Building intelligent sustainable Internet-based ecosystems

6 November 2023

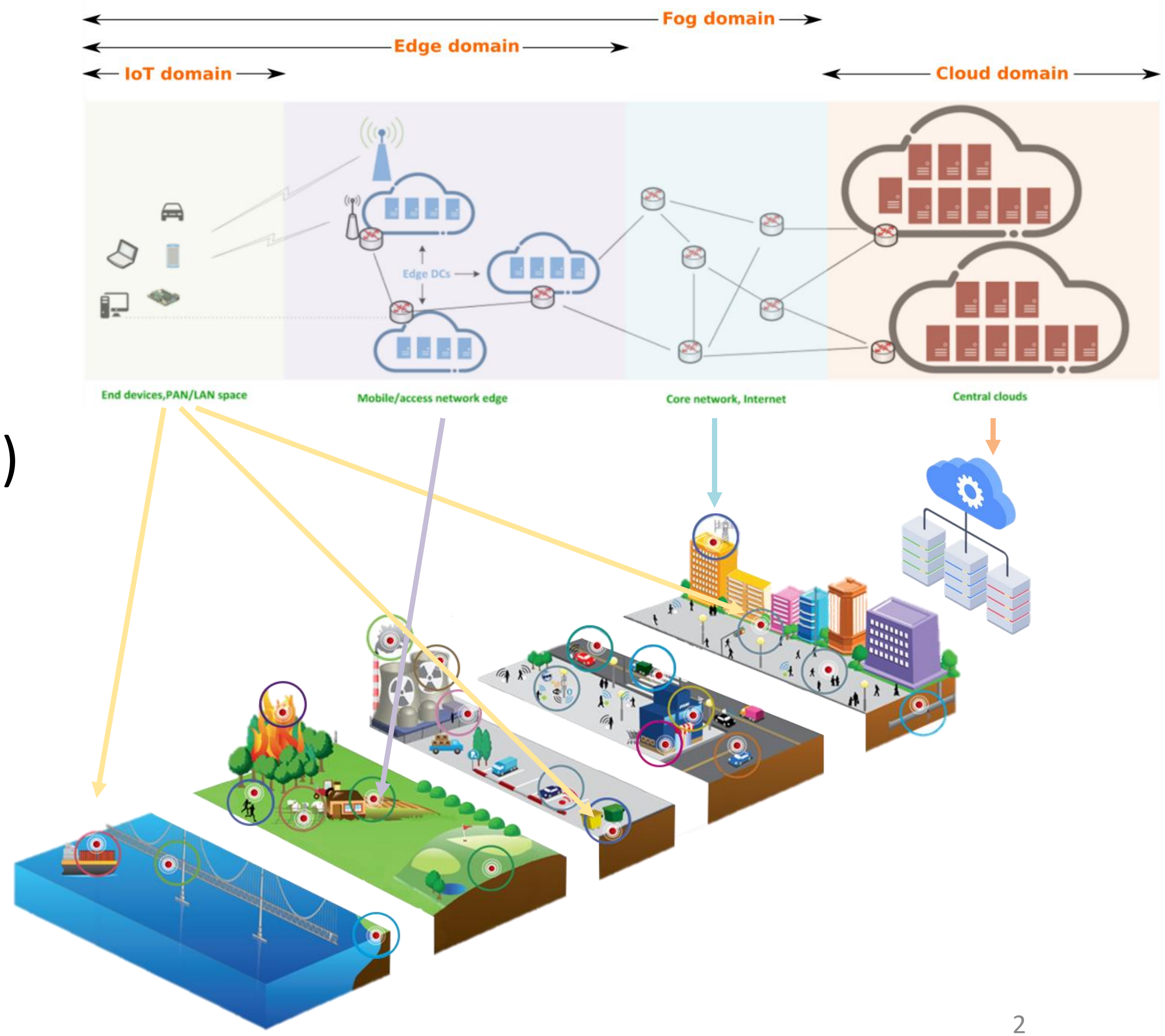
Schahram Dustdar

dsg.tuwien.ac.at

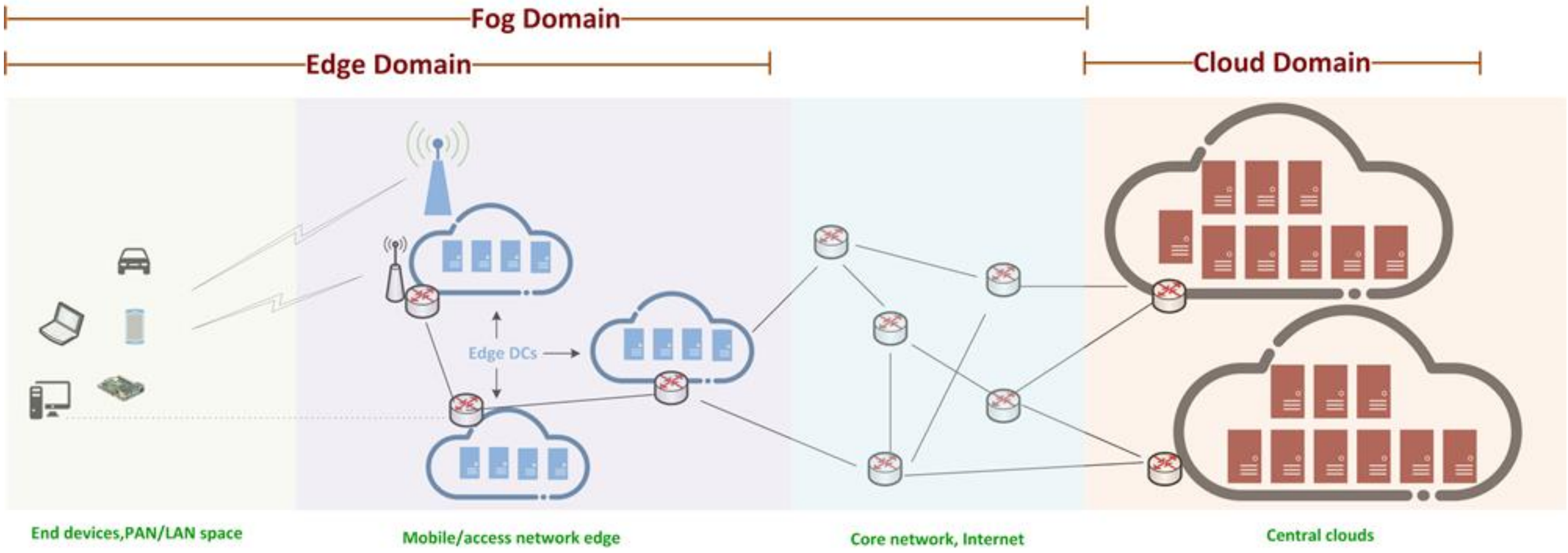
DISTRIBUTED SYSTEMS GROUP

Current State

- Distributed Systems are key to our society
- Underly our critical infrastructures and applications (Smart cities, Healthcare, Autonomous vehicles,…)
- Interconnectedness (fabric) of components (HW, SW, People) induces complexity
- We increasingly see fundamental issues we need to address



