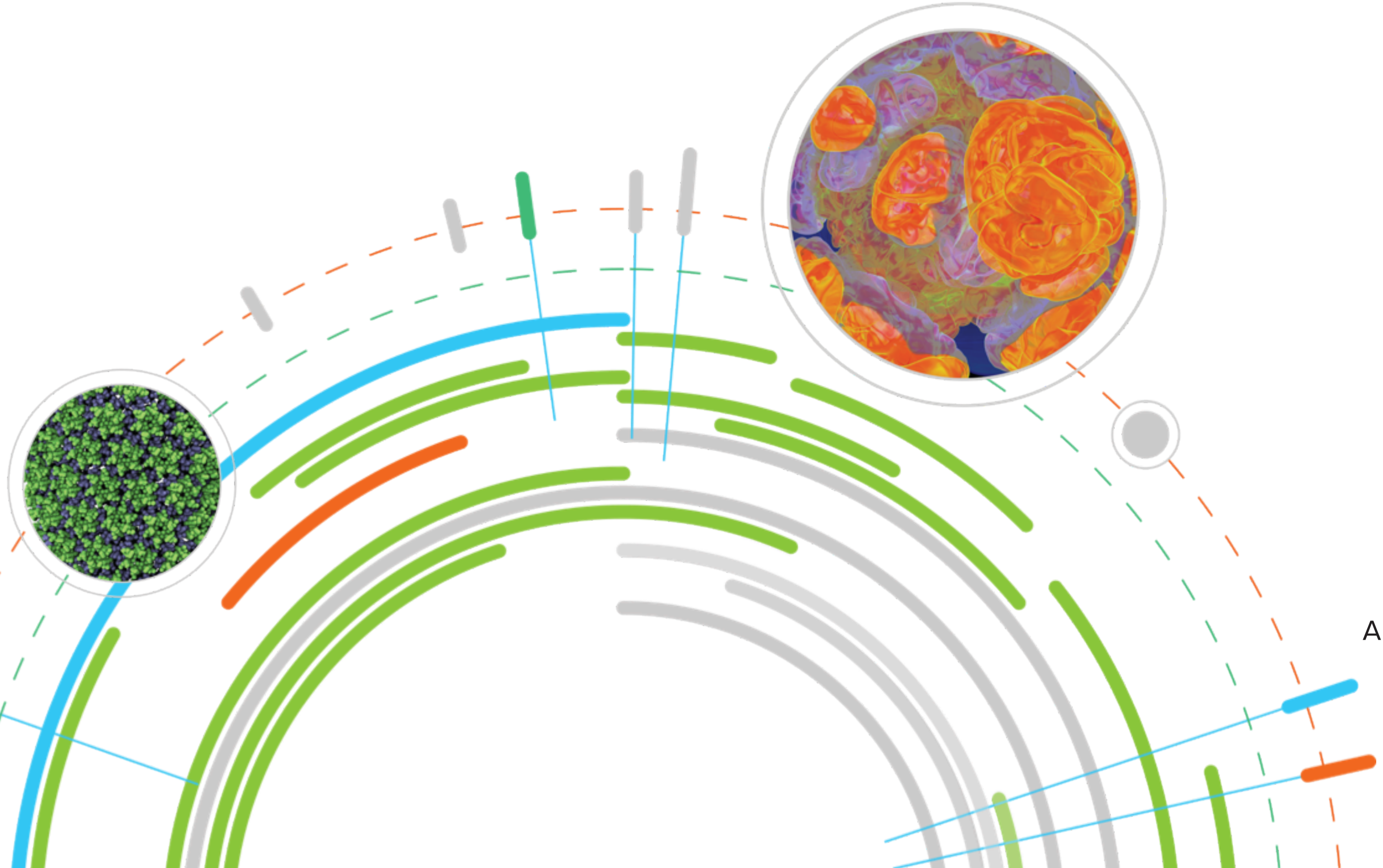
Application Power Profiling on IBM Blue Gene/Q

Sean Wallace*+, Venkatram Vishwanath+, Susan Coghlan+, John Tramm+, Zhiling Lan*, Michael E. Papka+#

*Illinois Institute of Technology, Chicago, IL, USA
+Argonne National Laboratory, Argonne, IL, USA
#Northern Illinois University, DeKalb, IL, USA

swallac6@iit.edu



Argonne Leadership Computing Facility

Argonne NATIONAL LABORATORY

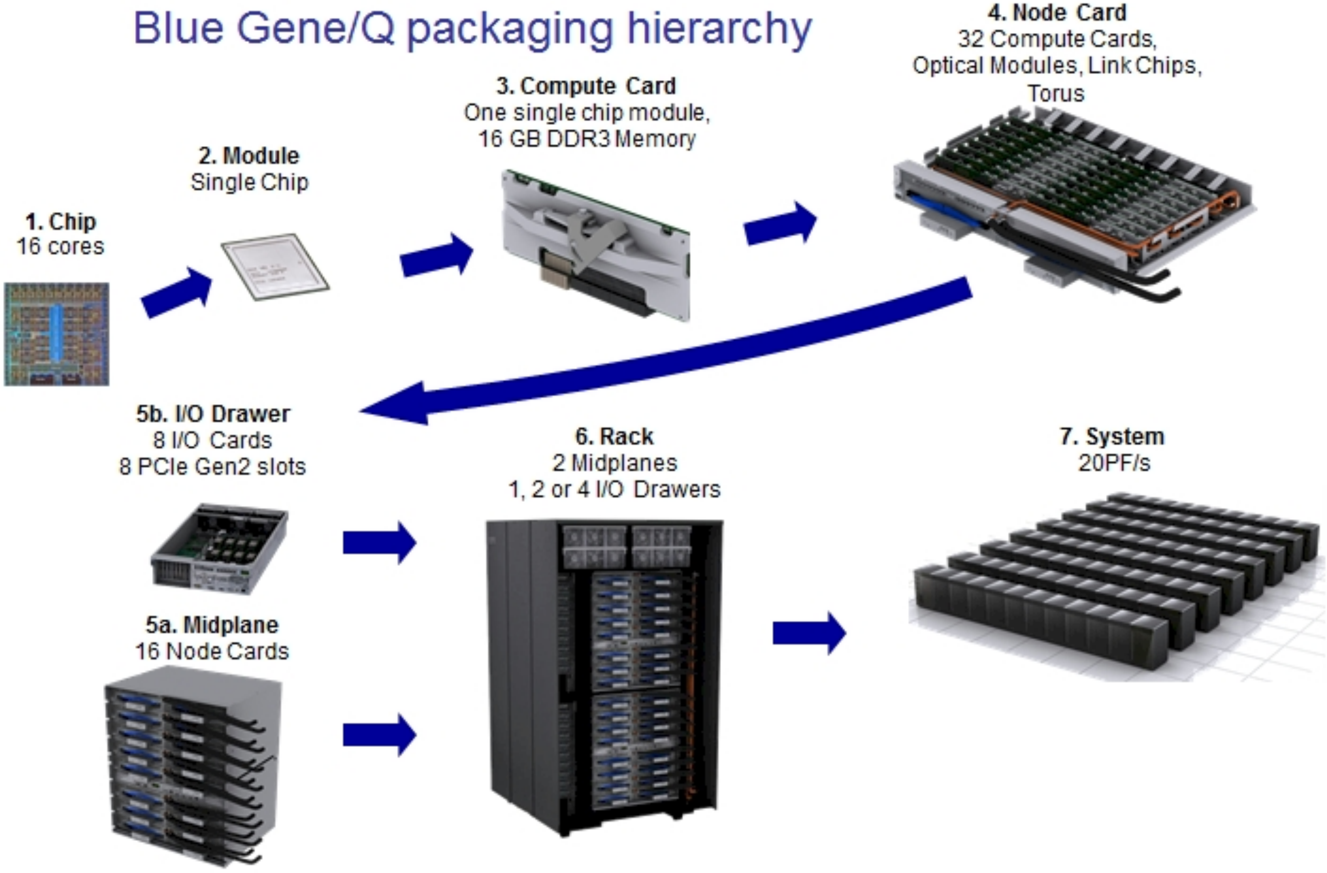
Motivation

- Power consumption is becoming an increasingly vital area of research.
- Projections show that exascale systems will be capped at power consumption of 20 MW.
  - To reach exascale then, current super computers need to scale by a factor of about **60** while increasing power by only a factor of **2**.
- Analysis of power consumption on state of the art supercomputers is imperative to understand how they differ from previous generations.
  - What’s gotten better? What’s worse? What questions can’t we answer?
- Hardware can not solve this problem alone, software has a huge role to play.

