bof121s1
Maximize Data Center Power Efficiency through Energy Aware Computing

Luigi Brochard, IBM
Chong Chen, IBM
Herbert Huber, LRZ
Aaron Andersen, NCAR
Terry Hewitt, Hartree Center

SC13 – November 2013

What we want to discuss in this bof

- Energy Aware Computing is not only about reducing power consumption but also performance and energy
- We'ld like to address several ways to achieve this goal

Several paths to Energy Aware Computing

- Use New Cooling
- Reduce Power Supply Loss
- Use High Watt/Flops processors
- Use Power & Energy Aware Tools
- Tune the application

Several paths to Energy Aware Computing

- Use New Cooling
- Reduce Power Supply Loss
- Use High Watt/Flops processors
- Use Power & Energy Aware Tools
- Tune the application

Infrastructure

Server

Software

IBM Platform LSF

Energy Aware Scheduling

Chong Chen
Senior Product Architect
chong@ca.ibm.com

SC13 – November 2013

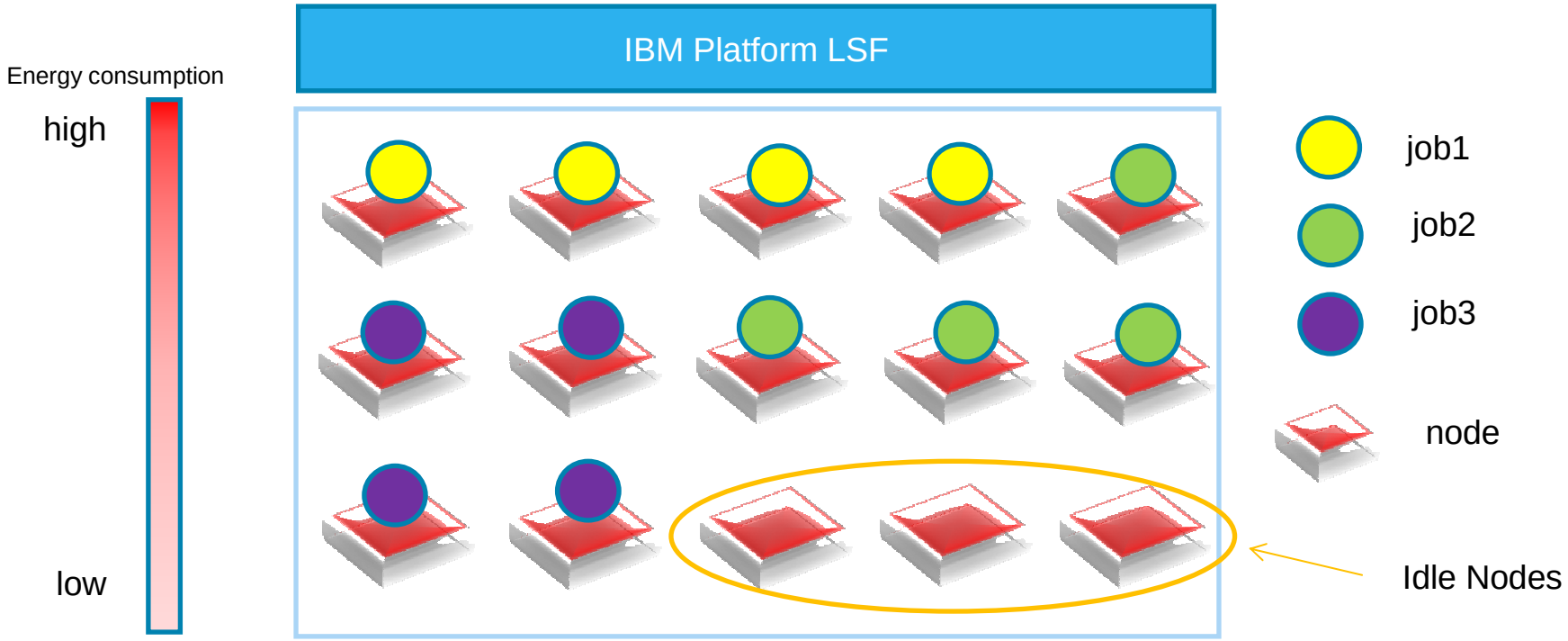
IBM Energy Aware Scheduling (Past and Current)

- In the past, IBM Platform LSF introduced a few “Green” solutions
  - Scheduling based on environmental factors, such as machine temperature
  - Power on/off idle nodes based on workload.

- In 2012, IBM EAS collaboration project with LRZ and implemented features in Loadleveler (LL)

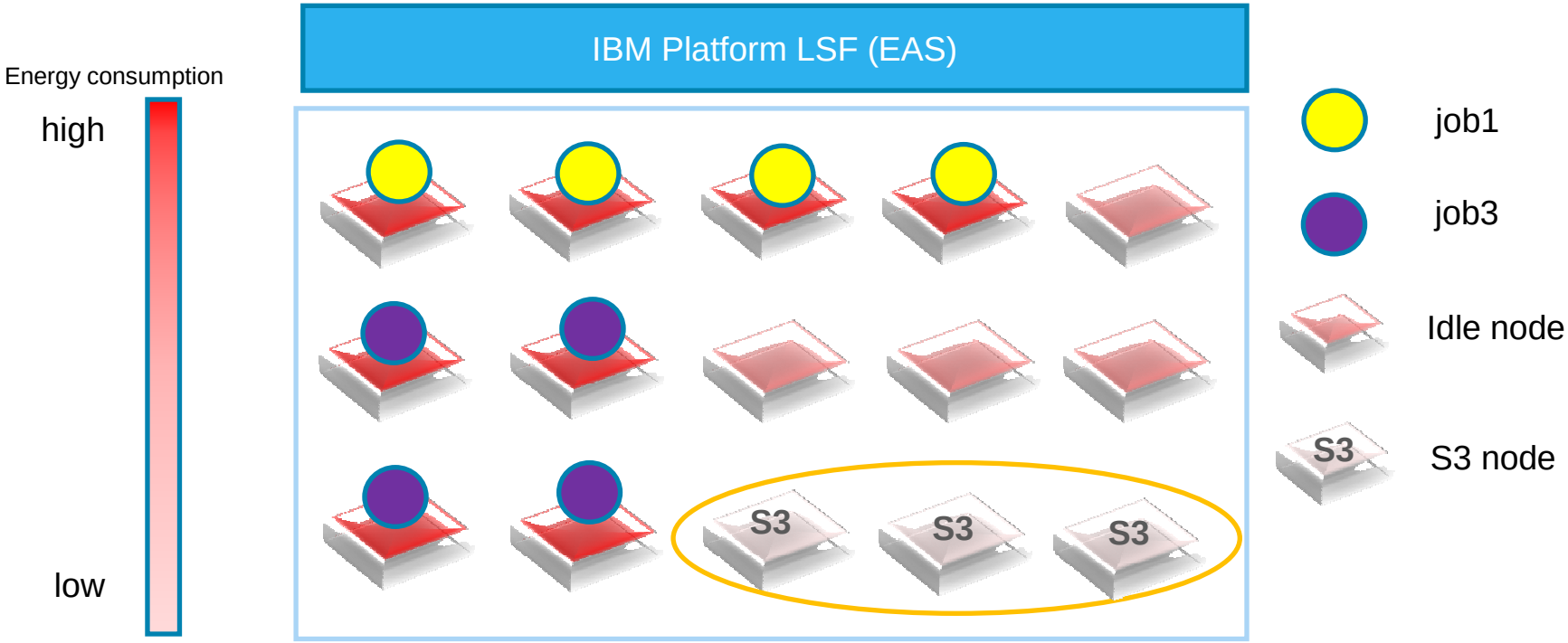
- LSF 9.1.2 release makes major enhancements and incorporates LL EAS work
  - **Features** :
    - Reduce power consumption of idle nodes
    - Optimize power consumption of active nodes for running workloads
    - Energy consumption accounting per job
  - **Hardware**: will support IBM iDataPlex and NeXtScale.
  - **Schedule**: Dec, 2013

Energy Consumption of Compute Nodes in HPC Data Center

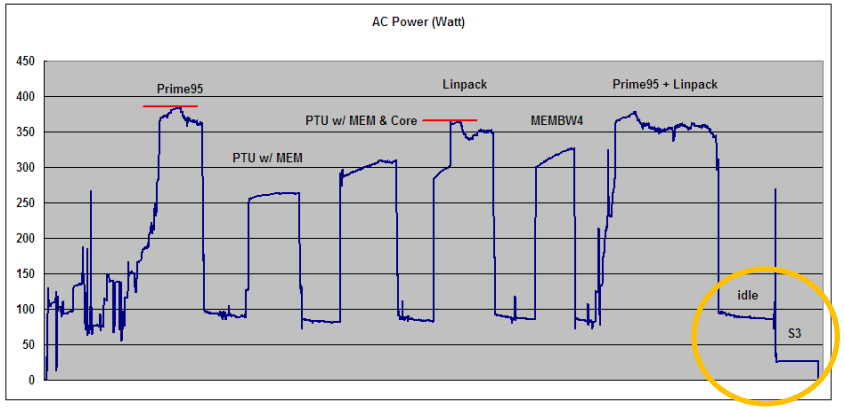


- All computer nodes runs under nominal frequency by default
  - Idle nodes consumes lots of energy. (Can we reduce power consumption for idle nodes?)
  - Active nodes run all jobs in full speed all the time. (Can we run different jobs in different cpu frequency?)

