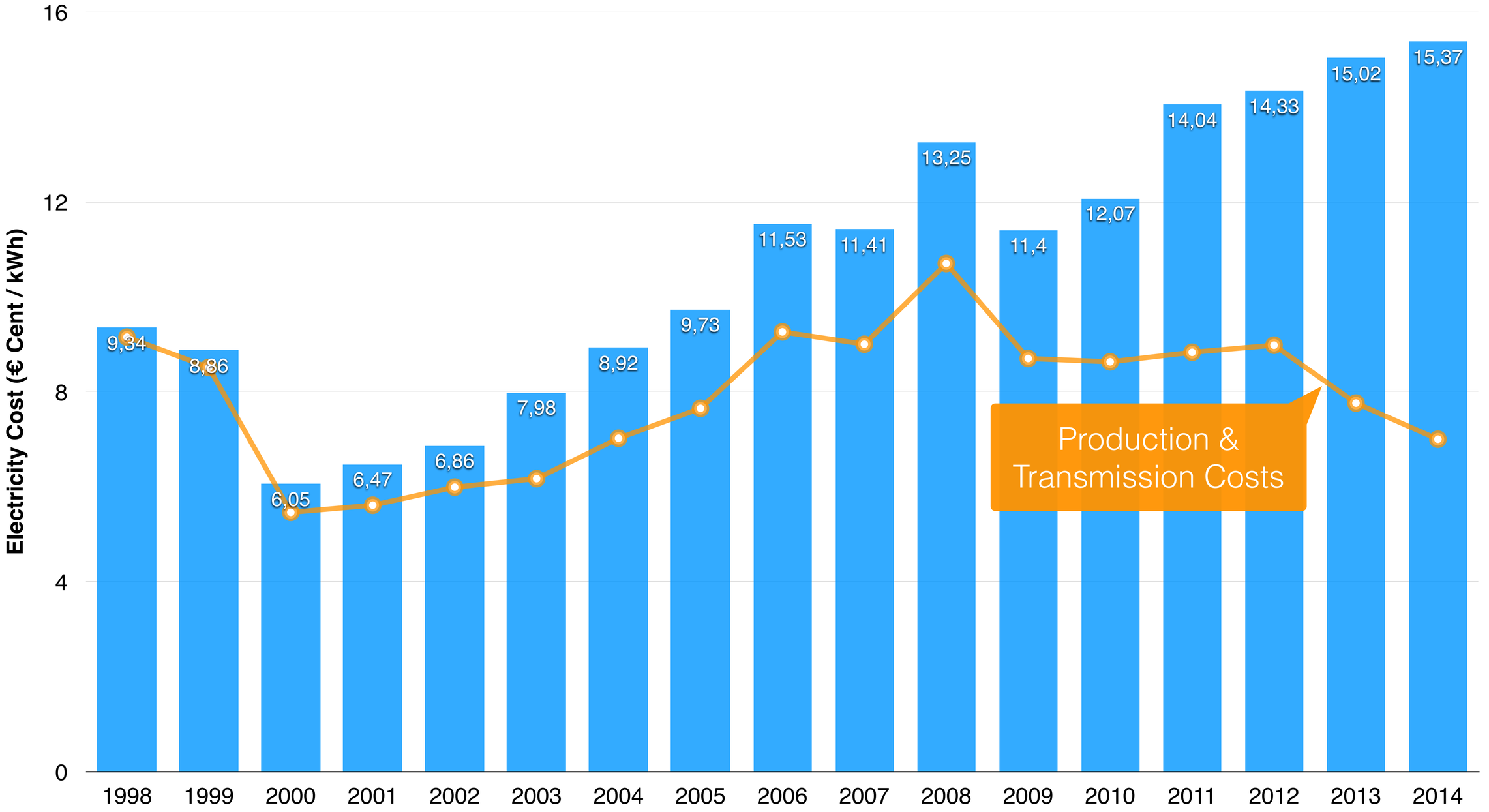
Energy Aware Scheduling @ LRZ

COLA Workshop – Ferrara – 25.02.2016

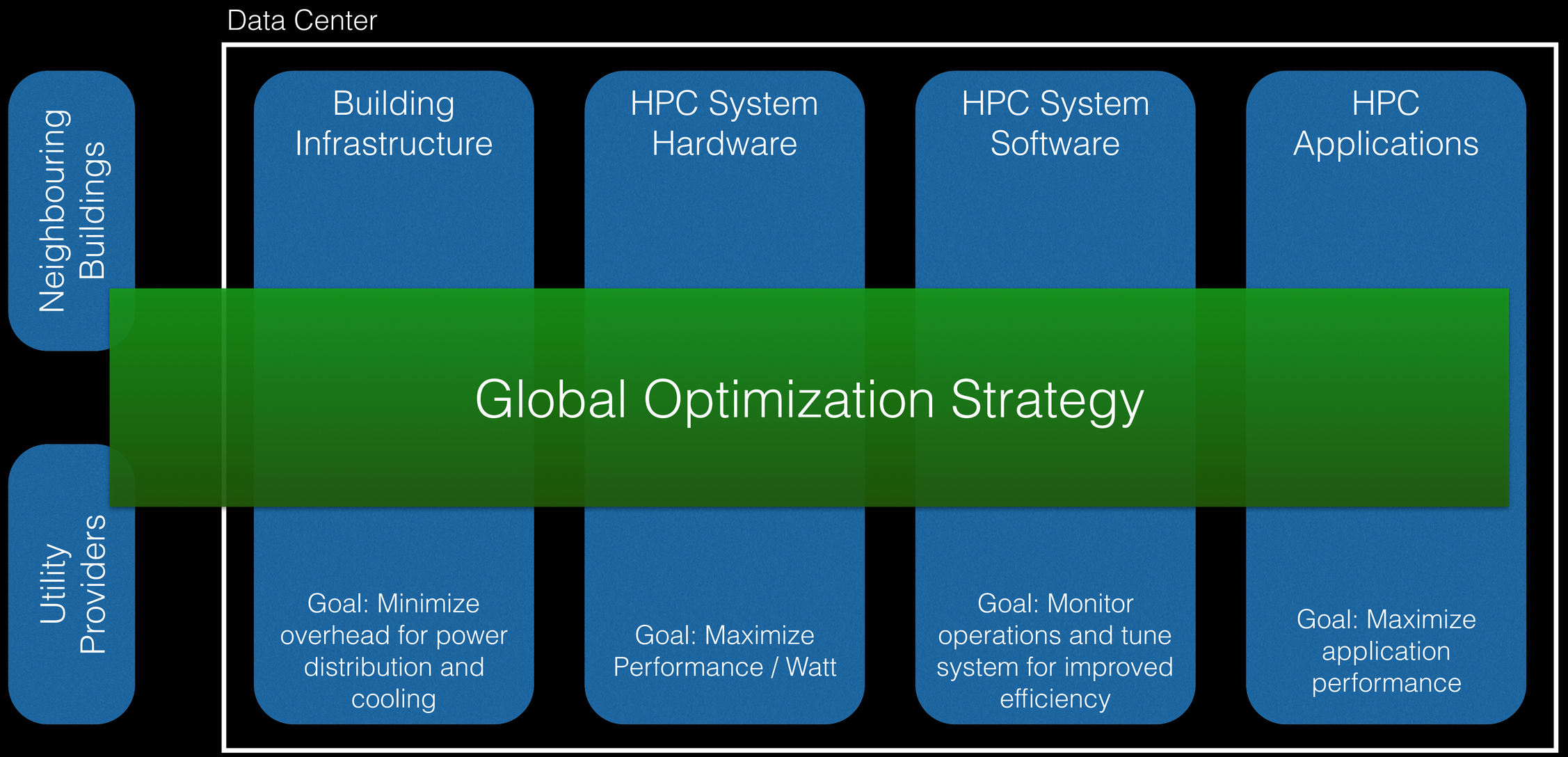
Daniele Tafani
(based on the slides by Axel Auweter)

