

Is Green Exascale Computing ... an Oxymoron?

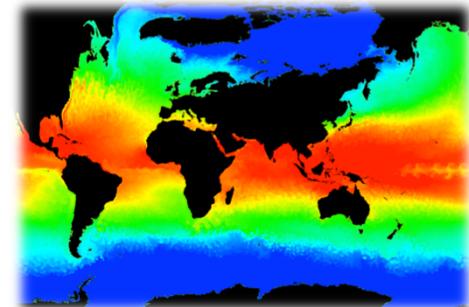
(a.k.a. “The Case of the Missing Supercomputer Energy”)

Wu FENG

Green500

Virginia Tech

- SEEC Center
- Dept. of Computer Science
- Dept. of Electrical & Computer Engineering
- Health Sciences
- Virginia Bioinformatics Institute



Performance Milestones of HPC Systems

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead

Keren Bergman

Shekhar Borkar

Dan Campbell

William Carlson

William Dally

Monty Denneau

Paul Franzon

William Harrod

Kerry Hill

Jon Hiller

Sherman Karp

Stephen Keckler

Dean Klein

Robert Lucas

Mark Richards

Al Scarpelli

Steven Scott

Allan Snavely

Thomas Sterling

R. Stanley Williams

Katherine Yelick

September 28, 2008



- Performance crosses a threshold of 10^{3k} operations per second, for some k .
 - 1997 : Terascale (10^{12}) → Intel ASCI Red
 - 2008 : Petascale (10^{15}) → IBM Roadrunner
 - 2015 : Exascale prediction (10^{18}) → ???

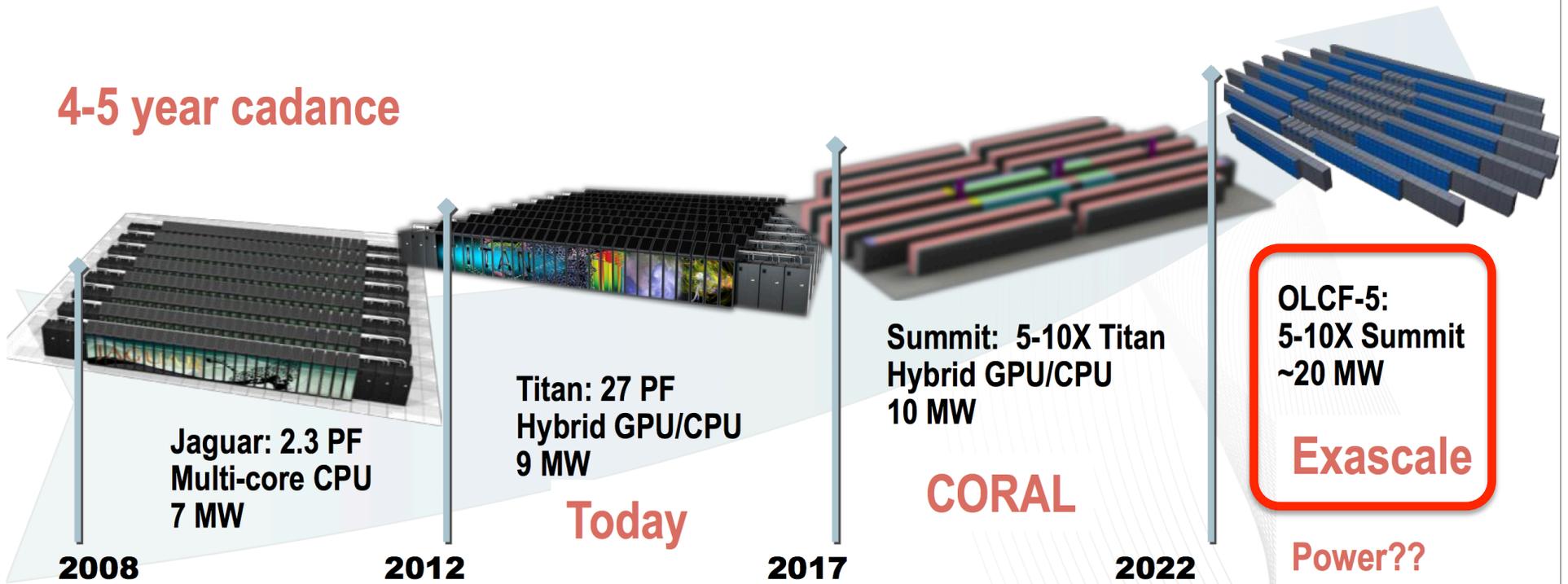


- Power?
20 MW (“impossible”) → 67 MW (“possible”)

Mission: Providing world-class computational resources and specialized services for the most computationally intensive global challenges

Vision: Deliver transforming discoveries in climate, materials, biology, energy technologies, etc

4-5 year cadance



3 AI Geist, Present & Future Leadership Computers at OLCF, DOE Data/Viz PI Mtg, Jan 2015



Why Does **Green** Exascale Appear Feasible Now?

1. **Timeline Changed!**

2015 → 2022 ... *as much as seven years of additional runway.*

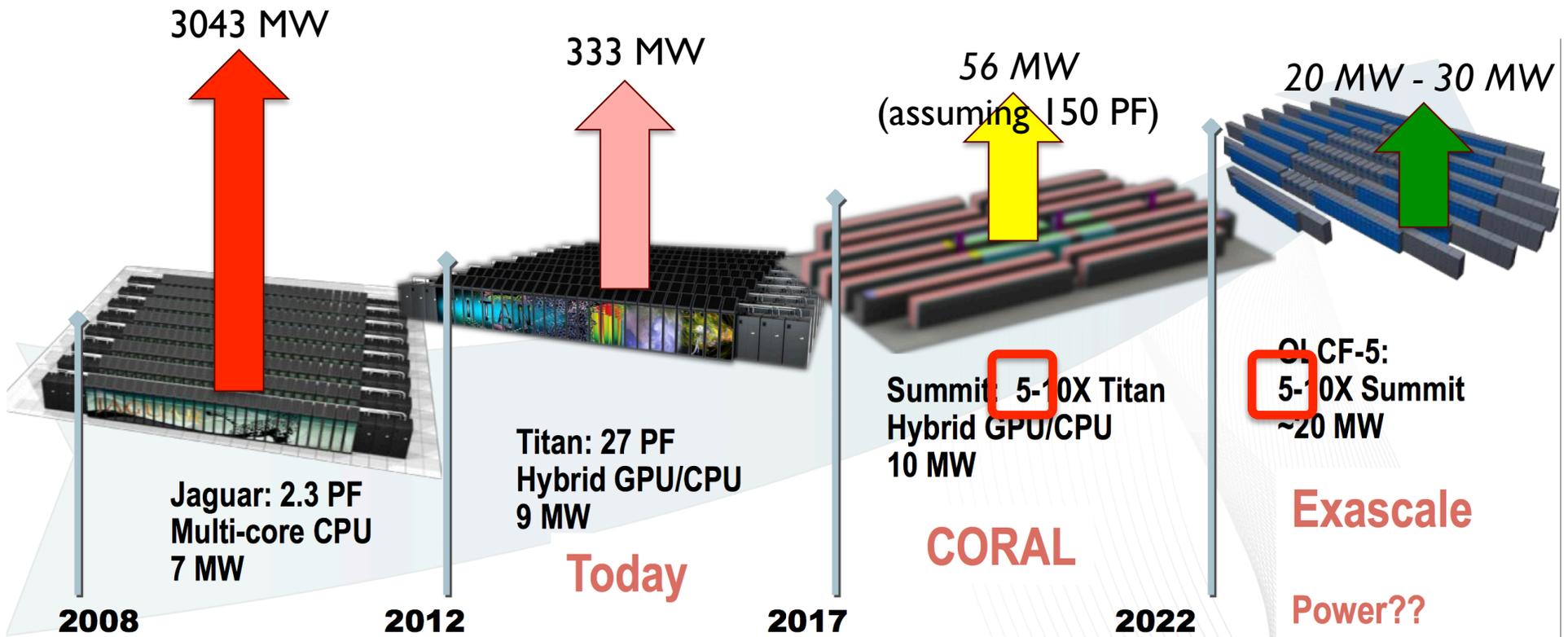
2. **Funding Investment by DOE Office of Science**

- **FastForward**
- **Design Forward and Design Forward 2**

with a focus on power and energy efficiency, i.e., greenness

We are now metering, monitoring, and measurement the greenness of systems from subsystems to nodes to entire supercomputing systems. (THIS WORKSHOP!)

