POW: System-wide Dynamic Reallocation of Limited Power in HPC

Daniel Ellsworth[1], Allen Malony[1], Barry Rountree[2], Martin Schulz[2]

UNIVERSITY OF OREGON [1]

Lawrence Livermore National Laboratory [2]

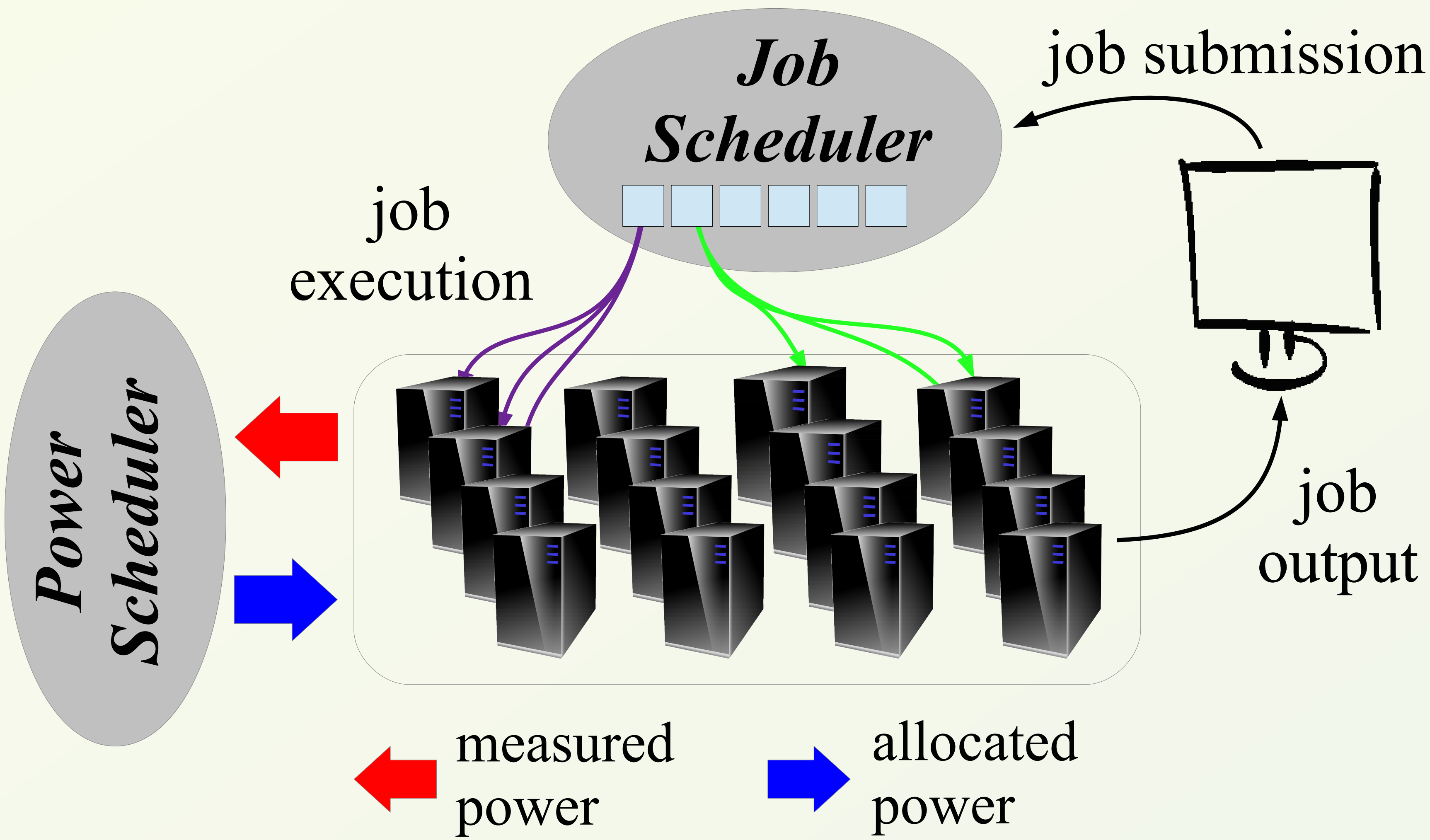
Power and Energy

- Different but related ideas
  - Rate vs Quantity
- Conversion:

$$1 \text{ Watt} = 1 \frac{Joule}{Second}$$

- 1 kWh = 3.6 megajoules
- Infrastructure required for 900 kWh over 1 hour is not the same as 900 kWh over 720 hours.