Outline

- Blue Gene/Q Architecture and Environmental Data Collection
- Argonne Leadership Computing Facility (ALCF) Science
- Environmental Power Analysis
- EMON/MonEQ Power Analysis
- MonEQ "How-To"

IBM Blue Gene/Q



Blue Gene/Q packaging hierarchy

1. Chip
16 cores

2. Module
Single Chip

3. Compute Card
One single chip module,
16 GB DDR3 Memory

4. Node Card
32 Compute Cards,
Optical Modules, Link Chips,
Torus

5b. I/O Drawer
8 I/O Cards
8 PCIe Gen2 slots

5a. Midplane
16 Node Cards

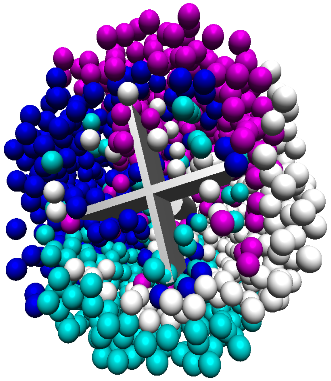
6. Rack
2 Midplanes
1, 2 or 4 I/O Drawers

7. System
20PF/s

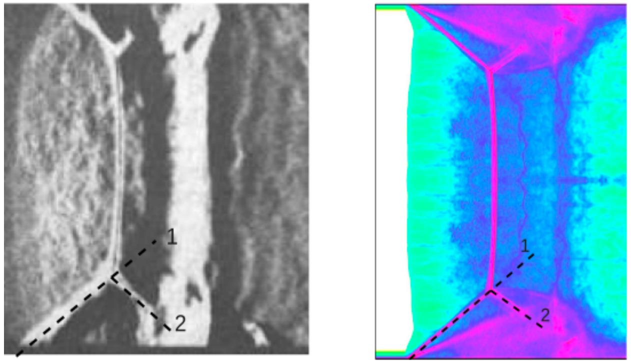
Mira

- 48-racks of Blue Gene/Q
  - Giving 768k cores.
- Five time more energy efficient than Intrepid, its Blue Gene/P predecessor.
- 10PF peak performance.

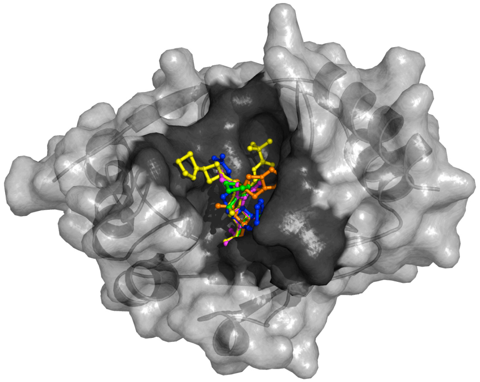
Advances achieved by ALCF users



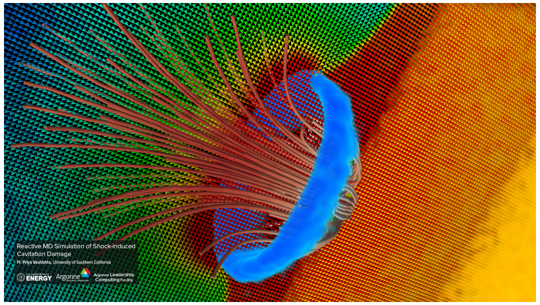
Snapshot of a simulation of a 4-blade vane rheometer with a suspension of hard spheres.



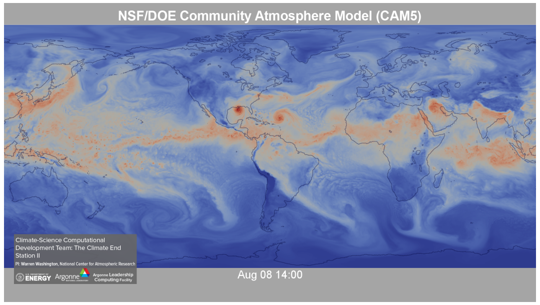
Shock bifurcation In CO2. Experiment on left, simulation on right. Simulations for H2O2 using the same code have for the first time agreed with Experiment and show the geometry of the shock for the first time



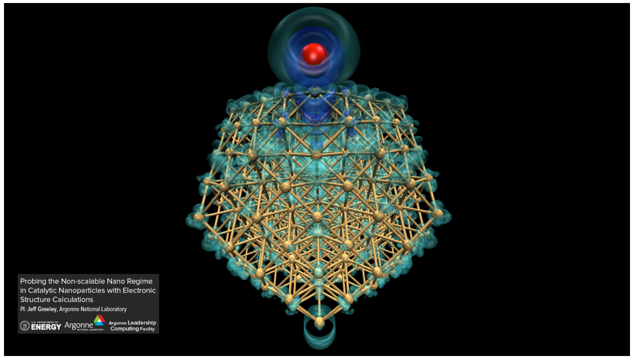
The NDM-1 enzyme's structure revealed a large cavity (dark gray) capable of binding a variety of known antibiotics and then destroying the compound's antibiotic activity.


Reactive MD Simulation of Shock-induced Cavitation Damage
PI: Priya Vashishta, University of Southern California

