Sustainable Computing

Ivona Brandić

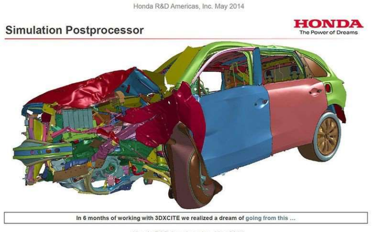
TU Wien

ivona.brandic@tuwien.ac.at

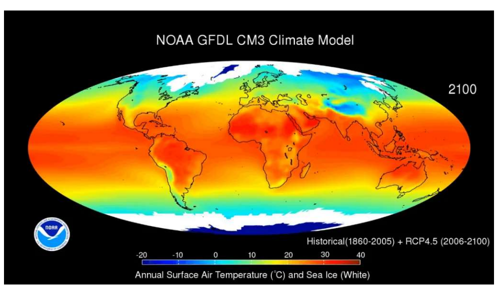
HPC Lab
High Performance Computing Laboratory

Computational Power

Simulation

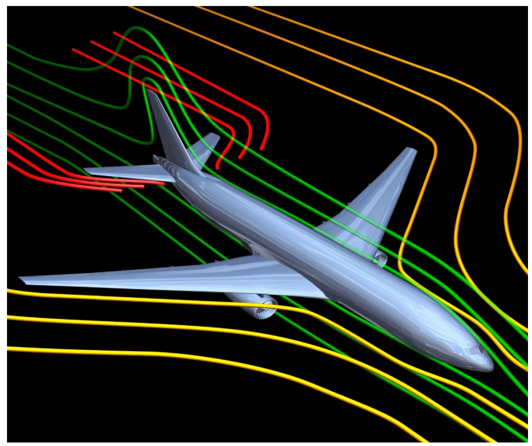


Mechanical Structure Simulation

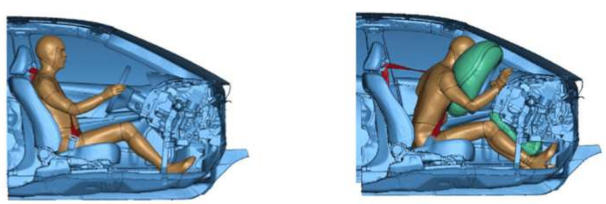


Climate Prediction

Optimization



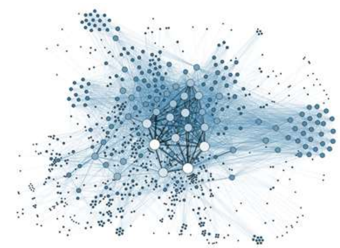
Airflow Optimization



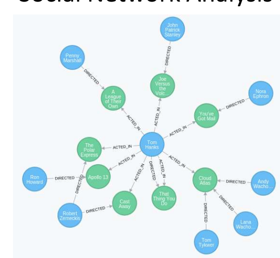
Finite Element Simulation
Hyper Parameter Optimization

Today: Analytics, AI

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

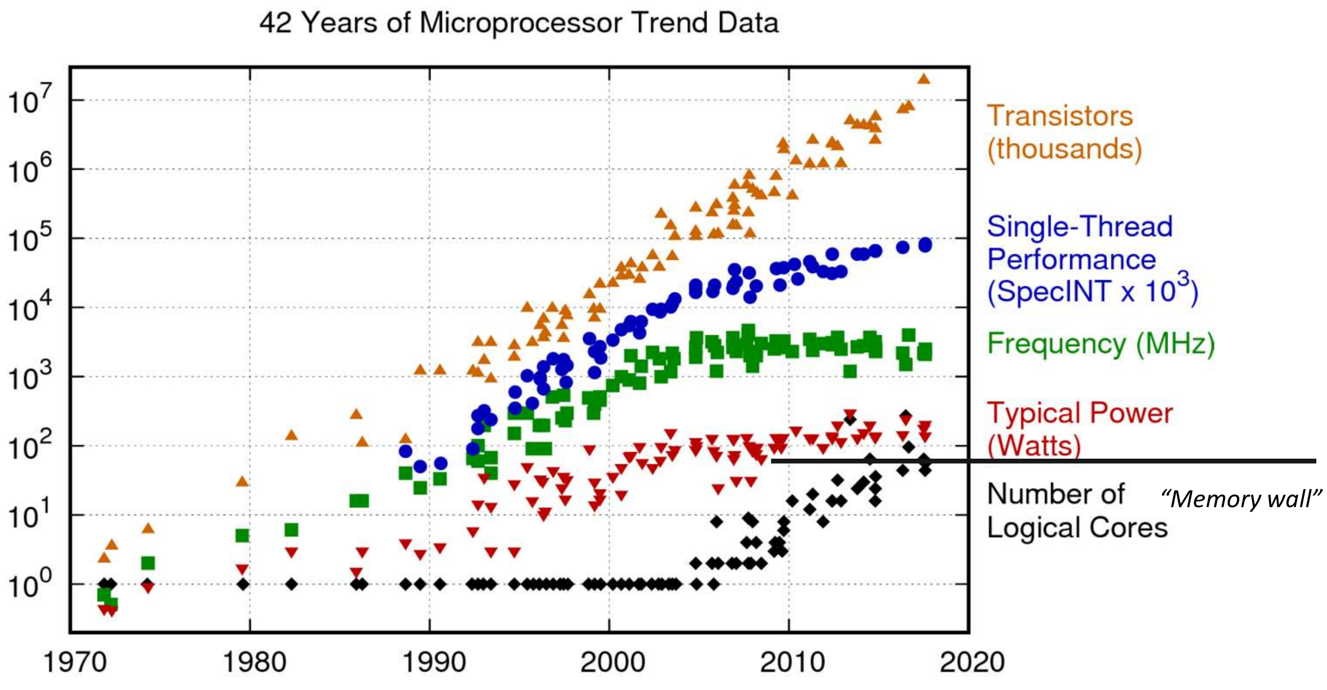


Social Network Analysis



Recommendation Engines

Problem 1: Practical Limitations



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

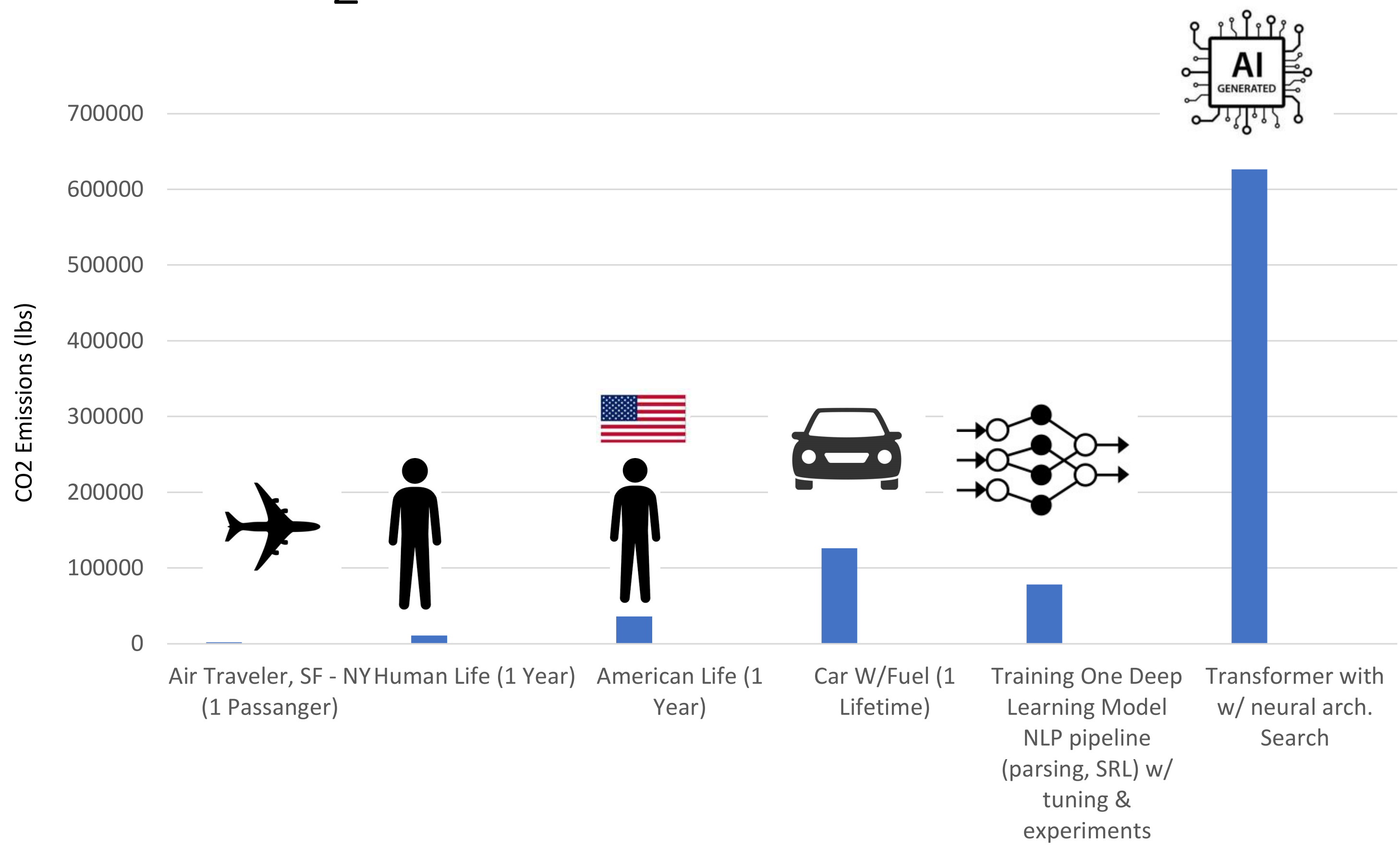
- #cores per chip doubles every 18 months instead of clock
- CPU-memory communication is becoming a bottleneck
- Too much heat is produced
- As transistors get smaller, power density increases because these do no scale with size anymore

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

Source: @mileszim on Twitter

C Lab
nce Computing Laboratory

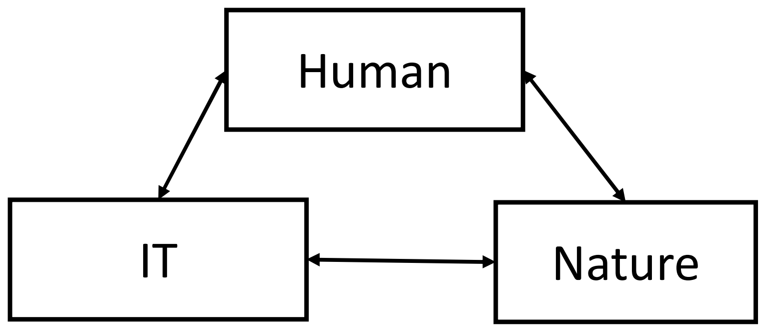
Problem 2: CO_2 Footprint of (generative) AI