Distributed Compute Continuum: A high level view



Low reliability

Volatility

Mobility

(Mostly) Wireless connectivity

Small form factor

Battery constraints

Mobile, IoT, smart home, vehicles, …

User/Service provider controlled

Edge of the (mobile) network

Low latency to end device

Close to/collocated with 4G/5G base stations

General purpose compute infrastructure

Standards-based architectures & management/orchestration stacks

Telecom operator controlled

“Unlimited” compute/storage resources

Full spectrum of cloud services

High availability

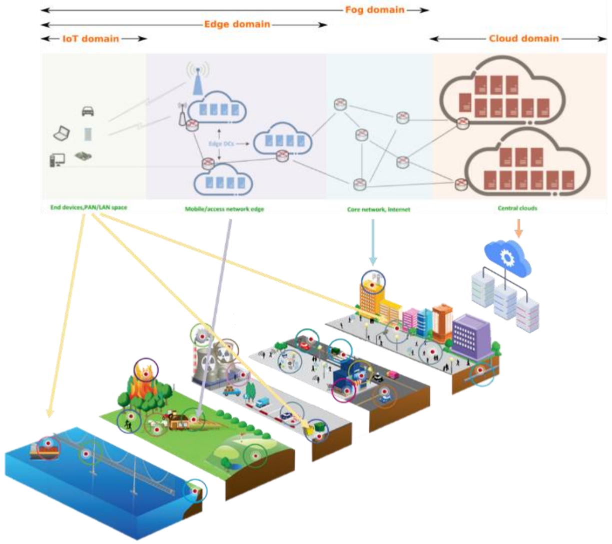
Lower cost

Higher latency vs. edge/fog

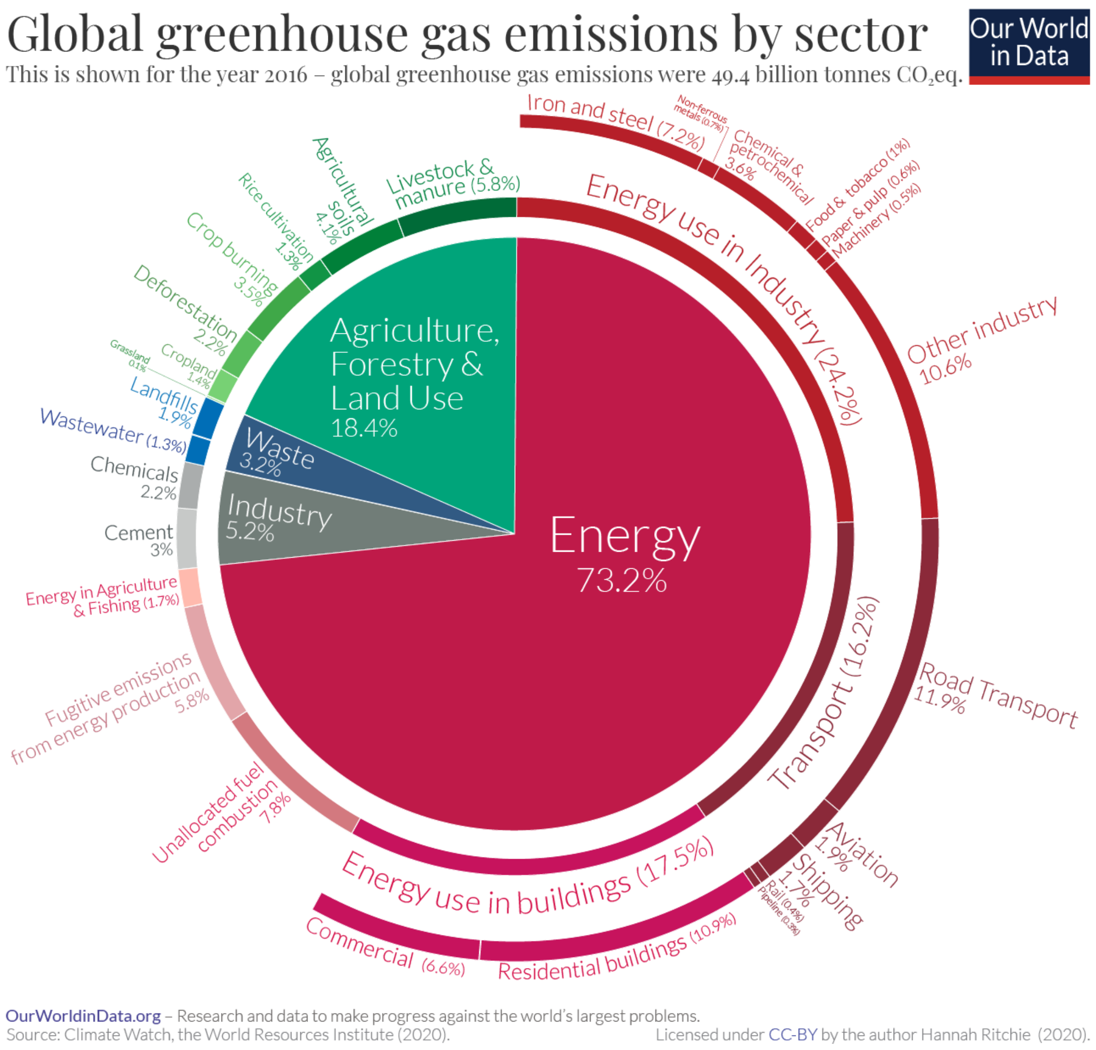
Cloud provider controlled

Distributed Computing Continuum Systems

Autonomous vehicles
eHealth
Industry 4.0
VR/AR
Resources (food, waste, energy…) management
…



- These applications will improve their current versions (imagine all vehicles driving to minimize consumption)
- BUT the distributed computing continuum will also require more energy.



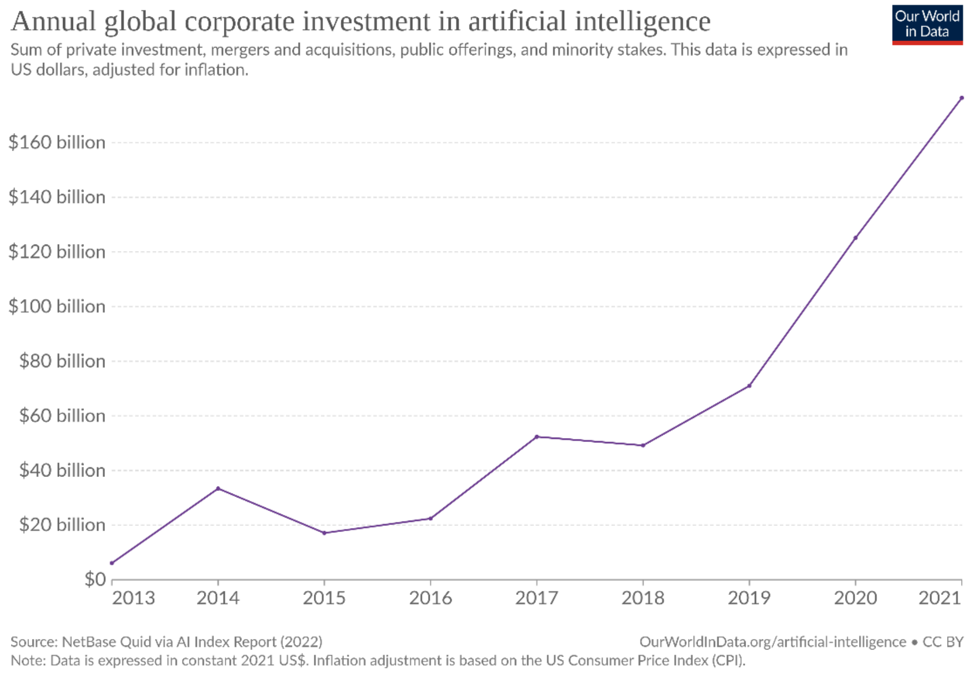
Hannah Ritchie, Max Roser and Pablo Rosado (2020) - "CO_2 and Greenhouse Gas Emissions". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/co2-and-greenhouse-gas-emissions' [Online Resource]

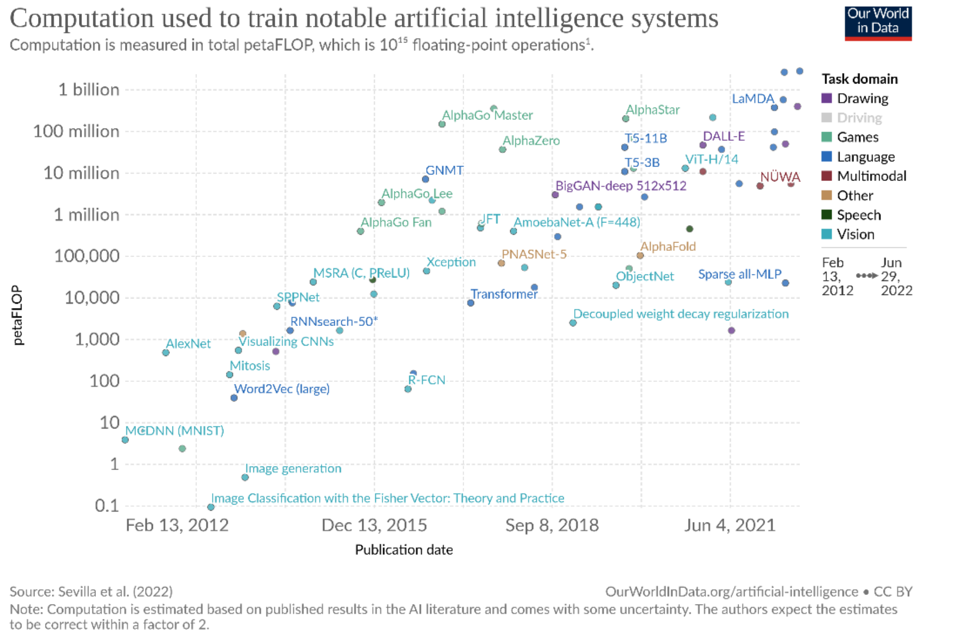
Computing energy demand growth

- Avg. human 5t CO2 per year [1]
- A Large Transformer model 285t CO2 per training (similar to a New York to San Francisco flight) [1]
- Train ChatGPT – 34 days in 1023 A100 GPUs (< 5 million $) [2]
- Run ChatGPT – 3 million $ per month [2]

[1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[2] “ChatGPT Statistics (2023) — Essential Facts and Figures,” Style Factory, Mar. 02, 2023. https://www.stylefactoryproductions.com/blog/chatgpt-statistics (accessed Mar. 06, 2023).