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities



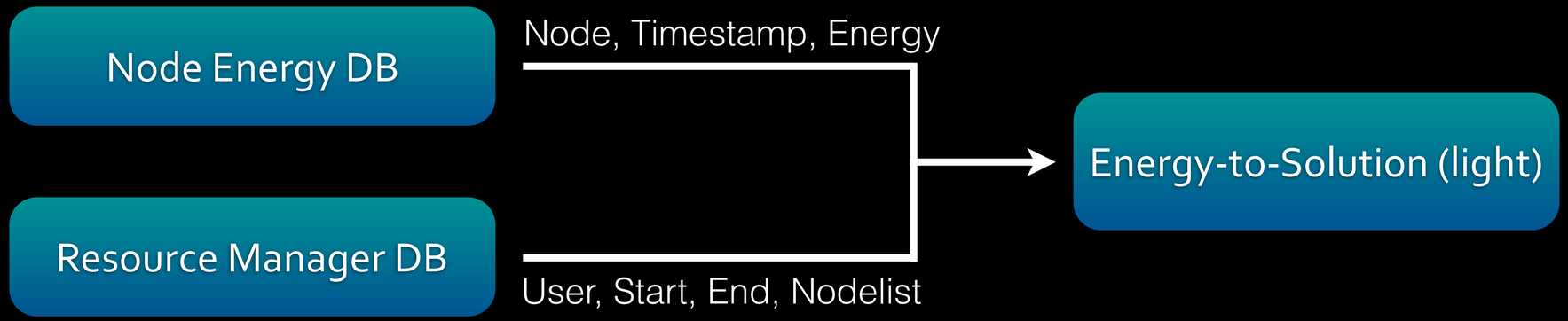
Average electricity costs for German industrial customers consuming up to 20.000 MWh / year.
Source: BDEW Strompreisanalyse 2014

