Sustainable Supercomputing

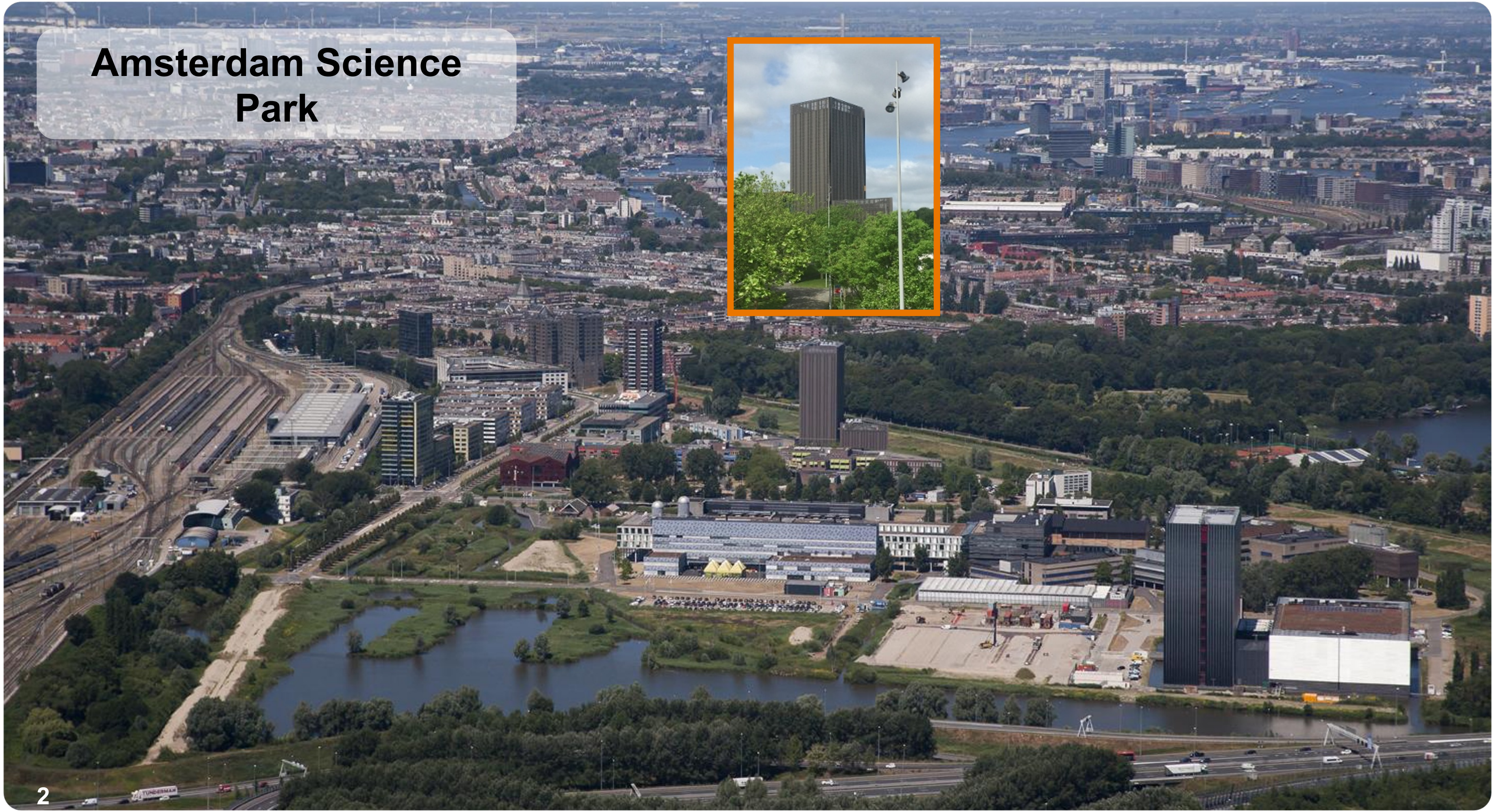
Journey from Idea to Production : Introducing green computing strategies for the National Supercomputer "Snellius"

Energy Efficient Computing Workshop

15th June, 2023

Sagar Dolas, Ben Czaja, Alessio Sclocco, Ana-Lucia Varbensco, Xavier Alvarez Farre, Carlos Teijeiro Barjas, Reinder Radersma

Amsterdam Science Park



SURF

Reliable, secure and innovative ICT infrastructure

Digital innovation and transformation of education and research

Knowledge exchange and trainings

Services development and integration with EU initiates

Academic Medical Centres

SURF Members

100+ and growing

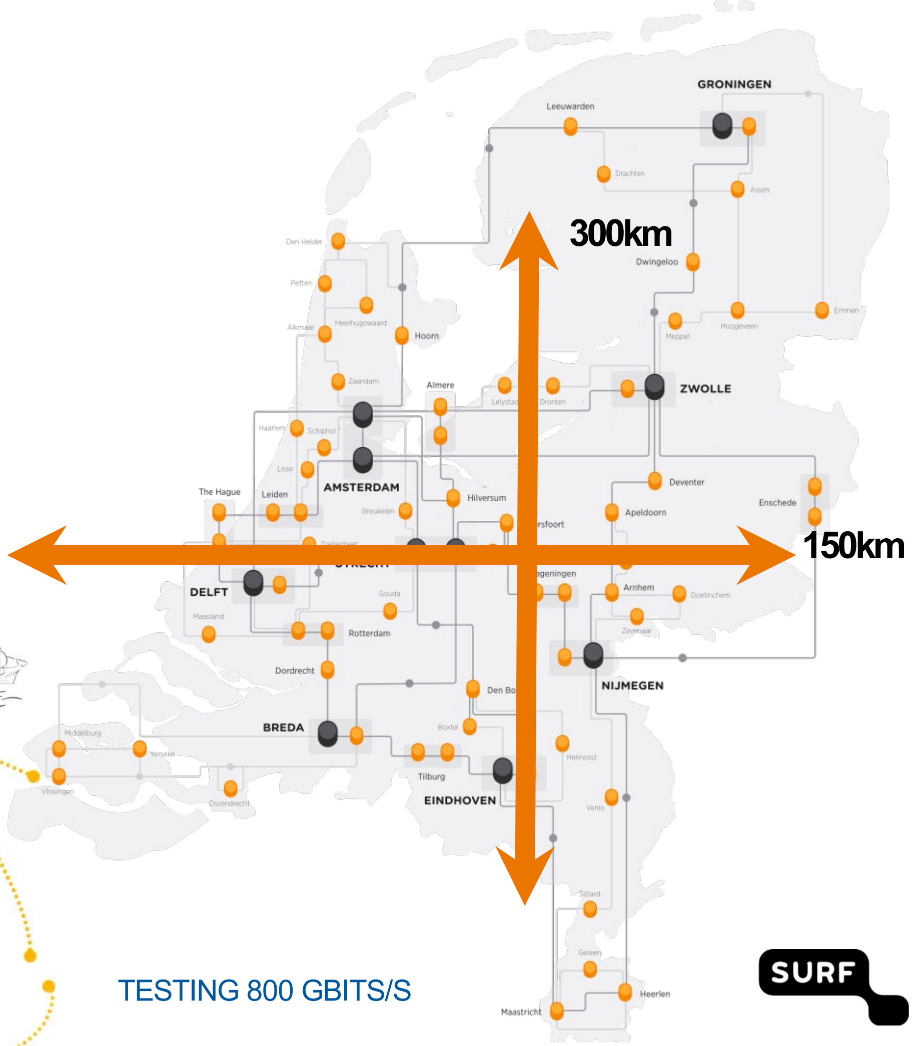
Universities for applied sciences

(+ Senior secondary vocational education institutions)

Universities

& other research institutes

SURF Network



TESTING 800 GBITS/S

SURF Research services

High bandwidth network
Dedicated light paths
Encrypted transfers
Cyber Security
AAI / SSO

> 400 Gbps

Snellius
Lisa
Data Processing
Custom Cloud Solutions
SURF Research Cloud
Public Cloud
Experimental zone

> 100.000 cores

Ceph
dCache
Storage DPU
Object Store
Preservation
EPIC PID
iRODs

> 150 PB

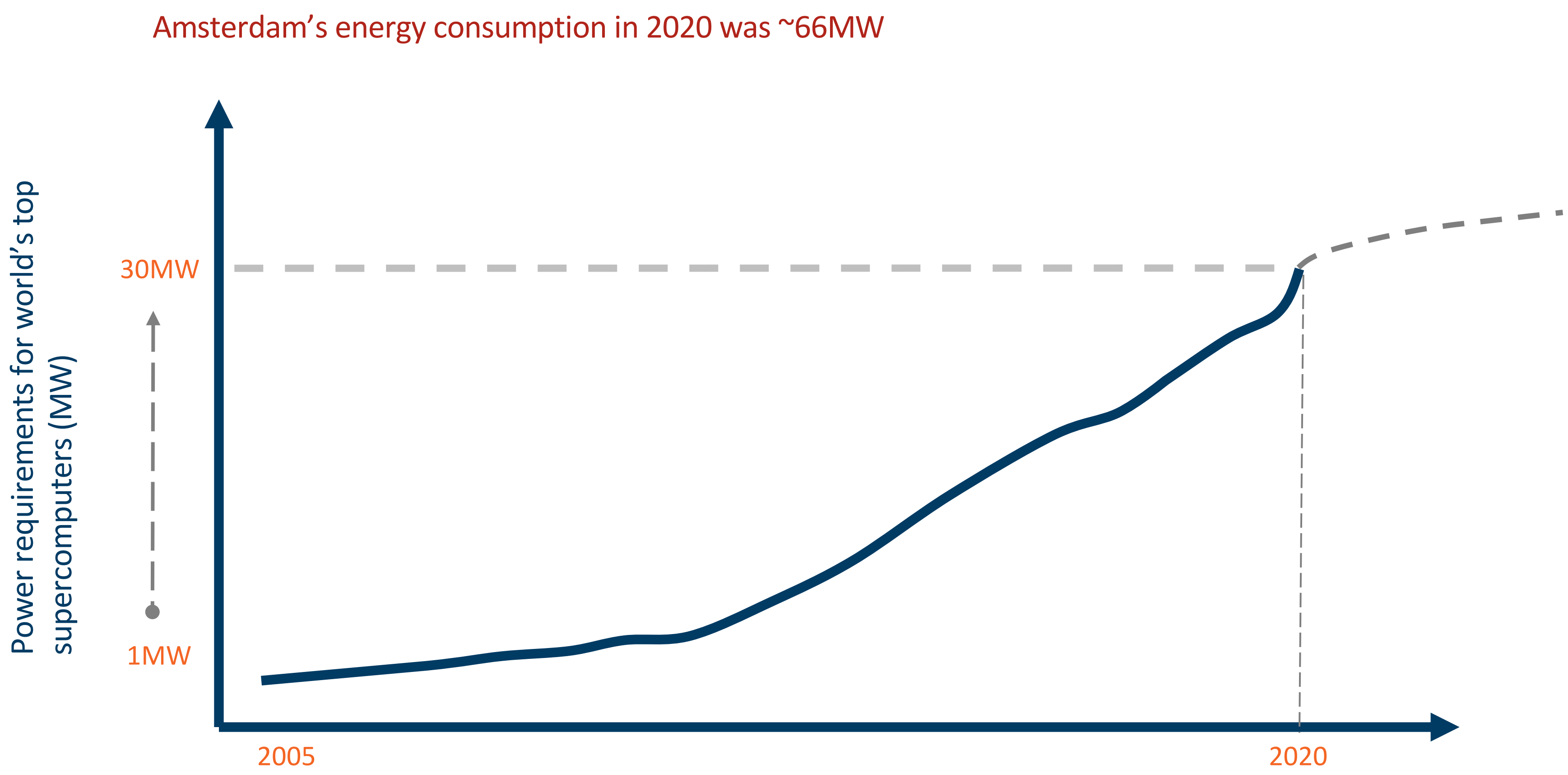
Support
Training
Visualisation
Optimisation
Expertise
Collaboration

