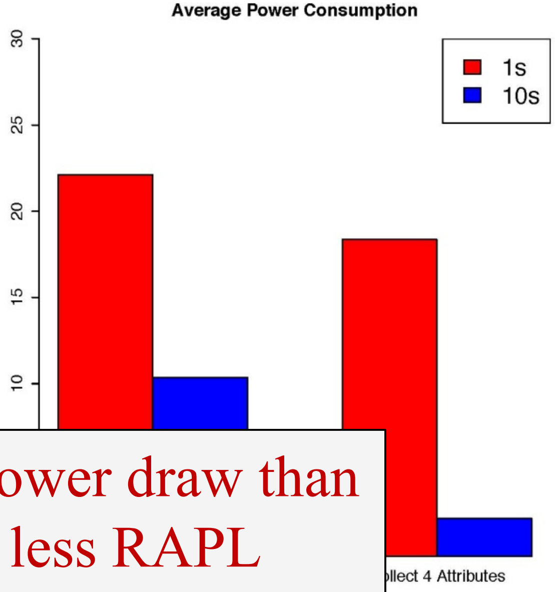
Overhead of Power Monitoring

- PAPI → MSR driver → RAPL attributes
- Measurement rate: 1s vs. 10s
- Measured RAPL attributes: 16 vs. 4

PACKAGE_ENERGY:PACKAGE0
DRAM_ENERGY:PACKAGE0
PACKAGE_ENERGY:PACKAGE1
DRAM_ENERGY:PACKAGE1



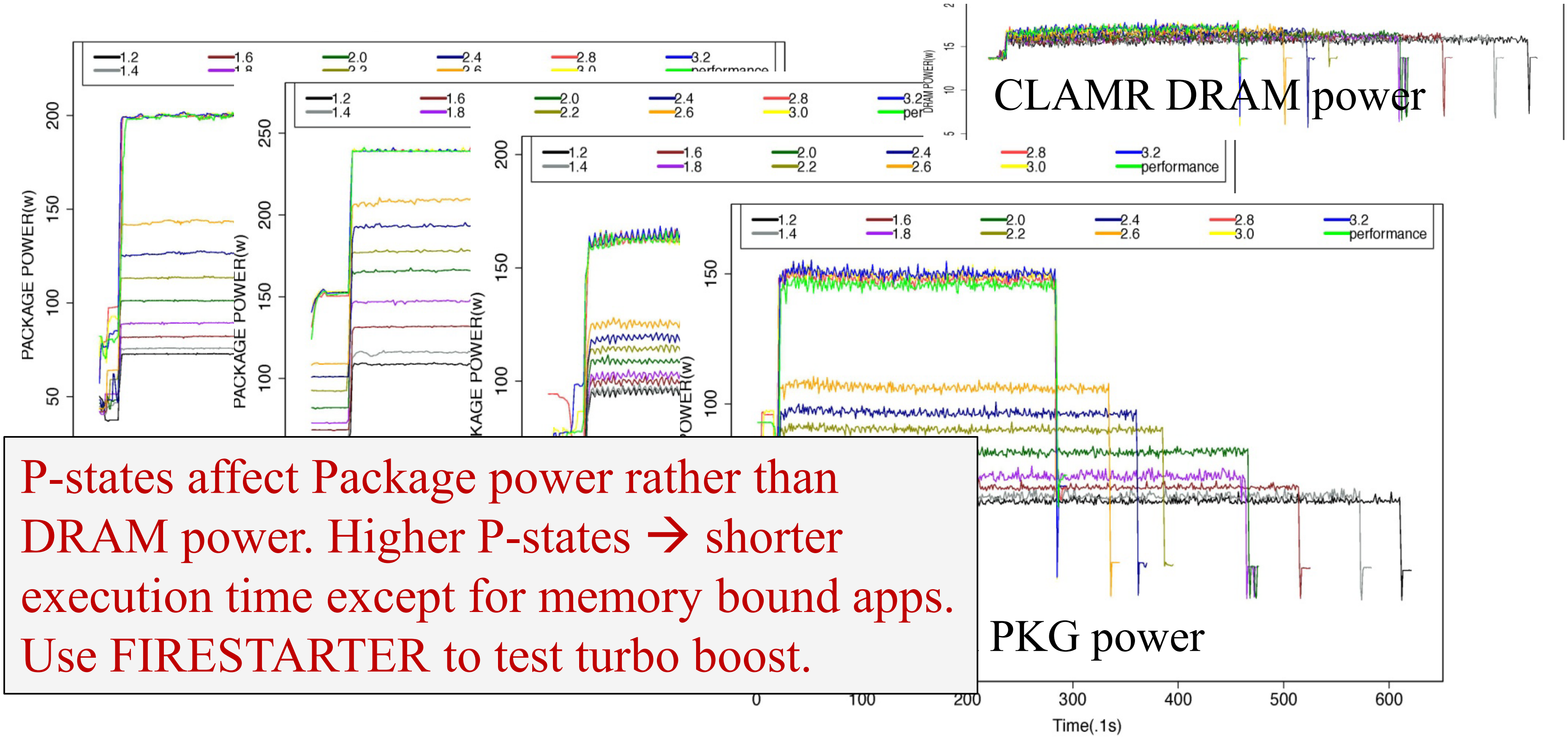




PAPI power monitoring causes up to 28.6% more power draw than the idle power. Using longer monitoring period and less RAPL attributes can reduce the overhead by 75%.

