TU Informatics

From Sustainable IT to the IT for the Sustainable World

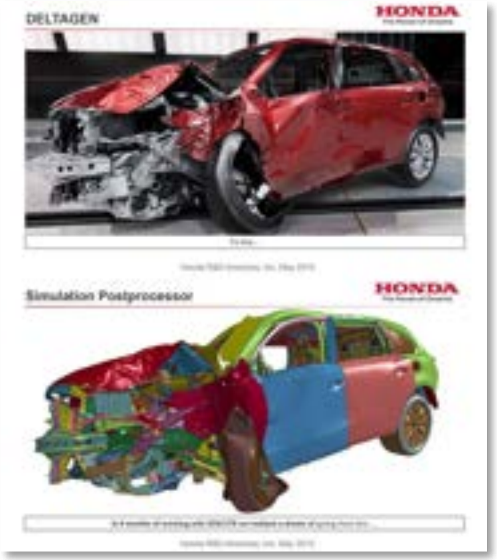
Ivona Brandić

TU Wien

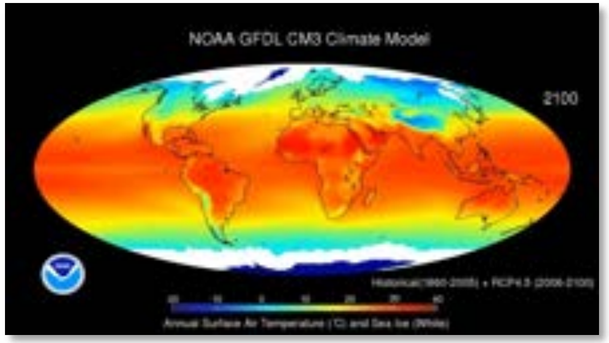
ivona.brandic@tuwien.ac.at

Computational Power

Simulation

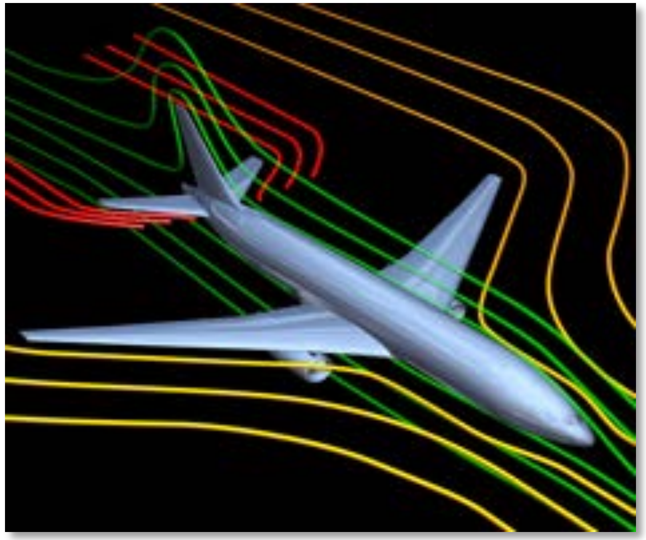


Mechanical Structure Simulation

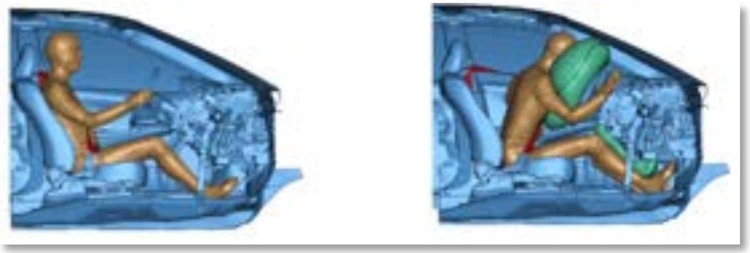


Climate Prediction

Optimization



Airflow Optimization



Finite Element Simulation
Hyper Parameter Optimization

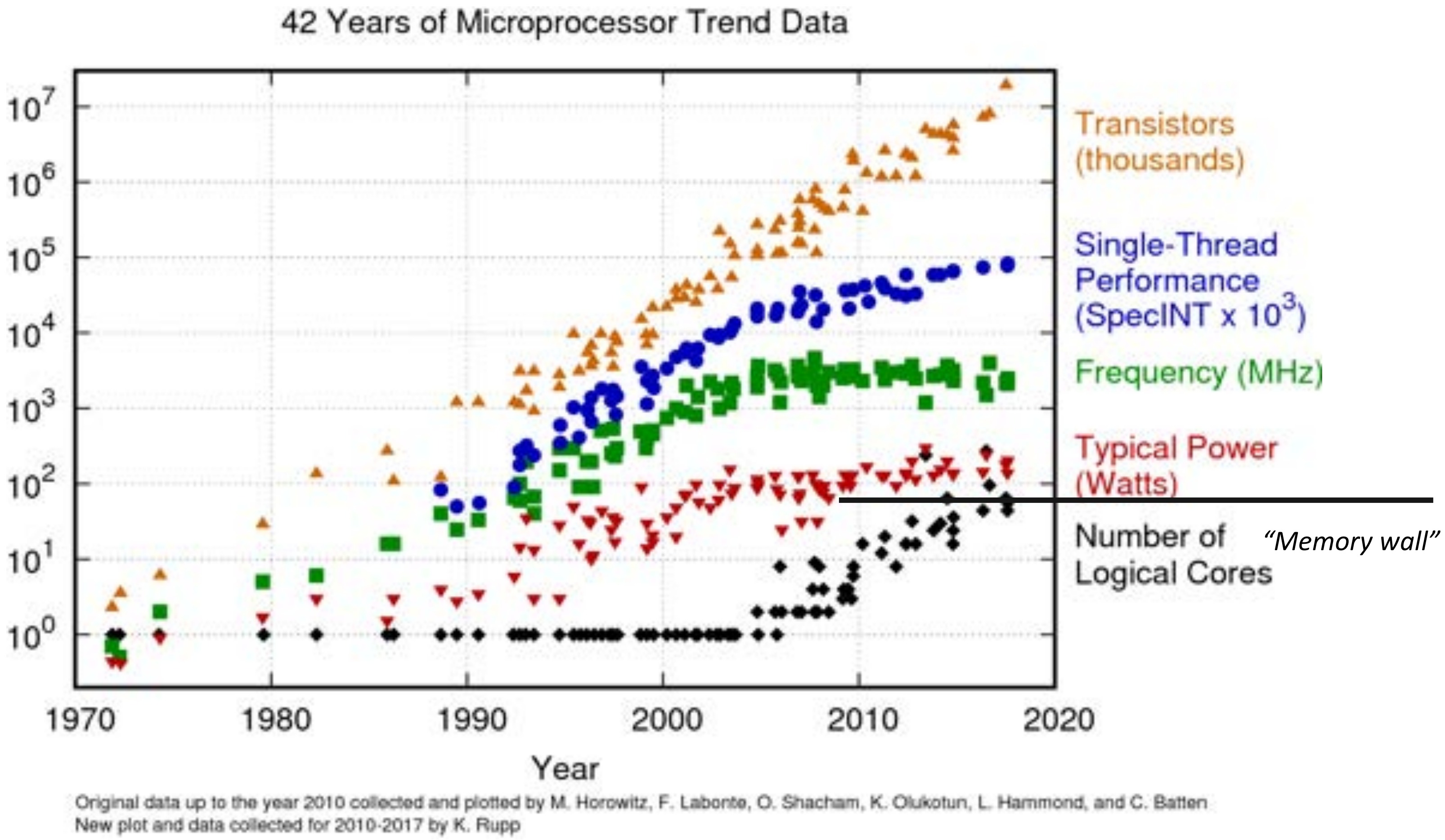
Today: Analytics, AI, LLMs,…

DNA Sequence Analysis
(e.g., Genomic sequencing of SARS-CoV-2)

Large Language Models

Recommendation Engines

Problem 1: Practical Limitations

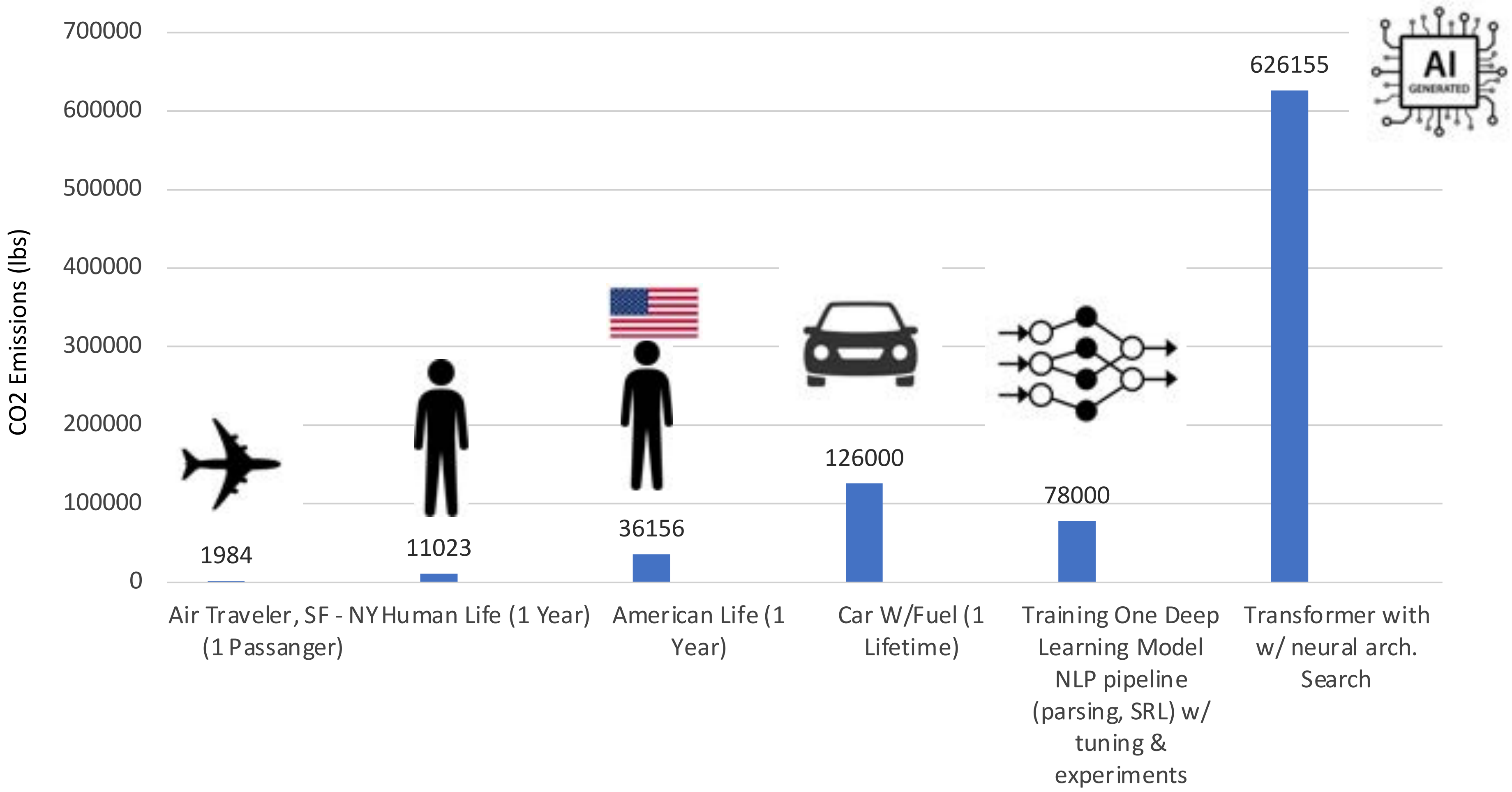


- #cores per chip doubles every 18 months instead of clock
- CPU-memory communication is becoming a bottleneck
- Too much heat is produced
- As transistors get smaller, power density increases because these do no scale with size anymore