Vashishta corrosion cracking project

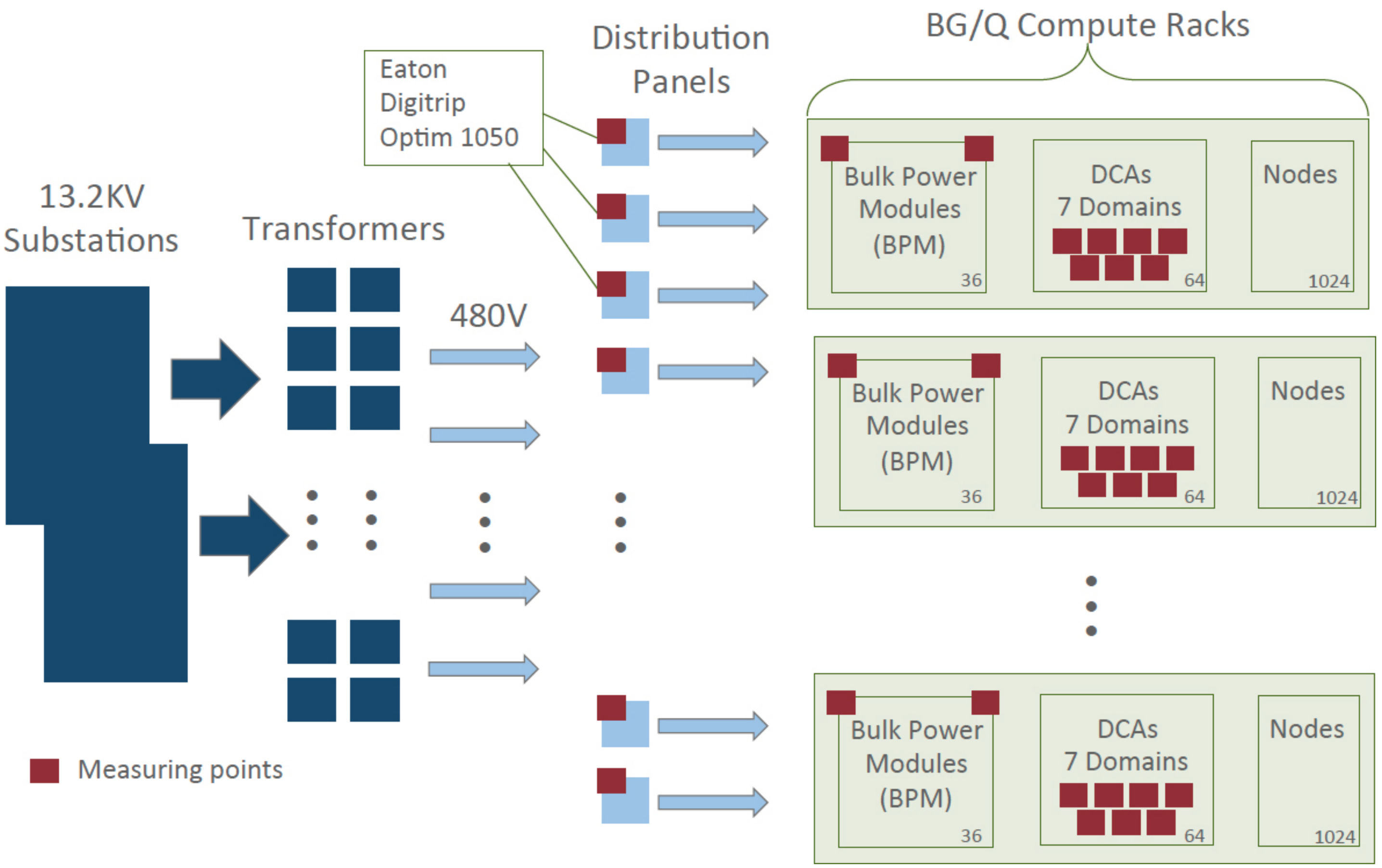

NSF/DOE Community Atmosphere Model (CAM5)
Climate-Science Computational Development Team: The Climate End Station II
PI: Warren Washington, National Center for Atmospheric Research
Aug 08 14:00

Washington


Probing the Non-scalable Nano Regime in Catalytic Nanoparticles with Electronic Structure Calculations
PI: Jeff Greeley, Argonne National Laboratory

Greeley catalysis CO CO2

Power Distribution



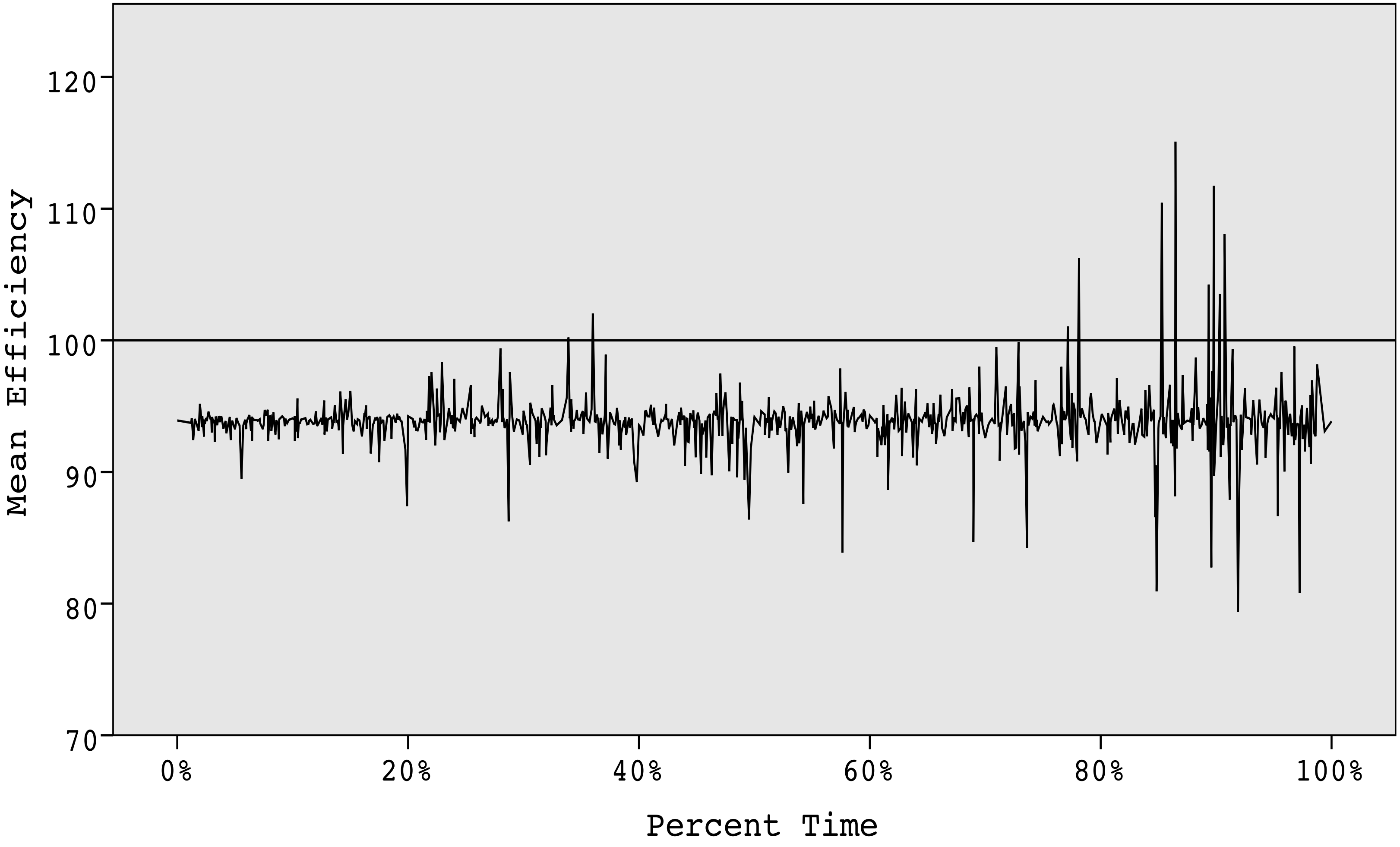
Node Board Power Domains

Domain ID	Description
1	Chip Core Voltage
2	Chip Memory Interface and DRAM Voltage
6	HSS Network Transceiver Voltage Compute+Link Chip
7	Chip SRAM Voltage
3	Optics
4	Optics + PCI Express
8	Link Chip Core

Environmental Database

- IBM DB2 relational database
- Data is populated by polling sensors every 4 minutes
- Power (watts/amperes) and temperature (degrees Celsius):
  - Bulk - AC/DC converter, power in both input and output
  - Node Board - Temperature
  - Node - Temperature
  - Link Card - Temperature
  - Service Card - Temperature
  - Coolant - Temperature sensors between inlet and outlet pipes