4 Pillars for Energy Efficient HPC



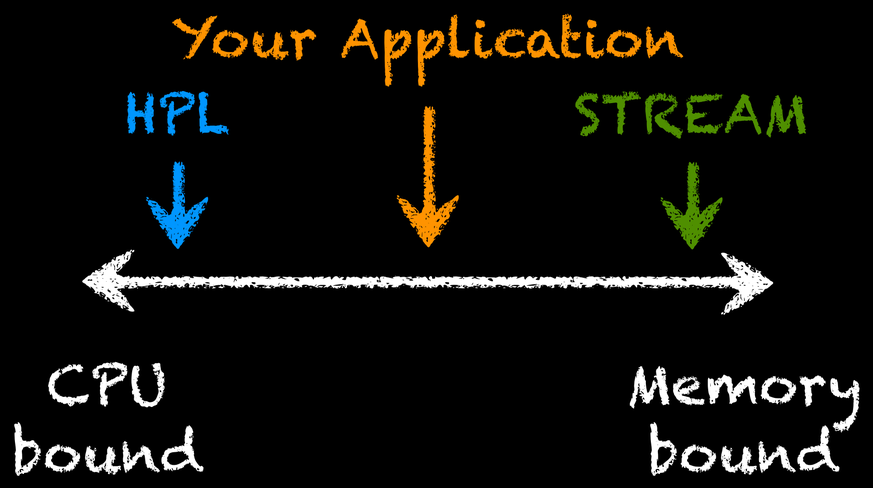
Energy-Aware Scheduling

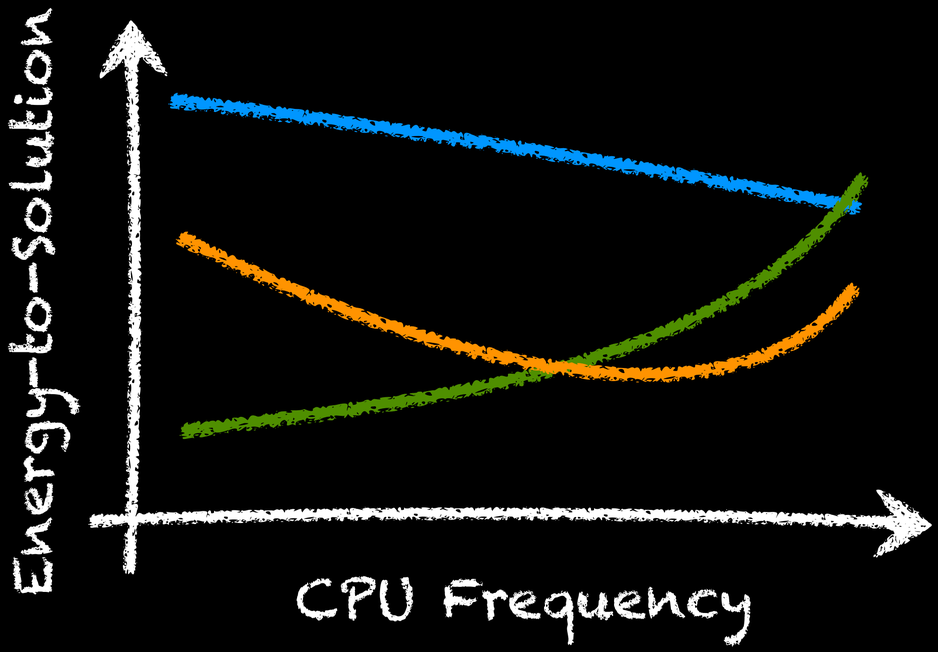
- Energy-to-Solution



Energy-Aware Scheduling

- Energy-aware Frequency Scaling

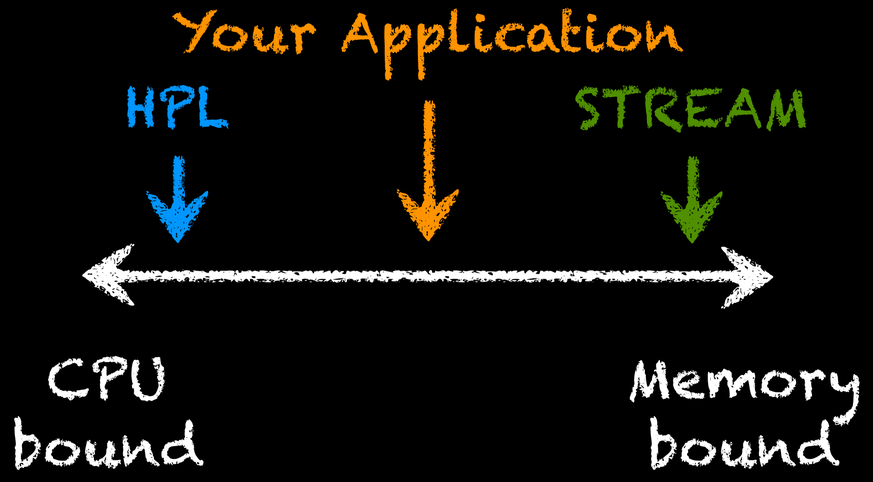