→ practical limitations to processor frequency to around 4 GHz since 2006

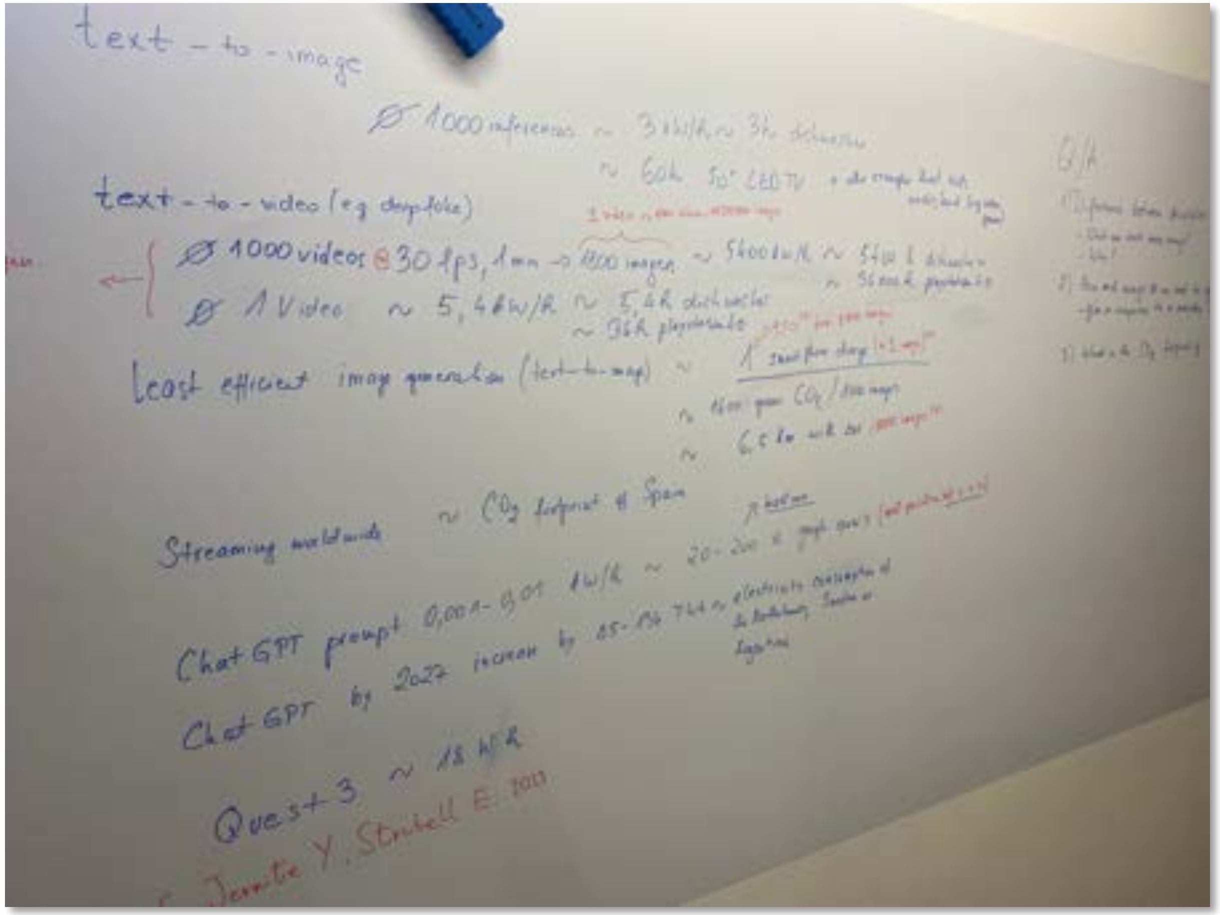
Source: @mileszim on Twitter

Problem 2: CO_2 Footprint of (generative) AI



Source: Emma Strubell, Ananya Ganesh, Andrew McCallum: Energy and Policy Considerations for Deep Learning in NLP, ACL (1) 2019: 3645-3650
Inspiration for Visualisation: Keren Bergman, Multicore World 2023 https://2019multicoreworld.files.wordpress.com/2023/02/bergman-keren-23.pdf

LLMs and generative AI



Asking ChatGPT Qs with 5-50 prompts[1] ≈

500ml of water

Search on LLM vs Google Search[1] ≈

2.9 Wh/ Search or up to 20x expensive

A text-to-image generation[2] (1000 inferences) ≈

3 kWh or 3hr of dishwasher

1 min video (e.g., midjourney) ≈ 5,4 kWh or 36hr playing Playstation 4.0

Slide courtesy: Shashi Ilager

TRUST! Explaining the increasing energy consumption of IT infrastructures (LLM Software, Chat GPT), call Science Communication for next Generation, FWF/ÖAW (2024)

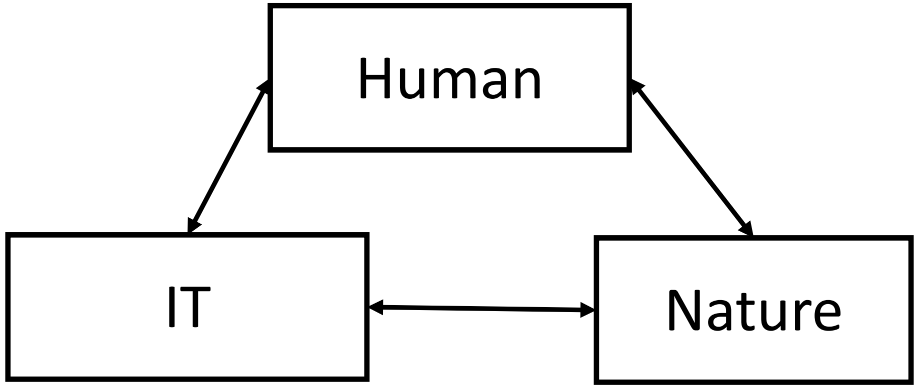
Image source: Generated from an LLM, text-to-image model, imagine.art

1. https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

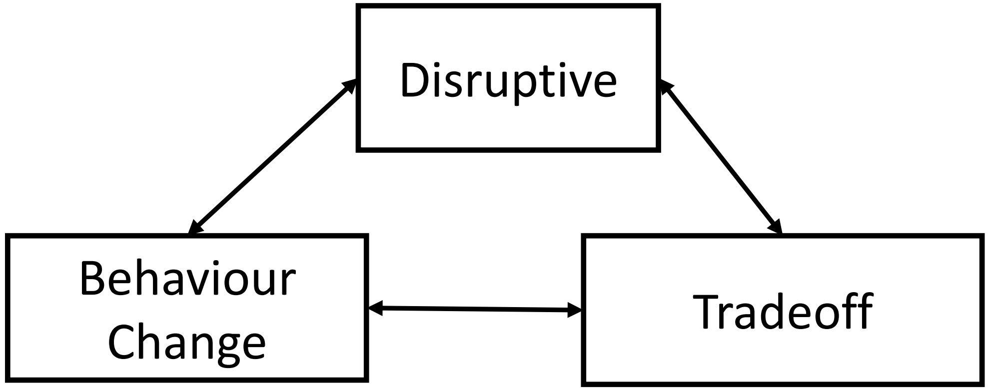
2. A. Luccioni, Y. Jernite, E. Strubell, Power Hungry Processing: Watts Driving the Cost of AI Deployment?, arxiv, 2023

Computational Sustainability

Actors:



Methods:



Computational Sustainability

Actors:

Human

IT

Nature

chist-era

Sustainable Watershed Management Through IoT-Driven Artificial Intelligence

FFG Forschung wirkt.

Virtual Shepherd

SFG

Agri Lorawan, Agricultural Testbench

Methods:

Disruptive

Behaviour Change

Tradeoff

HPQC

FFG Forschung wirkt.

High Performance Integrated Quantum Computing

TRUST! Explaining the increasing energy consumption of IT infrastructures

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

W|W|T|F

AI-supported Holographic Environmental Water

FWF

Triton: Transprecise Edge

Themis: Trustworthy and Sustaianble Code Offloading

Sources: https://www.youtube.com/watch?v=N4Ec_POLtes
https://www.youtube.com/shorts/O6QeeA6jbS0

Hi, AI (134 TWh)

© Ars Electronica Festival 2024, Julia Zukrigl
Song: https://www.jku.at/lit-robopsychology-lab/forschung/projekte/songs-about-ai/hi-ai/

Hi, AI (134 TWh)

Text+Music: Yasmo & Flip

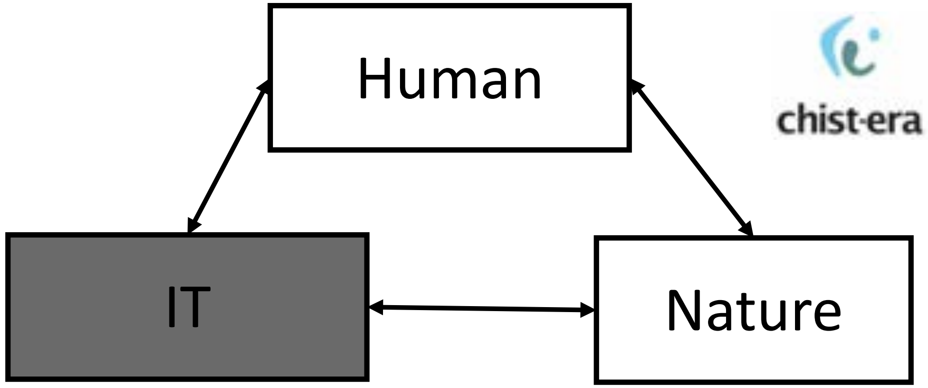
English translation: Christy Ellison Kemf and Andreas Kemf

STR1
Something here is changing, where should I begin?
Everything is moving, there's no option to stand still.
The Earth just keeps on turning, this I know, but to be frank,
It's also getting hotter, and we have ourselves to thank.
Oh my God, we're living in a science fiction film
Living in the future, should I take the red or blue pill?
Is it really all so bad, or do I just lack the skill
to understand what the future and the world from me would wi
Item 1 on the agenda would be well-being for all,
Point 2, Don't just watch the climate change as it takes its toll.
Point 3, not to incorrectly understand AI –
Maybe together find a way to hit all three in the bullseye.

HOOK
Hi, AI
You are pretty bright.
With the data that we feed you,
nothing's gonna break your stride.
But don't worry, we don't need your help for every single starter.
Does a conversation with you cost us half a liter water.

Computational Sustainability

Actors:

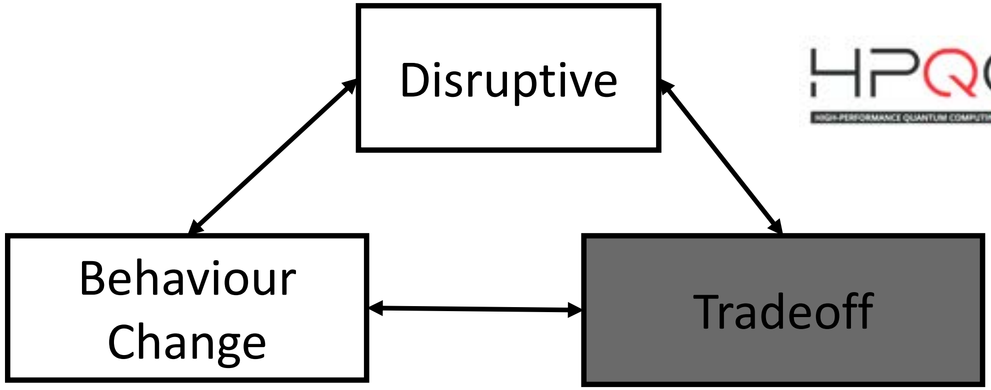


Sustainable Watershed Management Through IoT-Driven Artificial Intelligence

Virtual Shepherd

Agri Lorawan, Agricultural Testbench

Methods:



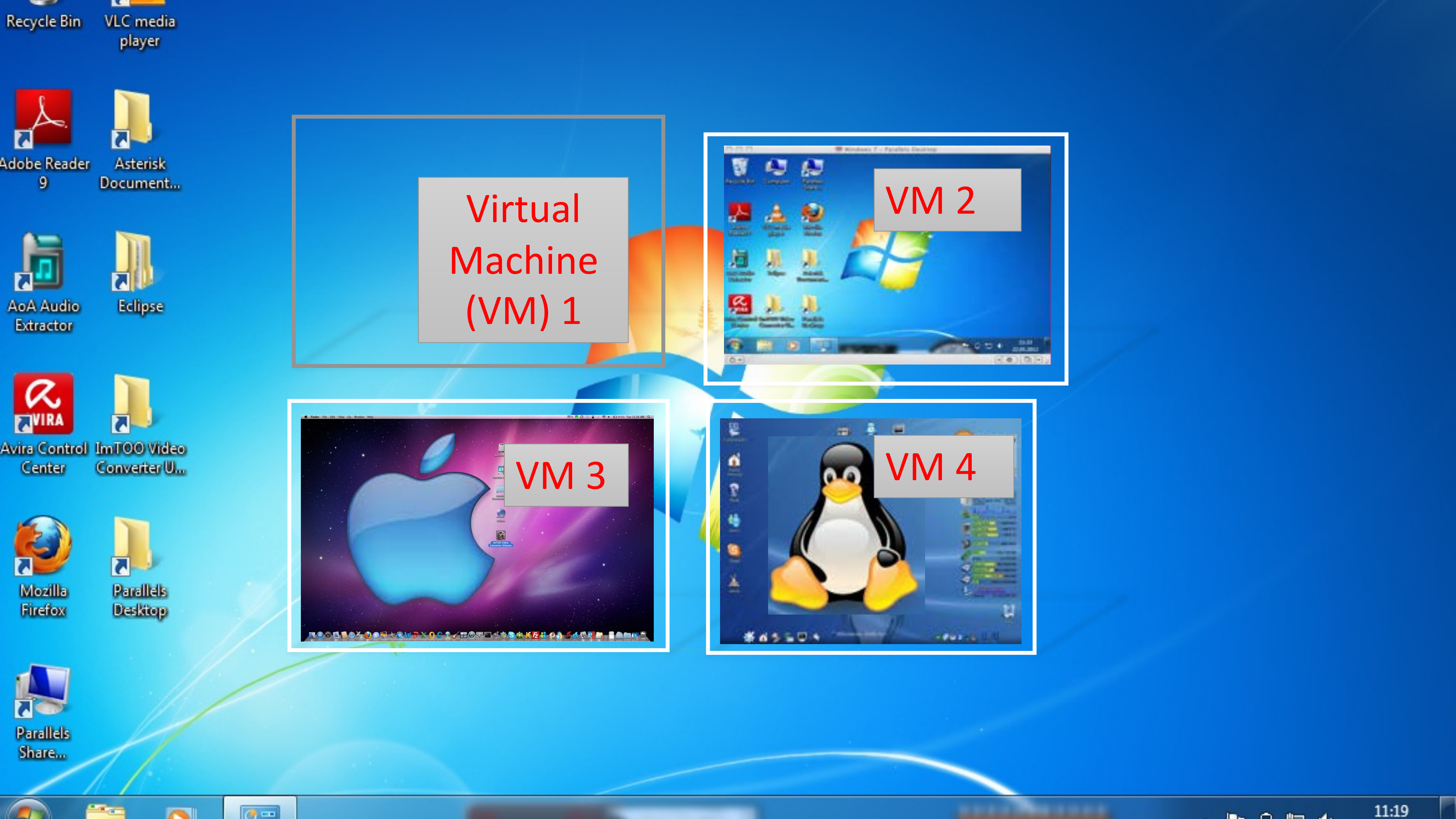
High Performance Integrated Quantum Computing