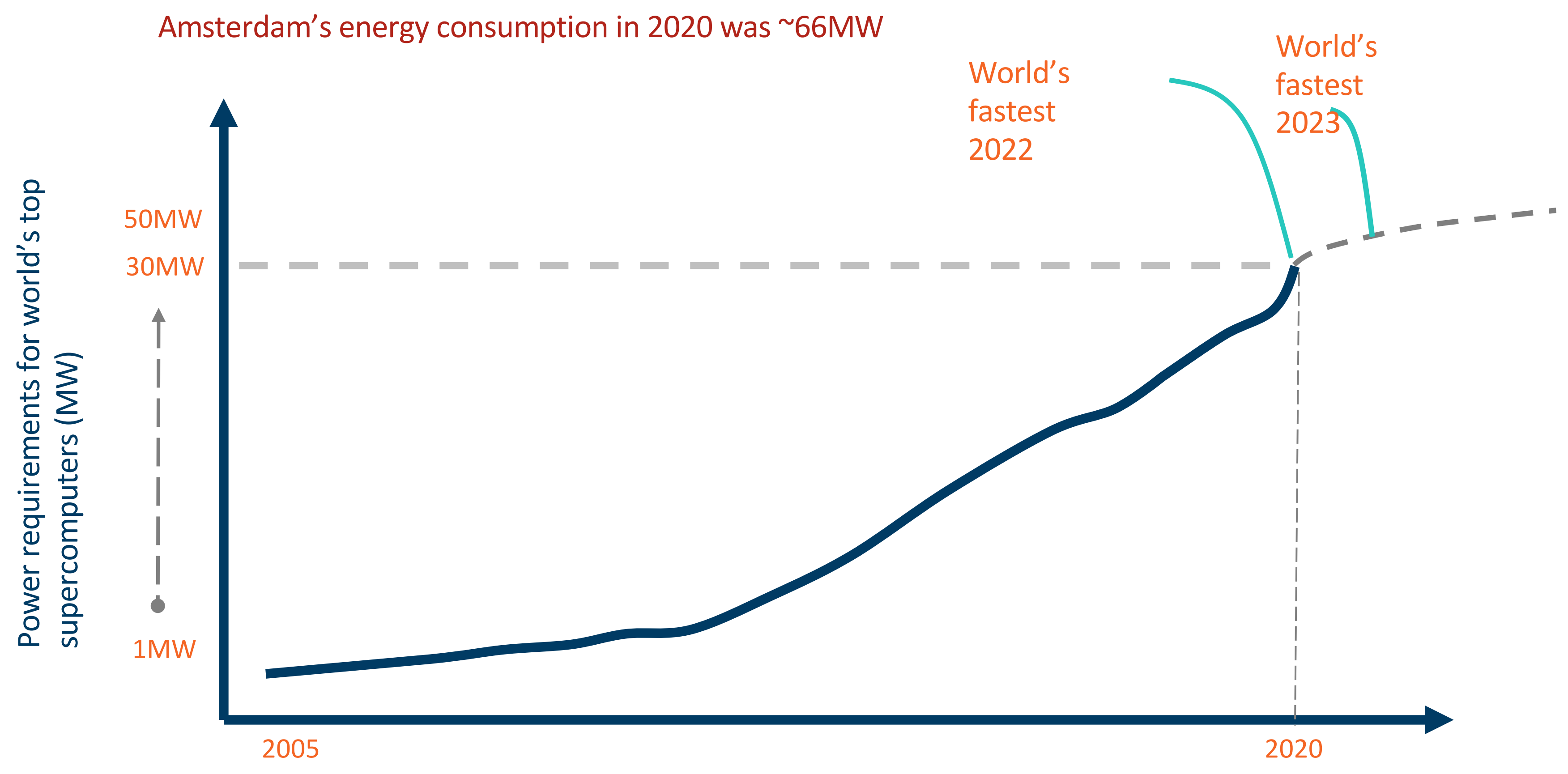
NL

Why we care ?

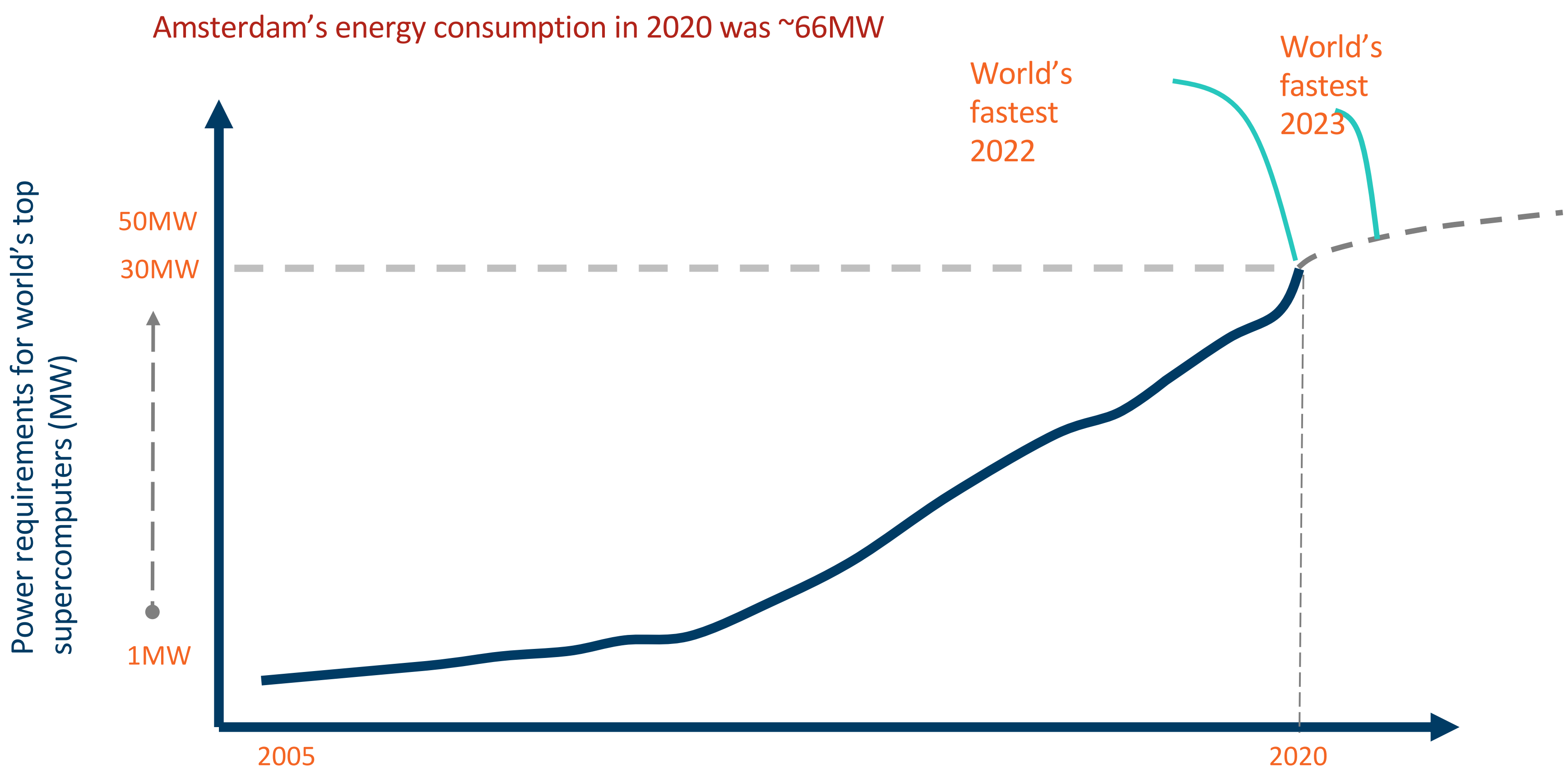
1. For **each of us** : Uphold value of energy sustainability
2. For our **partners** : Contribute to energy sustainability goals
3. For **society** : Alignment with United nation goals for Innovation, development of sustainable communities and mindful consumption of resources

