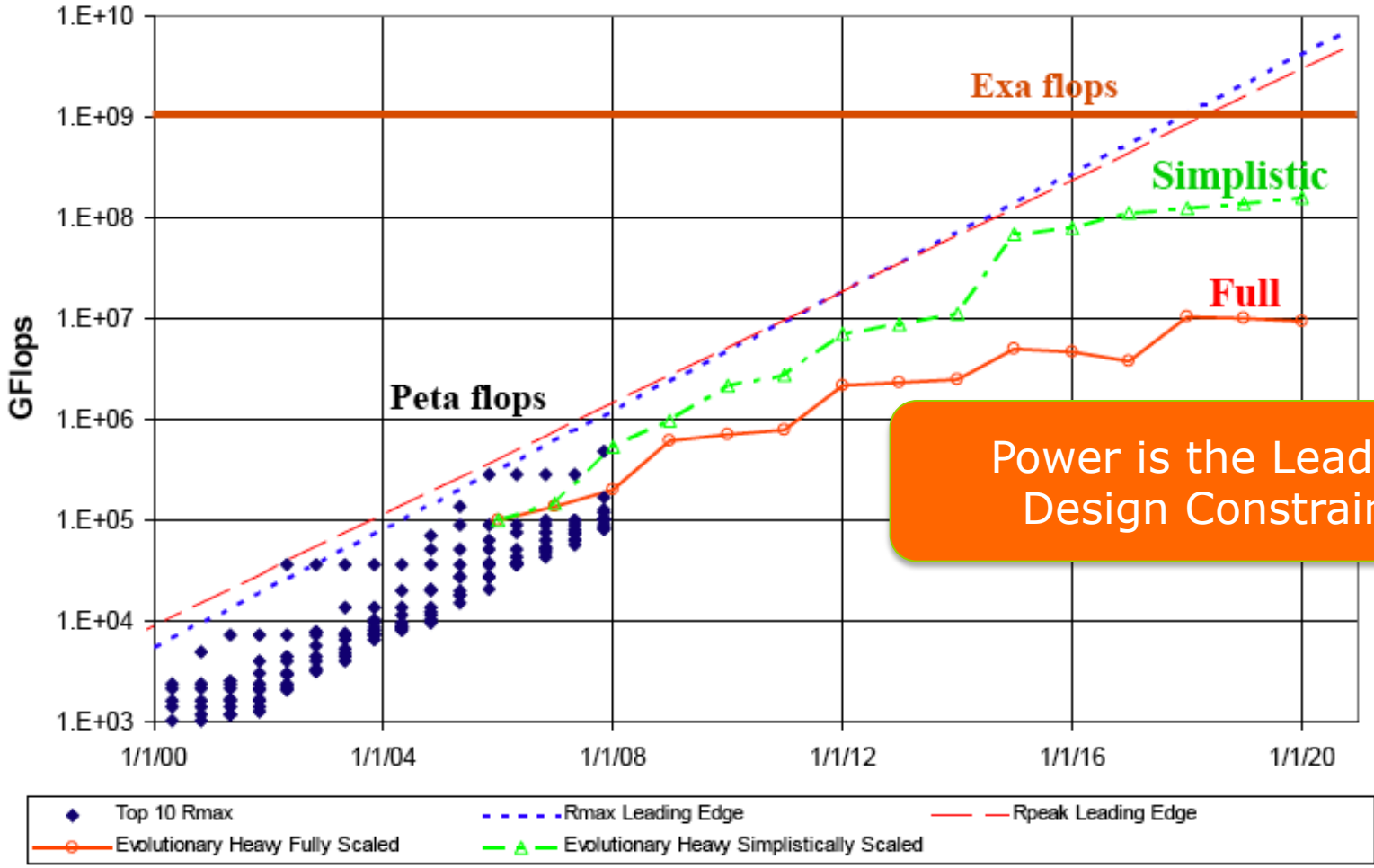
Towards Energy Efficient Exascale Computing

Natalie Bates,
Energy Efficient HPC Working Group

Technology roadmaps depart from historical gains



Power is the Leading Design Constraint

From Peter Kogge, DARPA Exascale Study

How to address this challenge?