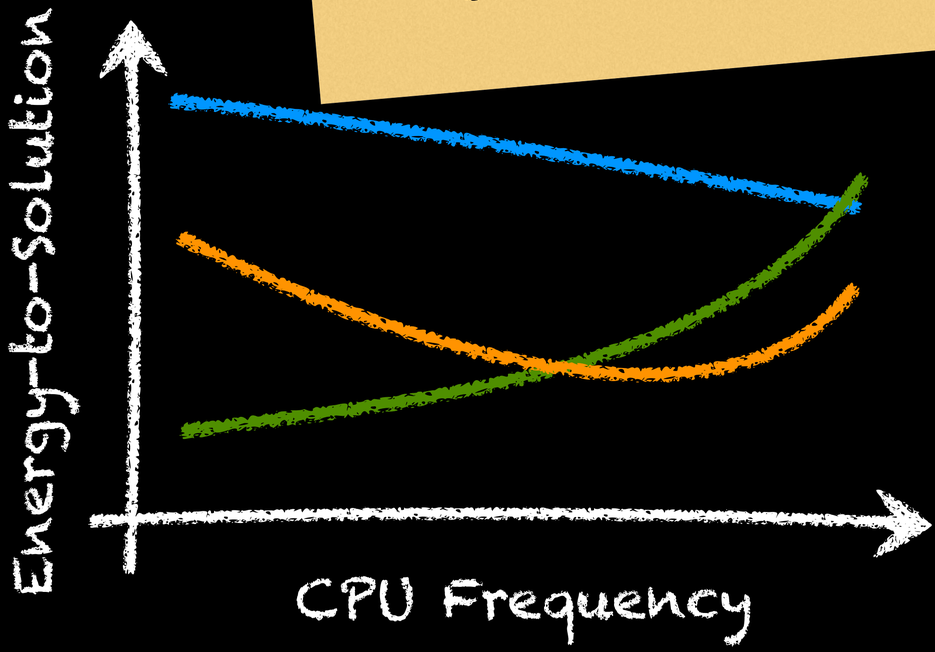


Energy-Aware Sch

- Energy-aware Frequency Scaling

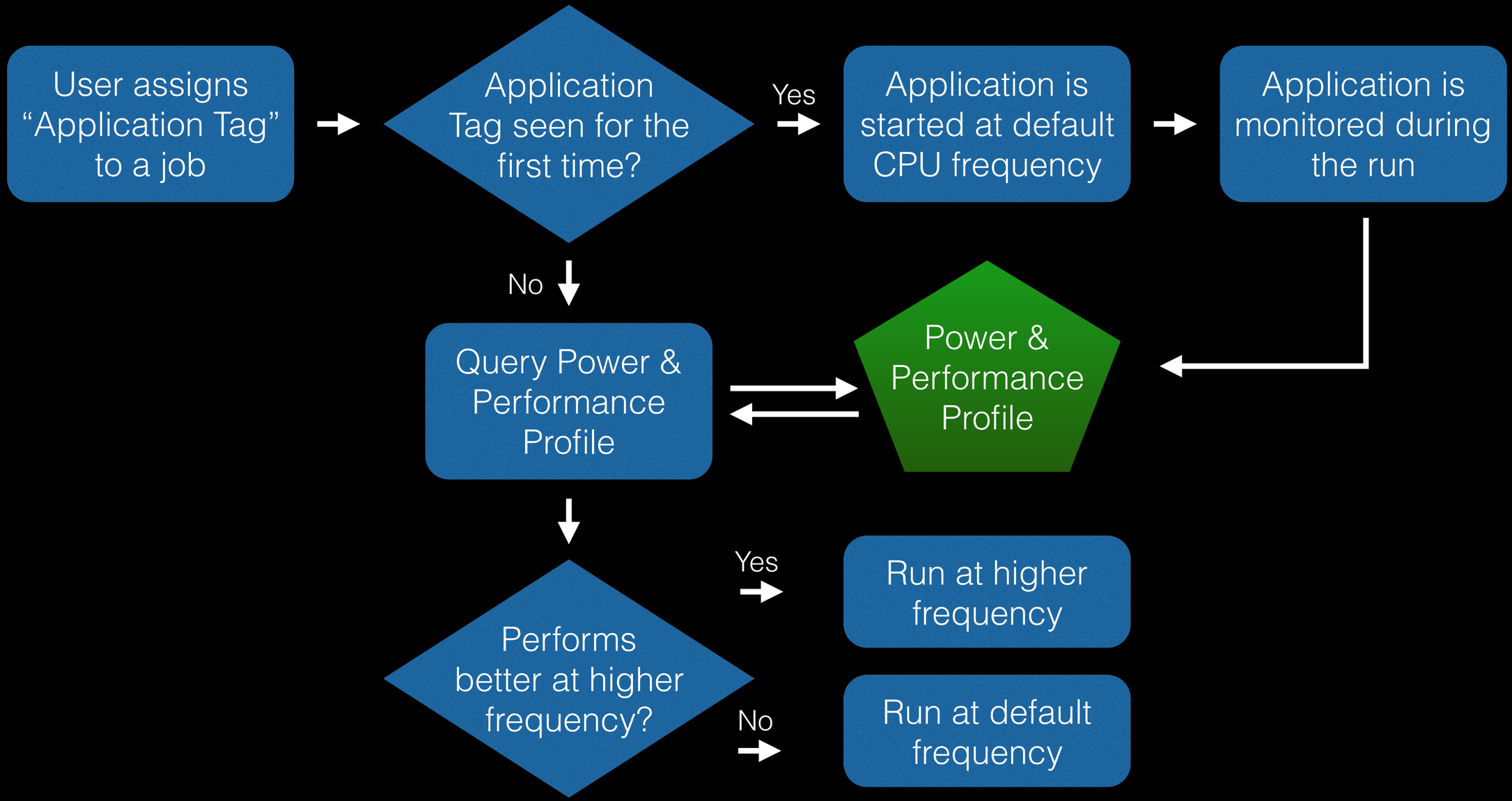
Many applications will not see a performance benefit from running at the highest CPU frequency!



