Gestion énergétique des data centers

October 4, 2022: Seminaire Master HPC

Luigi Brochard (Luigi.brochard@eas4dc.com), Energy Aware Solutions S.L.

The energy crisis

- Processors and servers consume more and more

- Electricity is becoming more expensive

- Carbon emissions need to be reduced

How to measure Power/Energy Efficiency

- PUE

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

  - **Power usage effectiveness** (**PUE**) is a measure of how efficiently a computer data center uses its power;
  - PUE is the ratio of total power used by a computer facility[1] to the power delivered to computing equipment.
  - PUE > 1; Ideal value is 1.0
  - It does not take into account how IT power can be optimised

- ITUE

$$ITUE = \frac{(\text{IT power} + \text{VR} + \text{PSU} + \text{Fan})}{\text{IT Power}}$$

  - **IT power effectiveness** (ITUE) measures how the node power can be optimised
  - ITUE >1; Ideal value if 1.0

- ERE

$$ERE = \frac{\text{Total Facility Power} - \text{Power}_{\text{reused}}}{\text{IT Equipment Power}}$$

  - **Energy Reuse Effectiveness** measures how efficient a data center reuses the power dissipated by the computer
  - ERE is the ratio of total amount of power used by a computer facility[1] to the power delivered to computing equipment.
  - If no Reuse, ERE = PUE, If all IT power is reused, ERE = PUE -1

Total Cost of Ownership: TCO

- $TCO = CAPEX + OPEX$

- $CAPEX = System\ acquisition\ and\ installation\ cost + Data\ Center\ installation\ cost$

where $Data\ Center\ installation\ cost$ includes the price to install or upgrade cooling equipment which have some importance in TCO

- $OPEX = Operational\ Cost + Energy\ Cost$

where $Operational\ Cost$ includes maintenance costs and floor space cost per sqm or sqt which have an impact in TCO when we factor in the density of servers

- $Energy\ Cost_{noreuse} = Total\ Facility\ Energy * Electricity\ Price$

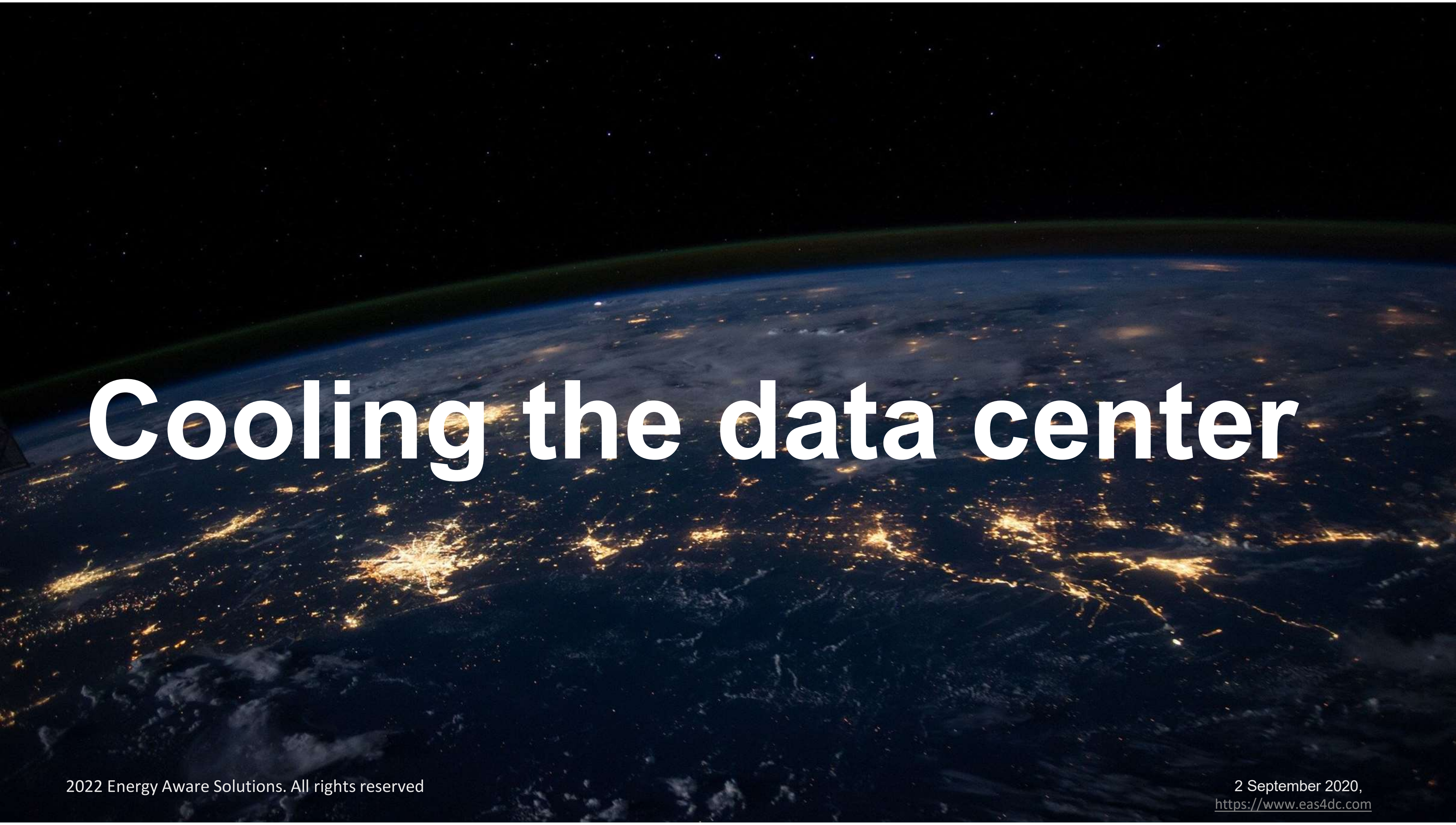
where $Energy\ Cost_{noreuse}$ is the Energy Cost when waste heat is not reuse, Total Energy is the amount of energy consumed by the computer facility over its life time and Electricity Price is the price of one kW/h.

Substituting PUE definition into equation above, we have :

- $Energy\ Cost_{noreuse} = IT\ Equipment\ Energy * PUE * Energy\ Price$

How to achieve Energy Efficiency ?