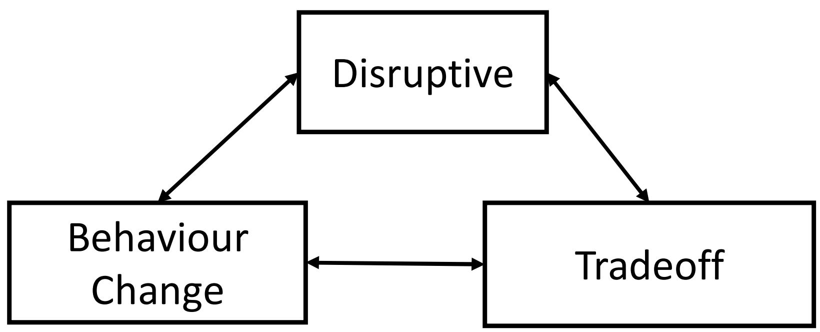
Source: Emma Strubell, Ananya Ganesh, Andrew McCallum: Energy and Policy Considerations for Deep Learning in NLP, ACL (1) 2019: 3645-3650
Inspiration for Visualisation: Keren Bergman, Multicore World 2023 https://2019multicoreworld.files.wordpress.com/2023/02/bergman-keren-23.pdf

Computational Sustainability

Actors:

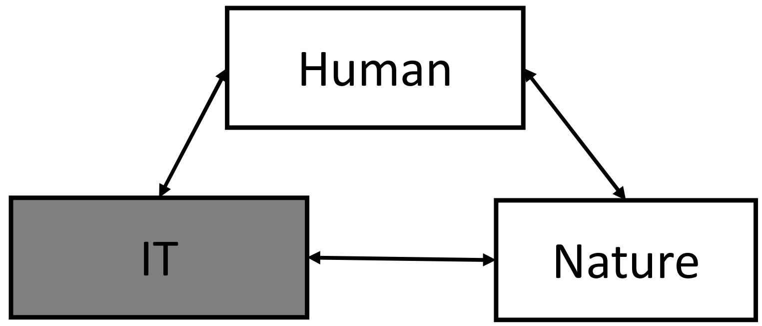


Methods:

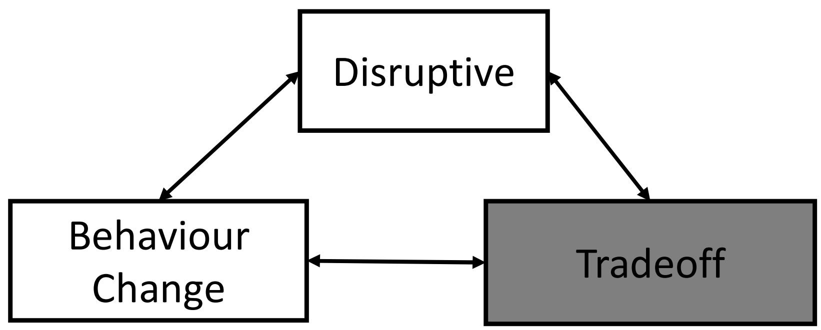


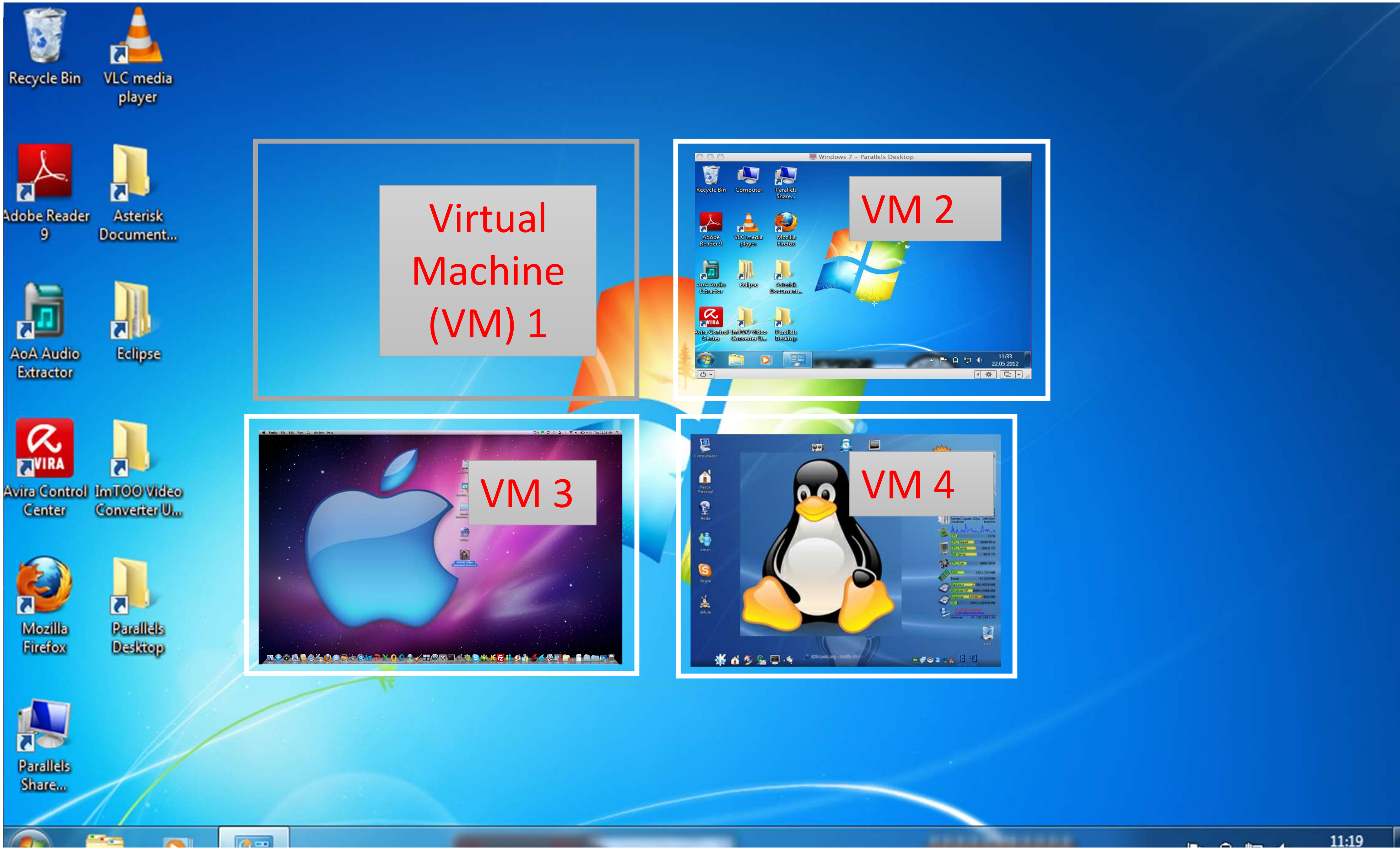
Computational Sustainability

Actors:

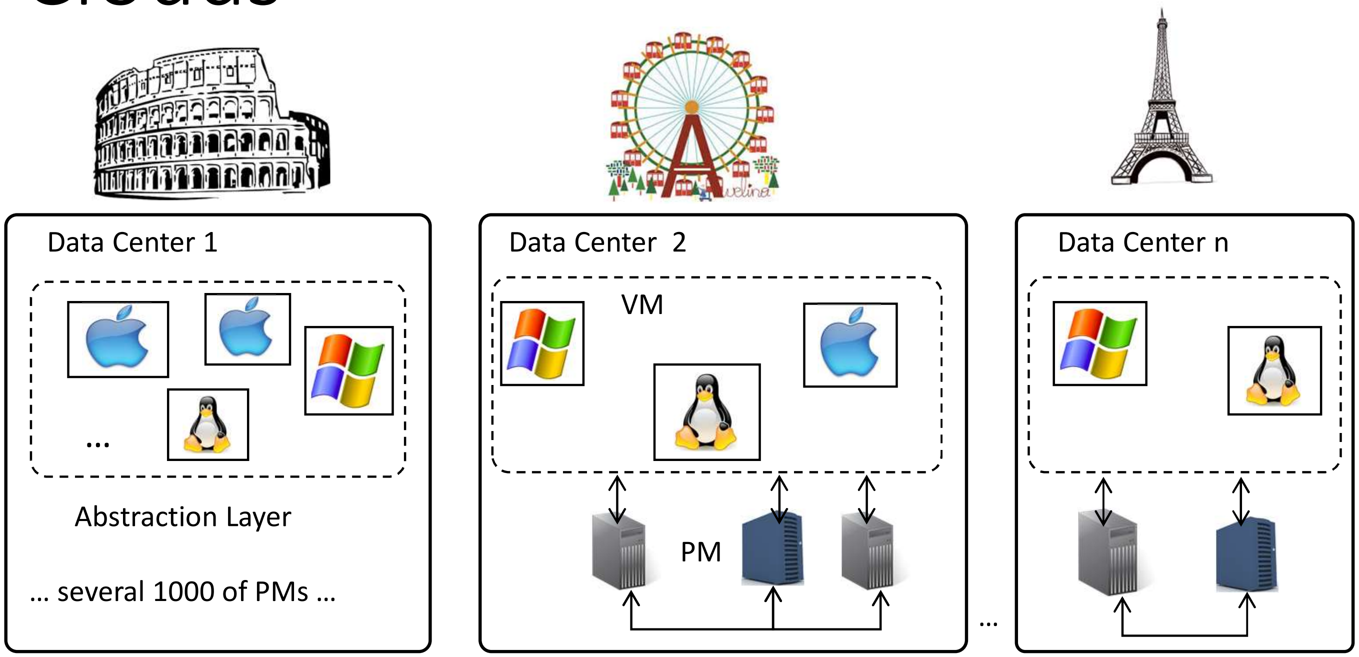


Methods:

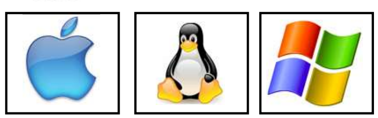



Recycle Bin
VLC media player
Adobe Reader 9
Asterisk Document...
AoA Audio Extractor
Eclipse
Avira Control Center
ImTOO Video Converter U...
Mozilla Firefox
Parallels Desktop
Parallels Share...
Virtual Machine (VM) 1
VM 2
VM 3
VM 4
11:19