BPM Efficiency Over Time



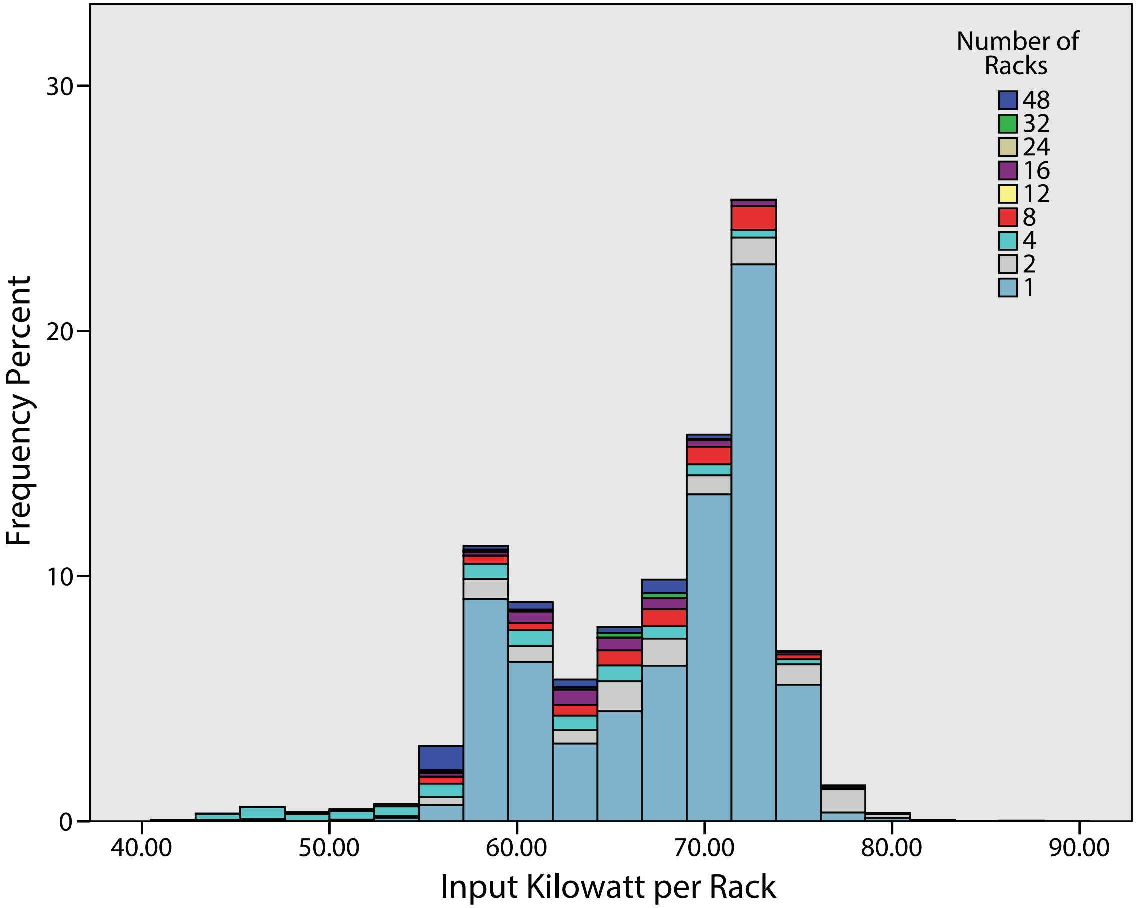
Power and Temperature Data

- Six month sample starting in December 2012
- Results provided from point of view of “average” job.

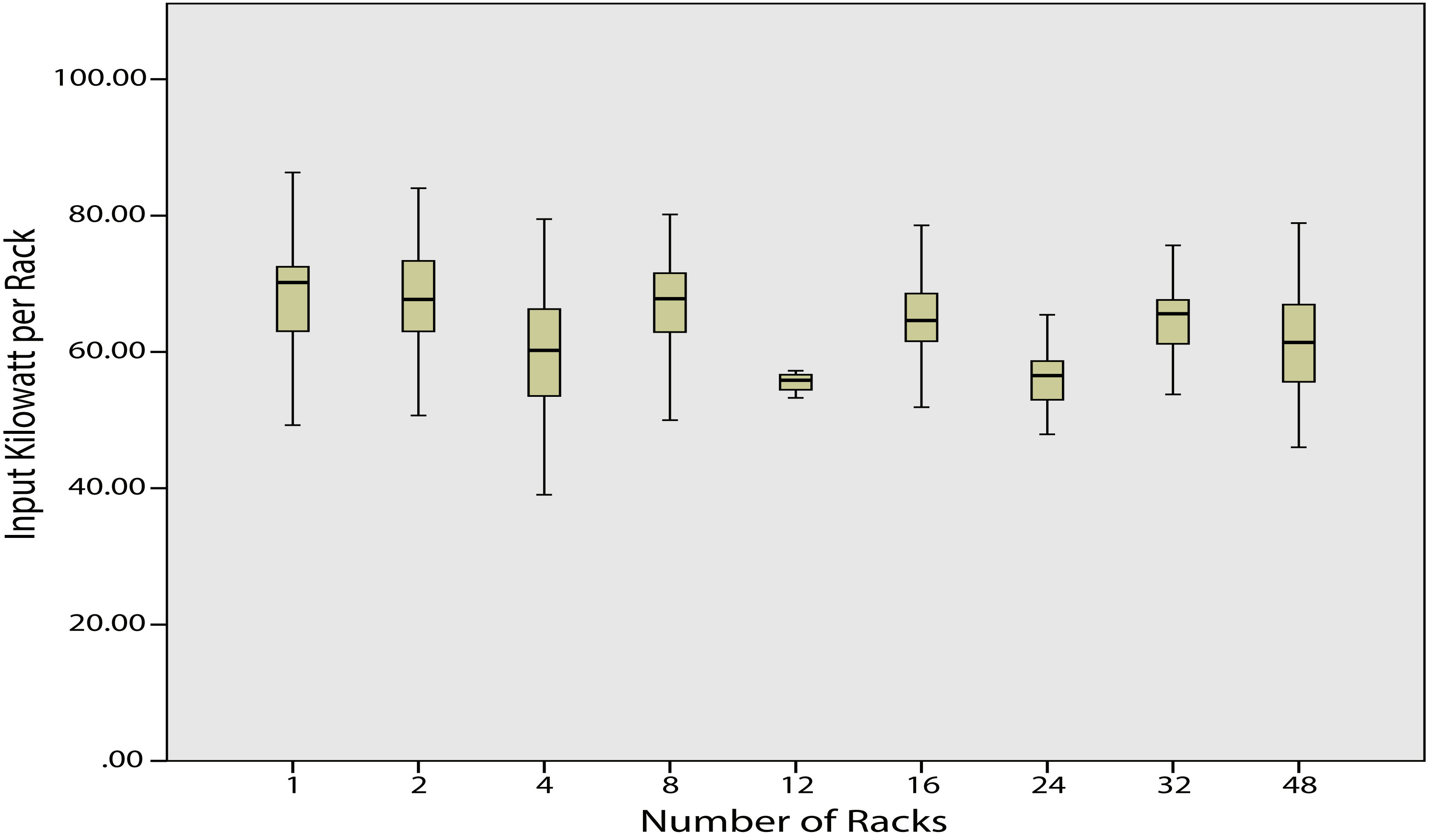
Number of Racks	Number of Jobs
1	14,520
2	2,278
4	1,771
8	1,203
12	5
16	739
24	38
32	218
48	1,006
Total	**21,778**