Linear Power Extrapolation to Exascale

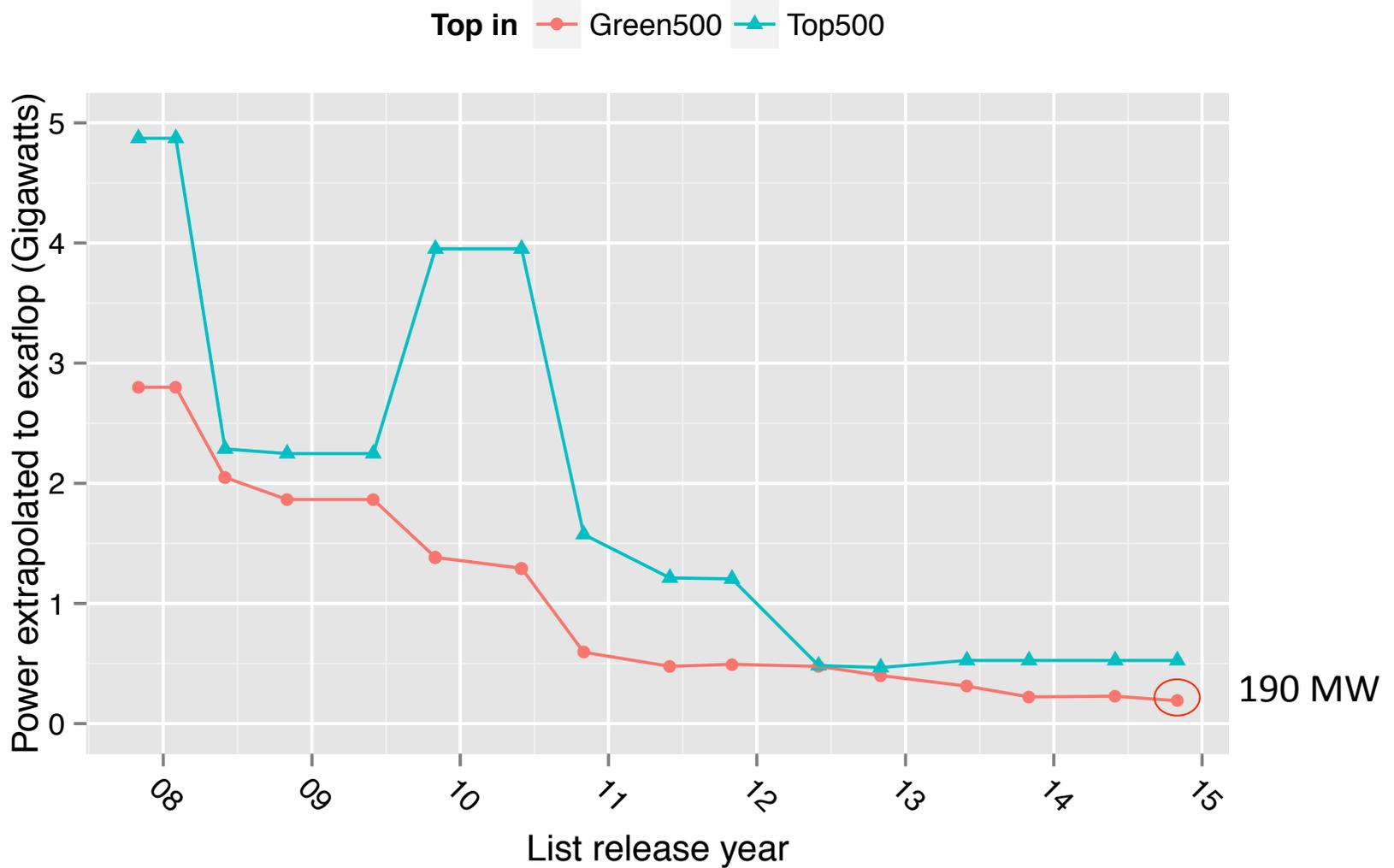


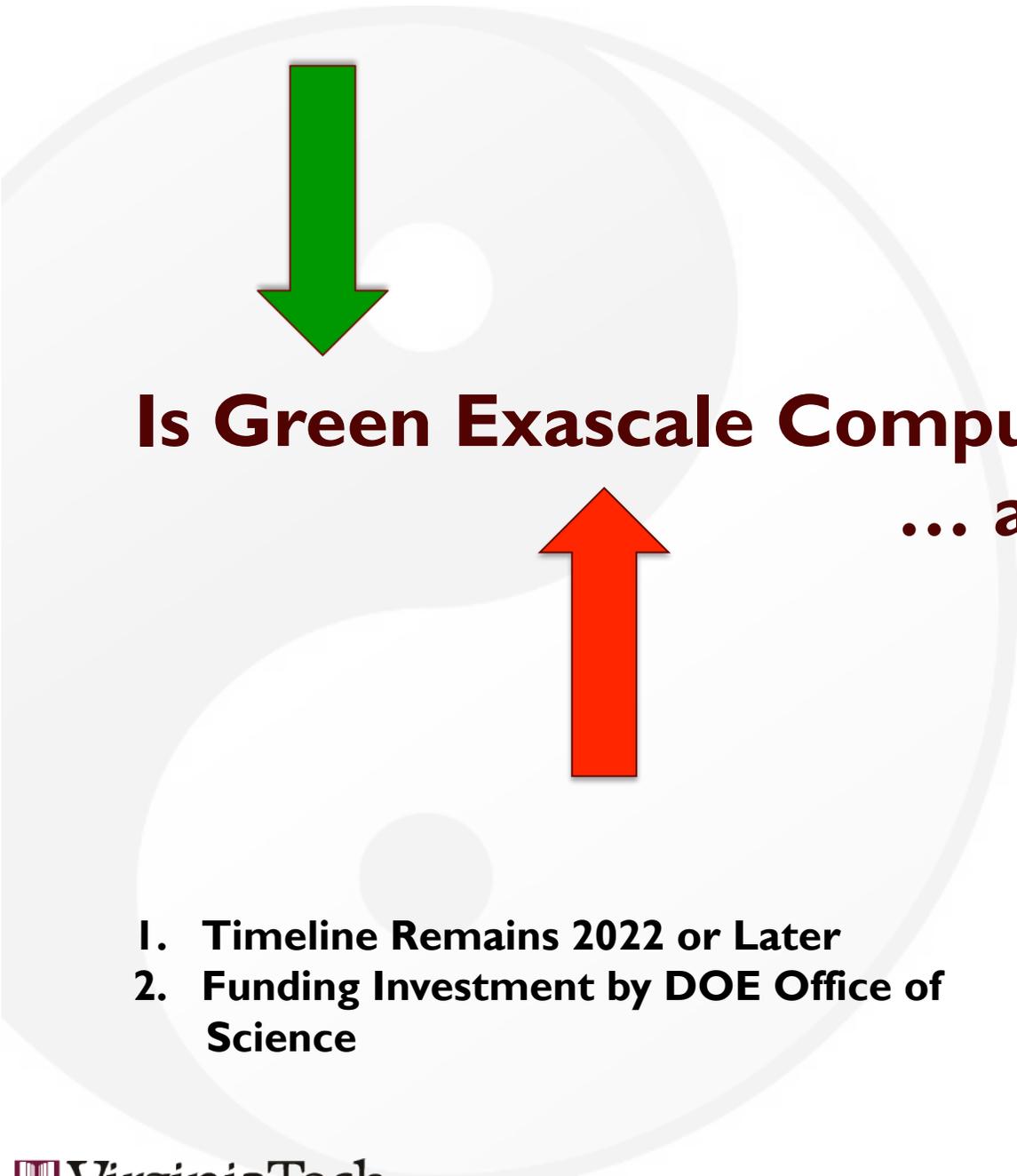
3 AI Geist, Present & Future Leadership Computers at OLCF, DOE Data/Viz PI Mtg, Jan 2015



We tend to overestimate what is possible in the near term (2-3 years)
 ... and underestimate what is possible in the next long term (10 years)

Trends: Extrapolating to Exaflop





Is Green Exascale Computing ... an Oxymoron?