HPC System


Job Scheduler
job submission
job execution
Power Scheduler
job output
measured power
allocated power

Power Scheduler Invariant

$$\forall t, \sum c_i^t \leq \sum a_i^t \leq L$$

L	System-wide power limit
n	Number of sockets
t	A time
c_i^t	Power consumed by socket i at time t
a_i^t	Power allocated to socket i at time t

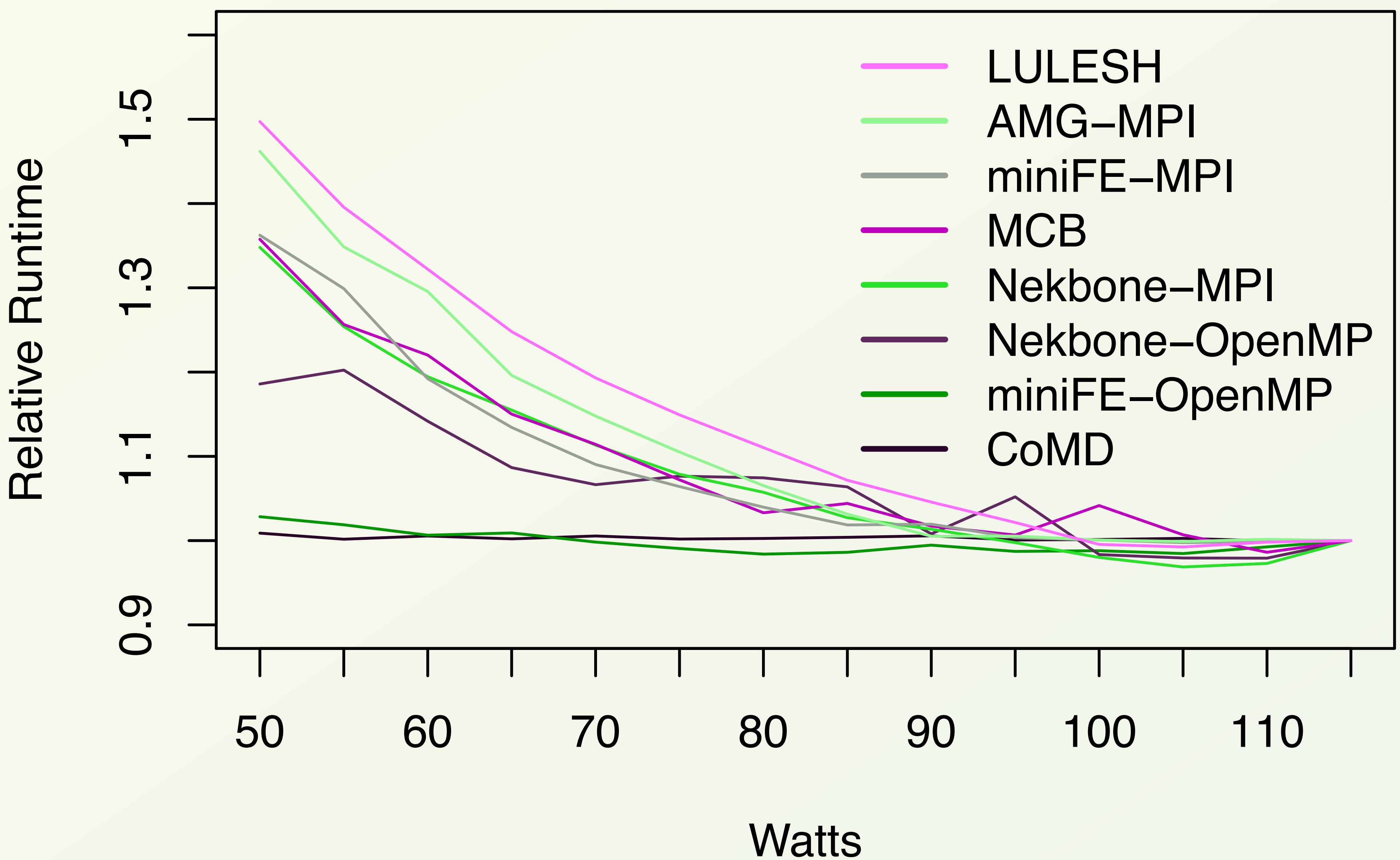
Naive Static Strategy

$$\forall t, \sum c_i^t \leq \sum a_i^t \leq L$$

L	System-wide power limit
n	Number of sockets
t	A time
c_i^t	Power consumed by socket i at time t
a_i^t	Power allocated to socket i at time t