Environmental Power

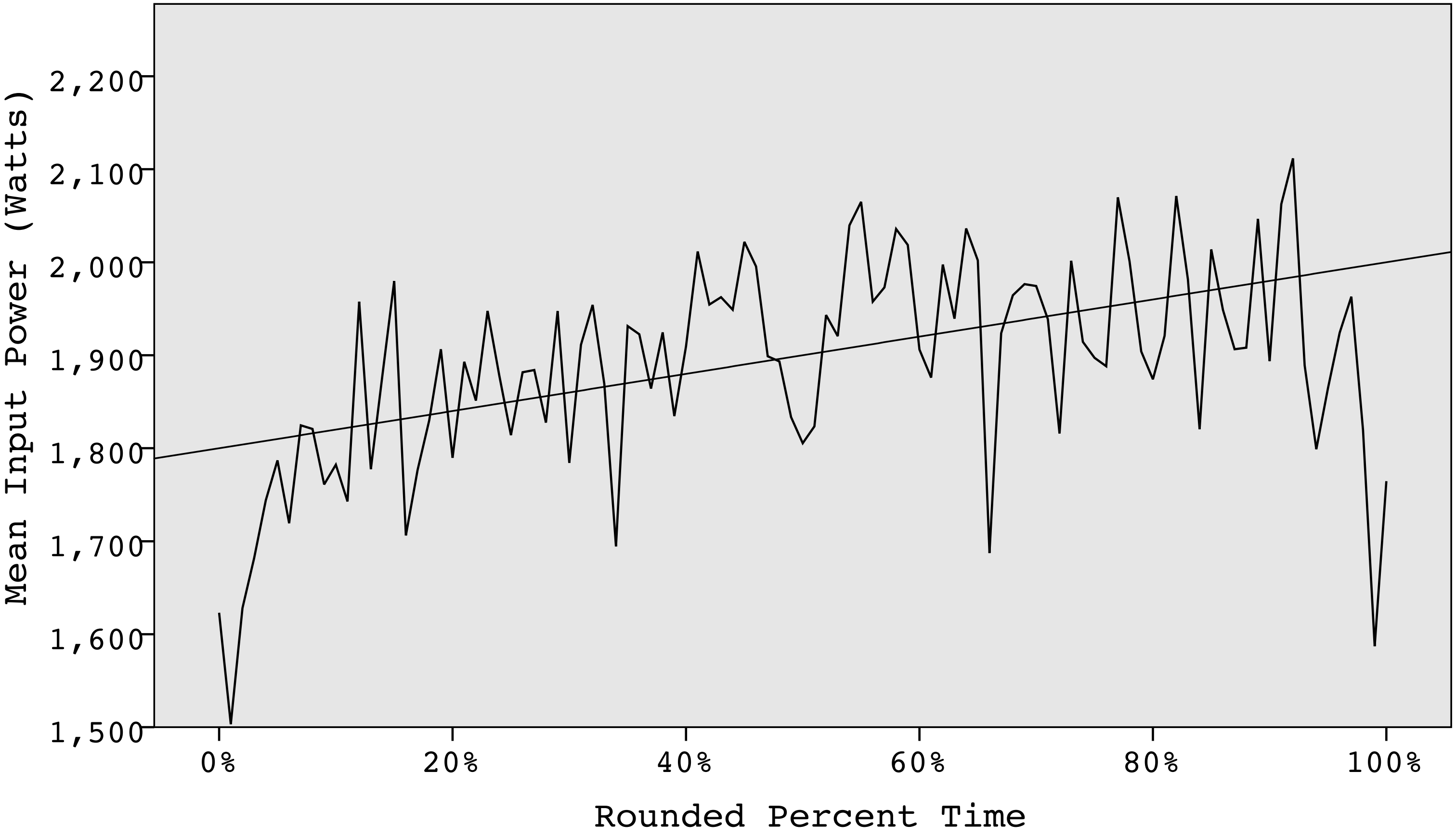
- Most jobs fall in 65 to 74 kW per rack bins. Very few jobs above 80 kW per rack.
- Large jobs (at or above 24 racks) tend to be in 52 kW bin.



Environmental Power

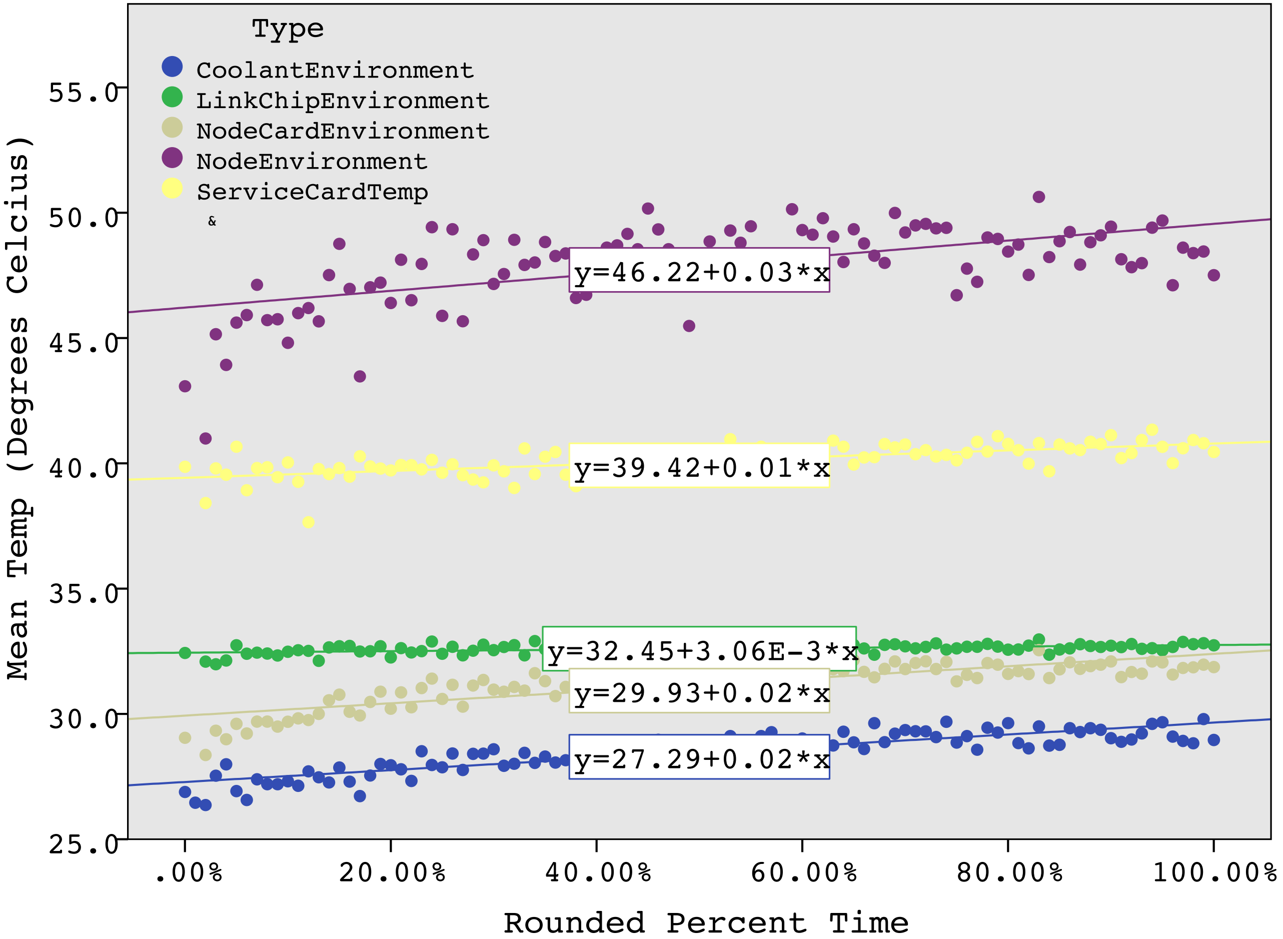


Power as a Function of Time



Environmental Temperature

- System gets “hotter” as jobs run. Most sensors indicate 2 to 3 degree increase.



EMON

- Environmental Monitoring (EMON) API that allows one to access power consumption data from code running on compute nodes at sub-second intervals.
- API by itself only returns total power consumption of all domains and does not contain any profiling functionality.
  - Thus, we developed MonEQ which allows us to read individual voltage and current data points.
- Not without faults:
  - Power information obtained is *total* power consumption from oldest generation of data.
  - Measurements not taken at precisely the same moment.
    - May result in inconsistent results in certain cases (such as when a piece of code stresses both the CPU and memory at the same time).
  - However, active research by IBM on these problems, so they might disappear entirely in a future software update.