"You can only improve
what you can measure"

…. but, this is easier said than done.

Driving HPC energy efficiency

- Energy Efficient HPC Working Group
  - Forum for sharing of information
    - Peer-to-peer exchange
    - Best practices and case studies
    - Collective action
  - Open to all interested parties

EE HPC WG Website
http://eehpcwg.lbl.gov
Email
energyefficientHPCWG@gmail.com
Energy Efficient HPC Linked-in Group
http://www.linkedin.com/groups?gid=2494186&trk=myg_ugrp_ovr

With a lot of support from Lawrence Berkeley National Laboratory

Agenda

- Apples to apples measurements
  - Flops per “whatt”?
  - Refining PUE with TUE
- Moving beyond the 1.0
  - Energy re-use
  - Carbon reduction

Improving energy efficiency HPC metrics

- Collaboration between Top500, Green500, Green Grid and EE HPC WG

- Evaluate and improve methodology, metrics, and drive towards convergence on workloads
- Form a basis for evaluating energy efficiency of individual systems, product lines, architectures and vendors
- Target architecture design and procurement decision making process

Workloads

- Leverage well-established benchmarks
- Must exercise the HPC system to the fullest capability possible
- Measure behavior of key system components including compute, memory, interconnect fabric, storage and external I/O
- Use High Performance LINPACK (HPL) for exercising (mostly) compute sub-system
- Use RandomAccess (Giga Updates Per second or GUPs) for exercising memory sub-system (?)
- *Need to identify workloads for exercising other sub-systems*

Power methodology complexities and issues

- Fuzzy lines between the computer system and the data center, e.g., fans, cooling systems
- Shared resources, e.g., storage and networking
- Data center not instrumented for computer system level measurement
- Measurement tool limitations, e.g., frequency, power verses energy
- dc system level measurements don’t include power supply losses

Proposed Improvement

- Current power measurement methodology is very flexible, but compromises consistency between submissions
- Proposal is to keep flexibility, but keep track of rules used and quality of power measurement
- 3 Levels of power measurement quality
  - Sampling rate; more measurements means higher quality
  - Completeness of what is being measured; more of the system translates to higher quality
  - Common rules for start/stop times
  - Vision is to continuously 'raise the bar' with higher levels

Power/Energy Measurement Methodology (Current Proposed)

Level	Aspect 1: Time Fraction & Granularity	Aspect 2: Machine Fraction	Aspect 3: Subsystems Measured
1	20% of run Power measurement ≥ 1 per second Report ≥ 1 avg. power measurement	(larger of) 1/64 of machine or 1kW	[Y] Compute nodes [ ] Interconnect net [ ]Storage [ ]Storage Network [ ]Login/Head nodes
2	100% of run Power measurement ≥ 1 per second Report ≥ 10 avg. power measurements	(Larger of) 1/8 of machine or 10kW	
3	100% of run Total integrated energy measurement Report ≥ 10 running total energy measurements	Whole machine	

