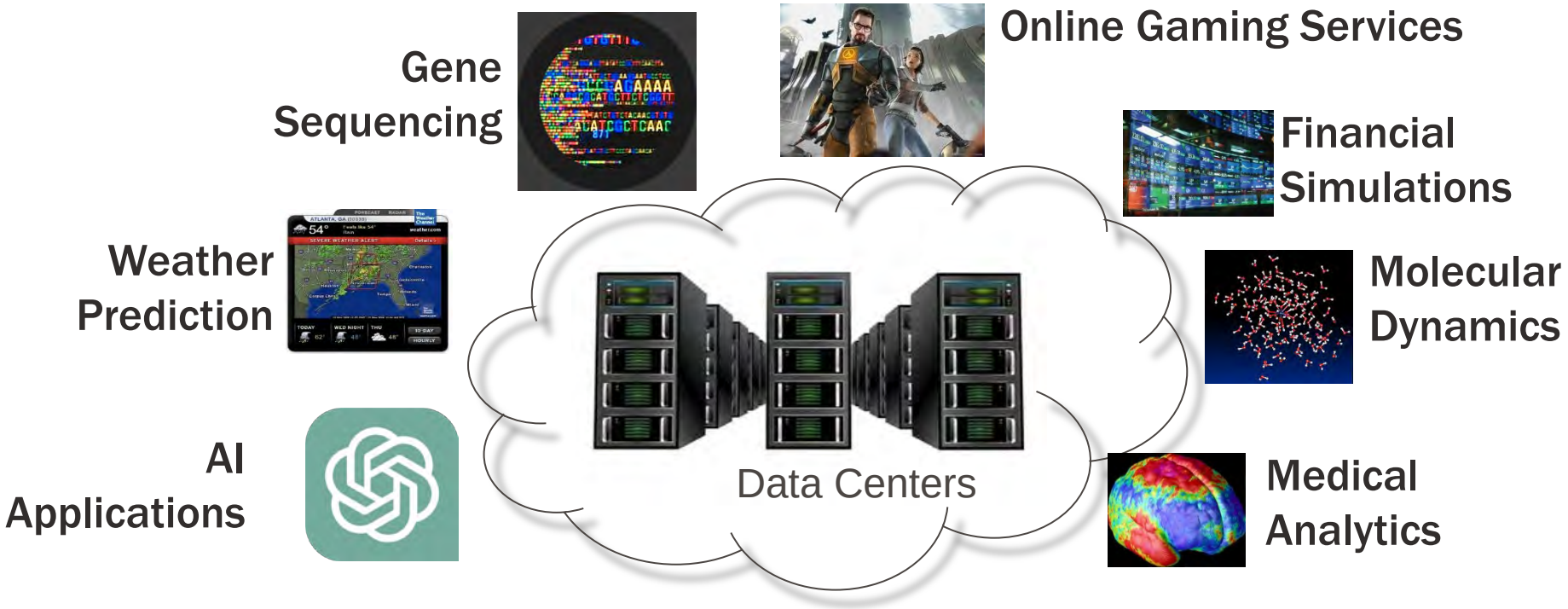
BEYOND THE BOTTLENECK: AI AS THE CATALYST FOR GREEN HPC

PROF. DAVID ATIENZA

EPFL, SWITZERLAND, DAVID.ATIENZA@EPFL.CH

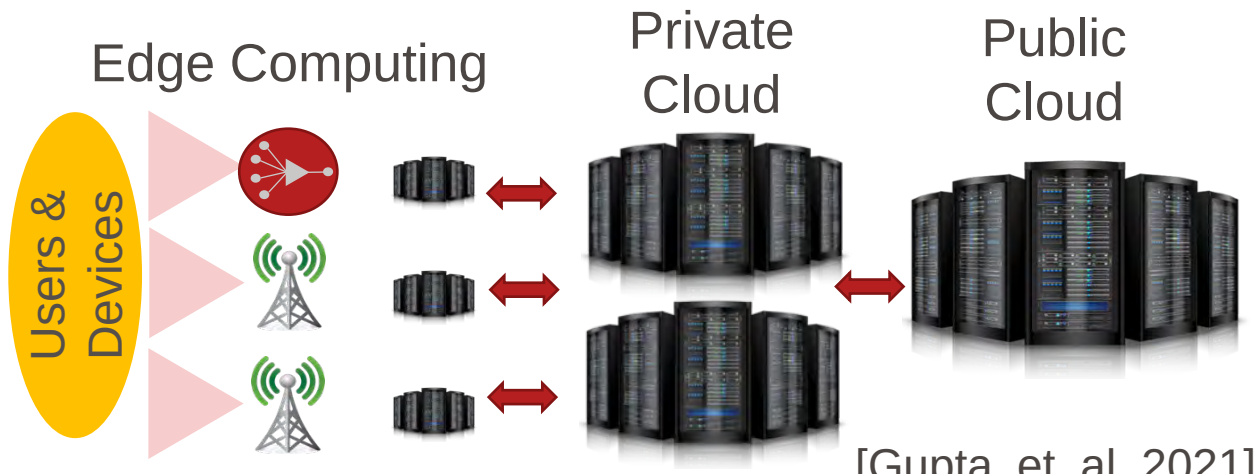
Computing Is Ever More Indispensable…

- Cloud is key in supply-chain of products/services
  - Boom on machine learning-based services: e-business, science, etc.



Multi-Layer Cloud Systems: Energy Keeps Growing!

- Cloud is an essential pillar in our digital economy
- Today, multi-scale computing beyond "classical cloud" (Public, private, and edge computing together)
- World's sustainability with IT?
  - Cloud growing: more services and datacenters, but **not sustainability-driven**
  - Cloud cannot keep up with new trends without **improving its efficiency**



[Gupta, et. al. 2021]

I use 17,000 times the amount of electricity than the average US household.

AI will run out of electricity and transformers in 2025. They're running out of transformers to run transformers.

Trend: DCs use 2% of global energy, they can reach 10% by 2030

Multi-Layer Cloud Systems: Energy Keeps Growing!

- Cloud is an essential pillar in our digital economy
- Today, multi-scale computing beyond "classical cloud" (Public, private, and edge computing together)
- World's sustainability with IT?
  - Cloud growing: more services and datacenters, but **not sustainability-driven**
  - Cloud cannot keep up with new trends without **improving its efficiency**

Users & Devices

SUSTAINABILITY

E

DeepMind AI reduces energy used for cooling Google data centers by 40%

Jul 20, 2016 · 4 min read

Rich Evans
Research Engineer, DeepMind

Jim Gao
Data Center Engineer, Google

Share

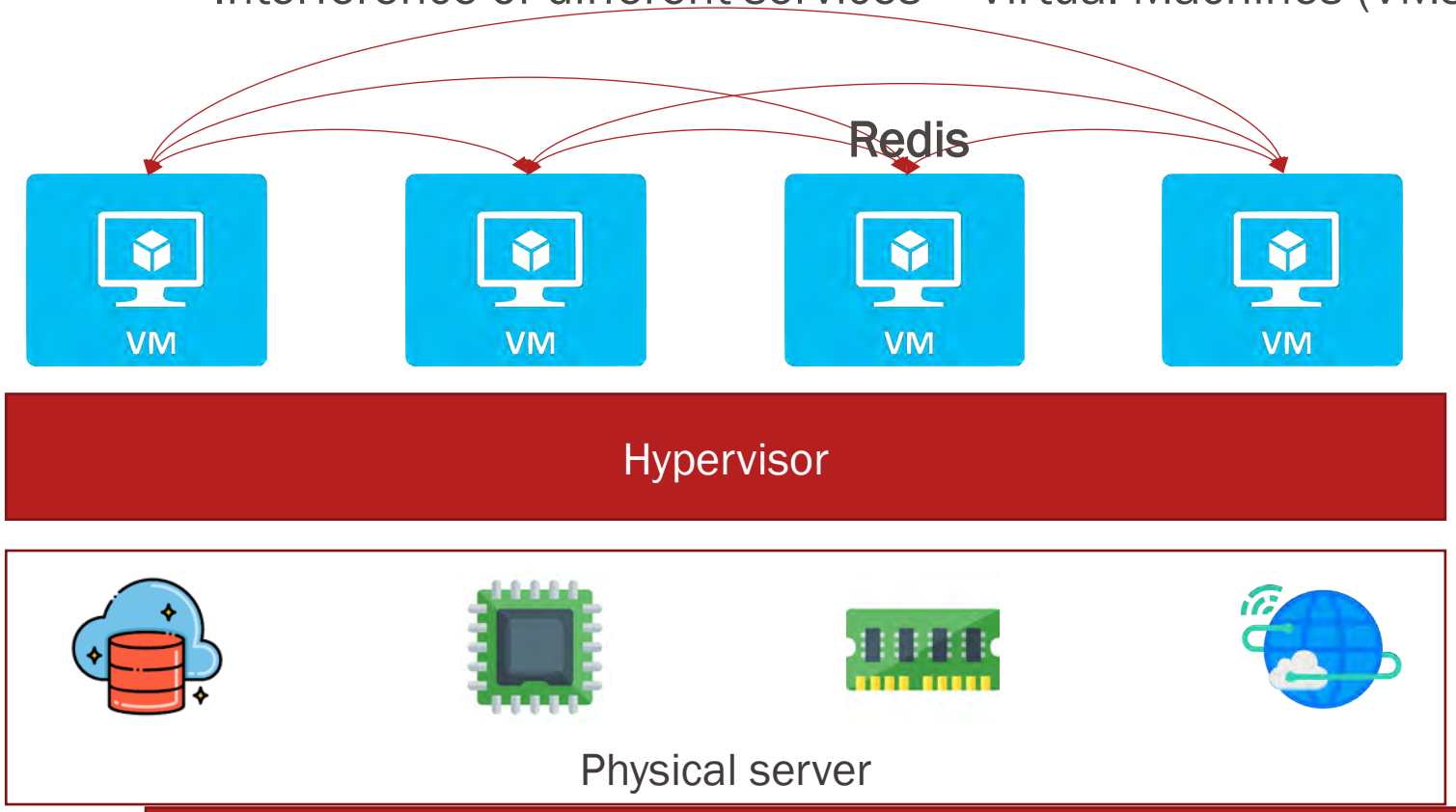
21]

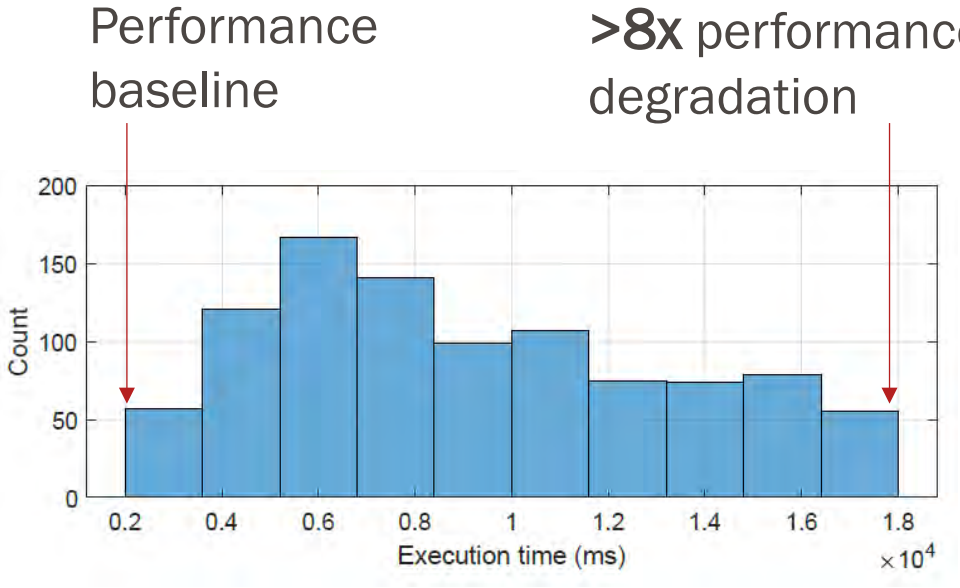
AI will run out of electricity and transformers in 2025. They're running out of transformers to run transformers.

Trend: DCs use 2% of global energy, they can reach 10% by 2030

Interference Problem on (Virtualized) Cloud Services

Interference of different services – Virtual Machines (VMs)





Performance of Redis benchmark

Collocated black-box VMs can suffer from severe performance degradation

Solution: Over-provisioning to “guarantee” performance in DCs:
Electricity and CO_2 emissions skyrocketing!

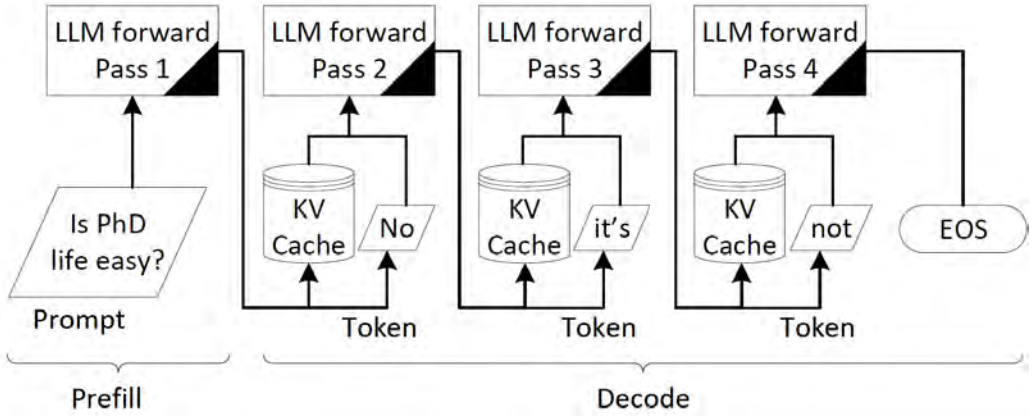
Key Observations: SLA Metrics & DVFS Mismatch

- **Two stages, two metrics:**
  - Prefill drives Time-To-First Token **(TTFT)**
  - Decode governed by Token-By-Token Delay **(TBT)**
    - Workload Throughput: Token-Per-Second **(TPS)**

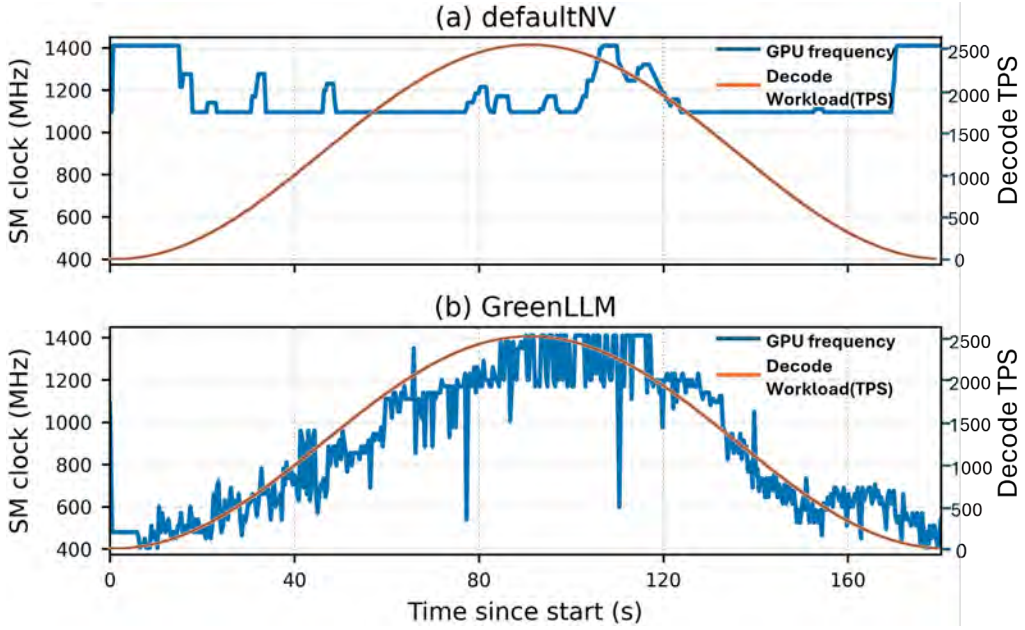
System for GreenLLM experiment

Component	Setting
GPUs	8× H100–SXM4 40 GB
CPU	AMD EPYC 7302, 16 cores
Framework	Dynamo v0.3.1

Reality today: Clear mismatch because default governors treat both phases the same way (wrong/reactive DVFS use, etc.)



LLM Inference Serving: Prefill and Decode



GPU Frequency vs. Decode TPS under defaultNV and GreenLLM

How to make CO_2 reduction economically sustainable?

Electricity: Berlin's shock plan to adapt to the weather

Companies may soon have to adapt their production to the strength of the wind and the duration of sunshine, in order to relieve the electricity networks, put to the test by the intermittency of renewable energies. This is the option proposed by the Ministry of Economy and Climate in a note published in July. Enough to trigger the ire of the business world.

Sustainability challenges in DCs

- Exa-scale amounts of data from AI, …
- Multi-node scalability
- Domain-specific computation

CO_2 Act - Paris Agreement: "Switzerland's target for 2030 is to reduce greenhouse gas emissions by <50% compared to 1990 "

Operational Footprint & Cost

CO2eq footprint from IT energy usage (computing, cooling, communications, etc.)

Embodied Footprint & Investment

Hardware manufacturing footprint (fabrication, transportation, etc.)