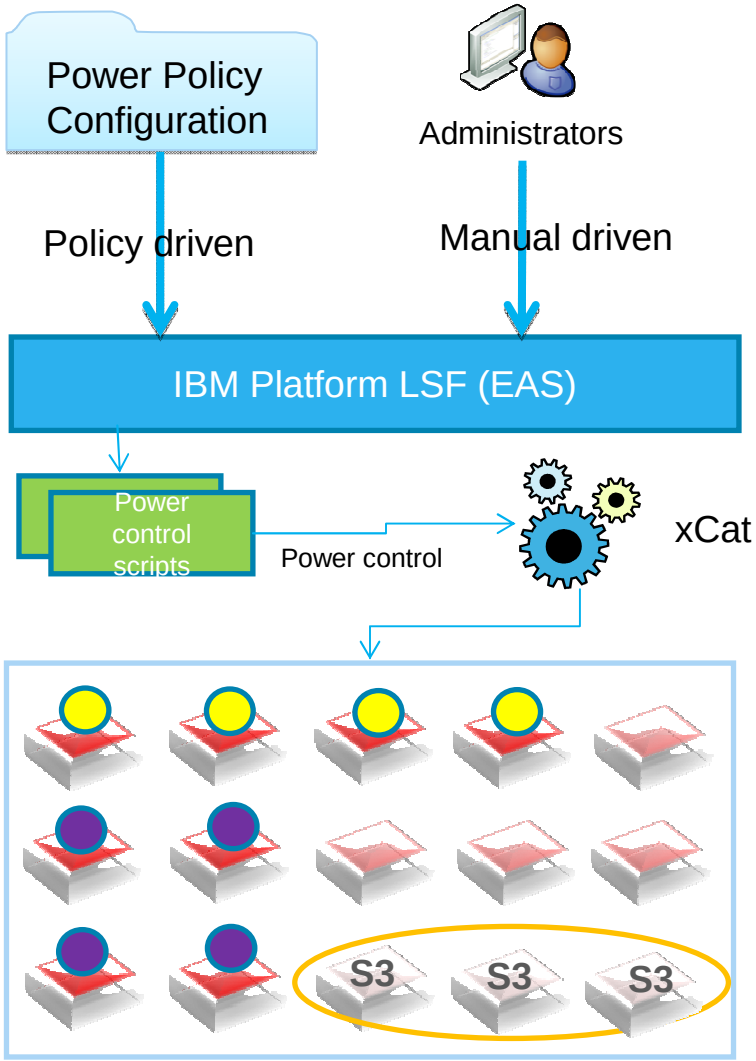
Reduce Power Consumption of Idle Nodes



- Automatically set "**ondemand**" CPUFreq Governor of nodes
- Intelligently put idle nodes into S3 state based on pre-defined policies (~60 W saved)



How LSF supports Host Power Management?



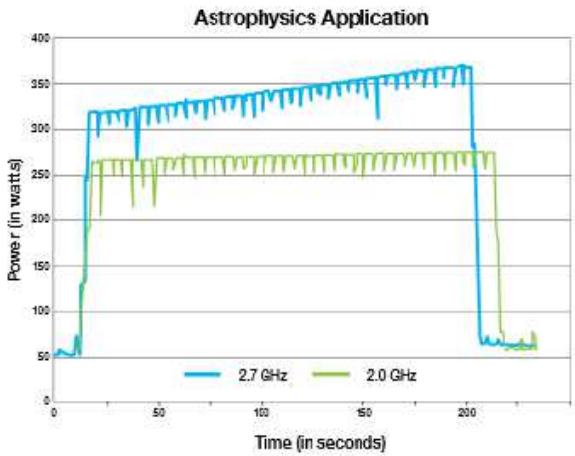
- **Manual management tools**
  - badmin hpower suspend/resume host
  - badmin hist
- **Policy Driven Power Saving**
  - Policy windows (i.e., 10:00 PM – 7:00 AM)
  - Node will be put in power saved (S3) if it is idle for a configurable period of time.
- **Power Saving Aware Scheduling**
  - Schedule jobs to use idle nodes first (Power saved nodes as last resort)
  - Aware of job request and wake up nodes precisely on demand
  - Safe period before running job on resumed nodes
- **Others**
  - customizable power control scripts
    - OOB xcat support
  - Power events tracking

Host Power Management Examples Output

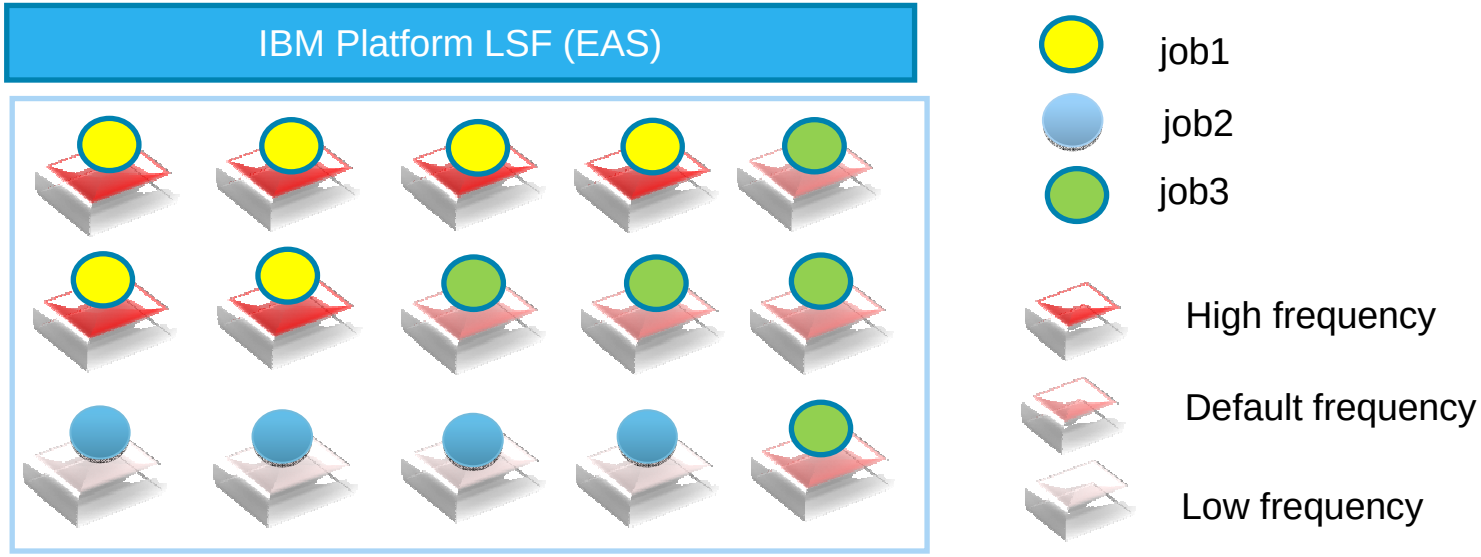
```
[lsflinux3]# badmin hpower suspend lsf_1
<lsf_1>: Host suspended

[lsflinux3]# bhosts -l lsf_1
HOST  lsf_1
STATUS           CPUF  JL/U    MAX  NJOBS    RUN  SSUSP  USUSP    RSV DISPATCH_WINDOW
closed_Power    60.00     -      2      0      0      0      0      0      -
 …
 POWER STATUS:          suspend
 IDLE TIME:             0m 44s
 CYCLE TIME REMAINING:  0m 0s

[lsflinux3]# badmin hpower resume lsf_1
<lsf_1>: Host resumed
[lsflinux3]# bhosts -l lsf_1
HOST  lsf_1
STATUS           CPUF  JL/U    MAX  NJOBS    RUN  SSUSP  USUSP    RSV DISPATCH_WINDOW
ok              60.00     -      2      0      0      0      0      0      -
…
 POWER STATUS: on
```

```
[lsflinux3]# badmin hhist lsf_1
Sun Jun 30 14:08:47: Host <lsf_1> suspend request from administrator <xyz>.
Sun Jun 30 14:08:48: Host <lsf_1> suspend request done.
Sun Jun 30 14:08:58: Host <lsf_1> suspend.
Sun Jun 30 14:10:47: Host <lsf_1> resume request from administrator <xyz>.
Sun Jun 30 14:10:48: Host <lsf_1> resume request done.
Sun Jun 30 14:10:58: Host <lsf_1> on.
```

Optimize Power Consumption of Active Nodes



- Set a default cpu frequency on nodes
- Ability to set specified frequency on core/node level for a given job/application/queue
- Intelligently tag the job and select optimal cpu frequency based on energy and performance predication and sit policies