Testing proposed improvements: early adopter phase

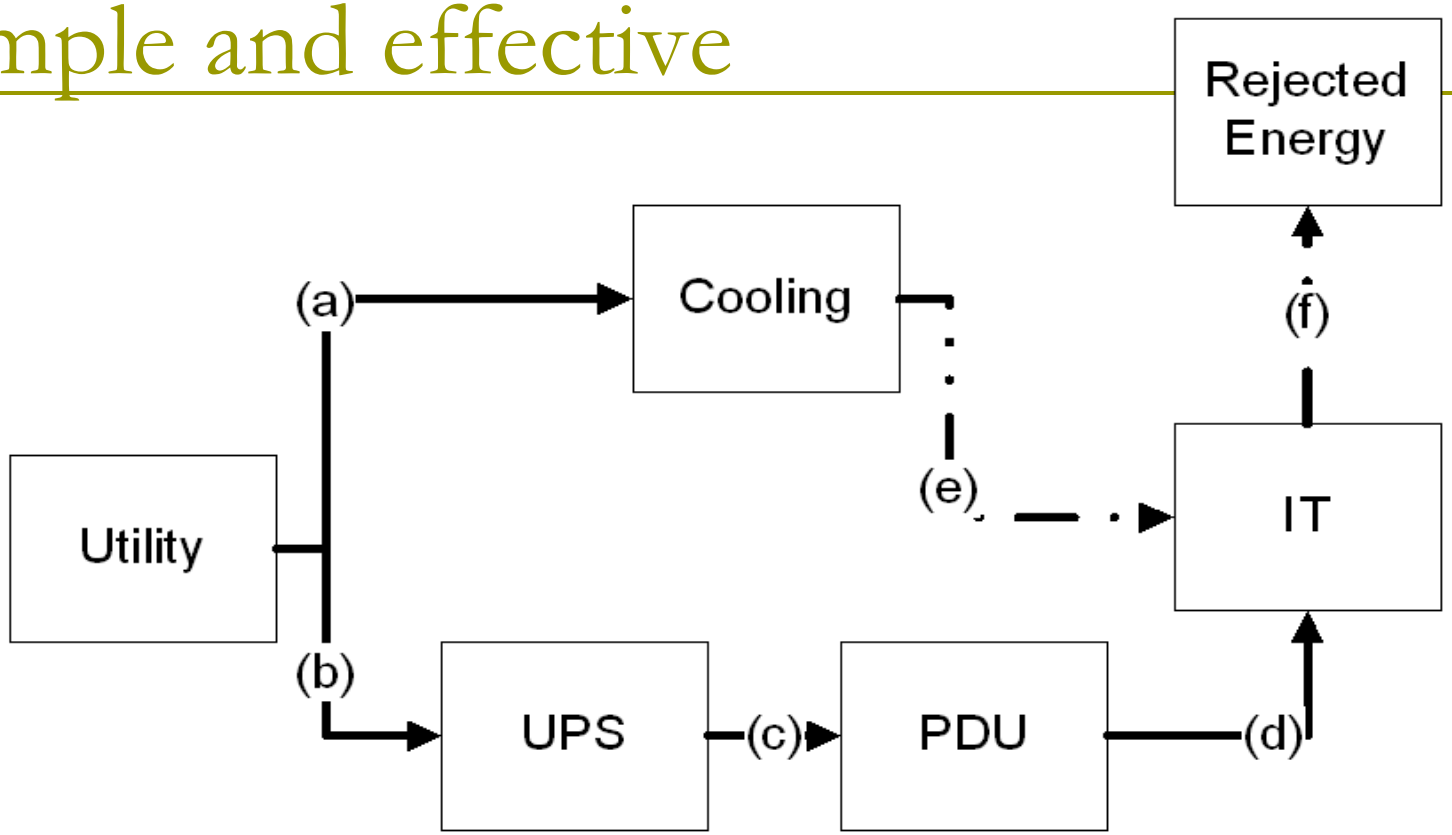
- 5 early adopters to use methodology for June'12 submissions to Top500 and Green500
  - Lawrence Livermore National Laboratory
  - Leibniz Supercomputing Center
  - Oak Ridge National Laboratory
  - Argonne National Laboratory
  - Université Laval, Calcul Québec, Compute Canada
- Seeking feedback
  - Early adopters
  - The Green Grid review
  - ISC Birds of Feather
- Next revision expected by September 2012

ISC 2012 Birds of Feather

- **Improving Power Measurement Methodology for Driving Energy Efficiency**
- **Tuesday, June 19, 2012**
  12:15 PM - 1:00 PM
- Join panelists Kim Cupps, Lawrence Livermore National Laboratory, Herbert Huber, Leibniz Supercomputing Center, Buddy Bland, Oak Ridge National Laboratory, and Susan Coghlan, Argonne National Laboratory as they review their experiences in taking power measurements while running HPL using a common methodology defined jointly by the Top500, Green500, Green Grid and the Energy Efficient HPC Working Group.