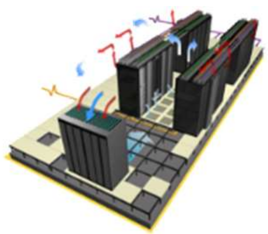
- Reducing the Cooling costs
  - Lower PUE
    - Better cooling technology
- Reducing the IT energy
  - More energy efficient servers (PSU, fans ….)
  - Higher GFlops/watt processor
  - Better Algorithm or Software to reduce the application power/energy
- Reusing waste heat energy
  - Lower ERE
    - Heat reuse



Cooling the data center

DataCenter Cooling Technologies

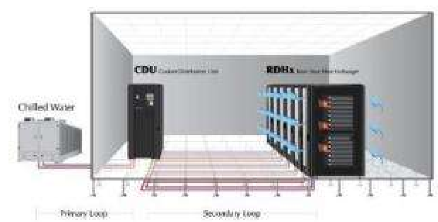
Air Cooled



- Standard air cooled systems with compressor chillers
- Fits in any datacenter
- Maximum flexibility
- Hot-Aisle/Cold-Aisle

PUE ~ 2-1.5

Rack Level Heat Exchangers



- Air cooled systems + RDHX
- Uses chilled water with economizer
- Close coupled aisle solutions
- Enables dense rack placement

PUE ~1.3

Direct Water Cooled

- Direct water cooled systems
- Higher watt/cm2
- Extreme energy efficiency & reuse
- Denser footprint
- Lower OPEX/CAPEX

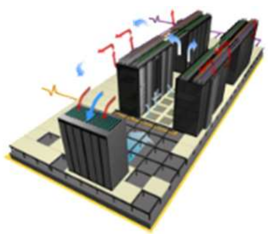
PUE <= 1.1

Servers have fans	Servers have no fans

Enterprise DataCenter ← → **HPC&AI DataCenter**

At what temperatures different coolings operate

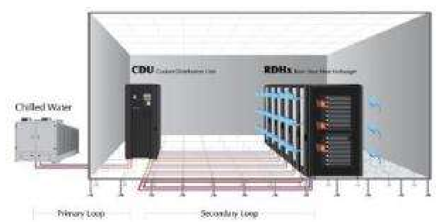
Air Cooled



- Standard air cooled systems
- Chilled Water
- Between 8° and 14°C

PUE ~ 2-1.5

Rack Level Heat Exchangers



- Air cooled systems + RDHX
- Chilled Water
- Between 12° and 20°C

PUE ~1.3

Direct Water Cooled

- Direct water cooled systems
- Chilled to Warm/Hot Water
- Between 18° and 45°C

PUE <= 1.1

Servers have fans	Servers have no fans

Enterprise DataCenter — **HPC&AI DataCenter**

Electricity Prices vs Servers Prices

After how many years does the electricity cost equal the server cost ?

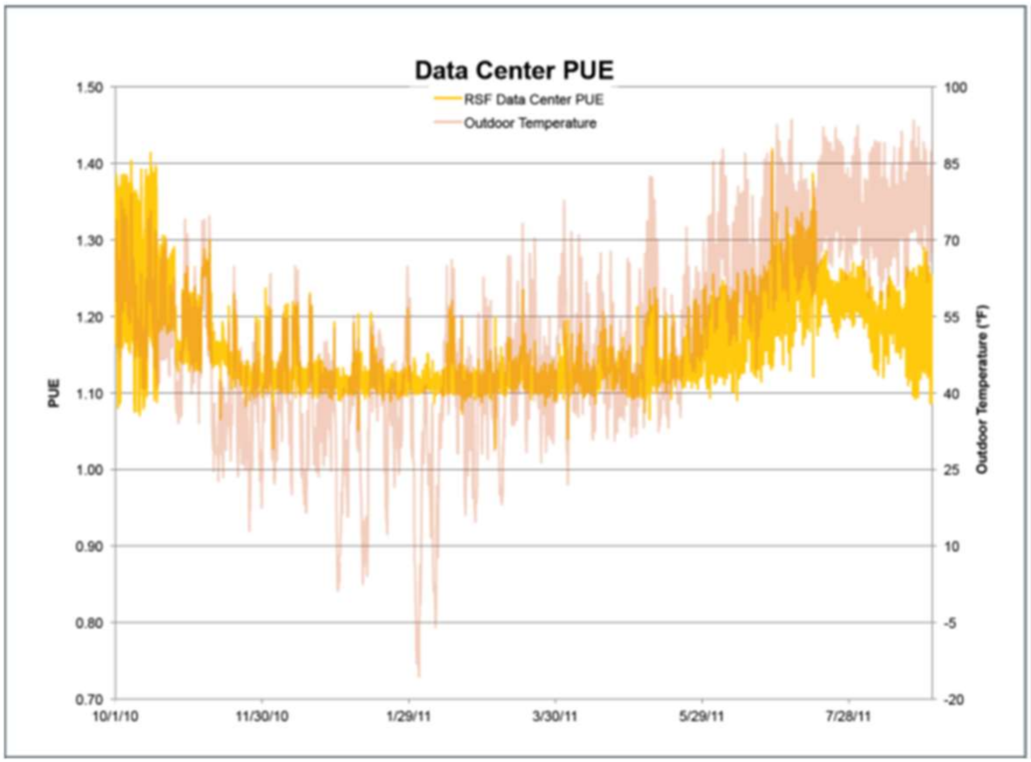
- With a PUE of 2.0
  - with 0.3 $/KWh => 2.1 years
  - with 0.2 $/KWh => 3.2 years
  - with 0.1 $/KWh => 6.4 years
- With a PUE of 1. 1
  - with 0.3 $/KWh => 3.9 years
  - with 0.2 $/KWh => 5.8 years
  - with 0.1 $/KWh => 11.6 years

What is the benefit of an increased temperature ?

- Higher temperature reduces the cost to cool air or water
  - Less electricity => reduced OPEX
  - Less chillers => potential reduced CAPEX
- Higher temperature can lead
  - Free cooling => No chillers => reduced OPEX and CAPEX
  - Heat reuse

Example of free cooling with air and RDHX

- National Rewable Energy Laboratory (NREL) in Colorado: PUE = 1.16



70°F = 21°C

40°F = 4°C

RSF hourly PUE over the first 11 months   free cooling 31% of the year   average PUE = 1.16

