Energy efficient computing and computing for energy

Ariel Oleksiak
Poznań Supercomputing and Networking Center

POLAND
61-139 Poznan
ul. Jana Pawła II 10
phone: (+48 61) 858-20-01
office@man.poznan.pl
www.psnc.pl

Question: How much energy has been consumed to display the „Gangnam style" video?

Hint: the most frequently displayed video on YouTube:

> 1.5 billion in 2013

Burundi ~ 300 GWh

Is energy an issue?

- Why such waste acceptable?
    - Economic gains

- Current AI hype requires attention to the energy efficiency and availability
    - Even for large companies problem with getting profits out of trained models
    - Development and use of AI models by smaller companies and science limited by high costs
    - Next year forecasts: 3.5 millions of NVIDIA H100 GPUs will consume around 13 TWh (yearly consumption of Guatemala or Lithuania)

- Energy costs
    - Sudden rise of energy prices in Europe after start of the war in Ukraine
    - Large energy costs of new HPC systems

- Sustainability goals & EU regulations
    - „European Green Deal"
        - making the EU climate-neutral by 2050
    - „Fit for 55" legislation package
        - reducing EU emissions by at least 55% by 2030 (compared to 1990)

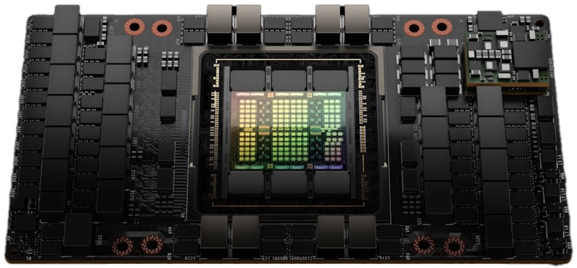
Some trends/predictions from GridLab 2012

- Move to **massive parallelism** using energy-efficient **manycore** architectures
  - From embedded devices market: many simple, low power cores (e.g. ARMs)
  - From highly specialized processors from gaming/graphics market space: GPU/Accelerator (e.g. NVidia Fermi, Tesla)
- **Hardware-software co-design**, customized hardware
- Huge infrastructures vs **micro-datacenters**
  - Soon a single rack will deliver 1PFlop!
  - Wide distribution of resources/providers as envisioned in grids?
- **Autonomous data center management**
  - Thermal-aware resource management, cooling system, energy supply
- Novel approaches to **managing energy supply**
  - Use of **renewable energy sources** and batteries
    - Automated distribution, reacting to fluctuations
  - Use of **direct current**
    - Reducing conversion losses
  - **Heat re-use**

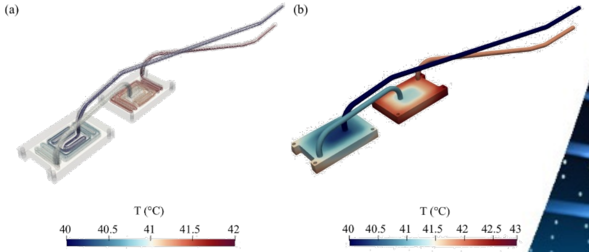
Edge computing?

Technologies and methods for improving efficiency

- Achievements so far
  - Development of CPU architectures
    - Multicore with specialized units
  - CPU and node level management (e.g. DVFS)
  - Virtualization
    - Even GPUs, e.g. VMware Bitfusion
  - Extensive use of accelerators and specialized hardware
    - Especially visible in cryptocurrency and AI applications



- Thermal management and cooling
  - Direct liquid cooling
  - Immersive cooling
  - 2-phase cooling

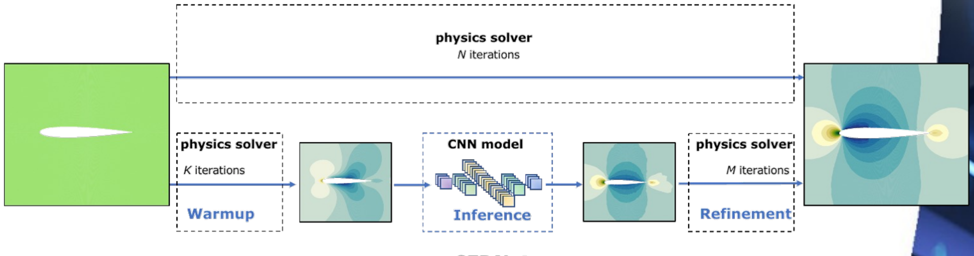


- Promising directions
  - Domain specific architectures, hardware-software co-design
  - Mixed precision computing
  - AI-supported simulations