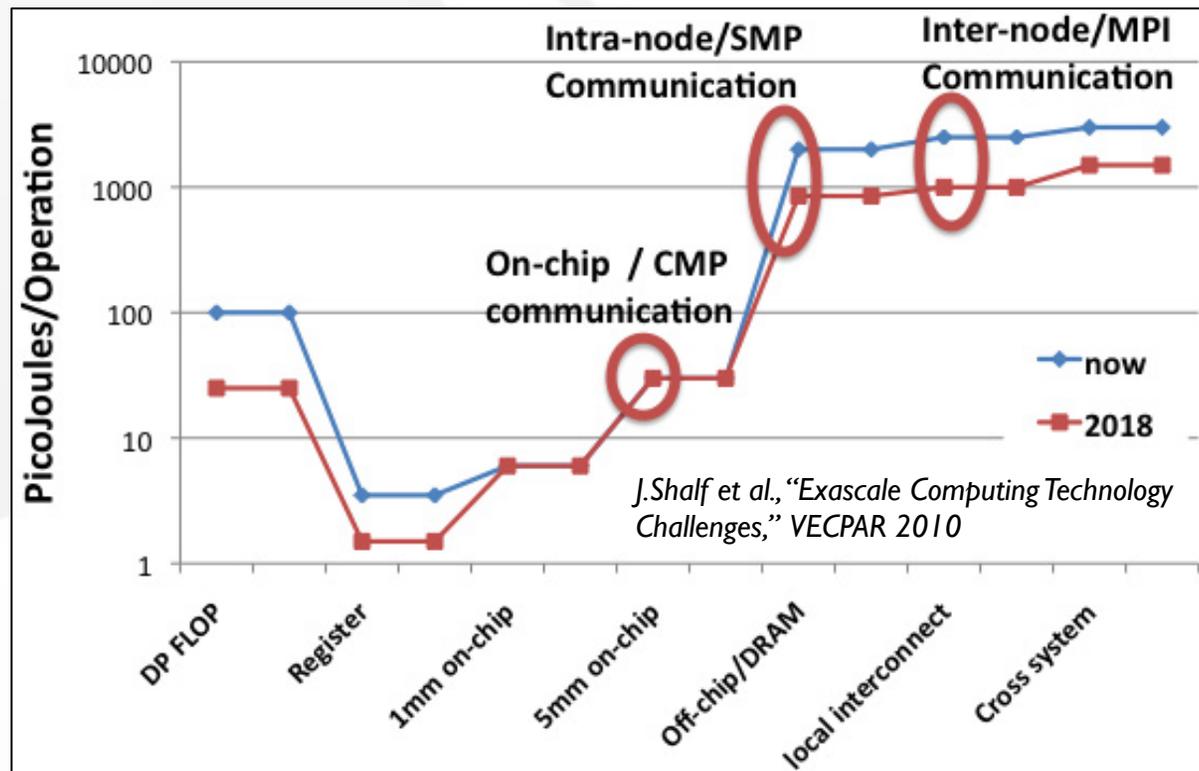
1. Timeline Remains 2022 or Later
2. Funding Investment by DOE Office of Science



Towards Green Exascale Computing

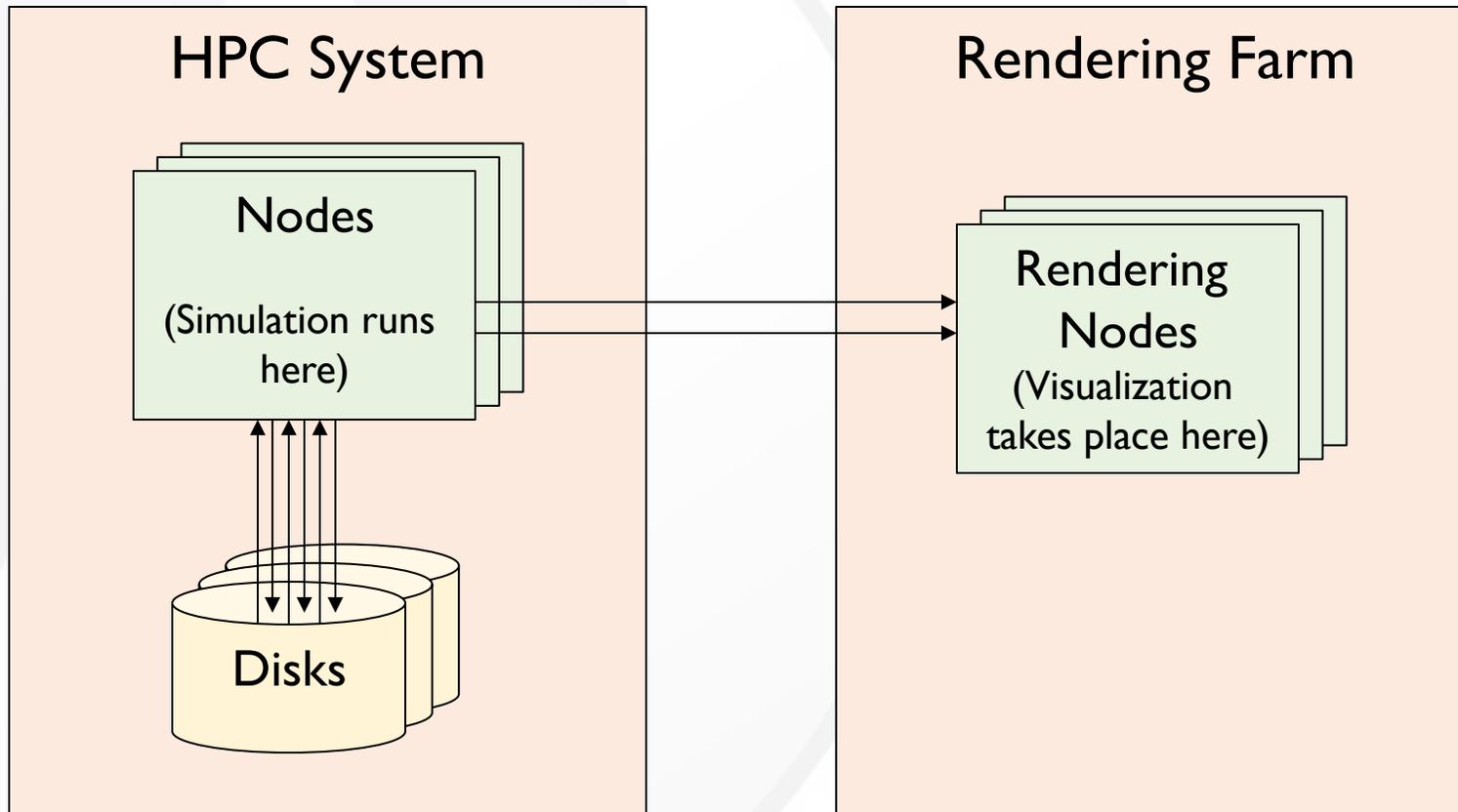
- A Few Prognostications
 - ... Towards Optimizing for Performance, Energy, and Power
 - **Minimize data movement**
 - Shifting focus to I/O rather than compute
 - Traditional visualization vs. in-situ visualization
 - Traditional storage vs. in-situ storage
 - **Address energy proportionality**
 - **Schedule for performance *and* power**
- How to Enable the Above?
 - Monitoring and measurement.
 - See Session 1: Metrics and Session 2: Monitoring Tools.

Minimize Data Movement



- Energy consumed for moving a bit increases as we move down the memory hierarchy
 - *Off-chip transfers cost nearly 100 times as much energy as on-chip transfers!*

Traditional “Post-Processing” Visualization



Solution: In-situ Visualization

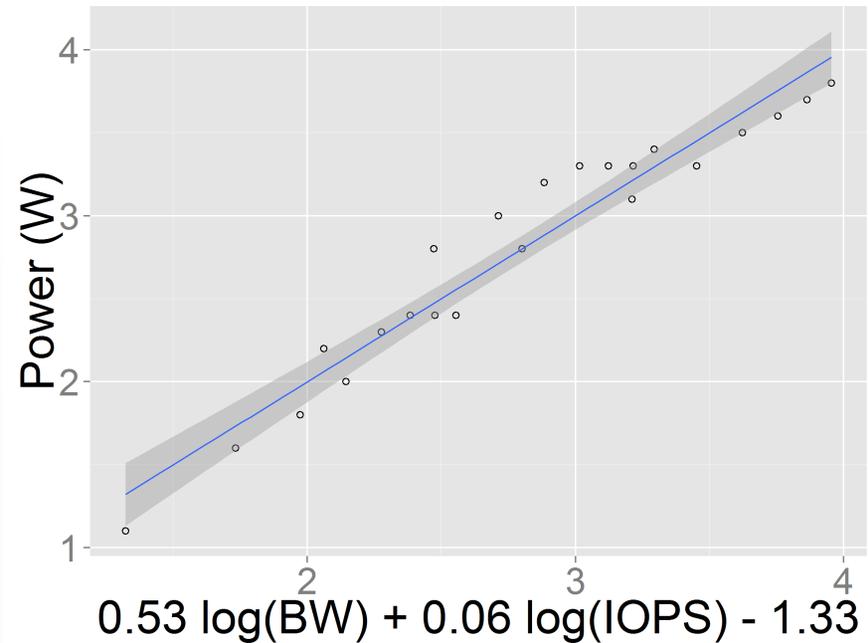
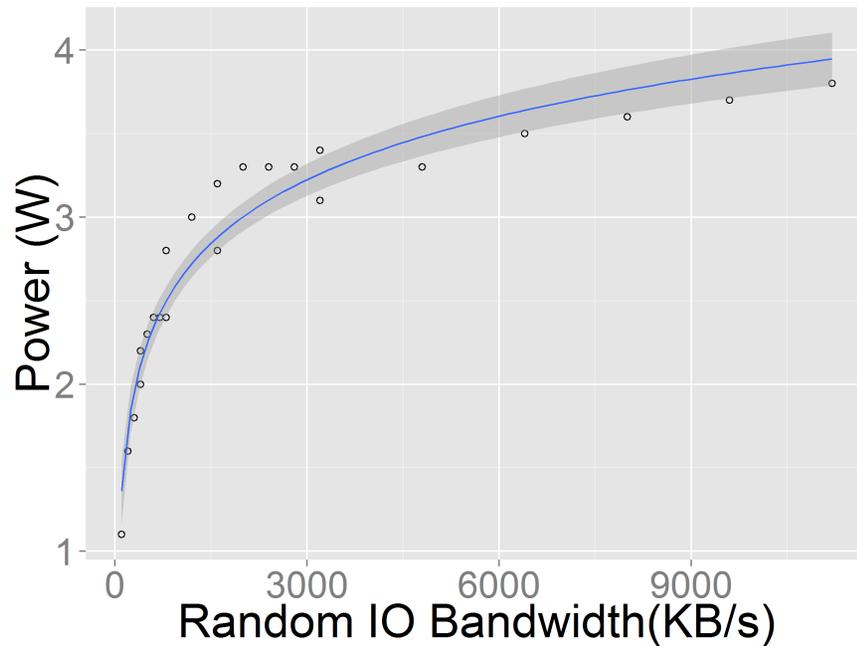
- Perform visualization alongside the simulation
 - Create an *image representation of data at end of each iteration directly* instead of writing raw data to disk
 - Visualize in-situ, e.g., GPGPU → GPU
 - Write the image representation (reduced size representation) to disk
 - May involve additional sampling strategies (e.g., spatial, temporal, etc.)