Water vs. Air Heat Capacity and Thermal Resistance

Water

1. High heat capacity

$$c_V \approx 1 \text{ Wh/(L}\cdot\text{K)}$$

2. Low thermal resistance

$$\Delta T = R_{th} \cdot \dot{q}''$$

$$R_{th} = 0.1 \text{ K cm}^2/\text{W}$$

$$\dot{q}'' = 50-100 \text{ W/cm}^2$$

$\Delta T \sim 5 - 10$ K

Air

1. Low heat capacity

$$c_V \approx 0.0003 \text{ Wh/(L}\cdot\text{K)}$$

2. High thermal resistance

$$\Delta T = R_{th} \cdot \dot{q}''$$

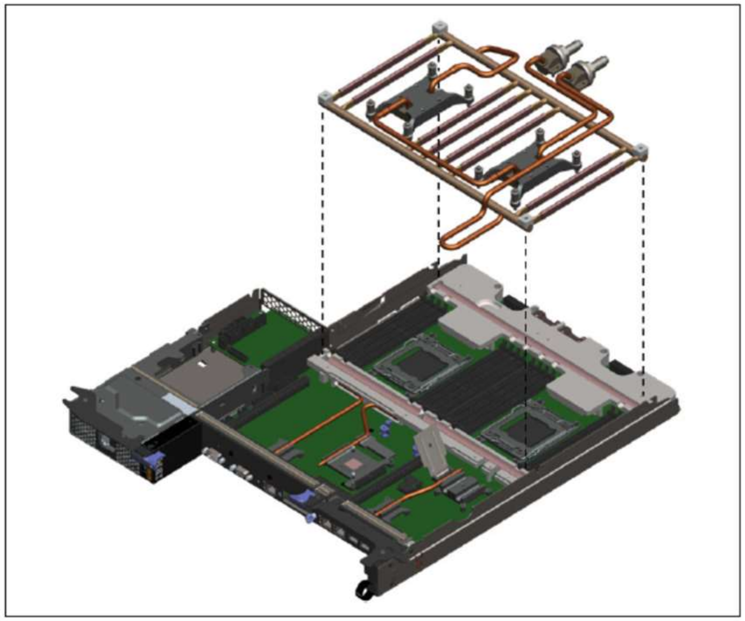
$$R_{th} = 1 \text{ K cm}^2/\text{W}$$

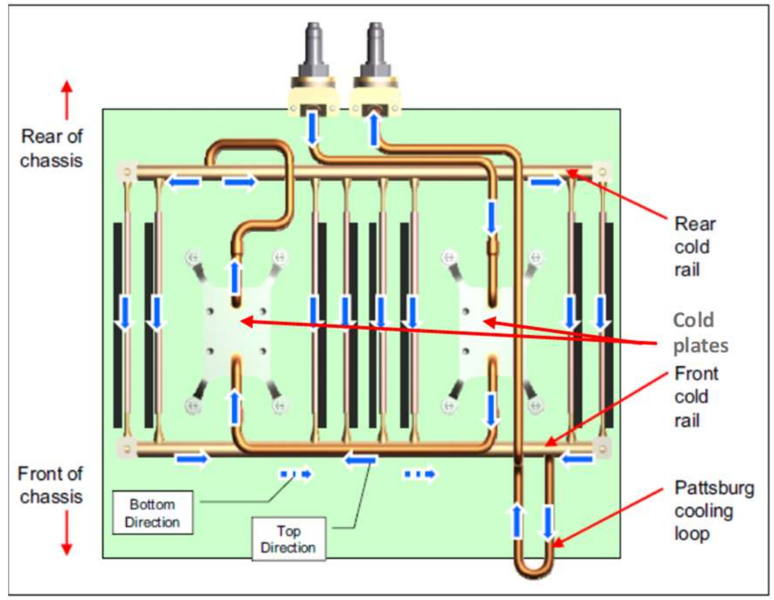
$$\dot{q}'' = 50-100 \text{ W/cm}^2$$

$\Delta T \sim 50 - 100$ K

Where cv is the heat capacity, q" is the heat flux , $\Delta T = R_{th} \cdot \dot{q}''$ is the 1D- representation of the heat flux equation from thermodynamics. The consequence is that water needs a much smaller delta between the processor temperature and the coolant temperature than air.

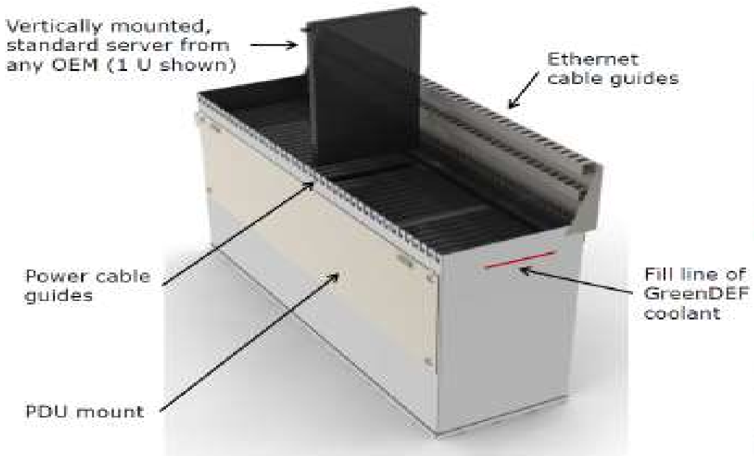
Example of a water cooled dense server



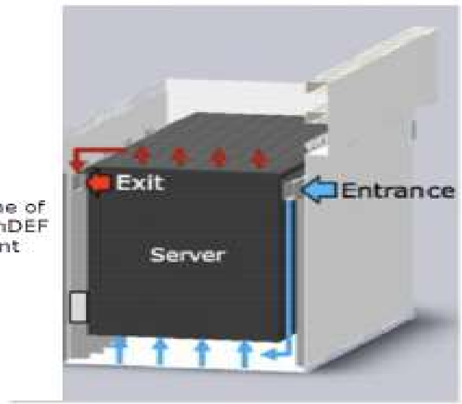


Water-Cooled IBM iDataPlex dx360 M3, 2012

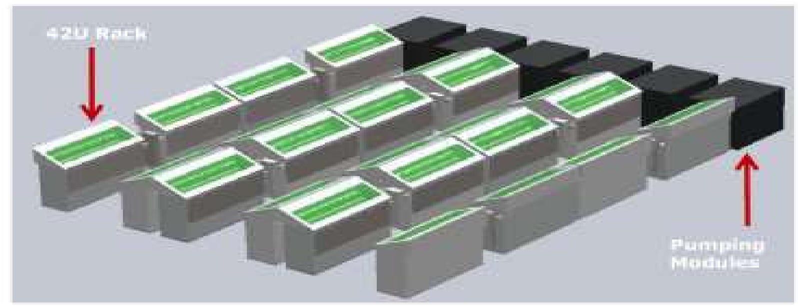
Green Revolution Direct Oil Immersion Cooling



Side View Cutaway



CarnotJet™ Example 24 Rack Install
(24-38 sq. feet per rack)



Equipment in Empty Tank

Server Prep For Immersion:

- Remove Fans
- Allow Server to run without fans
- Remove/replace thermal interfaces

Benefits:

- No Chiller (Very Low PUE)
- No Fan Power
- Can create waste water at 50C for reuse