Clouds

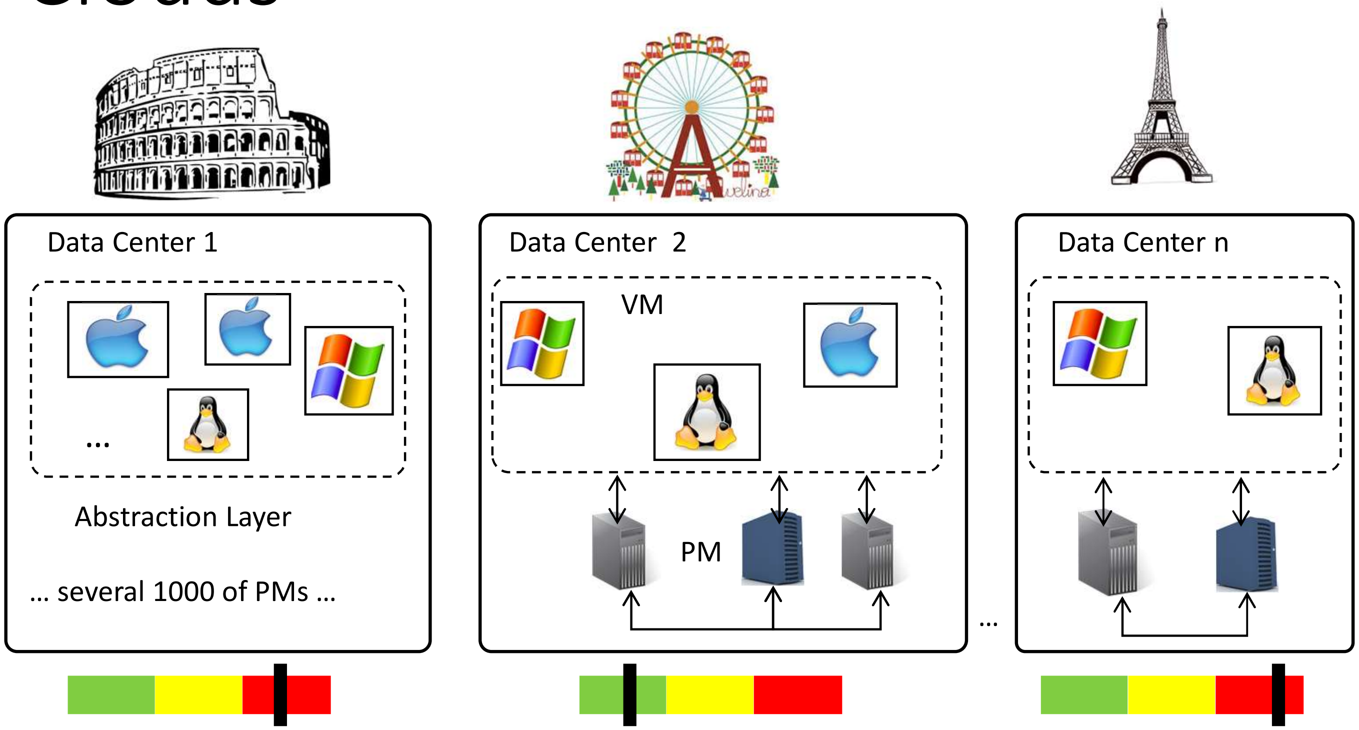


Physical Machine (PM)



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

Clouds

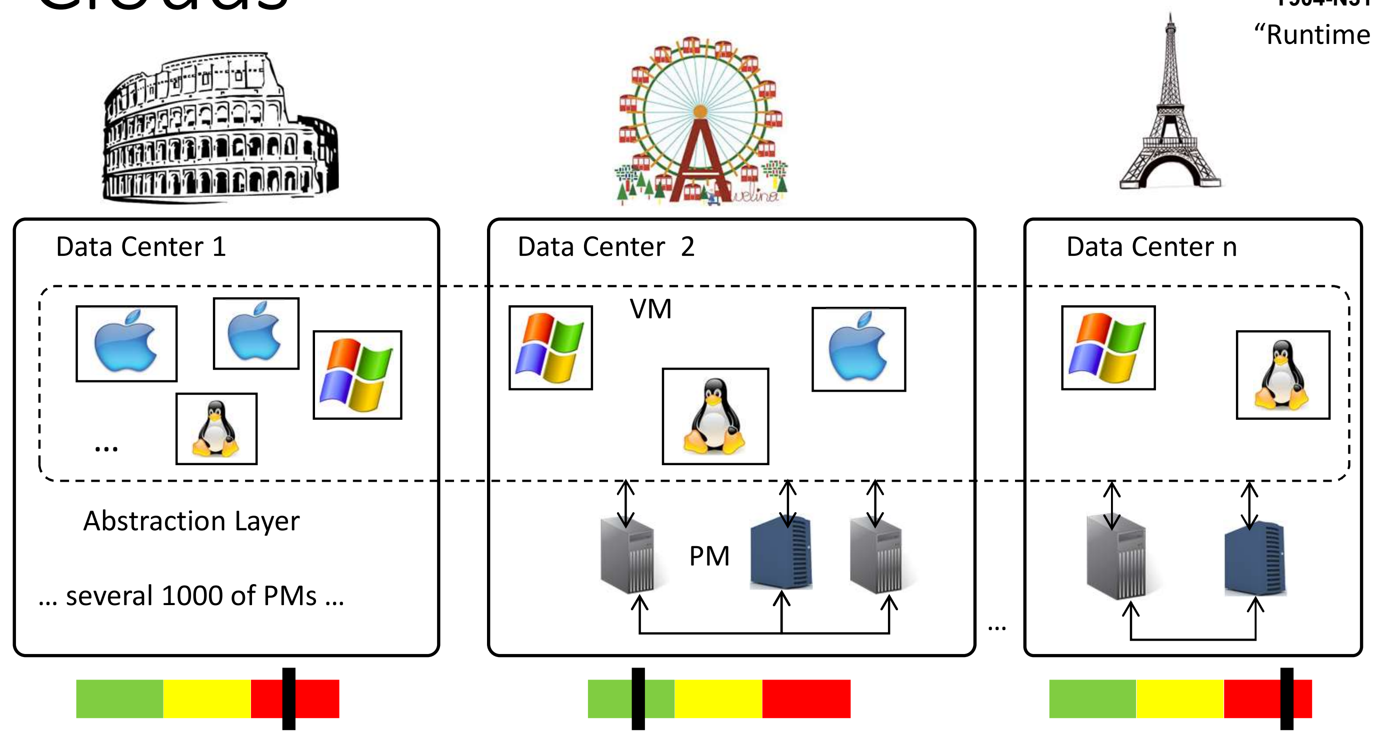


Physical Machine (PM)

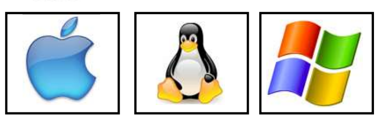
Virtual Machine (VM): Abstraction of a physical machine, *“simulation of a computer”*

Clouds

Y904-N31 START-Programme 2015
“Runtime Control in Multi Clouds”



Physical Machine (PM)

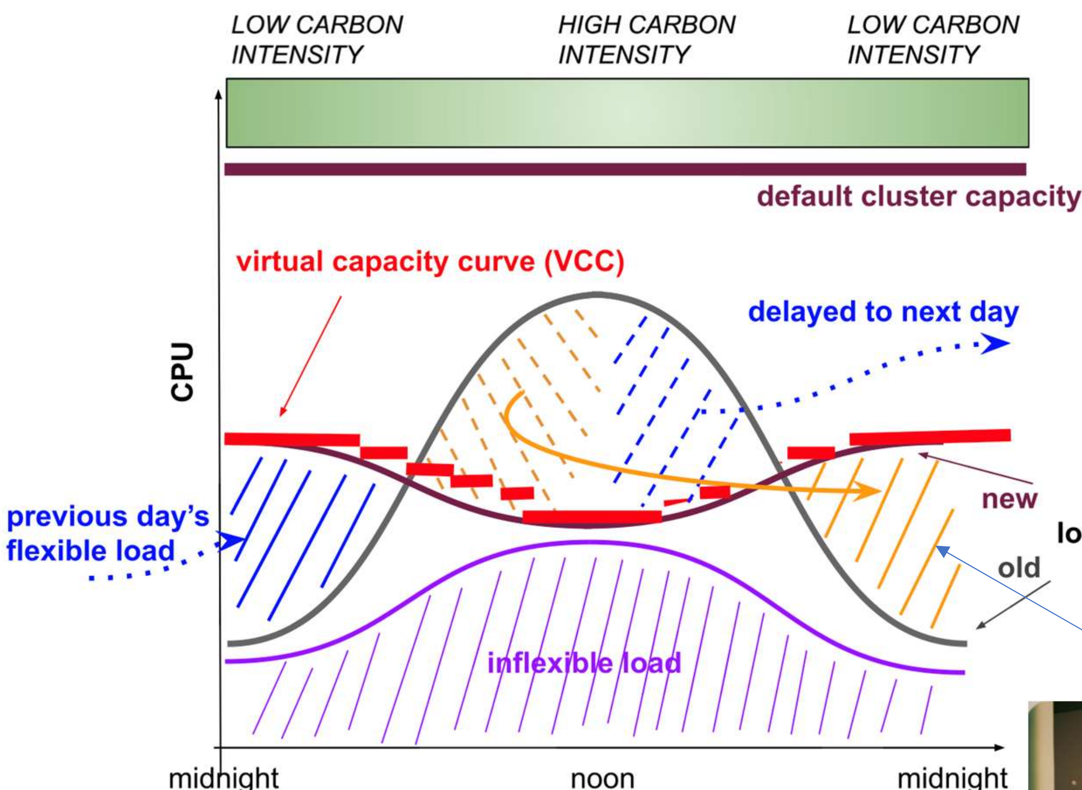


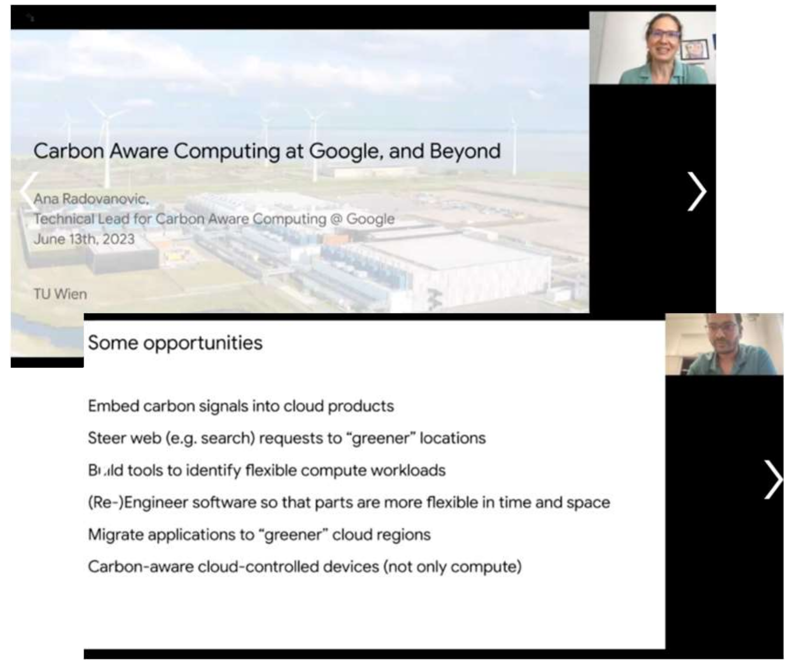
Virtual Machine (VM): Abstraction of a physical machine, *“simulation of a computer”*