Goal : Minimize Data Movement in Visualization

“Study the performance, power, and energy trade-offs among traditional post-processing, modern post-processing, and in-situ visualization pipelines”

- Detailed sub-component level power measurements within a node to gain detailed insights (e.g., RAPL)
 - Measure power consumption of CPU, memory, and disk
- Measurements at scale to understand problems unique to big supercomputers

Disk Power Model

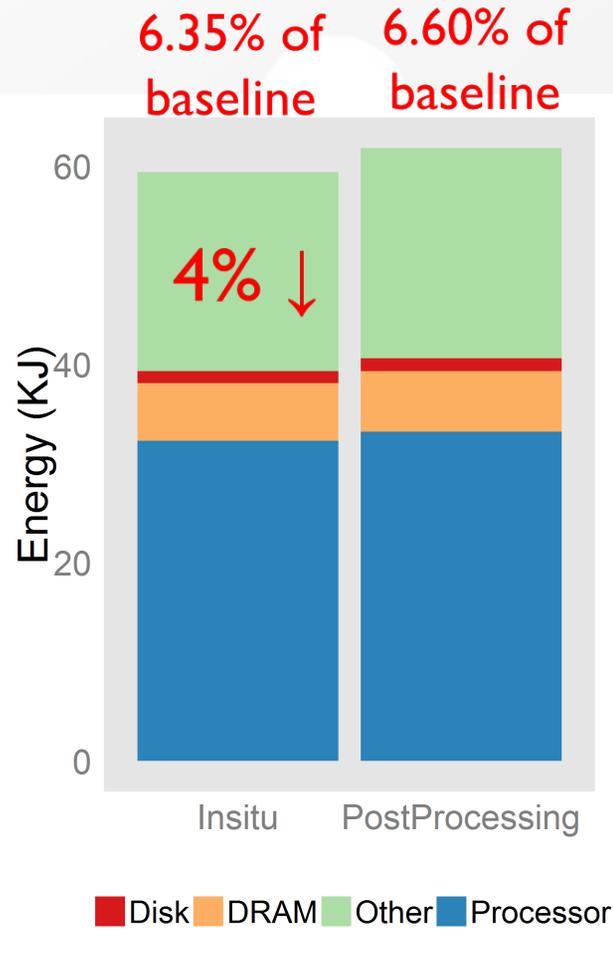


- I/O statistics collected from *iostat*
- Number of I/O operations and the amount of data written affects power consumption of the disk

Hardware Platform

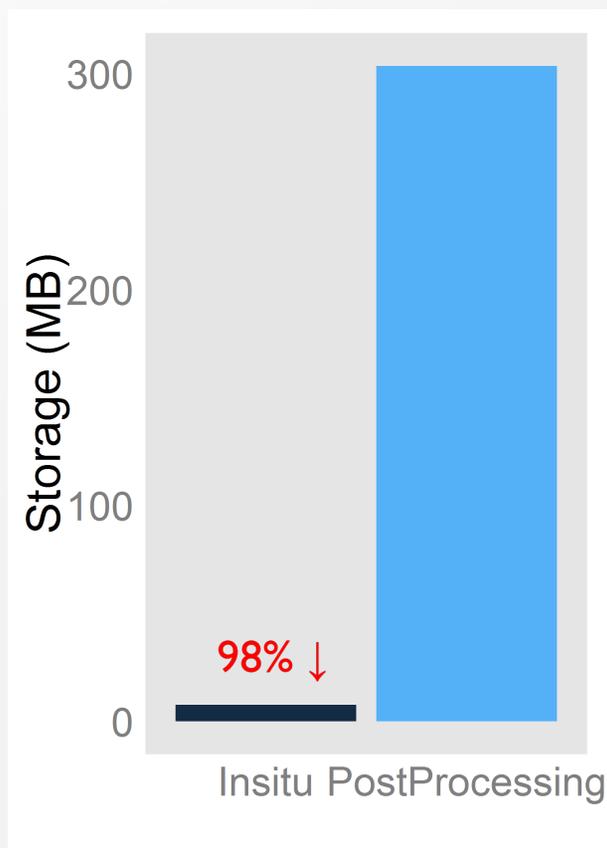
CPU	2x Intel Xeon E5-2665
CPU frequency	2.4 GHz
Last-level cache	20 MB
Memory	4x 16GB DDR3-1333
Memory size	64 GB
Hard disk	Seagate 7200rpm disk
Storage size	500GB
Disk bandwidth	6.0 Gbps

Results: Single-Node Energy Comparison



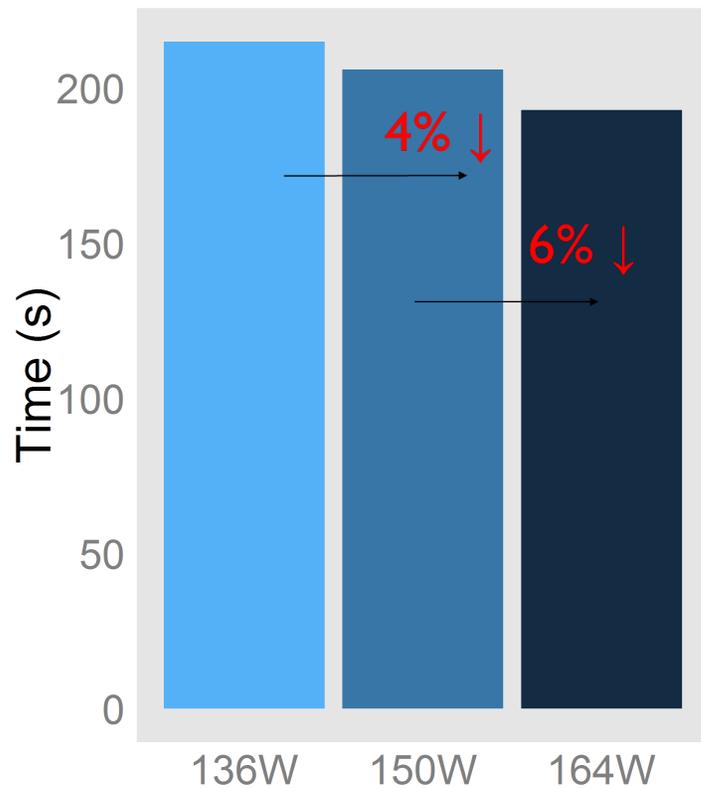
- In-situ consumes 4% less energy than *modern* post-processing
- Compared to *traditional* post-processing, both pipelines consume 93% lower energy

Results: Single-Node Storage Requirements



- 97.5% lower storage requirement for the in-situ pipeline
 - Implies smaller storage cluster
 - Implies lower power consumption