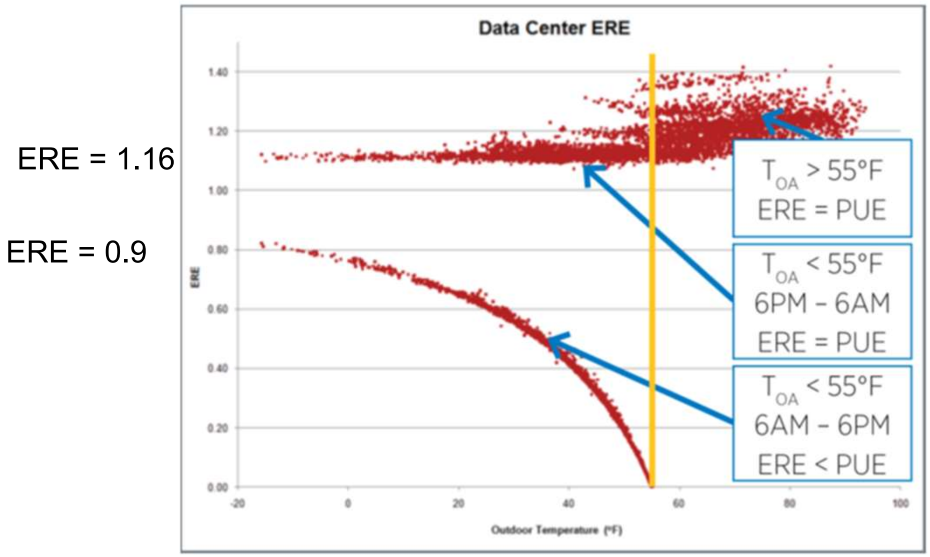

Reusing waste heat
2022 Energy Aware Solutions. All rights reserved
2 September 2020,
https://www.eas4dc.com

Waste heat reuse

- Reuse the heat to heat a building, a swimming pool
- Reuse the heat to cool a liquid and produce cold water

Example of waste heat reuse with air and RDHX

- National Rewable Energy Laboratory (NREL) in Colorado:



RSF ERE as a function of outdoor air temperature (TOA)

Reusing Waste H… duce Cold Water: CoolMUC2

TOP500
366 TFlop/s HPL Performance
Rank 262
(November 2015)

- **Lenovo NeXtScale Water Cool Technology (WCT) system**
  - ✓ Water inlet temperatures 50 °C
  - ✓ All season chiller-less cooling
  - ✓ 384 compute nodes
  - ✓ 466 teraflop/s peak performance

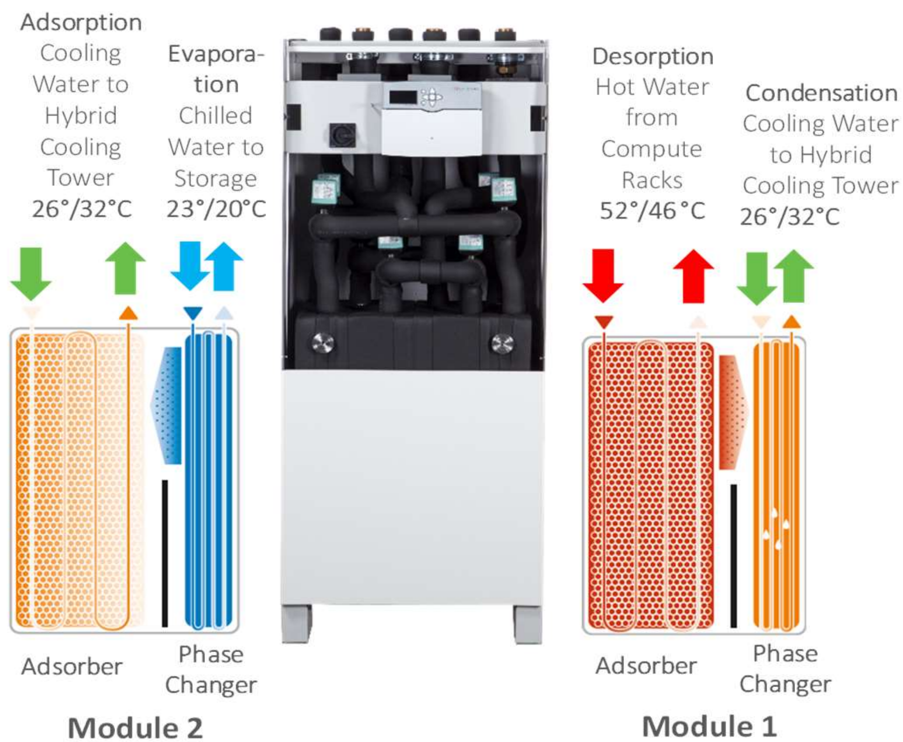
- **SorTech Absorbtion Chillers**
  - ✓ based of zeolite coated metal fiber heat exchangers
  - ✓ a factor 3 higher than current chillers based on silica gel
  - ✓ COP = 60%
  - ✓ Total electricity reduced by ~60%

ERE = 0.3

$$\text{ERE} = \frac{\text{Total Facility Power} - \text{Treuse}}{\text{IT Equipment Power}}$$