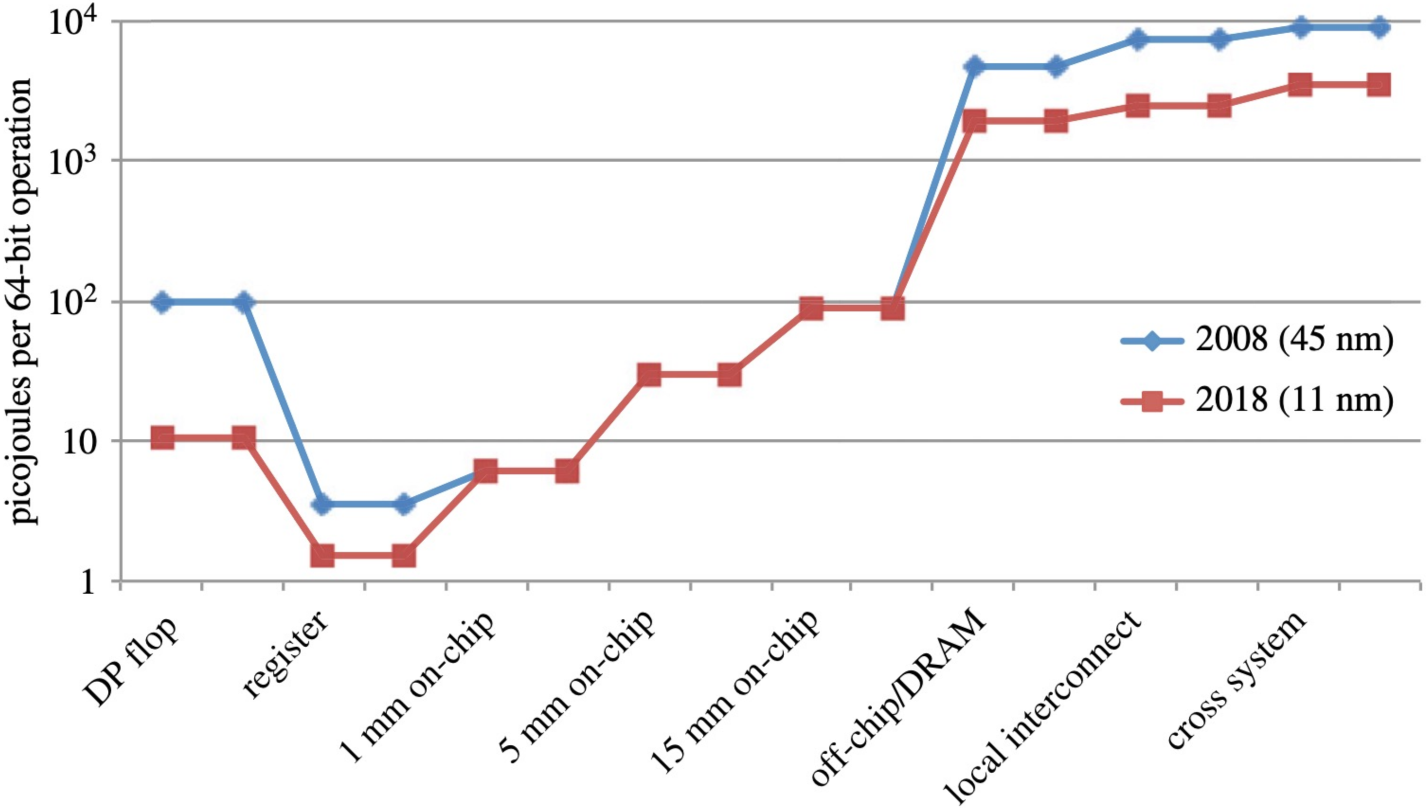
Why we care ?

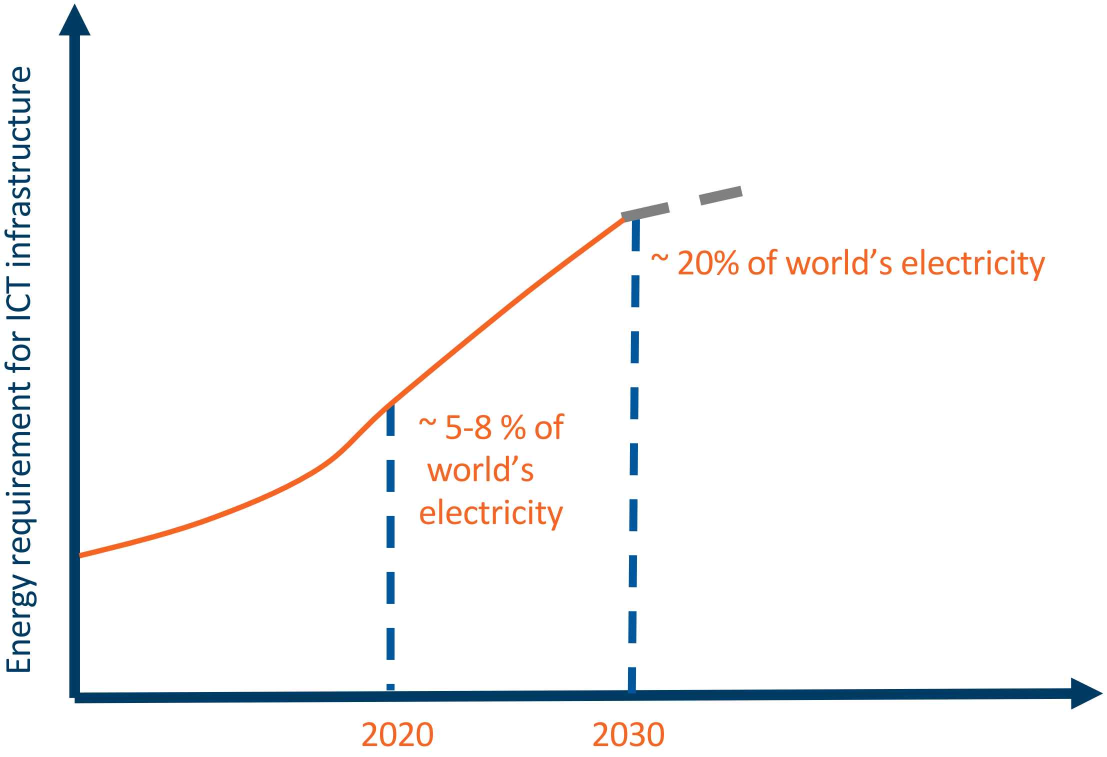

Amsterdam’s energy consumption in 2020 was ~66MW
Power requirements for world’s top supercomputers (MW)
30MW
1MW
2005
2020



Amsterdam's energy consumption in 2020 was ~66MW
World's fastest 2022
World's fastest 2023
Power requirements for world's top supercomputers (MW)
50MW
30MW
1MW
2005
2020


Amsterdam’s energy consumption in 2020 was ~66MW
World’s fastest 2022
World’s fastest 2023
Power requirements for world’s top supercomputers (MW)
50MW
30MW
1MW
2005
2020


10^4
10^3
10^2
10
1
picojoules per 64-bit operation
2008 (45 nm)
2018 (11 nm)
DP flop
register
1 mm on-chip
5 mm on-chip
15 mm on-chip
off-chip/DRAM
local interconnect
cross system



HTTPS://WWW.SRC.ORG/ABOUT/DECADAL-PLAN/

Snellius : Dutch National Supercomputer

Snellius

Phase 1:

- 76,832 cores (1.6 ×)
- 144 GPUs (3 Pflop/s, 14 ×)
- Total peak: 6.1 Pflop/s (3.4 ×)

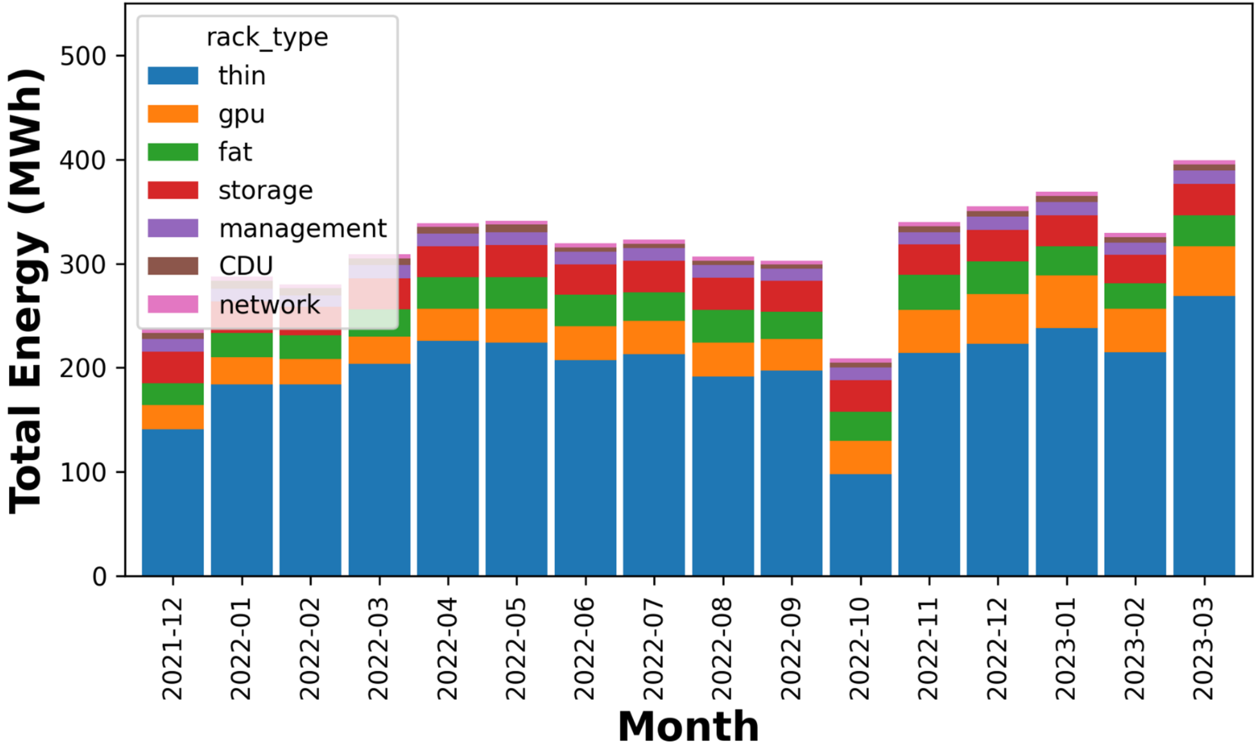
Phase 1 and 2

- Total peak: 11.2 Pflop/s (6.2 ×)

- Full system, based on choice for Phase 3:
  - > 200,000 cores (> 4 ×)
  - Total peak: 13.6 – 21.5 Pflop/s (7.6 – 11.9 ×)

Snellius – Energy Consumption

- HPL Energy Consumption (typical use: 85%, idle use: ~ 25%)
  - Snellius
    - Phase 1: 620 kW (0.7 ×)
    - Phase 1 and 2: 1200 kW (1.4 ×)
    - Full system (“worst case”: Phase 3 GPU): 1430 kW (1.6 ×)
- Average energy consumption based on phasing
  - 2021–2022: ~ 1 ×
  - 2023: 1.5 ×
  - 2024 and later: 1.6 ×