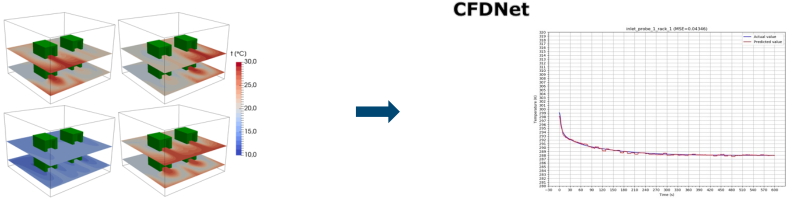
AI-supported simulations

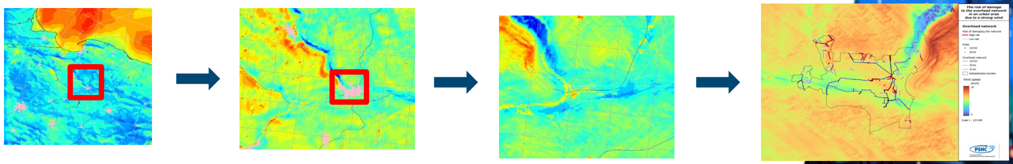
Examples

- Speed-up of CFD simulations
  - e.g. heat flow simulations using of OpenFoam
- Replacing heavy computations with surrogate AI model (after training on simulation data)
  - e.g. thermal models of data centers or other spaces
- Predictions based on simulations results
  - e.g. renewable energy production based on weather forecasts
- Many domain specific solutions!



CFDNet





ution:
100 m

Impact on efficiency

- Not always gains exceed the AI training effort
- Often high replicability of models needed
- Promising direction but need to find trade-off between AI costs and gains

Technologies and methods for improving efficiency

- Achievements so far
  - Developm
    - Multi
  - CPU and r
  - Virtualiza
    - Even
  - Extensive hardware
    - Especially visible in cryptocurrency and AI applications
- Thermal ma
  - Direct liq
  - Immersiv
  - 2-phase c
- Promising di
  - Domain s software co-design
  - Mixed precision computing
  - AI-supported simulations

Fast progress in energy efficiency – The fastests supercomputer:

25x more efficient in 2023 than in 2013

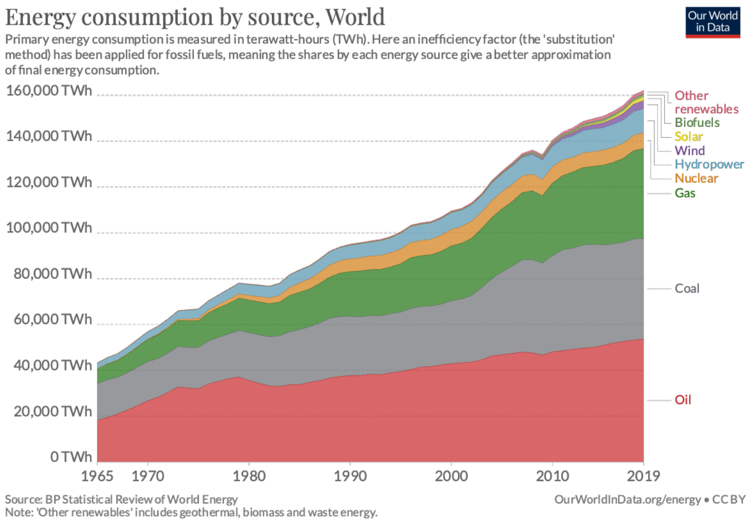
To continue or achieve large gains disruptive technologies needed:

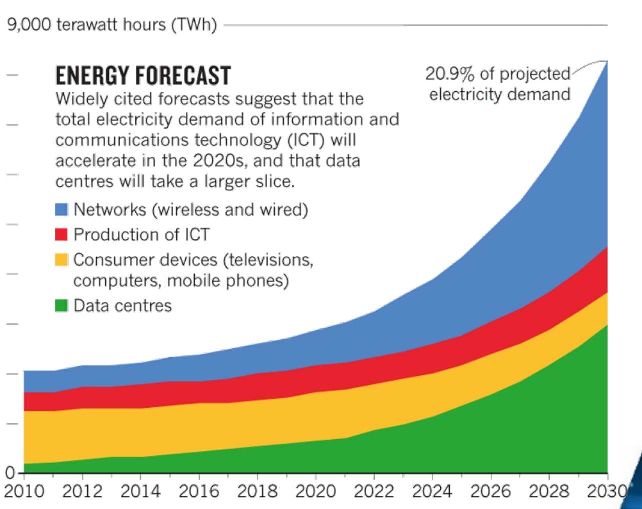
- Quantum computing
- Neuromorphic computing
- Other?

Are advances in energy efficiency enough?