Re-distributing Storage Power to Compute Nodes



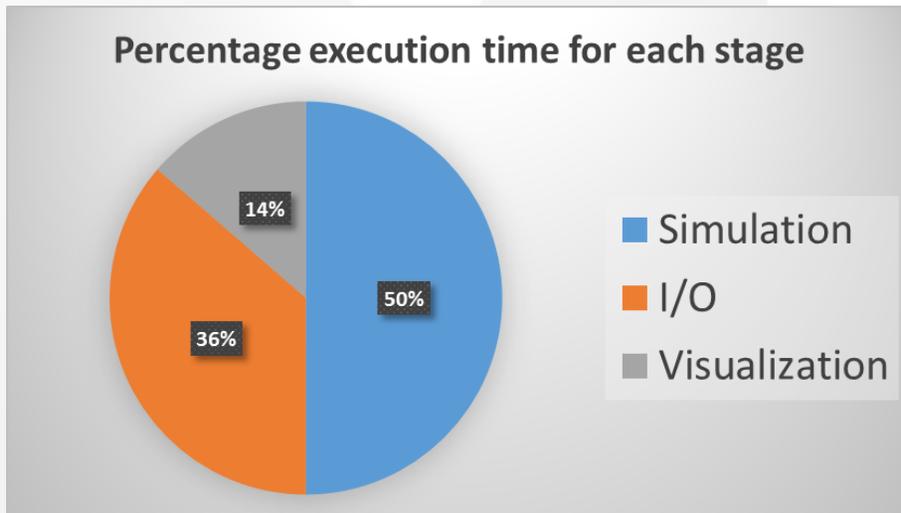
- Assuming reduced storage nodes results in 10% of total power redirected to compute nodes
 - Performance improves by up to 6% for MPAS Ocean Simulation
 - Data from power-capping experiments with RAPL

Results at Scale: Hardware Platform

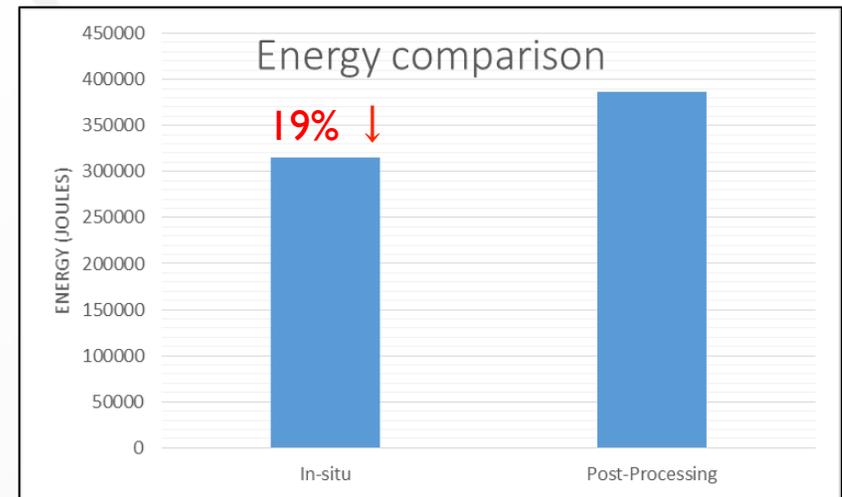
- Caddy supercomputer with a dedicated Lustre file system used for profiling
- Compute nodes
 - 64 nodes out of 150 nodes used in these experiments
 - Each node contains 2x Intel Xeon E5-2670 and 64 GB of RAM
 - Nominal power consumption
 - **6000 W (idle) to 20000 W (workload such as MPAS)**
- Storage nodes
 - 5 nodes configured as 1 master + 2 MDS + 2 OSS
 - 1 RAID storage per MDS and OSS
 - Nominal power consumption
 - **2500W (idle) to 2800W (active)**

Lacking in energy proportionality

Results at Scale: Energy Comparison



Real measurements on Caddy supercomputer at Los Alamos National Laboratory



Partial measurement on Caddy and extrapolation from spec sheets

Projections for Supercomputing

- Increased I/O wait time
 - Storage separated from compute by network
 - Longer execution time and corresponding increase in energy
- Additional energy consumption from data movement through the network
 - No data transfer via network cables in single-node system
- Power/energy overhead for storage higher
 - Separate cluster for storage → additional CPUs, memory, cooling, etc.
 - Storage sub-system shared with compute sub-system in single node

Findings : Minimizing Data Movement

(a.k.a. “The Case of the Missing Supercomputer Energy”)

Two Takeaways for “Missing Energy”

- Most energy savings come from *reducing system idling* (i.e., from reduced I/O wait time)
- Further savings possible if we can reduce the size of the storage nodes (or storage system)

Advantages of In-situ Visualization

- Reduced energy consumption
 - By reducing system idling or I/O wait time
- Reduced power
 - By using fewer storage nodes
- Improved performance
 - By reducing I/O wait time and by making more power available for compute nodes

Future Directions for Green Exascale Computing

- Enhancing HPC Systems
 - Flash buffers and SSDs can reduce I/O wait time
 - Downside: Introducing more components can increase power consumption as well as impact reliability
- Changing HPC System Design
 - Bringing storage nodes and compute nodes together
 - Similar to “Memory in Processor” or “Processor in Memory” concepts in the computer architecture community
- Changing Runtime System
 - Energy proportional computing and storage
 - Putting compute nodes to sleep states during I/O
 - Putting some storage nodes to deep sleep state when bandwidth and storage requirements are lower

Towards Green Exascale Computing

- A Few Prognostications Towards Optimizing for Performance, Energy, and Power ...
 - Minimize data movement
 - Traditional visualization vs. in-situ visualization
 - Traditional storage vs. in-situ storage
 - Address energy proportionality
 - Schedule for performance *and power*

The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

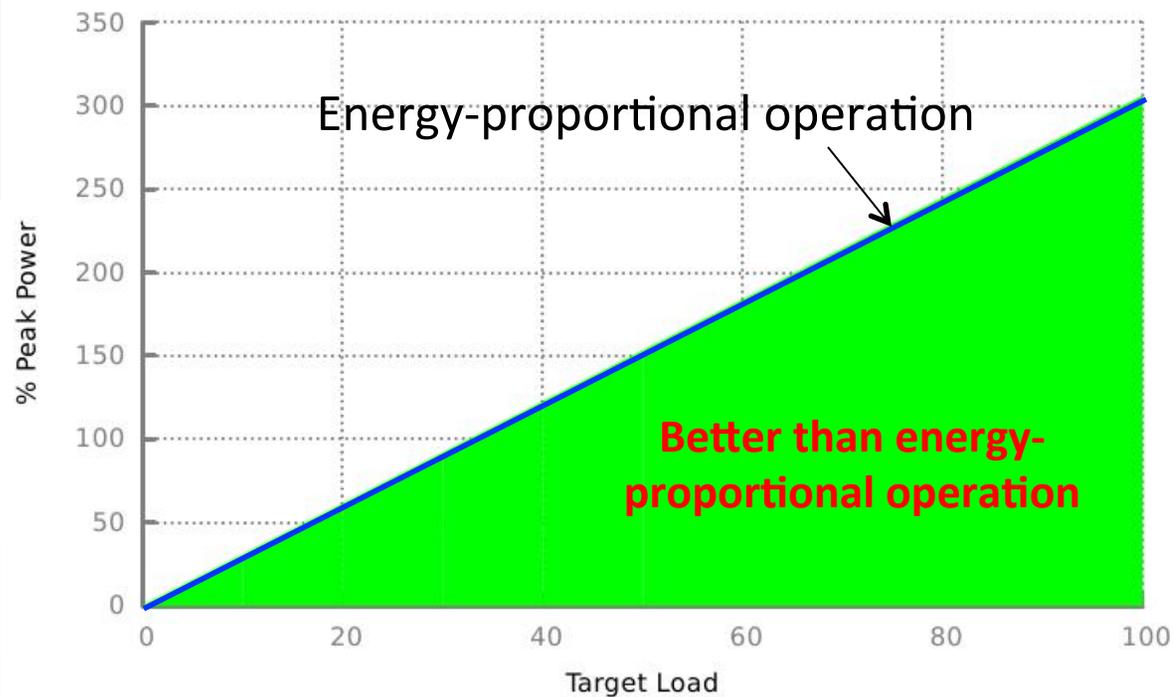
Consume power proportional to utilization (or load-level)