Refining PUE with TUE

Power Usage Effectiveness (PUE) – simple and effective



$$PUE = \frac{Total\ Energy}{IT\ Energy} = \frac{Cooling + PowerDistribution + Misc + IT}{IT} = \frac{a+b}{d}$$

Refining PUE for better comparison - TotalPUE

- PUE does not account for cooling and power distribution losses inside the compute system
- ITPUE captures support inefficiencies in fans, liquid cooling, power supplies, etc.
- TUE provides true ratio of total energy, (including internal and external support energy uses)
- TUE preferred metric for inter-site comparison

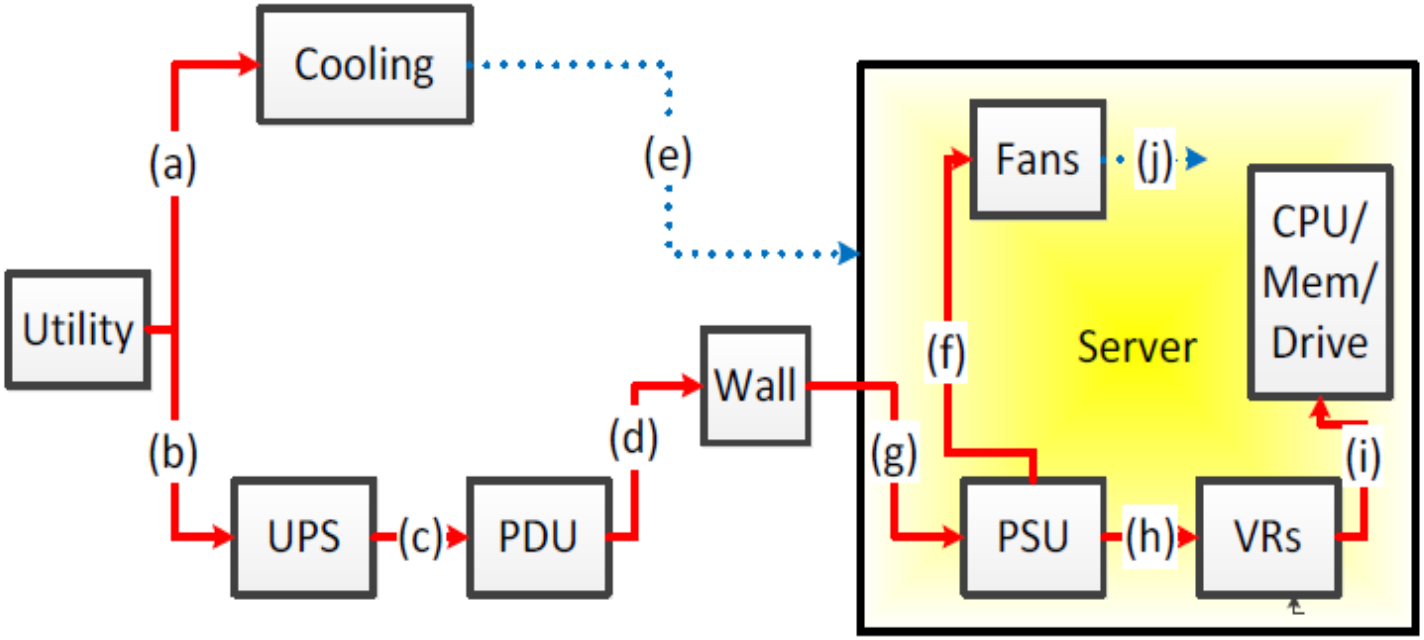
EE HPC WG Sub-team proposal

Example: PUE and Cooling

	PUE
Case A: Both building and IT fans.	Medium
Case B: Only IT fans.	Lowest
Case C: Only building fans.	Highest

PUE definition includes IT in numerator and denominator
=> Lower PUE if IT cooling fans

Combine PUE and ITUE for TUE



$$TUE = DCPUE \times ITPUE = \frac{a+b}{d} \times \frac{g}{i} = \frac{a+b}{i}$$

Challenges

- CPU, GPU, memory, memory controllers, MICs are all IT
- Fans, pumps, PSUs and VRs are all infrastructure
- But, what about disk drives, status lights and baseboard controllers?