TRUST! Explaining the increasing energy consumption of IT infrastructures

AI-supported Holographic Environmental Water Monitoring

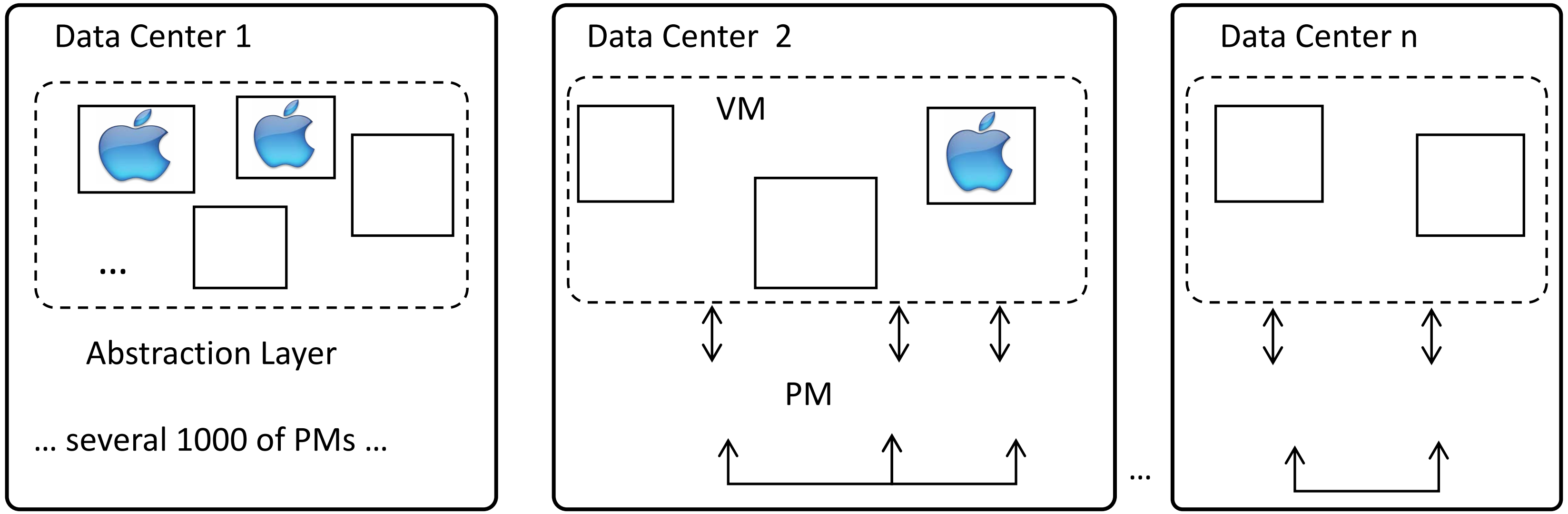
Triton: Transprecise Edge

Themis: Trustworthy and Sustaianble Code Offloading

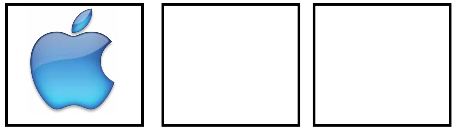


Recycle Bin
VLC media player
Adobe Reader 9
Asterisk Document...
AoA Audio Extractor
Eclipse
Avira Control Center
ImTOO Video Converter U...
Mozilla Firefox
Parallels Desktop
Parallels Share...
Virtual Machine (VM) 1
VM 2
VM 3
VM 4
11:19

Clouds

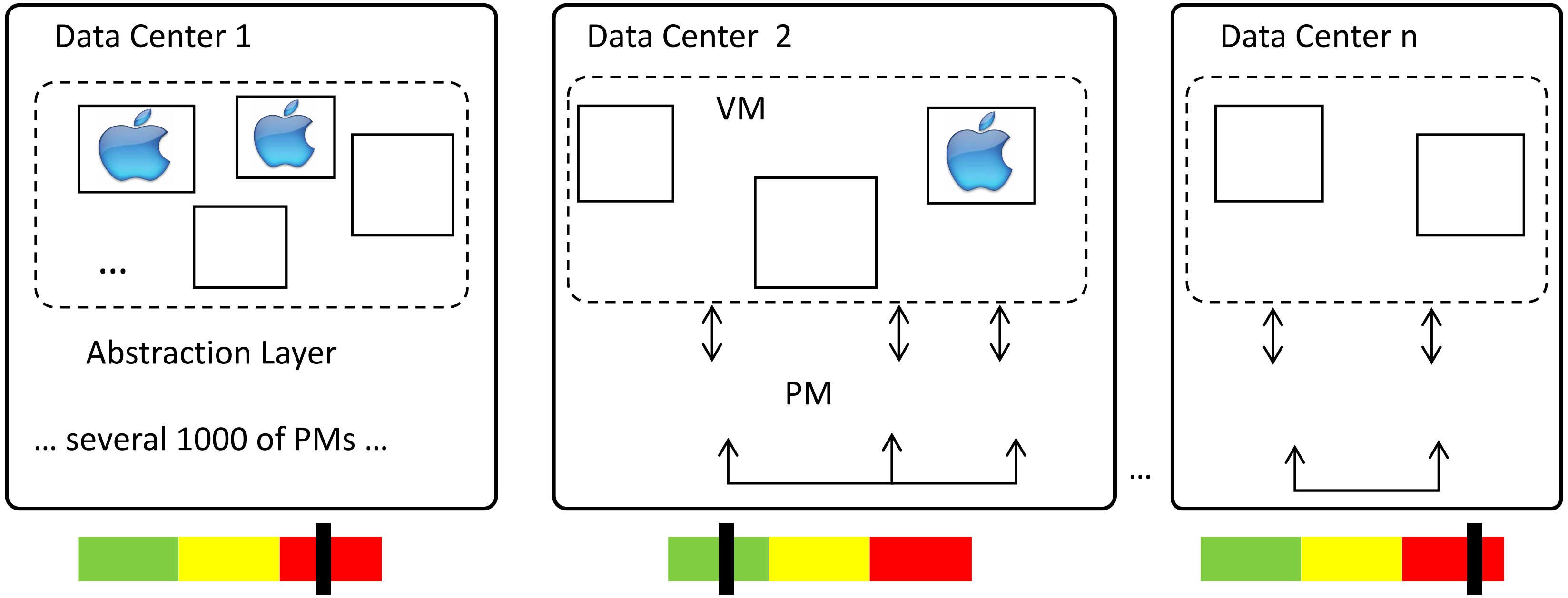


Physical Machine (PM)

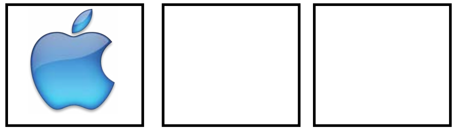


Virtual Machine (VM): Abstraction of a physical machine, *“simulation of a computer”*

Clouds



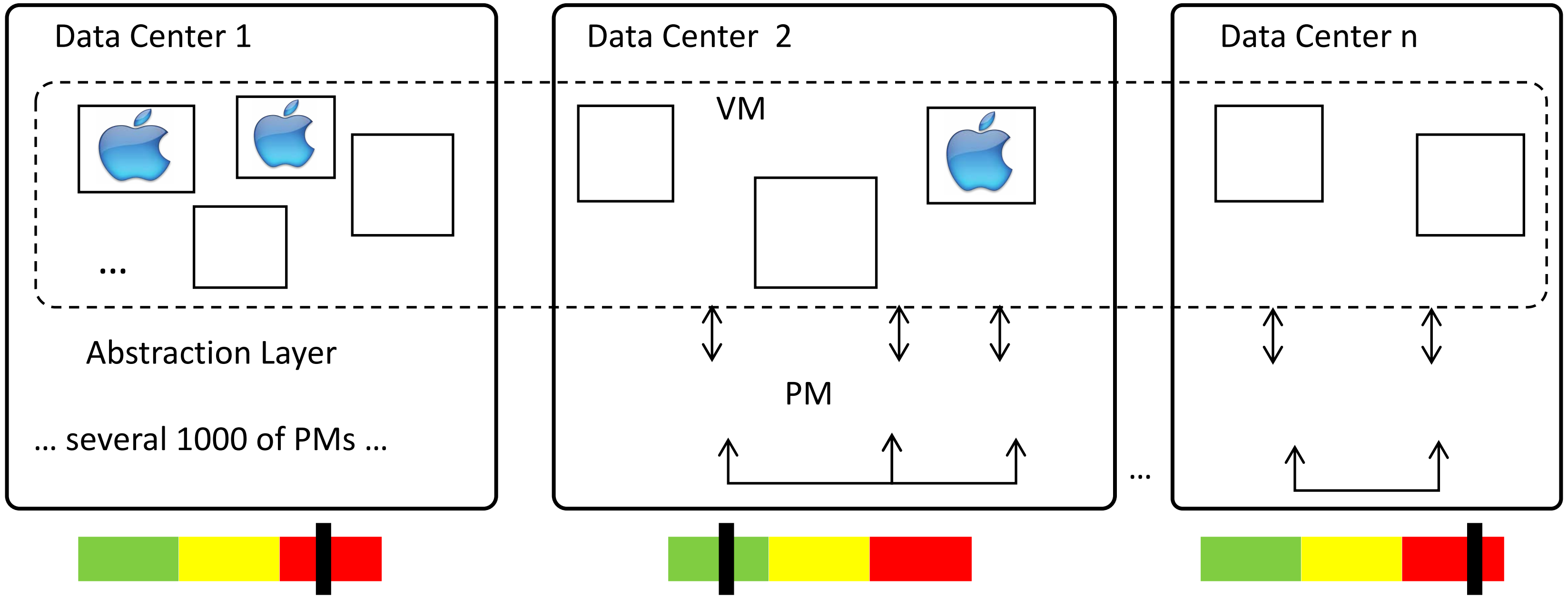
Physical Machine (PM)



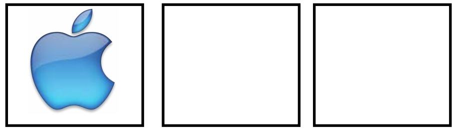
Virtual Machine (VM): Abstraction of a physical machine, "*simulation of a computer*"

Clouds

Y904-N31 START-Programme 2015
"Runtime Control in Multi Clouds"



Physical Machine (PM)