Example of Running jobs under Specified Frequency

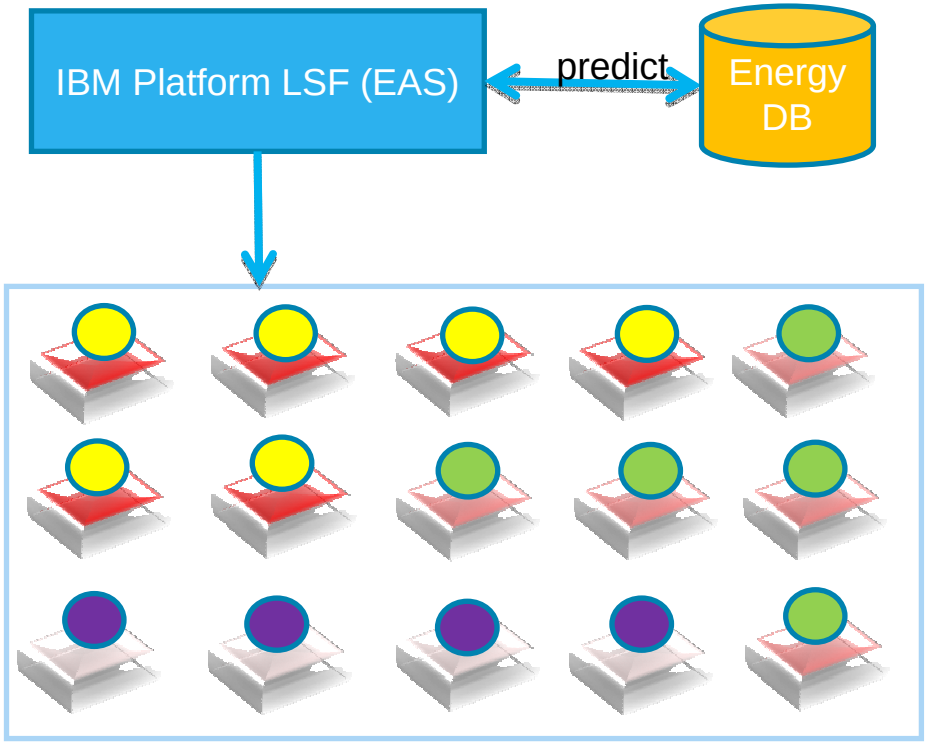
- CPU frequency can be set in:
  - -freq <frequency> as job submission option
  - CPU_FREQUENCY parameters in either LSF application profile or queue

```
[c909f04x24]$ cat parallel_job_example.cmd
#!/bin/sh
#BSUB –n 4 –R “span[ptile=2]”
#BSUB –x
#BSUB –freq 1.8GHz
#BSUB –network “mode=ip”
#BSUB –o /u/xyz/job/my.output.%J
#BSUB –e /u/xyz/job/my.error.%J

poe /tmp/llbld/bin/st.2.U 10

[c909f04x24]$ bsub < parallel_job_example.cmd
Job <1310> is submitted to default queue <production>
```

```
[c909f04x24]$ pwd
/sys/devices/system/cpu
[c909f04x24]$ find ./ -name  scaling_max_freq -exec cat {} \;
1800000
1800000
...
```

How LSF Automatically Select Optimal CPU frequency

- Step I: Learning
  - LSF evaluates the power profile of all nodes
  - calculates coefficients factors
  - save them in the energy database



- Step II: Set Default Frequency
  - System administrator defines cluster default cpu frequency (nominal or lower frequency)
- Step III: Tag the job first time
  - User submits the application with a tag
  - runs the job under default frequency
  - LSF collects energy consumption, runtime, hardware counters
  - Generates predication result and saves in database
- Step IV: Use predication
  - User re-submits the same application with the same tag and specifies energy policy
  - LSF selects the optimal cpu frequency for application based on predication result and policy setting.
  - Run the application under selected frequency

Supported Energy Policies

- Minimize Time to Solution
  - Goal: improve application performance and allow jobs to run in a higher frequency than default one.
  - How does it work?
    - Admin sets threshold value for each allowable higher frequency. The threshold value defines expected minimal performance improvement for application running on target frequency against default frequency.
    - LSF will select the highest frequency that meets performance threshold value for application

- Minimize Energy to Solution
  - Goal: save energy with configured maximum performance degradation <=x%
  - How does it work?
    - Admin sets threshold value to define maximum acceptable performance degradation. Say 10%
    - LSF will select the lowest frequency that meets threshold value for application.

Example of Automatically select optimal CPU Frequency

- Submit the application with tag first time (1.90 GHz as default frequency)

```
[c909f04x26]$ cat parallel_creat_tag.cmd
#!/bin/sh
#BSUB -n4 -R "span[ptile=2]"
#BSUB -x
#BSUB -a "eas(PARALLEL_STREAM, create)“
#BSUB -network "mode=ip"
#BSUB -o /u/xyz/job/my.output.%J
#BSUB -e /u/xyz/job/my.error.%J

poe /tmp/llbld/bin/st.2.U 10

[c909f04x26]$ bsub < parallel_creat_tag.cmd
Job <1327> is submitted to default queue <normal>.
```

```
[c909f04x26]$ bjobs -l

Job <1327>, User <xyz>, Project <default>, Status <RUN>, Queue <normal>,
                     Combined CPU Frequency <1.90 GHz>, Energy policy tag <xyz.
                     PARALLEL_STREAM>, Command <#!/bin/bash;#BSUB -n4 -R "span[
                     ...
Fri Nov 15 08:45:12: Started on 4 Hosts/Processors <c909f04x24> <c909f04x24> <c
                     909f04x26> <c909f04x26>, Execution Home </u/xyz>, Execu
                     tion CWD </u/xyz/job>;
...
```

Example of Automatically select optimal CPU Frequency

- Query predication result

```
[c909f04x26]$ bentags

  Energy Tag Name: PARALLEL_STREAM
           Job ID: 1327
             User: xyz
Default Frequency: 1.90 GHz
Node's Energy Use: 0.002586 kWh
          Runtime: 44 Seconds

CPU_FREQ(GHz) EST_USAGE(kWh) ENER_VAR(%) EST_RUNTIME(Sec) RUNTIME_VAR(%) POWER(W)
2.40          0.003084       19.27       42               -4.55          264.38
2.30          0.002950       14.07       42               -4.55          252.86
2.20          0.002767       7.01        41               -6.82          242.99
2.10          0.002761       6.76        43               -2.27          231.16
2.00          0.002633       1.81        43               -2.27          220.43
1.90          0.002586       0.00        44               0.00           211.60
1.80          0.002491       -3.69       44               0.00           203.78
1.70          0.002447       -5.38       45               2.27           195.76
1.60          0.002391       -7.56       46               4.55           187.10
1.50          0.002277       -11.95      46               4.55           178.21
1.40          0.002229       -13.82      47               6.82           170.71
1.30          0.002212       -14.49      49               11.36          162.48
1.20          0.002297       -11.19      53               20.45          156.01
```

Example of Automatically select optimal CPU Frequency

- Use Minimum Time Policy

```
Begin Min_Time
CPU_FREQ        RUNTIME_VAR
2000000KHz       -1
2400000KHz       -4
End Min_Time
```

```
[c909f04x26]$ cat parallel_minimize_time.cmd
#!/bin/sh
#BSUB -n4 -R "span[ptile=2]"
#BSUB -x
#BSUB -a "eas(PARALLEL_STREAM, minimize_time)"
#BSUB -network "mode=ip"
#BSUB -o /u/xyz/job/my.output.%J
#BSUB -e /u/xyz/job/my.error.%J

poe /tmp/llbld/bin/st.2.U 10

[c909f04x26]$ bsub < parallel_minimize_time.cmd
Job <1328> is submitted to default queue <normal>.
```

```
[c909f04x26]$ bjobs –l 1328

Job <1328>, User <xyz>, Project <default>, Status <RUN>, Queue <normal>,
                     Combined CPU Frequency <2.40 GHz(auto)>, Energy policy tag <xyz
                     .PARALLEL_STREAM>, Command <#!/bin/bash;#BSUB -n4 –R
                     ...
```

Example of Automatically select optimal CPU Frequency

- Use Minimum Energy Consumption Policy
  - THRESHOLD_RUNTIME_VAR=10

```
[c909f04x26]$ cat parallel_minimize_energy.cmd
#!/bin/sh
#BSUB -n4 -R "span[ptile=2]"
#BSUB -x
#BSUB -a "eas(PARALLEL_STREAM, minimize_energy)"
#BSUB -network "mode=ip"
#BSUB -o /u/xyz/job/my.output.%J
#BSUB -e /u/xyz/job/my.error.%J

poe /tmp/llbld/bin/st.2.U 10

[c909f04x26]$ bsub < parallel_minimize_energy.cmd
Job <1329> is submitted to default queue <normal>.
```

```
[c909f04x26]$ bjobs –l 1329

Job <1329>, User <xyz>, Project <default>, Status <RUN>, Queue <normal>,
                     Combined CPU Frequency <1.40 GHz(auto)>, Energy policy tag <xyz
                     .PARALLEL_STREAM>, Command <#!/bin/bash;#BSUB -n4 –R
                     ...
```