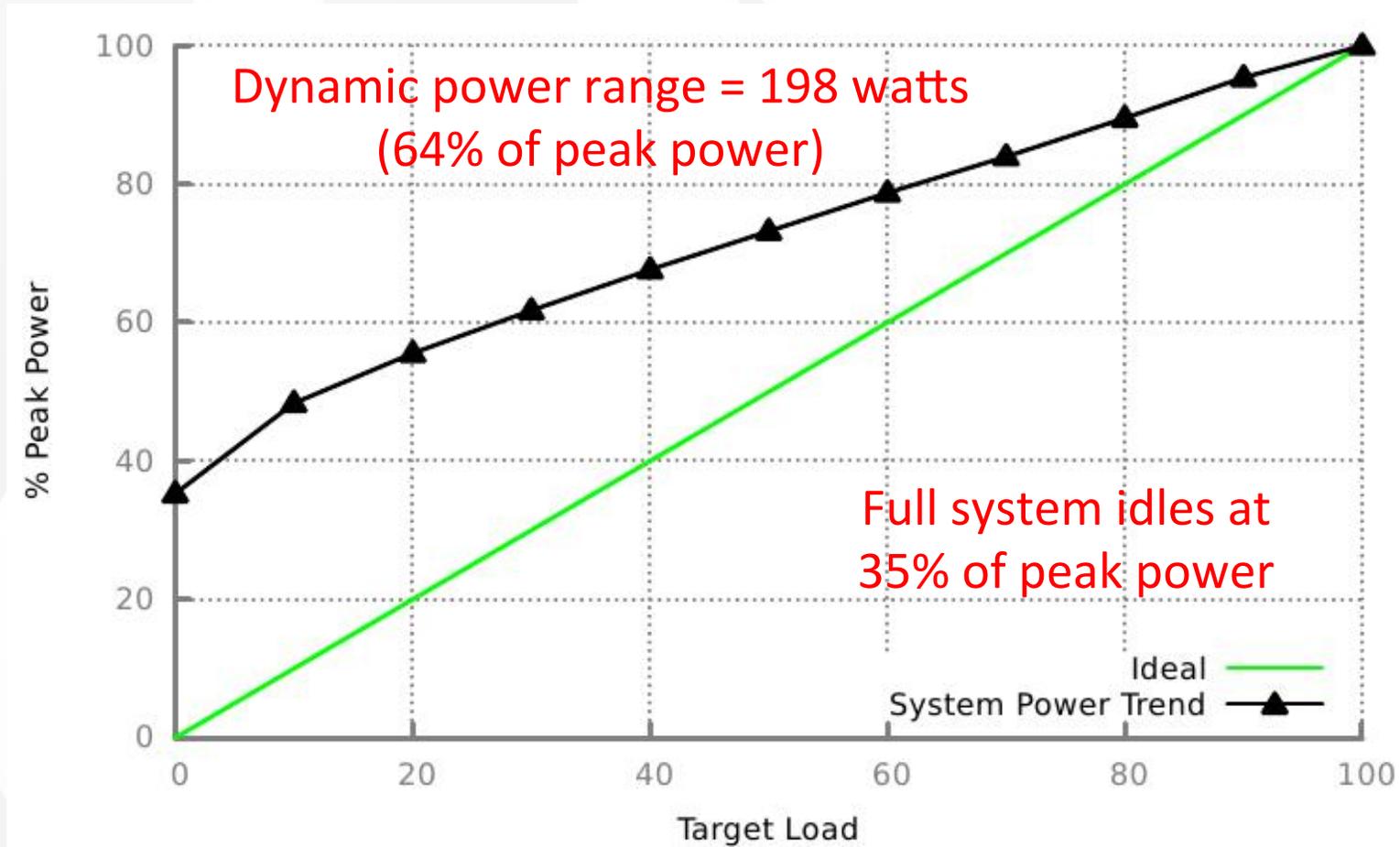
The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

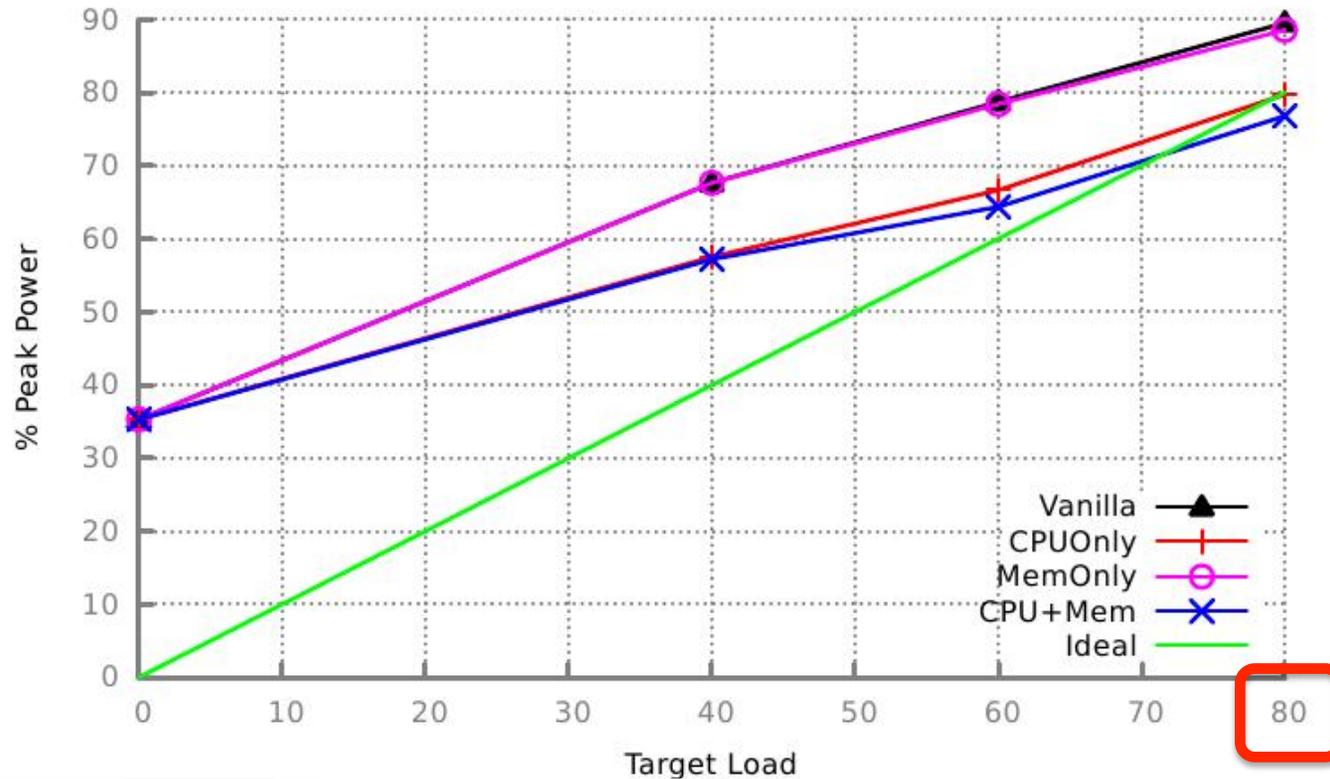
Consume power proportional to utilization (or load-level)



Energy Proportionality – Full System



Power Savings via RAPL – Full System



- CPU+Mem achieves best overall power savings (19% power saved)
- Energy-proportional operation for 80% load-level (via power capping)
- Achieving ideal non-peak power lessens as load-level decreases

Energy Proportionality

B. Subramaniam and W. Feng, “Towards Energy-Proportional Computing for Enterprise-Class Server Workloads,” *ACM/SPEC Int’l Conf. on Performance Engineering*, April 2013. *Best Paper Award*.

Future Work

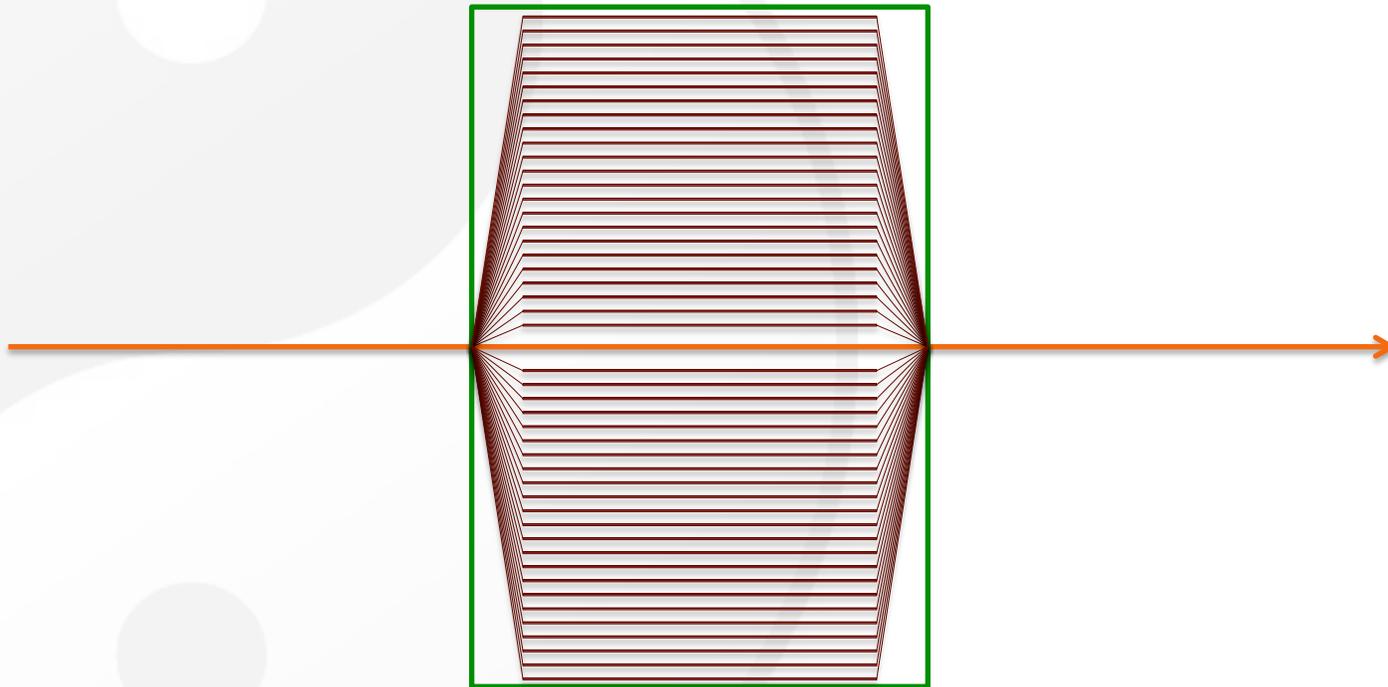
- Fully automate and apply to HPC workloads
- Power sloshing across a system
... rather than just within the CPU