Power Control Knob: P-States

Setting GHz	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.3
Measured (GHz)	1.2	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.9	2.9	2.9	2.9



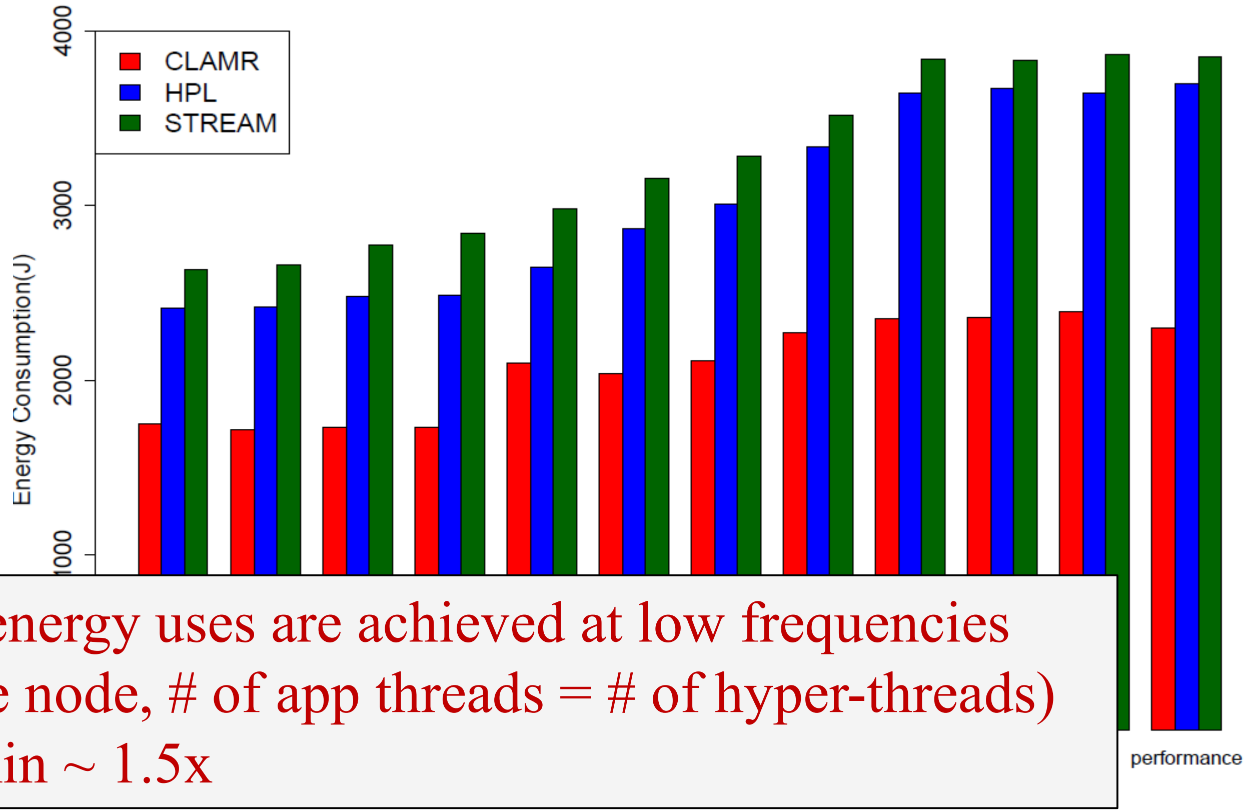
CLAMR DRAM power

PKG power

P-states affect Package power rather than DRAM power. Higher P-states → shorter execution time except for memory bound apps. Use FIRESTARTER to test turbo boost.