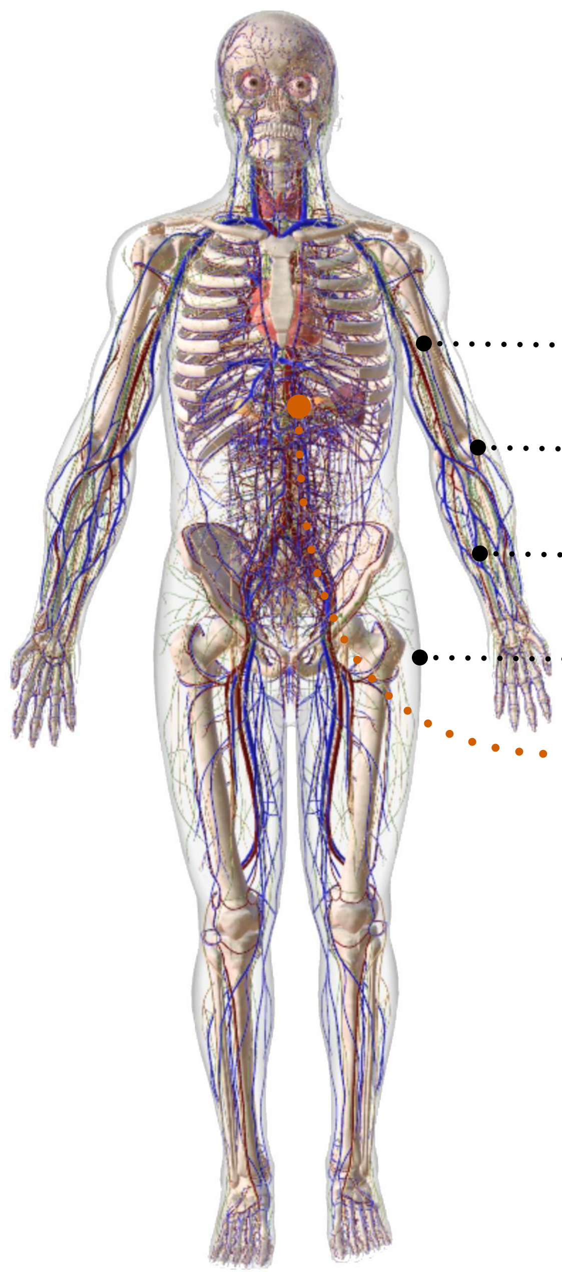


Hannah Ritchie, Max Roser and Pablo Rosado (2020) - "CO_2 and Greenhouse Gas Emissions". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/co2-and-greenhouse-gas-emissions' [Online Resource]

Towards Sustainable Distributed Computing Continuum Systems

- Energy awareness
  - Origin (green-renewal, battery, main distribution, …)
  - Usage (Computing, storing, data transfer, …)
  - Forecast (Consumption seasonality, computing peaks, …)

- Most of current research is currently on Energy-efficiency.
- Given a specific usage, new algorithms to reduce the recorded consumption are needed.
- Precise energy-awareness (specifically of the origin) is HARD to obtain.



Human Ecosystem

The human body is comprised of a series of complex systems, including:

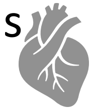
- Skeletal System
- Nervous System
- Cardiovascular System
- Lymphatic System
- Endocrine System

Nervous System, Cardiovascular System → Infrastructure Systems

→ Regulation Systems

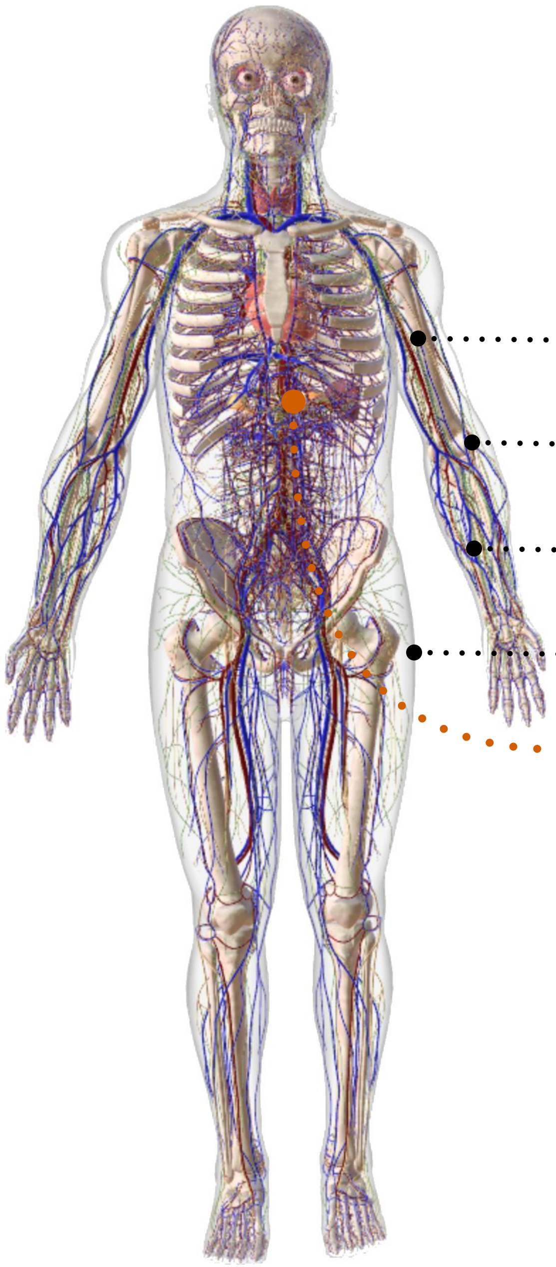
- Brain
- Spinal Cord
- Cranial Nerves
- Spinal Nerves

- Oxygen
- White Blood Cells
- Hormones
- Nutrients



Helping the body meet the demands (**40k neurons)**

Control Internal Environment, Memory and Learning (**86 billion neurons**)



Human Ecosystem

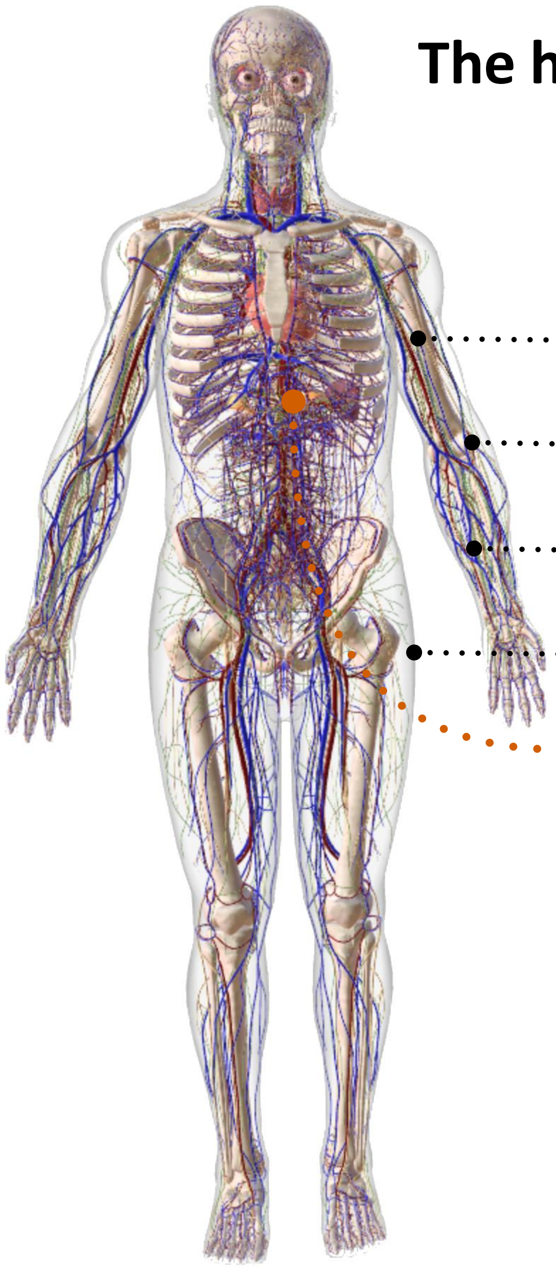
The human body is comprised of a series of complex systems, including:

- Skeletal System
- Nervous System
- Cardiovascular System
- Lymphatic System
- Endocrine System

→ Infrastructure Systems

→ Regulation Systems

The human body is comprised of a series of complex systems, including:

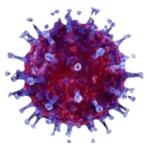


Human Ecosystem

- Skeletal System
- Nervous System
- Cardiovascular System
- Lymphatic System
- Endocrine System

Infrastructure Systems
- DeepSLOs
- Collaborative Learning
- Representation Learning