Average Dutch household in 2021

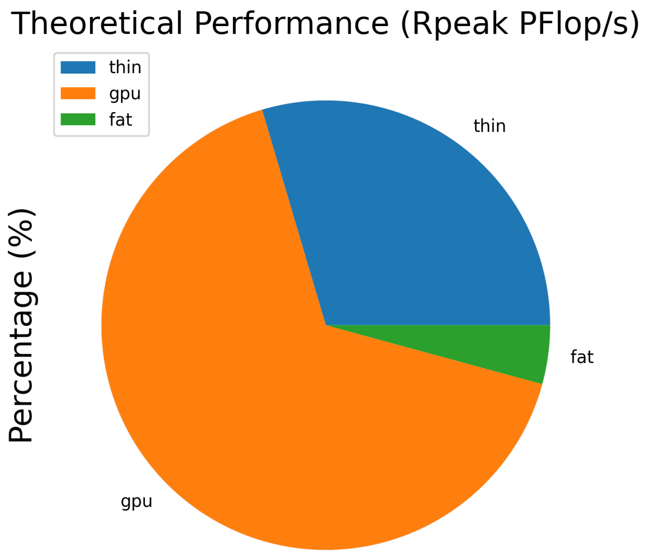
- 2.81 MWh

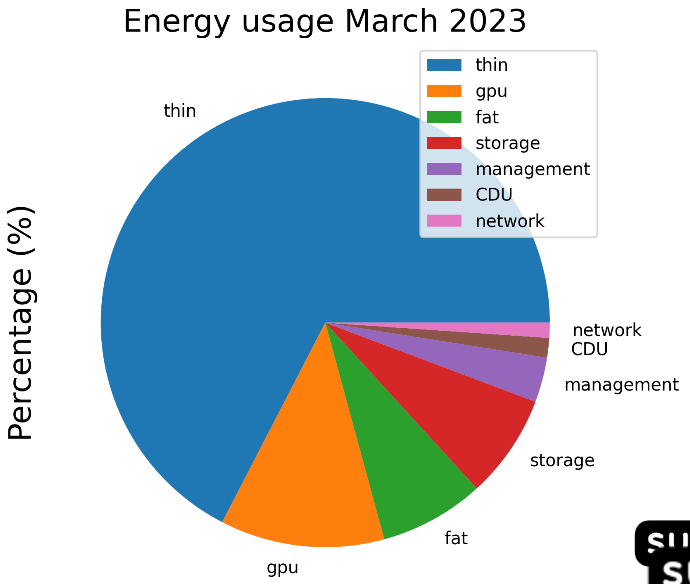
  2.04 MWh (Apartment)

  3.3 MWh (House)

Performance vs Energy

GPU ~ same compute for 1/5 the energy (March 2023)





SURF

An electric car consumes 25KWh for 100 miles ~ 0.9 million joules for 1 mile

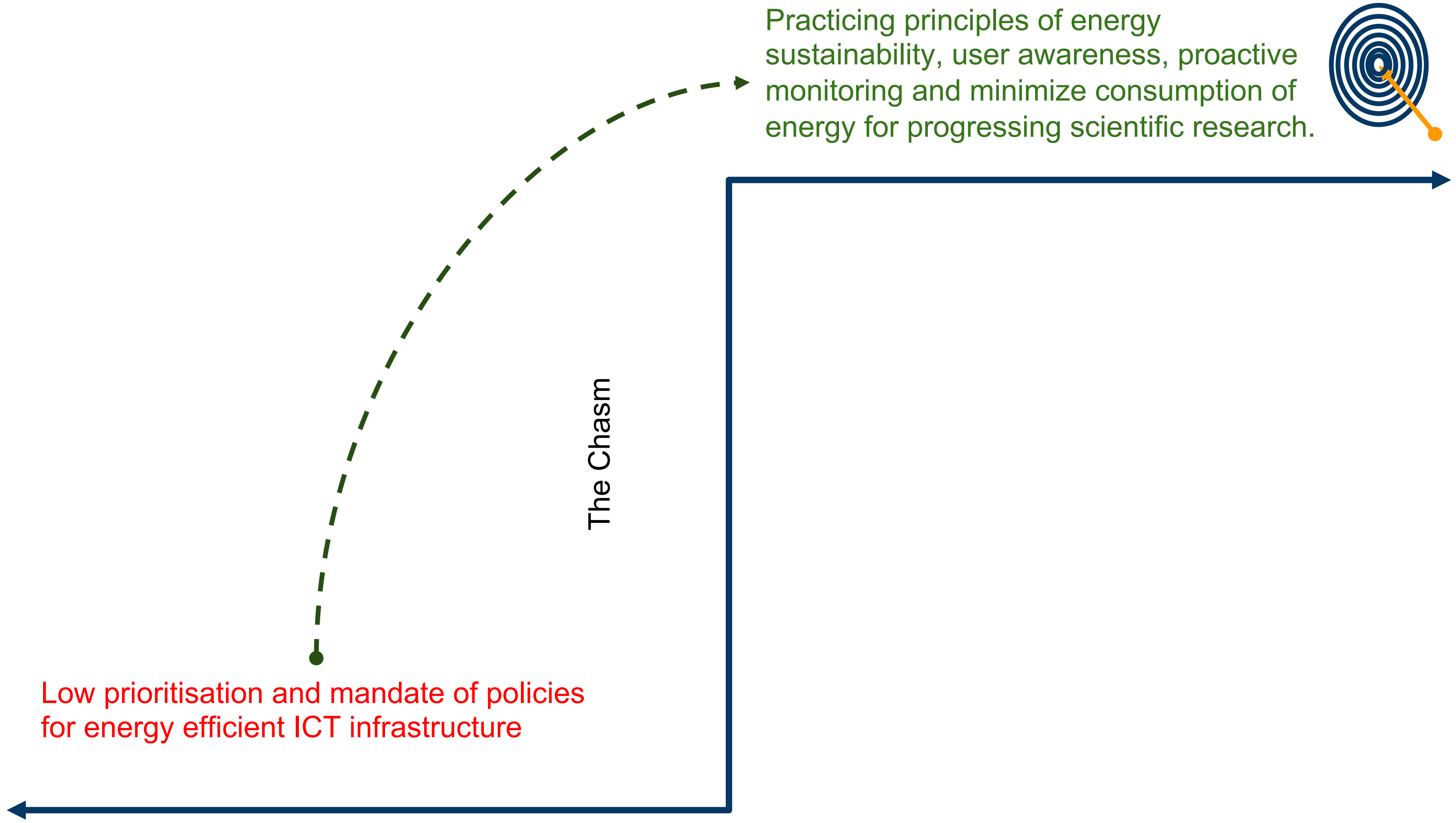
Moderately scaled Molecular dynamics simulation on CPUs can use up to 15 million joules or 15KW for 1000s ~ sufficient chunk of current energy envelop.

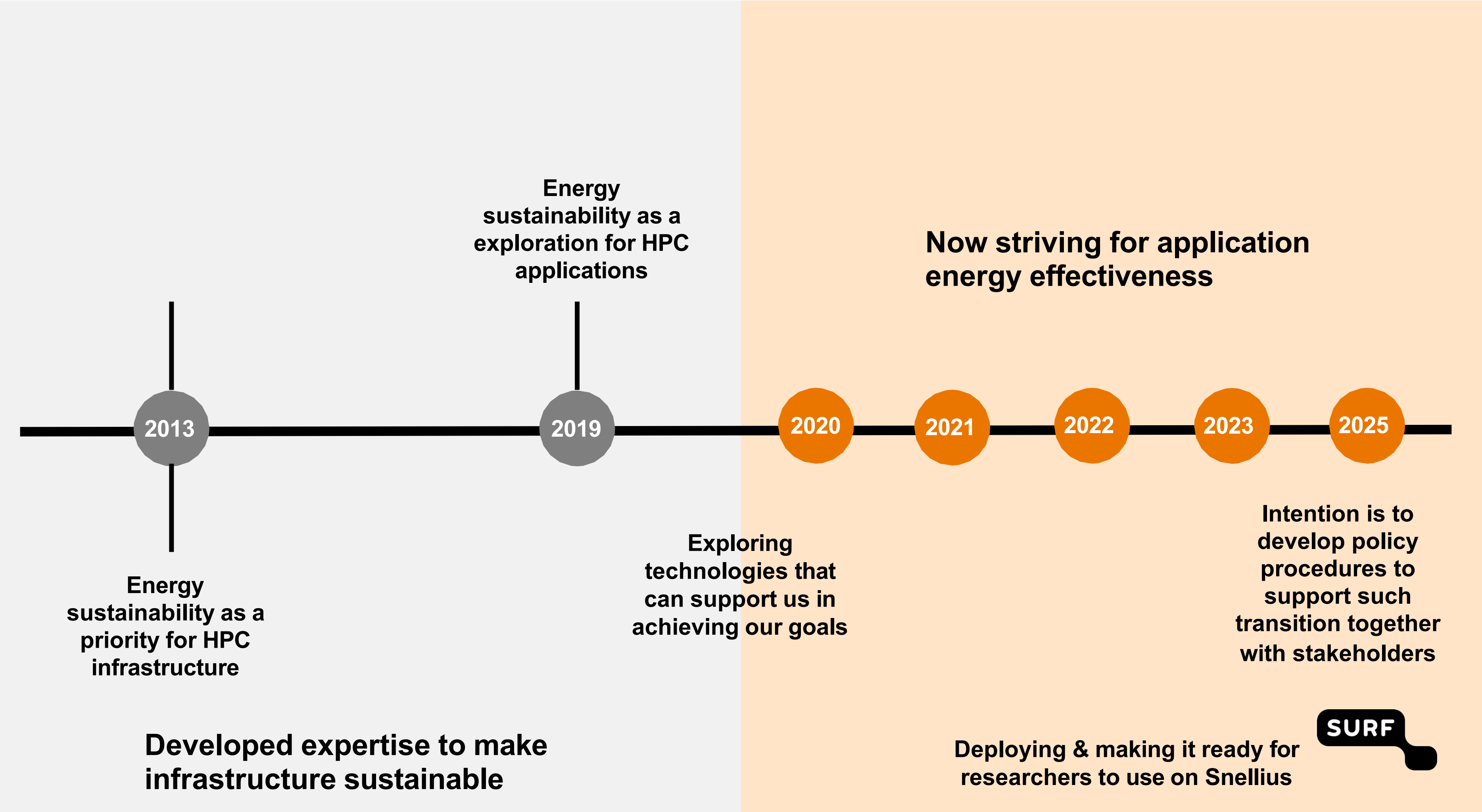
Moderately scaled Machine learning training on GPUs can use up to 4 million joules for 20KW for 3 minutes.