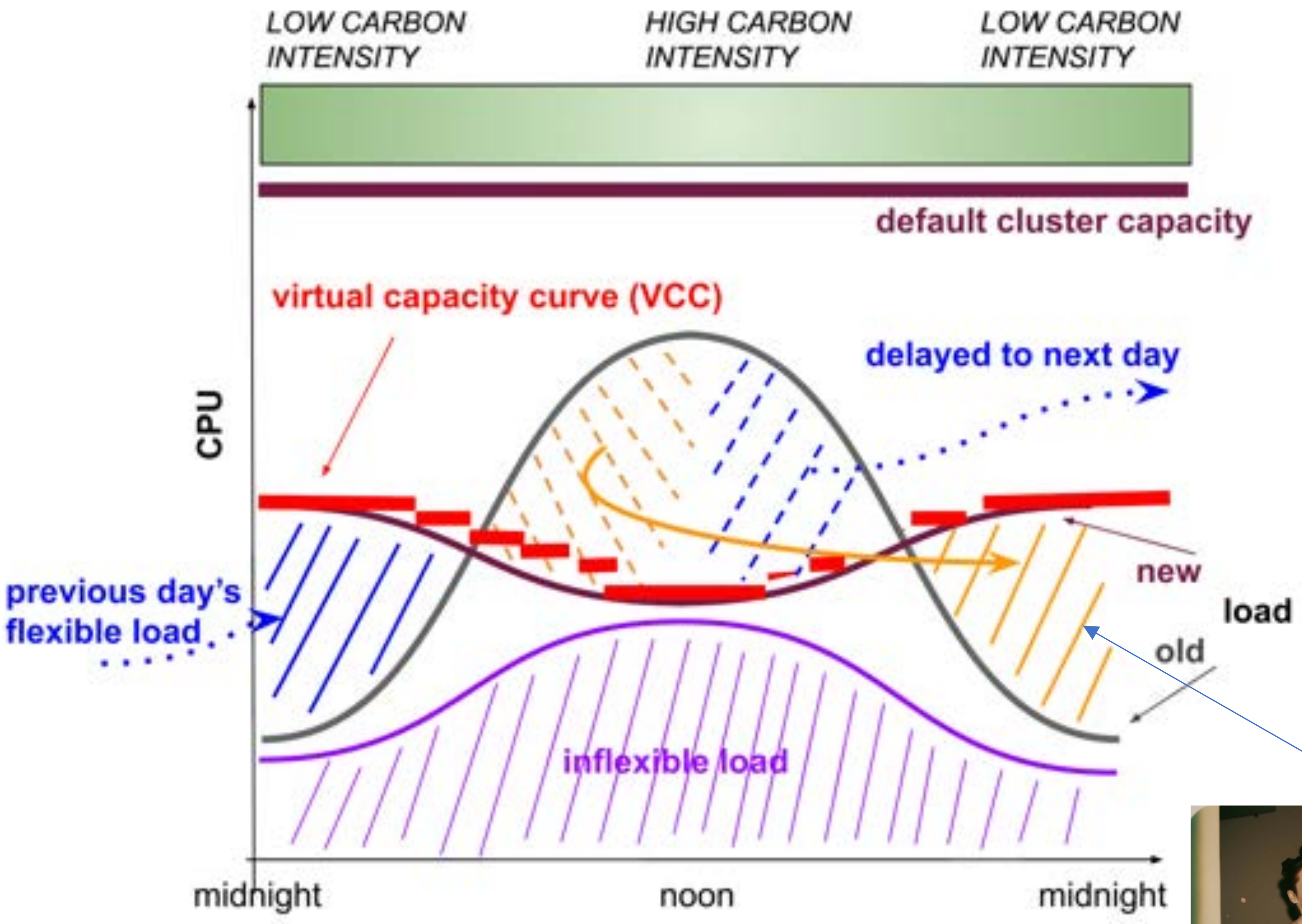


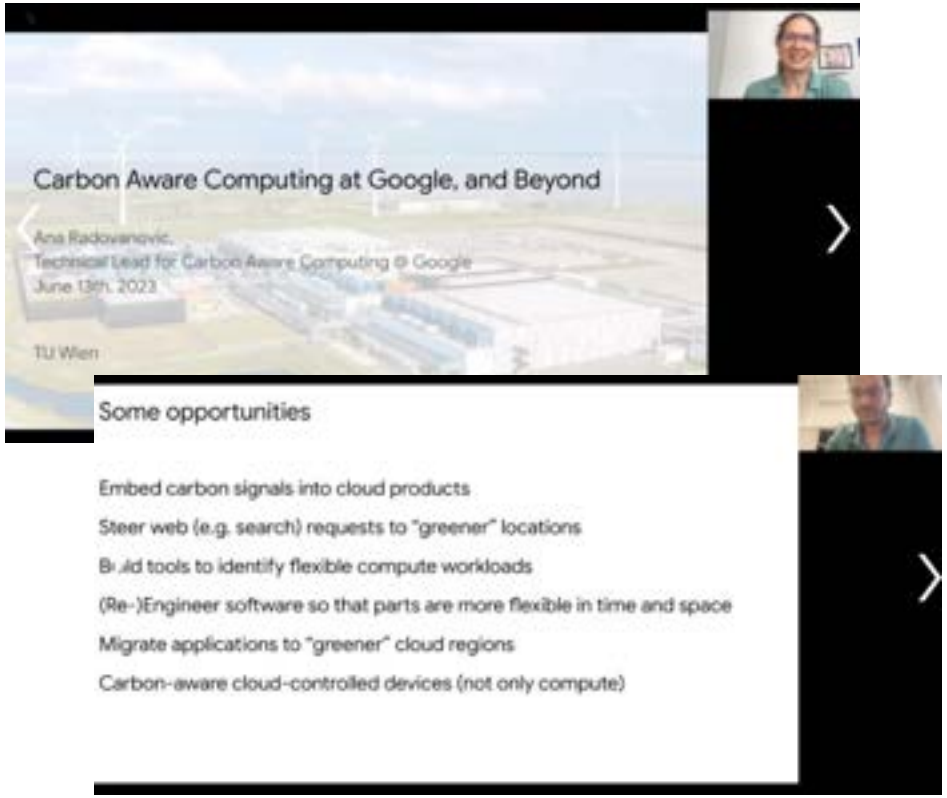
Virtual Machine (VM): Abstraction of a physical machine, "*simulation of a computer*"

Cloud: economic and ecological data center solutions

Source: Damien Borgetto, Michael Maurer, Georges Da Costa, Jean-Marc Pierson, Ivona Brandic: Energy-efficient and SLA-aware management of IaaS clouds. e-Energy 2012: 25

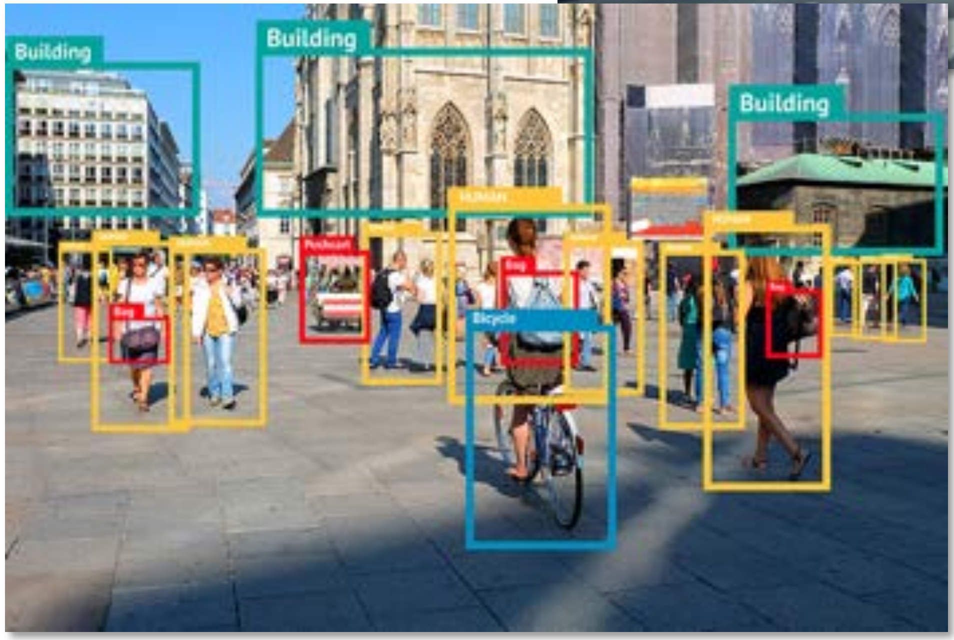
Workload Shift in Space and Time





Ana Radovanovic (Google) & Shashi Ilager (TU)
*Lecture at TU Wien: **"Data Intensive Computing"***

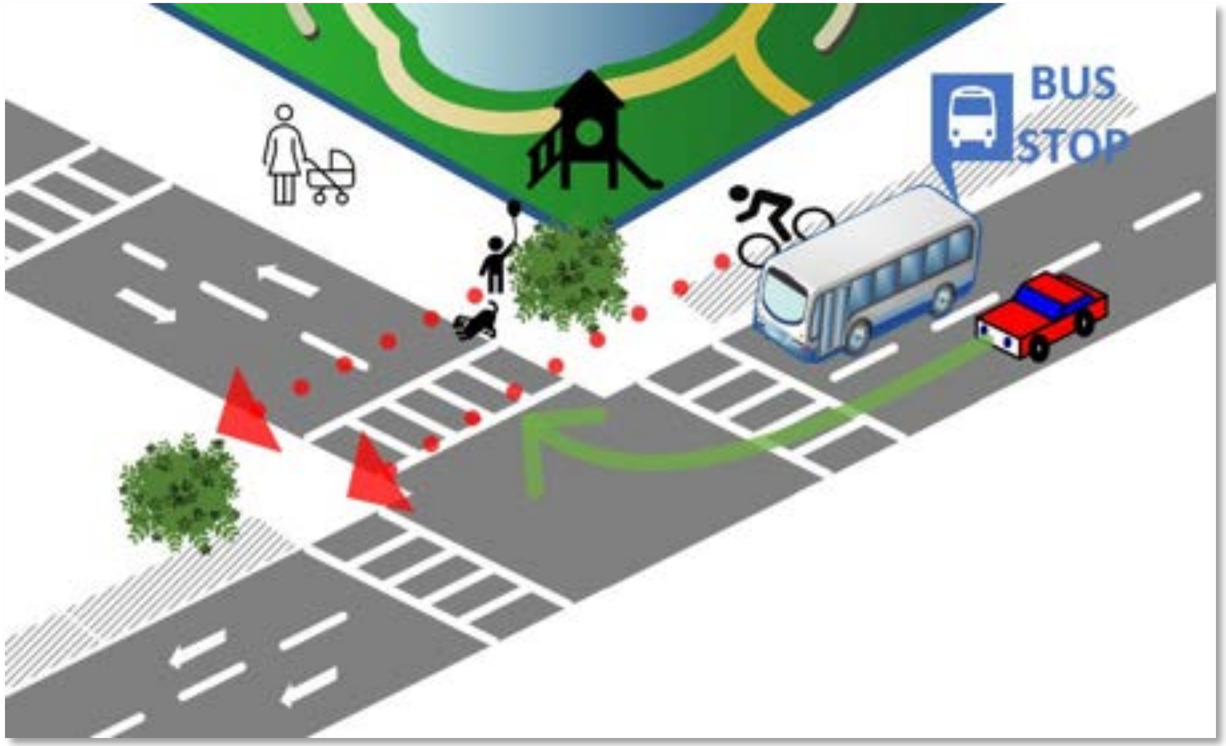
Source: Radovanović, Ana, et al. "Carbon-aware computing for datacenters." IEEE Transactions on Power Systems 38.2 (2022): 1270-1280.



Symbolic Data Representation

Edge Computing in Action: Smart City

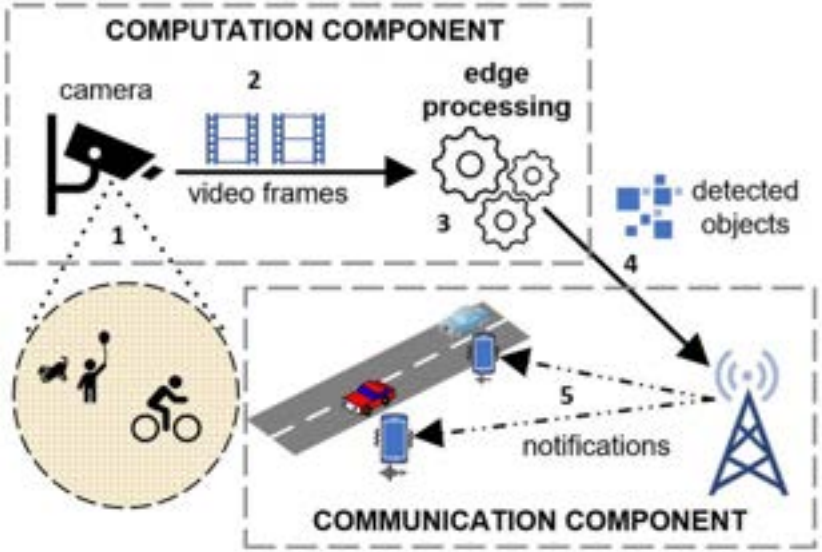
- Traffic accidents
- causing injuries and deaths
- Distractions, poor visibility (e.g., bad road and weather conditions), …
- Drivers’ brake reaction time
- 1500ms on average



Deaths among pedestrians and cyclists:
29% of all EU road deaths
ETSC (European Transport Safety Council) PIN Flash Report 38

Slide: courtesy Ivan Lujic (Ericsson Nikola Tesla d.d.)

Smart City



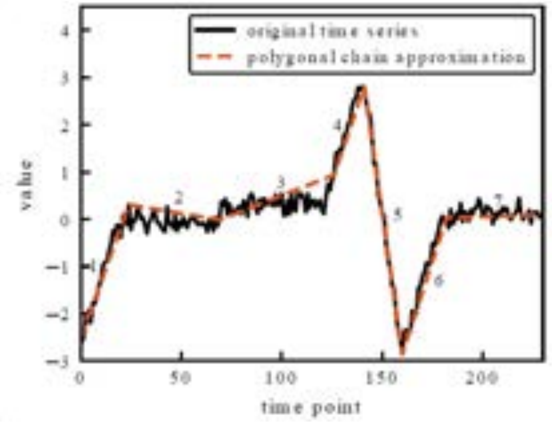
http://intrasafed.ec.tuwien.ac.at

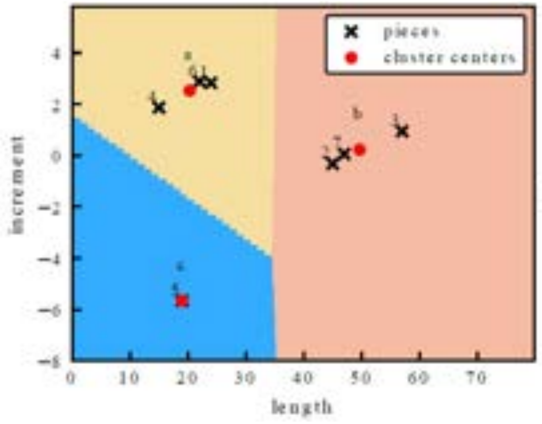
Slide: courtesy Ivan Lujic (Ericsson Nikola Tesla d.d.)

Source: Lujic, De Maio, Pollhammer, Bodrozic, Lasic, and Brandic, "Increasing Traffic Safety with Real-Time Edge Analytics and 5G," EdgeSys, pp. 19-24, 2021.