Regulation Systems
- Zero Trust

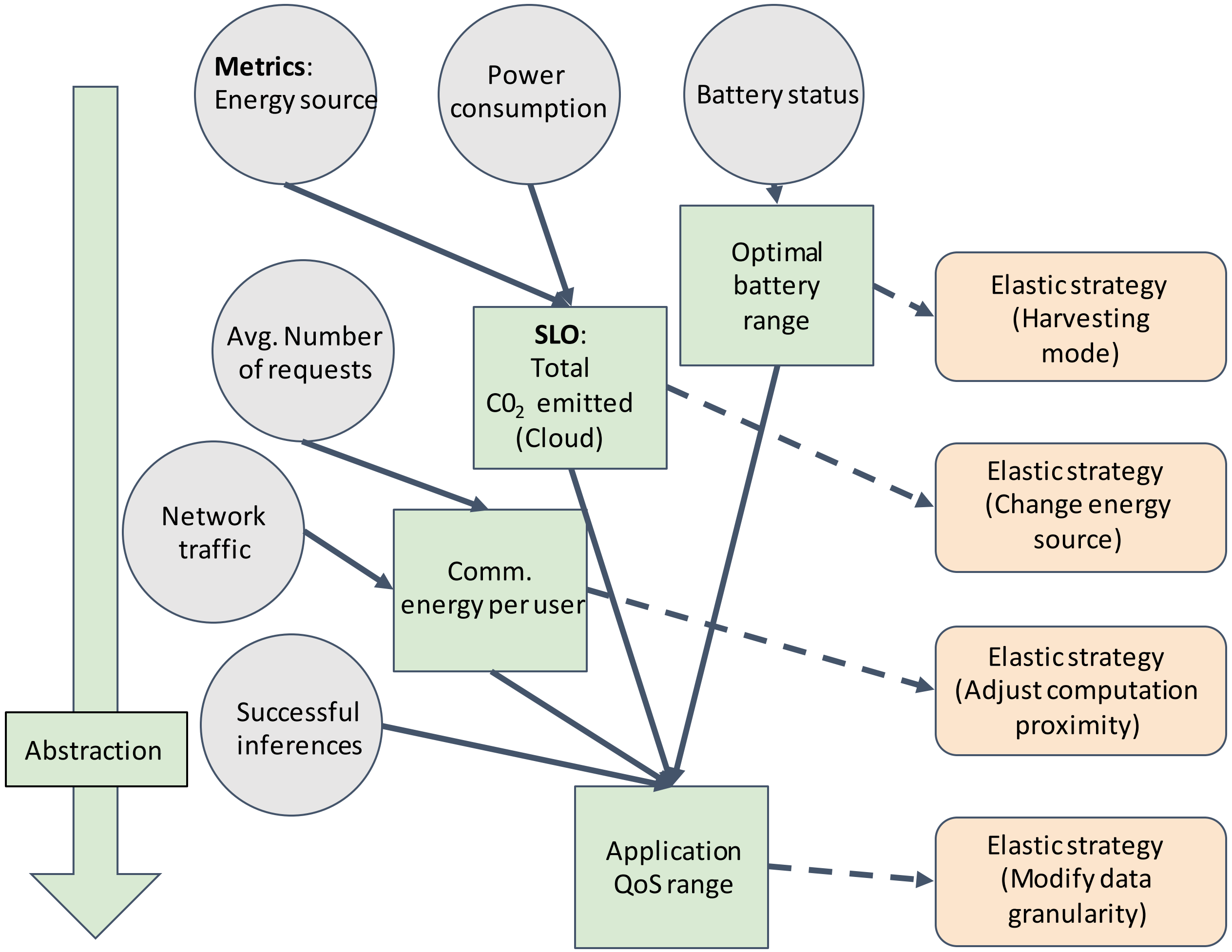


- Part of the immune system
- Protects your body against foreign invaders
- Control and coordinate your body's metabolism
- Response to injury, stress, and mood

Sustainable Distributed Computing Continuum Systems

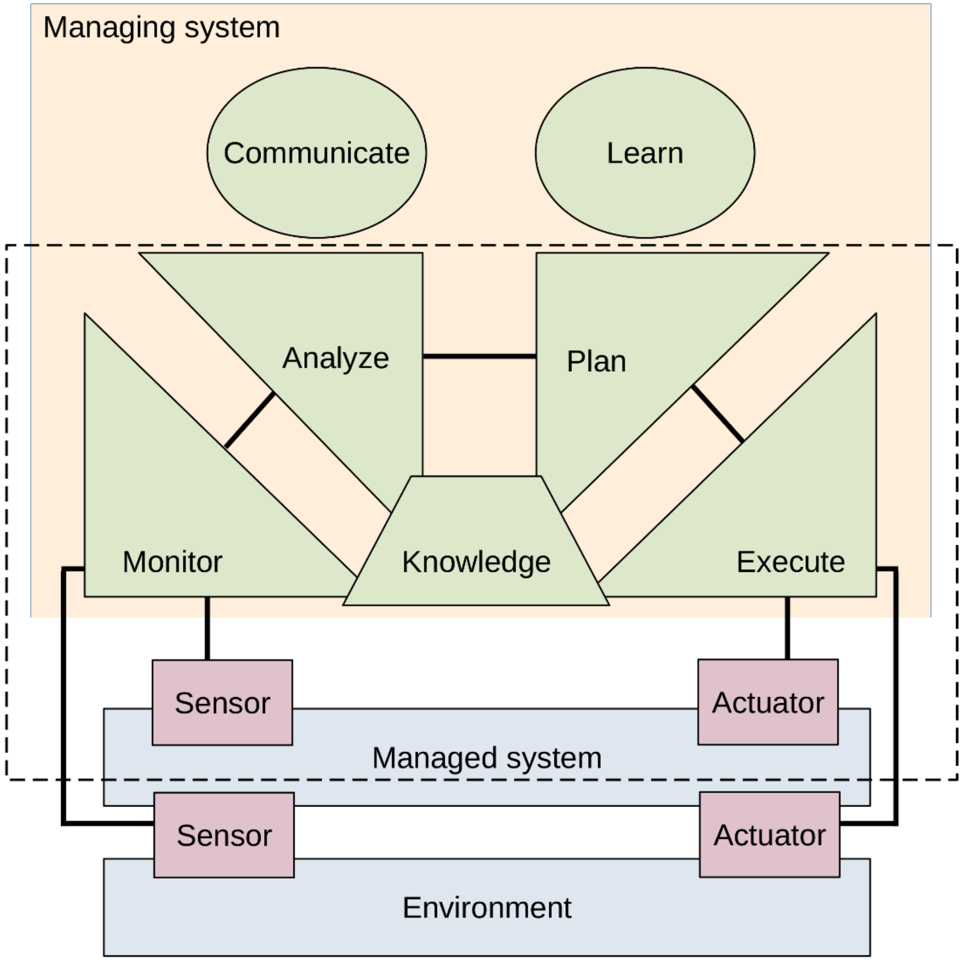
- Our vision aims at increasing the intelligence of the underlying computing infrastructure to provide the tools to handle energy-efficiency.

- We want to use hierarchically-structured set of SLOs (**DeepSLOs**) to acquire a layered energy profile of the system. This will allow to optimize energy efficiency at the stages which is more effective.



Sustainable Distributed Computing Continuum Systems

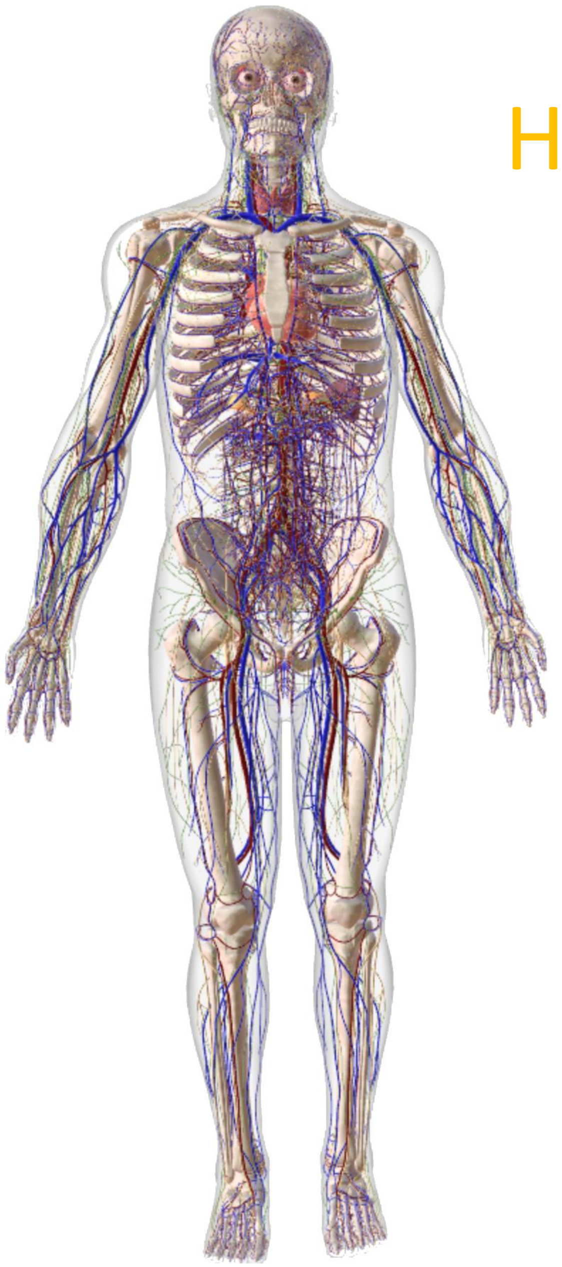
- Each SLO works following a MAPE-K (extended) schema.
- Higher abstracted SLOs can access policies from lower SLOs.
- Obtaining a loosely-coupled interaction between SLOs managing the system



Sustainable Distributed Computing Continuum Systems

- Is that enough?
- Does sustainability allow us to keep a continuous and steep increase on the computational requirements in our society?
- Similarly as it is done with CO_2, could computation have a limited usage?
- Can we develop systems with a fixed computational budget?

Homeostasis and Resilience in DCCS



Human Ecosystem

Nervous system

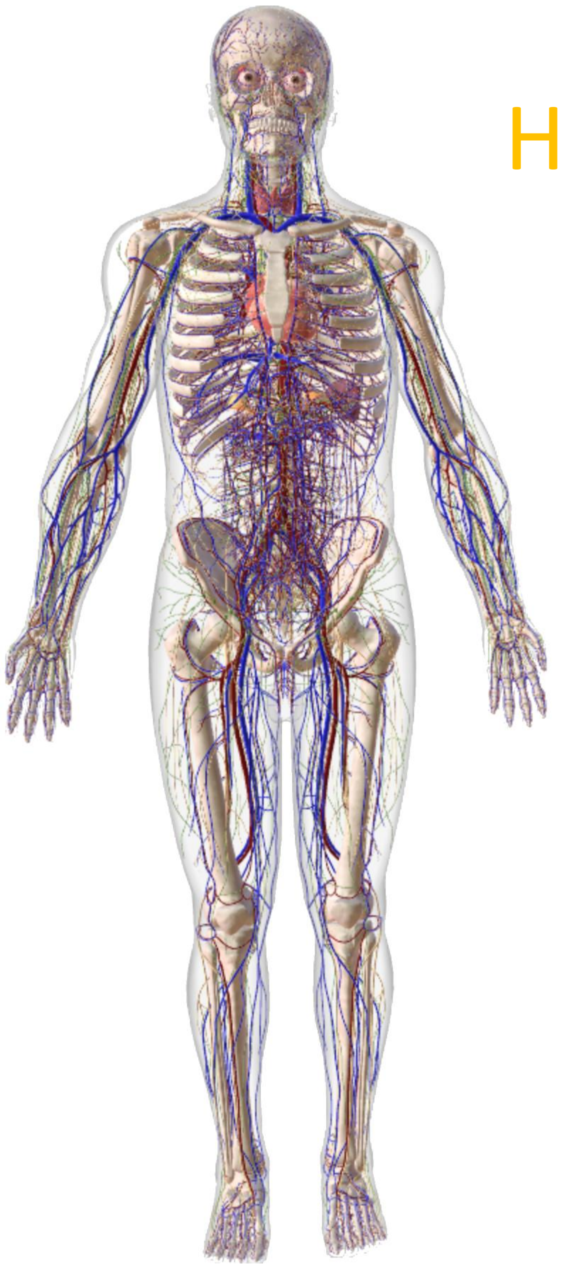
Human body self-regulates:

- Temperature
- Blood pressure
- …

Human body self-heals

Humans also learn how to maintain her/his needs satisfied.