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





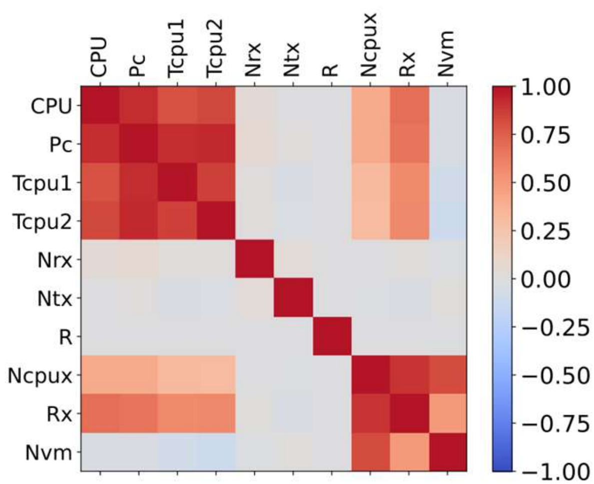
Ana Radovanovic (Google) & Shashi Ilager (TU)
*Lecture at TU Wien: **"Data Intensive Computing"***

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

Thermal Intricacies in Data Center

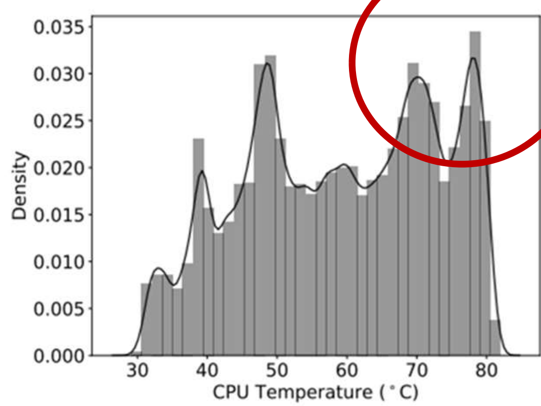
Metrics	Min		Max		Mean		SD	
	D1	D2	D1	D2	D1	D2	D1	D2
CPU	0.0	0.0	68.36	86.31	10.35	22.29	14.21	18.51
R_x	0.49	1.44	64.42	191.41	35.60	46.77	14.12	35.40
N_{Rx}	0	0	10.65+e8	52.63+e8	28.75+e5	25.59+e6	17.71+e6	20.94+e7
N_{Tx}	0	0	11.24+e8	59.97+e8	22.86+e5	14.55+e6	14.66+e6	19.41+e7
N_{vm}	0	0	54	261	9.6	9.01	5.58	32.91
N_{CPUx}	0	0	128	320.00	55.93	39.81	20.27	40.56
$fs_1 - fs_4$	280	-	13941.66	-	8804	-	2687.12	-
T_{CPU1}	29.14	26.66	82.01	79.23	57.00	59.52	11.06	11.60
T_{CPU2}	25.46	25.58	77.95	81.40	48.29	59.76	8.79	12.85
P	55.86	260.00	448	806.00	205.58	546.44	65.55	122.07
T_{in}	4	-	25.75	-	17.73	-	3.66	-

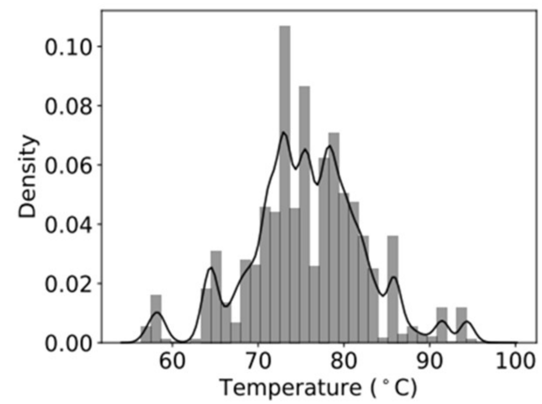
S. Ilager



Slide: cortesy Shashi Ilager

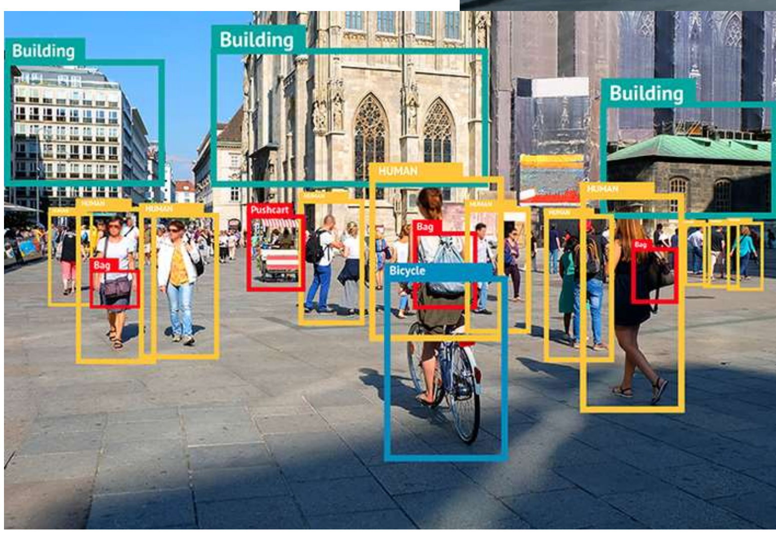
High number of hosts with peak temperature (although lower CPU utilization)





Source: S. Ilager, A. N. Toosi, M. Raj Jha, I. Brandic, R. Buyya, "A Data-driven Analysis of a Cloud Data Center: Statistical Characterization of Workload, Energy and Temperature", In Proceedings of the 16th IEEE/ACM International Conference on Utility and Cloud Computing (UCC2023), Vancouver, Messina, Italy, December 4-7, 2023.

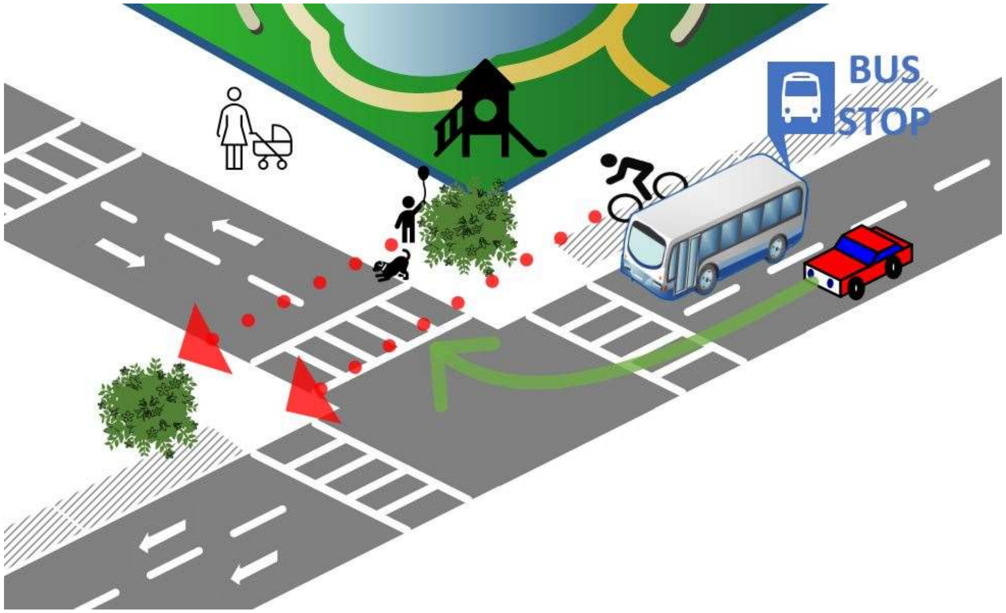
HPC Lab
High Performance Computing Laboratory



Symbolic Data Representation

Edge Computing in Action: Smart City

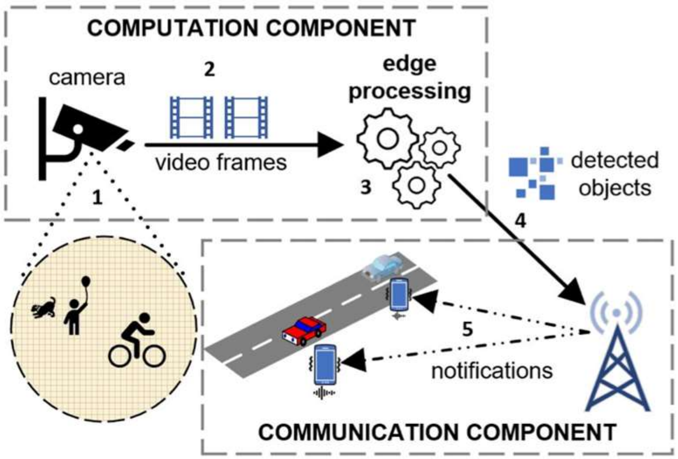
- Traffic accidents
- causing injuries and deaths
- Distractions, poor visibility (e.g., bad road and weather conditions), …
- Drivers' brake reaction time
- 1500ms on average



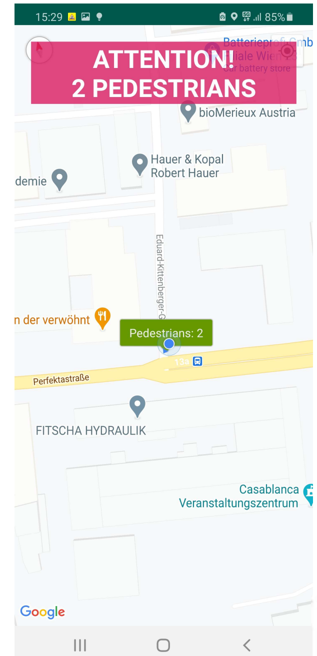
Deaths among pedestrians and cyclists:
29% of all EU road deaths
ETSC (European Transport Safety Council) PIN Flash Report 38

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

Smart City



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

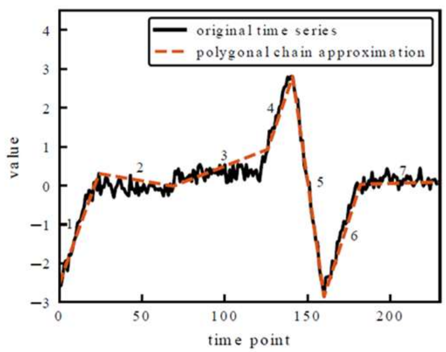


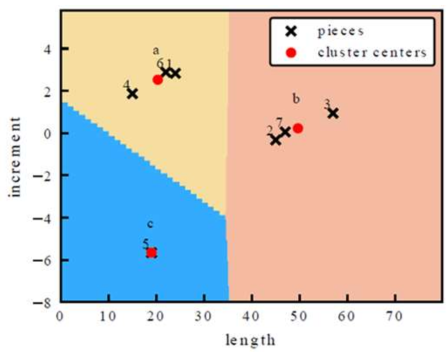
Slide: courtesy Ivan Lujic (Ericsson Nikola Tesla d.d.)