Towards Green Exascale Computing

- A Few Prognostications Towards Optimizing for Performance, Energy, and Power ...
 - Minimize data movement
 - Traditional visualization vs. in-situ visualization
 - Traditional storage vs. in-situ storage
 - Address energy proportionality
 - Schedule for performance *and power*

OpenMP Accelerator Behavior

`#pragma omp acc_region ...`



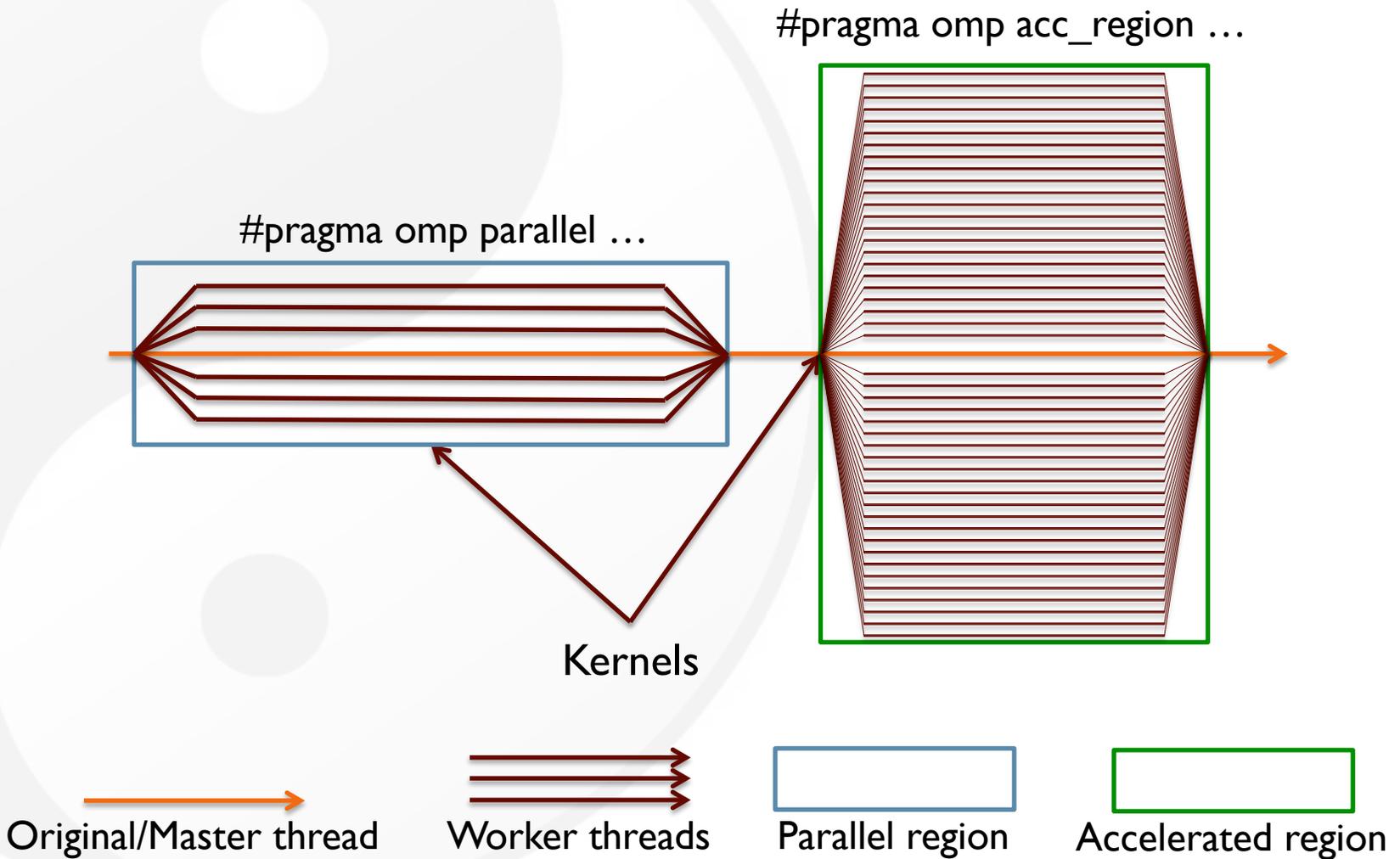
Original/Master thread

Worker threads

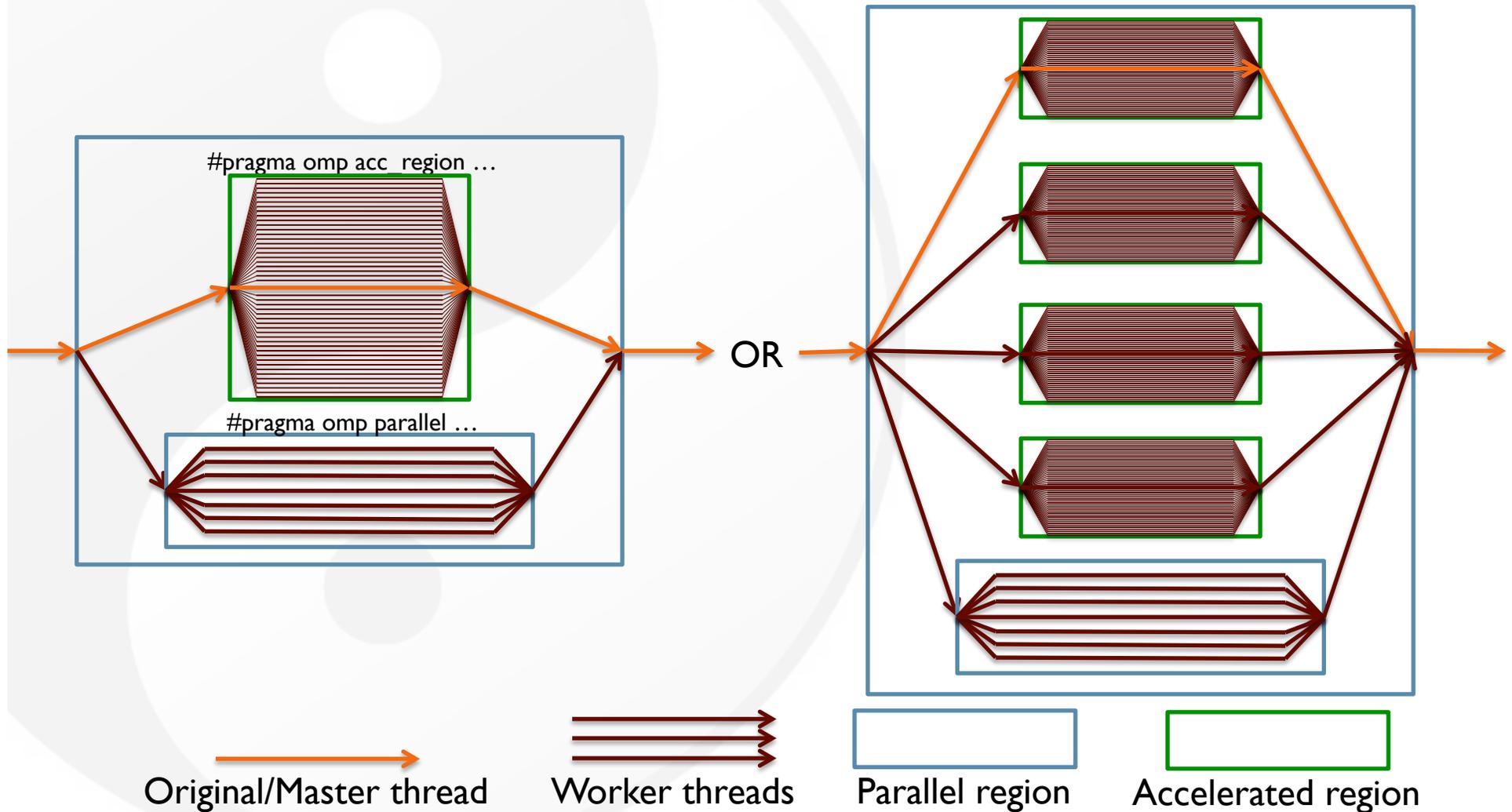
Parallel region

Accelerated region

OpenMP Accelerator Behavior



What We Want: Work-share a Region Across the Whole System



Automated Scheduling and Load-Balancing

- Measure computational suitability at runtime
- Compute new distribution of work
... through a linear optimization approach
- Re-distribute work before each pass