Real solutions: Minimize CO_2-eq emissions while maximizing return on investment in "sustainable technology" (i.e., incentives for companies)

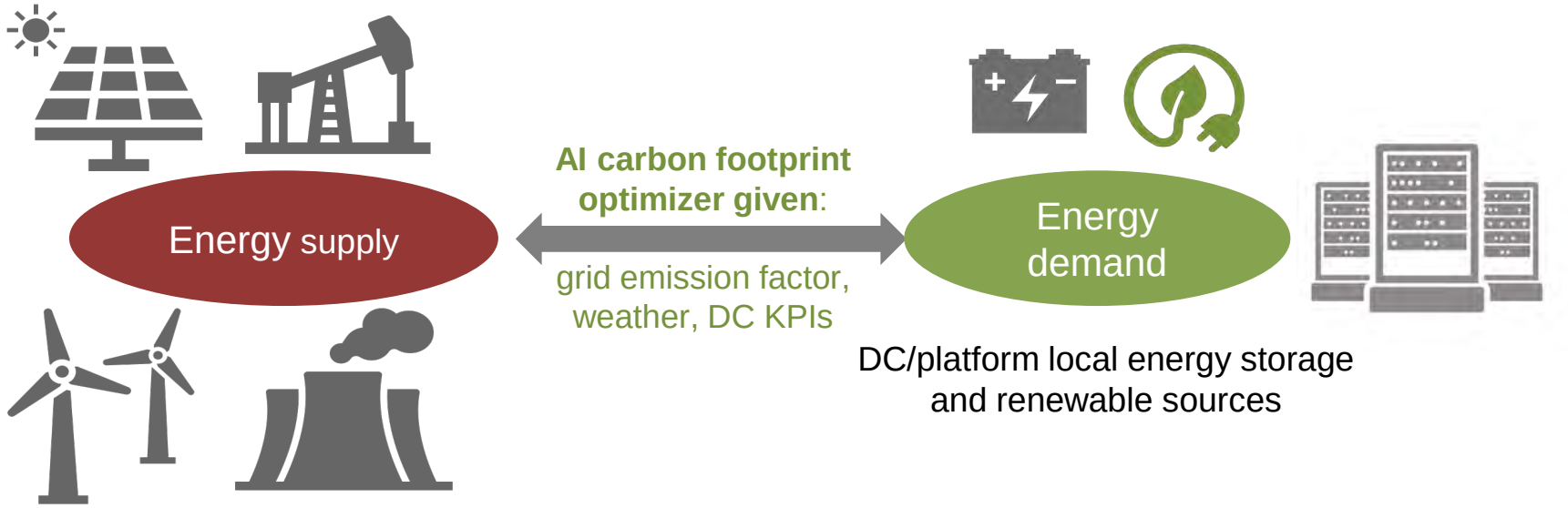
How to reduce the dominant factors for carbon emissions and cost in DCs?

$$Savings\ Emissions = Savings\ operational\ emissions - Embodied\ emissions$$

$$Savings\ Cost = Savings\ operational\ cost - Investments$$

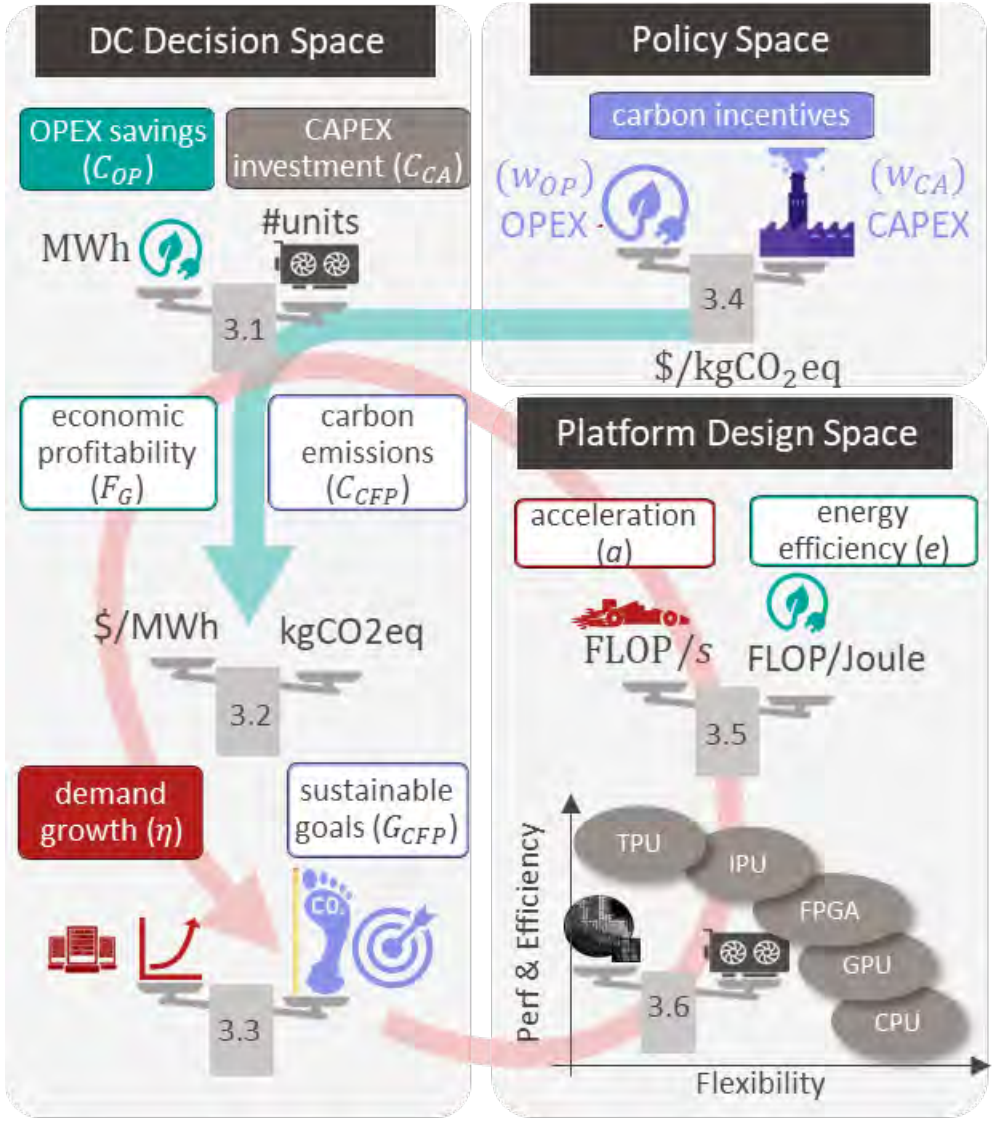
Build **optimized software** and **domain-specific accelerators** to make green computing economically feasible
And new ways to **design/manage DCs** (**for sustainability!**)

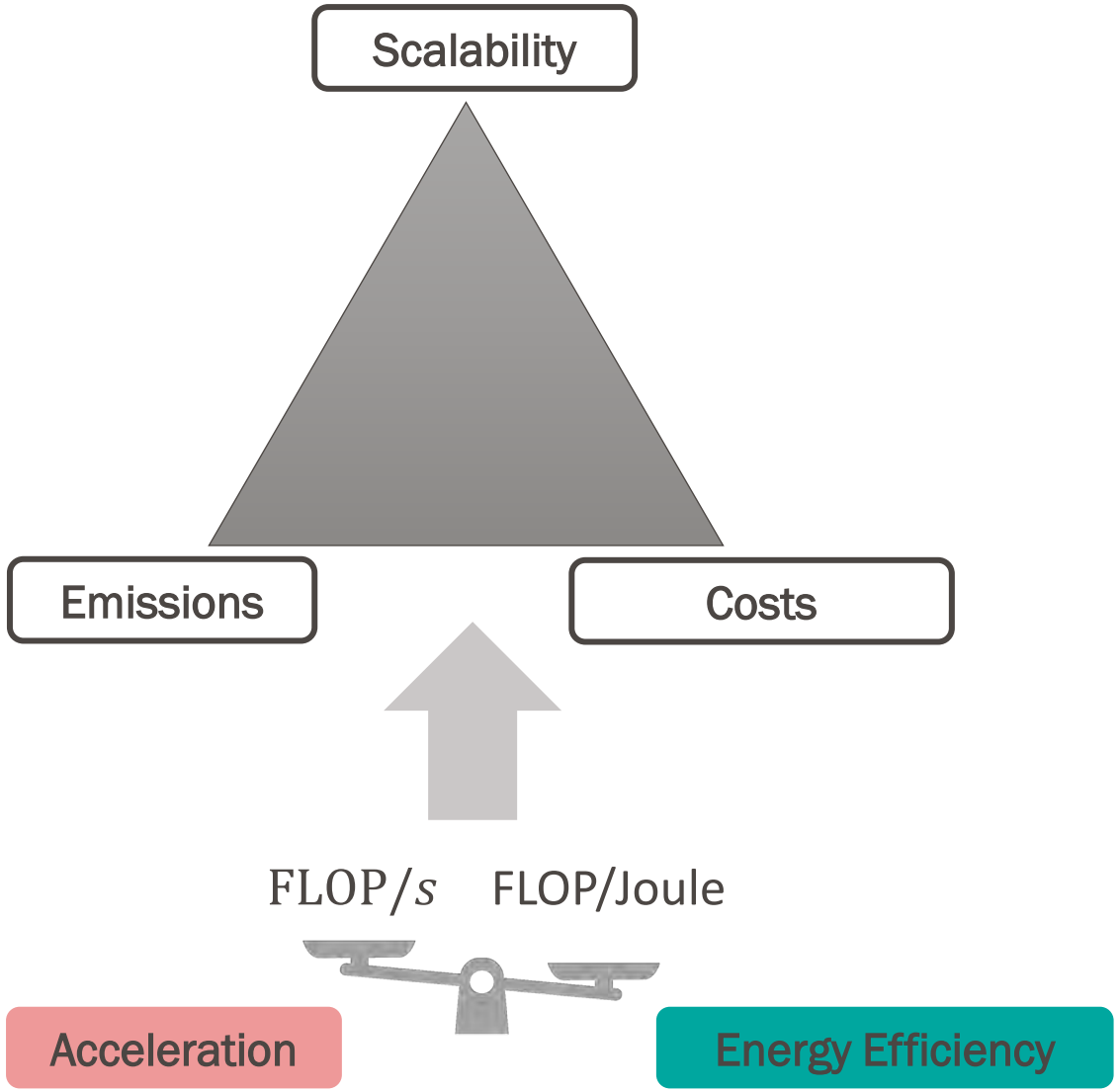
Key Research Questions in Sustainable Computing



- Which investments scenarios are economically viable to reduce the total DC CO2-eq by >50% before 2030?
- Which is the improvement factor (energy efficiency, acceleration, etc.) needed for a future platform to reduce CO2 without economic incentives?
- What is more economically sustainable, acceleration or energy efficiency to guide design space exploration for DCs in large case studies (AI, Astronomy and Genomics, etc.)?

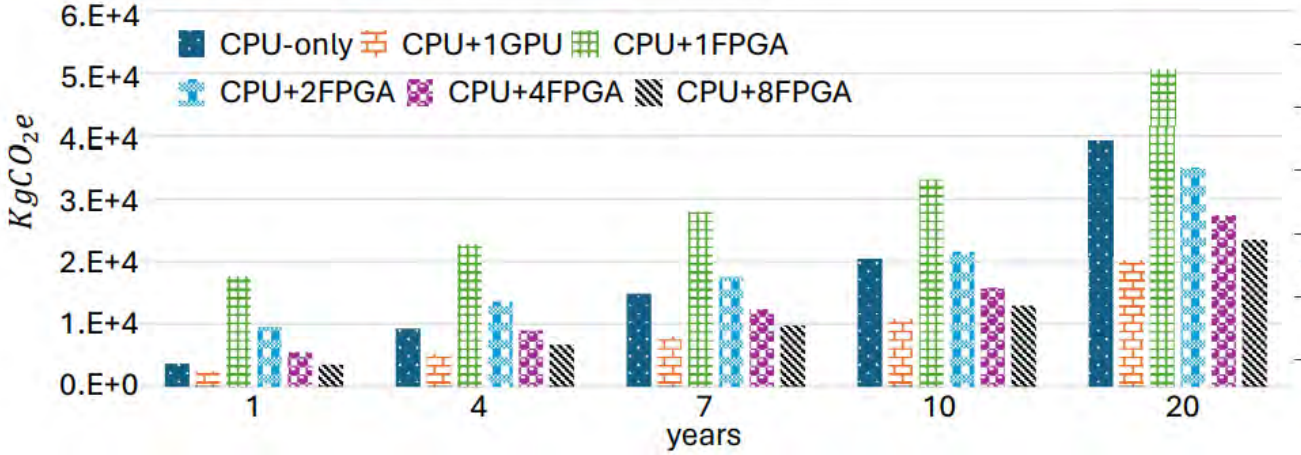
New CEO-DC Framework in a Nutshell



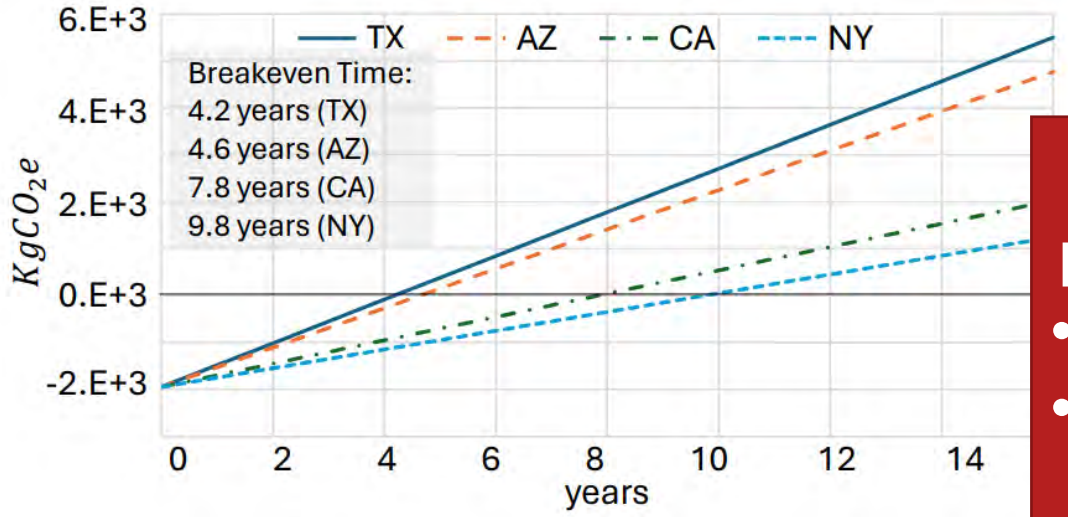


Rodríguez Álvarez, Rubén, Constantinescu, Denisa et al. "CEO-DC: An Actionable Framework to Close the Carbon Gap in HPC Data Centers." *arXiv preprint arXiv:2507.08923* (2025).