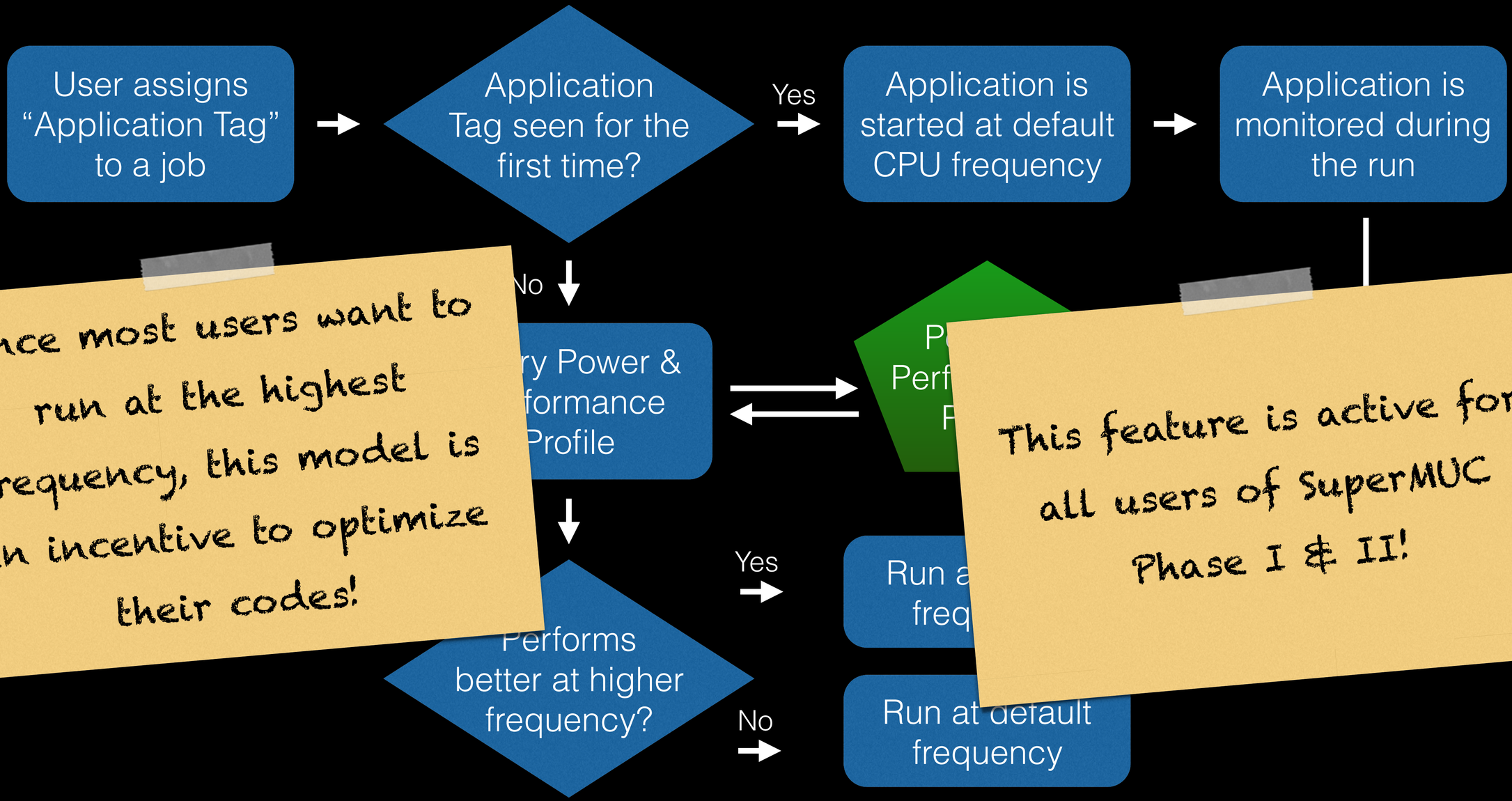


Disclaimer: heavily simplified for illustrative purpose - no real data…

LoadLeveler Implementation



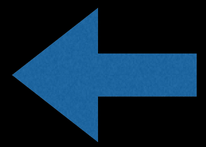
LoadLeveler Implementation



Power-Performance Profile

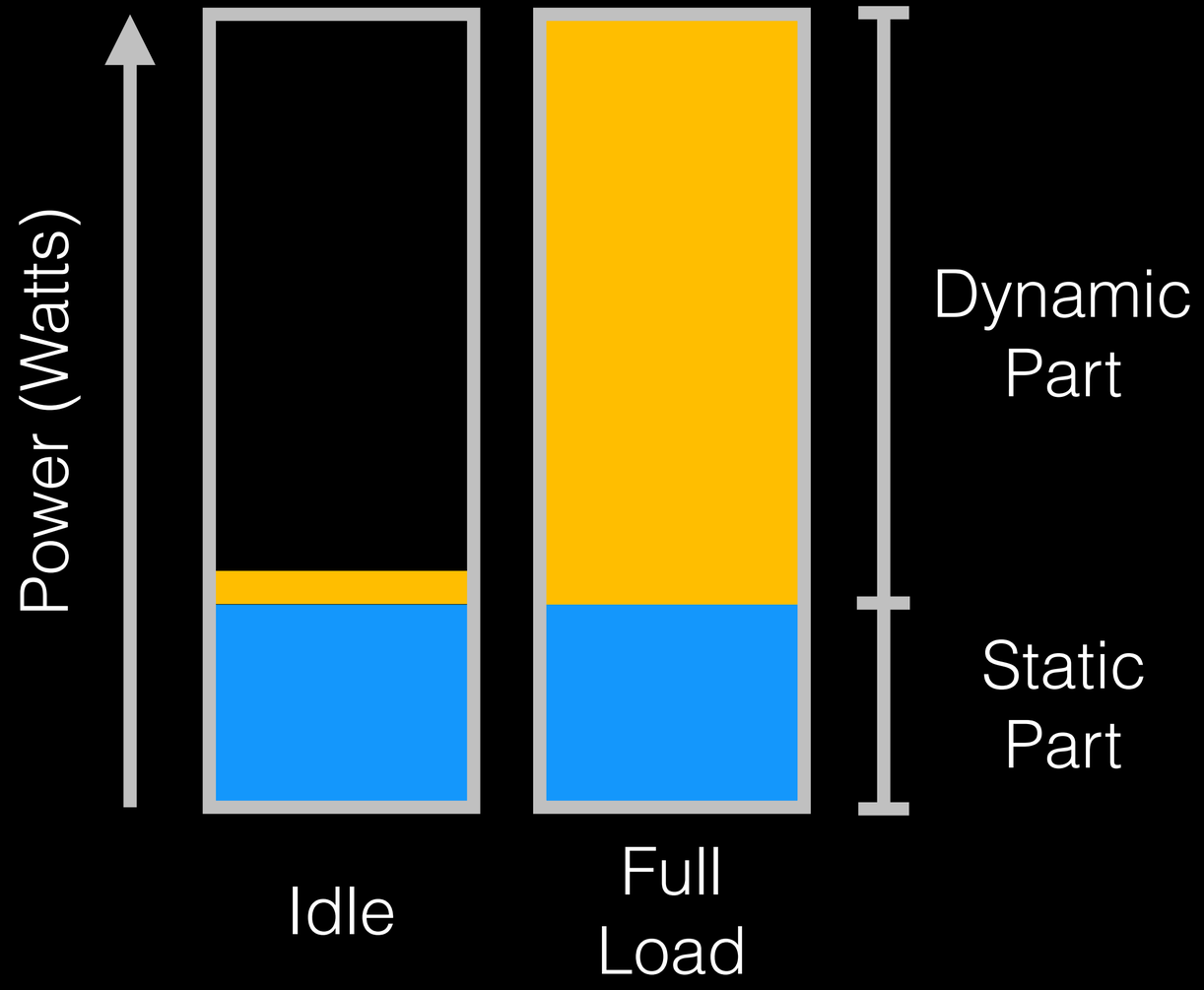
```
            Tag Name: qespresso1
        Generated by: srv03-ib.354607.0
      Last used Time: Thu Apr 17 11:11:31 2014
                User: lu32joj
   Nominal Frequency: 2,70 GHz
   Default Frequency: 2,30 GHz
Node's DC Energy Use: 0,018040 kWh
      Execution Time: 458 Seconds
```

Frequency(GHz)	EstDCEngCons(kWh)	DCEngVar(%)	EstTime(Sec)	TimeVar(%)	Power(W)
2,70	0,019446	7,80	400	-12,66	175,02
2,60	0,019000	5,32	412	-10,04	166,02
2,50	0,018628	3,26	426	-6,99	157,42
2,40	0,018283	1,35	440	-3,93	149,59
2,30	0,018040	0,00	458	0,00	141,80
2,20	0,017779	-1,44	474	3,49	135,03
2,10	0,017605	-2,41	494	7,86	128,30
2,00	0,017516	-2,90	516	12,66	122,21
1,90	0,017675	-2,02	548	19,65	116,11
1,80	0,017763	-1,53	578	26,20	110,64
1,70	0,017887	-0,85	611	33,41	105,39
1,60	0,018104	0,36	649	41,70	100,42
1,50	0,018426	2,14	693	51,31	95,72
1,40	0,018664	3,46	737	60,92	91,17
1,30	0,019064	5,68	789	72,27	86,98
1,20	0,019684	9,12	851	85,81	83,27



Reference run
@ 2.3 GHz

Modelling Power Consumption…



- Simplified Model for Compute Node Power
- Dynamic Part:
  - Processor
  - Memory
- Static Part:
  - South Bridge
  - NIC