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
Ericsson Nikola Tesla d.d.

TU WIEN

*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 CONFERENCE 2023

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"

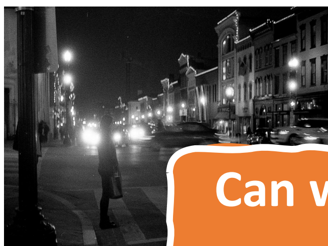
Daniel Hofstätter, Shashikant Ilager, Ivan Lujic and Ivona Brandic

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

Springer — 28 August – 1 September 2023 | Limassol, Cyprus — Lecture Notes in Computer Science

An Edge Computing Pedestrian Detection Scenario

Run-time Scenarios



Night | Few pe

Day | Crowd

Application Objectives

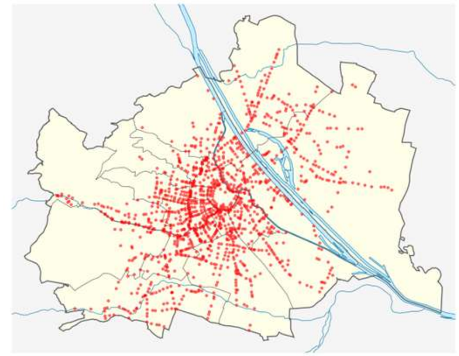
1. Being accurate
2. Consuming as little energy as possible

Can we determine the best application configuration to use?

Configurable Parameters

Parameter	Domain
Camera Resolution (R)	{1920x1080, 1280x720, 640x480}
Camera Frame Rate (FPS)	{1, 5, 10, 15, 20, 25, 30}
Object Detection Model (M)	{SSD MobileNet V1, EfficientDet-Lite0, ...}
Detection Threshold (T)	{0.1, 0,2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}
Use H/W Accelerator (TPU)	{true, false}

This slide has been designed using images from Flaticon.com



Source: Vienna Municipal Department 33, "Traffic lights with/without audible signal devices in Vienna," https://www.data.gv.at/, 2019, OpenData Österreich.

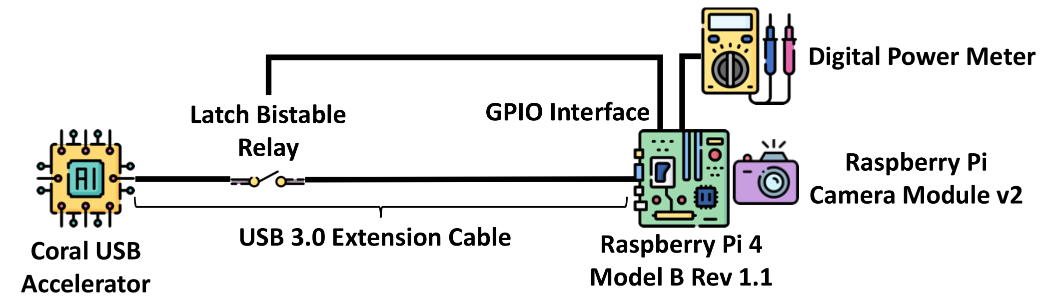
When is an H/W accelerator required?

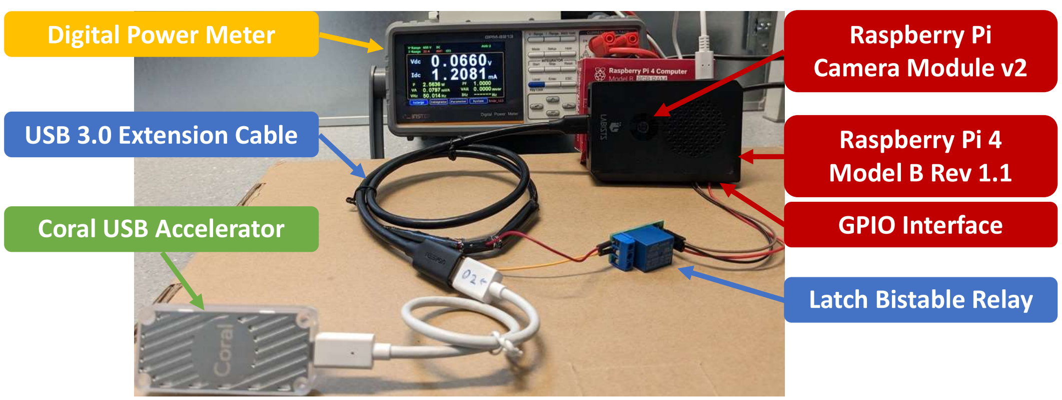
Which model?

How many frames per second?

Source: Alessandro Tundo, Marco Mobilio, Shashikant Ilager, Ivona Brandic, Ezio Bartocci, Leonardo Mariani. An Energy-Aware Approach to Design Self-Adaptive AI-based Applications on the Edge. 38th IEEE/ACM International Conference on Automated Software Engineering (ASE2023), Sep 11, 2023 - Sep 15, 2023, Luxembourg.

Evaluation Testbed





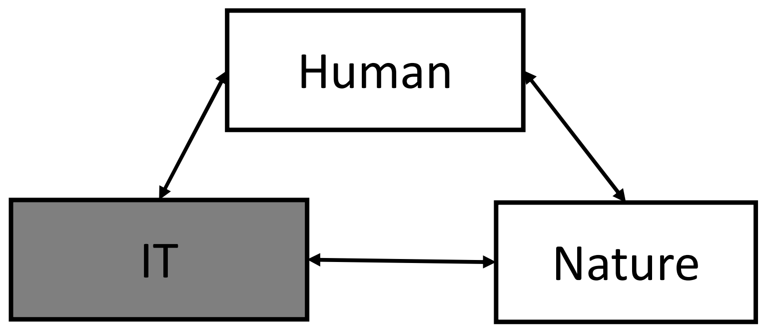
This slide has been designed using images from Flaticon.com

Source: Alessandro Tundo, Marco Mobilio, Shashikant Ilager, Ivona Brandic, Ezio Bartocci, Leonardo Mariani. An Energy-Aware Approach to Design Self-Adaptive AI-based Applications on the Edge. 38th IEEE/ACM International Conference on Automated Software Engineering (ASE2023), Sep 11, 2023 - Sep 15, 2023, Luxembourg.

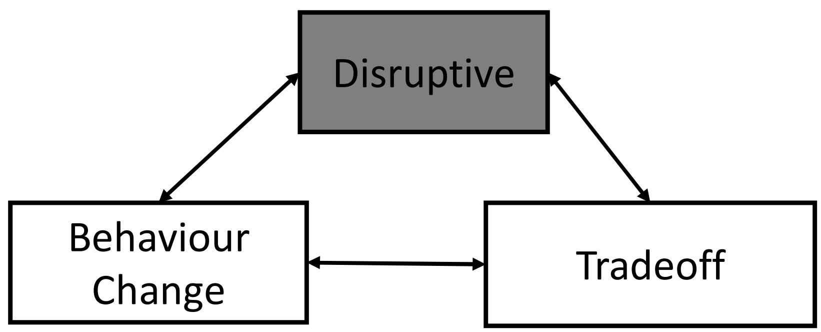
Slides: Cortesy Alessandro Tundo (University Milano-Bicocca)

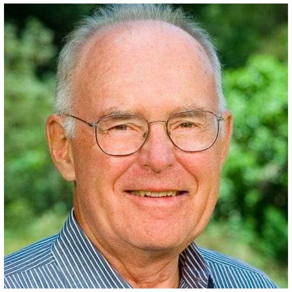
Computational Sustainability

Actors:



Methods:





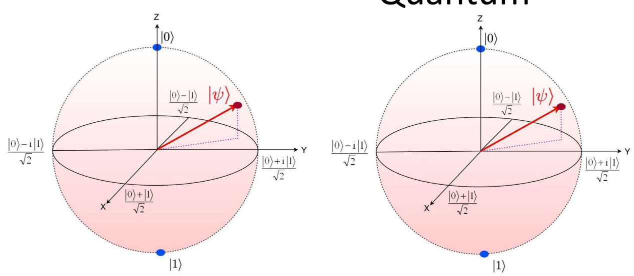
Gordon Moore:
Moore's Law
(1929 - 2023)

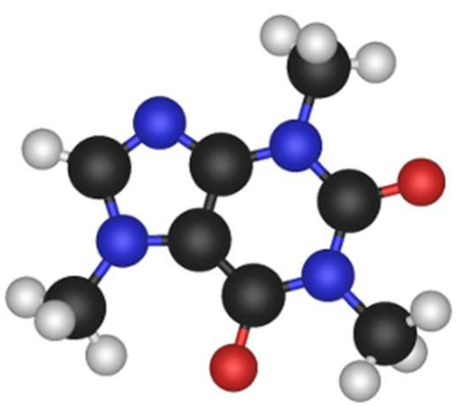
Data volumes are growing faster than the processing power

Alternatives:
- Neuromorphic Computing
- Quantum Computing

Beyond 0 and 1

Quantum





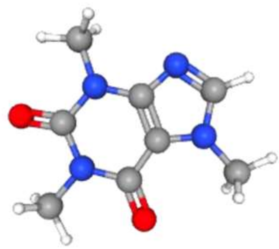
Von Neumann

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."