Power Control Knob: P-States

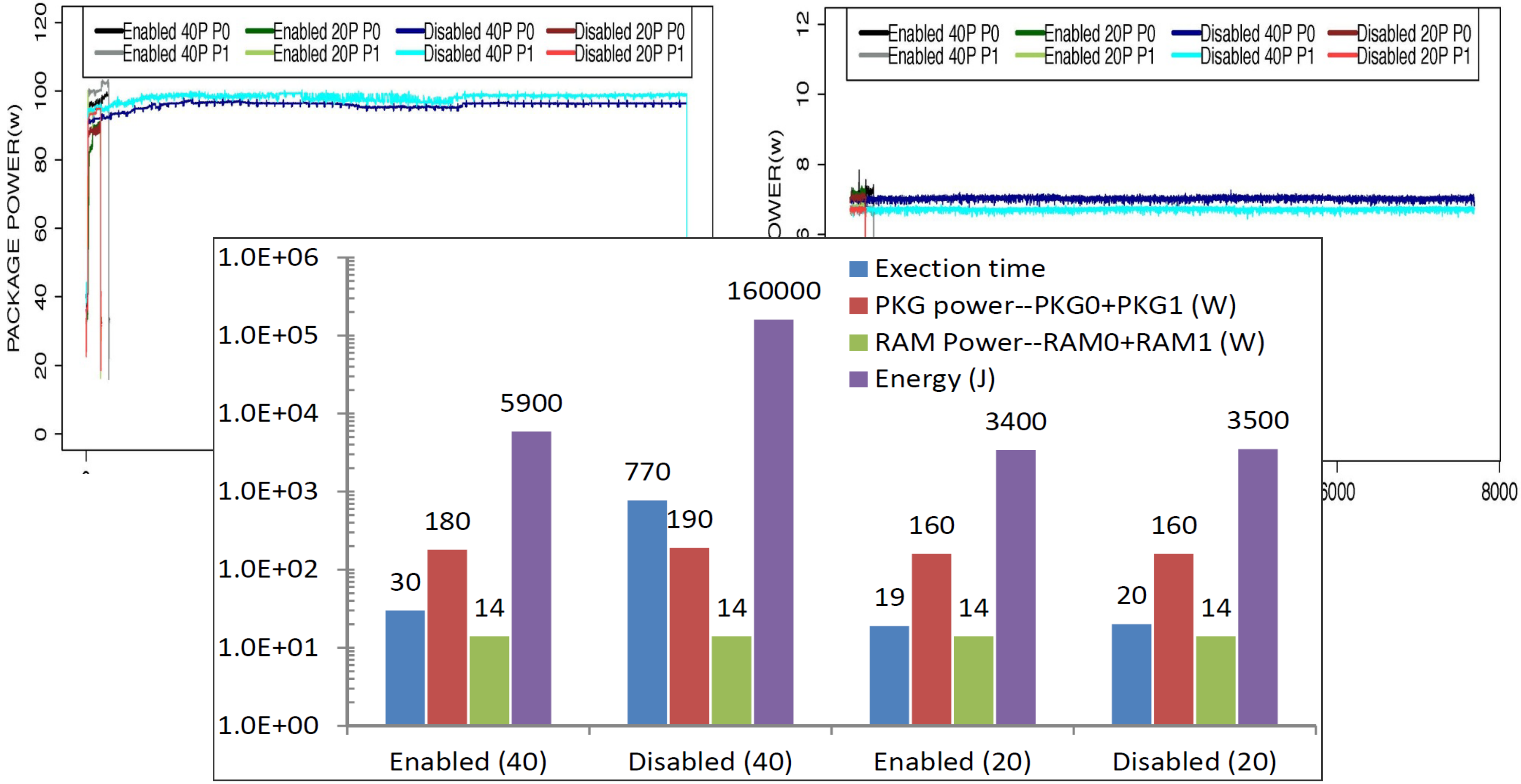
Setting GHz	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.3
Measured (GHz)	1.2	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.9	2.9	2.9	2.9



Least energy uses are achieved at low frequencies
(Single node, # of app threads = # of hyper-threads)
max/min ~ 1.5x

