TECHNISCHE UNIVERSITÄT DRESDEN

ZIH Center for Information Services & High Performance Computing

HAEC

DFG

READEX Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing

European Commission | Horizon 2020 European Union funding for Research & Innovation

DRESDEN concept

CPU Idle Loop Ordering Problem

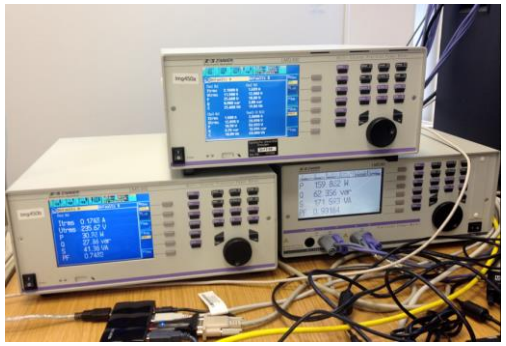
Rafael J. Wysocki
Intel

Thomas Ilsche
Center for Information Services and High Performance Computing (ZIH), TU Dresden

OSPM 2018 – Pisa

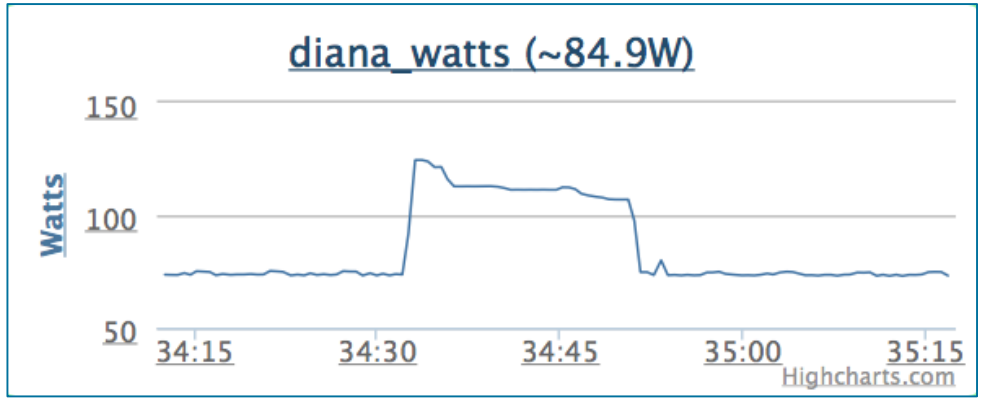
Observing Power Consumption Anomalies

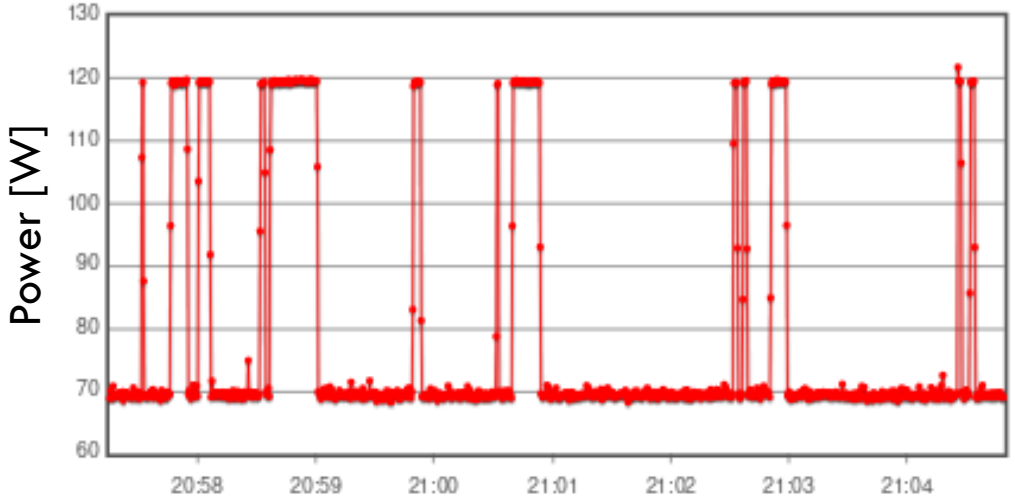
- Energy efficiency research
- Fine grained instrumentation (microsecond resolution)
- Large scale instrumentation (HPC system 1400 nodes)
- Tuned for low idle power consumption



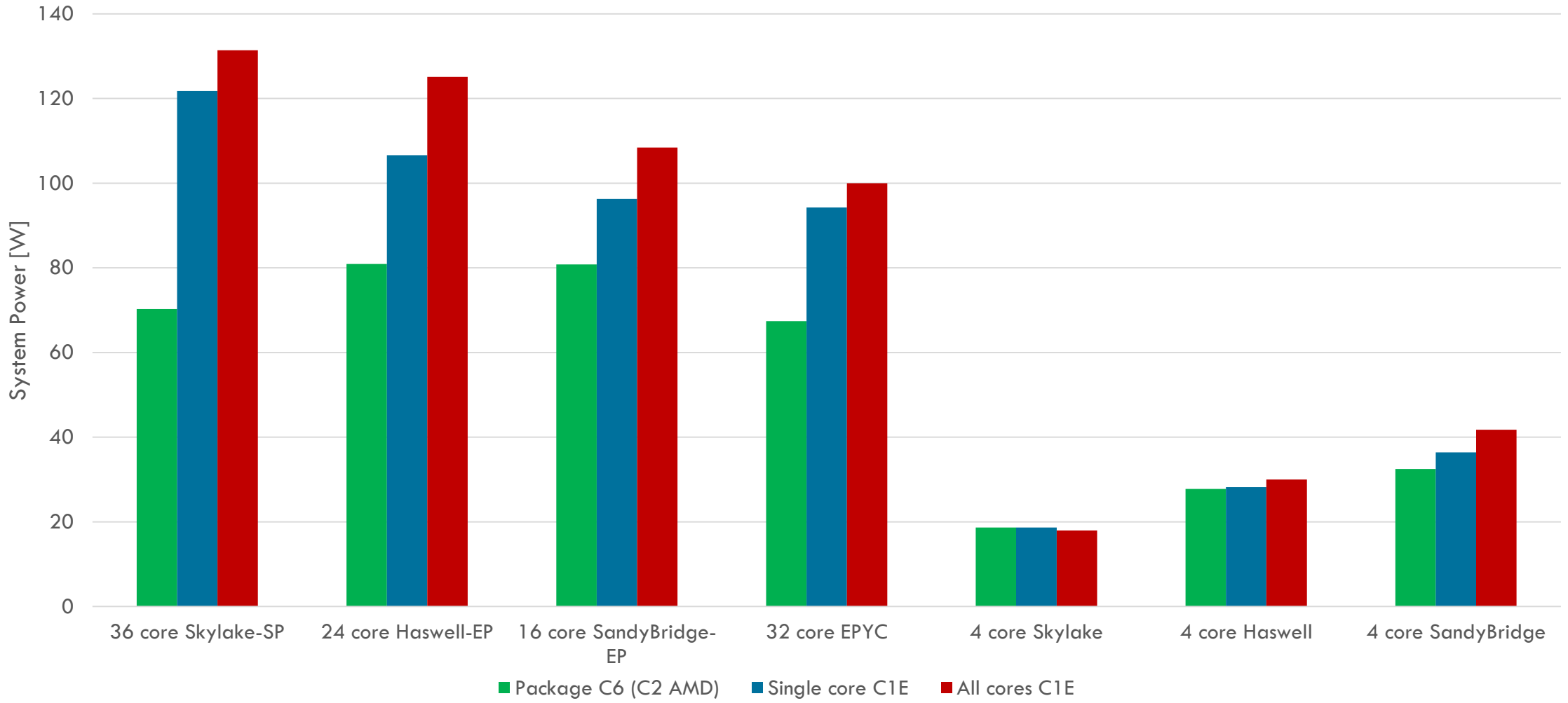
Observing Power Consumption Anomalies

- Prolonged phases of high power consumption during idle
- Disrupted power measurements
- Significant increase in idle power on Skylake 36 core system with stock Ubuntu installation

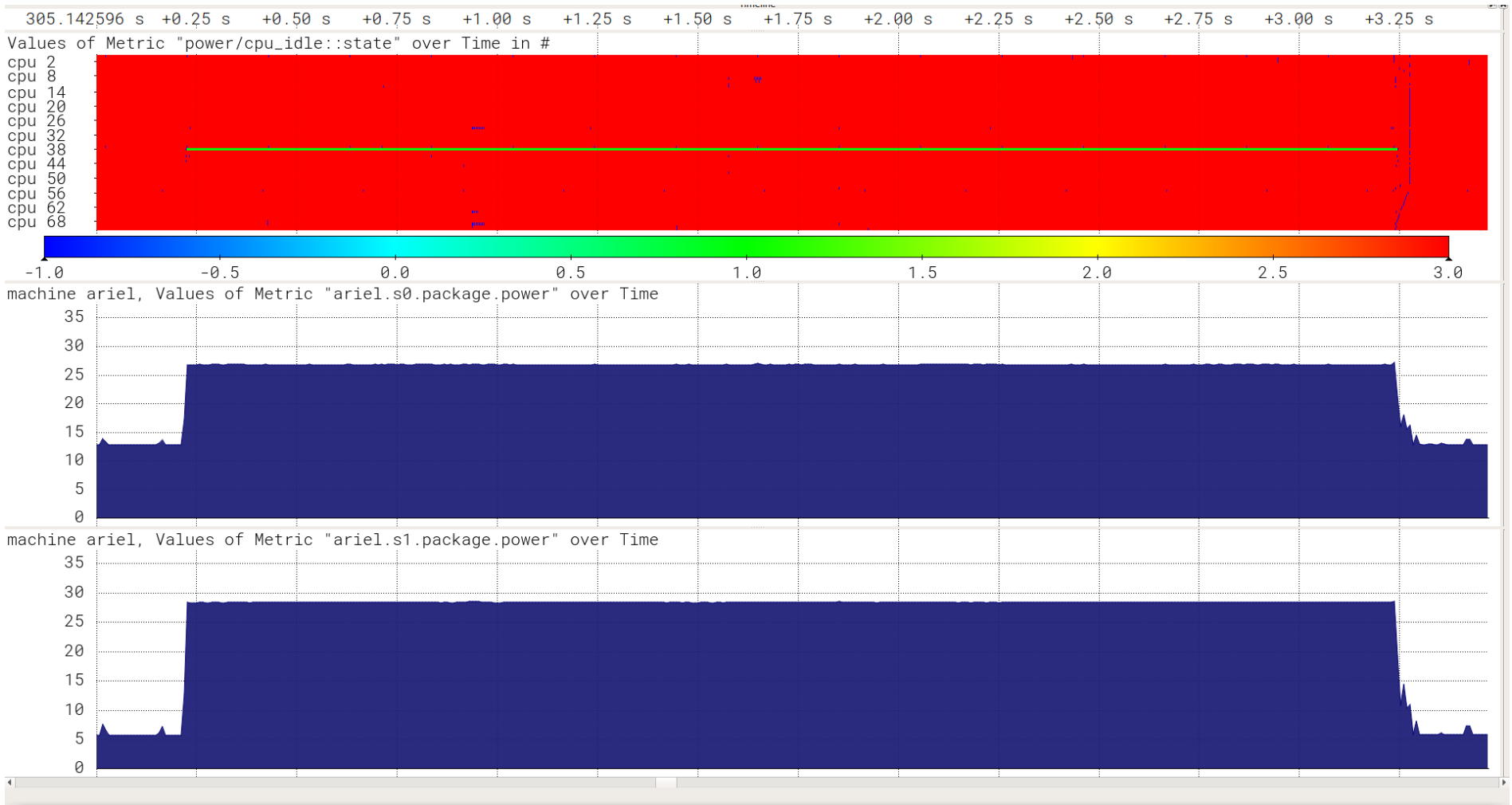




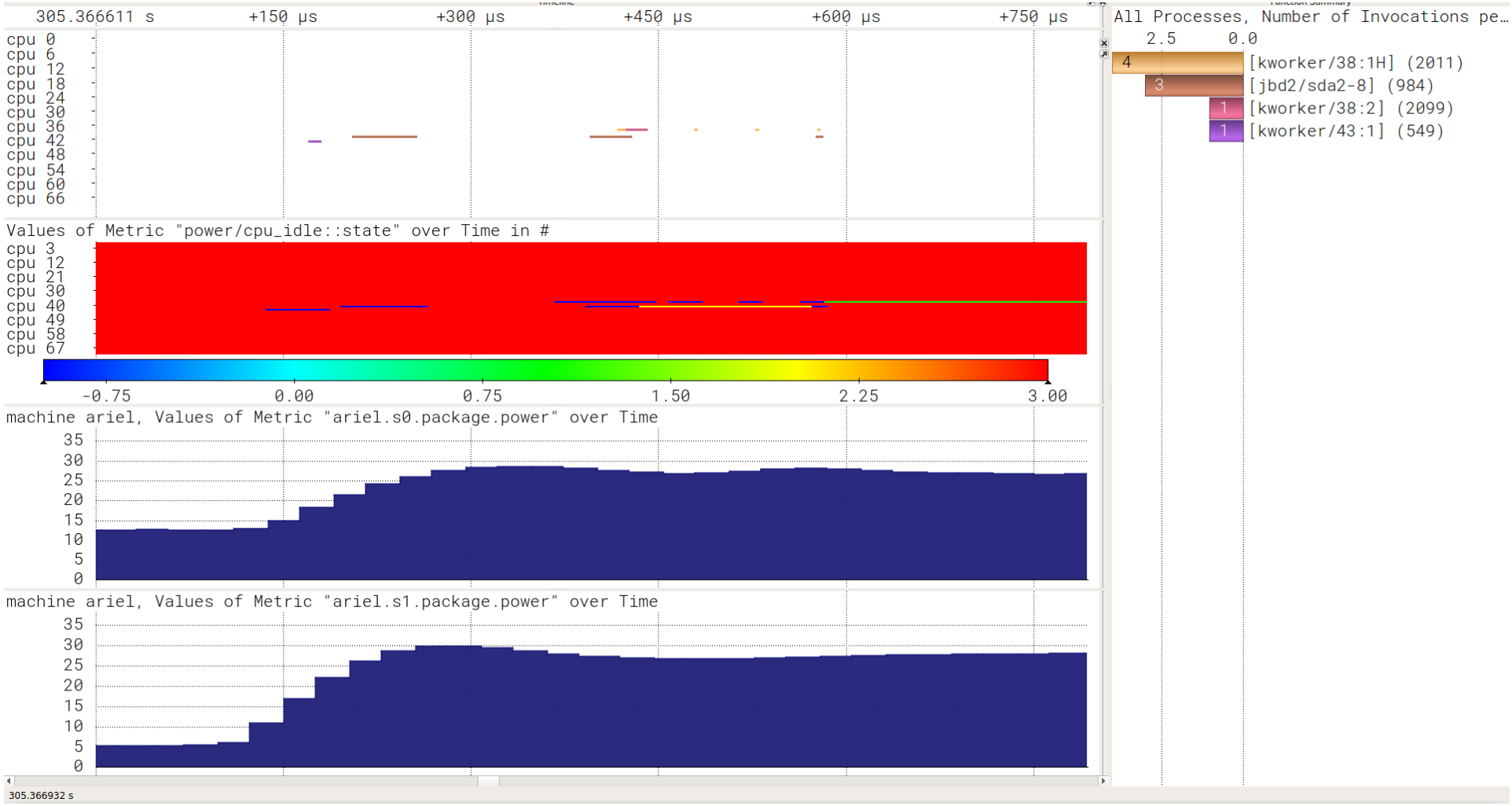
Impact of Package C States



Understanding Power Anomalies



Understanding Power Anomalies

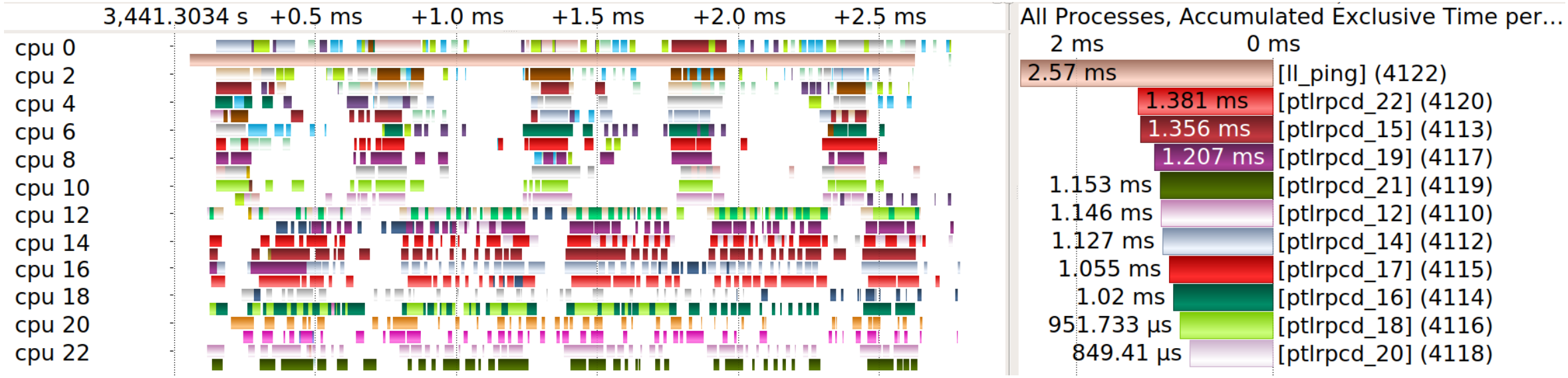


Cause, Trigger, Contributing Factors

- Cause
  - Menu governor heuristic underestimates sleep time
  - Uses repeatable interval detector with 8 data points
  - Non-optimal C state is selected
- Trigger
  - Short sleep phases on one core
  - Interaction between processes, e.g. kworkers, ssh/zsh/screen, lustre ping
  - Synthetic: burst sleep intervals
- Contributing factors
  - Long idle phase, no correction of wrongly selected C state
  - Stubborn heuristic
  - High impact of single core in wrong state

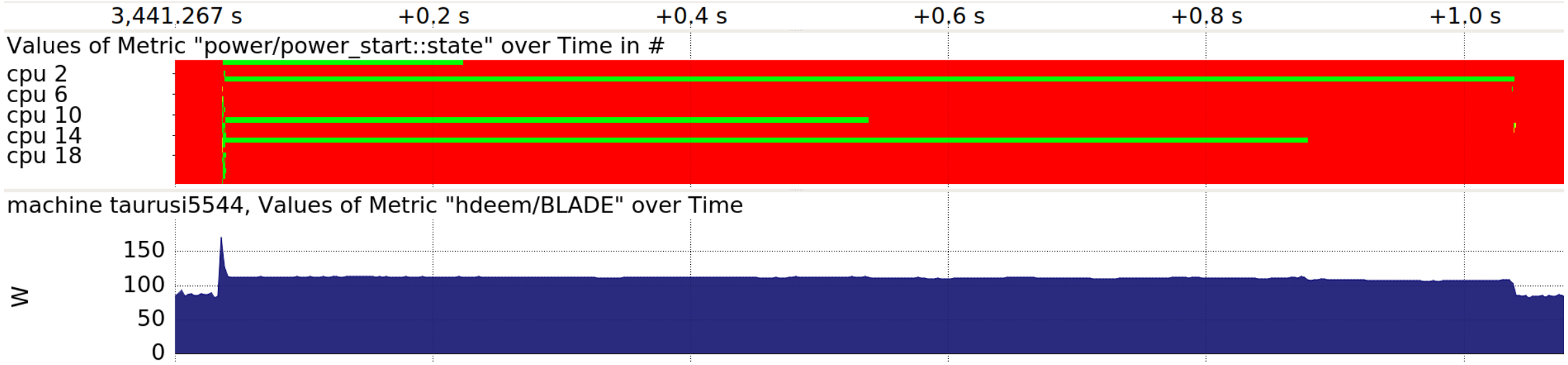
Understanding Power Anomalies (HPC System)

- Found on production HPC System with > 1400 nodes
- Lustre related pattern every 25 seconds

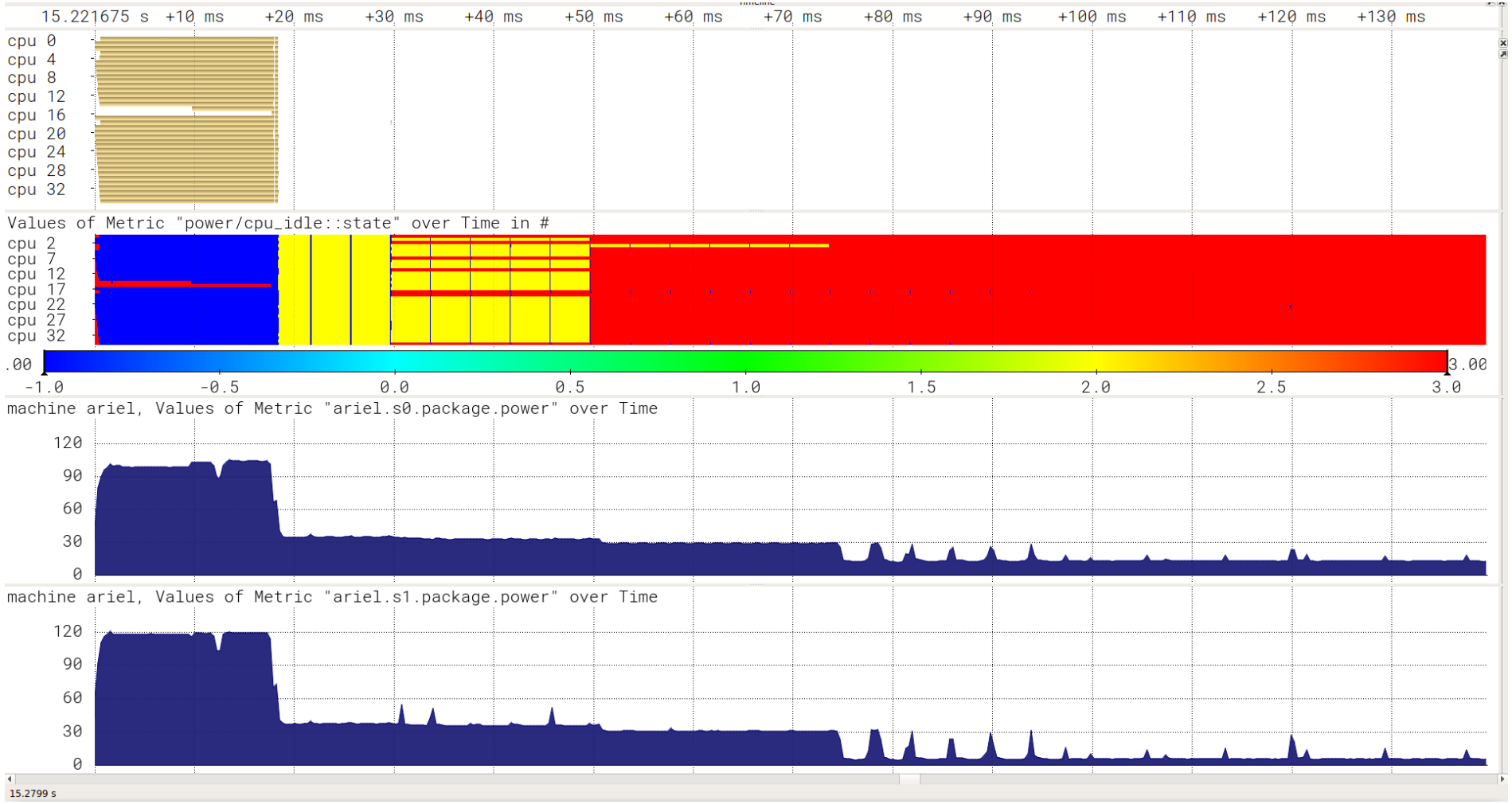


Understanding Power Anomalies (HPC System)

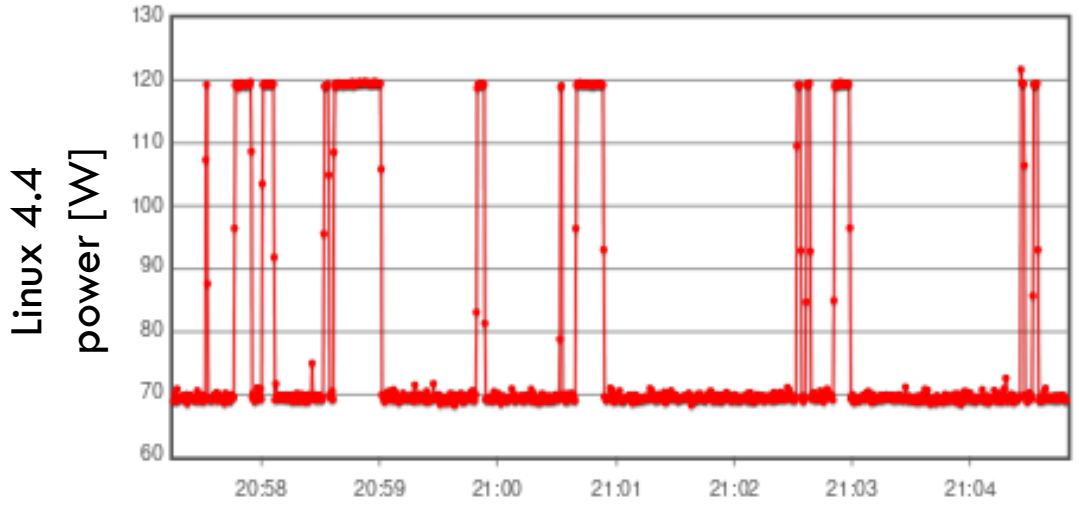
- Found on production HPC System with > 1400 nodes
- Lustre related pattern every 25 seconds
- Triggers up to one second Powernightmare
- 87 W → 131 W
- Lower impact due to regular background activity