Sample Output

Environmental Database

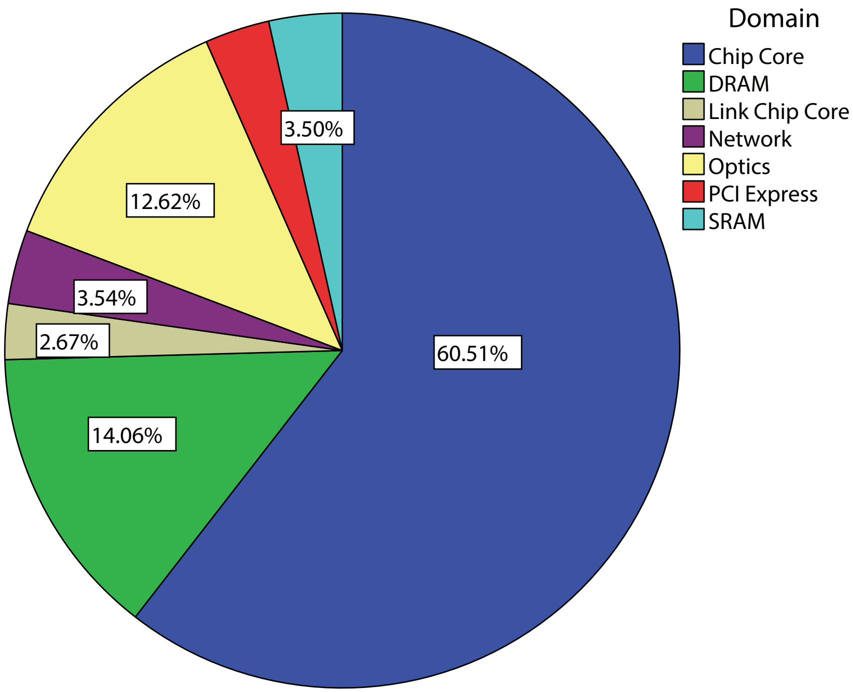
```
                                Location, Time, Input_Voltage, Input_Current, Output_Voltage, Input_Current
"Q1G-B-P0  ","2012-09-01-00.04.39.872955",+2.75906000000000E+002,+2.96900000000000E+000,+5.09180000000000E+001,+1.48120000000000E+001
"Q1G-B-P1  ","2012-09-01-00.04.39.873978",+2.76469000000000E+002,+2.96900000000000E+000,+5.09180000000000E+001,+1.47030000000000E+001
"Q1G-B-P2  ","2012-09-01-00.04.39.874556",+2.76875000000000E+002,+2.93800000000000E+000,+5.08980000000000E+001,+1.45780000000000E+001
"Q1G-B-P3  ","2012-09-01-00.04.39.875130",+2.75906000000000E+002,+3.17200000000000E+000,+5.09020000000000E+001,+1.58440000000000E+001
"Q1G-B-P4  ","2012-09-01-00.04.39.875765",+2.77344000000000E+002,+2.95300000000000E+000,+5.09380000000000E+001,+1.45780000000000E+001
"Q1G-B-P5  ","2012-09-01-00.04.39.876315",+2.78312000000000E+002,+2.95300000000000E+000,+5.08710000000000E+001,+1.44530000000000E+001
"Q1H-B-P0  ","2012-09-01-00.04.39.876965",+2.76500000000000E+002,+2.98400000000000E+000,+5.09100000000000E+001,+1.46250000000000E+001
"Q1H-B-P1  ","2012-09-01-00.04.39.878215",+2.75969000000000E+002,+2.98400000000000E+000,+5.09920000000000E+001,+1.45620000000000E+001
"Q1H-B-P2  ","2012-09-01-00.04.39.878872",+2.76281000000000E+002,+3.03100000000000E+000,+5.09450000000000E+001,+1.48120000000000E+001
"Q1H-B-P3  ","2012-09-01-00.04.39.879502",+2.77188000000000E+002,+2.95300000000000E+000,+5.09800000000000E+001,+1.44220000000000E+001
```

MonEQ

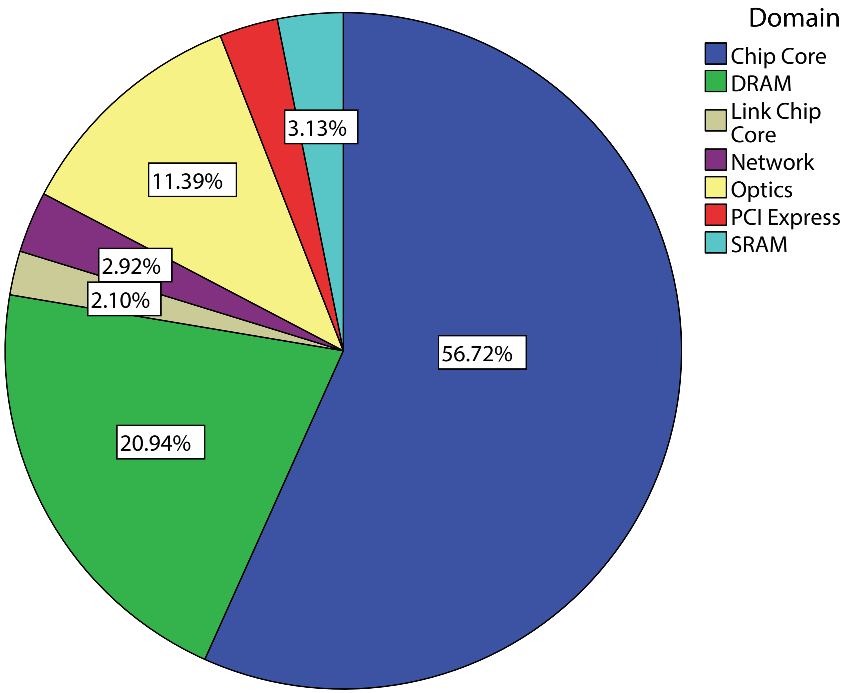
```
date_time, time_since_epoch, ticks, row, col, midplane, nodeboard, node_card_power, chip_core, dram, network, sram, optics, PCIexpress, link_chip_core
       Tue Dec 18 20:25:58 2012, 1355862358, 264569376610, 0, 0, 4, 1890.0816, 1042.8309, 435.4042,48.8761, 57.4618, 212.0369, 49.3437, 44.1282
      Tue Dec 18 20:25:58 2012, 1355862358, 273529419810, 0, 0, 4, 2259.0636, 1320.7986, 526.2582, 48.8679, 57.0442, 212.3337, 49.3437, 44.4173
      Tue Dec 18 20:25:59 2012, 1355862359, 282489320530, 0, 0, 4, 2235.8694, 1305.2437, 518.7294, 49.6738, 56.9985, 211.4175, 49.3437, 44.4629
      Tue Dec 18 20:25:59 2012, 1355862359, 291449678400, 0, 0, 4, 2230.2651, 1301.7340, 516.8031, 49.7253, 56.9922, 210.5338, 49.3437, 45.1330
      Tue Dec 18 20:26:00 2012, 1355862360, 300409284050, 0, 0, 4, 2238.8619, 1317.3014, 509.1803, 49.7253, 56.6683, 212.0369, 48.5413, 45.4084
```