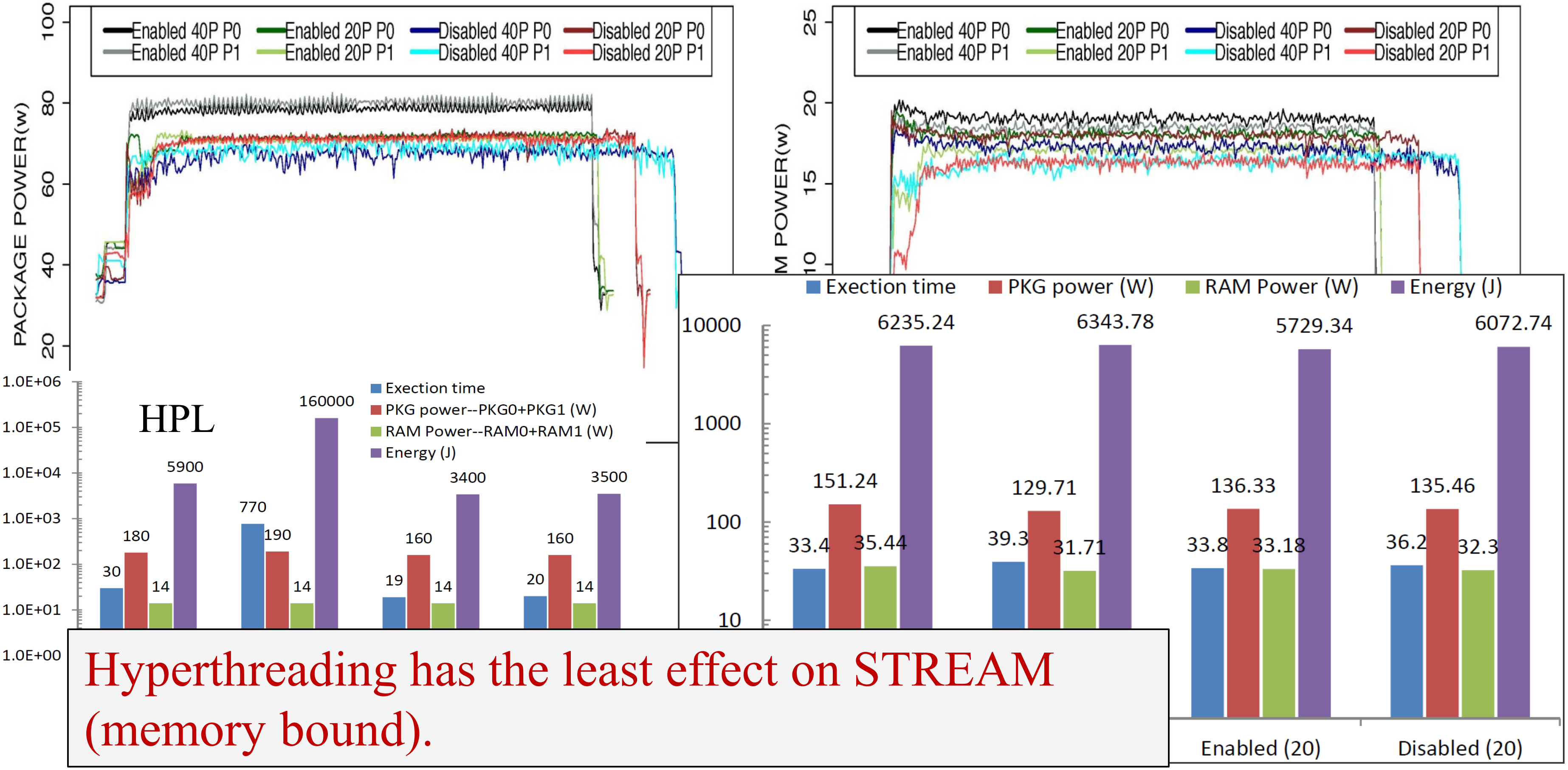
Power Control Knob: Hyperthreading

- HPL



Power Control Knob: Hyperthreading

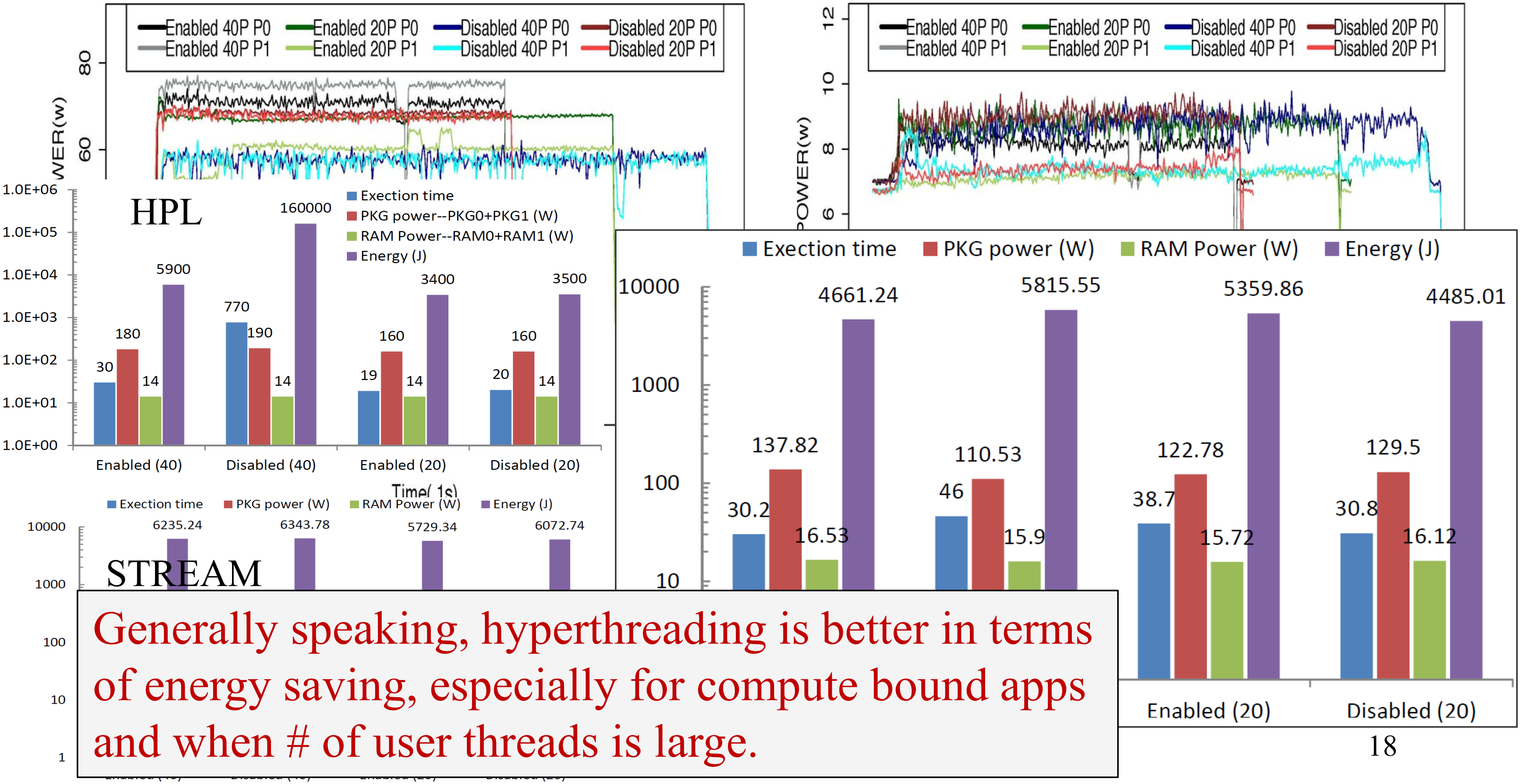
- STREAM



Hyperthreading has the least effect on STREAM (memory bound).