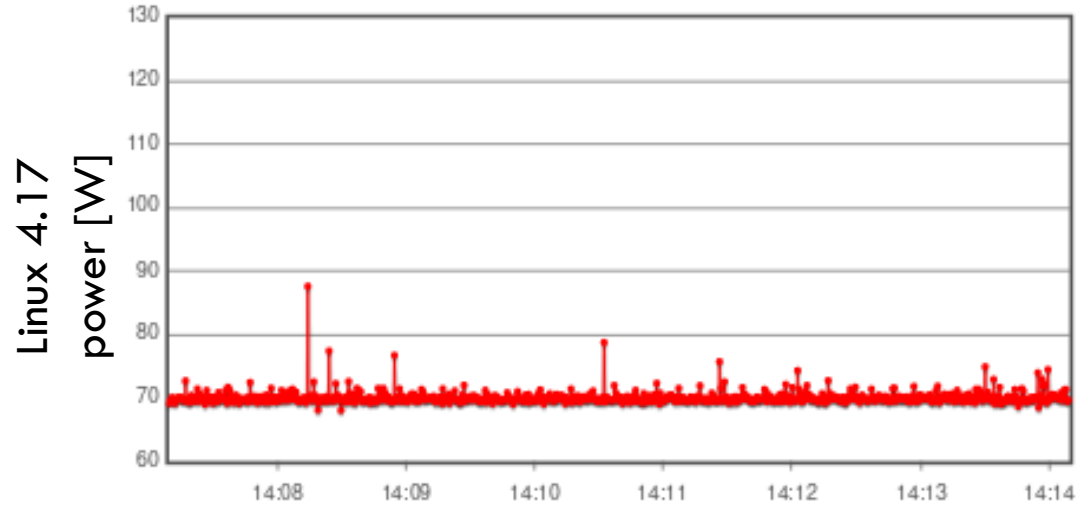
Synthetic trigger with idle-loop v9



Fixing Powernightmares for Linux 4.17



- Dual socket, 36 core SKL-SP system
- Default Ubuntu server installation, fully idle, no extra services
  - Frequent power spikes up to **120 W**
  - Average system power **78 W**



- Upcoming kernel 4.17
  - Constant, low idle power
  - Average system power **70 W** (-10.3%)

References

- Thomas Ilsche, Marcus Hähnel, Robert Schöne, Mario Bielert and Daniel Hackenberg.
  "Powernightmares: The Challenge of Efficiently Using Sleep States on Multi-Core Systems“
  In: 5th Workshop on Runtime and Operating Systems for the Many-core Era (ROME). 2017
- Thomas Ilsche, Robert Schöne, Mario Bielert, Andreas Gocht and Daniel Hackenberg.
  "lo2s – Multi-Core System and Application Performance Analysis for Linux“
  In: Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA). 2017. DOI: 10.1109/CLUSTER.2017.116
  🔗 https://github.com/tud-zih-energy/lo2s
- Thomas Ilsche, Robert Schöne, Joseph Schuchart, Daniel Hackenberg, Marc Simon, Yiannis Georgiou and Wolfgang E. Nagel.
  "Power Measurement Techniques for Energy-Efficient Computing: Reconciling Scalability, Resolution, and Accuracy" In: Second Workshop on Energy-Aware High Performance Computing (EnA-HPC). 2017