Domain Breakdown

MPI Bisection

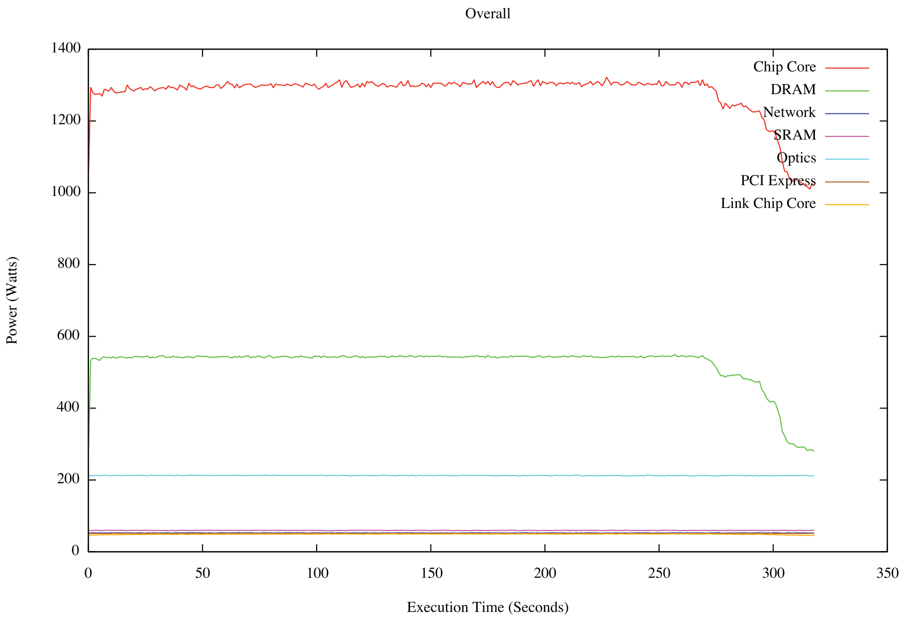


XSBench

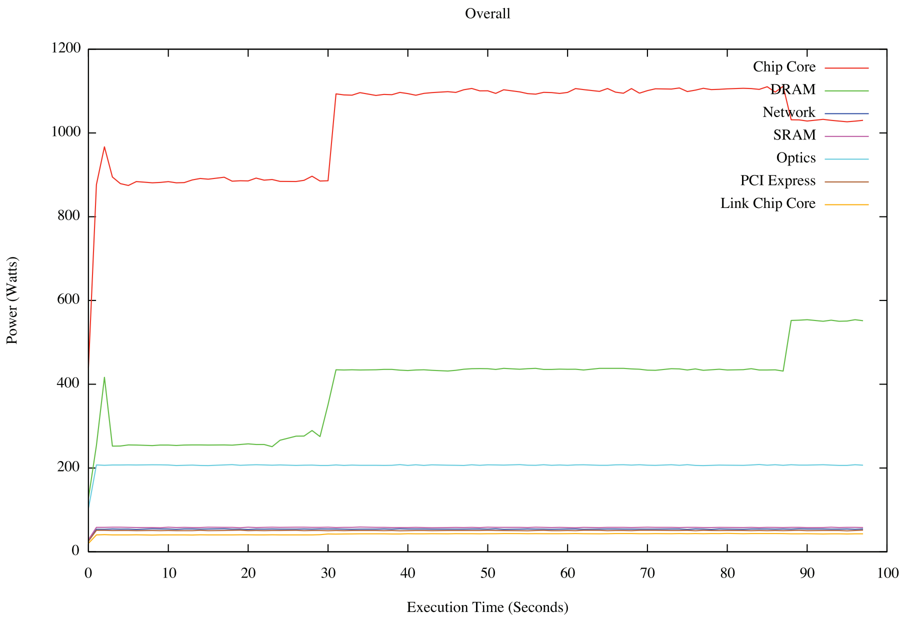


Domain Breakdown

DGEMM



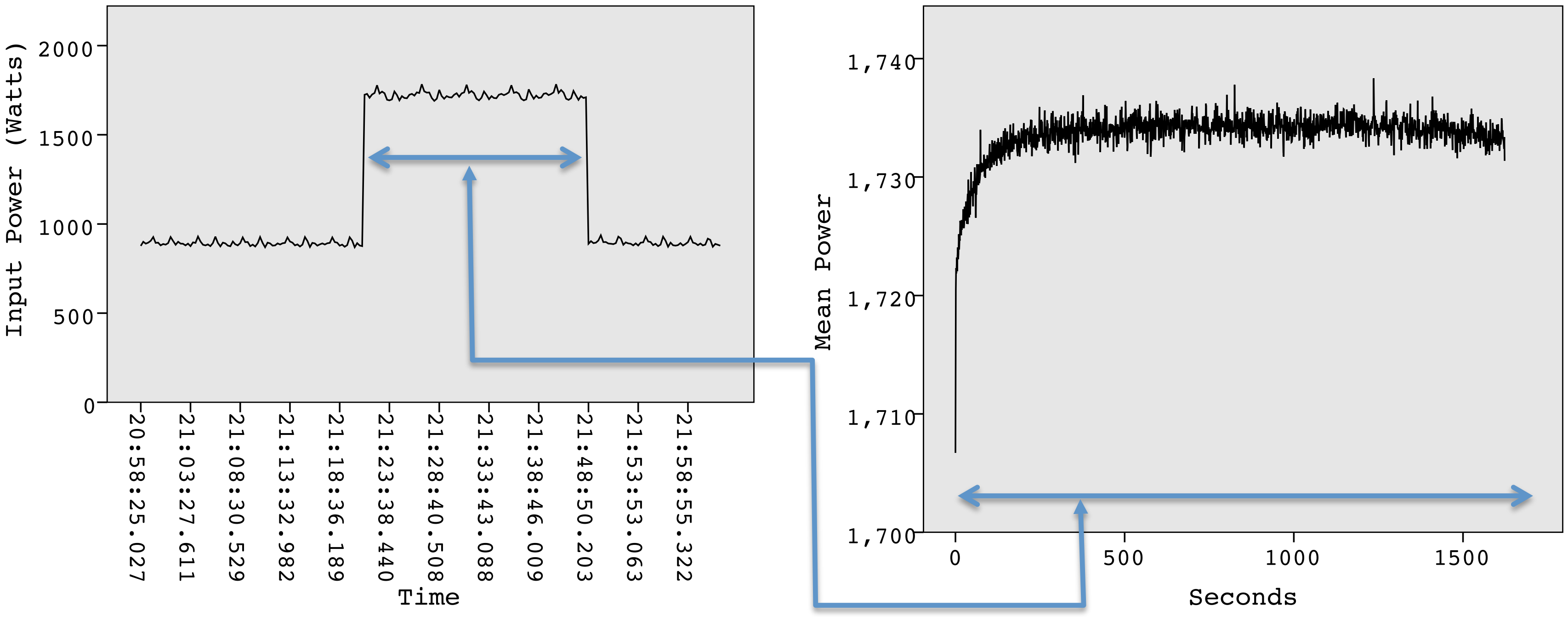
XSBench



Environmental Data VS. EMON Data

Environmental

EMON



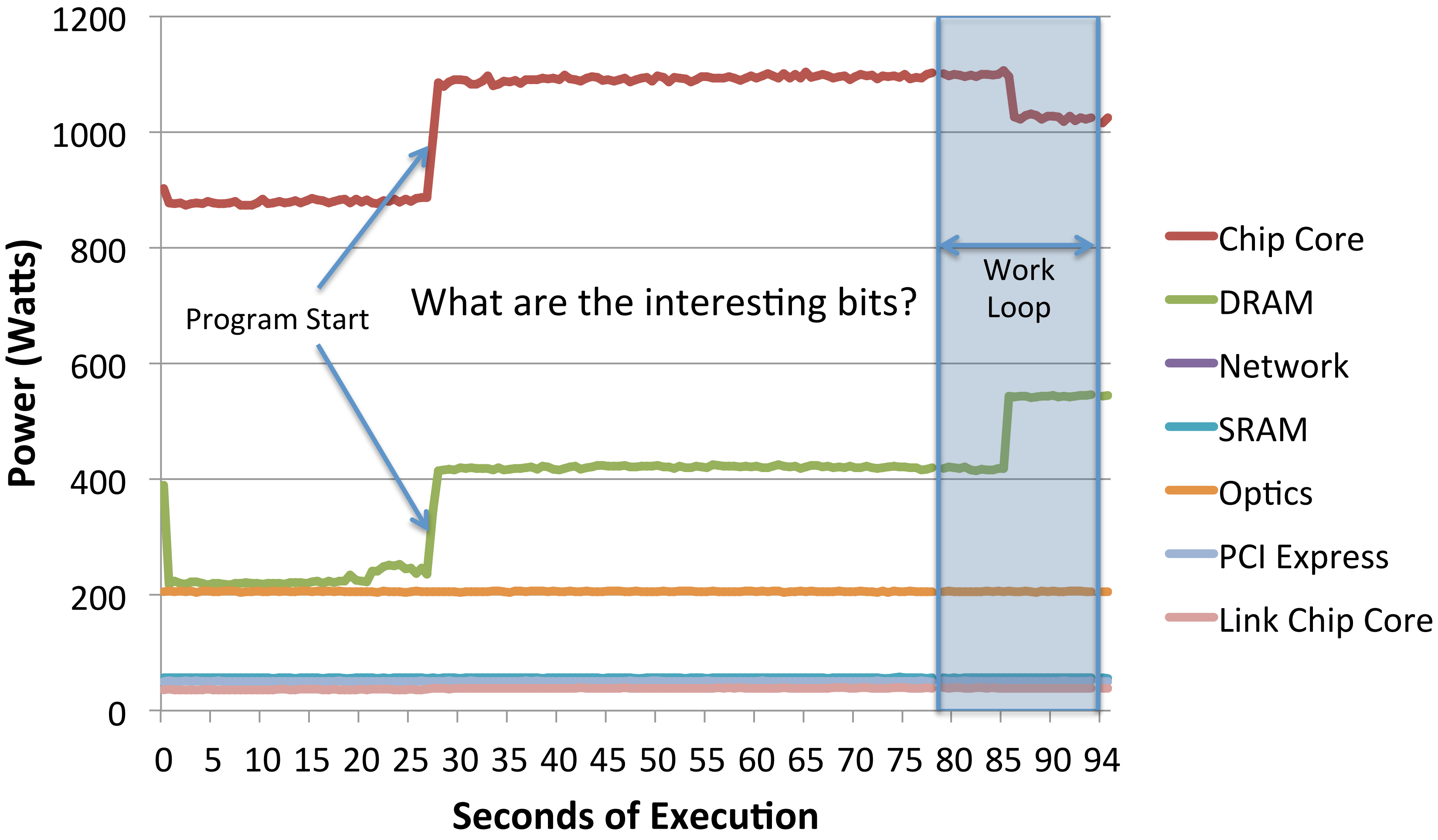
Simple MonEQ Example

```
int status, myrank, numtasks, itr;

status = MPI_Init(&argc, &argv);

MPI_Comm_size(MPI_COMM_WORLD, &numtasks);
MPI_Comm_rank(MPI_COMM_WORLD, &myrank);

/* Setup Power */
status = MonEQ_Initialize();


/* User code */


/* Finalize Power */
status = MonEQ_Finalize();

MPI_Finalize();
```