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

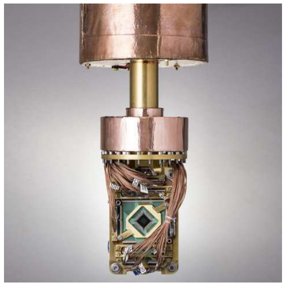
- Representing the energy configuration of a single caffeine molecule at a single instant requires approximately 10^{48} bits in a classical computer
- Can be done using 160 logical qubits on a quantum machine

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

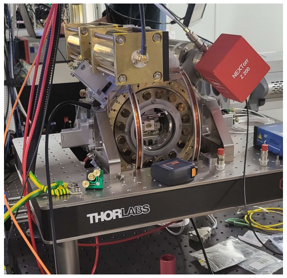
Courtesy Nico Einsiedler (IBM), TU Vienna Quantum Fall Fest 2022

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

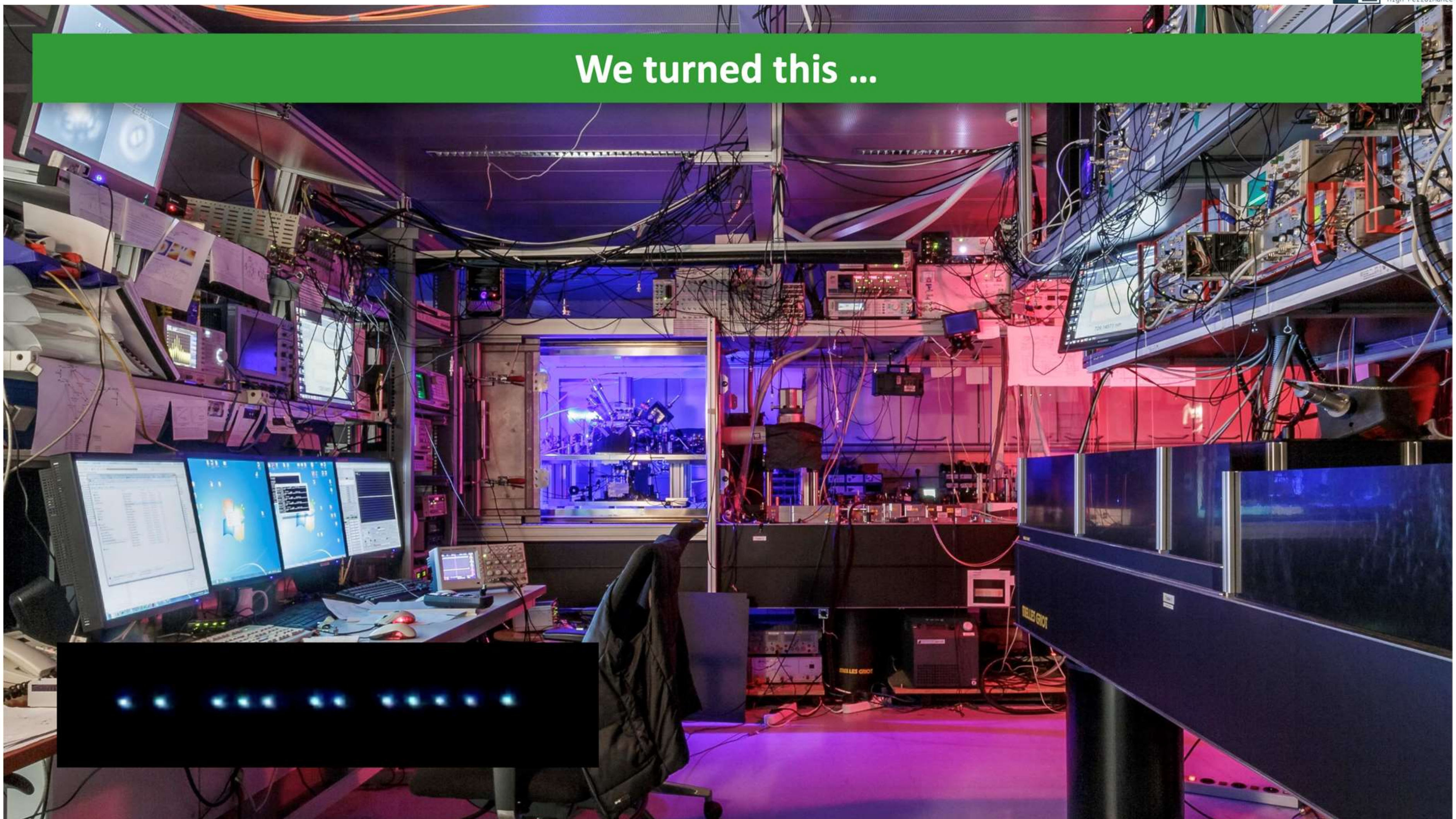


TRAPPED ION

SUPERCONDUCTING

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

We turned this …



Slide: courtesy Thomas Monz, Uni Innsbruck & AQT

AQT

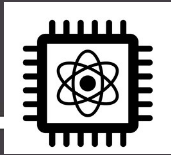
… into this

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:

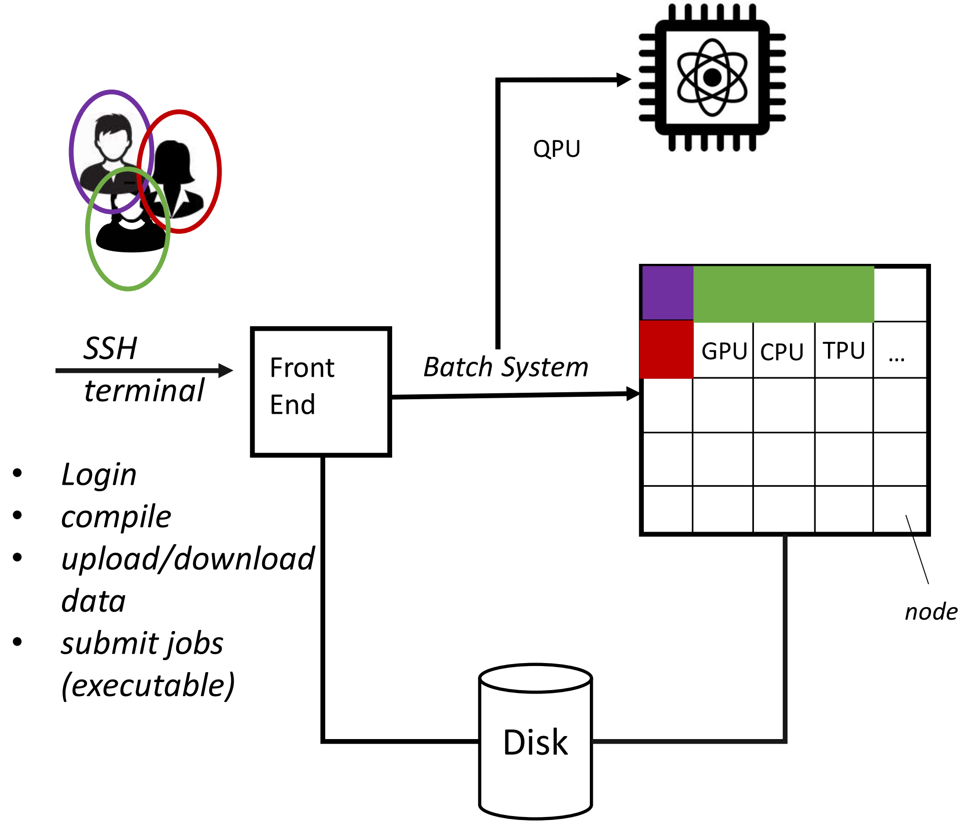
- Rack-mounted
- Cloud-accessible
- Data-center compatible


QPU

Slide: courtesy Thomas Monz, Uni Innsbruck & 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



FFG Forschung wirkt. HPQC HIGH-PERFORMANCE QUANTUM COMPUTING

HPQC Cluster

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

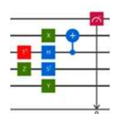
- Implement Hybrid Q-Libs
- Benchmark PoC use-cases in hybrid infrastructure

Direct access

Login console

aconet

HPC node



Circuit API

"string": "Hi",
"number": 2.5,
"boolean": true,
"null": null,
"object": { "name": "Kyle", "age": 24 },
"array": ["Hello", 5, false, null, { "key":
"arrayOfObjects": [
{ "name": "Jerry", "age": 28 },
{ "name": "Sally", "age": 26 }
]

Offloading API

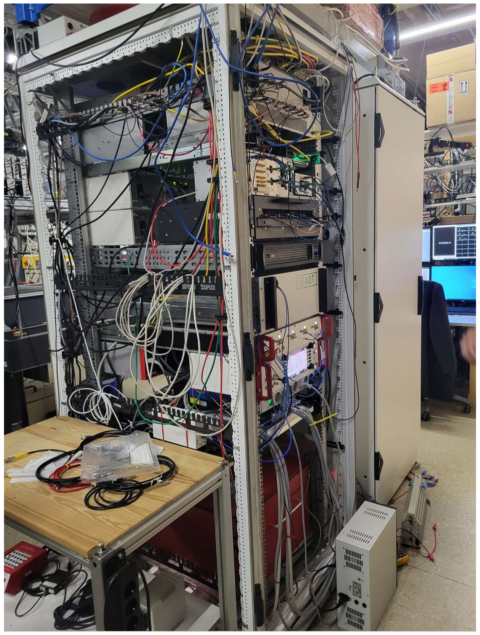
QC system

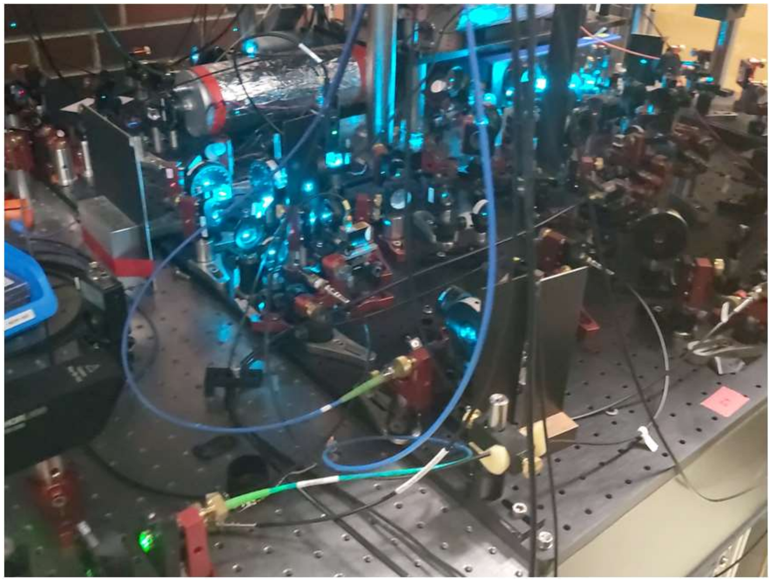
AQT JKU JOHANNES KEPLER UNIVERSITÄT LINZ universität innsbruck IBM MATHTEC TU WIEN FFG Forschung wirkt. HPQC HIGH-PERFORMANCE QUANTUM COMPUTING