Carbon savings of upgrading vs. non-upgrading servers: US case study



	Xeon 8180	Xeon 8375	V100	A100	ZCU102
Latency (ms)	217.98	176.68	2.96	1.84	32.72
Power (W)	205	300	250	175	25
Static Power (W)	10	10	39	53	1
Framework	ONNX	ONNX	TensorRT	TensorRT	HeatViT



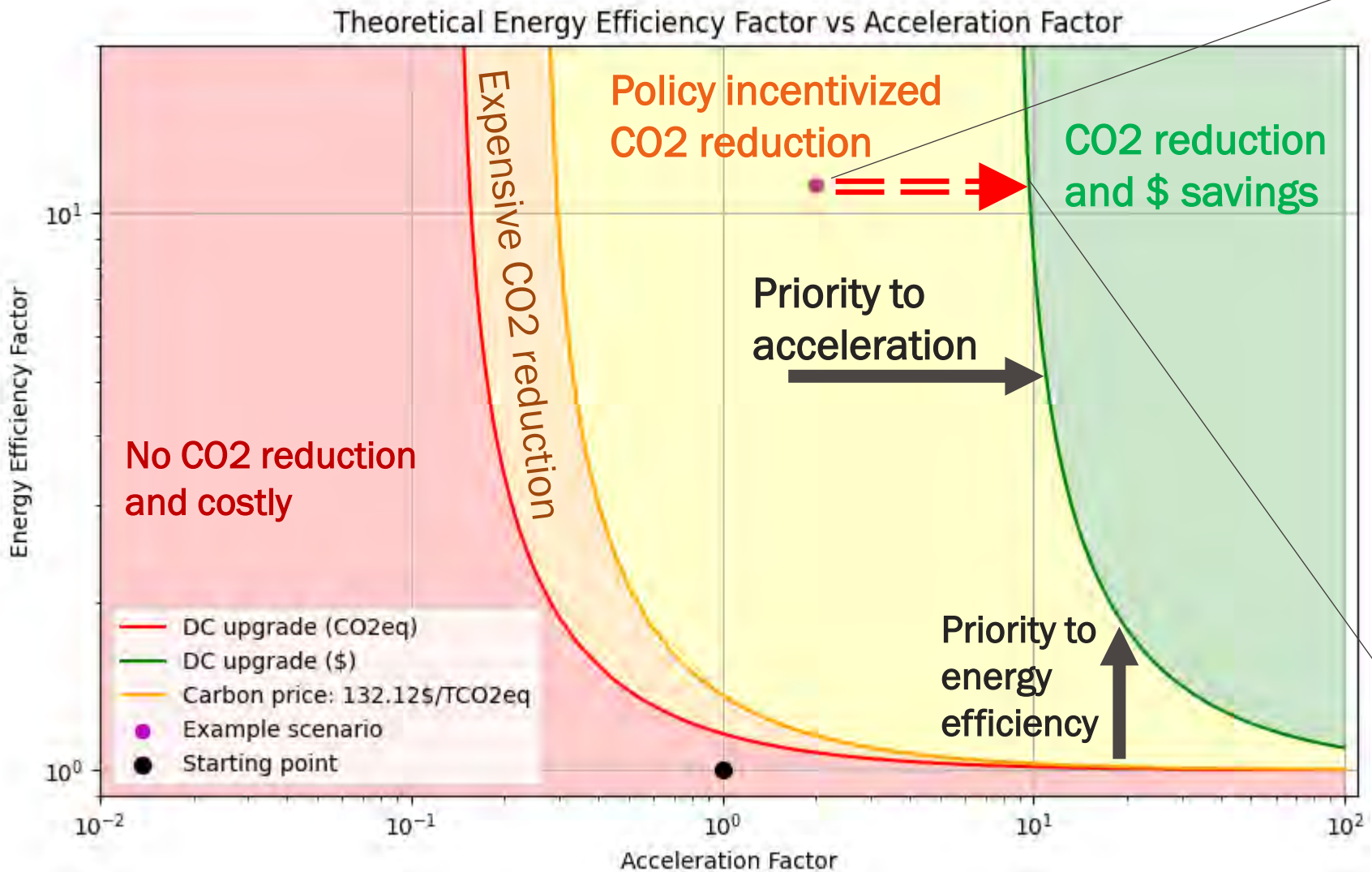
Carbon cost comparison among CPU-only, 1-GPU and 1-,2-,4-,8-FPGA servers in 4 different states in US

New server: lower energy consumption in oper. phase, but different carbon intensity per region:

- **4 to 10 years for breakeven point**
- **Two strategies (non-upgrading vs. upgrading) have the same overall carbon cost**

1. Ji, Shixin, et al. "SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators." *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2024. https://arxiv.org/pdf/2401.06270v3

Analysis of upgrading a Swiss DC: acceleration vs. energy efficiency



Large DC (50,000+ square feet)
Location: Switzerland
Tier 4 – Uptime 99.995%

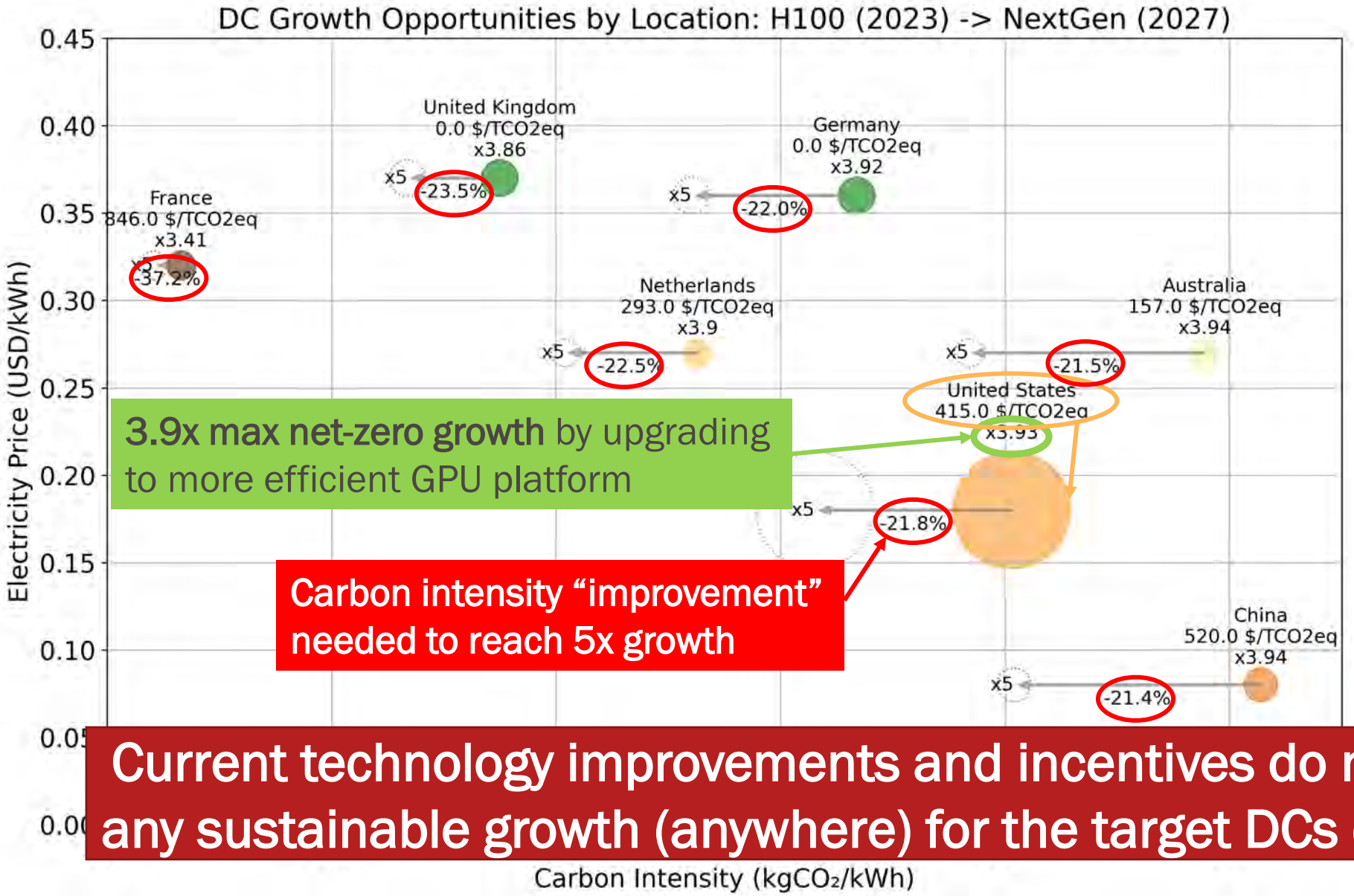
Custom acceleration
Accelerated workloads: 40%
Number of accel.: 509
Acceleration: x2
Energy efficiency: x11.2

CO2 [MTCO2eq]
(Emb) 0.01 < 0.16 (Op)
CO2 savings lifetime: 0.15
Global CO2 reduction: 33.62%

Initial Economic Analysis [M$]
Payback: 9.7 years (ROI: 0.1)
(Invest) 1.53 > 0.32 (Energy cost)
Gains lifetime: -1.21 M$

Trade-off [$/TCO2eq]
Carbon price: 132.12
Incentive required: 8.13

Carbon gap: energy efficiency vs. demand growth



Case study

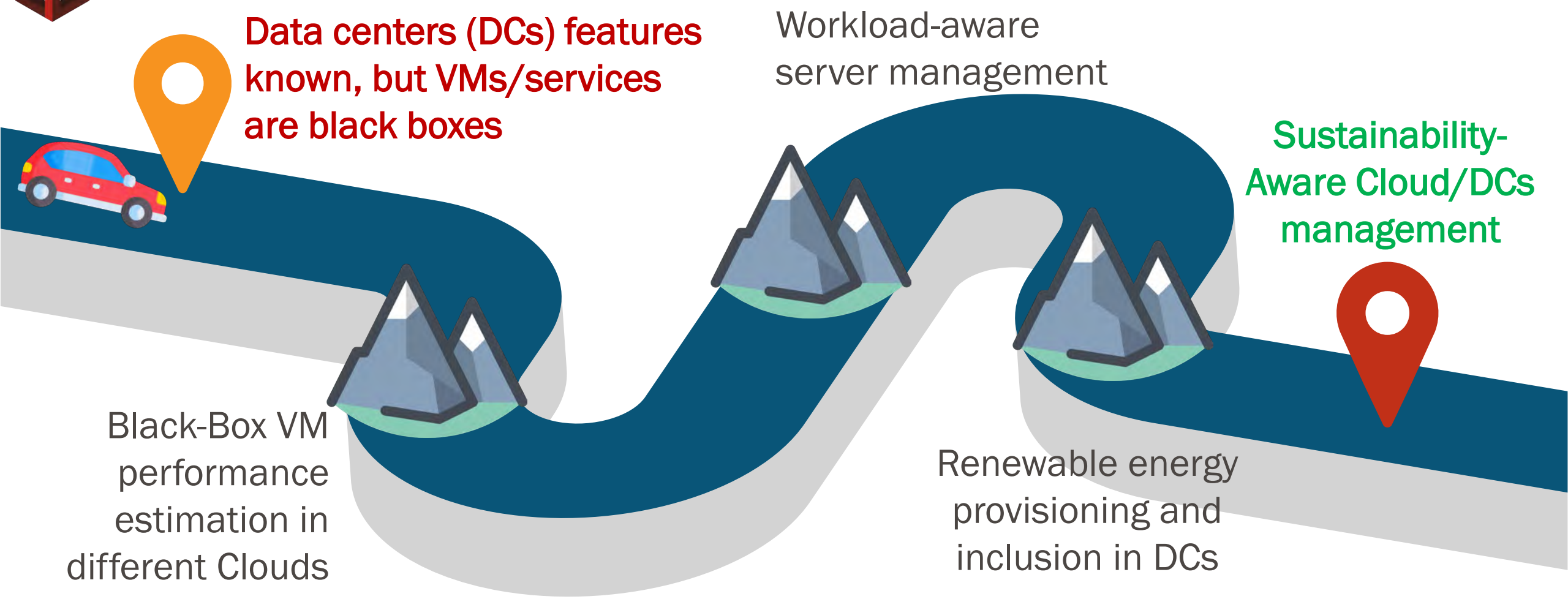
Close carbon gap for 5x growth

Means

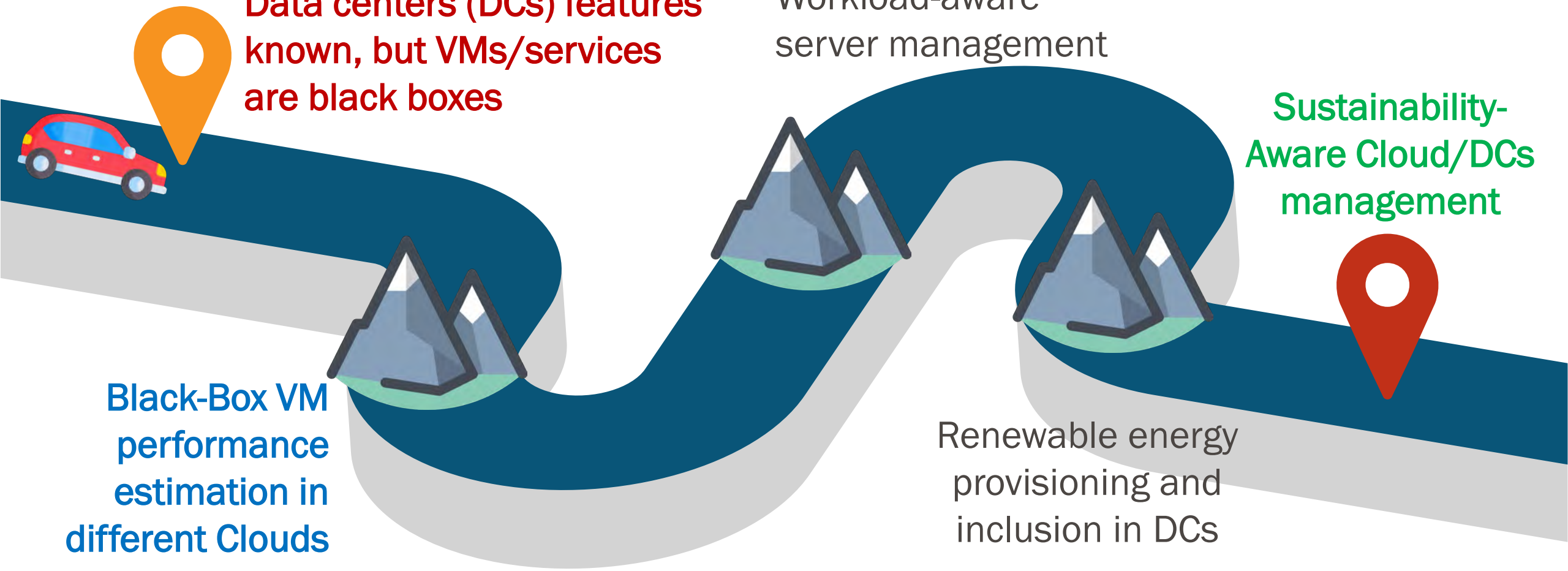
1. Upgrade to NextGen energy efficient platforms
2. Carbon tax incentives
3. Clean the electricity

Current technology improvements and incentives do not allow any sustainable growth (anywhere) for the target DCs demand!

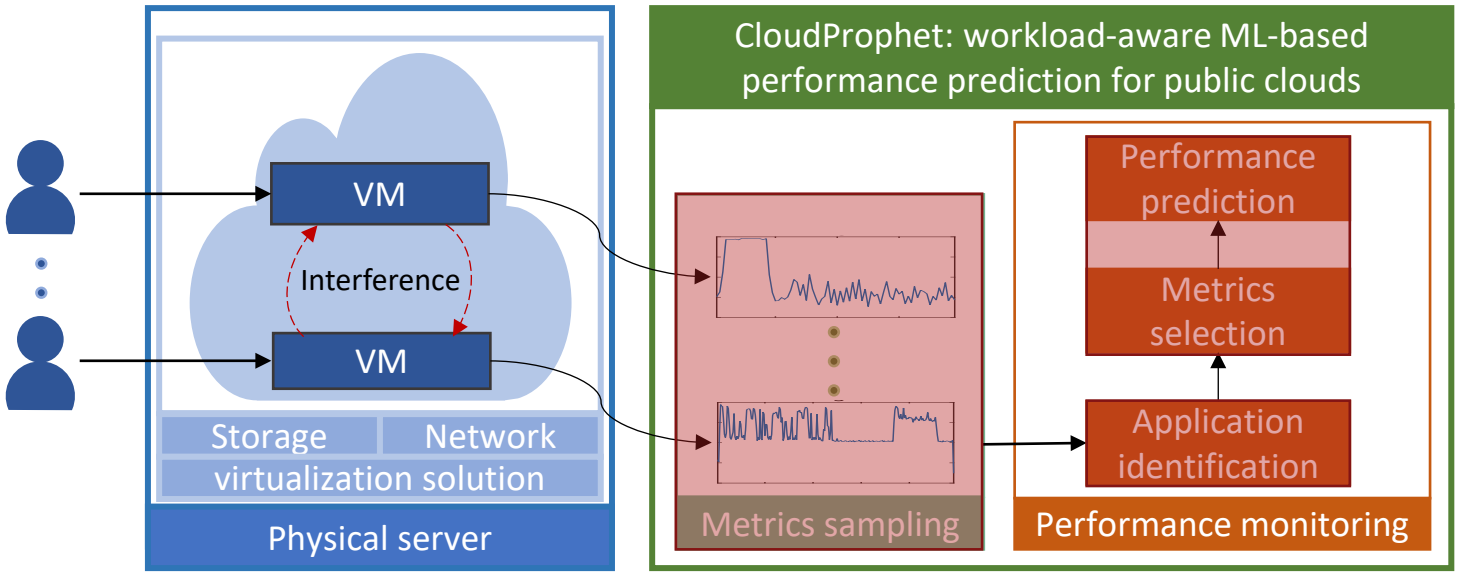
Challenges in our Path to a Sustainable Cloud



Challenges in our Path to a Sustainable Cloud



CloudProphet: Black-Box VM Performance Management



- Main steps:
  1. Monitoring data (black box)
  2. Application identification
  3. Performance prediction

CloudProphet on IEEE
[Huang et al., TSUSC 2024]

CloudProphet-Dataset repo

Monitoring data needs limited

- A few low-level hardware metrics are required

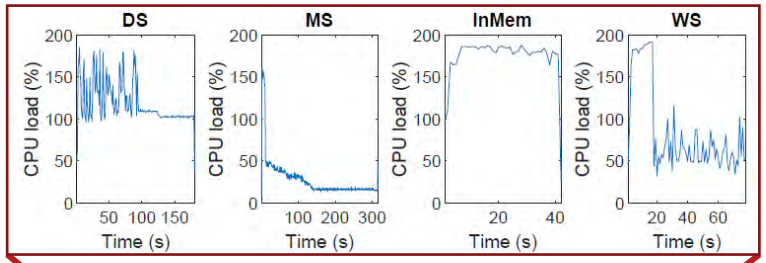
Example of monitored metrics

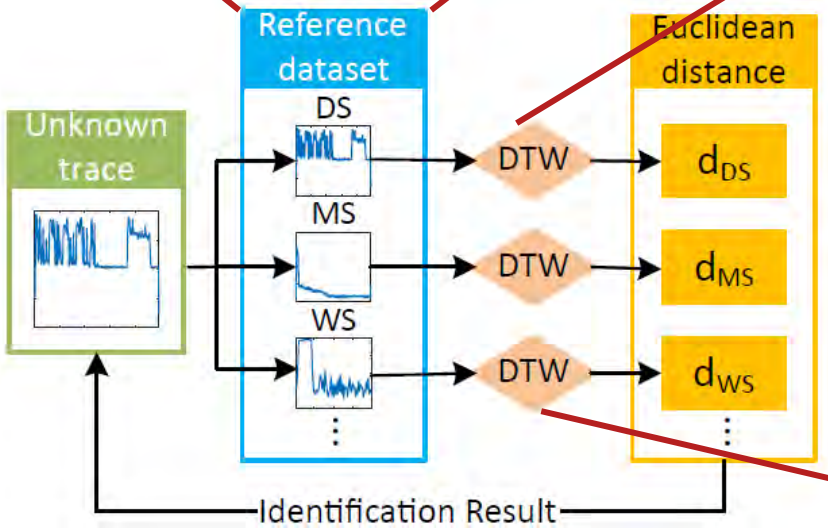
Category	Typical extracted metrics
CPU	CPU utilization level (%) Executed instructions (#)
Memory	LLC misses (#) Available memory space (KB) Read requests issued for disk usage (#)
Network	Received packets (Bytes) Sent packets (Bytes)