Power Control Knob: Hyperthreading

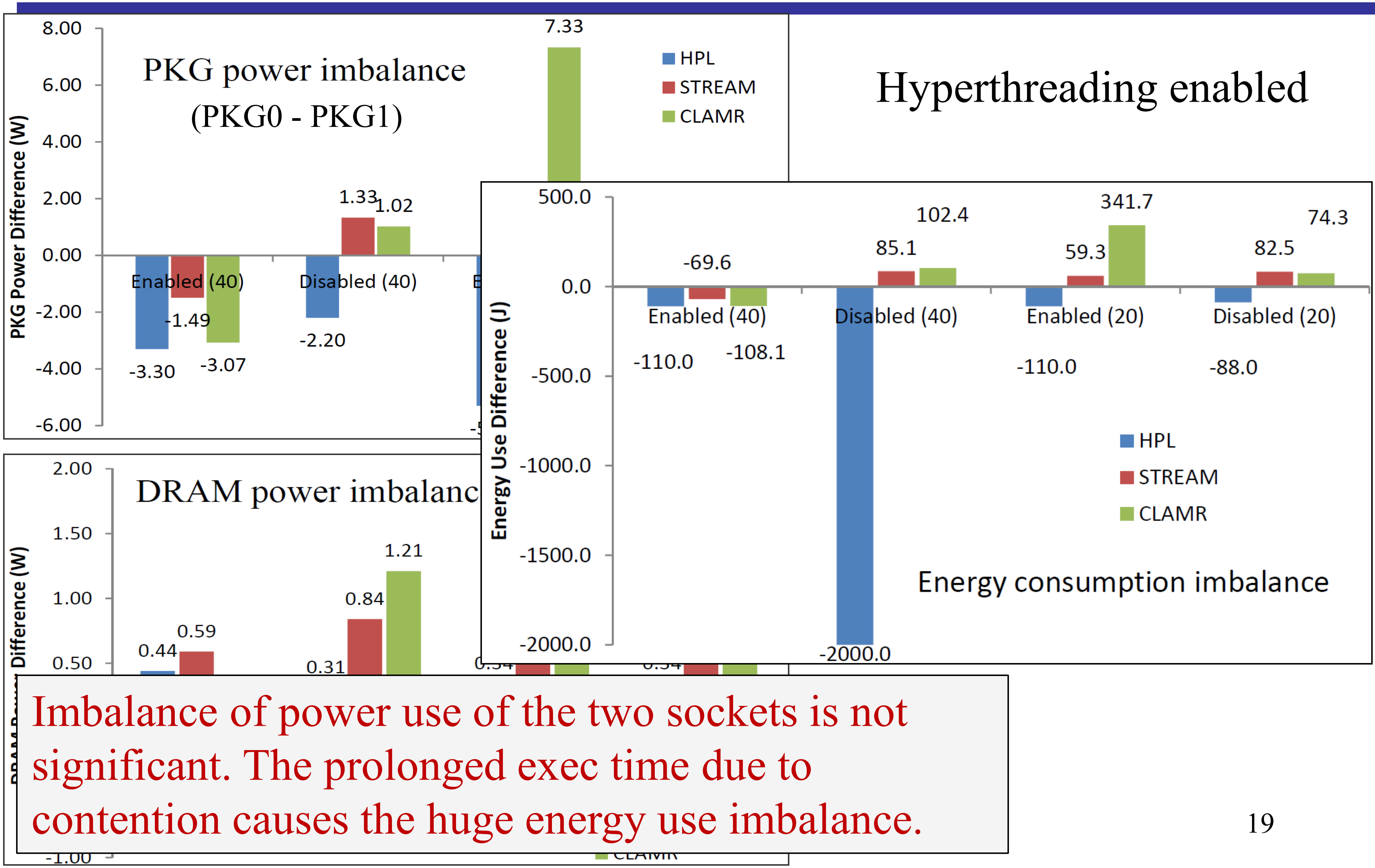
- CLAMR



Generally speaking, hyperthreading is better in terms of energy saving, especially for compute bound apps and when # of user threads is large.