Leibniz Supercomputing Centre

Adsorption Chillers

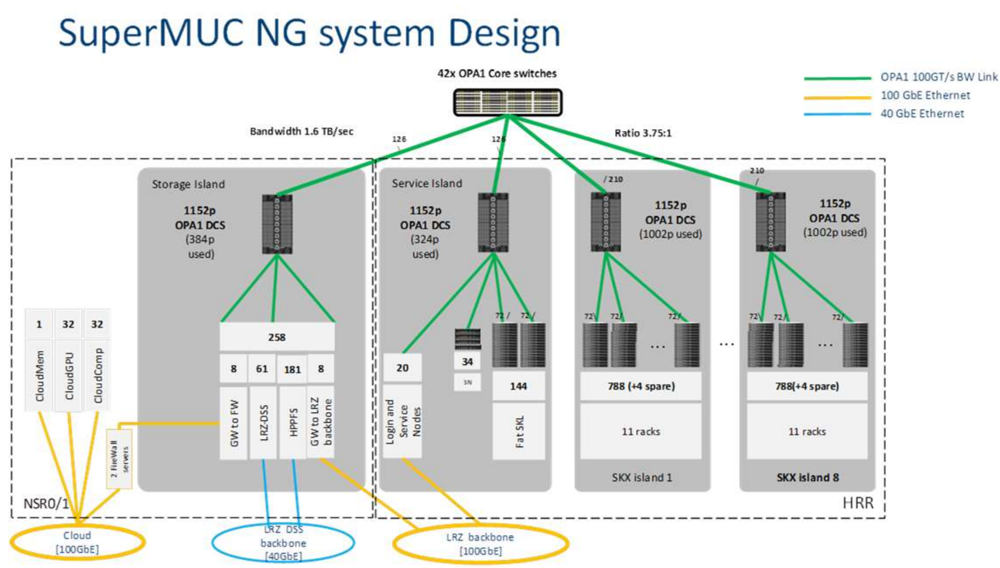


Adsorption chiller consists of two identical vacuum containers, each containing two heat exchangers:
Absorber: Coated with the adsorbent (silica gel or zeolite)
Phase Changer: Evaporation and condensation of water
During desorption (module 1) the adsorbent is heated up causing the previously adsorbed water vapor to flow to the condenser (red arrow), where it is condensed to liquid water.
During adsorption (module 2) the adsorbent is cooled down again causing water vapor to flow back (blue arrow) and evaporate in the evaporator generating cold. Water is evaporated at low temperatures, because the system is evacuated and hermetically sealed to the surroundings.

SuperMUC NG system at LRZ: number 8 on Top500, Nov 2018

Phase 1

- Based on Xeon Skylake
  - 6334 Nodes with 2 Intel SKL @205 W CPUs
  - HPL ~ 20 PetaFLOP/s
  - OPA island based Interconnect
  - Large File Space on IBM Spectrum Scale
    - Scratch : 51 PB, 500GigaByte/s IOR bw
    - …
- Energy Effective Computing
  - More efficienct Hot Water Cooling
  - Dynamic Energy Aware Run time
  - Waste Heat Reuse
- Best TCO and Energy Efficiency
  - overall estimated PUE ~1.08 and ERE = 0.7



Higher energy efficient processors

Flops per cycle across Xeon generations

Microarchitecture	Instruction Set	register lenght	FP execution units	SP Flops / cycle	DP Flops / cycle	DP FMA	DP ADD
Skylake	AVX512 & FMA	512	2 FP FMA 512	64	32	32	16
Haswell/Broadwell	AVX2 & FMA	256	2 FP FMA 256	32	16	16	8
Sandybridge	AVX	256	2 FP 256	16	8	16	8
Nehalem	SSE	128	2 FP 128	8	4	8	4

512 / 64 = 8
2 FP 512 = 16
2 FP FMA 512 = 32

Measured GFlops & GFlops/Watt on Xeon 6148

Xeon 6148; 2.4 GHz	Instruction set	DP Add	DP Mult.	DP FMA
GFlops	SSE2	382	305	763
GFlops	AVX2	762	763	1525
GFlops	AVX-512	1396	1400	2791

Xeon 6148; 2.401 GHz	Instruction set	DP Add	DP Mult.	DP FMA
GFlops	SSE2	492	407	984
GFlops	AVX2	828	828	1652
GFlops	AVX-512	1399	1397	2797

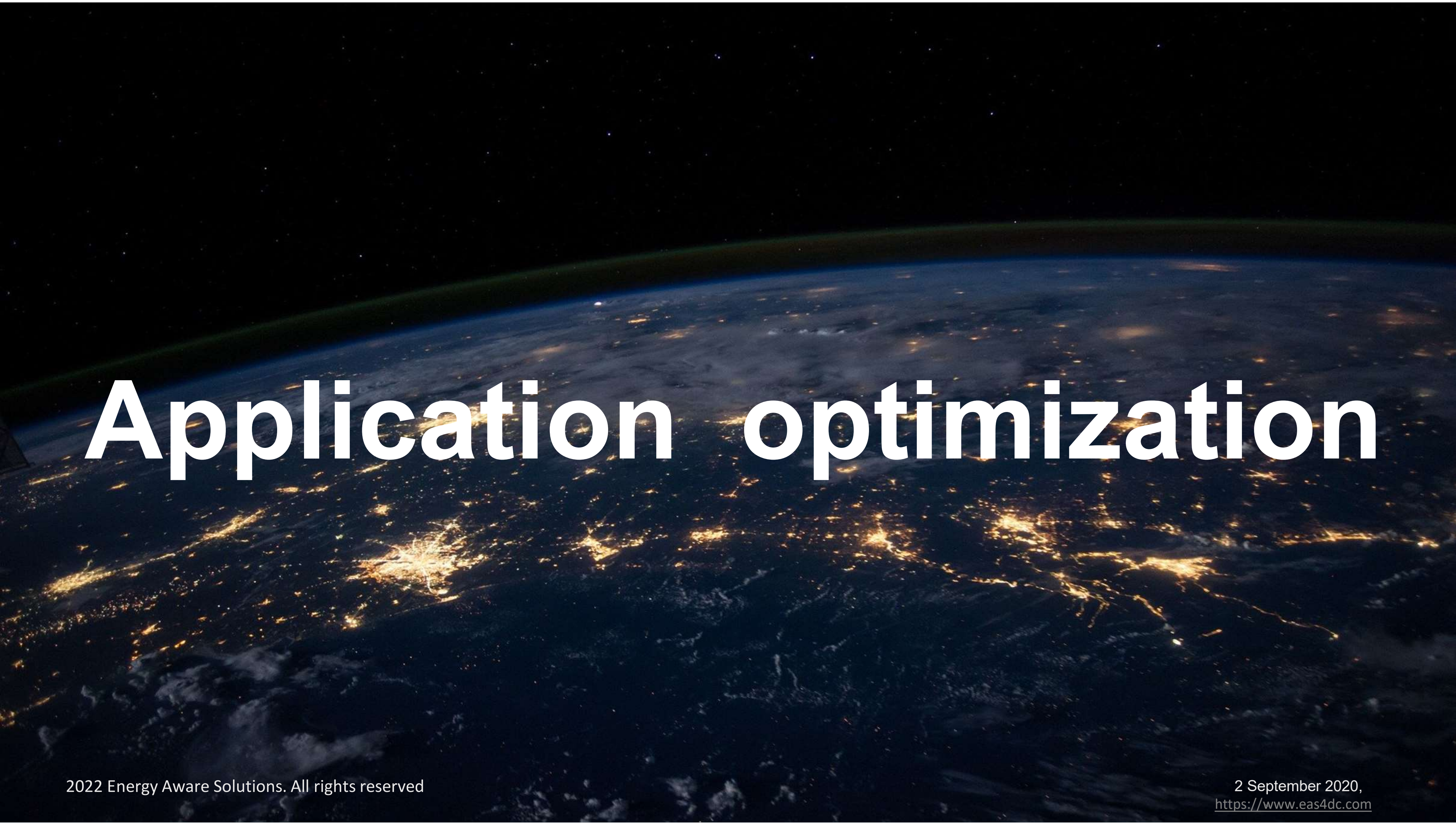
Xeon 6148; 2.4 GHz	Instruction set	DP Add	DP Mult.	DP FMA
GFlops/Watt	SSE2	1,54	1,33	3,01
GFlops/Watt	AVX2	2,86	2,93	5,67
GFlops/Watt	AVX-512	5,23	5,24	10,45