- Follow the black box assumption:
  - **No need to access the running application** inside the VM

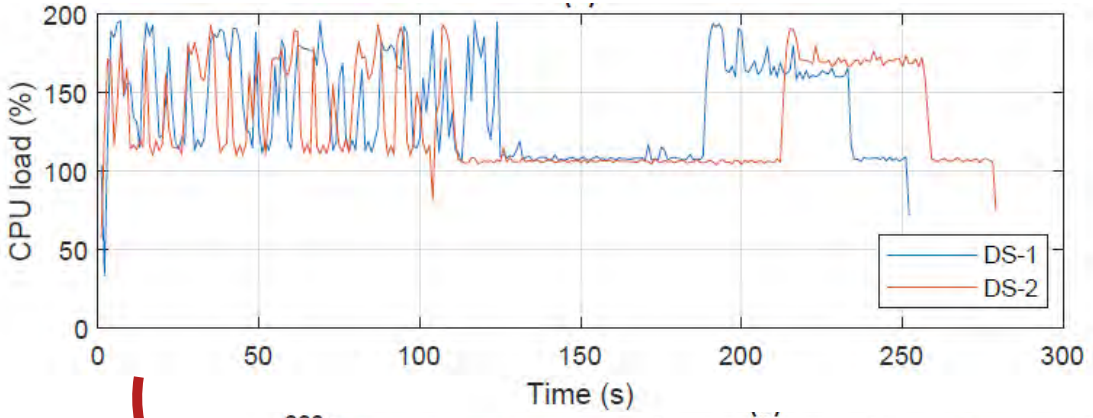
Application Identification

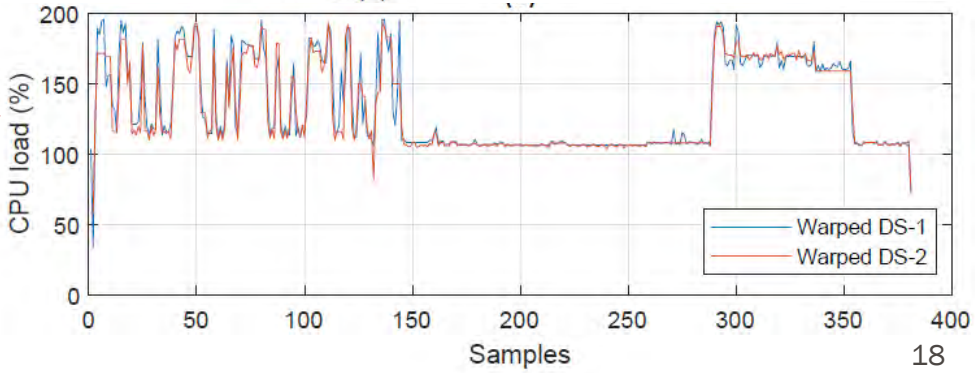
- Offline:
  - Create the reference dataset (Fingerprint)
- Online:
  - Dynamic time warping (DTW) -based identification



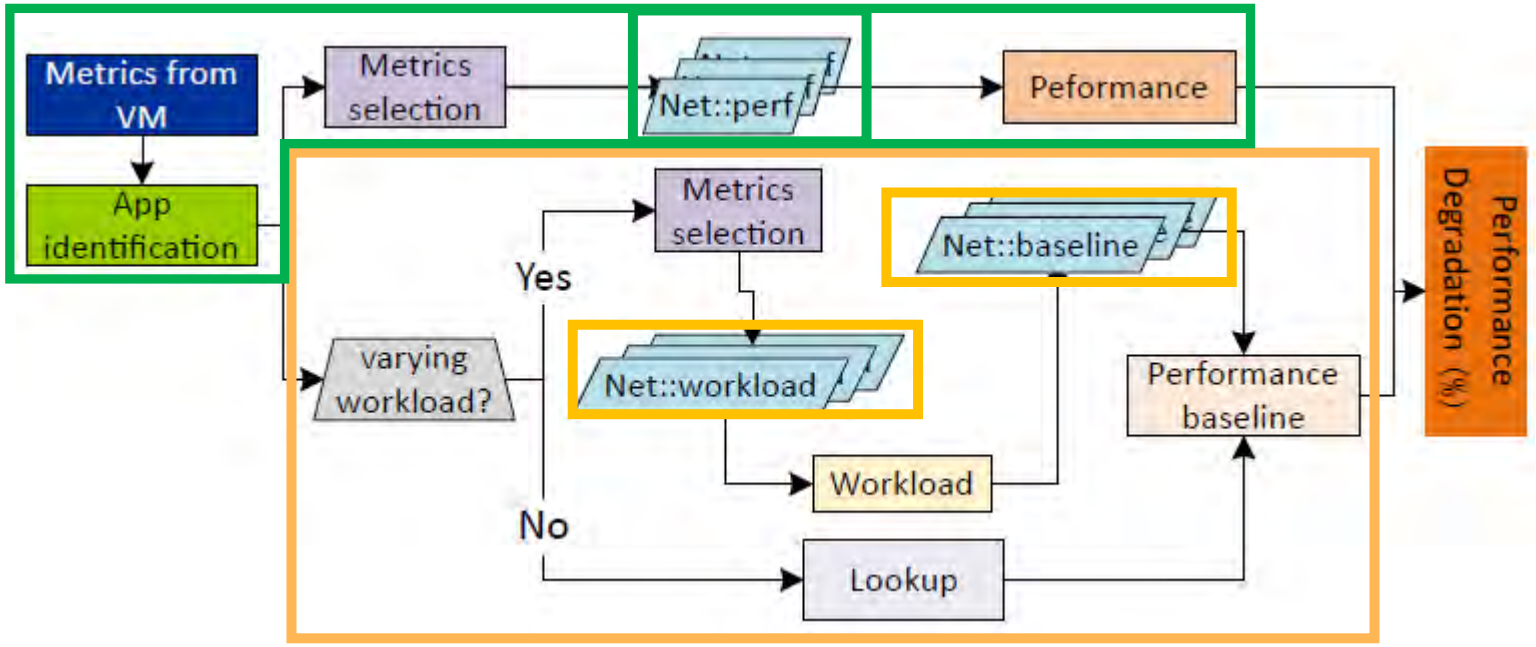


Dynamic time warping (DTW)

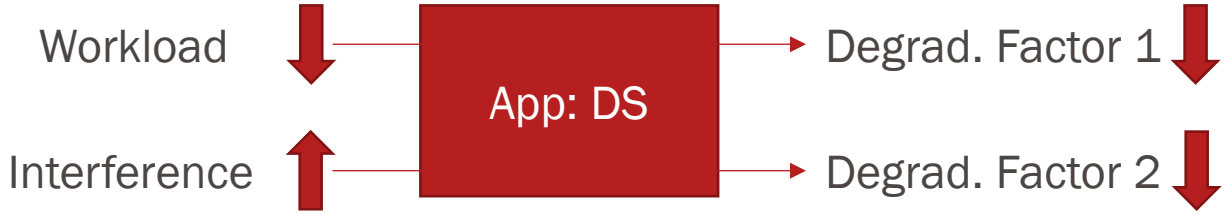




Workload-Aware Performance Prediction



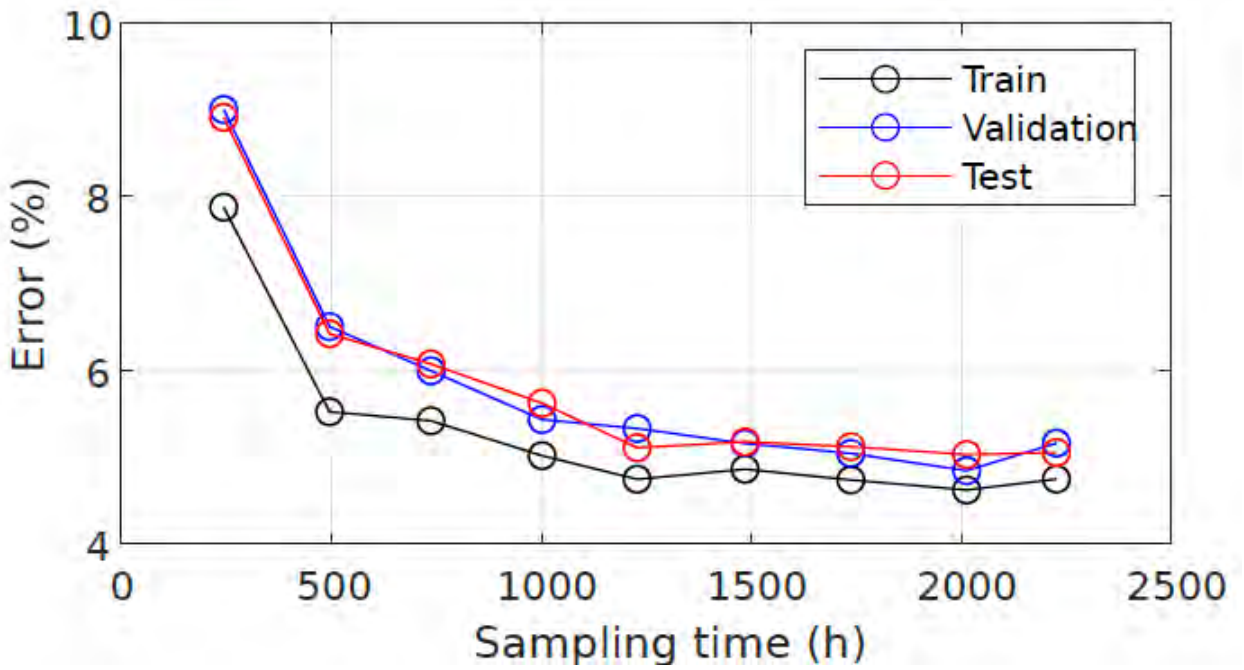
Both user interaction and interference influence the performance level of the application!



1. Performance Prediction
   - Metrics selection
   - **Neural network (NN) for each class of application**

2. Performance Degradation Prediction
   - Workload prediction
   - Initial baseline prediction dynamically readapted with an **additional NN for inference detection**

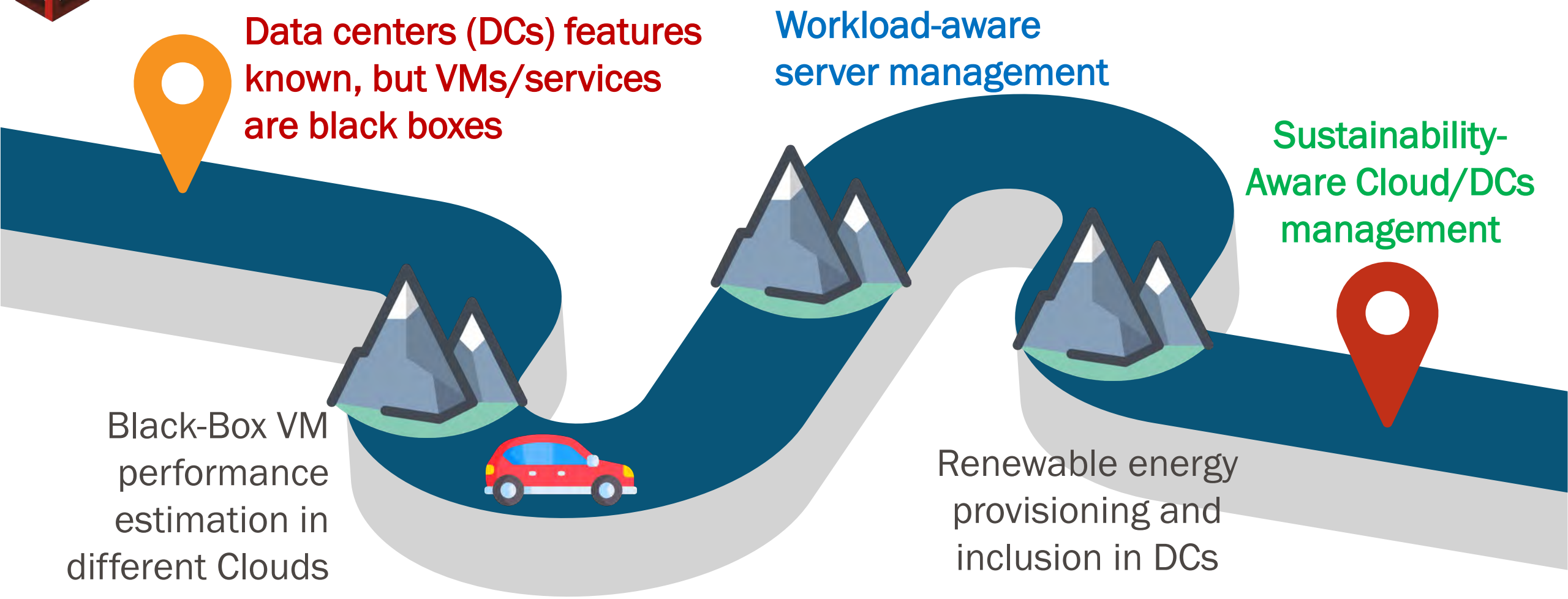
Accurate Performance Prediction of CloudProphet



Trade-off between sampling time and prediction accuracy

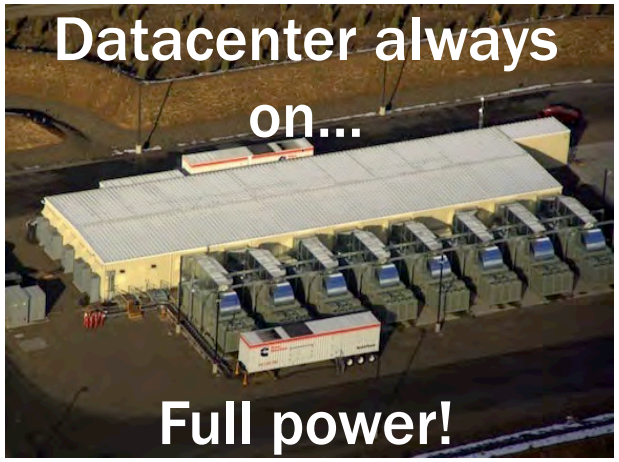
- **Less than 7% prediction error** after 20 days, better with more samples
- **5% performance prediction error** after 2 months of operation

Challenges in our Path to a Sustainable Cloud

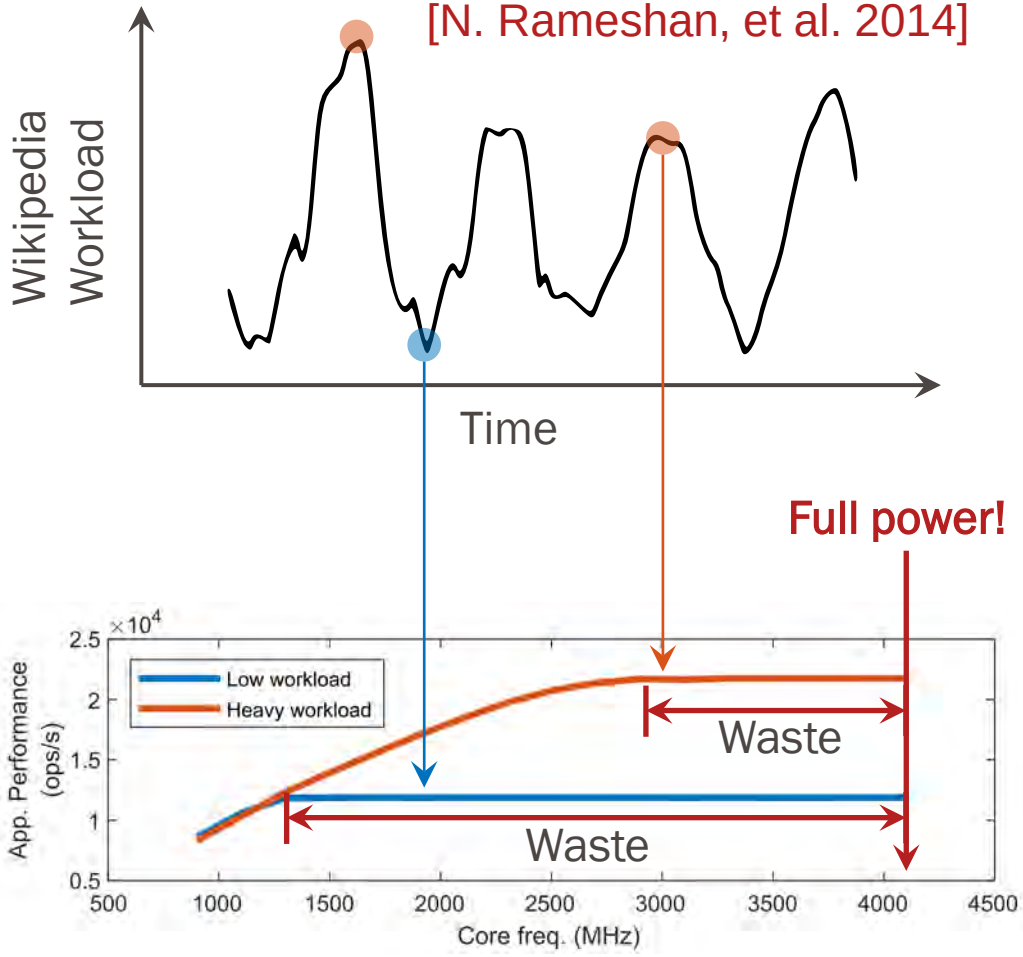


Large Resource Wasted in Cloud Designs!