I = total iterations available

i_j = iterations for compute unit j

f_j = fraction of iterations for compute unit j

p_j = recent time/iteration for compute unit j

n = number of compute devices

t_j^+ (or t_j^-) = time over (or under) equal

$$\min\left(\sum_{j=1}^{n-1} t_j^+ + t_j^-\right) \quad (7)$$

$$\sum_{j=1}^n f_j = 1 \quad (8)$$

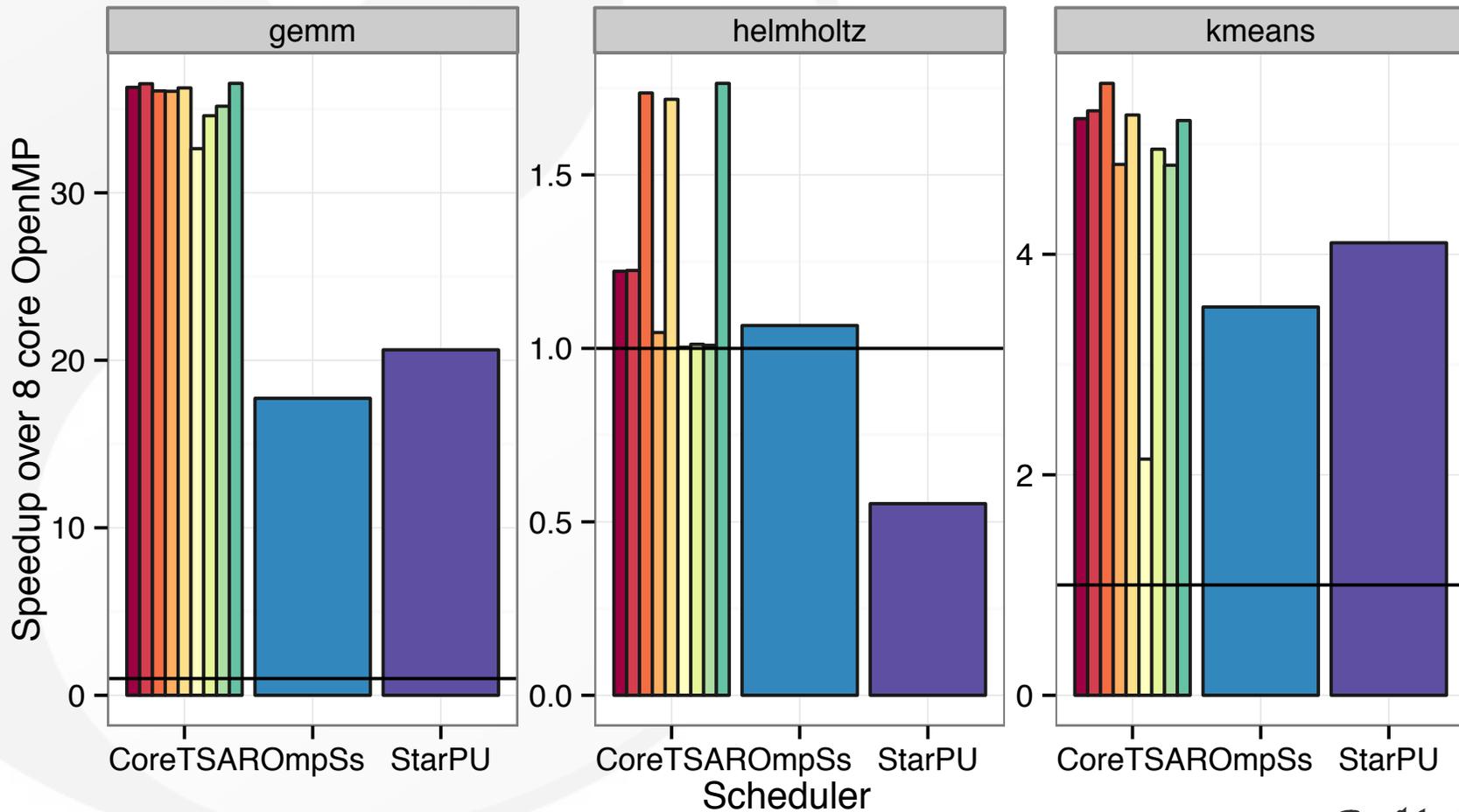
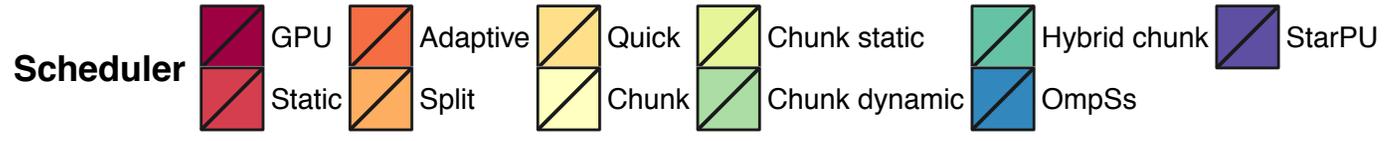
$$f_2 * p_2 - f_1 * p_1 = t_1^+ - t_1^- \quad (9)$$

$$f_3 * p_3 - f_1 * p_1 = t_2^+ - t_2^- \quad (10)$$

⋮

$$f_n * p_n - f_1 * p_1 = t_{n-1}^+ - t_{n-1}^- \quad (11)$$

Our CoreTSAR Results vs. OmpSs and StarPU



Scheduling and Load-Balancing by Adaptation

- Measure computational suitability at runtime
- Compute new distribution of work through a linear optimization approach

Transform from automated performance optimization ... to automated power (or energy) optimization at run time

(7)

(8)

(9)

n = number of compute devices

t_j^+ (or t_j^-) = time over (or under) equal

$$f_2 * p_2 - f_1 * p_1 = t_1^+ - t_1^- \tag{9}$$

$$f_3 * p_3 - f_1 * p_1 = t_2^+ - t_2^- \tag{10}$$

⋮

$$f_n * p_n - f_1 * p_1 = t_{n-1}^+ - t_{n-1}^- \tag{11}$$

Summary for CoreTSAR and Beyond

- Goal
 - Maximize performance (i.e., reduce execution time) of an application running on state-of-the-art HPC system (e.g., heterogeneous systems) under a given power budget by moving around power for various components of a HPC system.
 - Takeaway Message
 - Make the most efficient use of power across a HPC system (via monitoring)
- Power Management System for HPC Systems
 - Framework to decide the power budget for different components for different workloads → “TurboBoost” across a node & between nodes?
- Runtime System
 - Guarantee power limit *not* exceeding while maximizing performance for the given power budget

Conclusion:

Is Green Exascale Computing an Oxymoron?

- Green exascale computing is (optimistically) possible
 - ... with current timeline & continued funding investment
- Some research approaches (to complement the vendors)
 - Minimize data movement
 - Continue efforts on green compute but pay attention to other subsystems and their interactions, i.e., minimize data movement, via monitoring
 - Traditional visualization vs. in-situ visualization
 - The Case of the Missing Supercomputer Energy
 - Know where power and energy are going and address via scheduling, feature leveraging, and working with vendors
 - Automated scheduling (CoreTSAR) but for power (or perf & power)
 - Energy proportionality, not just for enterprise workloads but also for HPC, e.g., job submission scheduling.
 - Handful of nodes idle. Need to be energy proportional

Wu Feng, wfeng@vt.edu, 540-231-1192



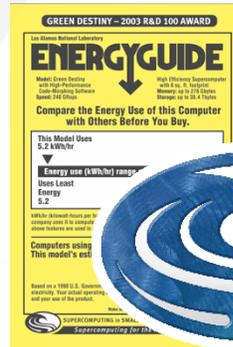
<http://synergy.cs.vt.edu/>



<http://www.chrec.org/>



<http://www.mpiblast.org/>



SUPERCOMPUTING
in SMALL SPACES

<http://sss.cs.vt.edu/>



<http://www.green500.org/>



<http://myvice.cs.vt.edu/>

"Accelerators 'R Us"

<http://accel.cs.vt.edu/>