Homeostasis and Resilience in DCCS



Human Ecosystem

Nervous system

Overall state - Top-bottom sensing.

From feeling *good-bad* to actual problem.

We also need this feature for DCCS due to their scale and interconnections.

Elasticity (Resilience)

(Physics) The property of returning to an initial form or state following deformation

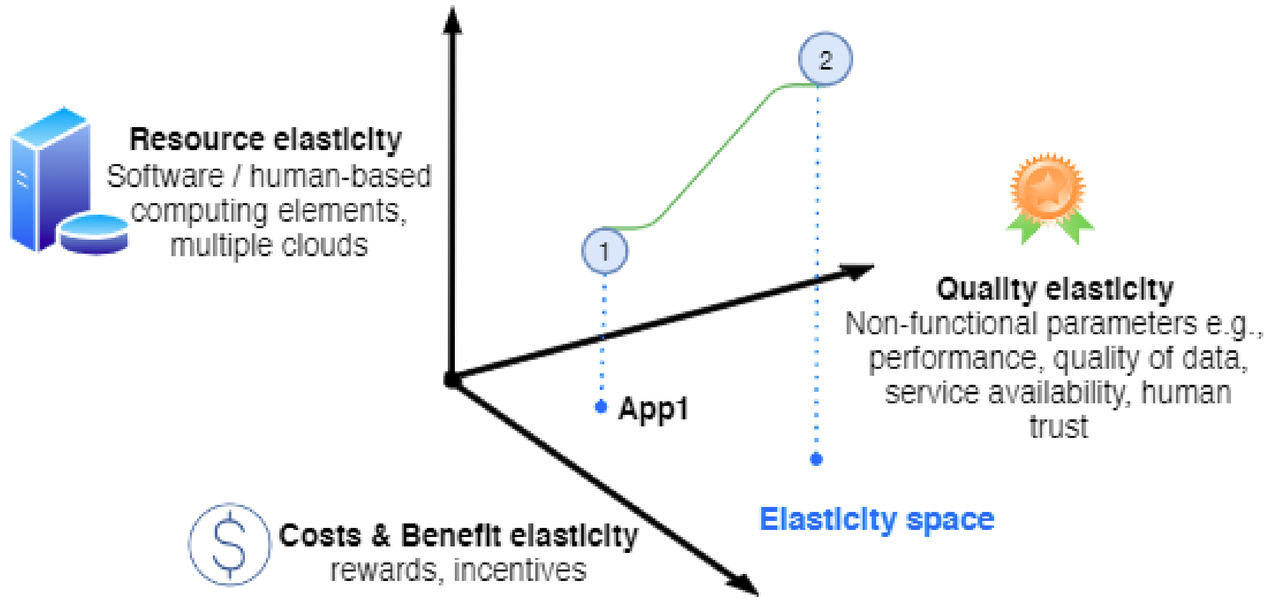
stretch when a force stresses them

e.g., ***acquire*** *new resources,* ***reduce*** *quality*

shrink when the stress is removed

e.g., ***release*** *resources,* ***increase*** *quality*

Elasticity > Scalability



Dustdar S., Guo Y., Satzger B., Truong H. (2012) [Principles of Elastic Processes](), IEEE Internet Computing, Volume: 16, [Issue: 6](), Nov.-Dec. 2012

High level elasticity control

#SYBL.CloudServiceLevel

Cons1: CONSTRAINT responseTime < 5 ms

Cons2: CONSTRAINT responseTime < 10 ms WHEN nbOfUsers > 10000

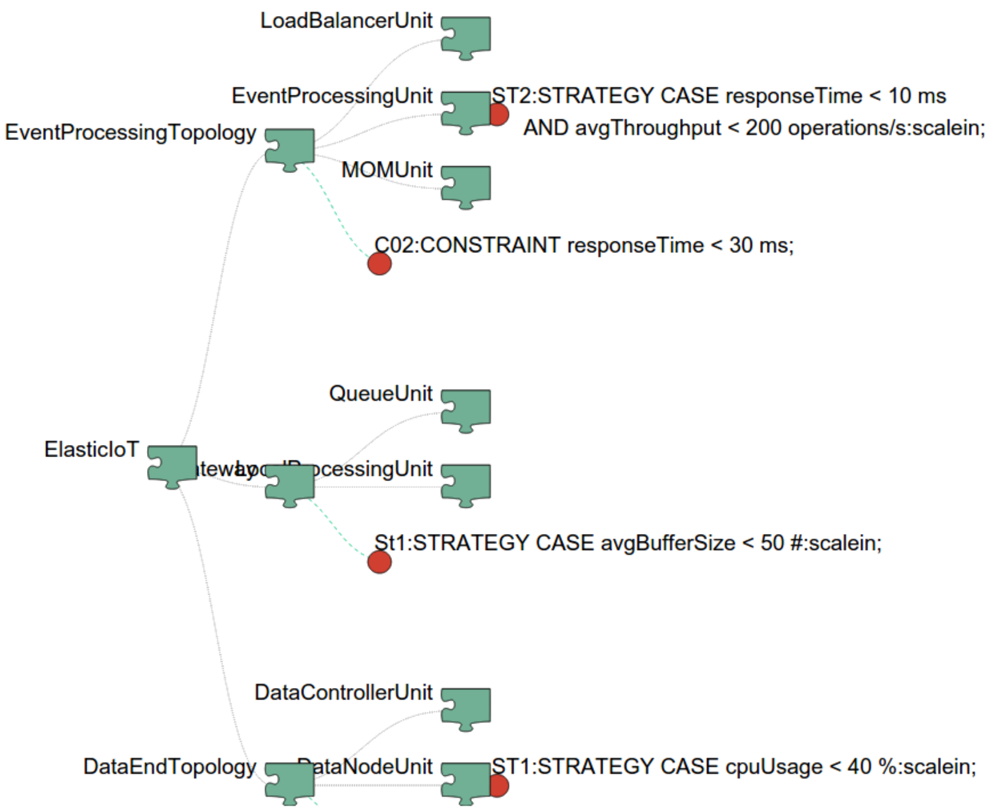
Str1: STRATEGY CASE fulfilled(Cons1) OR fulfilled(Cons2): minimize(cost)

#SYBL.ServiceUnitLevel

Str2: STRATEGY CASE ioCost < 3 Euro : maximize(dataFreshness)

#SYBL.CodeRegionLevel

Cons4: CONSTRAINT dataAccuracy>90% AND cost<4 Euro



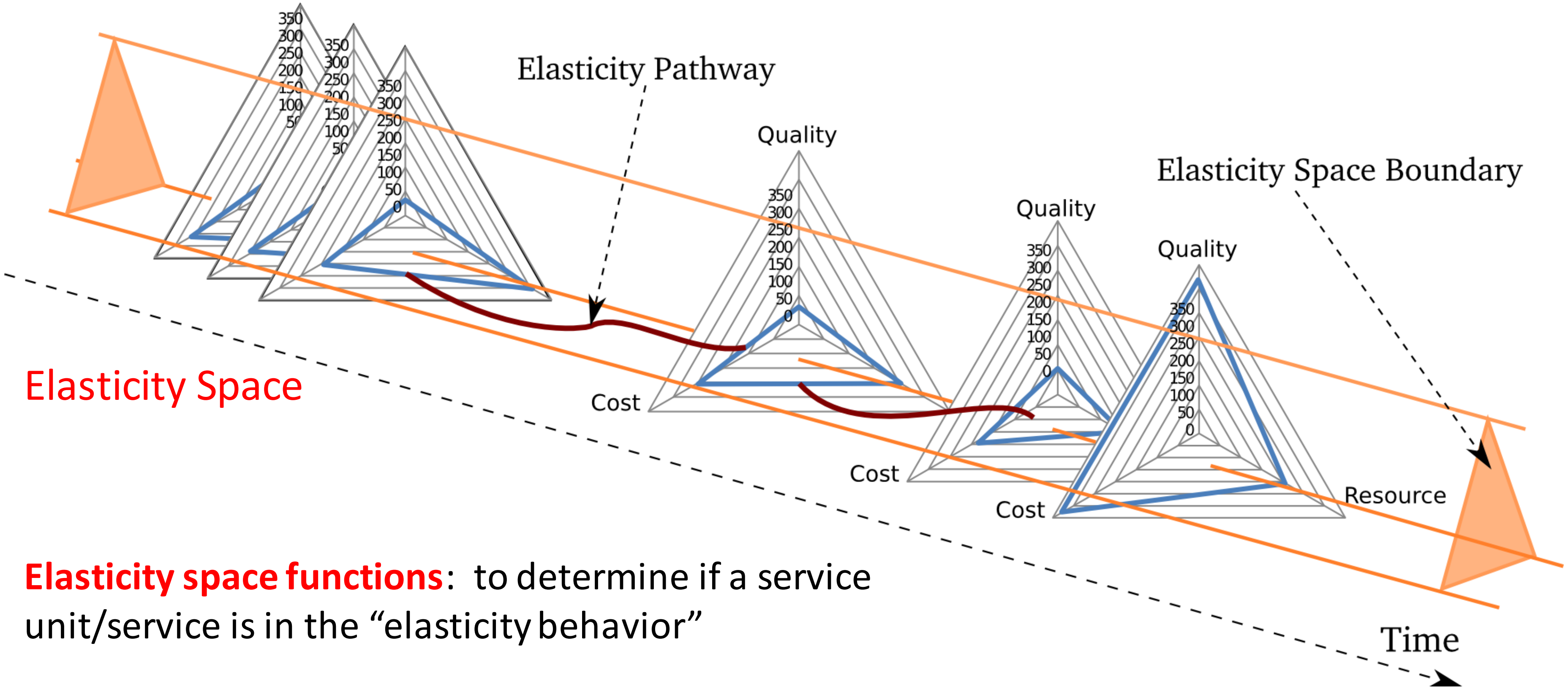
Georgiana Copil, Daniel Moldovan, Hong-Linh Truong, Schahram Dustdar, **"SYBL: an Extensible Language for Controlling Elasticity in Cloud Applications"**, 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), May 14-16, 2013, Delft, Netherlands