Adaptive and Online Symbolic Representation

P 36870- N 2023
“Transprecise Edge”





abbacab

D. Hofstätter

S. Ilager

I. Lujic

Ericsson Nikola Tesla d.d.

Daniel Hofstätter, Shashikant Ilager, Ivan Lujic, Ivona Brandic. ***SymED: Adaptive and Online Symbolic Representation of Data on the Edge****. 29th International European Conference on Parallel and Distributed Computing, 28 August - 1 September 2023 Limassol, Cyprus.*

EURO-PAR

University of Cyprus

29th International European Conference on Parallel and Distributed Computing

Chairs’ Appreciation Prize to a Junior Researcher

“SymED: Adaptive and Online Symbolic Representation of Data on the Edge”

Daniel Hofstätter, Shashikant Ilager, Ivan Lujic and Ivona Brandic

Prof. George A. Papadopoulos, Conference Chair
Prof. Marios Dikaiakos, Programme Chair
Prof. Rizos Sakellariou, Programme Chair

Springer

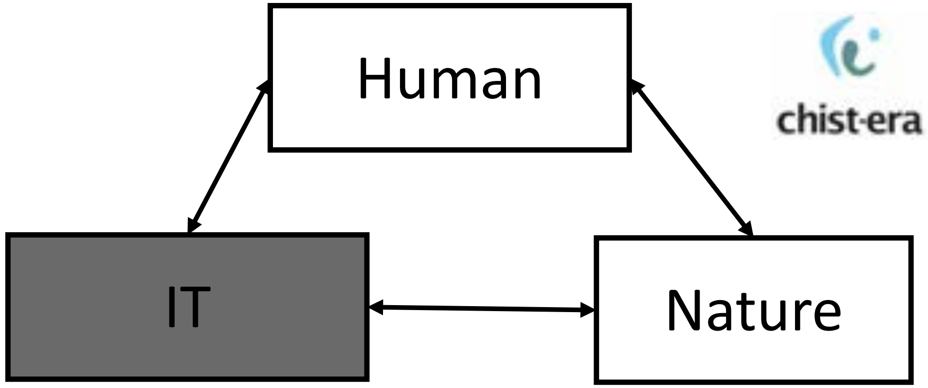
28 August – 1 September 2023 | Limassol, Cyprus

“Introduction to Computational Sustainability”

- Elective Course at TU Wien, current name “AI/Ml in the era of climate change”
- Part 1: Sustainable AI: Impact on sustainability by AI models
  - Hardware advancements, data explosion, and its energy impacts
  - Energy Challenge of AI models: Cost of Training and Inference
  - Large Language Models and their energy consumption
  - A path forward: Methods to address energy consumption of AI models
- Part 2: AI for sustainability: Using AI to combat the climate change issues
  - AI for climate change modelling use cases
  - Water supply, environmental monitoring, etc. → methods!

Computational Sustainability

Actors:

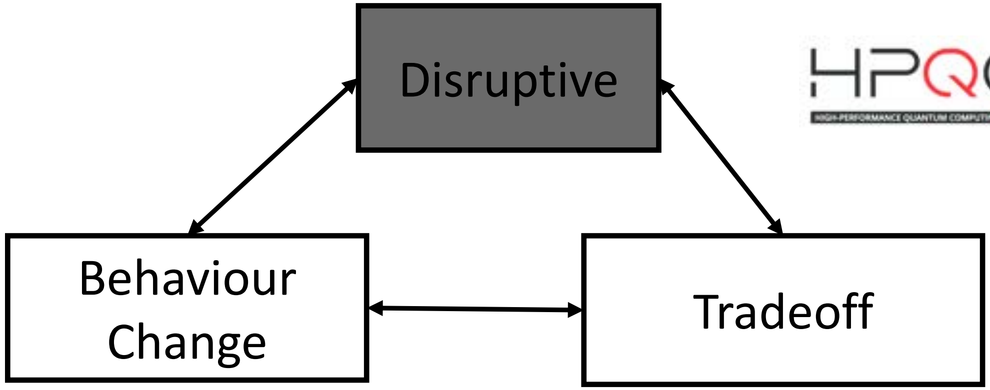


chist-era *Sustainable Watershed Management Through IoT-Driven Artificial Intelligence*

FFG Forschung wirkt. *Virtual Shepherd*

SFG *Agri Lorawan, Agricultural Testbench*

Methods:



HPQC FFG Forschung wirkt.

High Performance Integrated Quantum Computing

TRUST! Explaining the increasing energy consumption of IT infrastructures

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

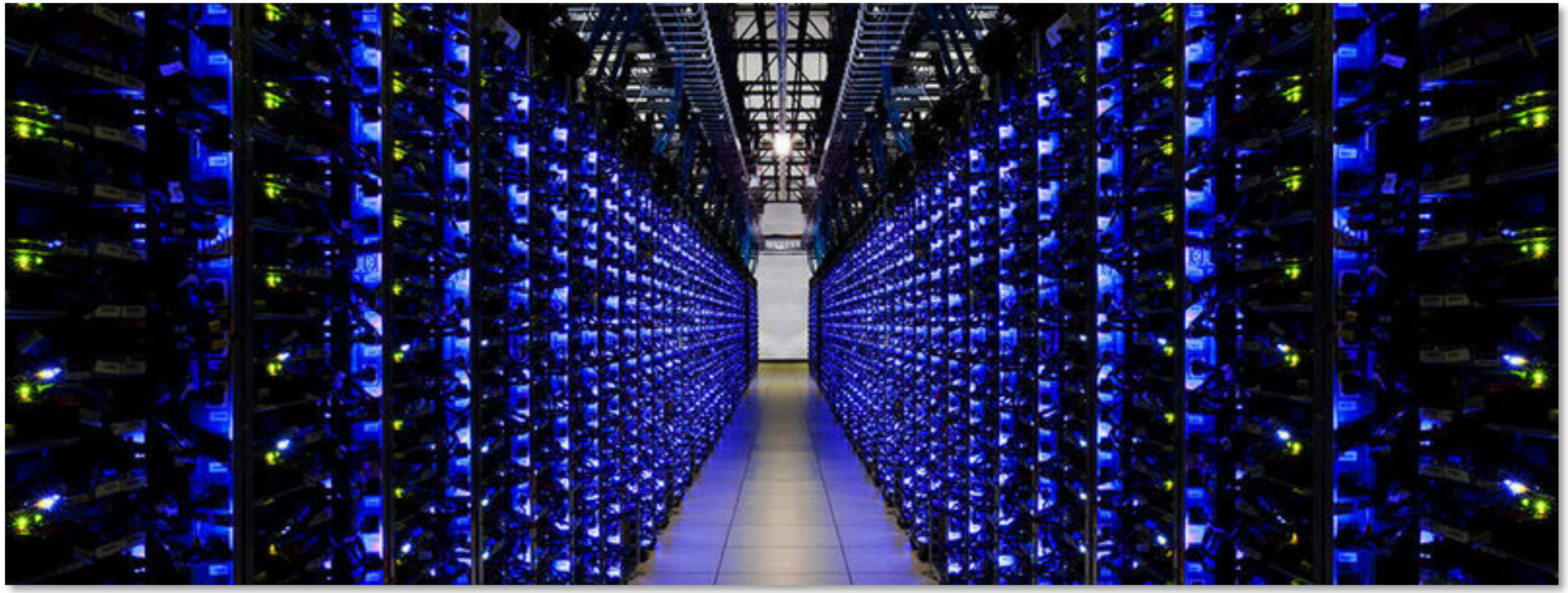
W|W|T|F

AI-supported Holographic Environmental Water Monitoring

FWF

Triton: Transprecise Edge

Themis: Trustworthy and Sustaianble Code Offloading



Gordon Moore:
Moore’s Law
(1929 - 2023)

Data volumes are growing faster than the processing power

Alternatives:

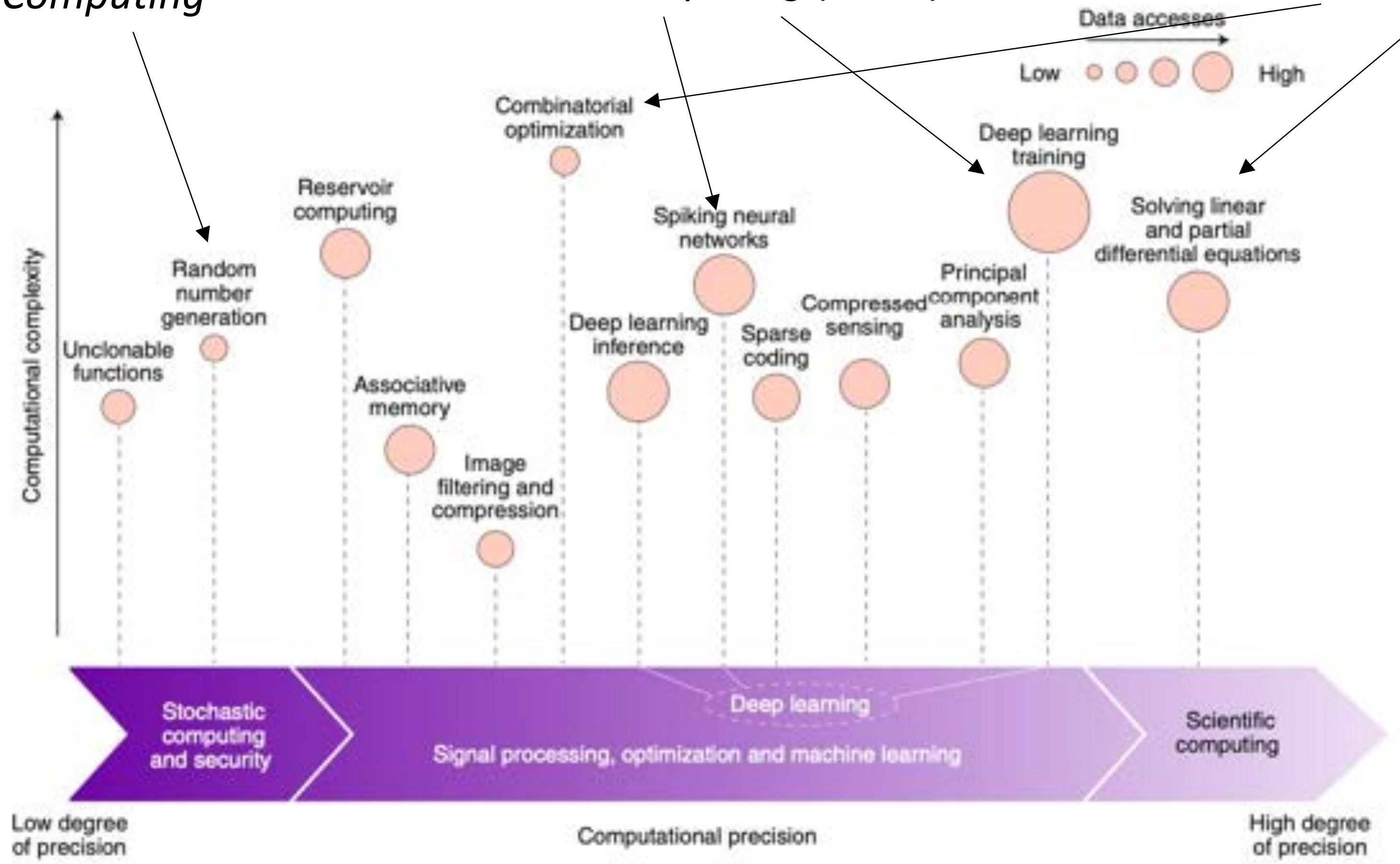
- Neuromorphic Computing
- Quantum Computing

Future: Hyper-Heterogeneous Architectures

Approximate Computing

Neuromorphic Computing (NPUs)

Quantum Computing (QPUs)

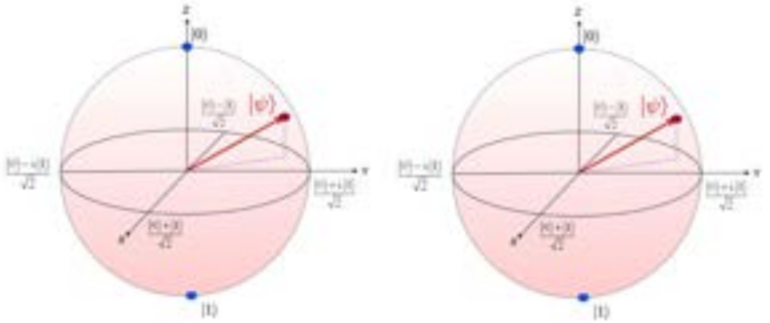


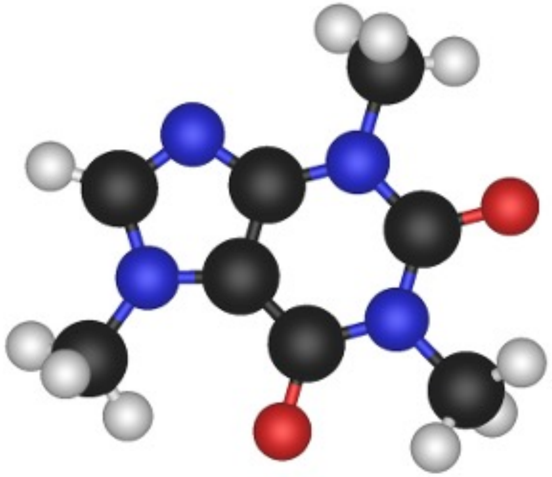
Hyper-heterogeneity, heavy geographical distribution
*- **More time is spent on communication!***

*Source: Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. et al. Memory devices and applications for in-memory computing. Nat. Nanotechnol. **15,** 529–544 (2020). https://doi.org/10.1038/s41565-020-0655-z*

Beyond 0 and 1

Quantum





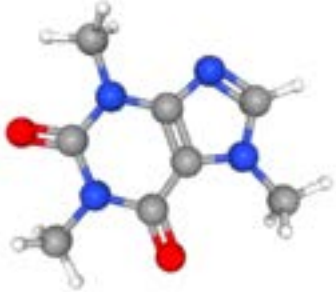
Von Neumann

Bottom up approach

- *Variational Quantum Linear Solver (VQLS)*
- *Quantum Eigenvalues –>* **Native 3d modeling of scientific applications**

Problem: Currently quantum systems can be used by quantum researchers only!