A physics-based fluid simulation can use 3.5 million joules or 6.5 KW for 10 mins

Supercomputing technologies are maturing
are becoming energy costly

How do we **envision**
Future of research & Education with ICT while
being **sustainable** ?


Practicing principles of energy sustainability, user awareness, proactive monitoring and minimize consumption of energy for progressing scientific research.
The Chasm
Low prioritisation and mandate of policies for energy efficient ICT infrastructure

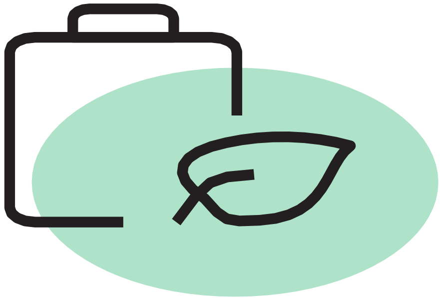

Energy sustainability as a exploration for HPC applications
Now striving for application energy effectiveness
2013
2019
2020
2021
2022
2023
2025
Energy sustainability as a priority for HPC infrastructure
Exploring technologies that can support us in achieving our goals
Intention is to develop policy procedures to support such transition together with stakeholders
Developed expertise to make infrastructure sustainable
Deploying & making it ready for researchers to use on Snellius
SURF

Awarding Requirements and their corresponding weights in the tendering process

Awarding Requirements	Weight
AR 1: Hardware	9%
AR 2: Storage, I/O and file system	9%
AR 3: Software	9%
AR 4: Operational	9%
AR 5: Corporate Social Responsibility	15%
AR 6: Applications Performance (benchmark)	40%
AR 7: Planning and phased growth, system image and partnership	9%
Total	**100%**

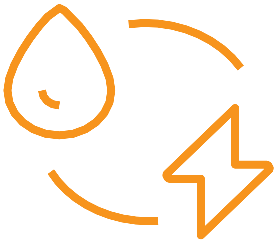
1

GREEN ENERGY



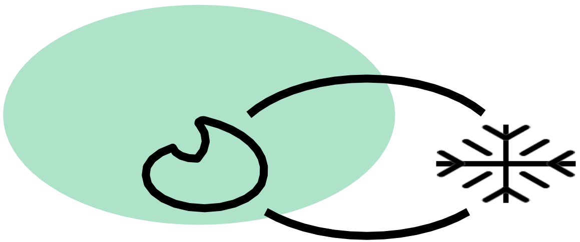
The supercomputer is powered by

100% renewable energy.

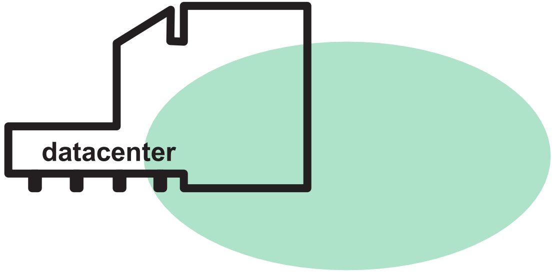


2

GREEN
Datacenter Infrastructure



The building has hot and cold water storage & isles for heating and cooling

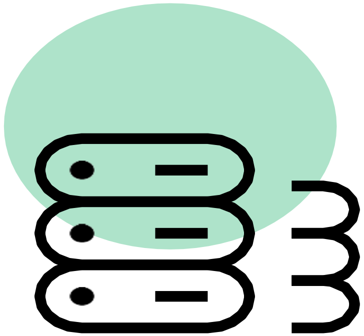


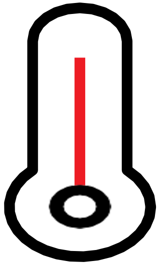
150m
under ground!

3

GREEN
SUPERCOMPUTER

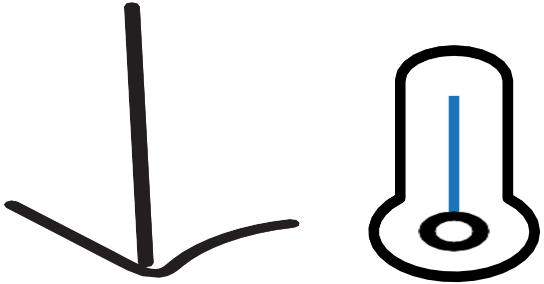
Advanced cooling techniques

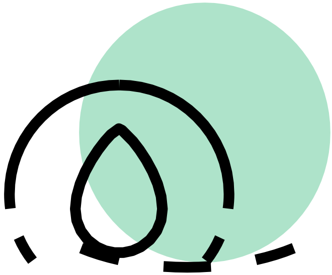




we use:

- Radiator Heat Exchange
- Direct liquid cooling





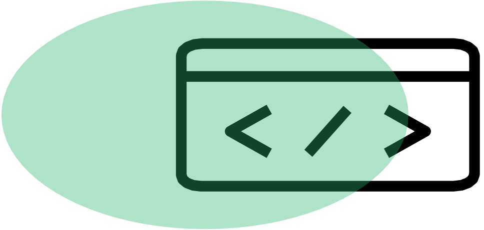
What is causing all that heat? Oh yes! The **billion scientific calculations** that this beast can do in seconds.

but what if we could
do it more efficiently and save energy?!

we can with →

4 GREEN HPC APPLICATIONS

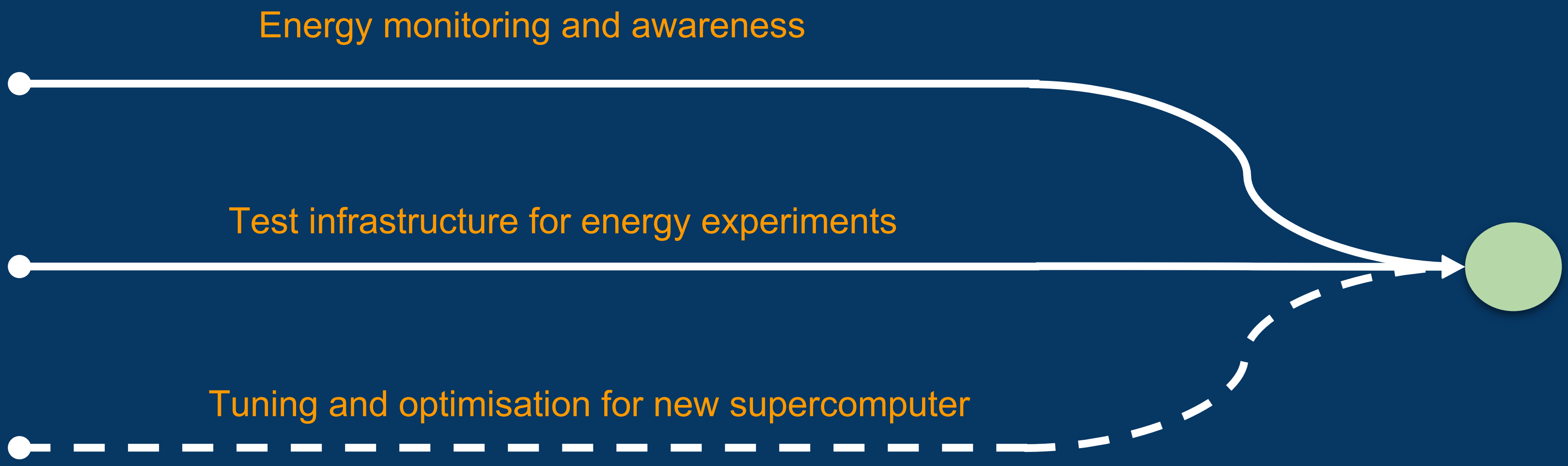
Using special tools to understand energy consumption & **optimize** the use of energy for scientific research



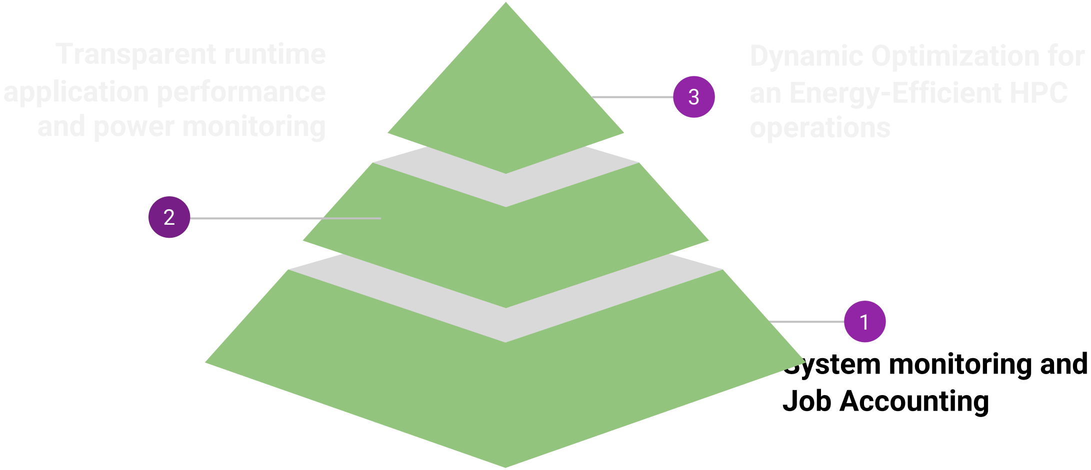
5

Advocating to Become a GREEN user!