Copil G., Moldovan D., Truong H.-L., Dustdar S. (2016). **rSYBL: a Framework for Specifying and Controlling Cloud Services Elasticity**. *ACM Transactions on Internet Technology*

Elasticity Model for Edge & Cloud Services

Moldovan D., G. Copil,Truong H.-L., Dustdar S. (2013). **MELA: Monitoring and Analyzing Elasticity of Cloud Service. CloudCom 2013**

Elasticity Pathway functions: to characterize the elasticity behavior from a general/particular view



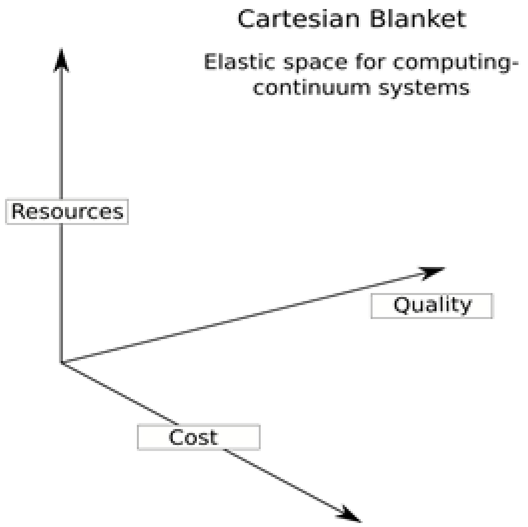
Elasticity Space

Elasticity space functions: to determine if a service unit/service is in the "elasticity behavior"

High-level state

Resources, Quality, Cost

- Highest-level description of system state from Cloud computing/elasticity work [1].
- DCCS have many different stakeholders with different interests, RQC can frame a common language.



Operational equilibrium

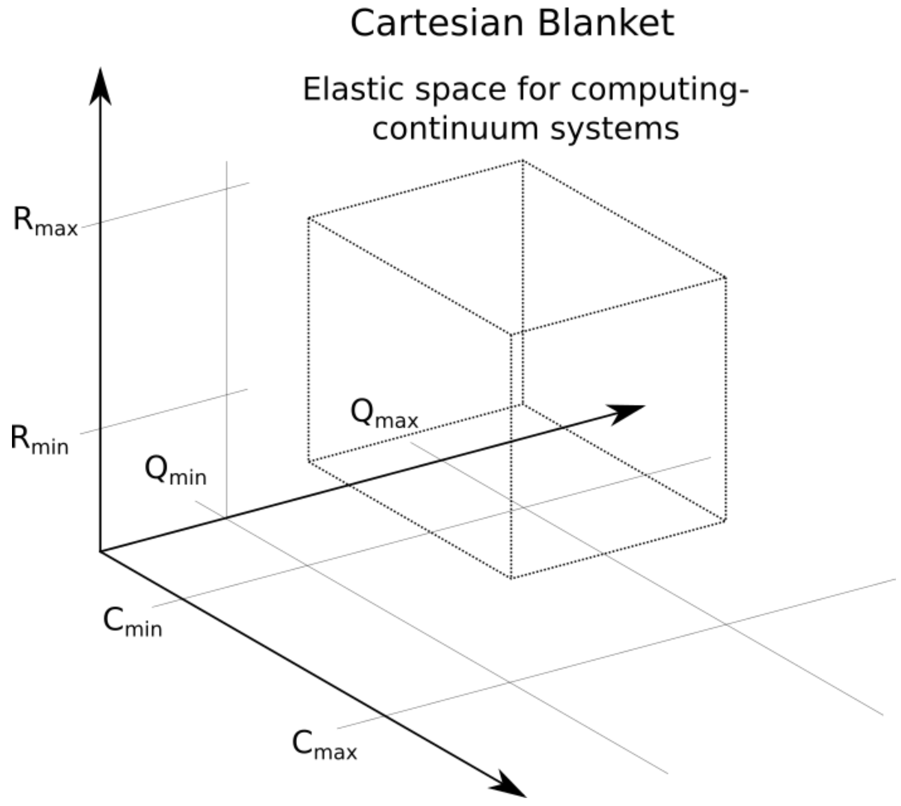
- Defined as an operational mode of the application, from the highest level state.
- Any system can have several operational equilibria, leading to different configurations of the underlying infrastructure

[1] S. Dustdar, Y. Guo, B. Satzger, and H. L. Truong, "Principles of elastic processes," *IEEE Internet Computing*, vol. 15, no. 5, pp. 66–71, Sep. 2011, doi: 10.1109/MIC.2011.121.

The Cartesian Blanket

Adapting elasticity in the continuum

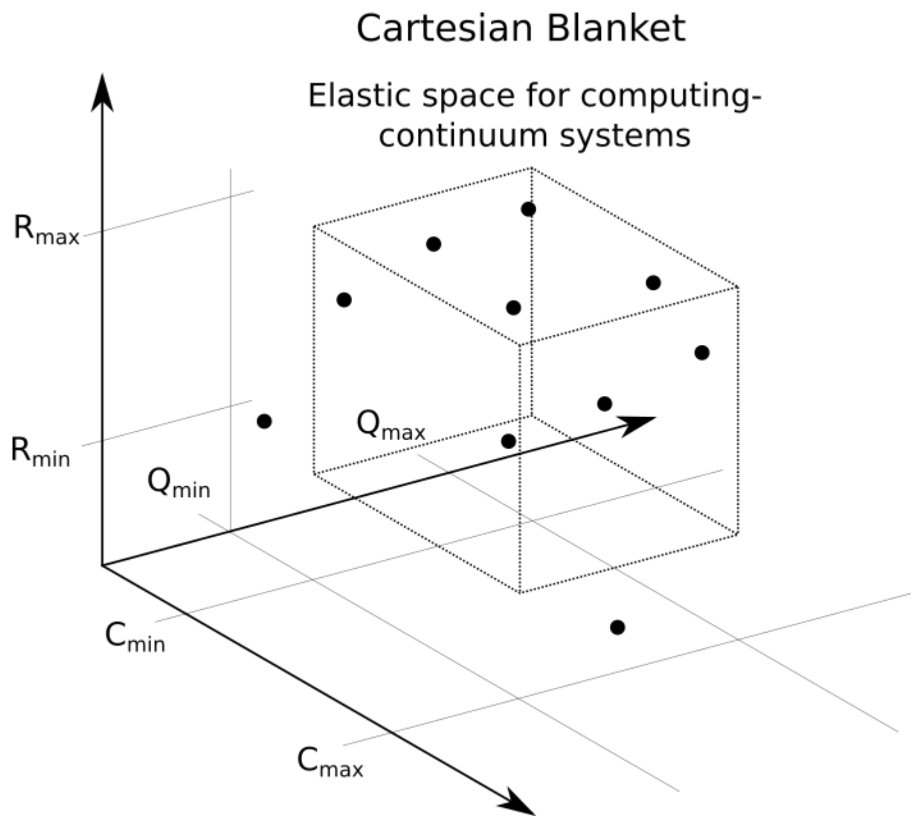
- System control based SLOs (Service Level Objectives)
- SLOs are represented as thresholds on the Cartesian space
- The system space is delimited within an hexahedron.
    - There is minimum and maximum value for each variable



The Cartesian Blanket

Adapting elasticity in the continuum

- The space is constraint to the actual infrastructure characteristics; not homogenous.
- The infrastructure is represented as points, not unlimited.
- The only valid infrastructure is the one **inside** the hexahedron.