A cup of coffee?


"Every time you add a qubit, you double your possible outcomes, With 20 qubits there are a million possible outcomes. With 100 qubits, you have more possibilities than there are bits in all the hard drives in the world. With 300 qubits—that's more possibilities than there are particles in the universe."

- Representing the energy configuration of a single caffeine molecule at a single instant requires approximately 10^{48} bits in a classical computer
- Can be done using 160 logical qubits on a quantum machine

1 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 vs. 160

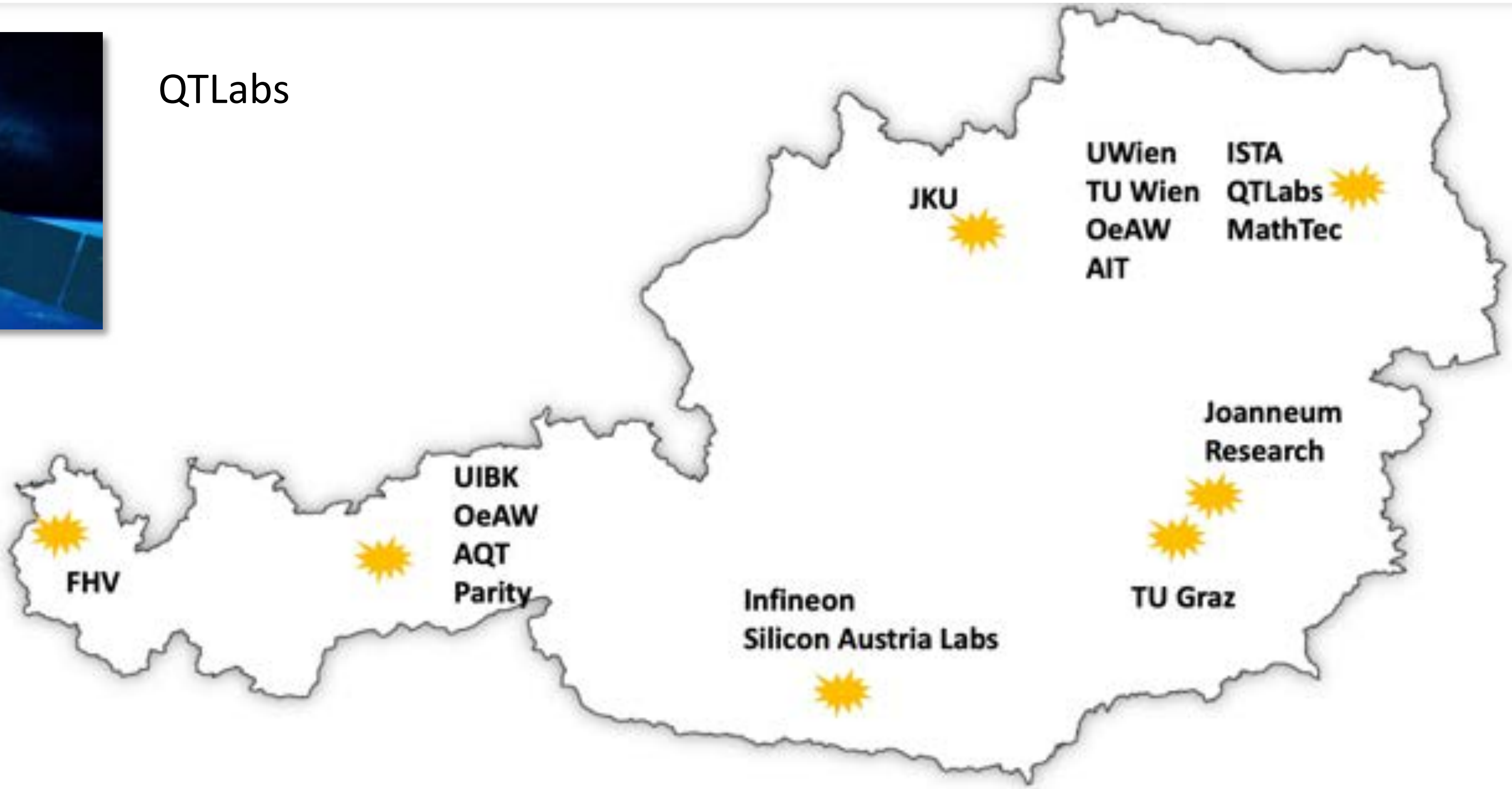
Courtesy Nico Einsiedler (IBM), TU Wien Quantum Fall Fest 2022

https://quantum.duke.edu/2020/10/16/more-possibilities-than-there-are-particles-in-the-universe/

Quantum Landscape in Austria

QTLabs

AQT

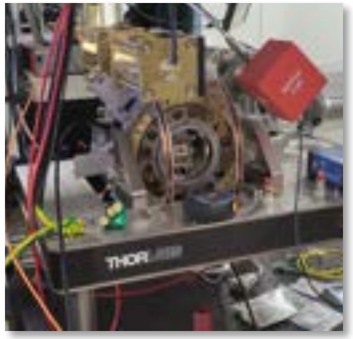


Slide: courtesy Thomas Monz, Uni Innsbruck & AQT

Known Quantum Speedup

- **Grover's algorithm** (unstructured search): $O(\sqrt{n})$ vs $O(n)$, developed 1996
- **Shor's algorithm** (finding the prime factors in integer): Polynomial vs Exponential, developed 1994
- Quantum ML
  - Bayesian Inference: quadratic
  - SVM: exponential
  - Reinforcement Learning: quadratic
- In reality
  - Lack of standardization
  - Data transformation / quantum state preparation
  - Decoherence
  - Noise

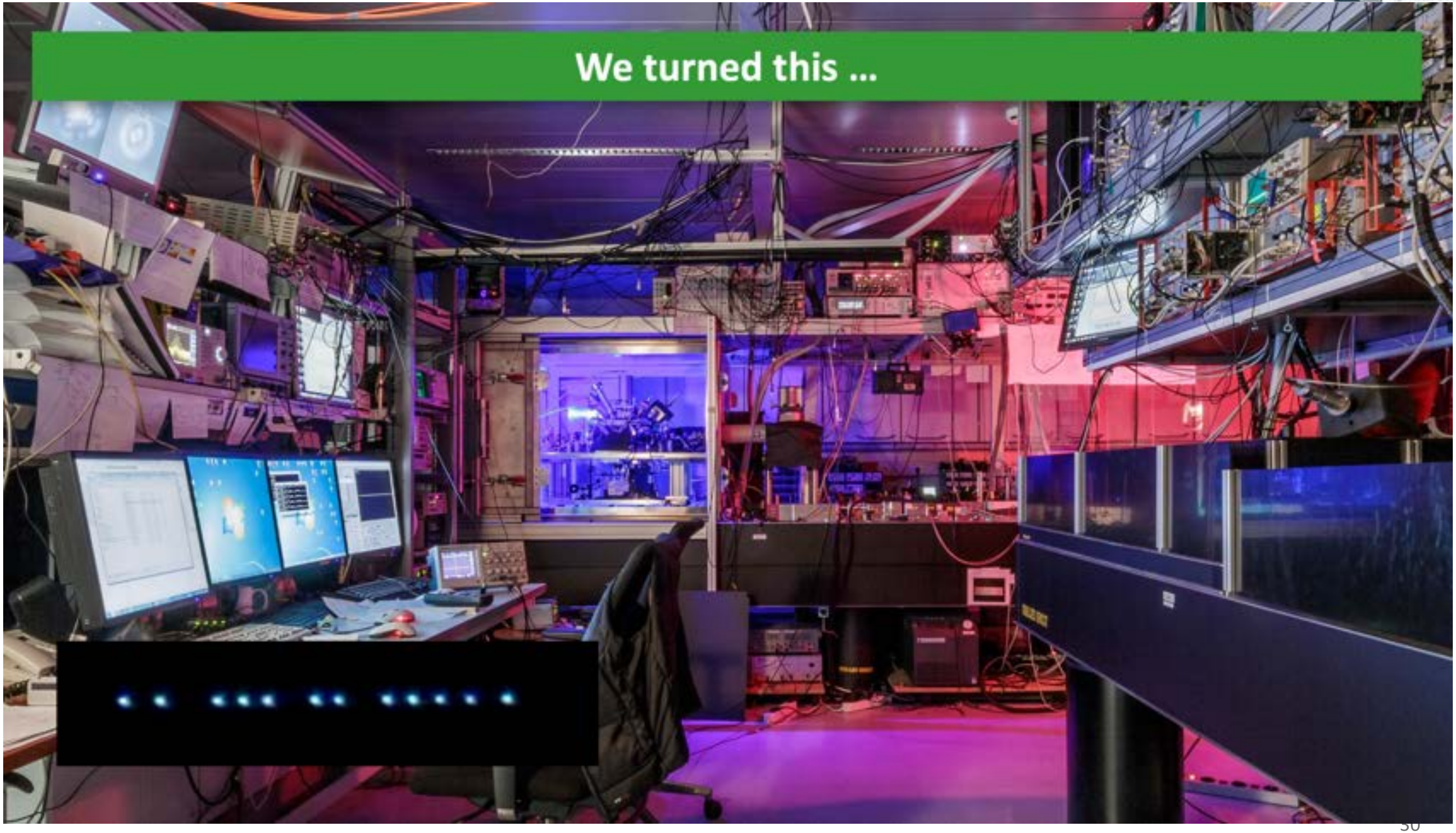
D-WAVE



TRAPPED ION

SUPERCONDUCTING

Source: *Machine Learning: Quantum vs Classical, Tariq M. Khan et al., IEEE Access, November 2020*



Slide: courtesy Thomas Monz, Uni Innsbruck

AQT

… into this

AQT DEMONSTRATED:

- 50+ ion-qubits
- 24-qubit entanglement
- Shor's algorithm
- Quantum Error Correction
- Fault-tolerant performance
- Demo'd finance applications
- Demo'd security applications
- Demo'd chemistry applications
- …

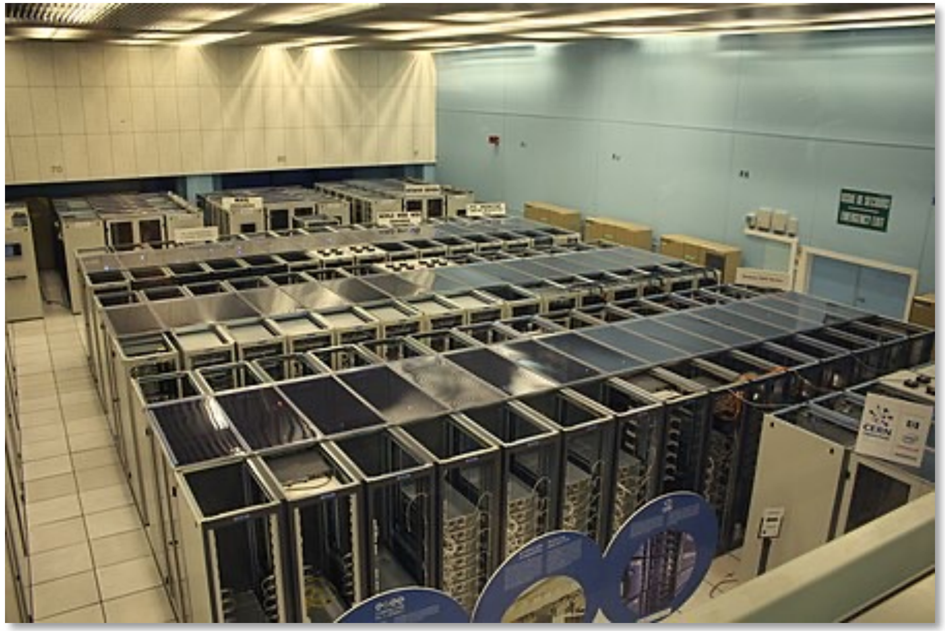
WITH OUR SYSTEM BEING:

- Rack-mounted
- Cloud-accessible
- Data-center compatible

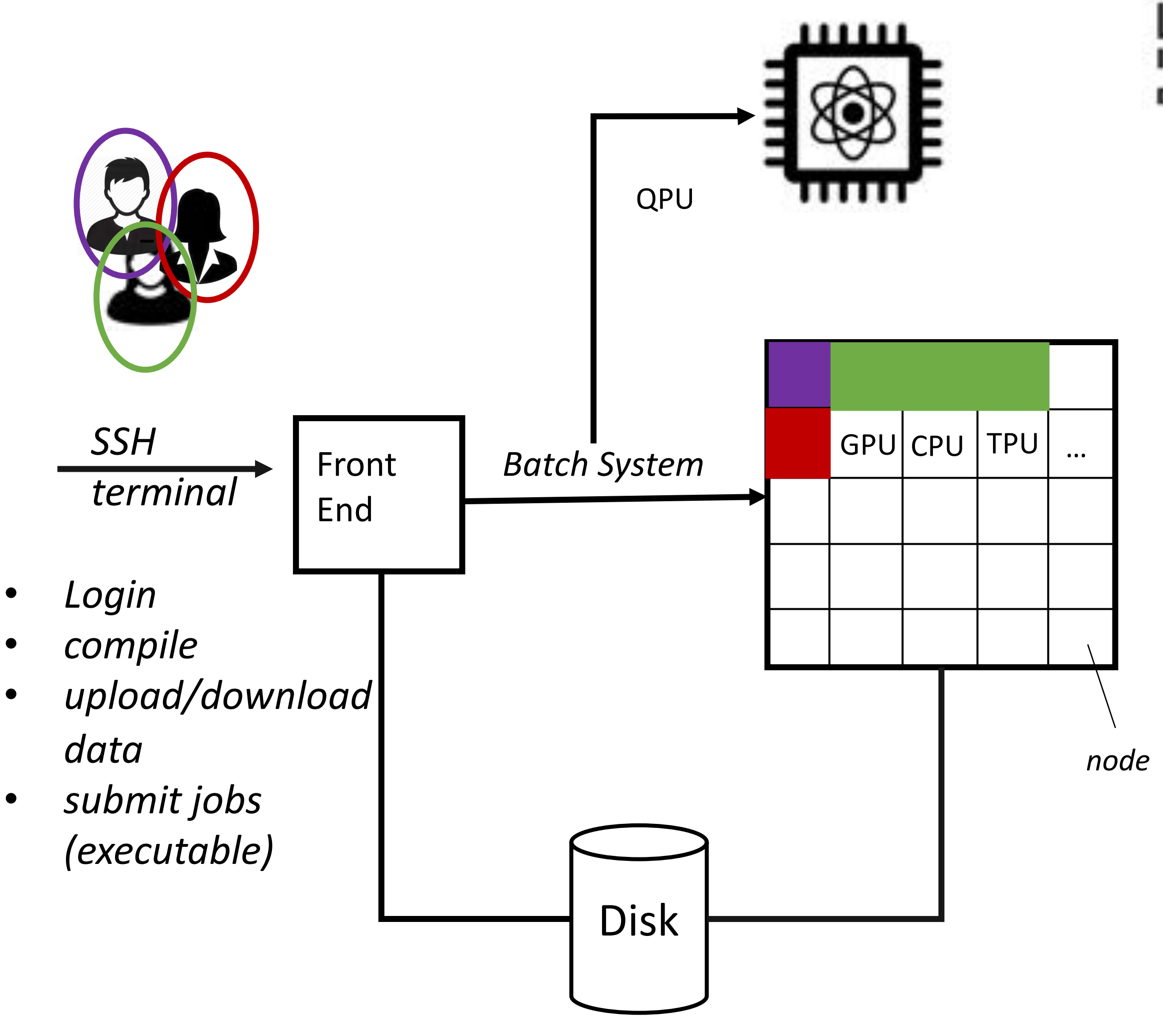
QPU

Slide: courtesy Thomas Monz, AQT

HPC Cluster



- Each “node” has its own operating system
- Nodes are interconnected with a network cable
- Higher performance demand more processors
- Accessed via front-end node/computer
- Shared with may users



HPQC Cluster

"Themis: Trustworthy and Sustainable Code Offloading"
Grant-DOI: 10.55776/PAT1668223

The system consists of
- 62 compute nodes with two 32-core Intel Xeon 8358 (Icelake) sockets each (i.e. 64 cores/node).
- Nodes are homogeneously networked with a high performance Infiniband (IB HDR100) interconnect.

- Improve QC performance
- Complement QEC capabilities
- Extend QC research with HPC

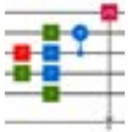
- Implement Hybrid Q-Libs
- **Benchmark PoC use-cases in hybrid infrastructure (HPC TU Wien)**

Direct access

Login console

HPC node

QPU



Circuit API

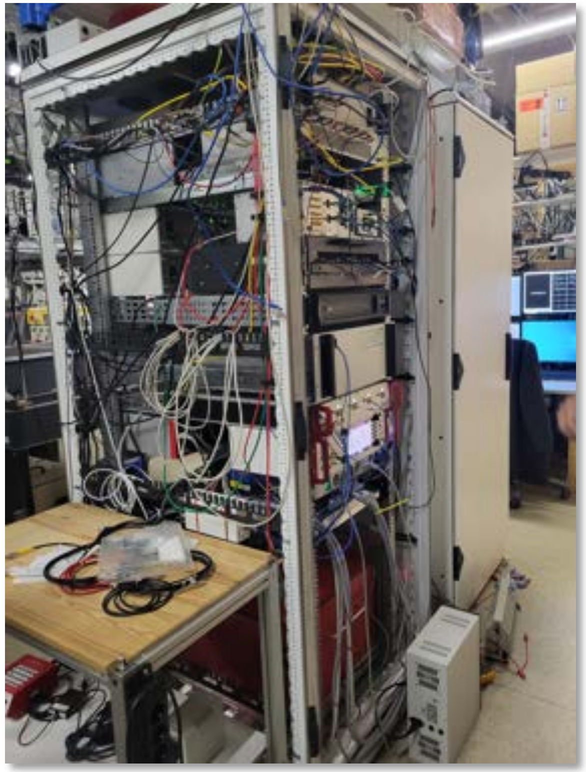
Offloading API

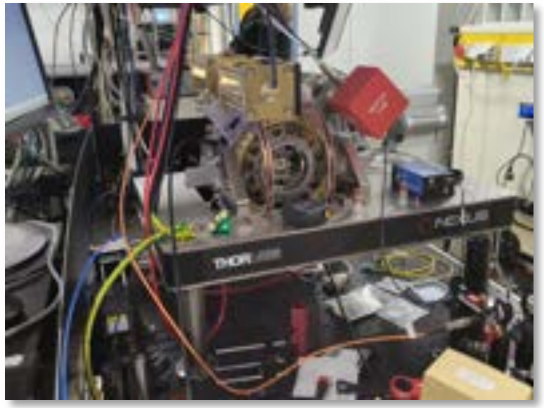
QC system

Assembly of a Quantum Processing Unit (QPU)

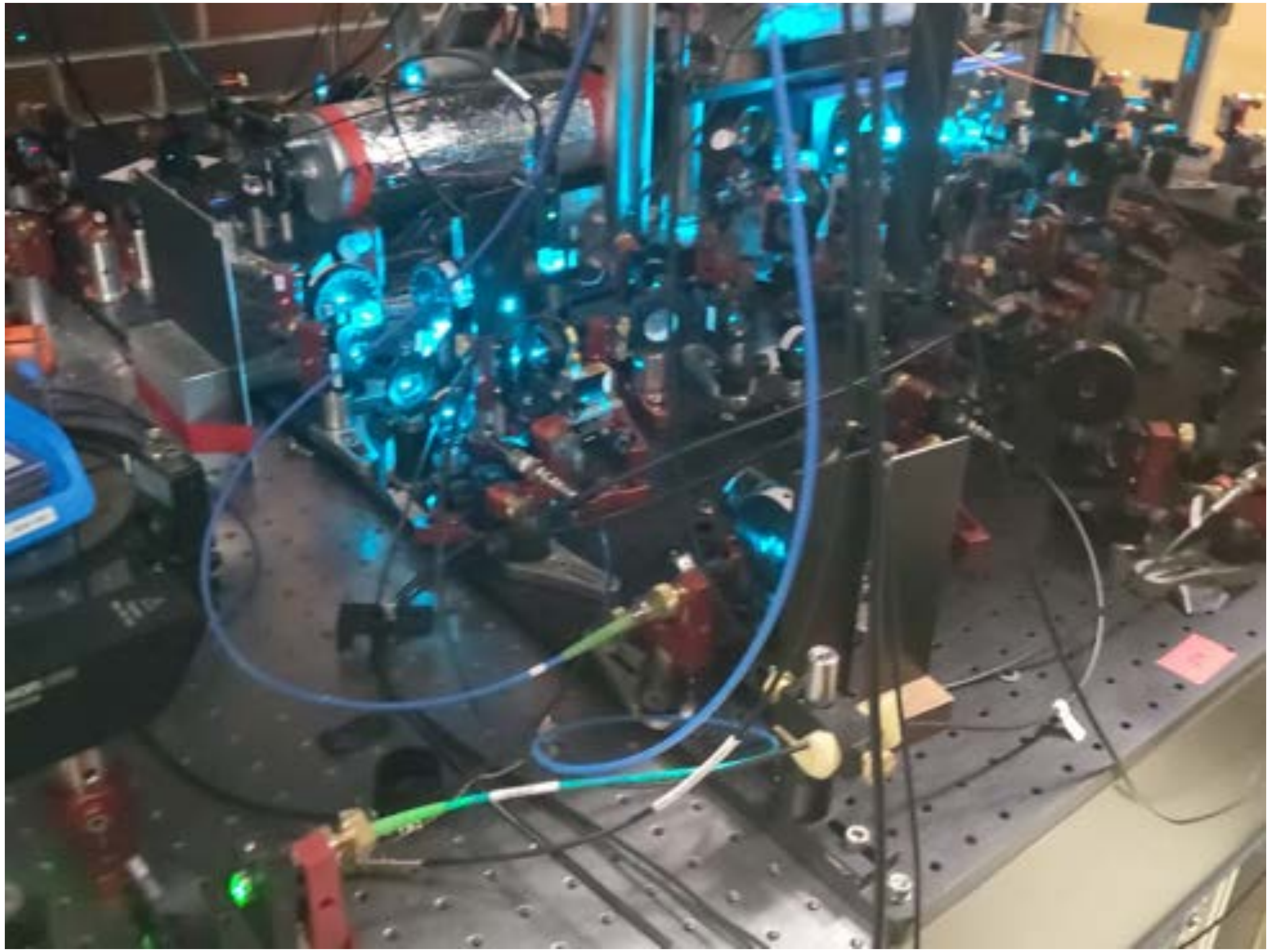
Picture: Courtesy AQT

- Move from Ca+ to Ba+
- New system with
  stage 1: 10x higher T_1
  stage 2: infinit. Higher T_1
- 2q error rate:
  legacy: < 10^{-2}
  target: < 10^{-3}
- Init error
  legacy: < 10^{-3}
  target: < 10^{-4}
- Readout error
  legacy: < 10^{-3} -> ~ 10^{-4}
  target: < 10^{-5}





Ion-trap quantum computer

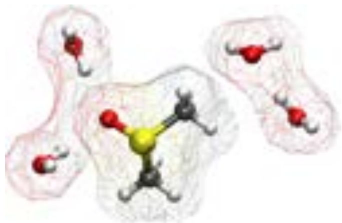


universität innsbruck

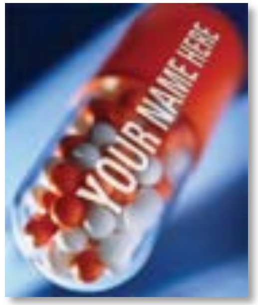
Courtesy Experimentalphysik, Univ. Innsbruck
Pictures, video: courtesy Vincenzo de Maio

Benchmarking on Hybrid Systems

P 36870- N 2023
"Transprecise Edge"