- Worse case resource provisioning paradigm: variable demand



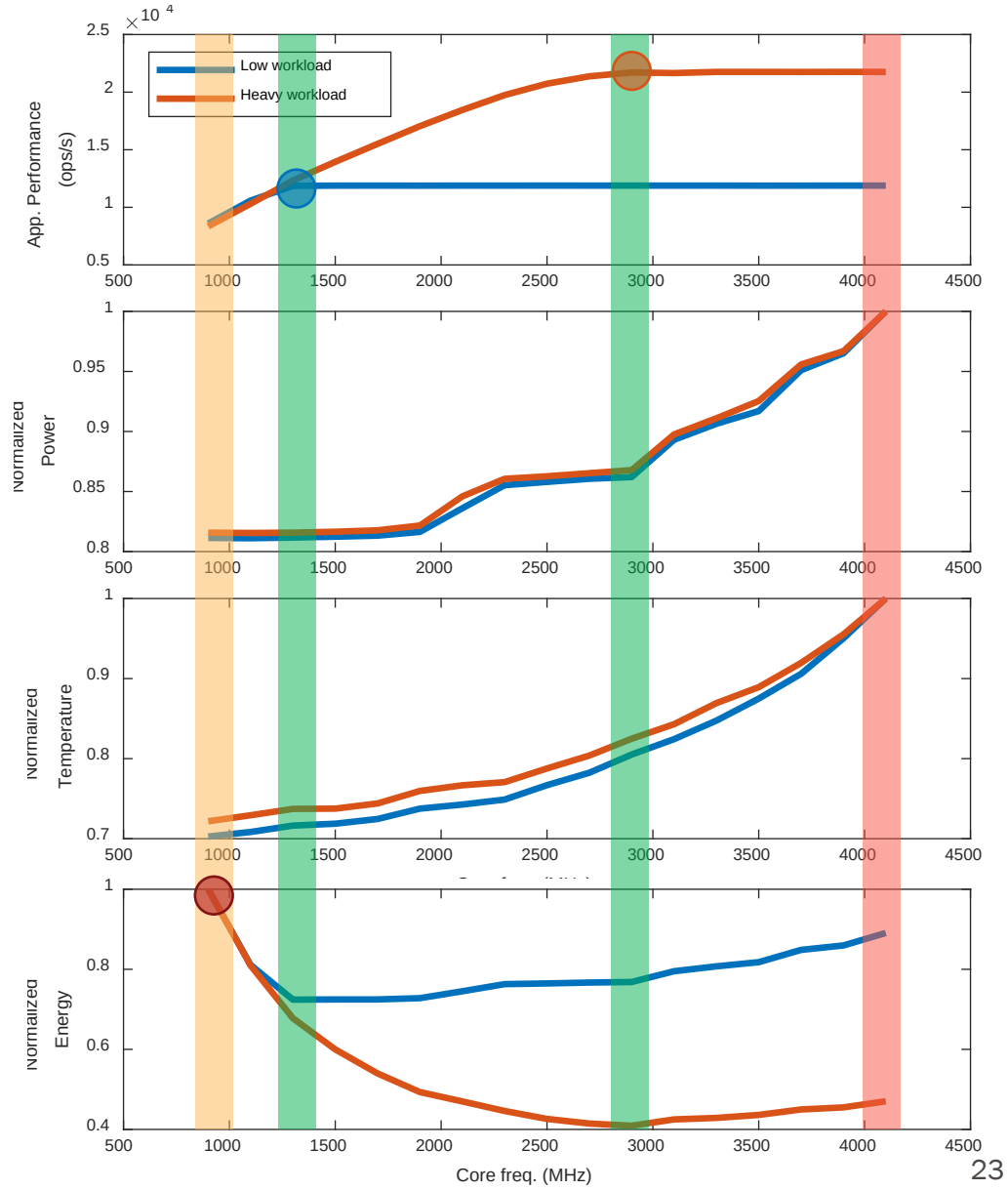
[N. Rameshan, et al. 2014]



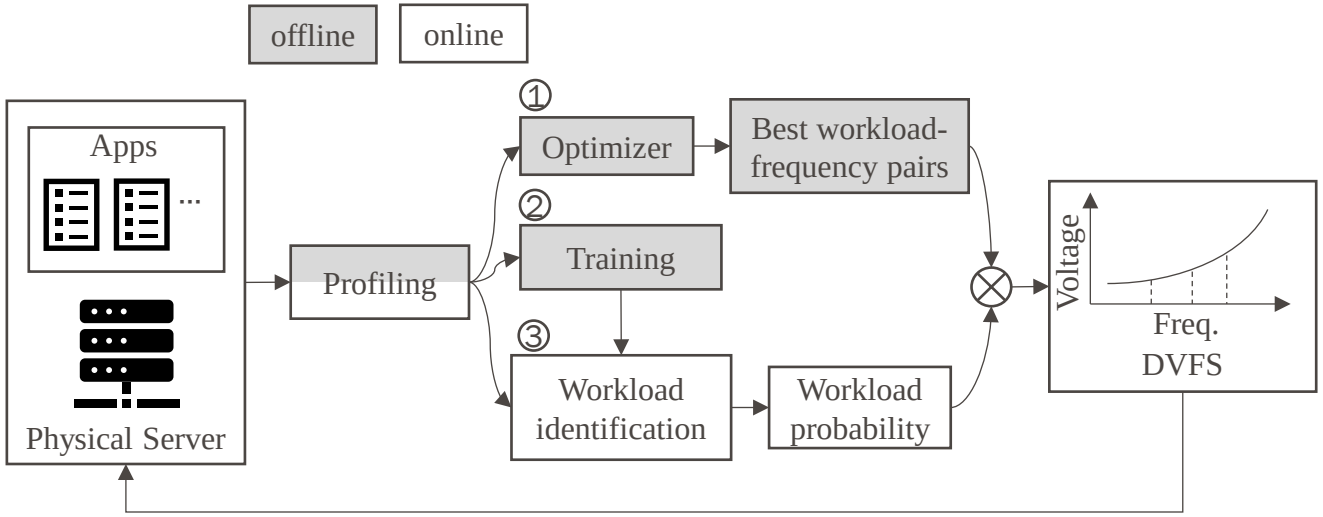
- Hint: Appropriate **frequency scaling approach** can significantly reduce energy use in data centers
  - But VDD scaling is required (simple cores!)

But Linux/Proprietary Scaling Governors Are Not Optimal

- *powersave:* 
- *Performance* and *intel*: 
- Take home messages:
  - **Linux/propr. scaling governors are clearly sub-optimal**
  - ***powersave* governor is the most energy-intensive one**



GreenDVFS: Workload-Aware Server Management



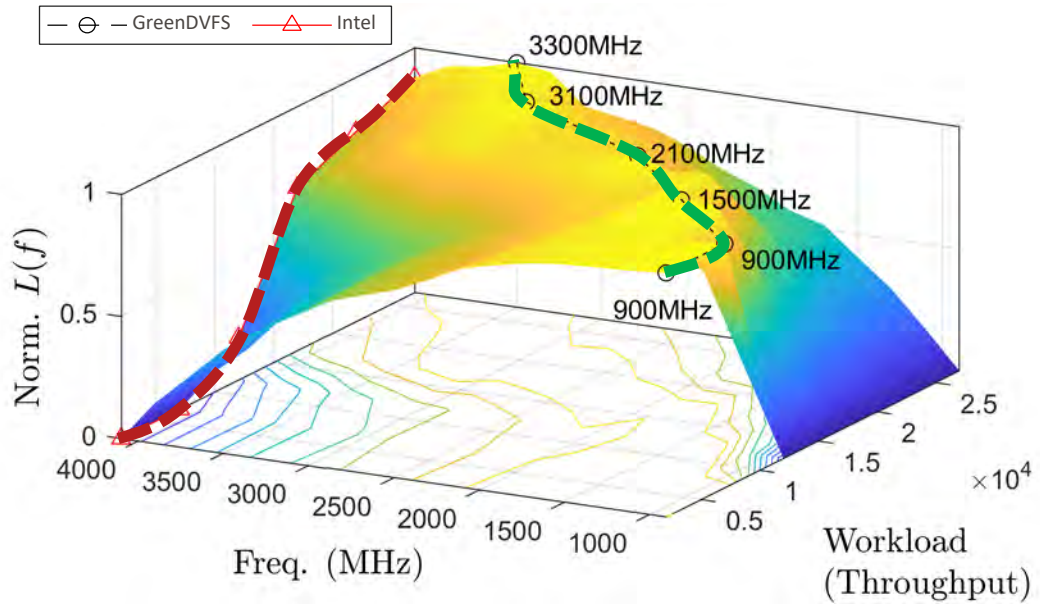
GreenDVFS (Camera-ready)

[Huang et al., CCGrid'24]

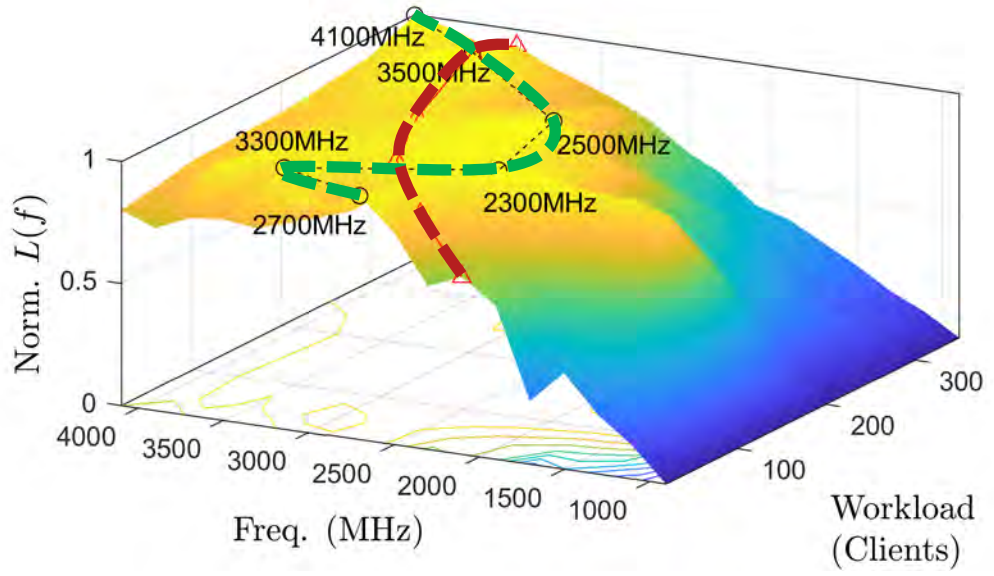
- An optimizer to select the best workload-frequency pairs
  - [offline] Tuned per server family (tech. dependent)
- Recurrent NN for management: Modified Long short-term memory (LSTM)
  - [offline] Customized training scheme
  - [online] Runtime workload identification (CloudProphet) and DVFS setting

Final Optimizer for Best Workload-Frequency Energy: Take it easy when going uphill!

- $L(f)$: optimizes performance, power, and temperature
  - Designed per server, fast tunable to different applications



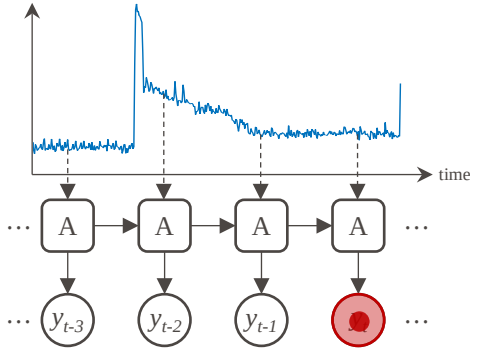
App: Data serving

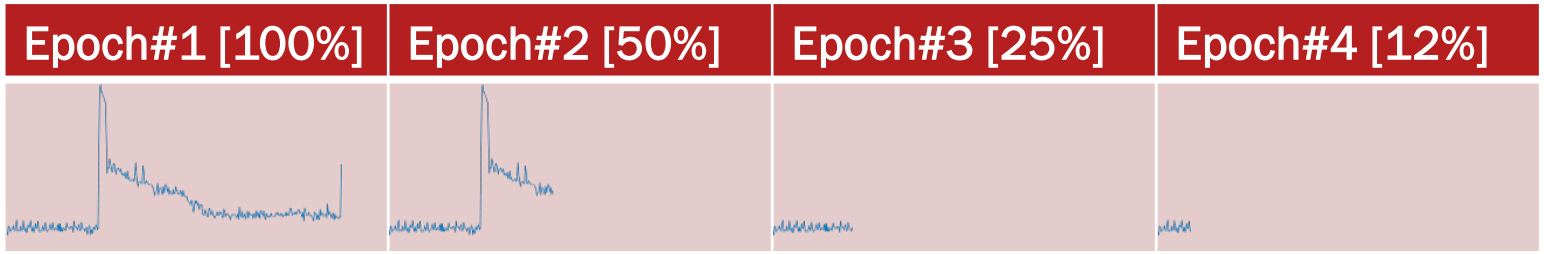


App: Web Search

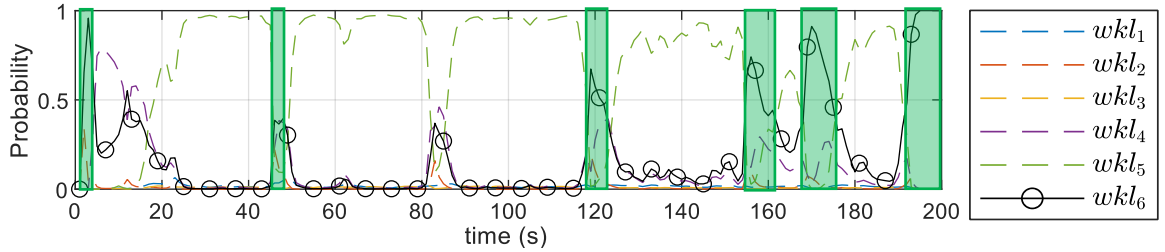
Adapted LSTM-based Early Workload (DVFS) Tuning

- Traditional LSTM training scheme puts much emphasis on the latest prediction results

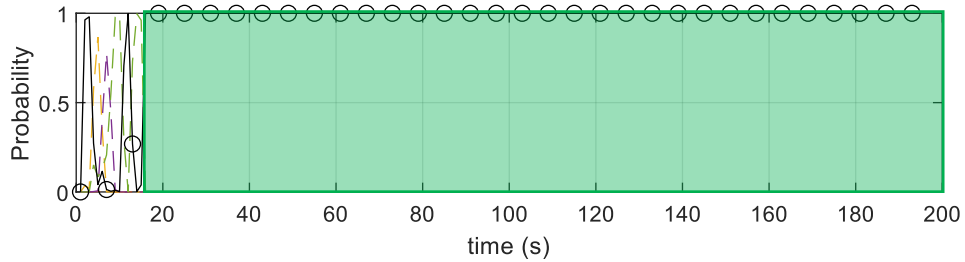




- New proposed LSTM training scheme: early phases are key
  - Keep only 50% previous training epoch for fast tuning with new data

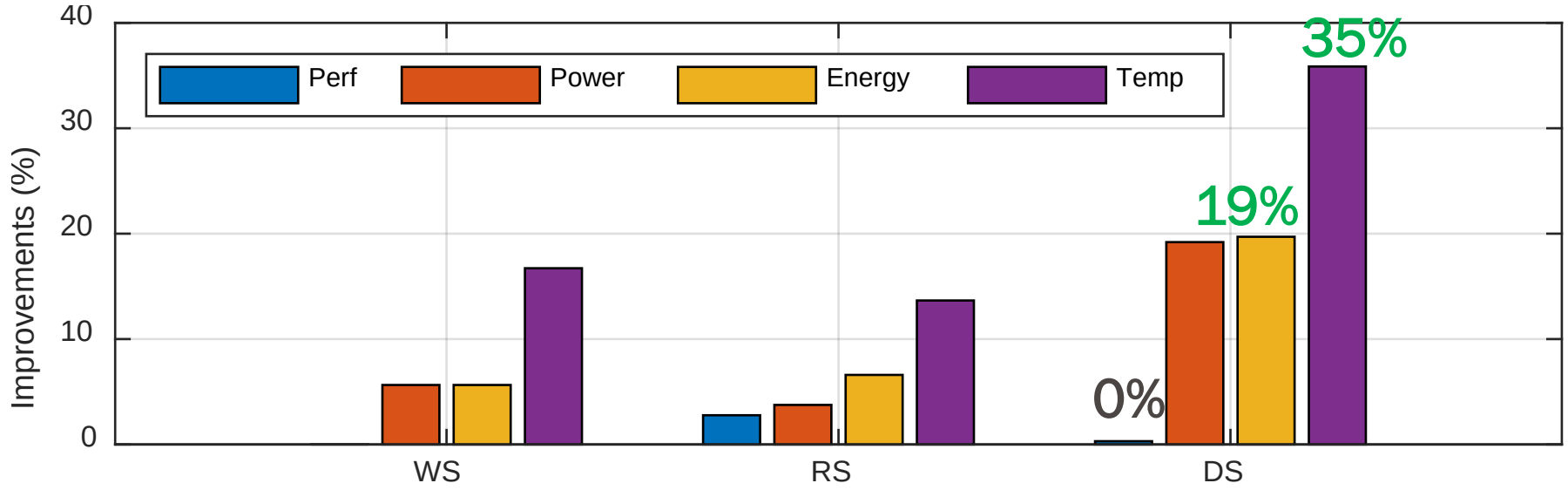


Traditional LSTM training



Customized LSTM training scheme

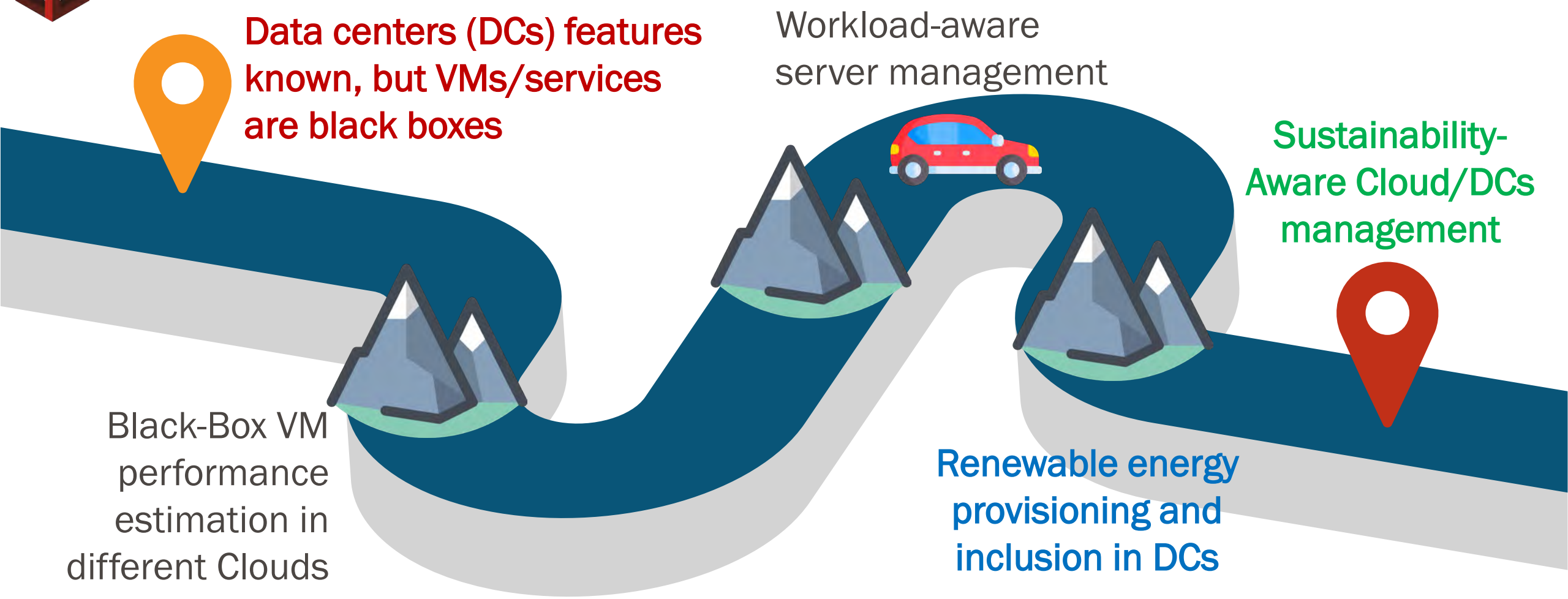
GreenDVFS for Energy-Efficient Server Management



- No performance loss
- Up to **19% less energy** consumed
- Up to **35% lower temperature** in operation

And additional savings possible if fine-grained and fast voltage scaling is possible: open-source RISC-V servers coming up (SwissChips!)

Challenges in our Path to a Sustainable Cloud



Rethink DC Design: New DC and Experimental Facility on Campus to Explore Sustainable Cloud Computing

- Merging EPFL central heating plant and DC
  - Financial support of AVP-CP/VPA, VPO, and donations from the industrial affiliates of EcoCloud

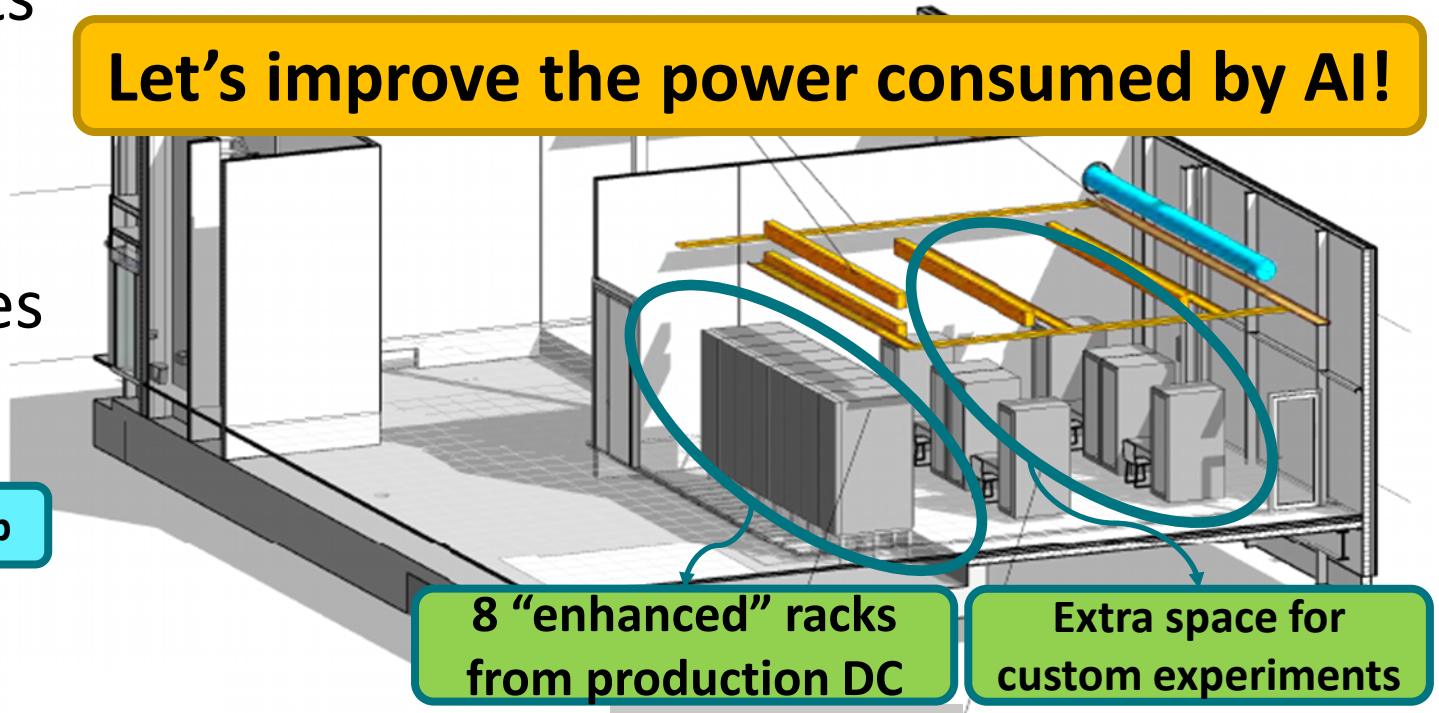
- Support multi-disciplinary research on energy-efficient DC and computing systems design: CS, EE, ME, etc. working together
  - Kuma: New supercomputer to enable cutting-edge and sustainable research
  - Heating Bits: DCs integrating heating and cooling supply of local districts
  - U

At EcoCloud-EPFL, we look forward to sharing this DC facility with new and interesting projects with the EPI partners!

EcoCloud Sustainable Experim. Computing Facility in EPFL DC

- ~150 m² of space for experiments on sustainable computing
  - Recycled racks/donations
- Experimental support: two spaces
  - 50KW per rack/2.5m rack
  - Monitoring: energy, tem
  - Cooling: air or water coo

Let's improve the power consumed by AI!

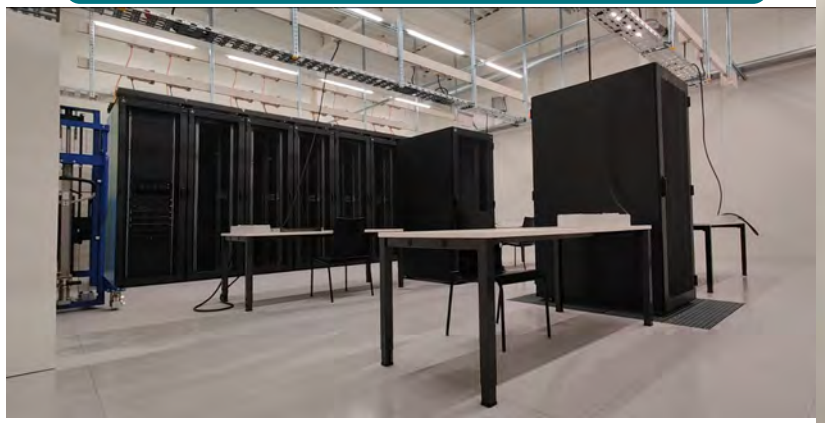


8 "enhanced" racks from production DC

Extra space for custom experiments

Controlled setup

Racks with air/water passive cooling



Underground water exchangers

Full supervision integrated with EPFL systems

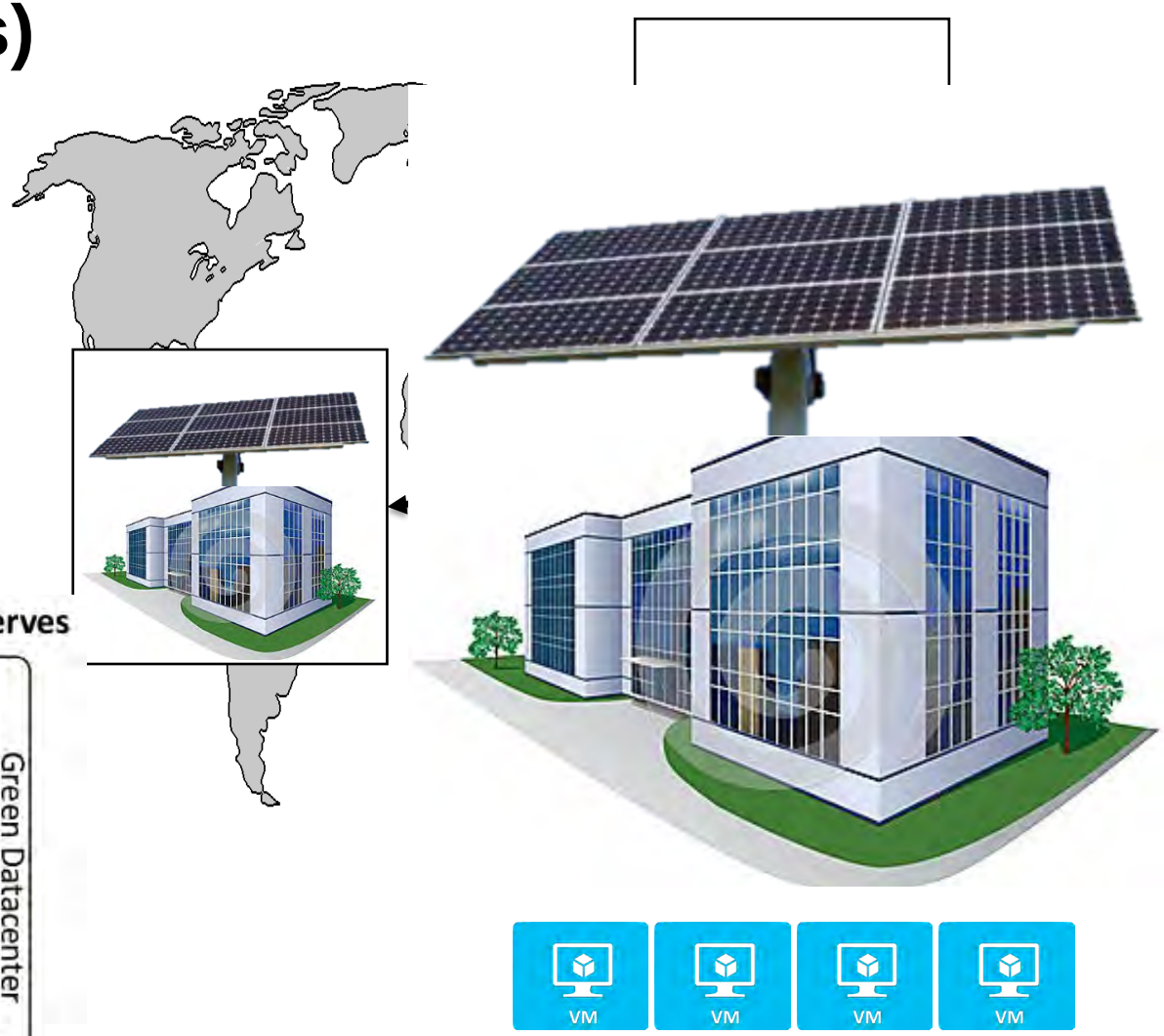
Rethink DC design: Detailed monitoring/manag. + Liquid Cooling for energy-efficient computing

- **RCP – Water-cooled doors for AI/ML research**
  - 383 GPUs - H100, A100, and V100 (55 nodes)
- **Kuma – EPFL's water-cooled supercomputer**
  - 336 H100 GPUs (84 nodes), Nvlink (900 GB/s)

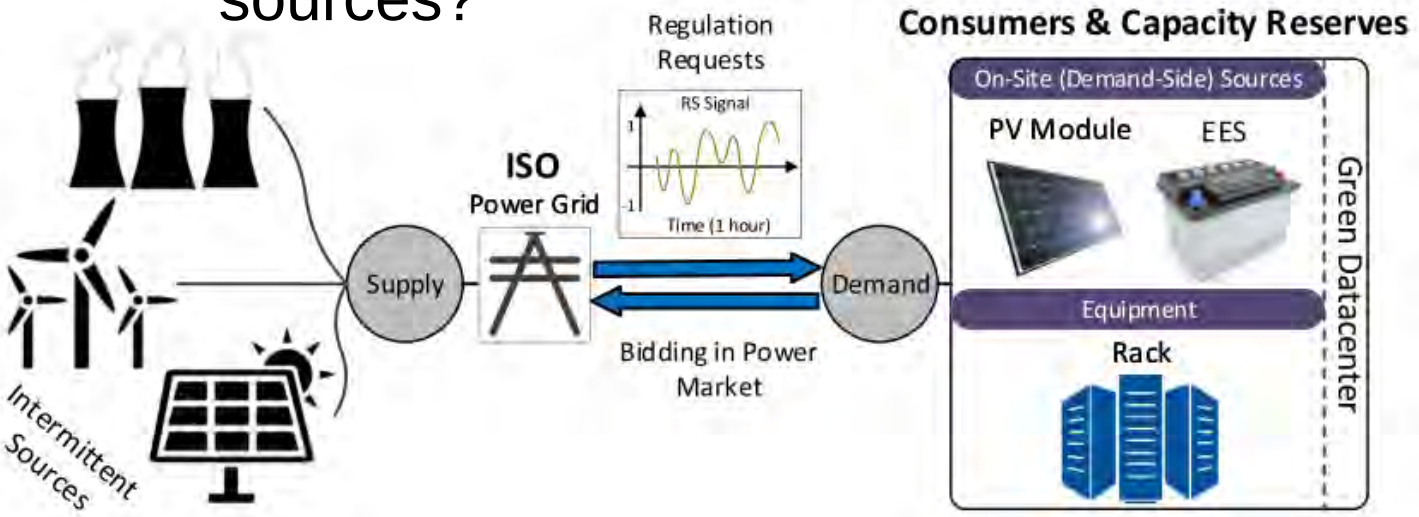
Ranked no. 23 in Green500: 54.9 Gflops / Watt
(#10 for academic institutions)

Include Sustainable Energy Sources: DCs Location

- **Geo-distributed data centers (DCs)**
  - Multiple DCs in different locations connected through network
  - How to allocate VMs to different DCs?

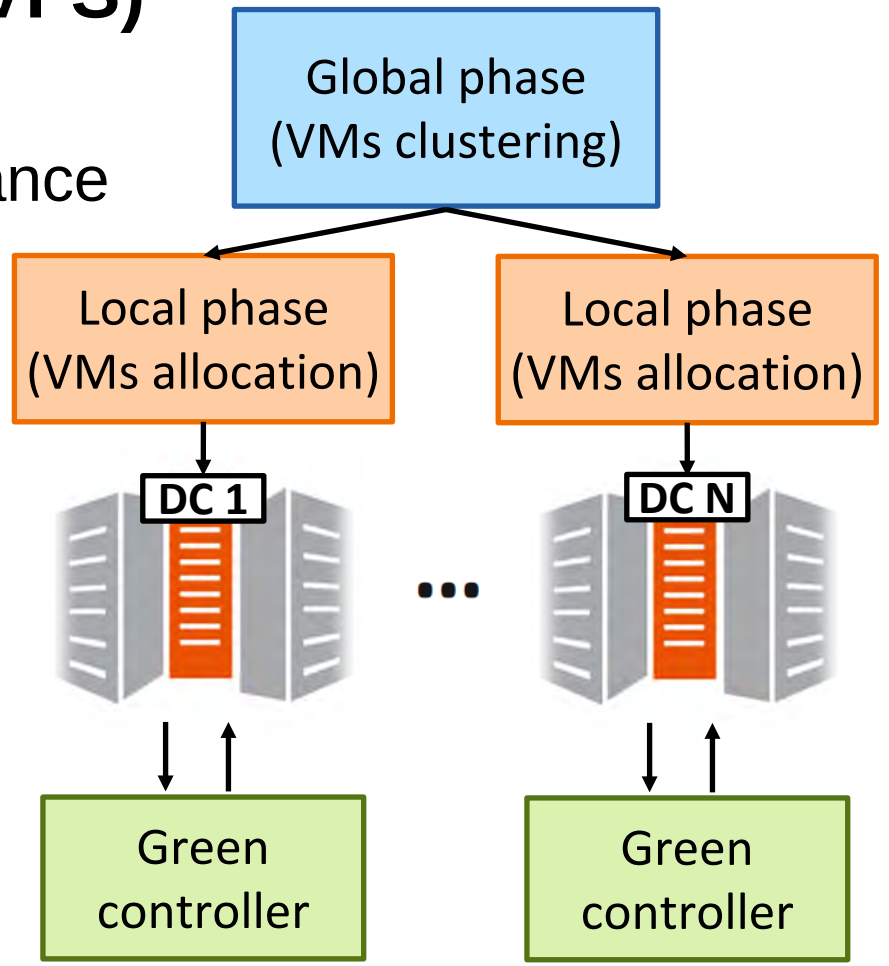


- **Ideal placement for green DCs**
  - How to manage renewable energy sources?



ECOGreen: Sustainability-Aware Renewable Energy Management

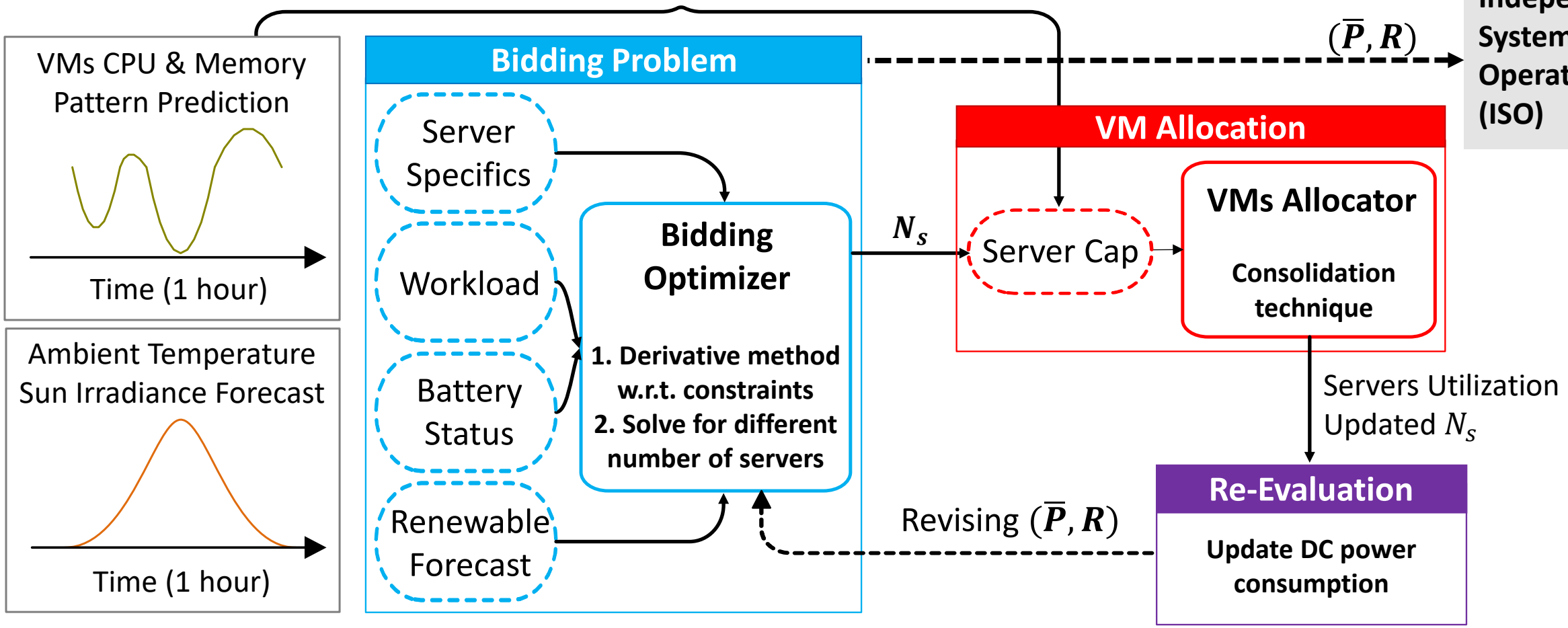
- **DCs/VM manag. (CloudProphet + GreenDVFS)**
  - Global phase: clustering VMs into DCs
  - Local phase: VMs allocation for minimum performance degradation
- **ECOGreen: Low-complexity green energy controller**
  - Management of renewable energy
  - Add batteries in DCs: charge / discharge decisions



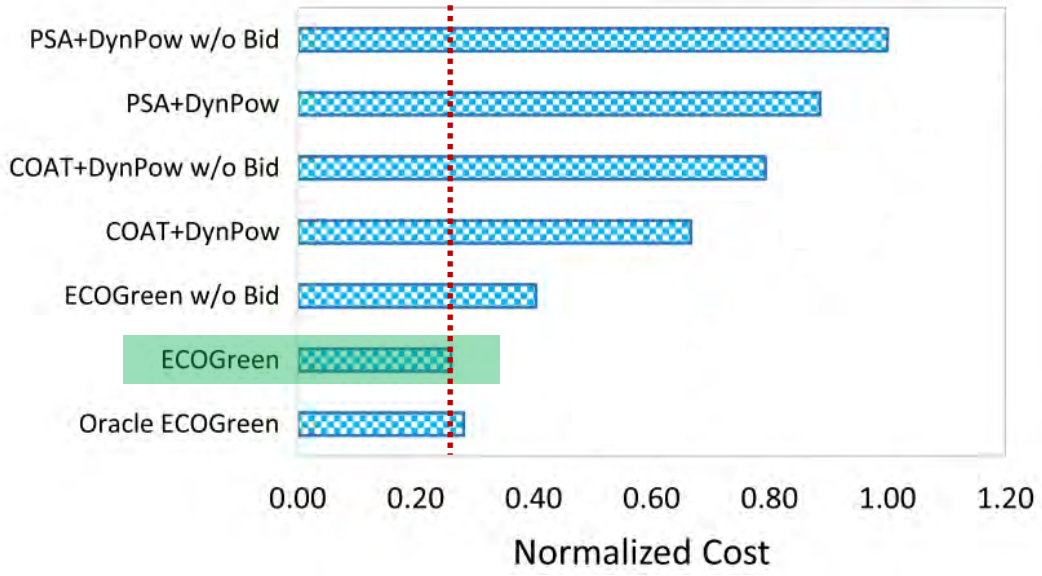
ECOGreen Energy Controller
[Pahlevan et al., TSUSC 2020]

ECOGreen: Proposed Strategy

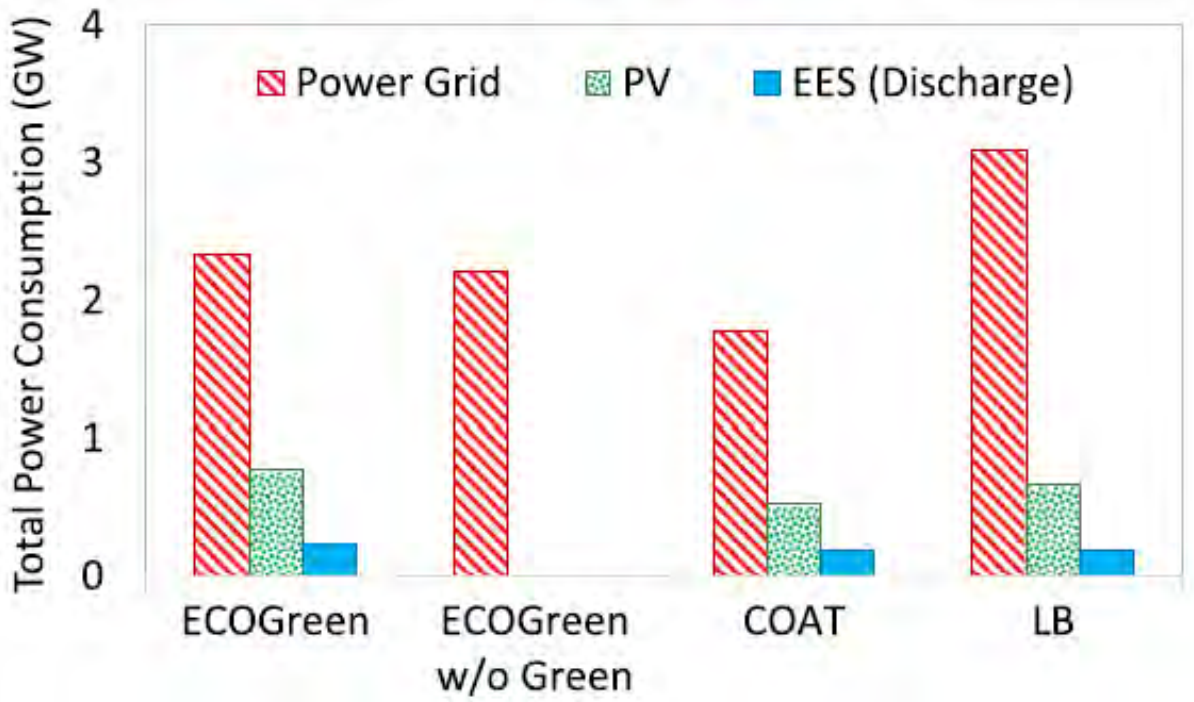
- **Hour-ahead power market (bidding)**



ECOGreen for Green and Sustainable DC



Normalized monetary cost (1-week time horizon)



Different power supply sources (1-week time horizon)

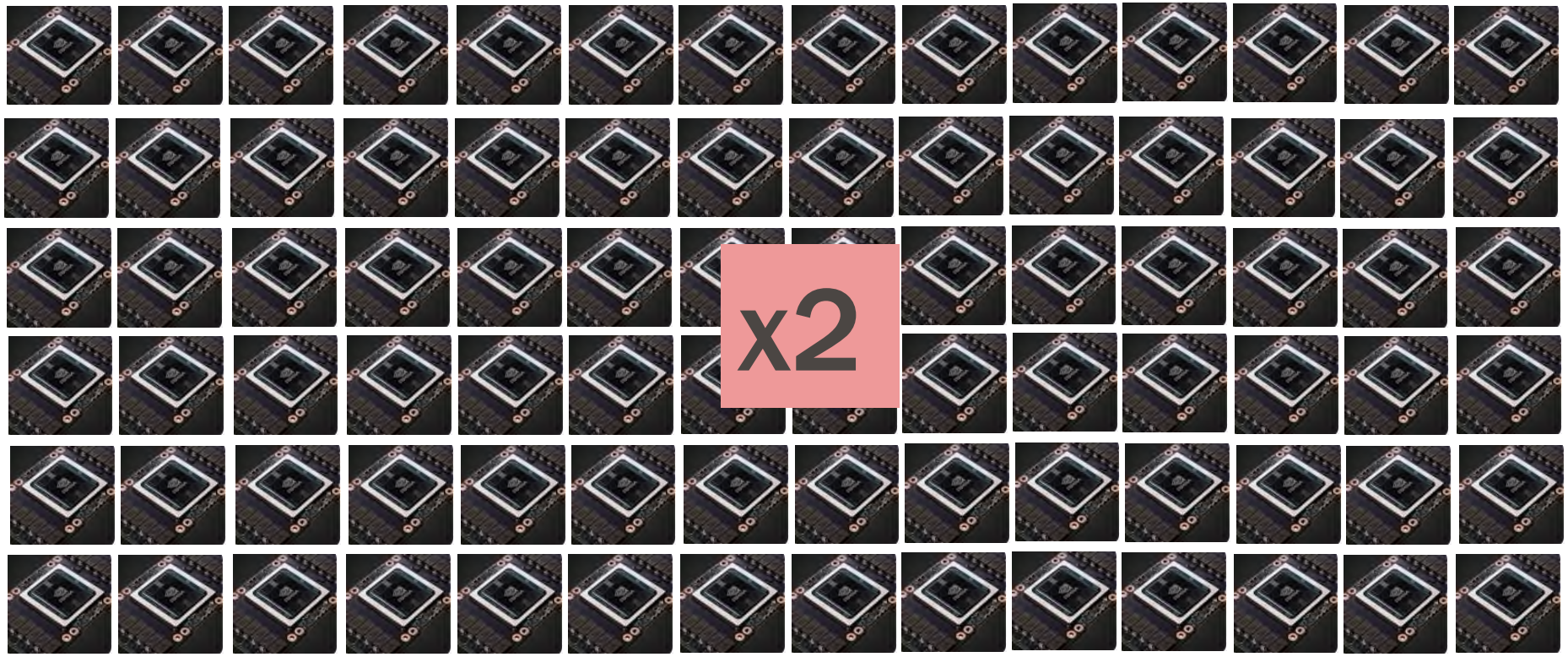
- In comparison to the-state-of-the-arts, ECOGreen
  - **71% reduction** of financial costs
  - **48% increase of use in** renewable energy (more sustainable!)

What's Next: Neural Architectures vs. GPUs

- Brain is ~160x better than our ICs (>1'000x more energy efficient)

Human brain (~20W), >10,000 TFLOPS

NVIDIA B100 (~30,000W), ~10,000 TFLOPS



Neural Architectures: 3D Liquid-Based Cooling and Powering

- Brain is ~160x better than our ICs: 3D + **Blood** (both cooling and energy supply)
- **PowerCool**: Use microfluidic fuel cells to generate power
  - Two electrolytes flowing in co-laminar regime, scalable for future 3D servers

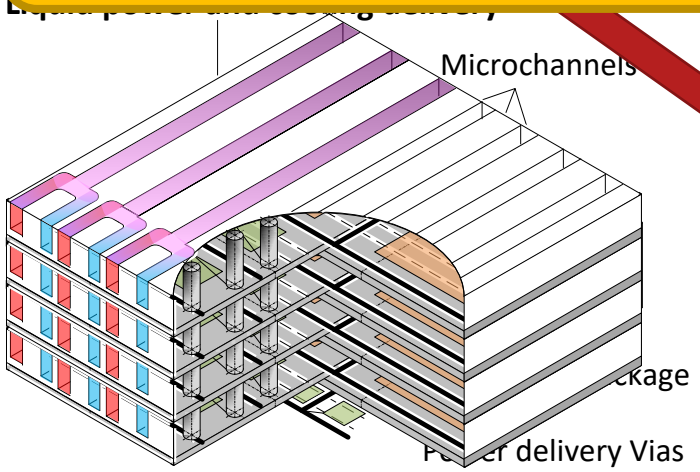
Human brain (**~20W**), >10,000 TFLOPS

NVIDIA H100 (**~30,000W**), ~10,000 TFLOPS

Power delivery

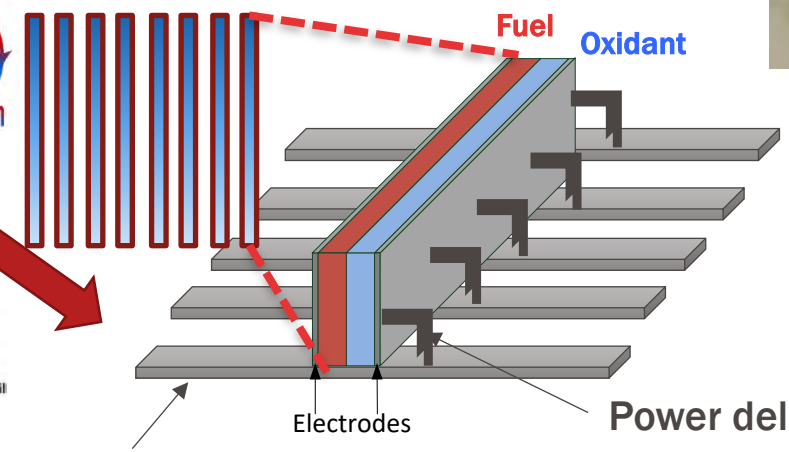
Efficient cooling

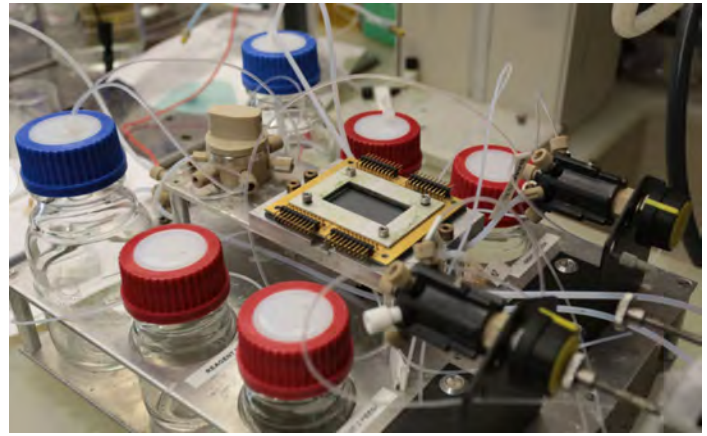
To generate electricity: the warmer the better!



Processing Architecture

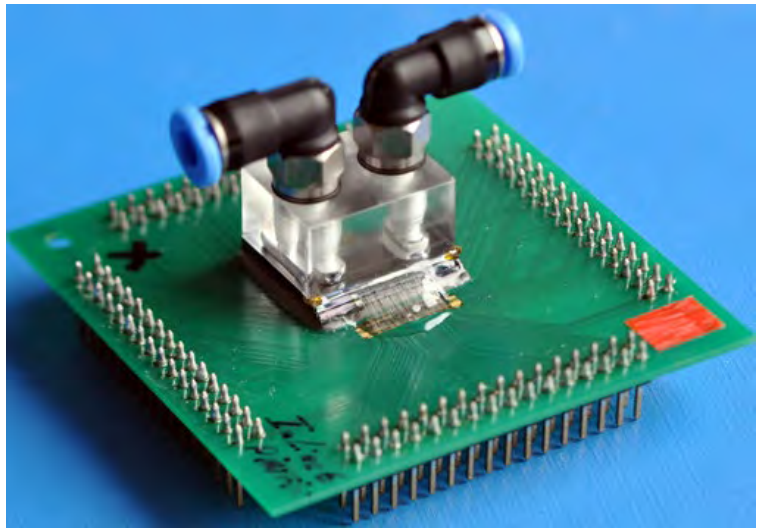
Integrated Cooling and Power Delivery

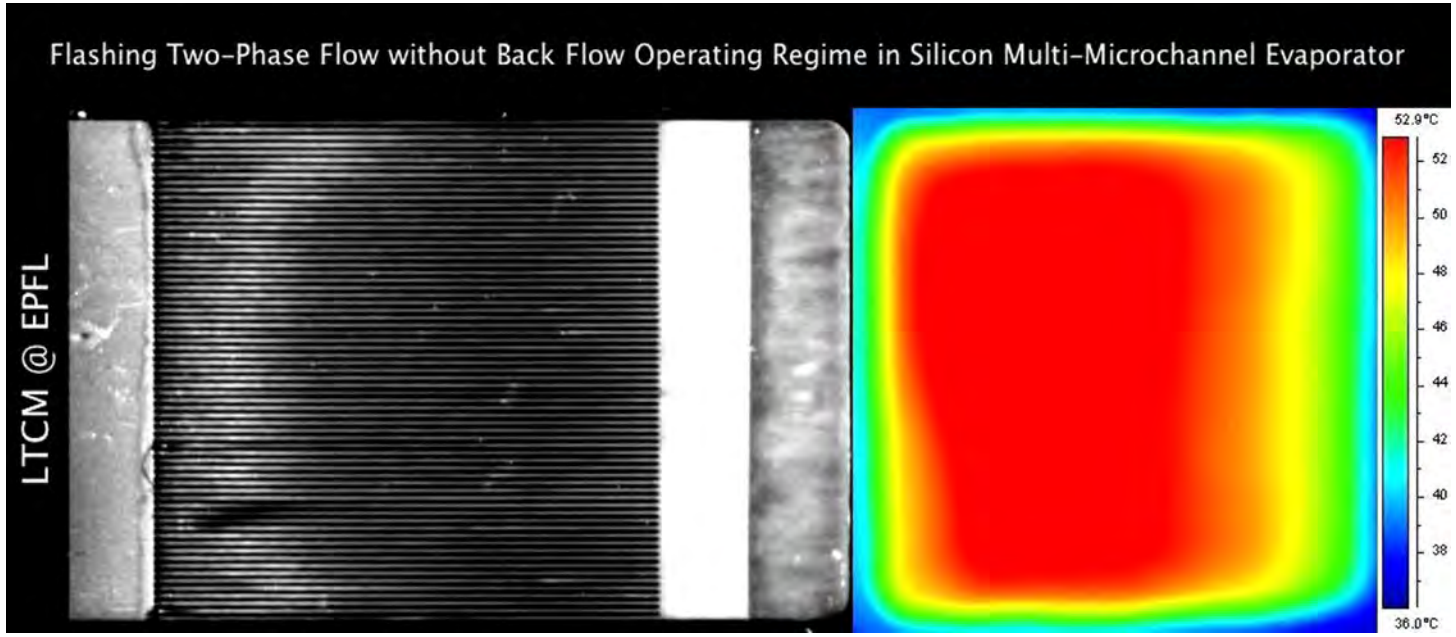




[Courtesy: IBM, "Electronic Blood", 2017]

Compusapien Chip: 5-Tier 3D AI Test Chip with Liquid Cooling Channels in Multiple Tiers (1000 W/cm^2 as NVIDIA B200)





Uniform temp. at 52-55º C, and 30% self-recovered energy possible (so heat finally could help!)

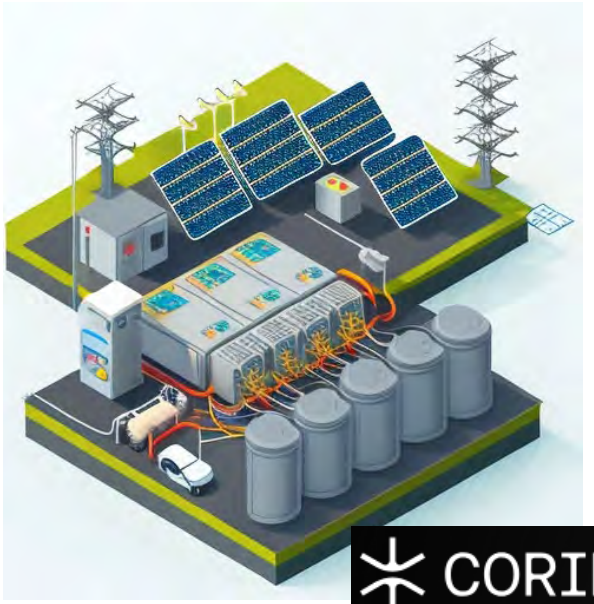
Energy-efficient 3D AI Computers are possible!

https://www.epfl.ch/labs/esl/research/thermal-modelling/fuel-cell-arrays/

Heating Bits: Renewable-Supplied DCs Integrating Heating and Cooling Supply of EPFL

HEATING BITS

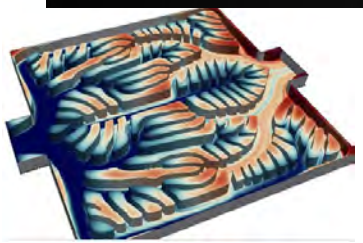
1. Increase DCs energy efficiency and operate them with the least CO_2 emissions
   - Power-aware computing
   - Optimize power supply: converters
   - Renewables and batteries integration
   - Reuse of waste heat in EPFL campus (heating and warm water)

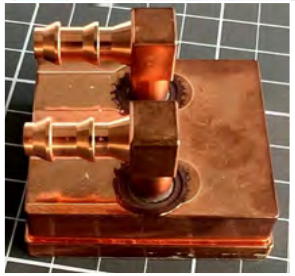


EcoCloud

CORINTIS

2. High-temp. liquid microcooling and electricity generation
   - Maximize servers efficiency with microfluidic cold plate
   - Transform heat back into electricity (Organic Ranking Cycle)





Funded by EPFL's Solutions for Sustainability (S4S) Initiative:
6 laboratories and EcoCloud Center, Microsoft is already convinced!

© EPFL

https://heatingbits.epfl.ch/

Heating Bits: Renewable-Supplied DCs Integrating Heating and Cooling Supply of EPFL

HEATING BITS

EcoCloud

CORINTIS

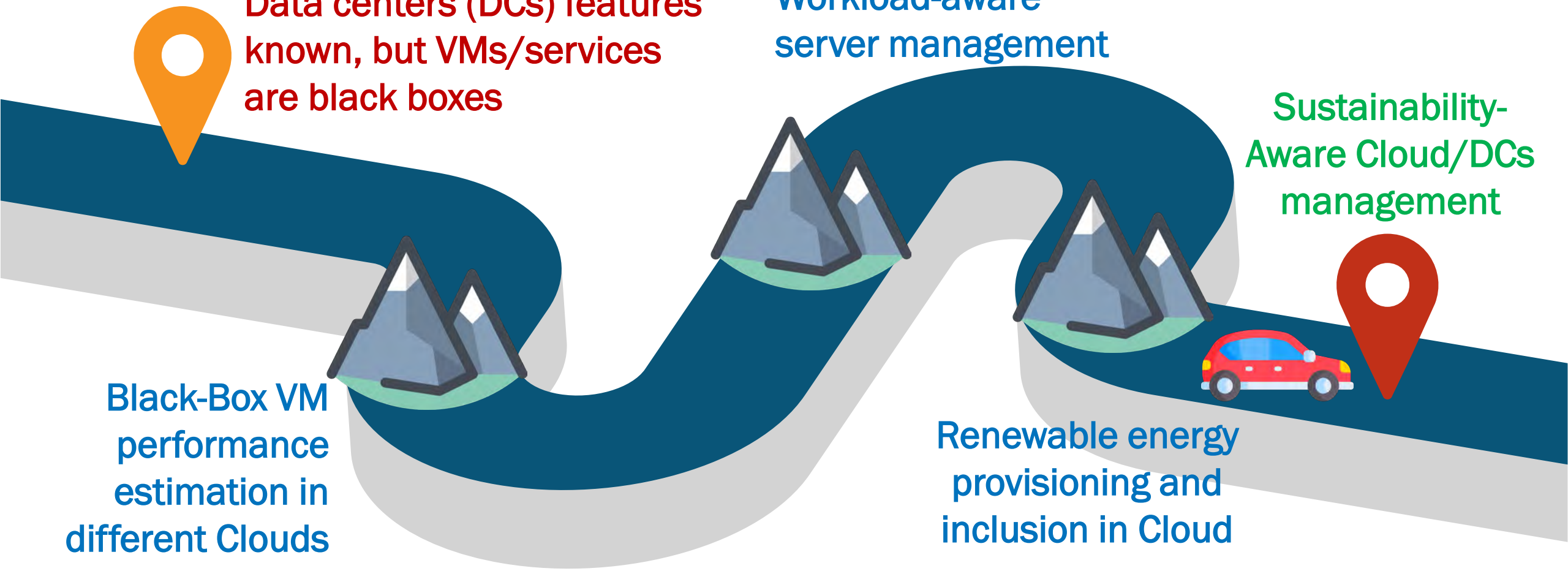
Photo by Dan DeLong for Microsoft

New family of Cobalt and Maia chips designed specifically to run Microsoft and customer workloads using this new cooling technology!

https://news.microsoft.com/source/features/innovation/microfluidics-liquid-cooling-ai-chips/

© EPFL

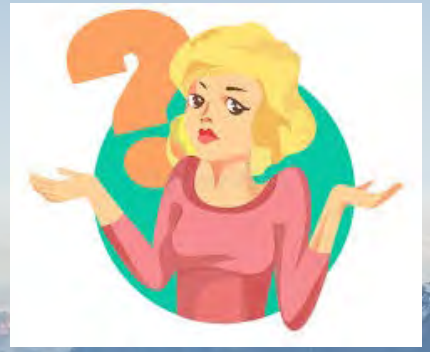
Challenges in our Path to a Sustainable Cloud



Conclusion

- IT/Cloud has enabled our progress for 50+ years
  - Multi-core servers and data centers are becoming more powerful
  - Big Data + IoT enabled us to conceive the new AI Era
- But current computing and cloud systems are **not sustainable for new AI Era!**
  - Very different and dynamic workloads from classical HPC
  - Severe performance interference among VMs and AI services collocated together
  - Computing systems and DCs are reaching heat and energy limits on supplies
- AI-based management of DCs **to the rescue for a sustainable cloud!**
  1. **CloudProphet:** Accurate and adaptive to new workloads (<7% error in accuracy)
  2. **GreenDVFS:** Higher energy efficiency per server (20% less energy, 35% less temp.)
  3. **ECOGreen:** Multi-DC management + renewables (48% increase of renewables)
- Next-gen. sustainable AI: **New brain-inspired cooling for servers and DCs**

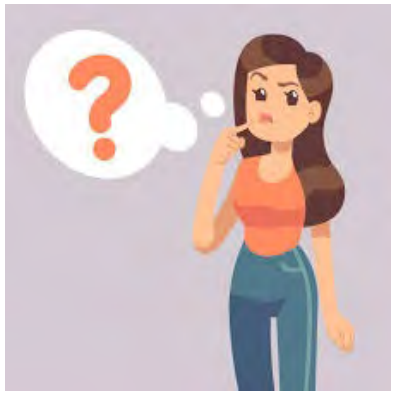
Thank you!
Questions?



David Atienza Alonso

EPFL - Embedded Systems Laboratory
david.atienza@epfl.ch

Questions?



Acknowledgements (alph. order): Dr. Denisa-A. Constantinescu, Prof. Ayse K. Coskun, Dr. Luis Costero, Dr. Darong Huang, Dr. Ali Pahlevan, Mr. Rubén Rodríguez Alvárez, Mr. Amirhossein Shahbazinia, Prof. Marina Zapater

ETH Board: UrbanTwin JI Action

Swiss National Science Foundation: SEAMS

European Commission

European Research Council

SERI:

© EPFL

Additional References and Bibliography

- D. Huang, L. Costero, et al., "CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds," IEEE TSUSC, January 2024
- D. Huang, L. Costero, et al., Is the *powersave* governor really saving power? IEEE/ACM Int. Symp. on Cluster, Cloud, and Internet Computing (CCGrid), 2024.
- D. Huang, A. Pahlevan, et al, “COCKTAIL: Multi-Core Co-Optimization Framework with Proactive Reliability Management”, IEEE T-CAD, February 2022.
- A. Iranfar, M. Zapater, et al., “Multi-Agent Reinforcement Learning for Hyperparameter Optimization of Convolutional Neural Networks”, IEEE T-CAD, December 2021.
- A. Pahlevan, M. Zapater, A. K. Coskun, and D. Atienza, “ECOGreen: Electricity Cost Optimization for Green Datacenters in Emerging Power Markets”, IEEE Transactions on Sustainable Computing (T-SUSC), April/June 2021.
- U. Gupta, et al. "Chasing carbon: The elusive environmental footprint of computing." IEEE HPCA. 2021.
- R. Bianchini, et al. "Toward ML-Centric Cloud Platforms." Communications of the ACM, 2020
- W. Simon, Y. Qureshi, et al., “BLADE: An in-Cache Computing Architecture for Edge Devices”, IEEE Transactions on Computers (TC), February 2020.
- K. Haghshenas, A. Pahlevan, et al., “MAGNETIC: Multi-Agent Machine Learning-Based Approach for Energy Efficient Dynamic Consolidation in Data Centers”, IEEE TSC, July 2019.
- L. Costero, A. Iranfar, et al., “MAMUT: Multi-Agent Reinforcement Learning for Efficient Real-Time Multi-User Video Transcoding”, Proc. of DATE, March 2019.
- A. Iranfar, M. Zapater, et al., “Machine Learning-Based Quality-Aware Power and Thermal Management of Multistream HEVC Encoding on Multicore Servers”, IEEE TPDS, October 2018.
- A. Pahlevan, X. Qu, et al., “Integrating Heuristic and Machine-Learning Methods for Efficient Virtual Machine Allocation in Data Centers”, IEEE T-CAD, August 2018.
- C. Lu, et al. Imbalance in the Cloud: an Analysis on Alibaba Cluster Trace, IEEE International Conference on Big Data (Big Data), 2017
- K. Kanoun, et al., “Big-Data Streaming Applications Scheduling Based on Staged Multi-Armed Bandits”, IEEE TC, Dec. 2016.
- A. Pahlevan, P. Garcia, et al., “Exploiting CPU-Load and Data Correlations in Multi-Objective VM Placement for Geo-Distributed Data Centers”, Proc. DATE, March 2016.
- A. Pahlevan, J. Picorel, et al., “Towards Near-Threshold Server Processors”, *Proc. of DATE*, March 2016.
- N. Rameshan, et al. "Stay-away, protecting sensitive applications from performance interference." Proceedings of the 15th International Middleware Conference. 2014.