Job Energy Consumption Accounting

- LSF collects energy consumption per job and writes into job output and accounting file

```
[xyz@bjidp1 job]$ cat my.output.320
...
Resource usage summary:

    CPU time :                  501.30 sec.
    Total Requested Memory : -
    Delta Memory :              -
    (Delta: the difference between Total Requested Memory and Max Memory.)
    Job Energy Consumption : 0.000921 kWh
...
```

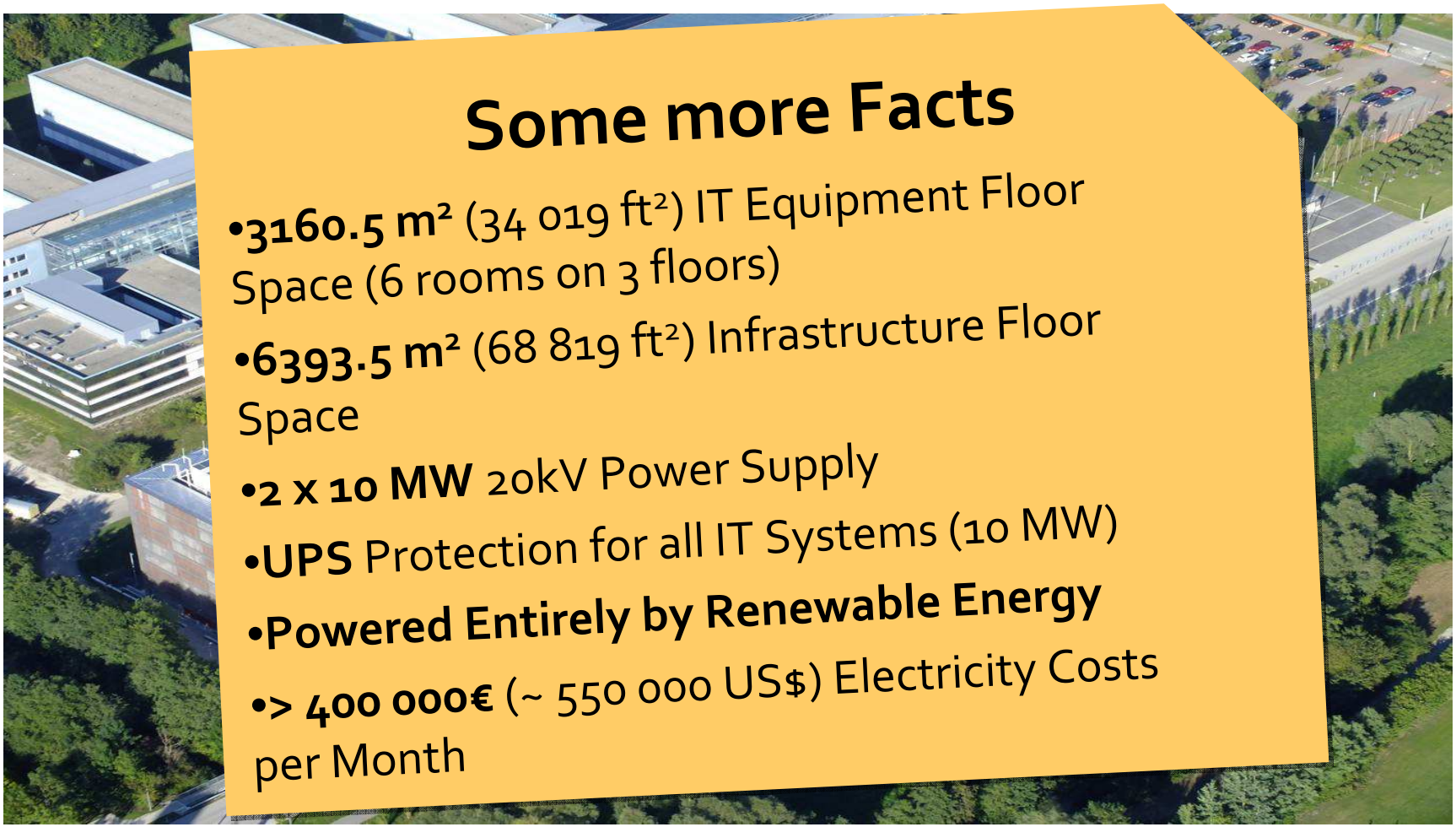
```
[xyz@bjidp1 job]$ bacct -l 320
...
Accounting information about this job:
     Share group charged </xyz>
     CPU_T     WAIT     TURNAROUND   STATUS     HOG_FACTOR    MEM    SWAP
    501.30       15             36     done        13.9249     0M      0M

Host based accounting information about this job:
                            HOST      CPU_T       MEM      SWAP
                          bjidp1     501.00        0M        0M

JOB ENERGY CONSUMPTION:
0.000921 kWh
```

Thank-You!

Visit us at:
ibm.com/technicalcomputing
ibm.com/platformcomputing

Energy Aware Scheduling on SuperMUC

First Experiences



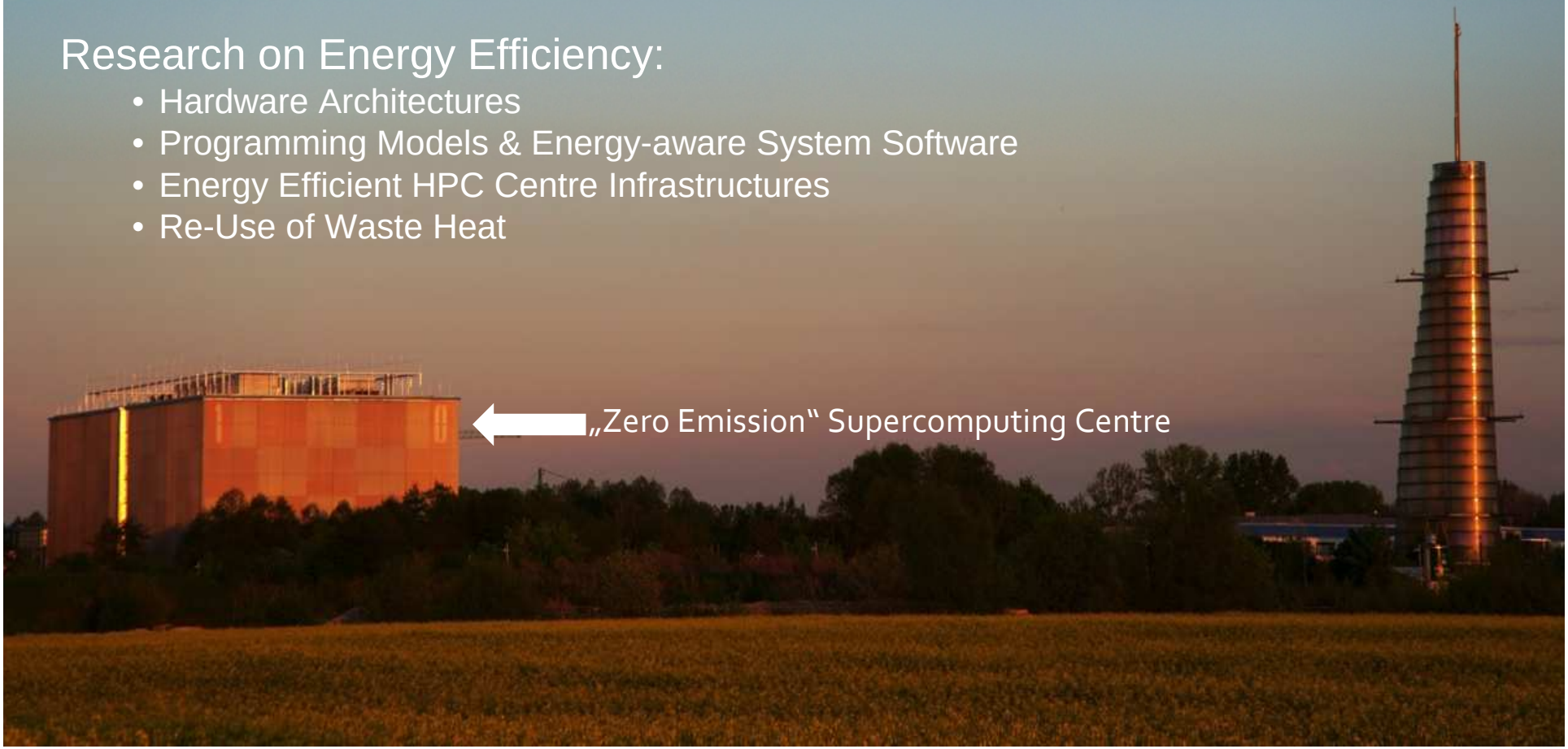
Some more Facts

- **3160.5 m²** (34 019 ft²) IT Equipment Floor Space (6 rooms on 3 floors)
- **6393.5 m²** (68 819 ft²) Infrastructure Floor Space
- **2 x 10 MW** 20kV Power Supply
- **UPS** Protection for all IT Systems (10 MW)
- **Powered Entirely by Renewable Energy**
- **> 400 000€** (~ 550 000 US$) Electricity Costs per Month

Energy Efficient HPC @ LRZ

Research on Energy Efficiency:

- Hardware Architectures
- Programming Models & Energy-aware System Software
- Energy Efficient HPC Centre Infrastructures
- Re-Use of Waste Heat



SuperMUC

TOP 500®
SUPERCOMPUTER SITES
2897 TFlop/s HPL Performance
Rank 4 (June 2012)
Rank 6 (November 2012)
Rank 9 (June 2013)
Rank 10 (November 2013)

24

SuperMUC: Warm Water Cooling

- Heat flux > 90% to water; very low chilled water requirement
- Power advantage over air-cooled node:
  - Warm water cooled ~10%
    (cold water cooled ~15%)
  - due to lower $T_{components}$ and no fans
- Typical operating conditions: T_{air} = 25 – 30°C, T_{water} = 18 – 45°C

25

SuperMUC Hardware & Software Features for EAS

- SuperMUC features node-level power measurements
- The workload manager (IBM Loadleveler) will not share nodes between users/jobs

➢ **We can measure the Energy-to-Solution of each job on SuperMUC!**

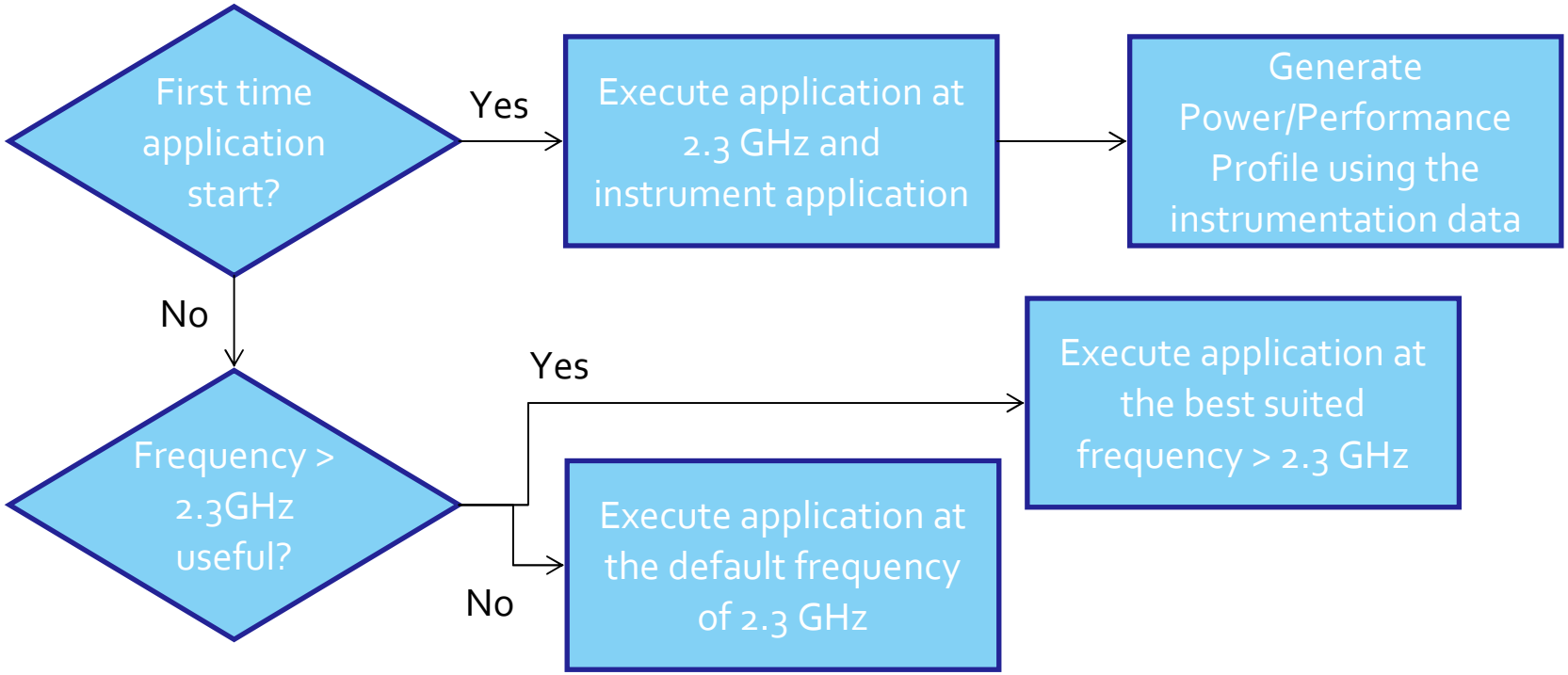
- SuperMUC's Intel Xeon (SandyBridge) CPUs can vary their clock frequency from 1.2 GHz to 2.7 GHz (and higher when using Turbo Mode)
- Using this, we can vary the Energy-to-Solution depending on the Power/Performance Profile of the application (e.g. memory-bound applications will draw more power but not see a benefit from higher clock frequencies)

➢ **Goal: Find a good (energy efficient) CPU frequency for the majority of jobs and avoid slowing down jobs that would run efficiently on higher frequencies**

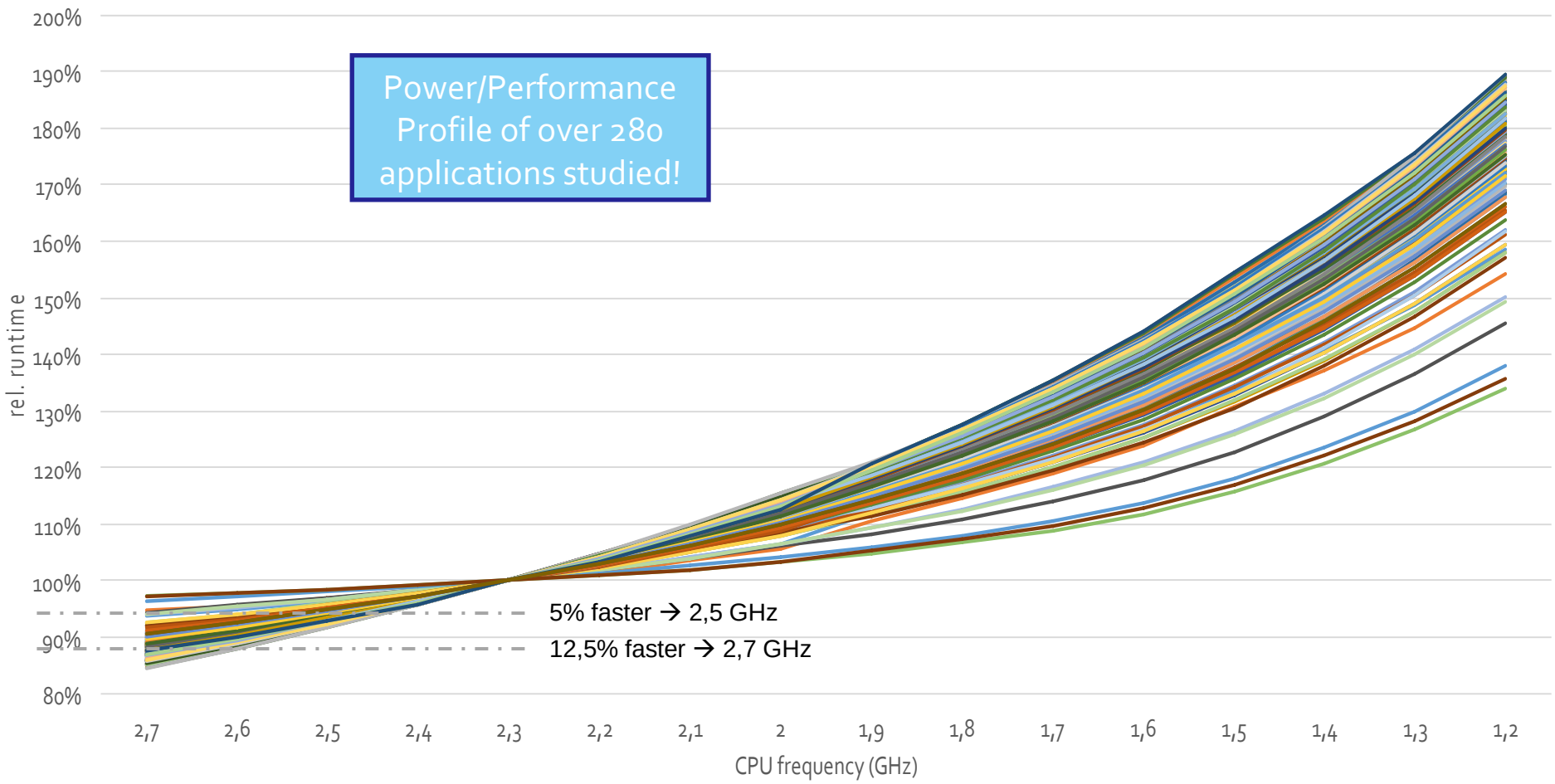
SuperMUC EAS Policy

- User tags an application
  - Loadleveler Keyword: energy_policy_tag = <some_user_assigned_tag>