Molecular Dynamic Application



H

Θ

$$\lambda_\theta = \langle \psi(\Theta) | H | \psi(\Theta) \rangle$$

Quantum Machine

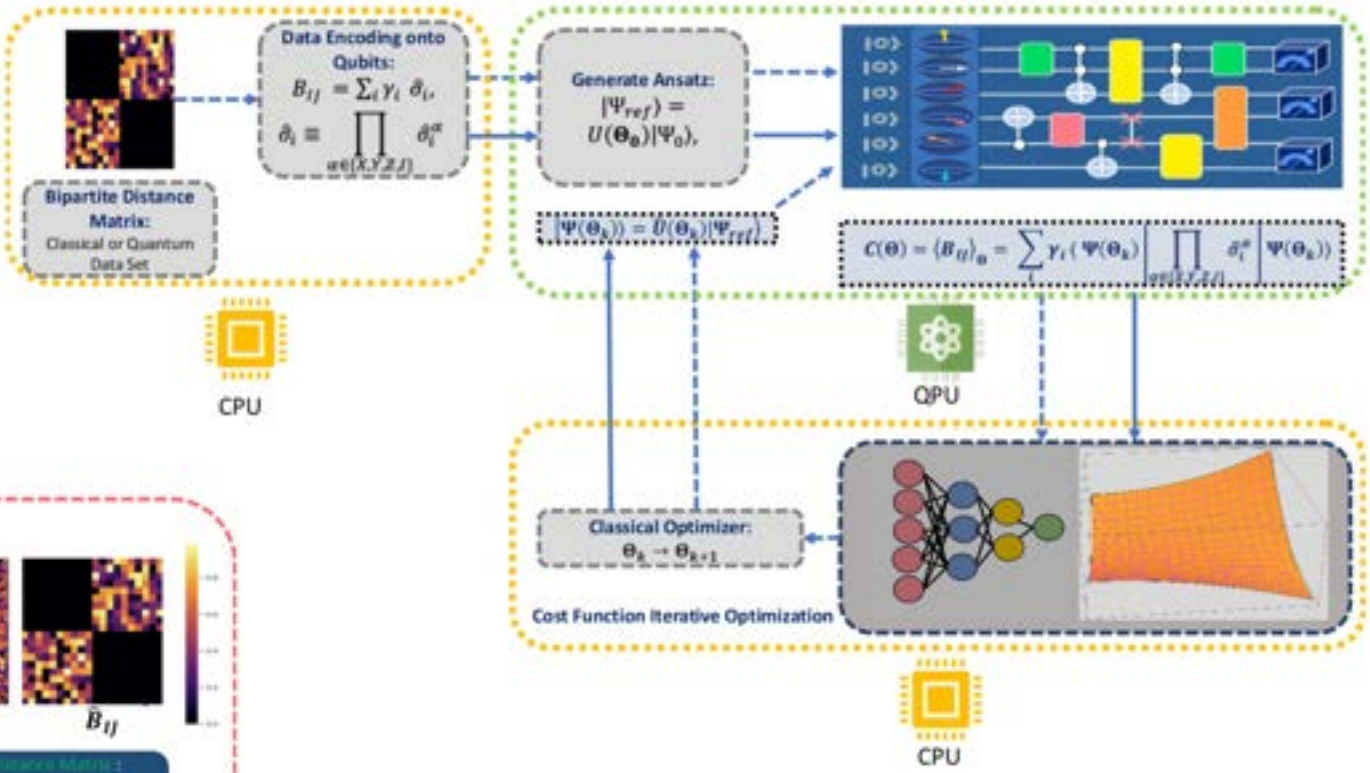
$C(\Theta)$

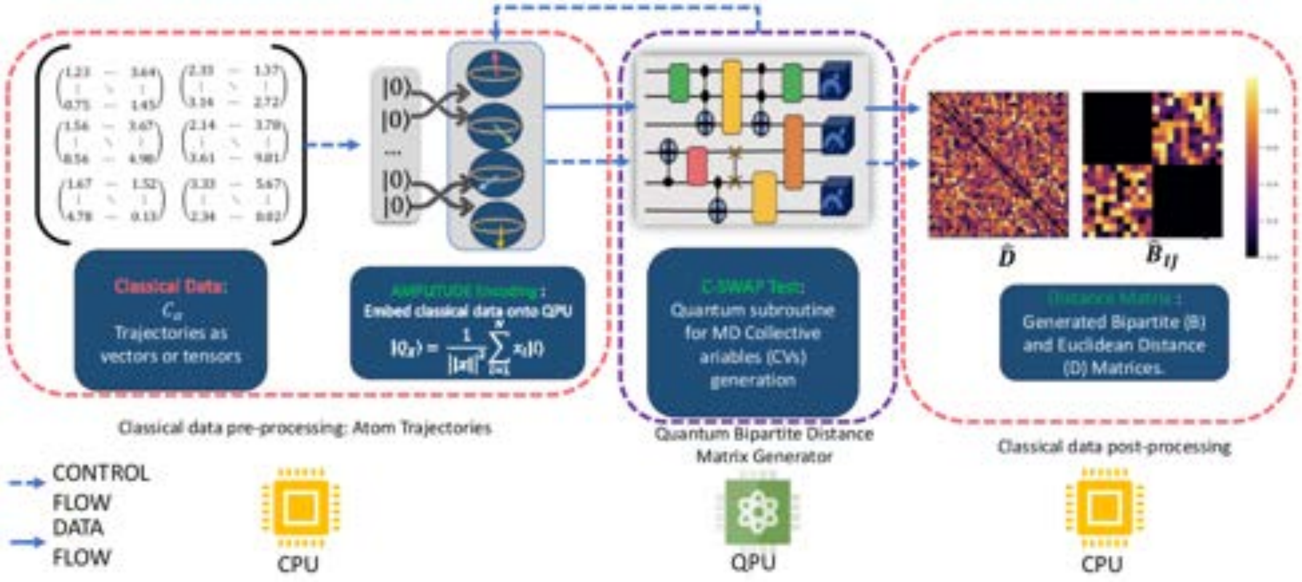
Optimizer

Classic Machine

Source: *Sandeep Suresh Cranganore, Vincenzo De Maio, Ivona Brandic, Tu Mai Anh Do, Ewa Deelman. Molecular Dynamics Workflow Decomposition for Hybrid Classic/Quantum Systems. IEEE eScience 2022, October 11-14, 2022 Salt Lake City, Utah, USA.*

Hybrid Classic Quantum Scientific Workflows

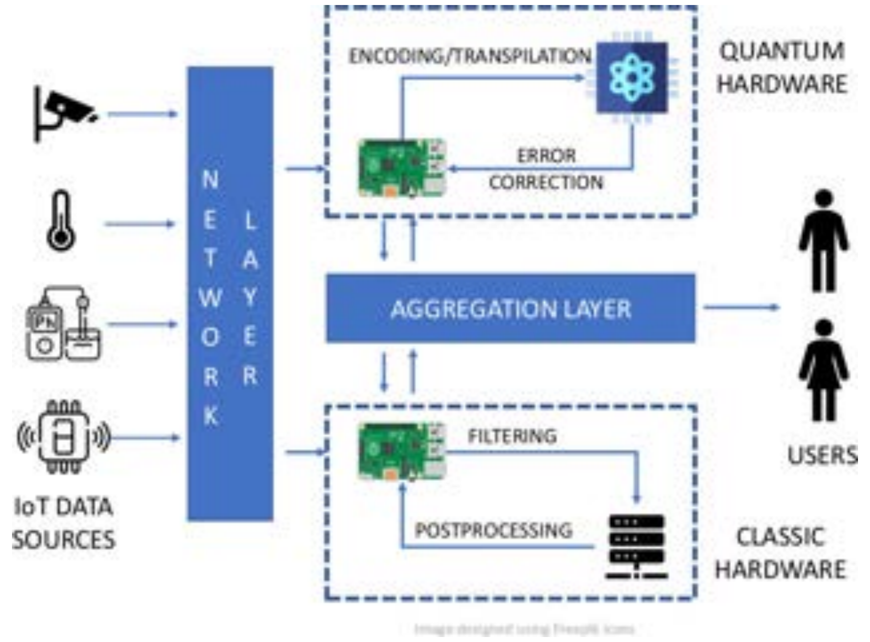




Source: Sandeep Suresh Cranganore, Vincenzo De Maio, Ivona Brandic, Ewa Deelman: Paving the way to hybrid quantum-classical scientific workflows. Future Gener. Comput. Syst. 158: 346-366 (2024)

Benchmarking Quantum Machine Learning

Triton – Transprecise Edge Computing P 36870- N 2023



Main hyperparameters:

- the **encoding** of the input into a quantum state
- the **structure** of the quantum circuit encoding the problem
- the **optimizer** that is applied to identify the optimal set of parameters
- **efficient execution of VQAs** requires fast communication between classic and quantum systems, and eventually offloading some parts of the computation

→ Preliminary results on IBM machines

S. Herbst

***Siemens Excellence Sholarship** for her Master Thesis: Beyond 0's and 1's: Exploring the Complexities of Noise, Data Encoding, and Hyperparameter Optimization in Quantum Machine Learning*

https://quantumzeitgeist.com/vienna-university-researchers-develop-quantum-edge-for-efficient-iot-data-streaming/

***S. Herbst**, V. De Maio, I. Brandic, "Streaming IoT Data and the Quantum Edge: A Classic Quantum-Machine Learning Use Case", International Workshop on Urgent Analytics for Distributed Computing, co-located with EuroPar 2023*

Outreach and Teaching

Co-organisation/sponsorship by CS TU Wien, Physik TU Wien, AQT

N. Friis

F. Zilk

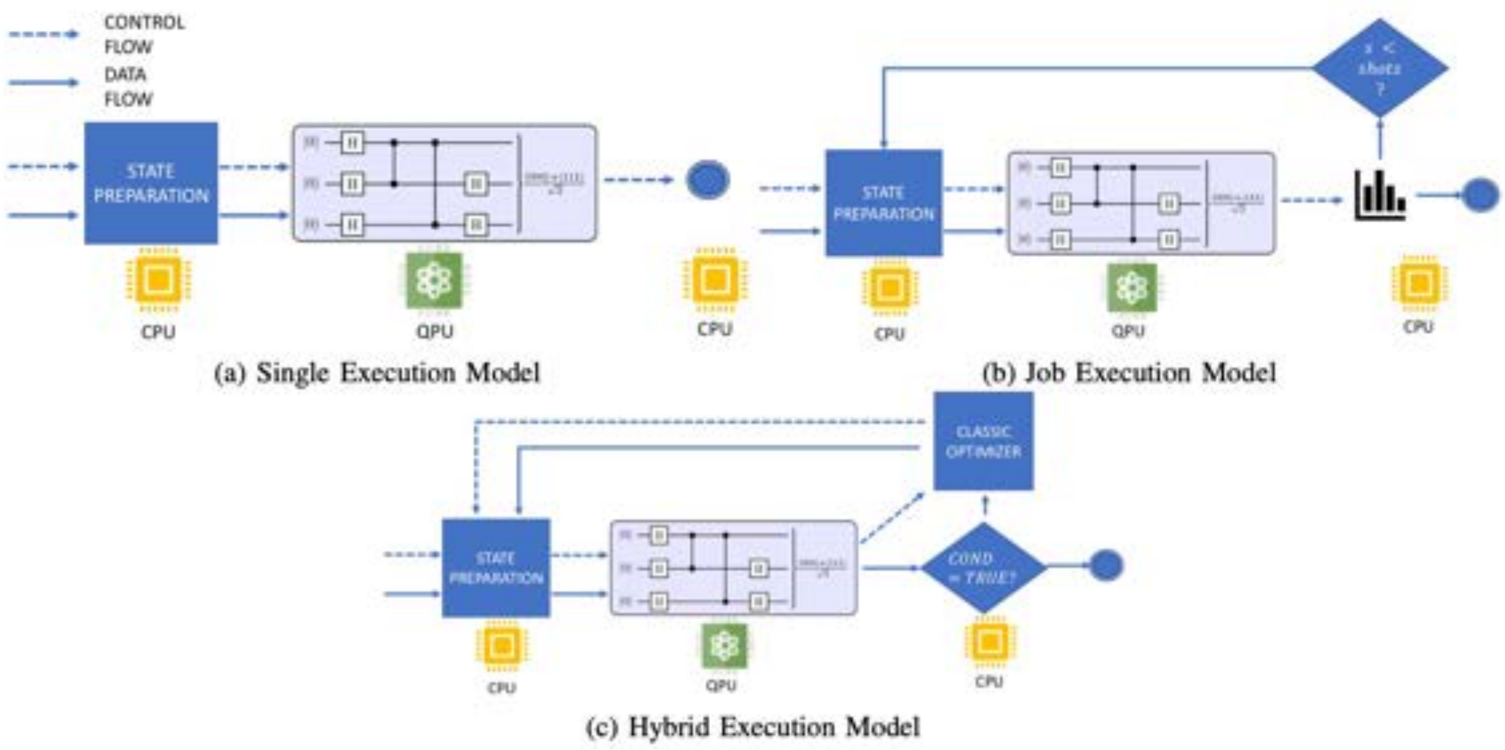
M. Kanatbekova

T. Guggemos

V. De Maio

To our knowledge first **joint lecture** on ‘Hybrid classical-quantum systems’ with a focus on applications currently attended by about 30 students
Organised by TU CS, TU Physics, Uni Wien Photonics

Lecture “Hybrid Classic-Quantum Systems”



Exercise 1: Design of classical K-Means algorithm. The goal of this exercise is to design the baseline implementation of classical K-means that will be used as a comparison for quantum-augmented K-means.

Exercise 2: Quantum-Augmented K-Means. The goal of this exercise is to implement a version of K-Means that is augmented with quantum tasks.

Exercise 3: Quantum SVM.

Source: Vincenzo De Maio, Meerzhan Kanatbekova, Felix Zilk, Nicolai Friis, Tobias Guggemos, and Ivona Brandic. Training Computer Scientists for the Challenges of Hybrid Quantum-Classical Computing. The 24th IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing. Philadelphia, May 6-9, 2024. (to appear).

Conclusion

- Trade off between multiple dimensions
  - Accuracy
  - Maintainability
  - Modularity
  - Energy consumption
  - …
- QPU for very specific operations
  - Chemistry
  - ML
- Challenge of integrating hybrid systems
- Mindset and education
- Limited hardware availability
  - Importance of simulators for teaching and engineering
  - Importance of benchmarking on real machines
- Focus on telescope technology

- You have to move out of the comfort zone!

Thanks to funding agencies and my team

TU WIEN

European Commission | Horizon 2020 European Union funding for Research & Innovation

cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

chist-era

QUANTUM AUSTRIA

W|W|T|F

Stadt Wien

MARIE CURIE ACTIONS

FWF

FFG Forschung wirkt.

SFG

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN