SAND2017-12440C

Power API User Experiences

Ryan E. Grant, James H. Laros III, Steve Martin, Lee Ward

Outline

Sandia National Laboratories

- Overview of Power API uses in 2017

- Steve Matin
  - Cray experiences in deploying and using the Power API

- Lee Ward
  - Experiences using the Power API for monitoring and control of production applications

Overview

Sandia National Laboratories

- Power API 2.0 Implementation
  - Reference Implementation:

    100s of downloads in 2017

    Deployed in the Tri-lab Operating System – many significant sized clusters supported at LLNL, LANL and SNL
  - Cray Implementation:

    Complete and functional

    Works with CapMC

    Performance optimized for Cray XC systems

Use of Power API at Cray

- Cray implemented the Power API for XC series systems
- Used to test, evaluate and tune Power API implementation
- Several improvements discovered while optimizing library
  - Opportunities for tuning statistics interface for greater efficiency
  - Additional attributes to add to the specification

Figures

Use of Power API on Large Systems
-Lee Ward-

Workloads

- MiniMD
  - A molecular dynamics mini app representing LAMMPS
  - Supports Lennard-Jones pair interactions, equivalent to LJ liquid simulations
- LULESH
  - Unstructured mesh Lagrangian explicit shock hydrodynamics mini app
  - Representing DOE hydrodynamics apps, especially ALE3D
- MiniFE
  - Implicit finite element (condition simulation using a conjugate gradient solver
  - Rectangular problem, represents broad class of FE apps, CG is generic

Job-wide Aggregate Information

- Fastest clock speeds are almost universally the most performant
  - Exception is HSW running MiniFE where non-turbo is slightly better
- FOM/Watt is not always straightforward
  - No information when job phase performance is related to energy usage
  - MiniFE, HSW shows 1.2Ghz best FOM/Watt but total runtime != FOM

		Figure of Merit Per Node			**Average Watts Per Node**			**Figure of Merit Per Watt**		
		Volta Ivy Bridge	Trinity Haswell	Trinity Phi KNL	Volta Ivy Bridge	Trinity Haswell	Trinity Phi KNL	Volta Ivy Bridge	Trinity Haswell	Trinity Phi KNL
MiniMD	Turbo	**2.08e7**	**3.01e7**	**5.92e7**	269	334	246	7.73e4	9.01e4	2.41e5
	No Turbo	1.84e7	2.56e7	5.66e7	213	236	228	**8.64e4**	**1.08e5**	2.48e5
	1.2 GHz	9.45e6	1.39e7	4.93e7	138	142	194	6.85e4	9.79e4	**2.54e5**
LULESH	Turbo	**1.36e4**	**1.85e4**	**1.51e4**	291	346	218	46.7	53.5	69.3
	No Turbo	1.24e4	1.75e4	1.43e4	236	295	208	**52.5**	59.3	68.8
	1.2 GHz	6.75e3	1.09e4	1.25e4	156	175	180	43.3	**62.3**	**69.4**
MiniFE	Turbo	**1.24e4**	1.42e4	**3.00e4**	185	212	138	67.0	67.0	217
	No Turbo	1.23e4	**1.43e4**	2.93e4	145	152	133	**84.8**	94.1	**220**
	1.2 GHz	8.37e3	1.41e4	2.76e4	104	104	127	80.5	**136**	217

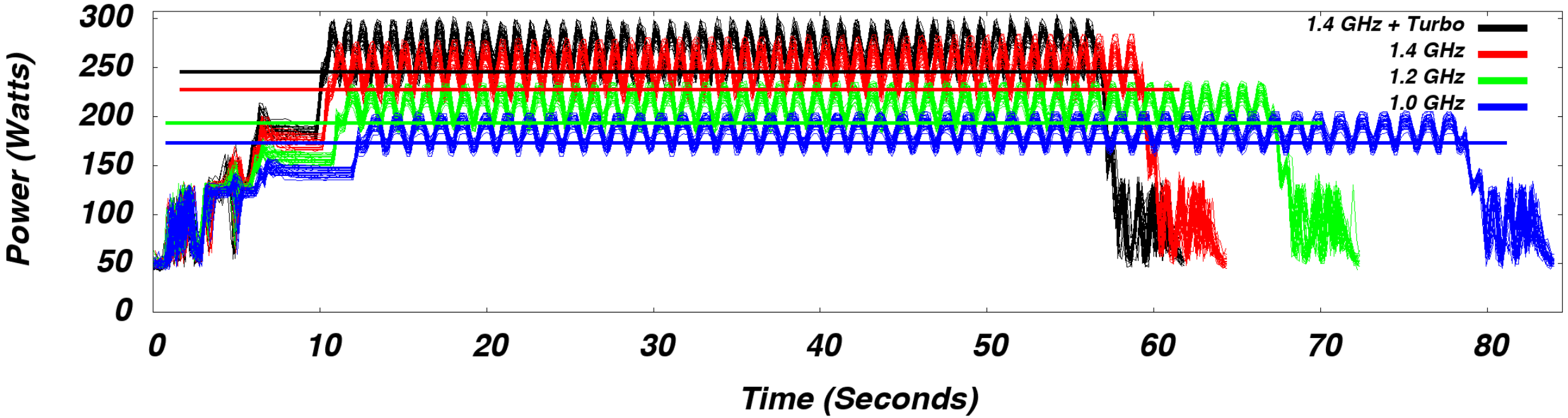
Overhead of Application Profiling

- Power profiling can be done through in-band measurement
  - Ex: calling PowerAPI or using RAPL directly
- The overhead of in-band measurement can be significant
  - Need to effectively quantify overhead
- Overhead found with in-band with region sampling of MiniMD and LULESH
  - 4-8% is significant for HPC apps
  - Overhead could increase running at extreme scale due to increased noise

		Power + Energy Region Profiling	Timestamps Only
MiniMD	Turbo	6.84%	-0.08%
	1.4 GHz	7.49%	-0.08%
	1.2 GHz	7.71%	0.08%
	1.0 GHz	8.15%	0.07%
LULESH	Turbo	4.84%	0.24%
	1.4 GHz	4.94%	0.35%
	1.2 GHz	5.23%	0.22%
	1.0 GHz	4.73%	-0.08%
MiniFE	Turbo	-1.22%	0.15%
	1.4 GHz	-0.59%	-0.95%
	1.2 GHz	-1.50%	-1.42%
	1.0 GHz	-1.26%	-1.96%

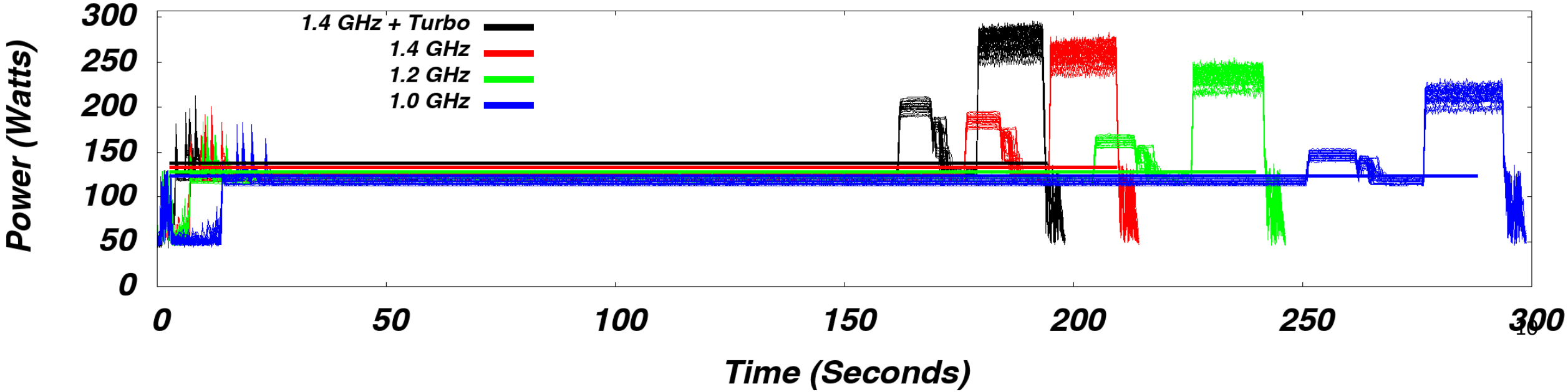
Out-of-Band Periodic Sampling- MiniMD

- Illustrates expected behavior in main solve region
- Periodic power consumption corresponds with known solver phases
- Each P-state shows the expected number of phases
- Lower P-state lengthens phases but does not alter power trends