Imbalance Between Two Sockets

Hyperthreading enabled

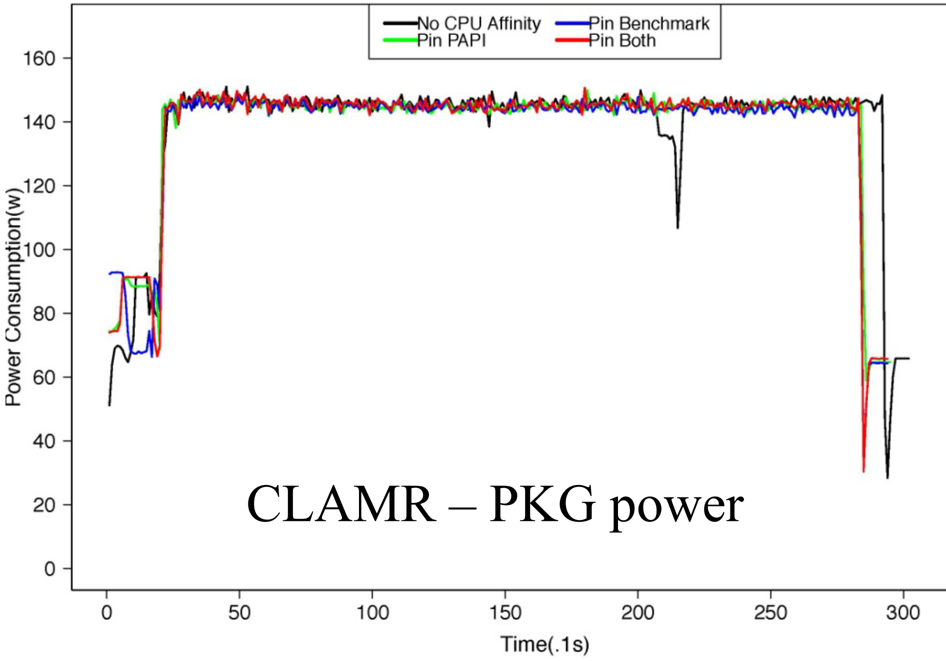


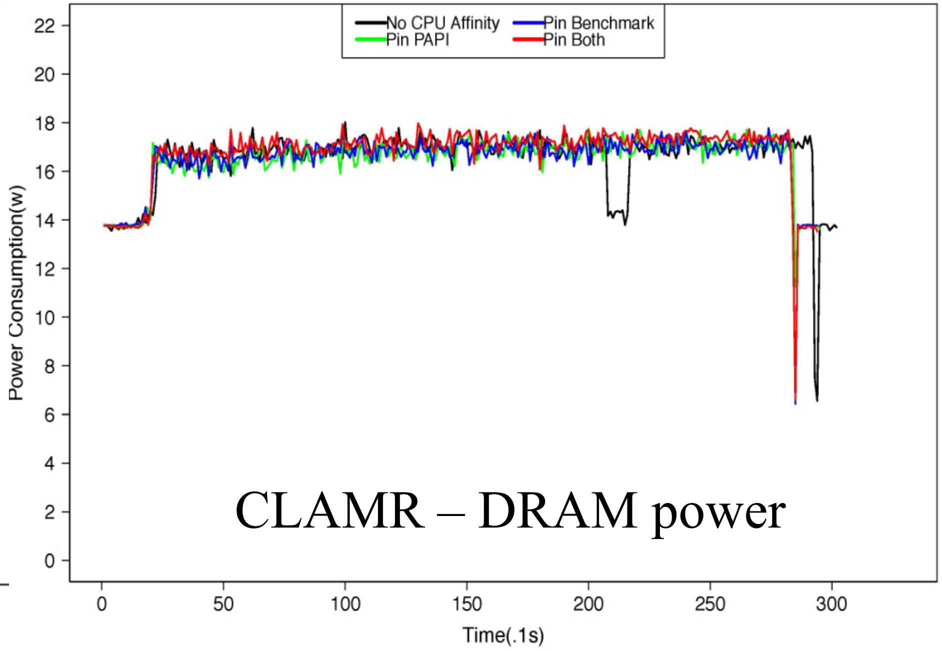
Imbalance of power use of the two sockets is not significant. The prolonged exec time due to contention causes the huge energy use imbalance.

Power Control Knob: Core Affinity

(App threads, PAPI thread)

(floating, floating)	(floating, pinned)
(pinned, floating)	(pinned, pinned)


CLAMR – PKG power

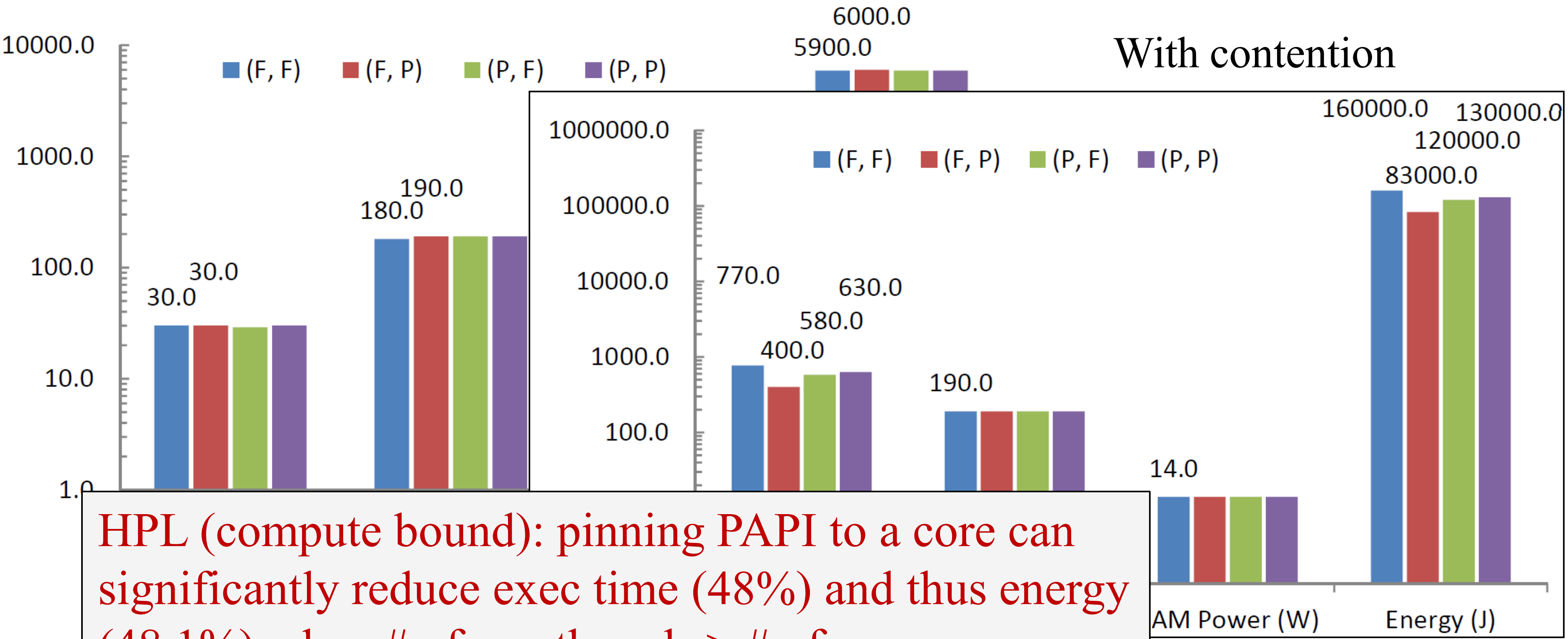

CLAMR – DRAM power

Power Control Knob: Core Affinity

(App threads, PAPI thread)

(floating, floating)	(floating, pinned)
(pinned, floating)	(pinned, pinned)

HPL



With contention

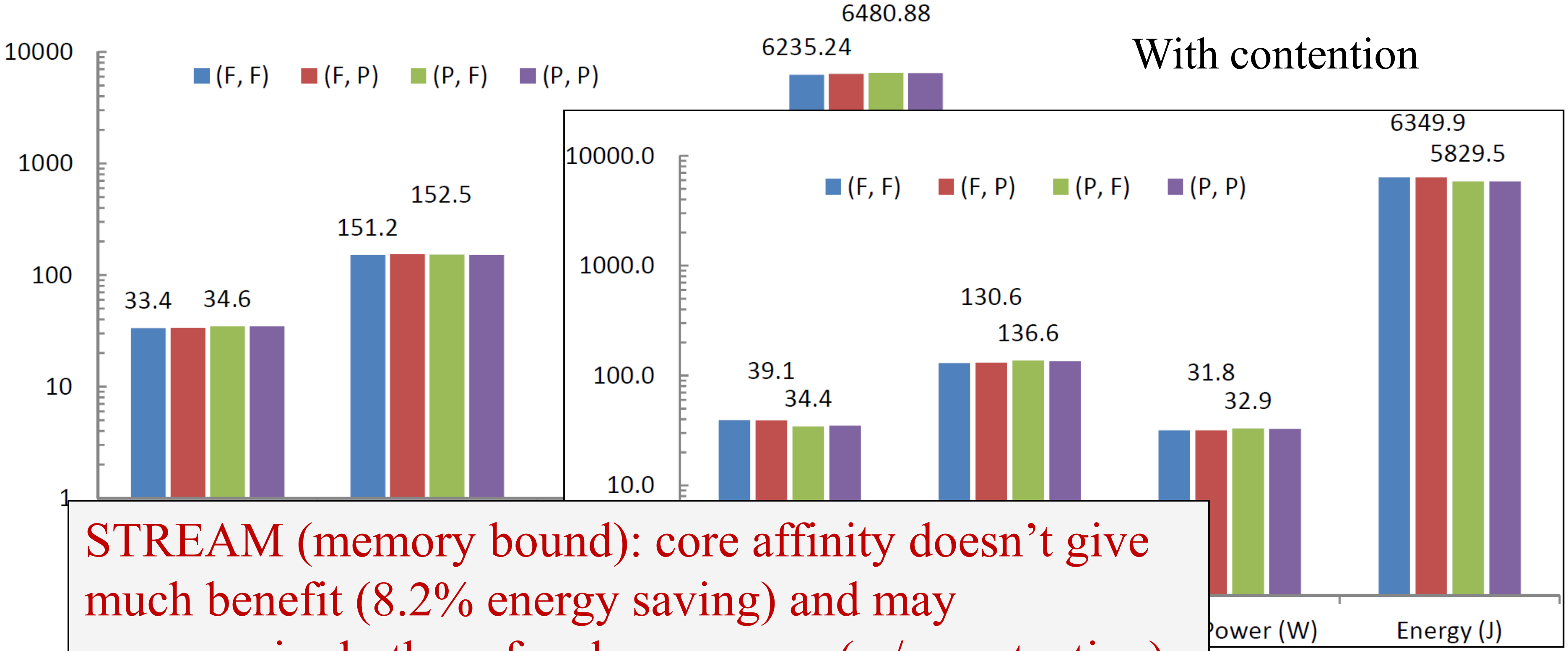
HPL (compute bound): pinning PAPI to a core can significantly reduce exec time (48%) and thus energy (48.1%) when # of app threads > # of cores.

Power Control Knob: Core Affinity

(App threads, PAPI thread)

(floating, floating)	(floating, pinned)
(pinned, floating)	(pinned, pinned)

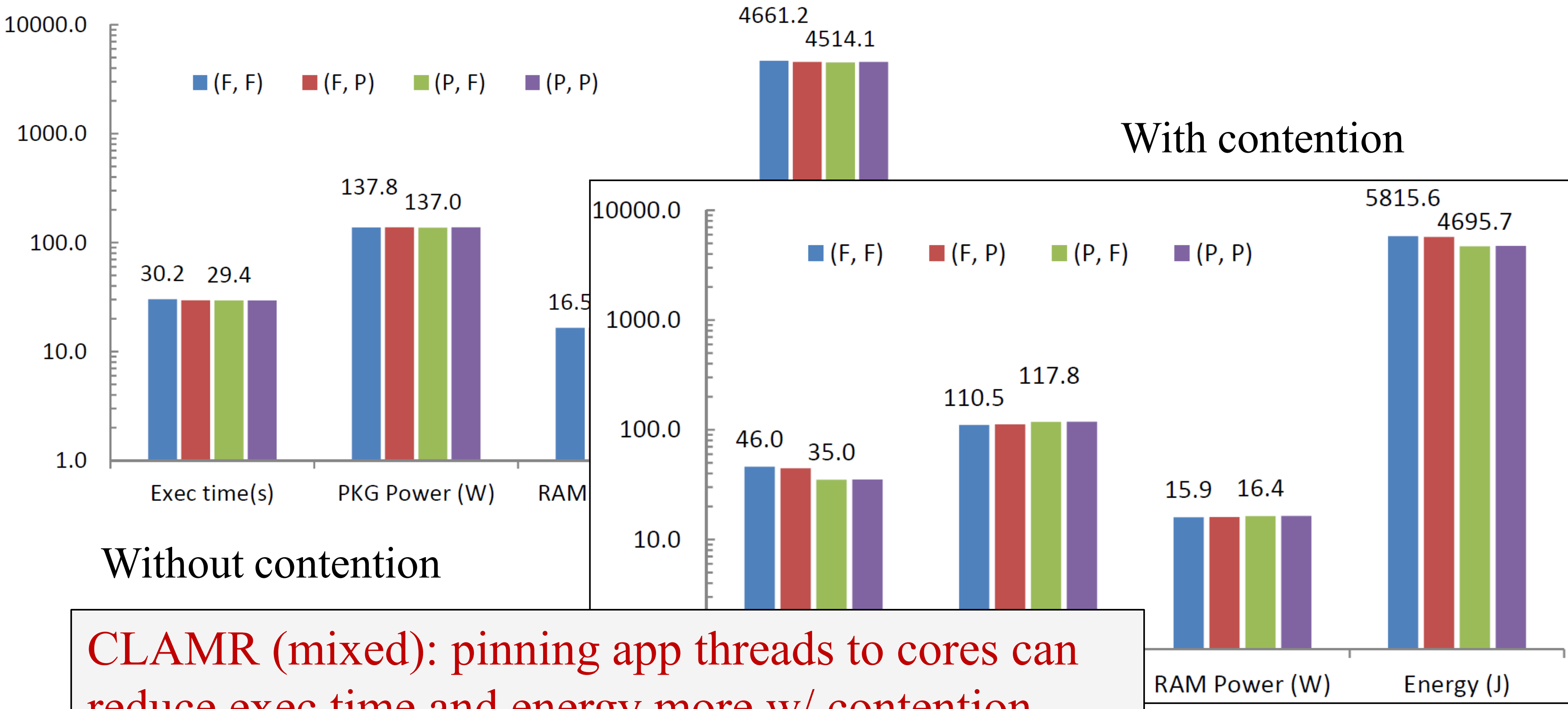
STREAM



STREAM (memory bound): core affinity doesn't give much benefit (8.2% energy saving) and may compromise both perf and energy use (w/o contention).

Power Control Knob: Core Affinity

CLAMR



With contention

Without contention

CLAMR (mixed): pinning app threads to cores can reduce exec time and energy more w/ contention (24% and 19.3%).

Summary and Future Plans

- **Goal**: to understand and characterize power-performance of Haswell EP for HPC.
- Test four benchmark programs on PowerEdge R730 with control knobs
  - P-states (most effective)
  - Hyperthreading (effective for compute bound app)
  - Core affinity (good if contention exists)
- Plans
  - Test more applications and use more control knobs
  - Use more compute nodes
  - Examine the combined effects of control knobs
  - Power capping

Acknowledgements

I would like to thank

Michael Lang, Scott Pakin, Nathan DeBardeleben, Sean Blanchard, Bradley Settlemyer, Qiang Guan, Hsing-Bung (HB) Chen, Gary Grider, John Bent …

and the summer students

Measurement and Characterization of Haswell Power and Energy Consumption

Thank you!

Questions & Suggestions?

LA-UR-15-25938