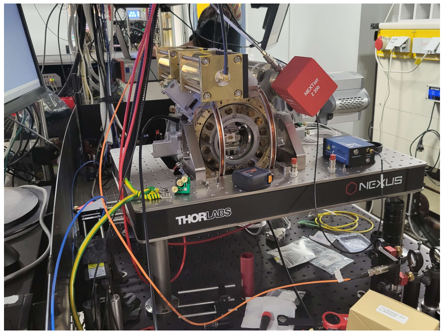
Installation of HPQC

Courtesy Experimentalphysik, Univ. Innsbruck

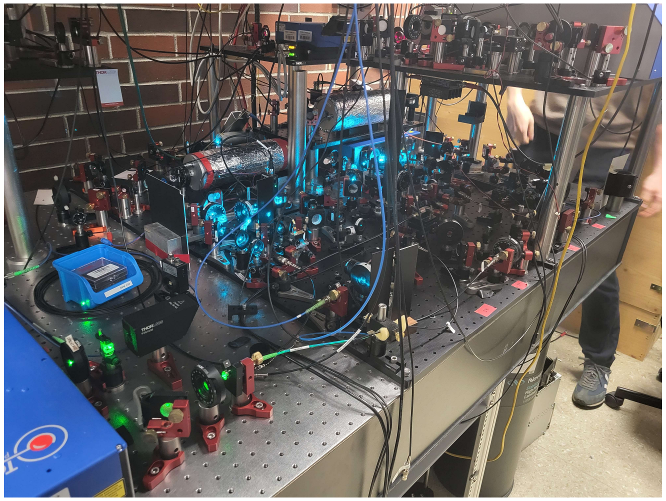
- Move from Ca+ to Ba+
- New system with
  stage 1: 10x higher T_1
  stage 2: infinit. Higher T_1
- 2q error rate:
  legacy: $< 10^{-2}$
  target: $< 10^{-3}$
- Init error
  legacy: $< 10^{-3}$
  target: $< 10^{-4}$
- Readout error
  legacy: $< 10^{-3}$ -> ~ 10^{-4}
  target: $< 10^{-5}$







Ion-trap
quantum
computer



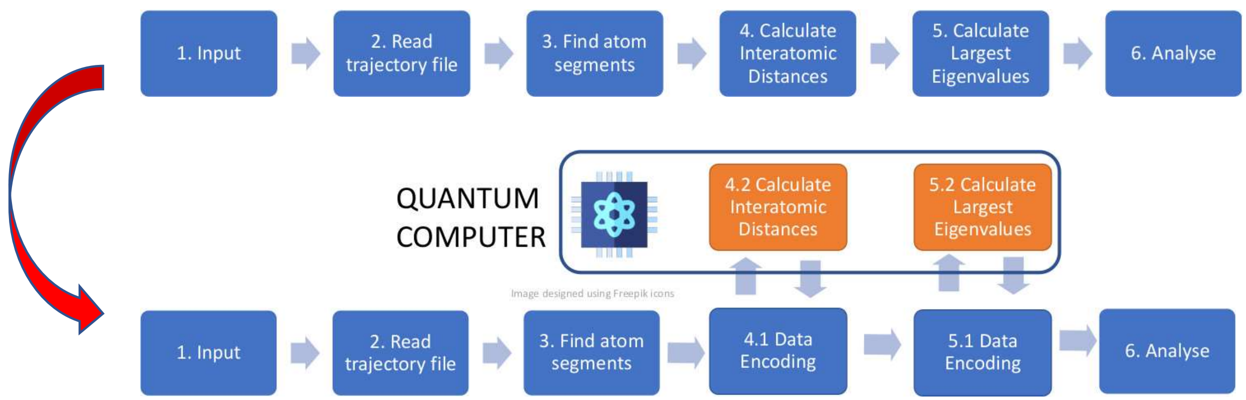
universität
innsbruck

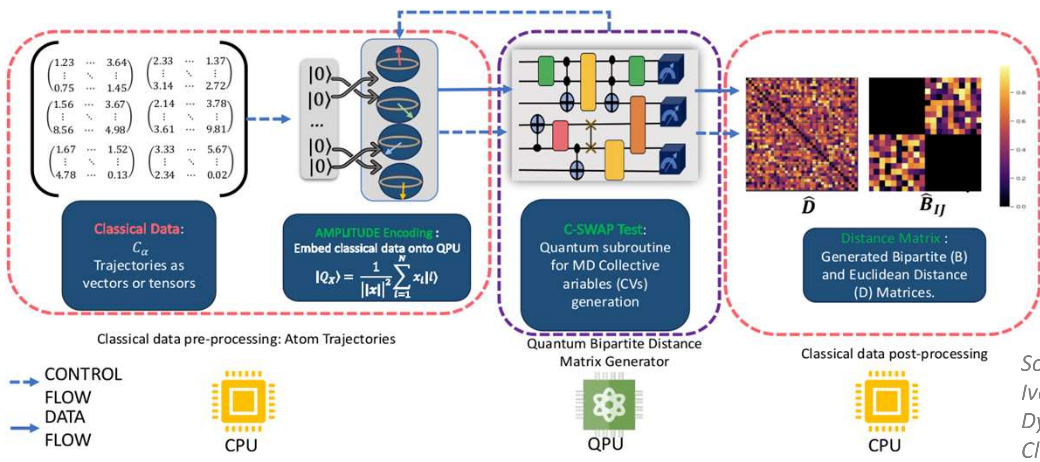
Courtesy Experimentalphysik, Univ. Innsbruck
Pictures, video: courtesy Vincenzo de Maio

HPQC
HIGH-PERFORMANCE QUANTUM COMPUTING

HPC Lab
High Performance Computing Laboratory

Benchmarking Molecular Dynamics Application



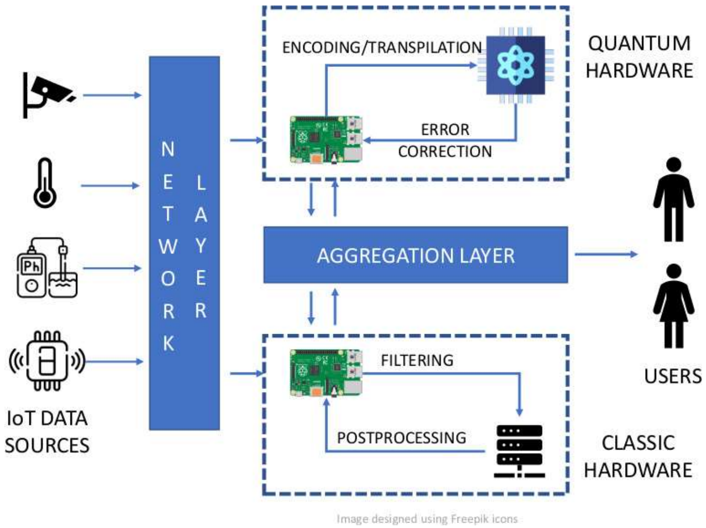


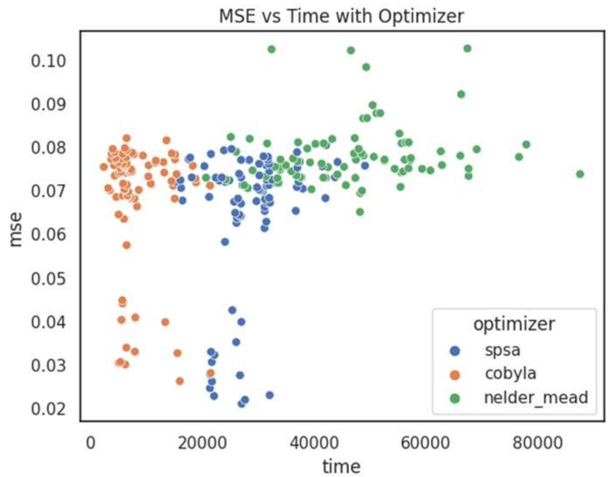
https://pegasus.isi.edu

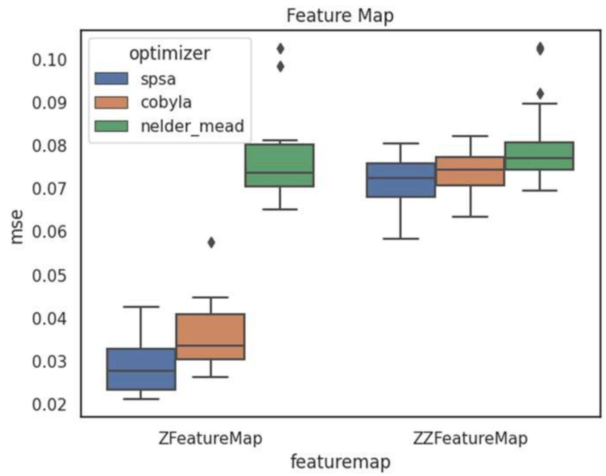
- → Identification of tasks that can be executed on quantum hardware;
- → Benchmarking on IBM Quantum machines;

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.

Benchmarking Quantum Machine Learning







- → Benchmarking of Quantum Regression on typical IoT dataset
- → Conceptual design of a Edge-enhanced pipeline for QML
- → Preliminary results on IBM machines

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

HPC Lab
High Performance Computing Laboratory

What’s next: Protein Folding

V. De Maio

Submission

Team A1 | Team A2 | Team A3

Team Members

1. Akshay Masetty (masettyakshay2002@gmail.com)
2. Ayan Barui (ayanbarui86@gmail.com)
3. Laia Domingo Colomer (laia.d.c@hotmail.com)
4. RAJUL SHARMA (rajulsharma001@gmail.com)
5. Rahul Dev Sharma (sci94tune@gmail.com)
6. Sravani Yanamandra (sravani.yanamandra@research.iiit.ac.in)
7. Srushti Patil (srushtitukarampatil@students.iisertirupati.ac.in)
8. Vincenzo De Maio (vinc.demaio@gmail.com)

ingenii

Protein Folding Using QC: PROJ- 00121

Source: https://quantumopenai.com/protein-folding-using-qc-proj-00161/

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

Conclusion

- Trade off
  - 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

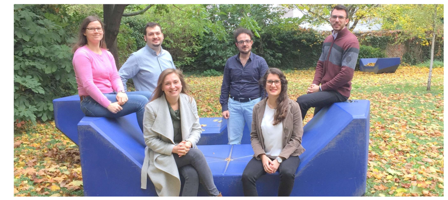
Thanks to funding agencies and my team

HPC Lab
High Performance Computing Laboratory

2010

2013

2023



2017

2021

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

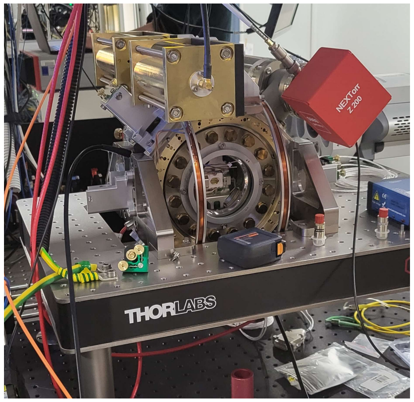
FFG Forschung wirkt.