… with Linear Regression

$$PWR(f_n) = A_n * GIPS(f_0) + B_n * GBS(f_0) + C_n$$

Predicted node power consumption at frequency n

Giga instructions per second at ref. frequency f_n

Gigabytes per second transferred at ref. frequency f_n

Platform-specific power coefficients for predicting the node power at frequency f_n

(Example)

Generating Coefficients

$$PWR(f_n) = A_n * GIPS(f_0) + B_n * GBS(f_0) + C_n$$

- Assemble a list of compute kernels with a large spectrum of GIPS and GBS characteristics
- For each kernel: measure *GIPS* and *GBS* at f_0
- For each kernel, for each CPU frequency f_n:
  - Execute kernel
  - Measure average node power $PWR(f_n)$
- Approximate A_n, B_n, C_n for each frequency f_n to satisfy the equation for all kernels

(Example)

Predicting Runtime

$$CPI(f_n) = D_n * CPI(f_0) + E_n * TPI(f_0) + F_n$$

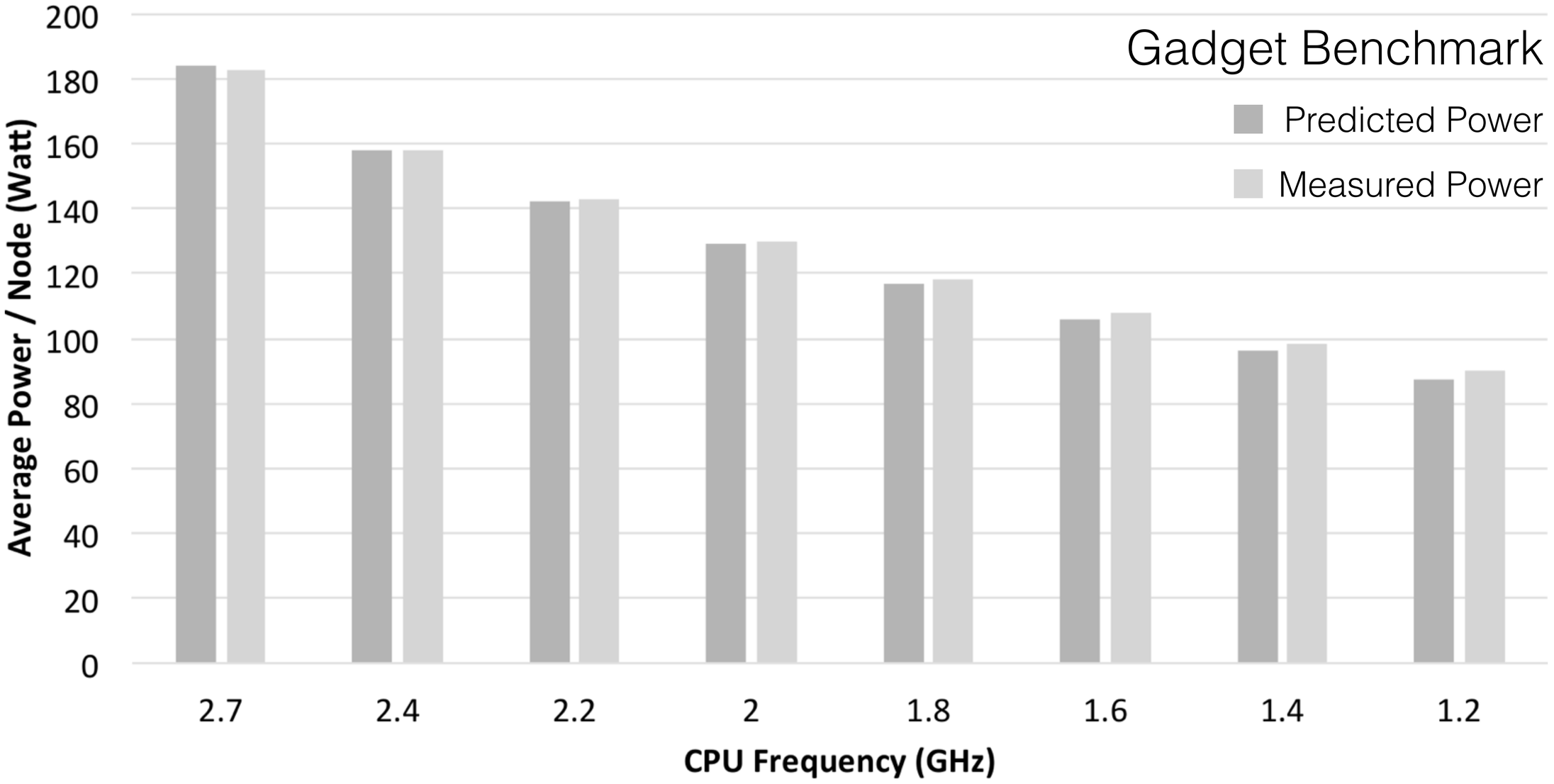
Cycles per instruction at frequency fn

Cycles per instruction at ref. frequency fo

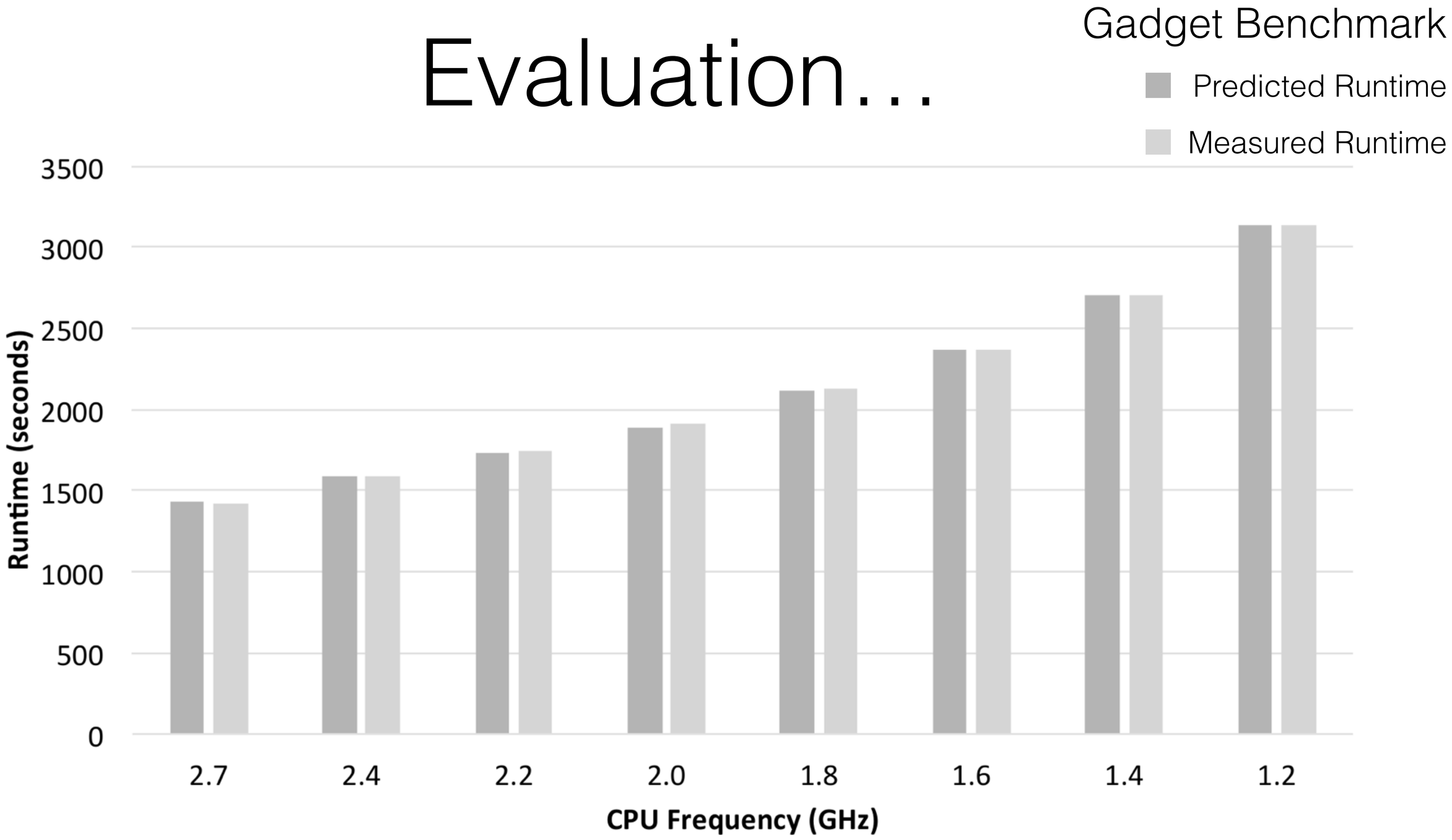