Xeon 6148; 2.401 GHz	Instruction set	DP Add	DP Mult.	DP FMA
GFlops/Watt	SSE2	1,53	1,37	3,00
GFlops/Watt	AVX2	2,90	2,93	5,65
GFlops/Watt	AVX-512	5,24	5,24	10,09

DP Add with 2 AVX512 is 16 Flops/cycle per core
Xeon 6148 has 20 cores / processor * 2 processors
Its nominal frequency is 2.4 GHz … but
its base AVX512 frequency with 20 cores loaded is 2.2 GHz
16* 2.2 GHz= 35.2 GFlops ; 35.2 * 40 = 1408 Gflops
This DP Add loop is reaching 99 % of peak @ 2.2 GHz,
And only 91 % of peak @ nominal

CPU and GPU peak performance and performance per Watt

	TDP (Watt)	SP Tflops	SP Gflops/W	Tensor Tflops	Tensor Gflops/W
Intel Skylake 8180	205	4.5	21.9	NA	NA
NVIDIA Volta V100	300	14.9	49.6	125	416.7



Application optimization

Application performance and energy optimization

- Recompile/rewrite the algorithm to improve performance and perf/watt
  - Not easy
  - Need skills and tools
- Tune the cpu/gpu frequencies at run time to improve the performance/watt
  - No application modification,
  - Need tools

Energy optimization with EAR runtime

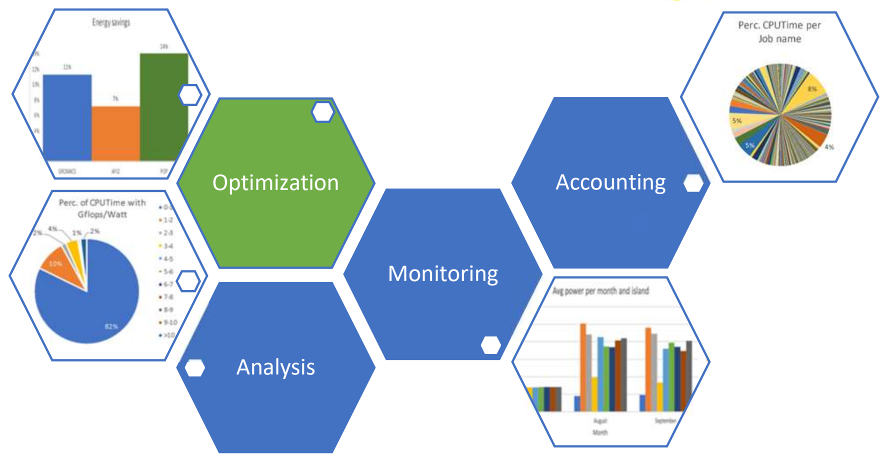
2 September 2020,
https://www.eas4dc.com

Energy Aware Runtime and Energy Aware Solutions

- EAR is an open source energy management software developed by BSC through a BSC-Lenovo collaboration since 2016
  - EAR documentation is available from BSC web site
    - https://gitlab.bsc.es/ear_team/ear/-/wikis/home
- EAS is a spin-off of BSC created by EAR authors in September 2020
- EAS provides **services** to control/reduce **data center energy** through EAR
  - EAR installation/training/support
  - **Energy optimization** and **analysis services**

EAR main features and values

- **Monitoring & Accounting**
  - **System** monitoring
    - **Nodes** temperature, power**,** effective frequency …
    - Automatic **reporting** of run time **hardware issues**
  - **Application**
    - Performance metrics **monitoring** at **job/loop level**
    - Energy **accounting**
      - Granularity: jobid, stepid, user, node
      - Scope: Application average and at run time
    - **To be used for application** analysis and optimization
- **Optimization**
  - Runtime **application energy optimization**
    - **Transparent**, **dynamic** and **lightweight runtime** library with **no user intervention** required
    - **Automatic energy savings** according to energy policies
  - **Cluster Energy** and **Power Capping**



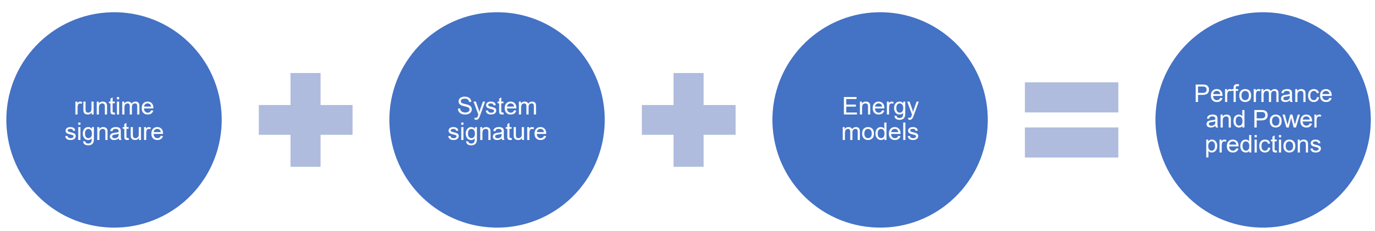
- In **production** at LRZ SuperMUC_NG **6480 nodes** since **August 2019** and at Surf SARA Snellius **576 AMD and 23 Intel + NVIDA nodes** since **February 2022**

JC14

Signatures: Application and System

JC1

- Application signature is a set of metrics computed at runtime by EAR. Describe application basic characteristics:
  - CPI: Cycles per Instruction
  - GBs: Main memory bandwidth
  - Node Power , Iteration time , %AVX512 instructions, %MPI, Input/Ouput (MBs)
  - GPU metrics
  - used also for the application classification
- System signature is a set of metrics to describe the hardware characteristics regarding power and performance
- Both signatures are used in the energy models