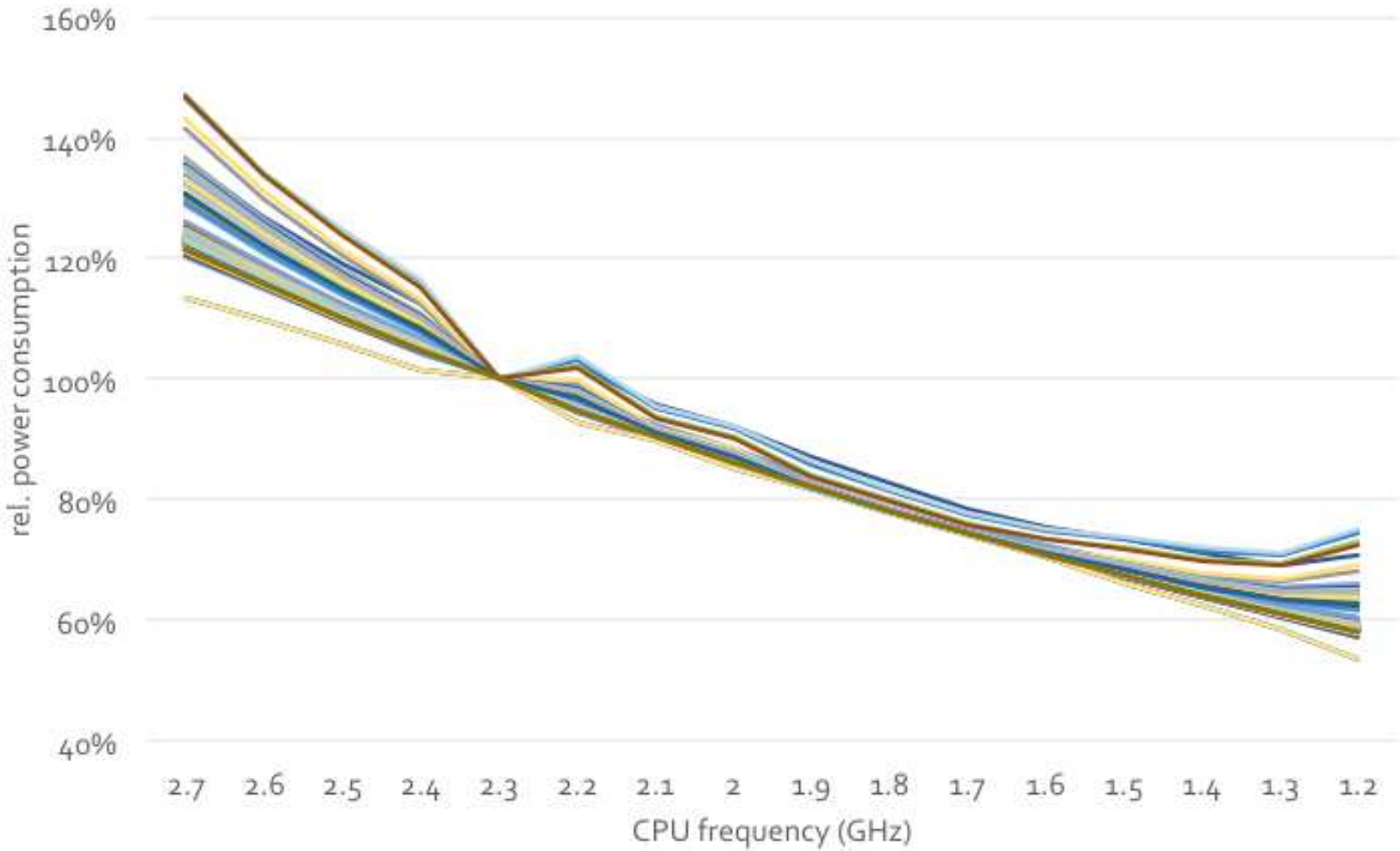
- Workflow during application execution:



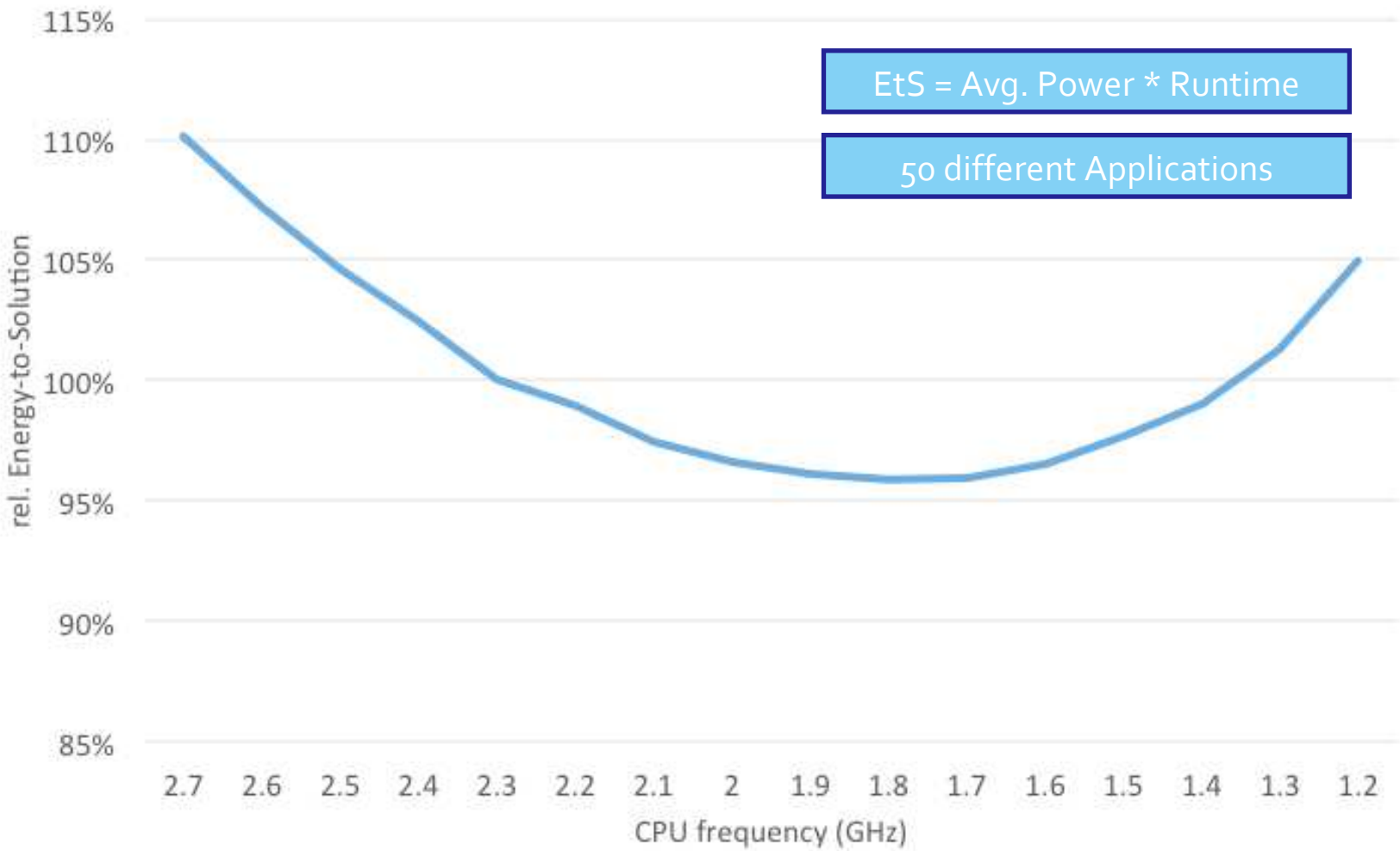
Runtime Profile of SuperMUC Applications