Beyond the 1.0

PUEs: Reported and Calculated

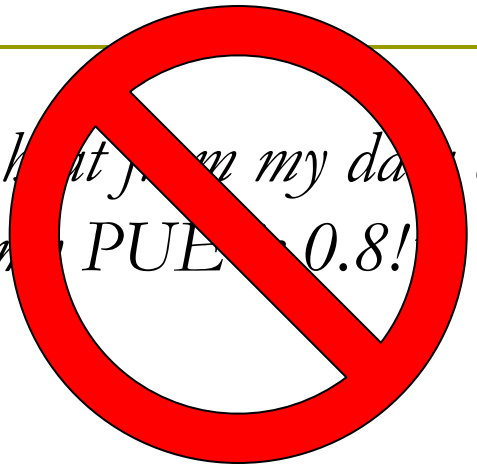
	PUE
EPA Energy Star Average – reported in 2009	1.91
Intel Jones Farm, Hillsboro	1.41
ORNL CSB	1.25
T-Systems & Intel DC2020 Test Lab, Munich	1.24
Google	1.16
Leibniz Supercomputing Centre (LRZ)	1.15
Na h (NCAR)	1.10
Ya	1.08
Fa	1.07
Na NREL)	1.06

It's all about the "1".

1.06? Should we….
focus on driving the 0.06 down?
Or work on the 1.0?

"I am re-using waste heat from my data center on another part of my site and my PUE is 0.8!"

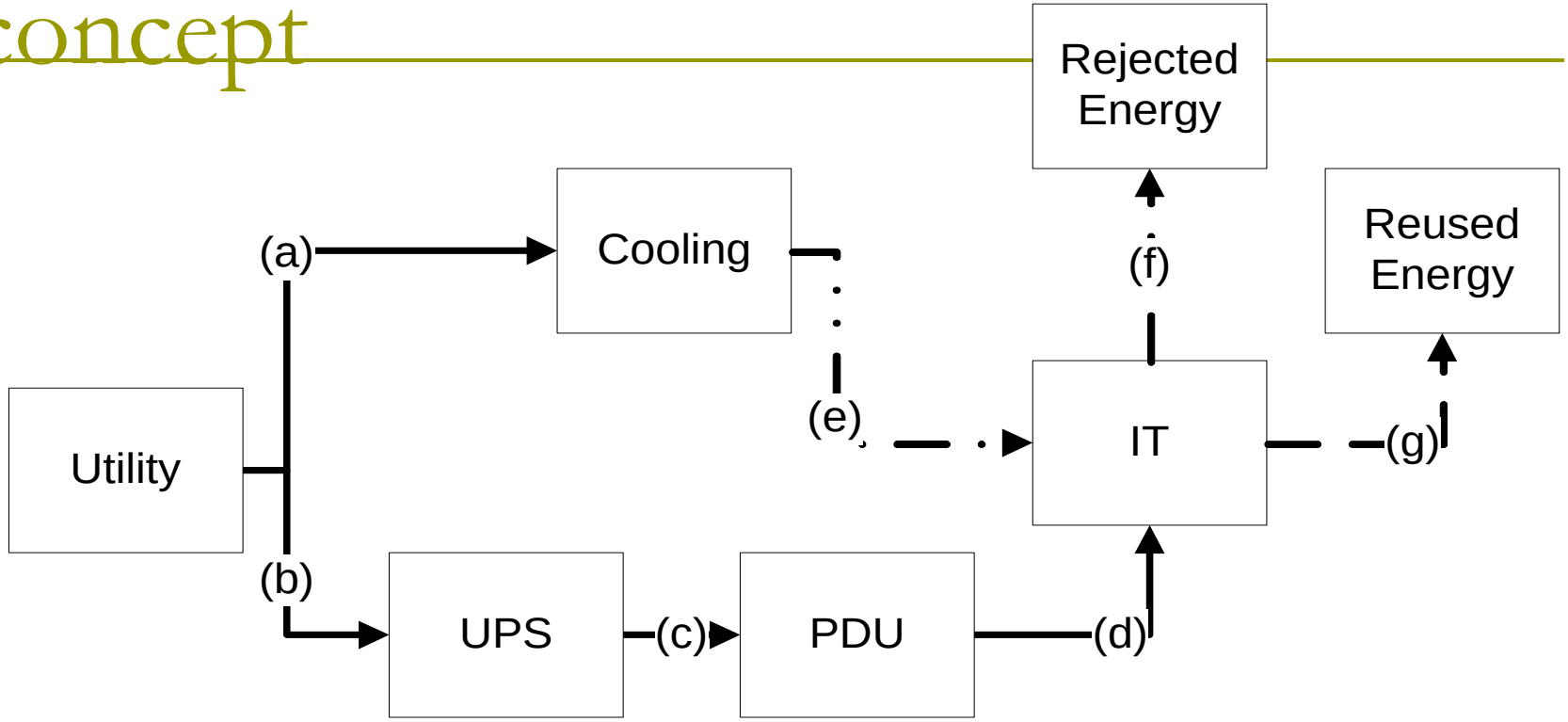
"I am reusing waste heat from my data center on another part of my site and my PUE is 0.8!"