86% of my group are third party funded – Thank you!

Quantum Hardware

D-WAVE

SUPERCONDUCTING



TRAPPED ION*

- No “best” technology at the moment
- No standards
- Every machine different architecture
- Integration →

FFG Flagship Project High Performance Integrated Quantum Computing (HPQC)

**Courtesy of University of Innsbruck, department of experimental physics*

HPQC
HIGH-PERFORMANCE QUANTUM COMPUTING

Cloud DC - Temperature Evolution

Understanding Thermal Behaviour in DC

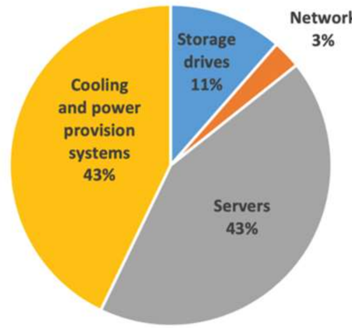
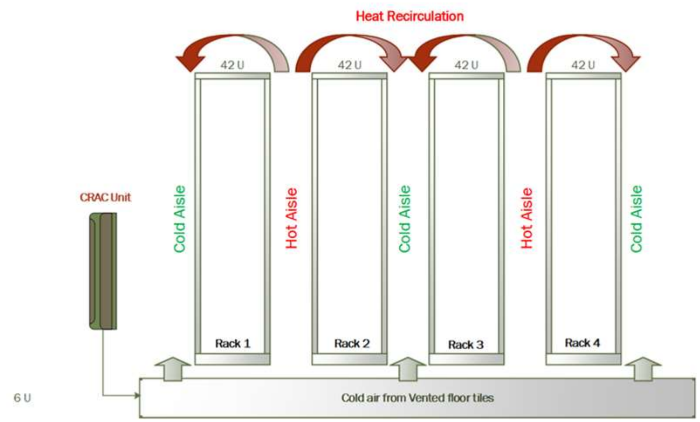


Figure 1. Fraction of U.S. data center electricity use in 2014, by end use. Source: Shehabi 2016.

S. Ilager

Features	Definition
CPU	CPU Load (%)
R	RAM- Random Access Memory (mb)
R_x	RAM in usage (mb)
N_{CPU}	Number of CPU cores
N_{CPUx}	Number of CPU cores in use
N_{Rx}	Network inbound traffic (Kbps)
N_{Tx}	Network outbound traffic (Kbps)
P_c	Power consumed by host (Watts)
T_{CPU1}	CPU 1 temperature (℃)
T_{CPU2}	CPU 2 temperature (℃)
fs_1	fan1 speed (RPM)
fs_2	fan2 speed (RPM)
fs_3	fan3 speed (RPM)
fs_4	fan4 speed (RPM)
T_{in}	Inlet temperature (℃)
N_{vm}	Number of VMs running on host

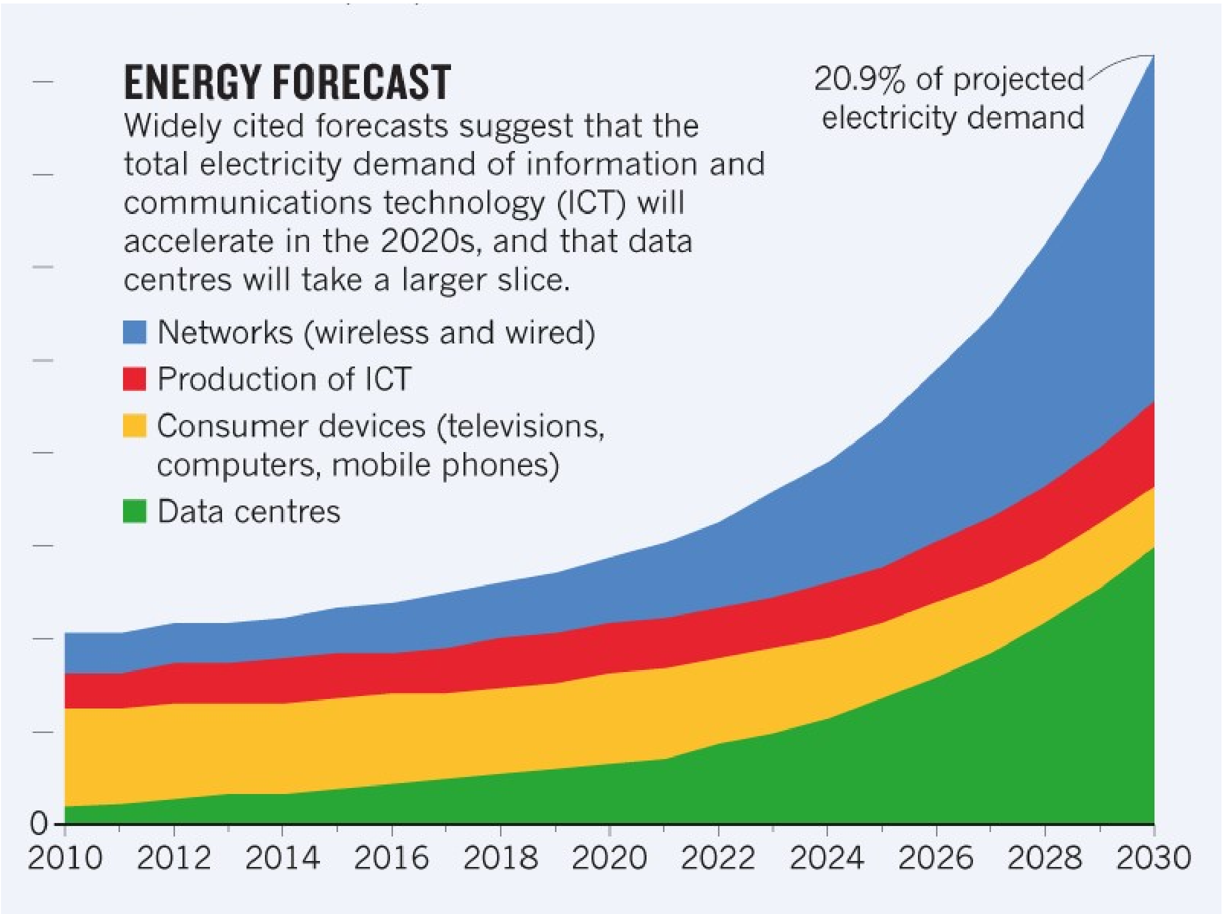


Data Centre Rack Layout

Source: S. Ilager, K.Ramamohanarao, and R. Buyya, Thermal Prediction for Efficient Energy Management of Clouds using Machine Learning, IEEE TPDS, Volume 32, No. 5, Pages: 1044-1056, USA, May 2021.
https://energyinnovation.org/2020/03/17/how-much-energy-do-data-centers-really-use/

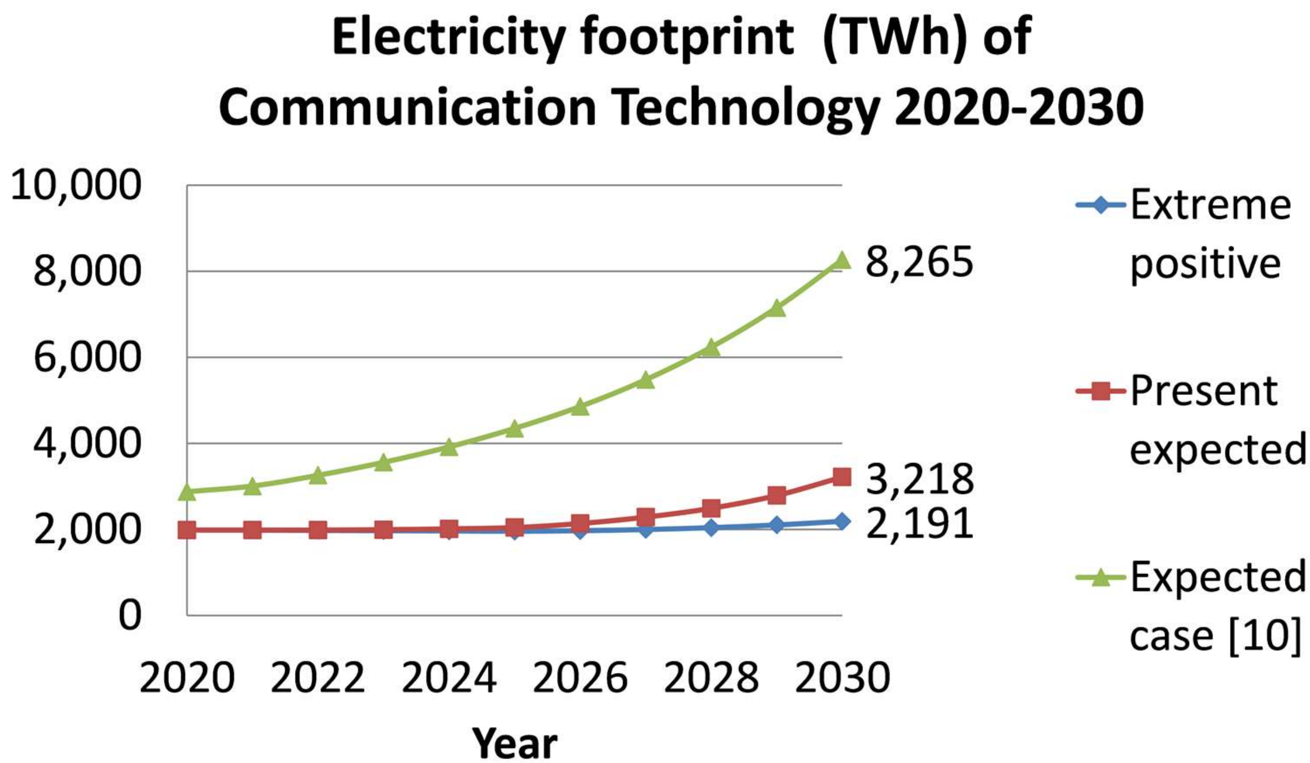
Slide: cortesy Shashi Ilager

Problem 2: Explosion of Energy Demands



Nicola Jones. How to stop data centres from gobbling up the world's electricity, Nature, 12. Sep, 2018.

Problem 2: ~~Explosion of~~ Increasing Energy Demands



Source: SG Andrae, Anders. "New perspectives on internet electricity use in 2030." Engineering and Applied Science Letter 3.2 (2020): 19-31.