No

The Jevons Paradox:
increase in efficiency of resource use **=** increase in resource consumption





Power usage of the fastest supercomputer:

2023	2013
22,7 MW	**17,8 MW**

PSNC 30
30 YEARS OF POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

Computing & Energy systems integration

Fact 1: Computing system produce large amounts of heat

- Energy used to information processing is negligible (Energy is „borrowed”)

→ Let’s use this heat
And transport it with low overhead

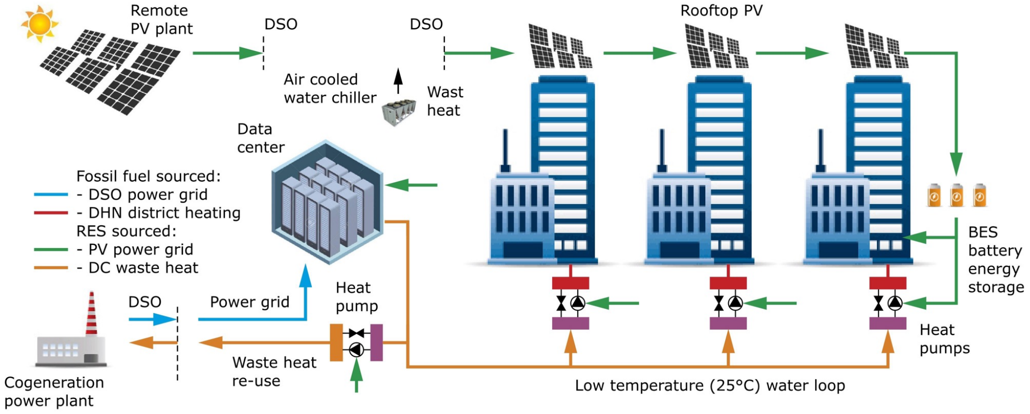
Fact 2: Computing systems consume large amount of Energy

- The source of energy and time of use matters

→ Use the cheap and clean Energy
Use it when it is cheap and clean

Heat re-use example at campus

- Current state:
  - Heat from a data center used for PSNC offices heating
  - Existing water loops and District Heating network
- Plans to provide excess heat to nearby campus
  - 30 000 GJ heat per year could be re-used
  - Qualitatively, between 25% and 60% of the demand is usually satisfied
  - Individual values can get close to 100% (e.g. during the holidays)
- Simulation models built to optimize the architecture and configuration

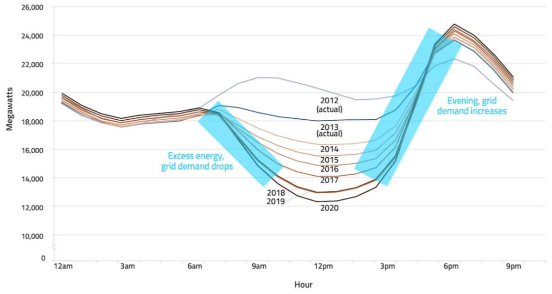


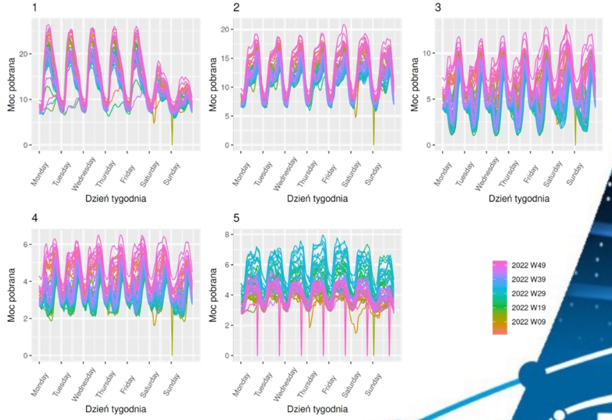
Use of renewable energy for computing

- PSNC laboratories at the airfield ~70km from Poznań
- PV installation ~1MWp of power
- Energy storage ~500kWh
- Planned small data center up to 240/480kW power
- The problem:
  - Maximize solar energy use (minimize costs of energy)
  - Based on:
    - Energy production (and its prediction)
    - Battery state
    - Load of computing system
    - Prices of energy (and its prediction)

Computing for Energy

- Power grid – can be treated as a big energy storage
  - The same problem as in the small scale
  - And many more to ensure its reliability
- Time aspect – The Duck Curve
  - Production from renewable energy sources does not fit the demand
- Energy distribution network
  - A large graph with a large number of nodes
    - ~38 000 power stations in regional distribution network
  - Large time series data
    - 1 billion events monthly
- Distributed energy generation
  - Revising the concept of power grids
  - Towards a digital twin of the network