- While re-using excess energy from the data center can be a good thing to do, it should not be rolled into PUE. The definition of PUE does not allow this.

- There is a new metric to do this; ERE

ERE – adds energy reuse to the PUE concept



$$ERE = \frac{Total\ Energy - Reuse\ Energy}{IT\ Energy}$$

$$= \frac{Cooling + PowerDistribution + Misc + IT - Reuse}{IT} = \frac{a + b - g}{d}$$

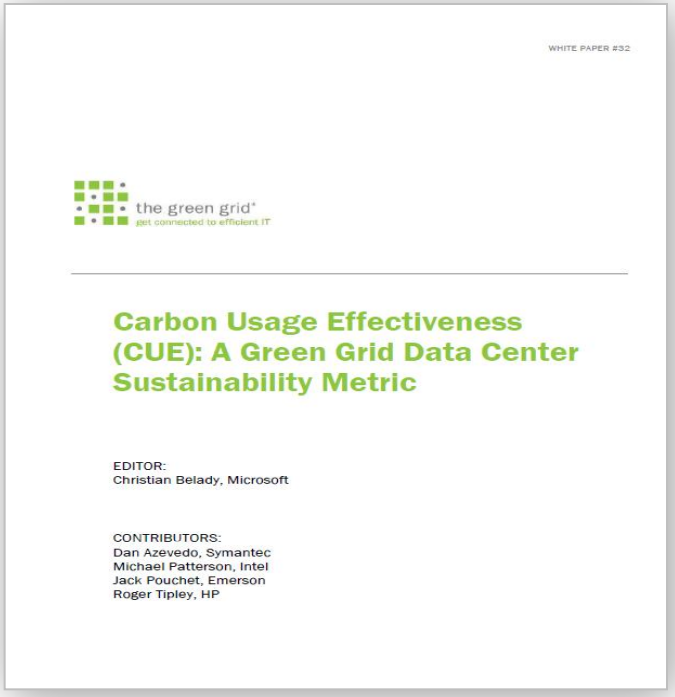
PUE & ERE resorted….

	PUE	Energy Reuse
EPA Energy Star Average	1.91	
Intel Jones Farm, Hillsboro	1.41	
T-Systems & Intel DC2020 Test Lab, Munich	1.24	
Google	1.16	
NCAR	1.10	
Yahoo, Lockport	1.08	
Facebook, Prineville	1.07	
Leibniz Supercomputing Centre (LRZ)	1.15	☑ ERE <1.0
National Renewable Energy Laboratory (NREL)	1.06	☑ ERE <1.0

Carbon Usage Effectiveness (CUE)

$$CUE = \frac{Total\ CO\ emissions\ caused\ by\ the\ Total\ Data\ Center\ Energy}{IT\ Energy}$$

- Ideal value is 0.0
- Example, the Nordic HPC Data Center in Iceland is powered by renewable energy – CUE ≃ 0.0



Thank you!