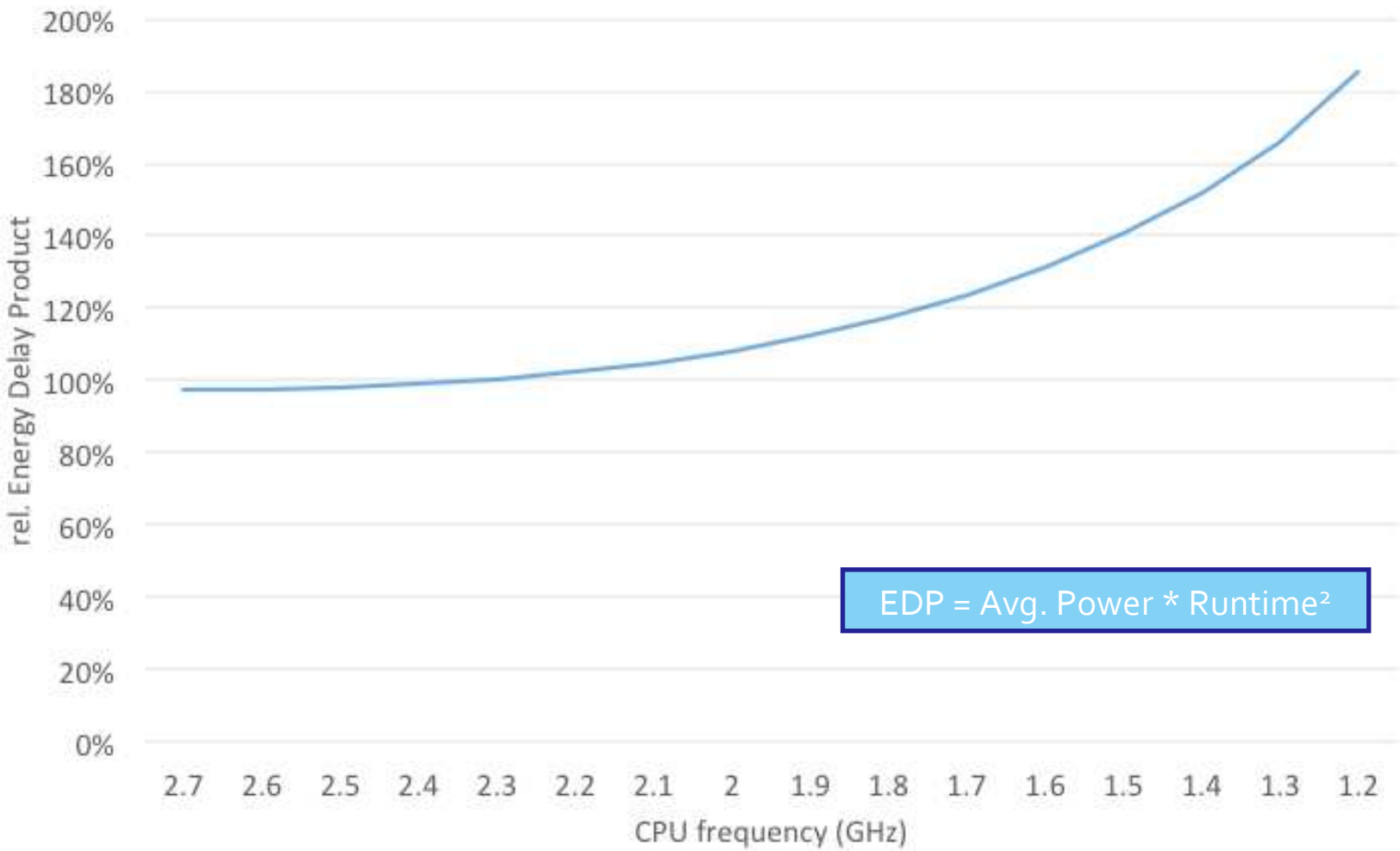
Power Profile of SuperMUC Applications



Average Energy-to-Solution Profile on SuperMUC



Average Energy-Delay-Product on SuperMUC



EDP = Avg. Power * Runtime²

Many Thanks for your Attention!

Computational & Information Systems Laboratory

CISL

Optimizing Energy Efficiency at the NCAR Wyoming Supercomputing Center

Aaron Andersen
Deputy Director Operations & Services
November 2013

NCAR

NCAR

General Facility Information

- 153,000 Gross Sq.Ft.
- 2 - 12,000 Sq.Ft. Data Halls
- 4 – 4MW electrical modules 16MW possible
- 98~99% annual hours Evaporative Water Side Economizer
- Critical Systems are on UPS and Generator Backup
  - Network, Cybersecurity, Storage, Archive
  - HPC Nodes are not on UPS
- 10 ft. Raised Floor System
  - Primary reason was large liquid cooled systems
  - Secondary reason limit copper distances
- Heat Pump System Recover Heat from Computing for Building Use

NCAR

Big Picture Focus On Efficiency

Utilize the regions cool, dry climate to minimize energy use

- Very low pressure drop
  - Minimize bends
  - Oversized pipe
- Elevated chilled water temp
  - 65 degree

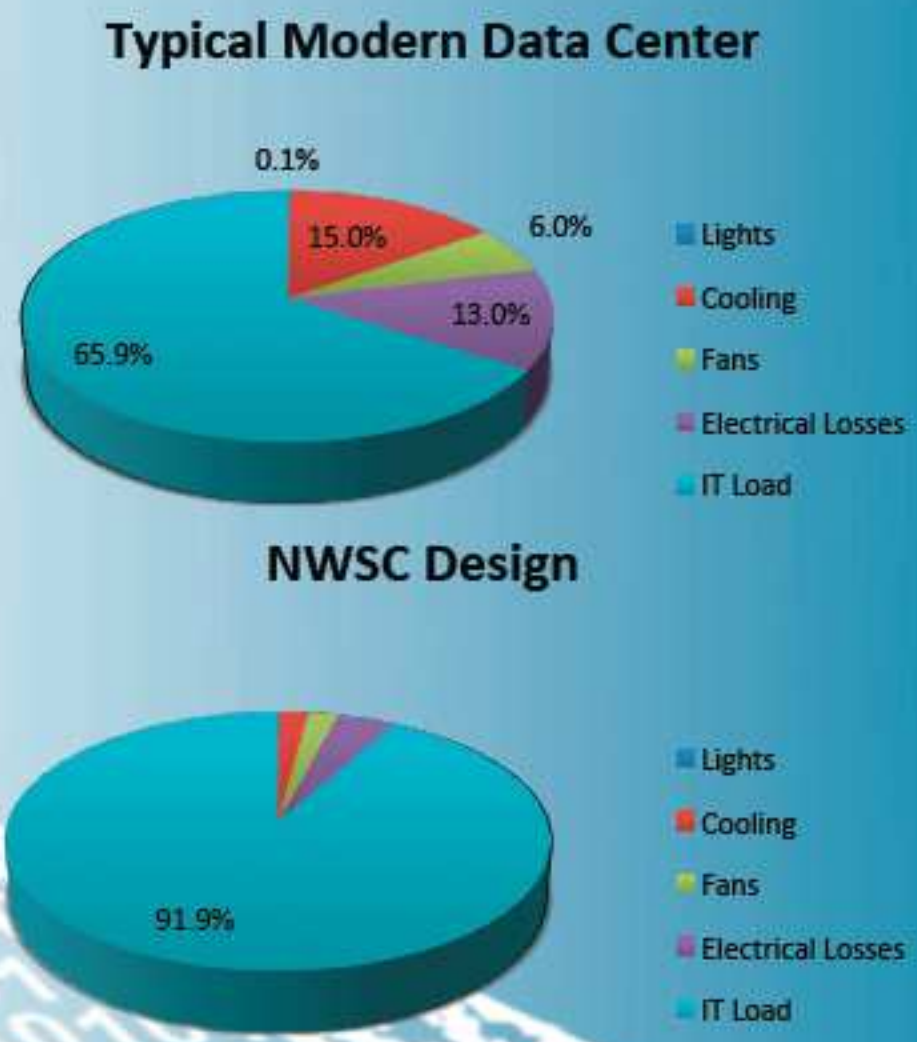
Utilize liquid cooled computer solutions where practical

- HPC Systems

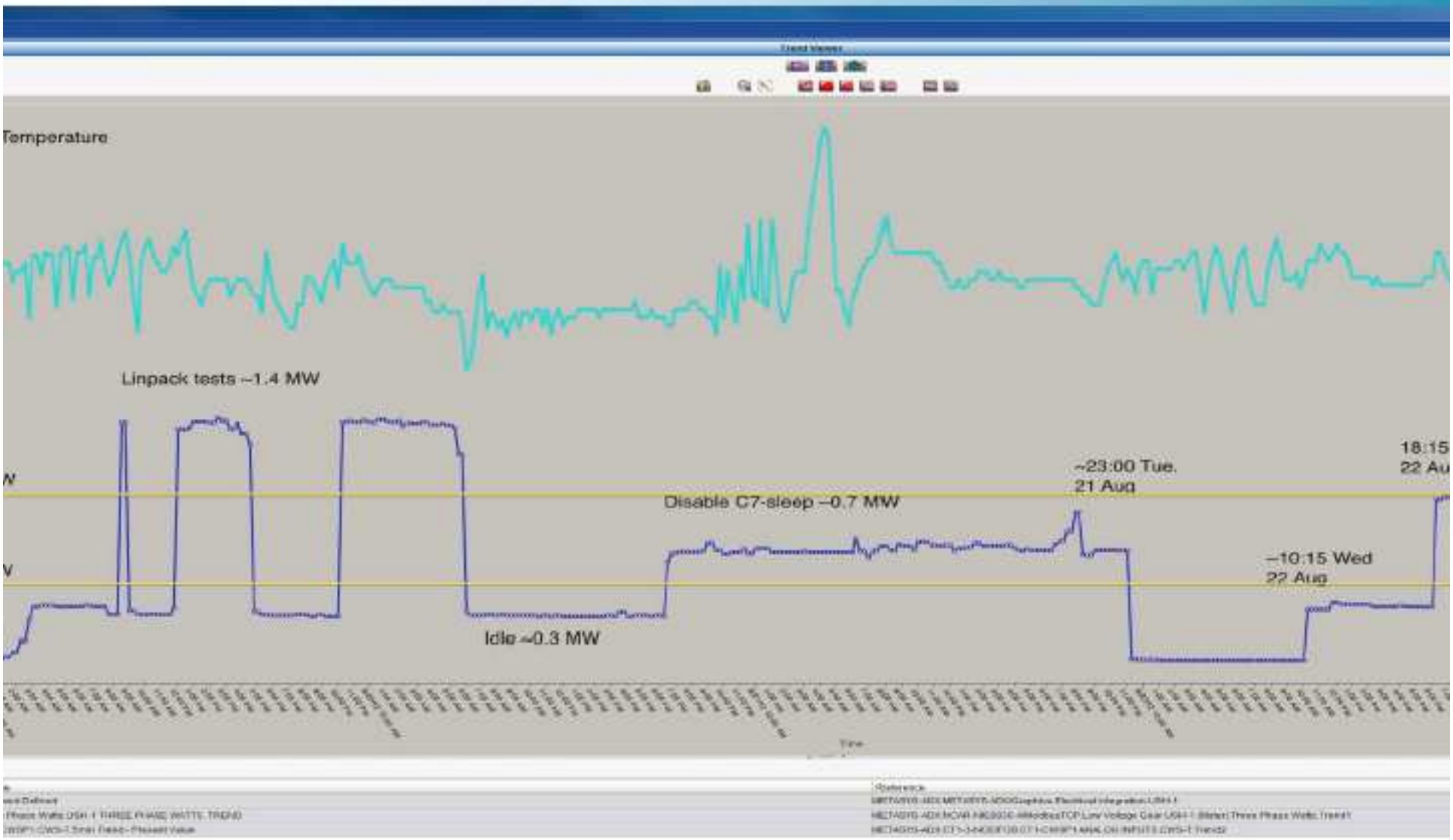
Utilize hot aisle containment for commodity equipment