Memory transactions per instruction at ref. frequency fo

$$TIME(f_n) = TIME(f_0) * \frac{CPI(f_n)}{CPI(f_0)} * \frac{f_0}{f_n}$$

Evaluation…


Gadget Benchmark
Predicted Power
Measured Power
Average Power / Node (Watt)
200
180
160
140
120
100
80
60
40
20
0
2.7
2.4
2.2
2
1.8
1.6
1.4
1.2
CPU Frequency (GHz)

Evaluation…


Gadget Benchmark
Predicted Runtime
Measured Runtime
Runtime (seconds)
3500
3000
2500
2000
1500
1000
500
0
2.7
2.4
2.2
2.0
1.8
1.6
1.4
1.2
CPU Frequency (GHz)

Conclusion

- LRZ Policy on SuperMUC is now:
  - No application tag: run @ default frequency (2.3 GHz)
  - With application tag:
    - Execute at 2.4 GHz if performance gain > 2.5%
    - Execute at 2.5 GHz if performance gain > 5%
    - Execute at 2.6 GHz if performance gain > 8.5%
    - Execute at 2.7 GHz if performance gain > 12%
- Applies to all jobs on SuperMUC
- Estimated energy savings ~5 %
- Big incentive for scientists to improve their codes!

Thanks for the attention! Questions?