More Complex MonEQ Example

```
const int buf_size = (1024 * 1024);
/* Setup Power      */
MonEQ_DisableAutoCollection ();
status = MonEQ_Initialize();

/*     User Code */

/*     Report the Current Power */

/*     -------------Create The Array    -------------- */
arr = (int*) malloc ( sizeof(int) * buf_size);
if (0 == arr) {
       printf("Error allocating Array \n");
       fflush(stdout);
}
memset(arr, 0, buf_size * sizeof(int));

/*     -------------Populate the Array  -------------- */
for (itr = 0; itr < buf_size; itr++) {
       arr[itr] = 7 +  itr;
}

if (0 == myrank){
       MonEQ_PrintDomainInfo ();
       MonEQ_PrintVTMRatio ();
}

if (MonEQ_MonitorAgentOnRank()) {
       tm1 = GetTimeBase();
       power =  MonEQ_GetPower();
       tm2 = GetTimeBase();

       lat = ((double)tm2 - (double) tm1) / 1600e6;

       printf (" Power is %f w, call latency: %f sec \n", power, lat);
}

/* Finalize Power   */
status = MonEQ_Finalize();
```

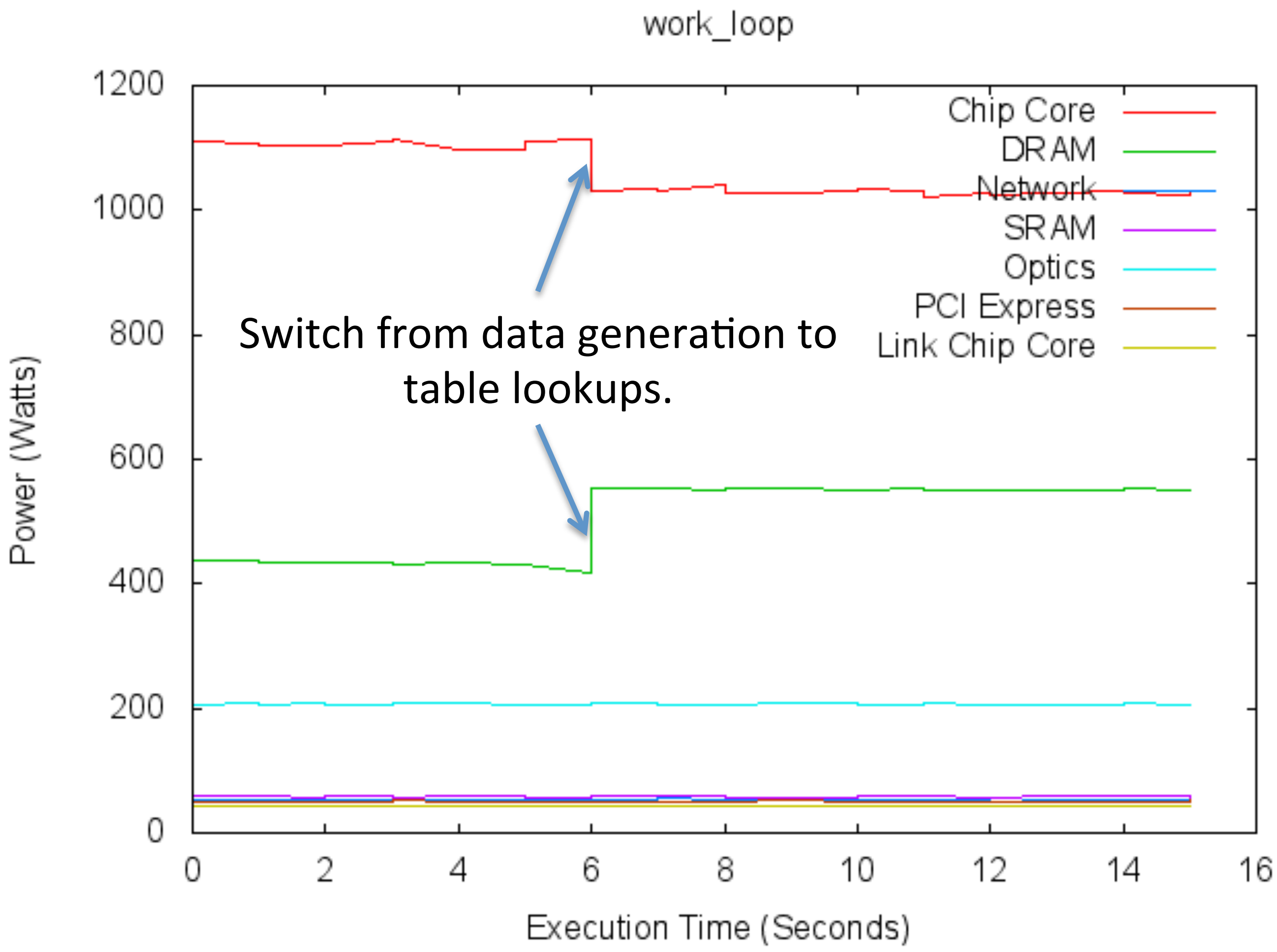
Domain Profiling Results



Tagging

```
/* Initialize MonEQ power monitoring */
status = MonEQ_Initialize();

/* Add tag */
MonEQ_StartPowerTag("for_loop");

for (i = 0; i <= ...) {
...
}

MonEQ_EndPowerTag("for_loop");

...

/* Finalize MonEQ power collection */
status = MonEQ_Finalize();
```

Tagging Results



Switch from data generation to table lookups.

Conclusion

- Evaluated existing power monitoring capabilities of an IBM Blue Gene/Q system.
  - While designed for environmental monitoring, also very useful for profiling applications at course grain.
- MonEQ, which utilizes EMON API, reports same data as in environmental database but at sub-second intervals across several domains.
  - Unlike environmental data, accessible to end users!
- Looking forward:
  - Much more profiling of benchmarks and applications.
  - Power aware scheduling?

Acknowledgements

- This research has been funded in part and used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357. The research has been funded in part by the Center for Exascale Simulation of Advanced Reactors (CESAR) Co-Design center, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

- The authors would also like to thank Paul Coteus and Christopher M. Marroquin from IBM for their help in clarifying results as well as providing essential information of the inner workings of the system.

QUESTIONS?