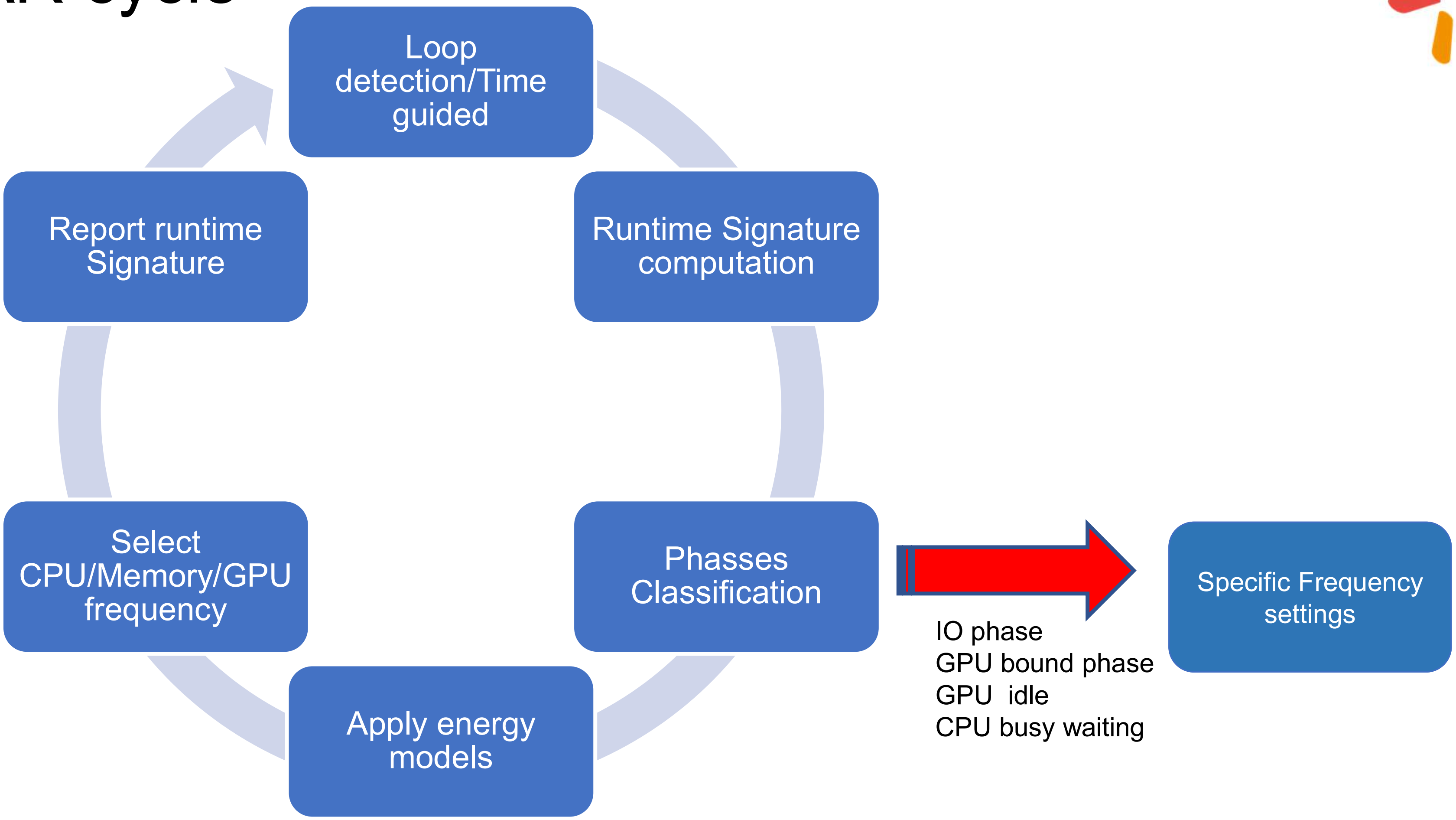


Slide 30

JC13 Julita Corbalan; 18/08/2022

JC14 I would remove this picture
Julita Corbalan; 18/08/2022

EAR cycle


Loop detection/Time guided
Runtime Signature computation
Phasses Classification
Apply energy models
Select CPU/Memory/GPU frequency
Report runtime Signature
IO phase
GPU bound phase
GPU idle
CPU busy waiting
Specific Frequency settings

What to do for an existing data center ?

JC1

01 Be energy aware

- Measure your system power consumption
- Measure the applications energy and performance

02 Be energy efficient

- Optimize and control your system energy consumption
- Minimize your electricity bill and carbon footprint

Slide 32

JC1 I will remove the figures here and in the previous, slide

Julita Corbalan; 17/08/2022

What does EAS propose ?

01 **Be energy aware**

- EAR Energy Detective
- EAR Energy Detective Pro

02 **Be energy efficient**

- EAR Energy Optimizer
- EAR Energy Optimizer Pro

EAS offers

- **Energy Detective:**
  - EAR node/cluster monitoring and basic job energy accounting
  - Installation and training remote
  - Installation of new EAR versions
- **Energy Detective Pro:**
  - EAR Detective + advanced job energy and performance accounting
- **Energy Optimizer:**
  - Detective Pro +
  - **Energy job optimization**
  - **Cluster energy monitoring**
- **Energy Optimizer Pro**
  - Optimizer +
  - **Power capping**
  - **Energy capping**

Example 1 : Average metrics (CPU only)

This a very well tuned application using MPI

Skylake

- Low CPI, Mid-high GFlops, low MPI percentage.
- Avg DC Node Power: 334 W.
- Energy efficiency (Gflops/Watts) = 0,23.
- Energy = 1.148.626 J

CPU Freq (Ghz)	CPI	GFLOPS	%MPI
2.28	0.44	78.77	7%
MEM Freq (Ghz)	GBS (GB/s)	IO MBS (MB/s)	Time (s)
2,4	52.75	0.24	3439

Remember that Skylake DP Add (or Mult) is 5.24 GFlops/Watt

Icelake

CPU Freq (Ghz)	CPI	GFLOPS	%MPI
2.37	0.41	192	8%
MEM Freq (Ghz)	GBS (GB/s)	IO MBS (MB/s)	Time (s)
2,18	112	0.6	1420

- Low CPI, Mid-high GFlops, low MPI percentage and high memory bandwidth.
- Same per-process memory bandwidth
- Avg DC Node Power: 678 W.
- Energy = 962.760 J
- Energy efficiency (Gflops/Watts) = 0.28 (higher Energy-efficiency than in Skylake)

Example 2: : Average metrics (CPU only)

This a Python application with no MPI

Skylake

CPU Freq (Ghz)	CPI	GFLOPS	%MPI
2.29	1.71	0.04	0%
MEM Freq (Ghz)	GBS (GB/s)	IO MBS (MB/s)	Time (s)
2,39	1.4	13.87	3924