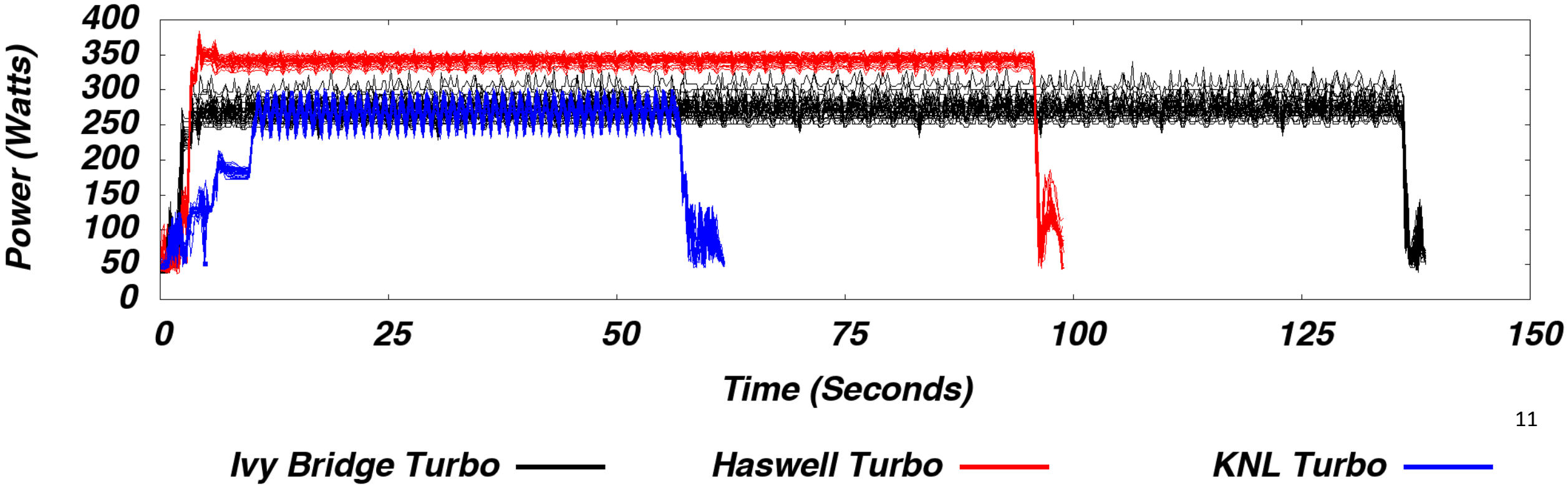
Out-of-Band Periodic Sampling- MiniFE

- MiniFE shows 2 phases well
  - Long, low-power Assembly phase
  - Short, high-power problem Solve phase (CG)
- P-state for Assembly has little effect power, but large effect on runtime
  - FOM calculated only on Solve phase



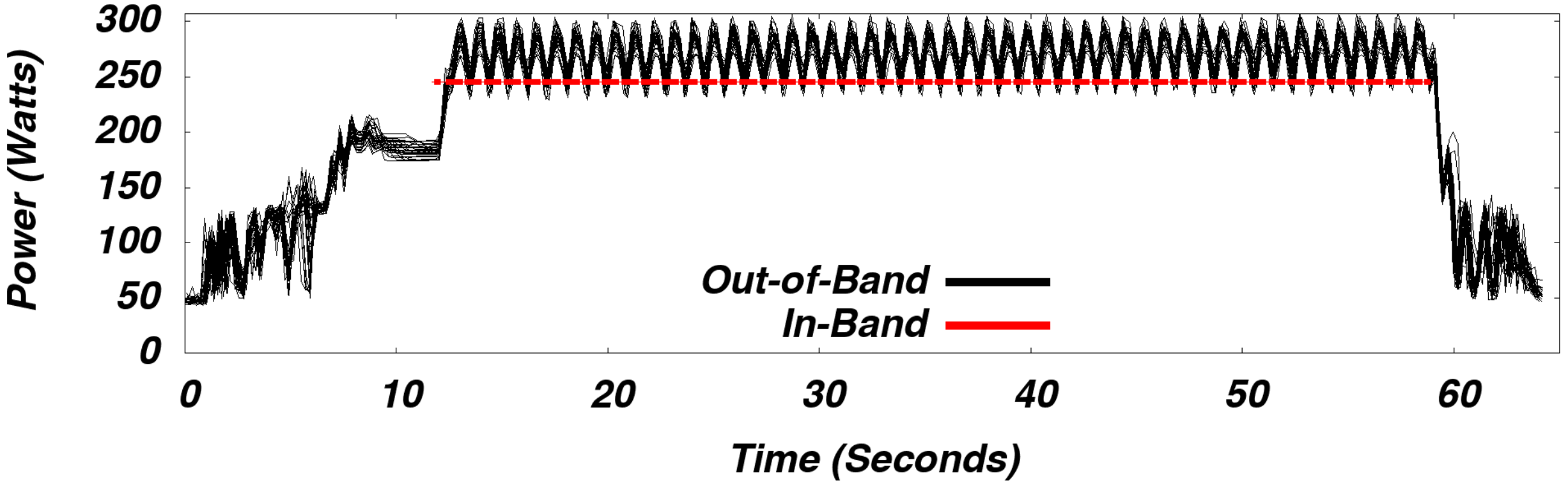
Sampling across 3 Platforms - MiniMD

- Can also compare all 3 platforms using out-of-band measurement
- See MiniMD periodic behavior in KNL, but IVB has more noise



In-band vs Out-of-Band Region Profiling

- Region profiling w/ timestamps and in-band power/energy profiling paired with the collection of out-of-band data allows for better insight
- See issue of resolution with in-band data sampled at region entry/exit
  - Periodicity missed with in-band sampling but seen in out-of-band sampling



Conclusion

- Detailed power profiling is possible on large—scale HPC systems
- The combination of application region profiling and out-of-band power measurement provides an accurate view of application power profiles with negligible overhead

Power API Meeting

- Power API Specification Community Meeting!
  - December 4th-5th, 2017
  - At Intel's Ronler Acres Campus in Hillsboro OR, USA
- Purpose
  - This group will propose, review and vote on changes to the Power API specification as a community effort
  - Consider the first proposals for changes to the specification
  - Interaction between, gov labs, vendors, and academia
- Agenda
  - Introductory tutorials on the Power API
  - Formulation of working group rules and procedures
  - Infrastructure discussion

powerapi.sandia.gov