Become **aware** of the energy footprint of your research & optimize the consumption

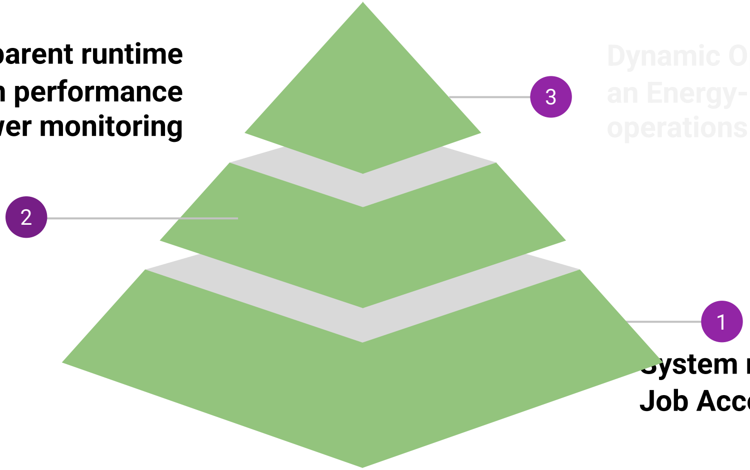

Energy monitoring and awareness
Test infrastructure for energy experiments
Tuning and optimisation for new supercomputer

Design focus



Design focus

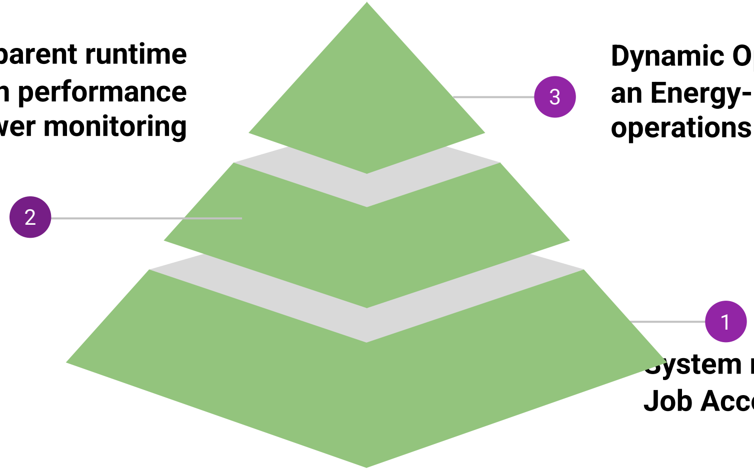
Transparent runtime application performance and power monitoring



Dynamic Optimization for an Energy-Efficient HPC operations

System monitoring and Job Accounting

Design focus

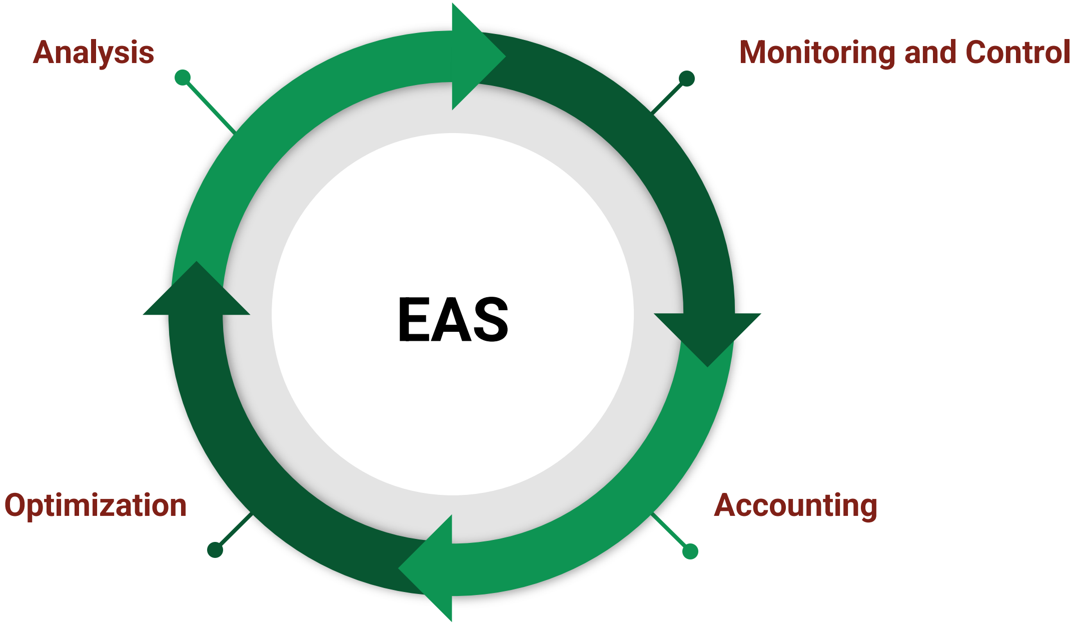


Transparent runtime application performance and power monitoring

Dynamic Optimization for an Energy-Efficient HPC operations

System monitoring and Job Accounting

Energy-efficiency process



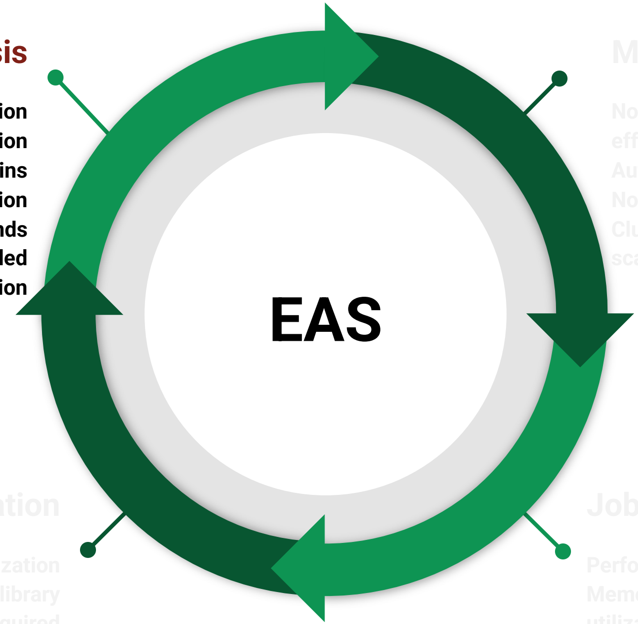
Energy-efficiency process

Analysis

Hints for application analysis and optimization
Hints for system analysis and optimization
Reporting system, extensible based on plugins
Trace file generation
Energy accounting and reporting commands included
Graphana visualization

Monitoring and Control

Nodes temperature, power (Node, DRAM. Package, GPU), effective frequency …
Automatic reporting of run time hardware issues
Node powercap: heterogeneous nodes, power balance
Cluster powercap : Hierarchical architecture for system scalability



Optimization

Runtime application energy optimization
Transparent, dynamic and lightweight runtime library with no user intervention required
Automatic energy savings according to energy policies

Job Accounting

Performance metrics monitoring: CPI, CPU Gflops, Memory bandwidth, MPI hints, IO bandwidth, GPU utilization
Granularity: jobid, stepid, “loop”, user, node

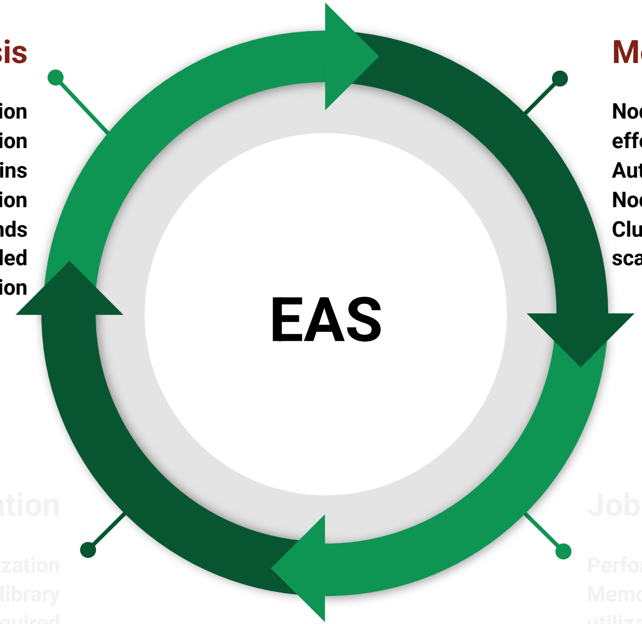
Energy-efficiency process with EAS technology

Analysis

Hints for application analysis and optimization
Hints for system analysis and optimization
Reporting system, extensible based on plugins
Trace file generation
Energy accounting and reporting commands included
Graphana visualization

Monitoring and Control

Nodes temperature, power (Node, DRAM. Package, GPU), effective frequency …
Automatic reporting of run time hardware issues
Node powercap: heterogeneous nodes, power balance
Cluster powercap : Hierarchical architecture for system scalability



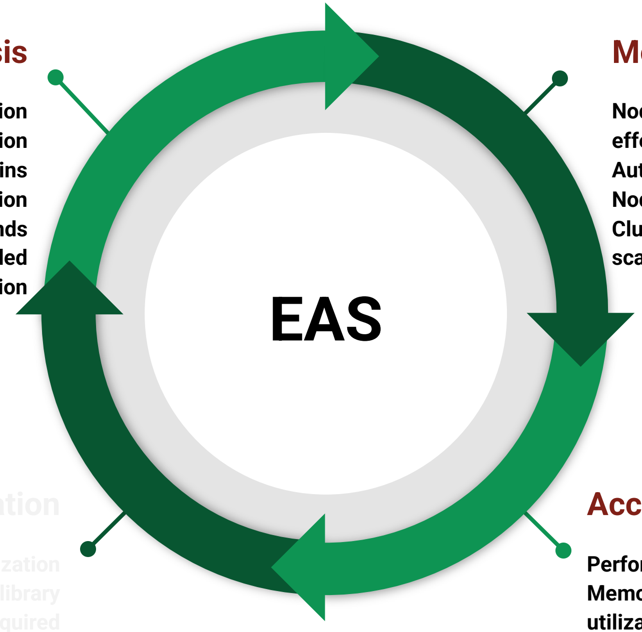
Optimization

Runtime application energy optimization
Transparent, dynamic and lightweight runtime library with no user intervention required
Automatic energy savings according to energy policies

Job Accounting

Performance metrics monitoring: CPI, CPU Gflops, Memory bandwidth, MPI hints, IO bandwidth, GPU utilization
Granularity: jobid, stepid, “loop”, user, node

Energy-efficiency process with EAS technology



Analysis

Hints for application analysis and optimization
Hints for system analysis and optimization
Reporting system, extensible based on plugins
Trace file generation
Energy accounting and reporting commands included
Graphana visualization

Monitoring and Control

Nodes temperature, power (Node, DRAM. Package, GPU), effective frequency …
Automatic reporting of run time hardware issues
Node powercap: heterogeneous nodes, power balance
Cluster powercap : Hierarchical architecture for system scalability

Optimization

Runtime application energy optimization
Transparent, dynamic and lightweight runtime library with no user intervention required
Automatic energy savings according to energy policies

Accounting

Performance metrics monitoring: CPI, CPU Gflops, Memory bandwidth, MPI hints, IO bandwidth, GPU utilization
Granularity: jobid, stepid, “loop”, user, node

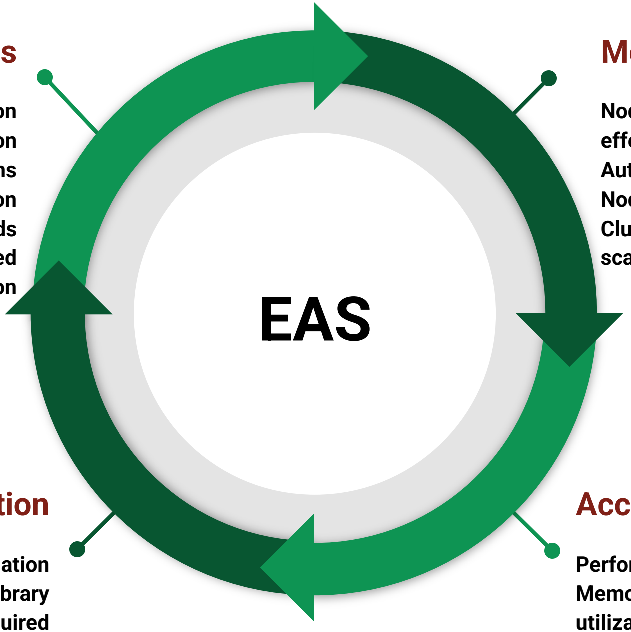
Energy-efficiency process with EAS technology

Analysis

Hints for application analysis and optimization
Hints for system analysis and optimization
Reporting system, extensible based on plugins
Trace file generation
Energy accounting and reporting commands included
Graphana visualization

Monitoring and Control

Nodes temperature, power (Node, DRAM. Package, GPU), effective frequency …
Automatic reporting of run time hardware issues
Node powercap: heterogeneous nodes, power balance
Cluster powercap : Hierarchical architecture for system scalability

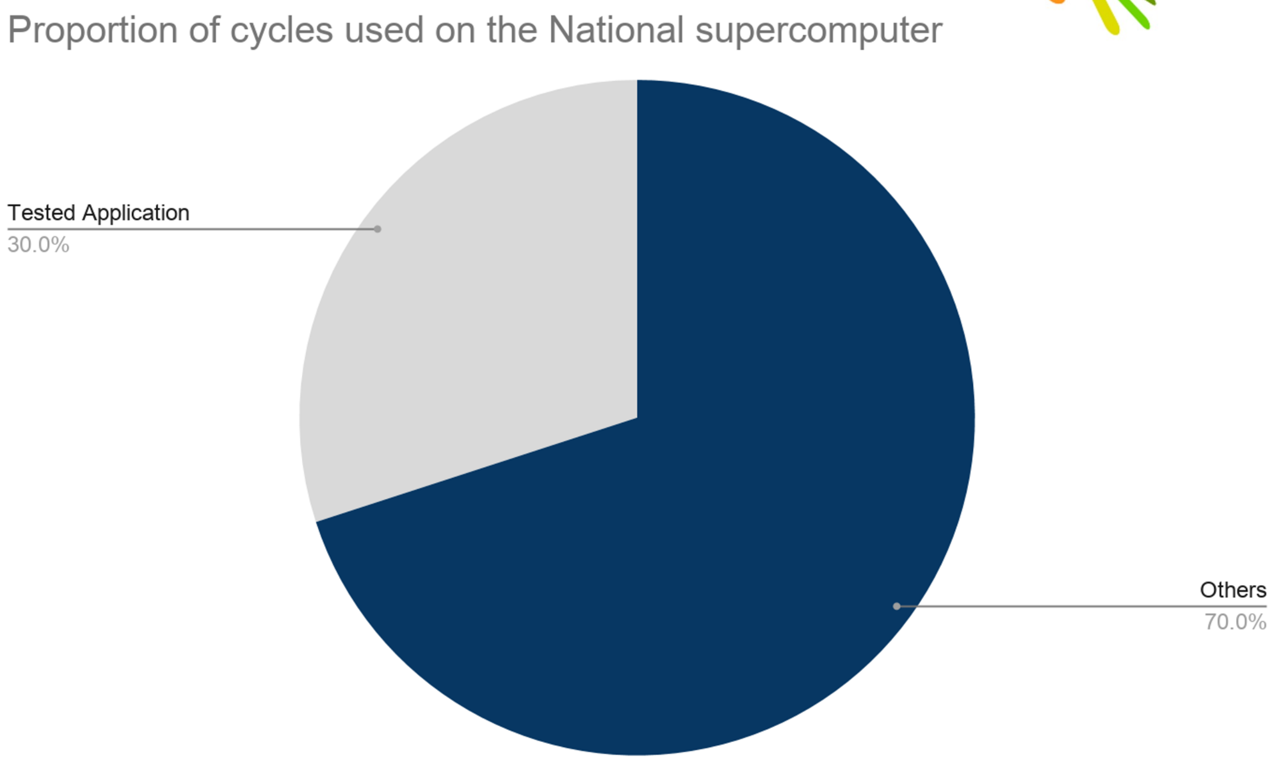


Optimization

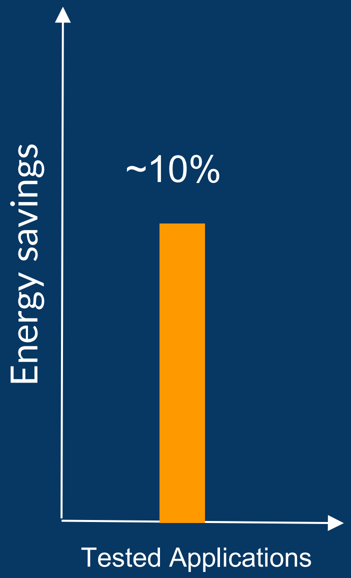
Runtime application energy optimization
Transparent, dynamic and lightweight runtime library with no user intervention required
Automatic energy savings according to energy policies

Accounting

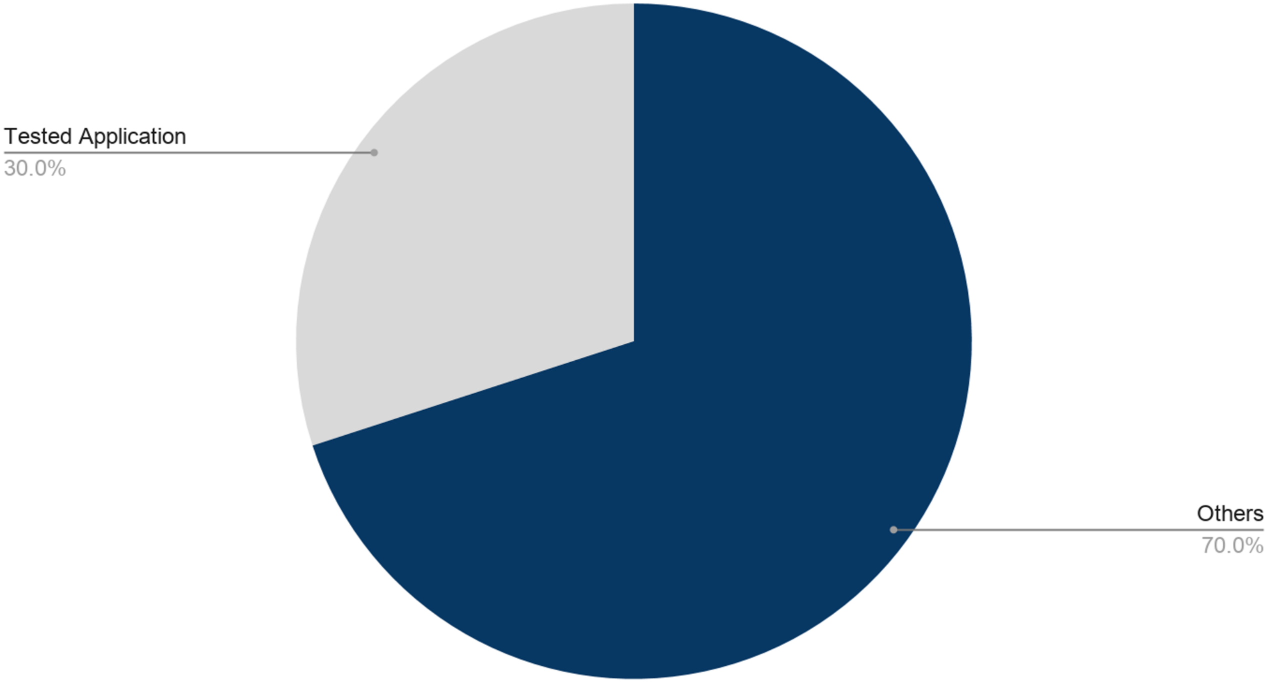
Performance metrics monitoring: CPI, CPU Gflops, Memory bandwidth, MPI hints, IO bandwidth, GPU utilization
Granularity: jobid, stepid, “loop”, user, node


Proportion of cycles used on the National supercomputer
Tested Application
30.0%
Others
70.0%

Early studies in collaboration with EAR-BSC



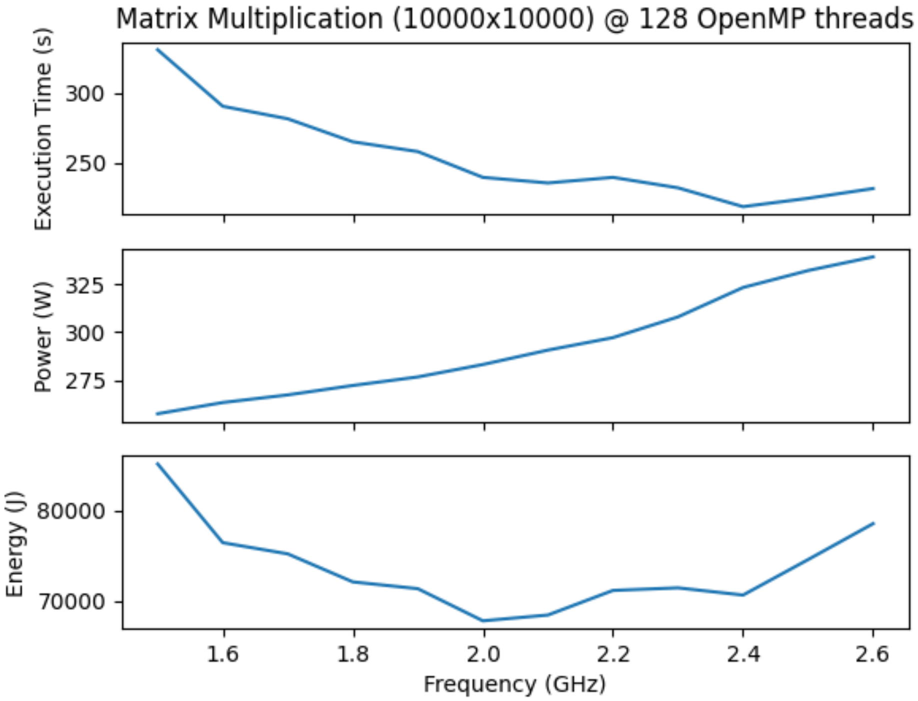
Proportion of cycles used on the National supercomputer



Visualisation Prototype

What are we going to do?