$$a_i^t = \frac{L}{n} \implies \sum c_i^t \leq L$$

Power and Runtime



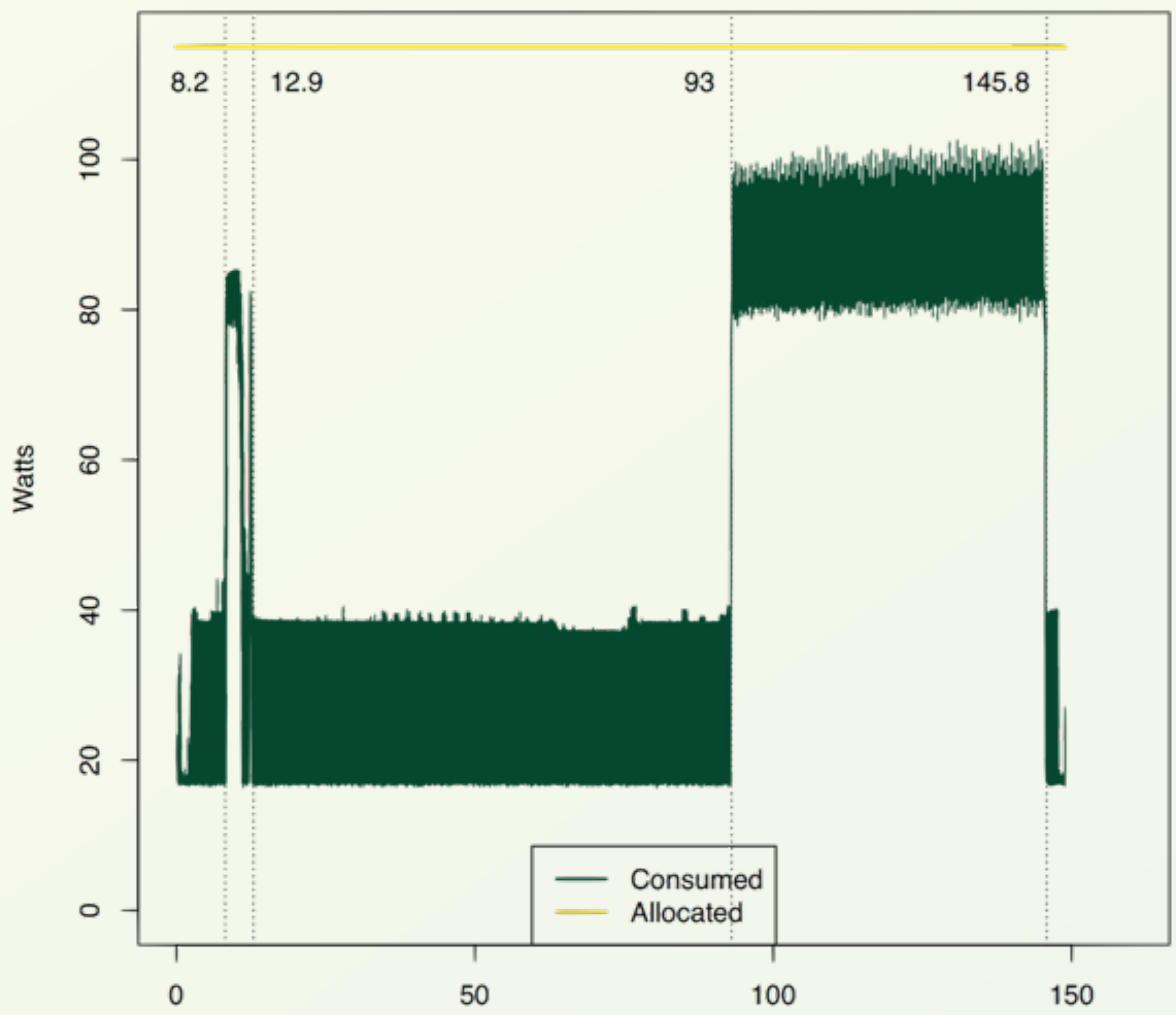
Job Static Strategy

$$\forall t, \sum c_i^t \leq \sum a_i^t \leq L$$

L	System-wide power limit
n	Number of sockets
t	A time
j_+	Maximum power consumed by a socket for job j
j_n	Number of sockets in job j
c_i^t	Power consumed by socket i at time t
a_i^t	Power allocated to socket i at time t

$$\forall j, \sum j_+ j_n \leq L \implies \sum c_i^t \leq L$$

Power and Energy



Naive Dynamic Strategy

$$\forall t, \sum c_i^t \leq \sum a_i^t \leq L$$

L	System-wide power limit
n	Number of sockets
t	A time
w_i^t	Waste power for socket i at time t
c_i^t	Power consumed by socket i at time t
a_i^t	Power allocated to socket i at time t

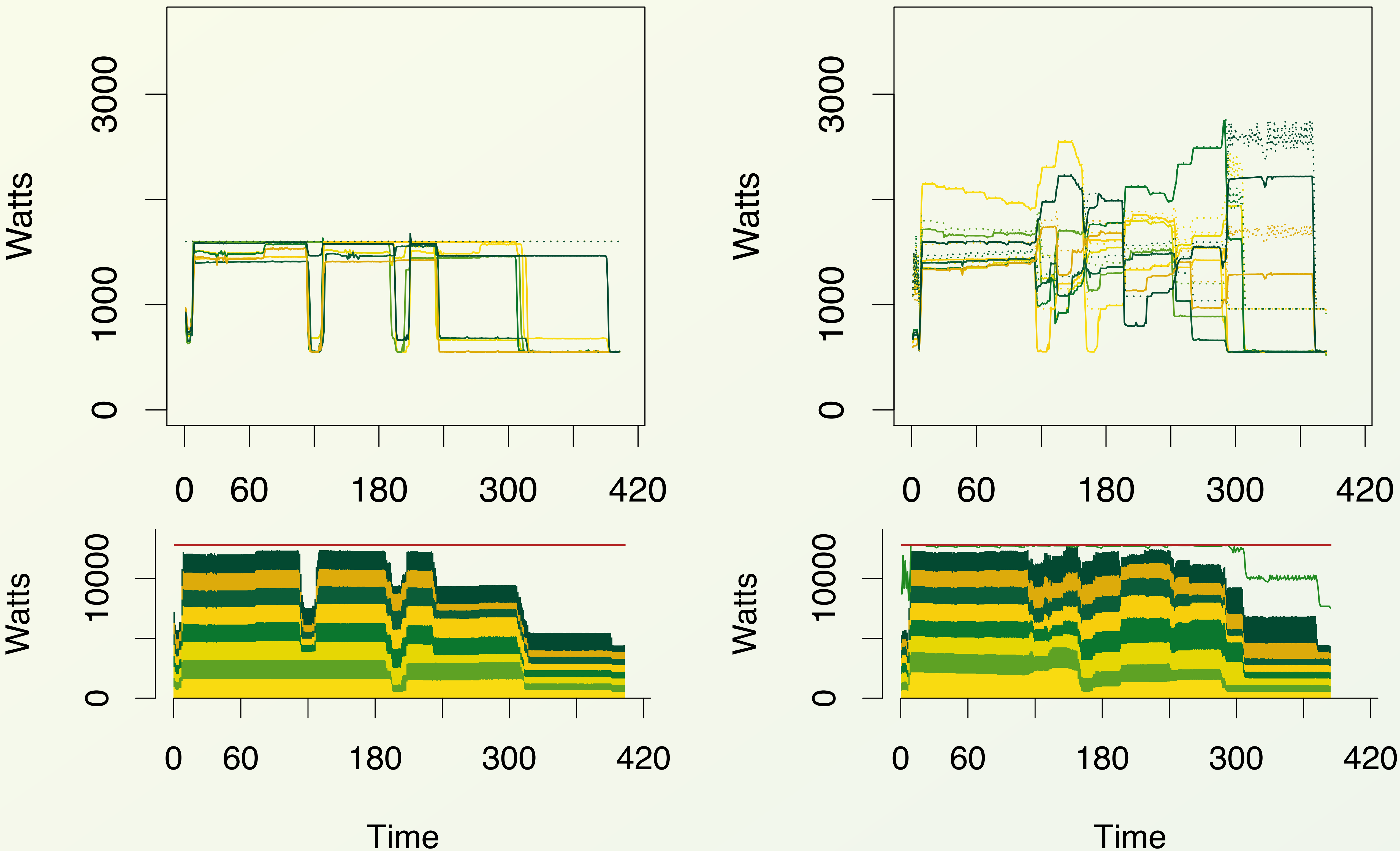
$$\forall t, L = \sum a_i^t \qquad c_i^t + w_i^t = a_i^t \qquad c_i^t \approx c_i^{t+1}$$

$$w_i^{t+1} \approx \frac{1}{n} \sum a_i^t - c_i^t \implies a_i^{t+1} \approx c_i^t + w_i^{t+1}$$

POWsched

```
procedure MAIN
    while True do
        GETREADINGS                                   ▷ Phase 1
        ALLOCDOWN                                     ▷ Phase 2
        ALLOCUP                                       ▷ Phase 3
        sleep rest of interval
    end while
end procedure
```

50W Static and Dynamic



Static vs Dynamic

Experiment	Runtime	Stddev	Improvement
115W static	278.26	9.57	
115W dynamic	276.24	4.84	0.7%
90W static	284.63	3.20	
90W dynamic	277.13	5.04	2.6%
70W static	323.83	4.90	
70W dynamic	278.02	4.97	14.1%
50W static	407.21	18.00	
50W dynamic	371.92	13.23	8.7%

In Summary

- Power Optimization != Power Bound Enforcement
- Static power allocation may not be optimal
- Dynamic power reallocation can reduce time to solution

Funding Acknowledgement

- Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-669276).

- Work by the University of Oregon is supported by the DOE Office of Science, through a Sub-Contract No. 3F-32643 from the University of Chicago, Argonne, LLC (as operator of Argonne National Laboratory), under Prime Contract No. DE-AC02-06CH11357.