Focus on the biggest losses

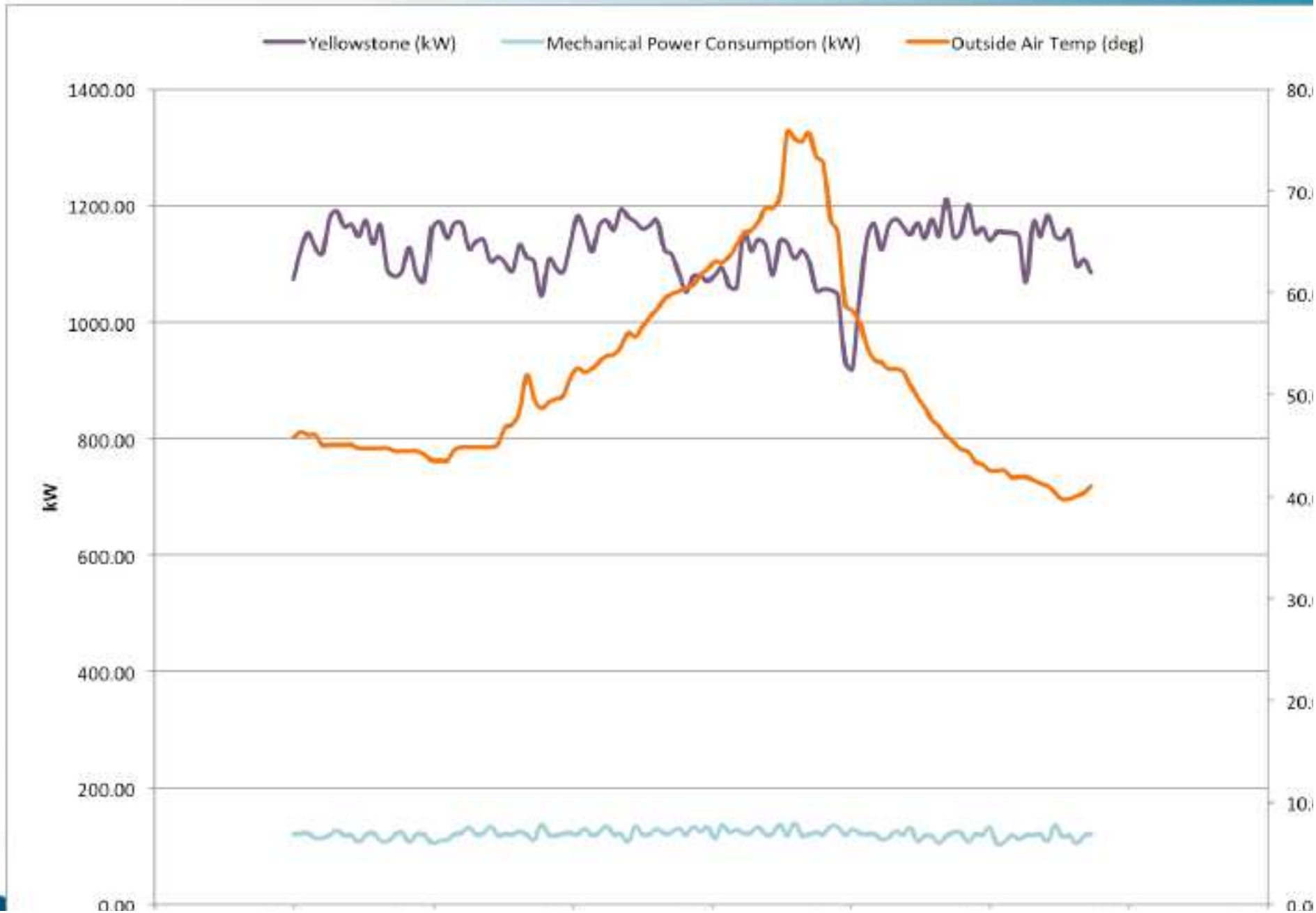
- Compressor based cooling
- Transformer losses



Highly Dynamic Compute System Loads



Characterizing the Building Response



Now Focused on the Left Side of PUE Decimal
The Logical Progression of Efficiency

Facilities Efficiency
- Facility Engineering
- Operations & Automation

System Efficiency
- Hardware Vendor
- Procurement Methods

Scheduling Efficiency
- Energy Aware Schedulers
- Energy Information/Charging

Application Efficiency
- Algorithms
- Data Movement

NCAR

Thermal-Aware Heterogeneous Computing”

NSF Funded Proposal

- Colorado State University
  - Sudeep Pasricha
  - HJ Siegel
  - Mark Oxely
- NSF Funded Project
- Looking at energy aware resource scheduling

Future Focus and Issues

Focus

- Reporting on Energy Use & Accounting
  - Informational for User
  - Given the focus of our research simply that may be highly effective
- Evaluate energy aware scheduling in our test environments

Issues

- User community already fi runtime variability concern
- Getting users to think of jo runs and optimization of energy use
  - Focus for most has been go
- Are the benefits worth the complexity and the cost
  - Simple payback
  - Longevity

Attribution

- Prius – *This Wikipedia and Wikimedia Commons image is from the user Chris 73 and is freely available at //commons.wikimedia.org/wiki/File:Toyota_Prius_side.jpg under the creative commons cc-by-sa 3.0 license.*

- Formula 1 - This image was originally posted to Flickr by KamalSell at http://flickr.com/photos/49502597@N00/378125991. It was reviewed on 20 June 2007 by the FlickreviewR robot and was confirmed to be licensed under the terms of the cc-by-2.0.

Energy Efficient Computing

The Hartree Centre
A Research Collaboratory in Association with IBM

Professor Terry Hewitt
Delivery Project Executive
(terry.hewitt@stfc.ac.uk)

Energy Efficient Computing

- The Chancellor of the Exchequer (George Osborne) visited Daresbury on February 1, 2013 and announced a £19M capital investment to finance research into energy efficient computing.
- “to firmly establish the UK as the world leader in energy efficient supercomputer software development to meet big data challenges.”

44

Result

- Note: £19M is capital

- Establish an Energy Efficient Research Programme
- Mission Statement is:
  - Working with academia and industry the Scientific Computing Department will utilise our expertise, computing infrastructures and relationships to:
  - Create and contribute to the development of new software
  - Optimise and performance tune existing software

  - This will increase the UK's scientific and economic impact by avoiding waste and maximising energy efficiency.
- This will be financed by Grant and Commercial Income

45

£19M

- £10M to IBM under the existing contractual arrangements
- £2.3M refurbish Machine Hall
- £7M Open procurement
  - £1.0M – novel cooling technology demonstrator
  - £1.0M – data flow
  - £0.7M – novel low power processors
  - £4.0M – Big data and data analytics
- Procurement underway so can’t say much!
- All subject to BiS/Cabinet Office approval

46

- Tenders go out 13th November
- Tenders back to us – 9th December
- Orders on 16th December

47

Evolution vs Disruptive

- Evolution
  - Big Data & Data Analytics
  - Novel Cooling
- Disruptive
  - Dataflow
- Evolution or Revolution?
  - Low Power Processors
- £4M on
  - Focus on Software (with some hardware)

Key Areas - Our Perspective

- Energy to solution – “flops per Watt”
- Energy - efficient data centre techniques
- Energy - efficient software

- Holistic approach
  - Bring together lots of independent efforts
    - Batch queueing
    - Machine hall power limits
    - Power down unused assets
    - Minimize energy
      - Per job
      - Per rack
      - Per hall

49

http://www.stfc.ac.uk/hartree

http://community.hartree.stfc.ac.uk

hartree@stfc.ac.uk
01925 603 444

Thank you for listening

QUESTIONS

Thank-You!