1. Large number of experiments will be carried out to fully understand tunable parameters
2. Assessing impact, operational and user requirements



Matrix Multiplication (10000x10000) @ 128 OpenMP threads
Execution Time (s)
300
250
Power (W)
325
300
275
Energy (J)
80000
70000
1.6
1.8
2.0
2.2
2.4
2.6
Frequency (GHz)

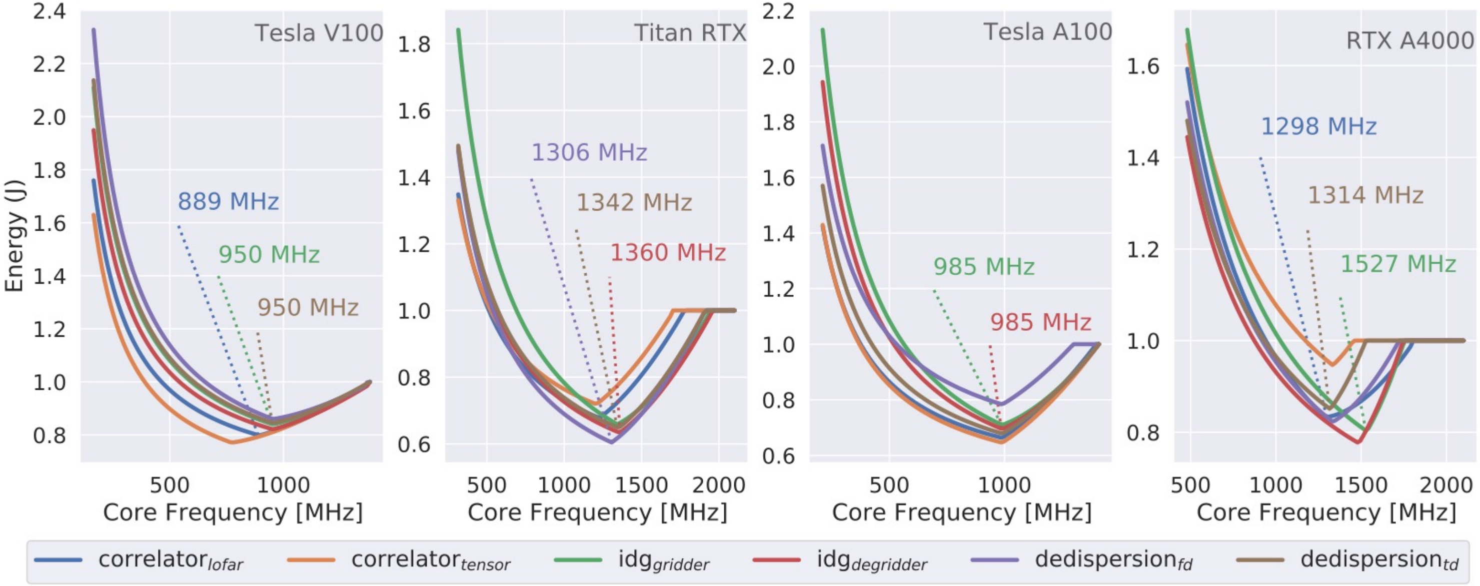


Fig. 10: Modelled energy usage (J) with power consumption model for core clock frequencies (MHz) of LOFAR kernels for the Tesla V100, Titan RTX, Tesla A100 and RTX A4000 GPUs.

Our approach published in European Research Magazine for computer science & mathematics (ERCIM)

Making Scientific Research on Dutch National Supercomputer Energy Efficient

by Sagar Dolas (SURF), Ana Verbanescu (Twente University) and Benjamin Czaja (SURF)

Energy is an emerging topic in the scientific computing ecosystem and is becoming a design point for the future of research. Science relies increasingly on digital research computing as a tool for analysis and experimentation. Exponential increase in demand for computing means that classically designed ICT infrastructure will soon become unsustainable in terms of its energy footprint [1]. We need to experiment with energy-efficient methods, tools, and algorithms and hardware technologies. In the Netherlands, we are working towards zero energy waste for high performance computing (HPC) applications on the national supercomputer "Snellius". It involves discussing challenges, proposing new research directions, finding opportunities to engage the user community, and taking steps for responsible use of software in research.

Traditionally, supercomputing focuses on improving latency or throughput, which are of massive importance for applications such as drug discovery or climate simulations. For many decades we developed infrastructure, algorithms, and software tools to obtain improvements. Given the rapid increase in energy usage for ICT services, further emphasised by the imminent energy crisis, it is a priority to understand and optimise the energy consumption of research computing applications [2].

Specifically, our initiative is about working with three stakeholders that need to collaborate to reduce the energy impact: application developers, system integrators, and system operators:

1. *Application developers* are responsible for improving the energy efficiency of their own code, making use of algorithmic, programming, and hardware tools at their disposal. Ideally, applications should be able to adapt to the available system resources and use them effectively. Research into programming models and tools that enable such flexibility is accelerating.
2. *System integrators* are responsible for offering the right resources for the application developers and system operators. These resources must include efficient hardware – e.g., different GPGPUs, CPUs (Central processing unit), or even FPGAs (Field Programmable Gate array) – to enable different application mixes. Research into procuring systems and provisioning applications with the right resources is mandatory.
3. *System operators*, with their holistic view, are responsible for efficiently scheduling workloads on system resources and potential energy harvesting where resources/systems are massively underutilised. Research into tools for energy-efficiency resource management and scheduling, as well as energy harvesting, is ongoing.

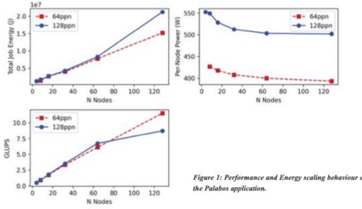


Figure 1: Performance and Energy scaling behaviour of the Palabos application.

To address the interests and concerns of all these stakeholders, we follow three lines of actions:

1. Define methods for application characterisation, towards building a detailed application signature in terms of resource utilisation, performance, and energy consumption.
2. Co-design system-level tools and platforms to allow operators to formulate recommendations to optimise overall energy consumption, and further shape green computing policies for the supercomputing systems.
3. Co-design frameworks to assess and configure systems procurement and resource provisioning for high-efficiency, low-waste application deployment.

Case study

We performed a strong scaling study on the computational fluid dynamics solver Palabos [L1], which is based on the Lattice Boltzmann Method. Palabos, we believe, serves as a typical use case of a memory-intensive HPC application and through performance/energy analysis can serve as a template for energy usage for other similar HPC applications.

As shown in Figure 1, the application scales linearly with increasing the number of nodes. We observe a flattening of the scaling behaviour on larger node counts and identify that the memory bandwidth limits the code. By lowering the processes per node (ppn) of the application, we could maximise the memory bandwidth available to the application, thus resulting in much lower energy usage. The analysis on Palabos represents the value of including energy as a metric and traditional performance analysis. It explains how the resources of a cluster can be adjusted for an HPC application to maximise performance and minimise energy usage.

Also, as shown in Figure 2, we are able to get iteration level information of the application using the system level tool Energy Aware Runtime [3]. With this fine-grained information, we can profile the application in order to identify when and where the application is using the most energy. In Figure 2, we identify that Palabos uses the most energy during the "simulation" phase, where the value of GFLOPS is also very high.

The analysis on Palabos represents the value of including energy as a metric and traditional performance analysis. It explains how the resources of a cluster can

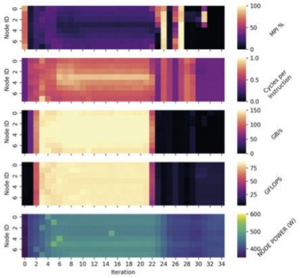


Figure 2: Heatmap analysis of Palabos.

We are also exploring Neuromorphic computing for better energy efficiency

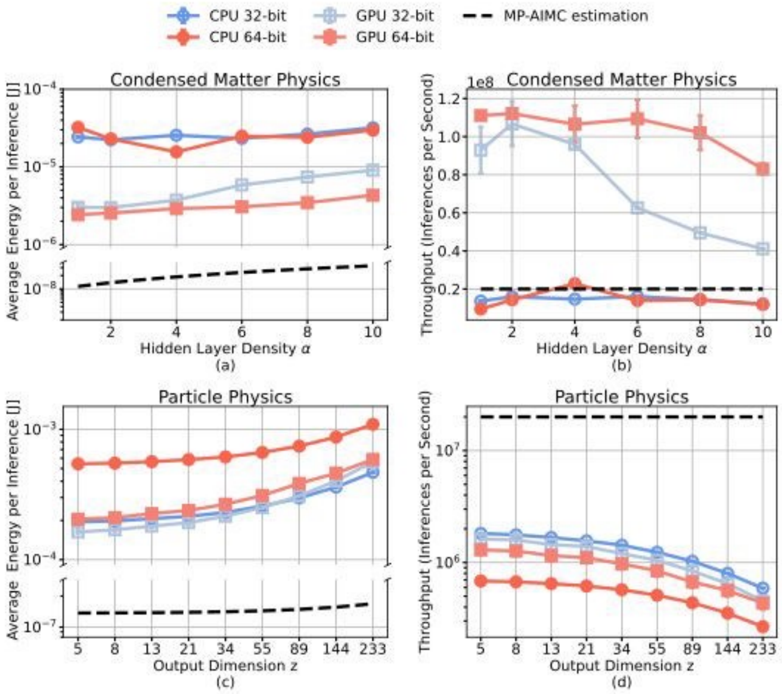


FIG. 4. Results benchmarks on CPU (Intel Xeon Gold dual socket) and GPU (NVIDIA V100) and estimations for the proposed AIMC architecture. (a) The average energy per inference (or average energy per state) in Joule and (b) the throughput inferences per second (or states per second) for the condensed matter physics use case. (c) The average energy per inference (or average energy per event) in Joule and (d) the throughput inferences per second (or events per second) for the particle physics use case. The measurements on the CPU and GPU are repeated 10 times and the standard deviation is included in the plots. The batch size used for the condensed matter physics use case is equal to all states with zero magnetization (12870) and the batch size used for the particle physics use case is 10^6. Horizontal axes indicate a measure defining the network complexity/size; this is the hidden layer density α for the condensed matter physics use case and the output dimension z for the particle physics use case.

Let’s make “***Energy as a design philosophy***” for the future of Computing.

netherlands
eScience center

EAR

University of Amsterdam

UNIVERSITY OF TWENTE.