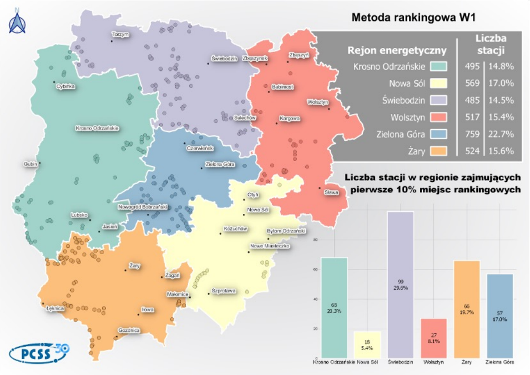


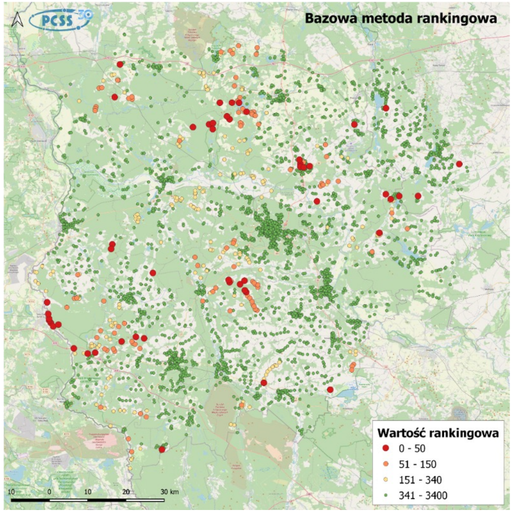


30 YEARS OF POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

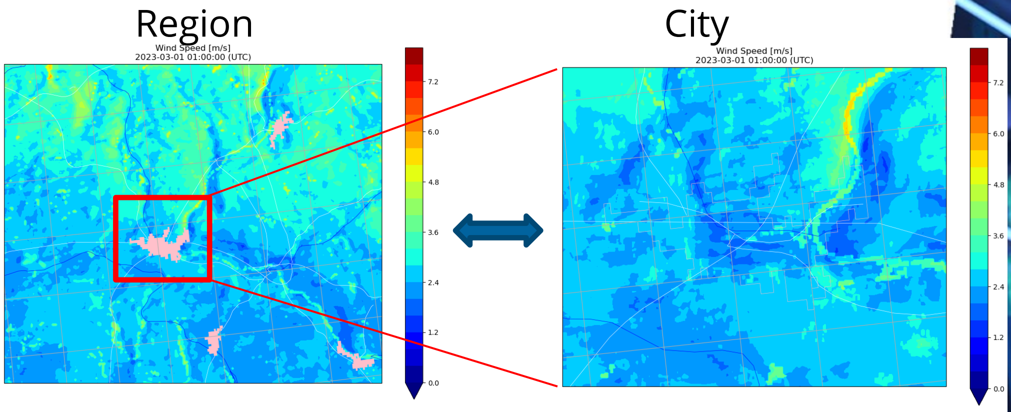
Computing tools for energy network operators

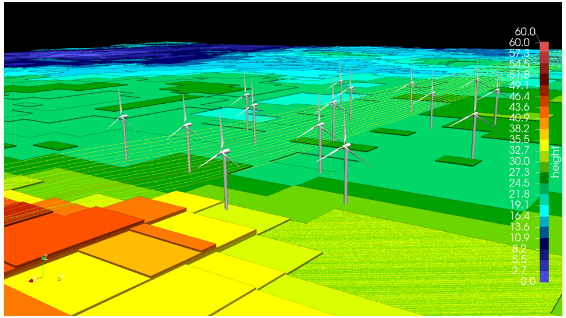
Optimisation of investments allocation

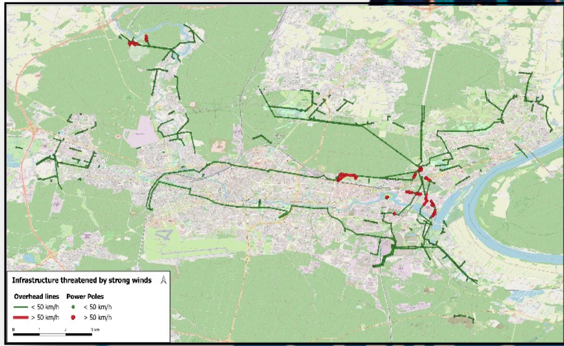




Predictions based on weather forecasts







Summary & Main messages

- Technologies for energy efficiency
  - Lot's of improvements especially on hardware level
  - Some promising approaches going on
  - But disruptive changes needed
    - e.g. quantum computing
  - Still not enough due to large demand and problem-specific improvements
- Integration of computing and energy systems
  - Heat re-use and efficient heat transfer
  - Maximization of renewable energy consumption
- One of missions of computing for the next decade:
  - Support digitalization and transformation of power grids and energy systems

Thank You!