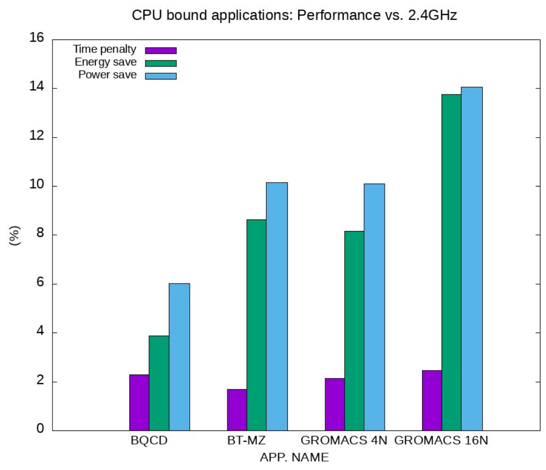
- High CPI, very low GFlops. Very low GB/s. Some IO but low values.
- Node Power: 178 W.
- Energy efficiency (Gflops/Watt) = 0,00022
- Energy = 698.472 J

Icelake

CPU Freq (Ghz)	CPI	GFLOPS	%MPI
2.37	1.22	0.05	0%
MEM Freq (Ghz)	GBS (GB/s)	IO MBS (MB/s)	Time (s)
2,18	0.95	20.56	2647

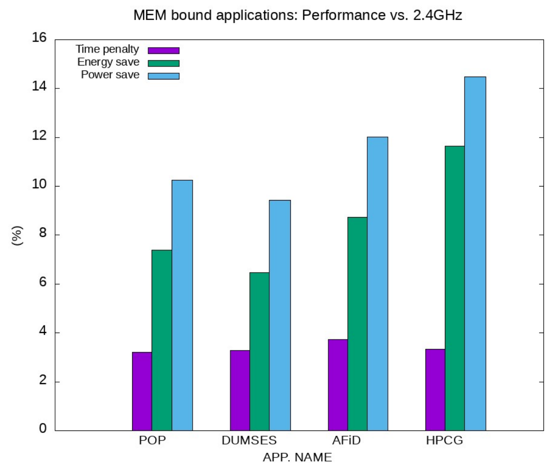
- High CPI, very low GFlops and Memory bandwidth.
- Node Power: 340 W.
- Energy efficiency = 0,00014
- Energy = 899.980 J

EAR energy optimization results on Intel & NVDIA



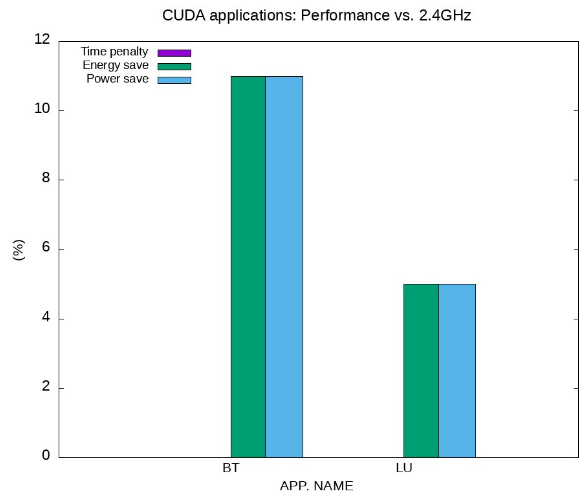
Compute bound applications

- Reducing UNC freq. when not needed (BT-MZ)
- Reducing CPU freq. in AVX512 apps (GROMCACS)



Memory bound applications

- Reducing CPU freq. for memory bound apps.
- Reducing UNC freq. when not needed (HW is too conservative)



CUDA applications

- Reducing CPU&UNC freqs. during busy waiting periods

EAS installations

- **LRZ SuperMUC_NG, Germany**
  - 6480 nodes Intel Skylake since 2019
  - Energy Optimizer Pro
- **SURF Snellius, Netherlands**
  - 36 nodes Intel Icelake with 4 NVIDIA 100 GPUs
  - 576 nodes AMD Rome
  - Energy Optimizer
- **CentraleSupelec, France**
  - 180 nodes Intel Skylake
  - Energy Detective Pro
- **Institut de Physique du Globe de Paris (IPGP), France**
  - 60 nodes AMD Rome + 4 Intel Skylake nodes with NVIDIA A100 GPUs
  - Energy Detective Pro
- **Bordeaux University, France**
  - 340 Intel Skylake nodes
  - Energy Detective Pro

EAS installations

- **POC en cours**
  - EDF/DER
    - Cronos : 1880 Intel Cascade lake nodes , 115 Cascade lake + NVIDIA Quatro and V100
  - CERFACS
    - Kraken : Intel Skylake, Icelake and NVIDIA A30
- **Installations à venir en 2023**
  - Marenostrum 5 à BSC, phases 1 et 2
  - Phase 2 SURF Snellius with AMD Genoa
  - Phase 2 LRZ SuperMUC_NG with Intel SPR and PVC
  - LRZ Innovation Partnership for ExaMUC

EAS propose des stages

JC16

- EAR et EAS sont au coeur des problemes environementaux d'aujourd'hui
- EAS est en pleine croissance et a besoin de talents
- Nous proposons des stages:
  - en remote ou sur le campus du Barcelona Supercomputing Center à Barcelone
  - en collaboration avec l'équipe de BSC et EAS qui développe EAR
  - renseignement et candidature
    - luigi.brochard@eas4dc.com

Slide 40

JC16 It's the same for all the packages
Julita Corbalan; 17/08/2022

EAS
Energy Aware Solutions

For more information:

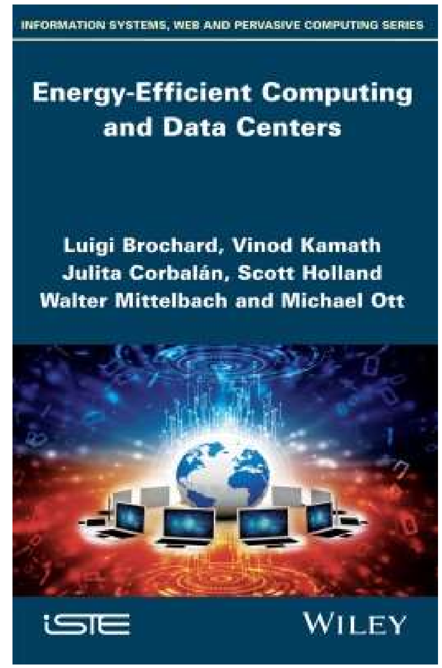
Luigi Brochard (luigi.brochard@eas4dc.com), Energy Aware Solutions S.L.

Contact contact@eas4dc.com

Browse www.eas4dc.com

Read the book : https://onlinelibrary.wiley.com/doi/book/10.1002/9781119422037

Read the article : https://www.techniques-ingenieur.fr/base-documentaire/energies-th4/energie-economie-et-environnement-42593210/vers-des-data-centers-zero-emission-et-autonomes-en-energie-be6003/



www.eas4dc.com