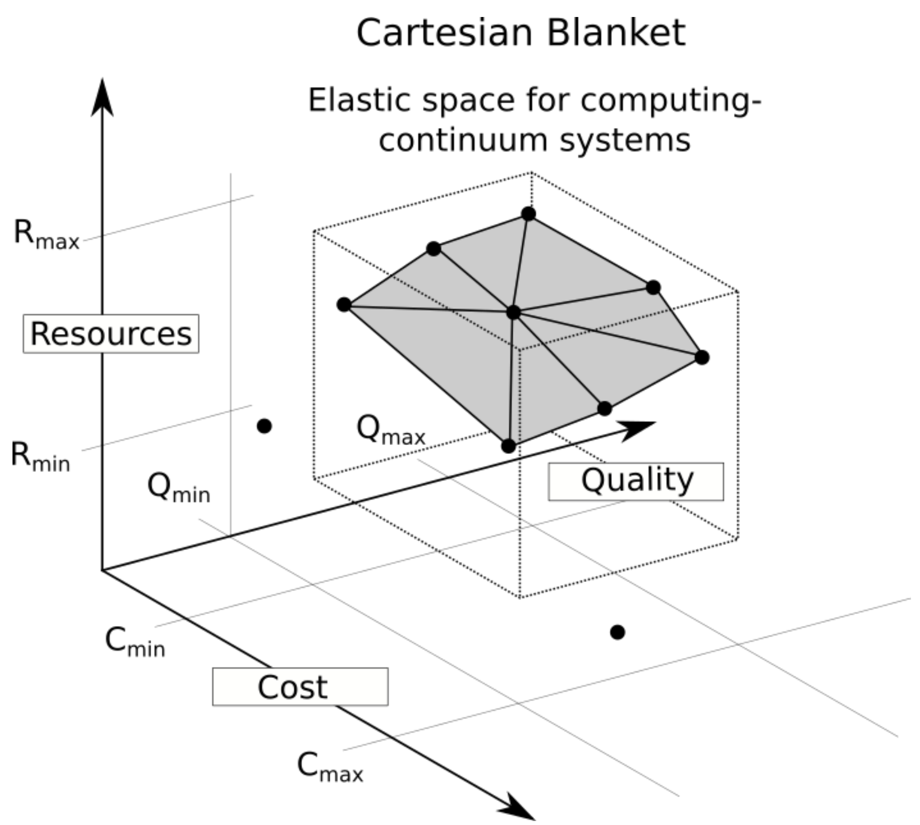
The Cartesian Blanket

Adapting elasticity in the continuum

- The system space possible configurations can be visualized as a stretched blanket over the infrastructure points.
  - Assuming linear interpolation on the space between the infrastructure components.

- Now we have the system represented, but

How can this representation help on the design and management of the distributed computing continuum systems?



Markov Blanket

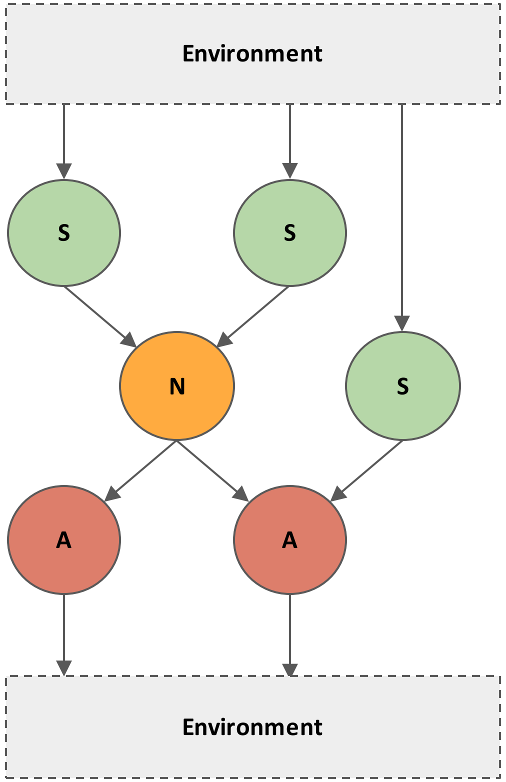
Statistical perspective [1]

The Markov Blanket provides conditional independence to its central variable. Hence, its central variable can be inferred only by the values of its Markov Blanket.

Ontological perspective [2]

Separates a thing from all its environment due to conditional independence. Defines 4 types of nodes:

- The internal node (N): the thing.
- The external nodes (E): The environment.
- The Markov Blanket states (S,A):
  - The sensory nodes (S): Receive input from the E and act on N.
  - The action nodes (A): Receive input from N and act on E.



[1] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc.

[2] K. J. Friston, · Klaas, E. Stephan, and · K E Stephan, "Free-energy and the brain," *Synthese 2007 159:3*, vol. 159, no. 3, pp. 417–458, Sep. 2007, doi: 10.1007/S11229-007-9237-Y.

Markovian Blanket for DCCS

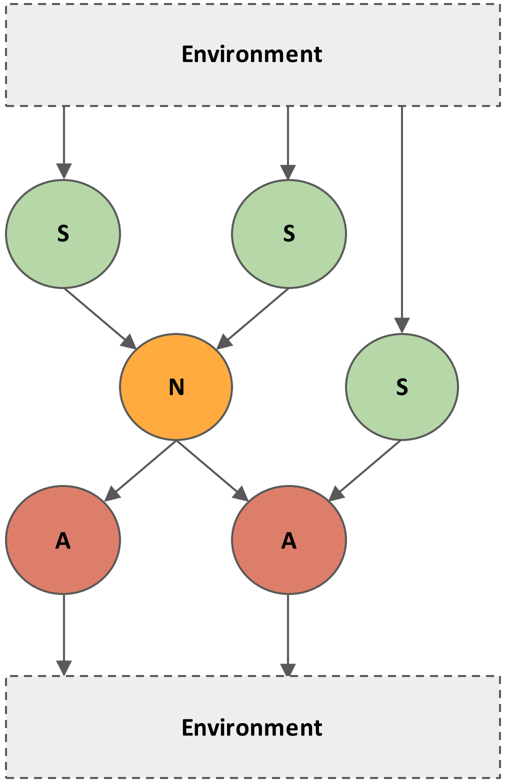
We aim to define DCCS based on the Markov Blanket abstraction with different granularities due to its nesting capacity.

Coarsest granularity:

- Central nodes are Resources-Quality-Cost. Highest abstraction level SLOs are influencing them.
- Overall configuration options (operational equilibriums) are defined to adapt the system at that level.

Finest granularity:

- A single SLO, influenced by a subset of metrics from the infrastructure.
- Affects a subset of action states able to precisely affect infrastructure state.



Markovian Blanket for DCCS

We aim to define DCCS based on the Markov Blanket abstraction with different granularities due to its nesting capacity. From an application perspective

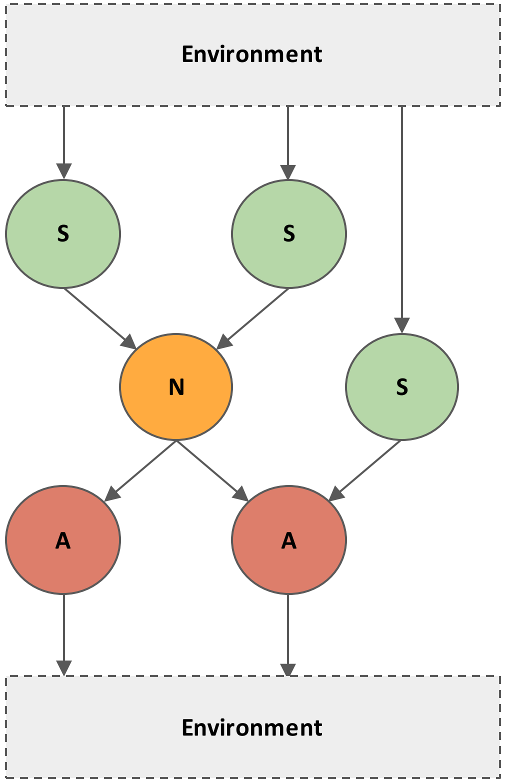
Coarsest granularity:

- The entire application, i.e. managing all mobility of autonomous vehicles in a smart city

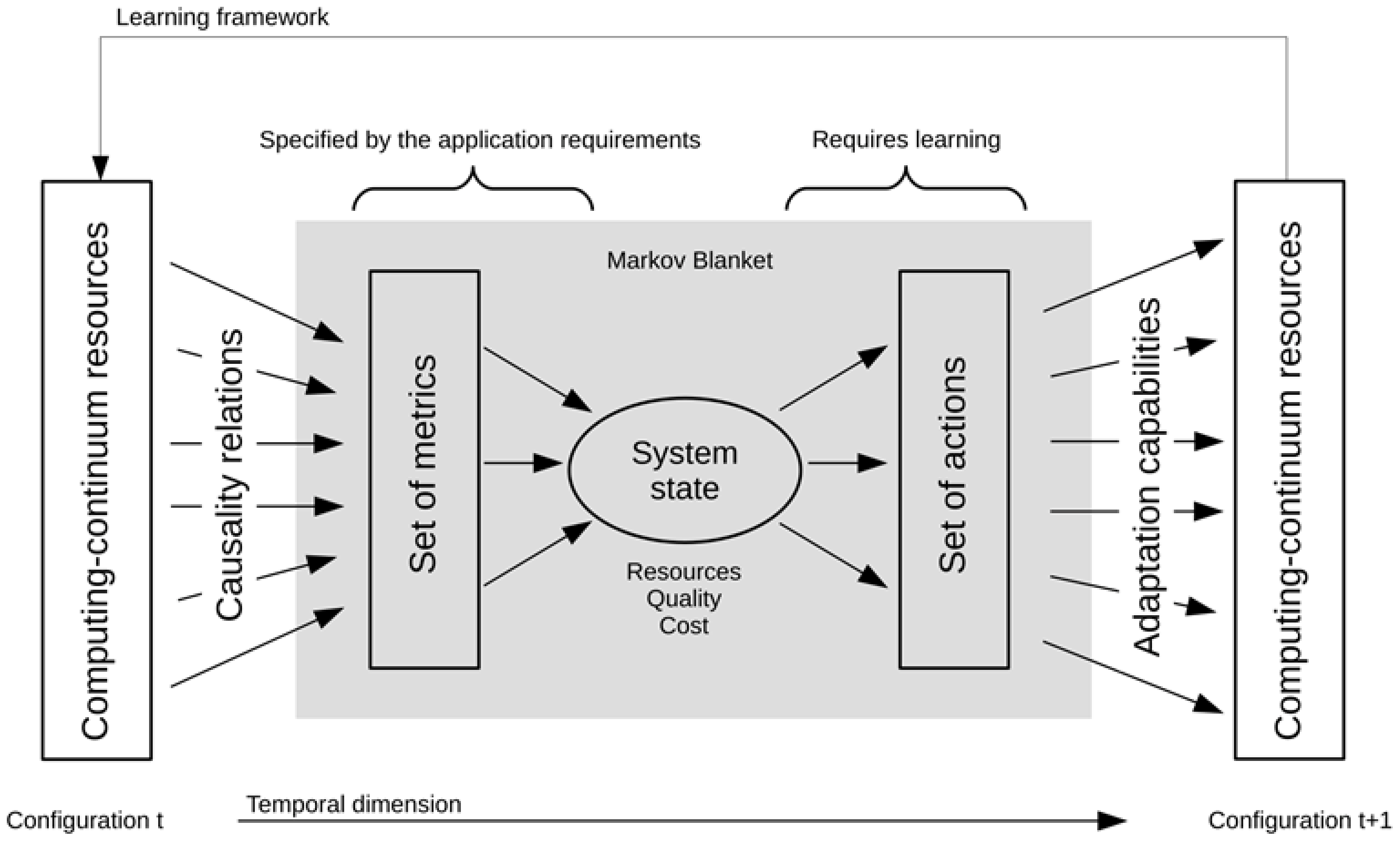
Finest granularity:

- A service to assess traffic congestion.

Nested capacity can be cast as a causality filter to focus on the most relevant autonomic component.



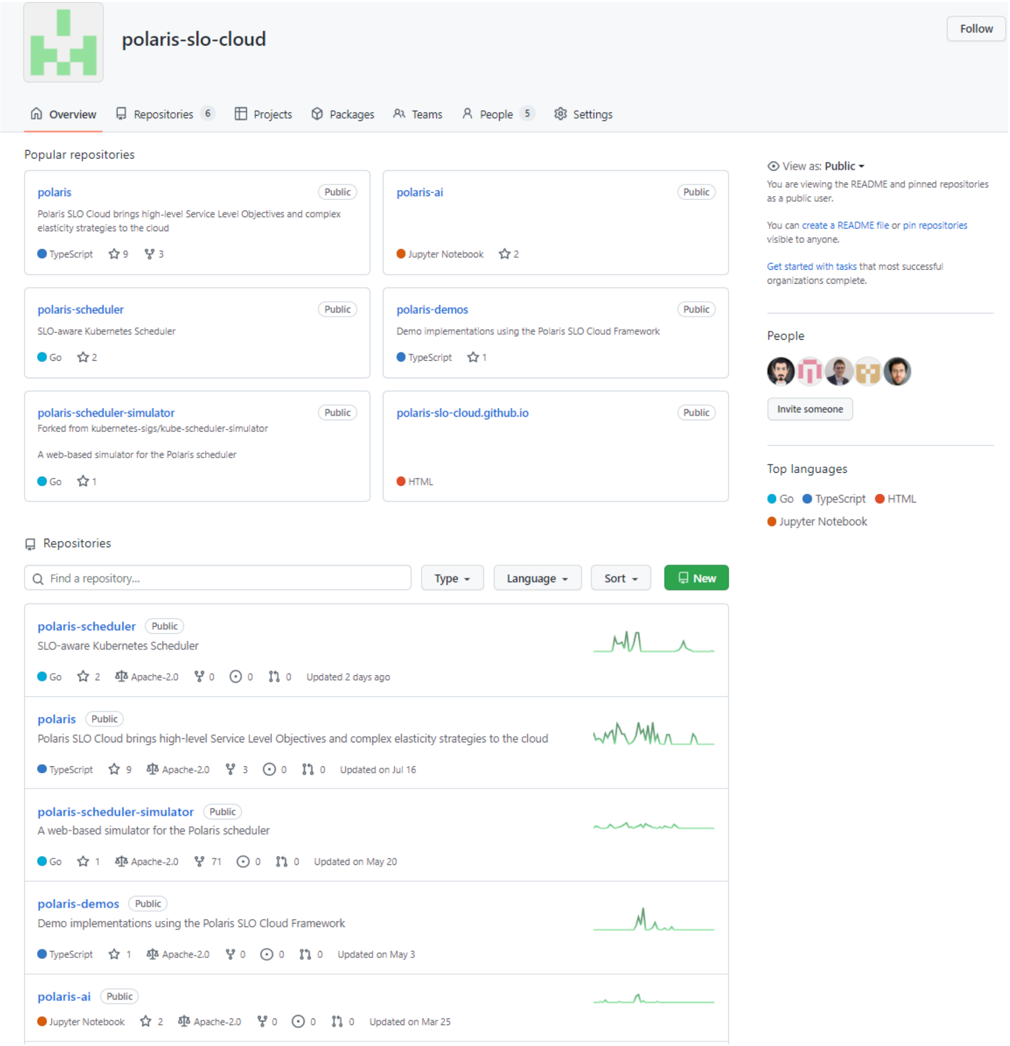
Markovian Blanket for DCCS – Big Picture



SLO Management with Polaris SLO Cloud

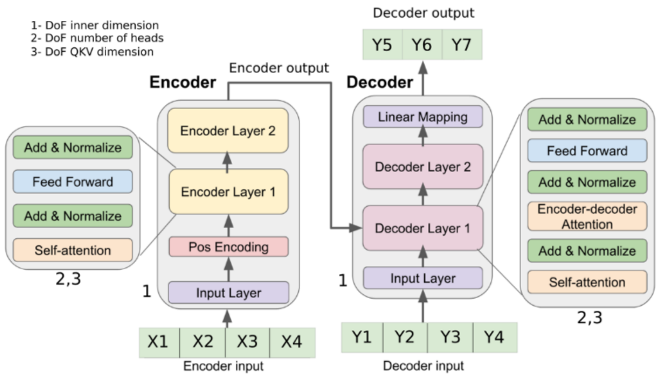
https://polaris-slo-cloud.github.io/polaris/

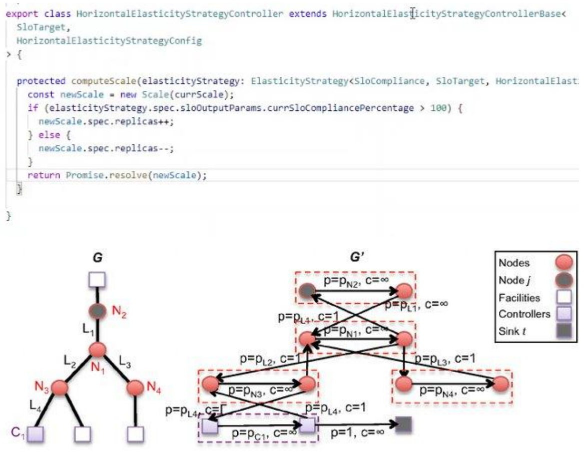
- Management of SLOs in Edge-Cloud native systems
- Project between TU Wien/DSG and Futurewei USA
- Fully Open-Source project carried by Linux Foundation since Jan 2021
- Core concept -> Polaris SLO Controllers (custom Kubernetes controllers but not limited to), enabling:
    - Specifying custom SLOs (based on TypeScript)
    - Monitoring of SLOs
      (2 models for predictive based on LSTM enabling high-level SLOs)
    - Resource monitoring
    - Enforcing SLOs during at runtime (Elasticity control strategies e.g., for modifying topologies etc.)



Polaris Controllers Very High-Level Overview

Monitoring → SLO → Control Elasticity Strategies





Nastic, S., Morichetta, A., Pusztai, T., Dustdar, S., Ding, X., Vij, D. and Xiong, Y., 2020. SLOC: Service level objectives for next generation cloud computing. IEEE Internet Computing, 24(3), pp.39-50.

Pusztai, T., Morichetta, A., Pujol, V.C., Dustdar, S., Nastic, S., Ding, X., Vij, D. and Xiong, Y., 2021, September. SLO script: A novel language for implementing complex cloud-native elasticity-driven SLOs. In *2021 IEEE International Conference on Web Services (ICWS)* (pp. 21-31). IEEE.

Pusztai, T., Morichetta, A., Pujol, V.C., Dustdar, S., Nastic, S., Ding, X., Vij, D. and Xiong, Y., 2021, September. A novel middleware for efficiently implementing complex cloud-native SLOs. In 2021 IEEE 14th International Conference on Cloud Computing (CLOUD) (pp. 410-420). IEEE.

Research line - Model

Markovian models

- Markov blanket (DAG)
- Markov fields (non directed graphs)
- Markov chains

Deep neural networks

- Federated learning
- Graph neural networks

Agent based

- Active inference
- Reinforcement learning

- How to deal with a multimodal environment?

Incorporate data from video sources, results from video processing units, quality of the predictions, overall system cost…

- How to model relations?

The shortage of computing power on an edge device will affect overall control system, but how much?

- How to treat abstraction?

Include concepts of cost or quality along with basic infrastructure metrics, i.e. number of drivers detected at the phone and GPU usage in the same framework.

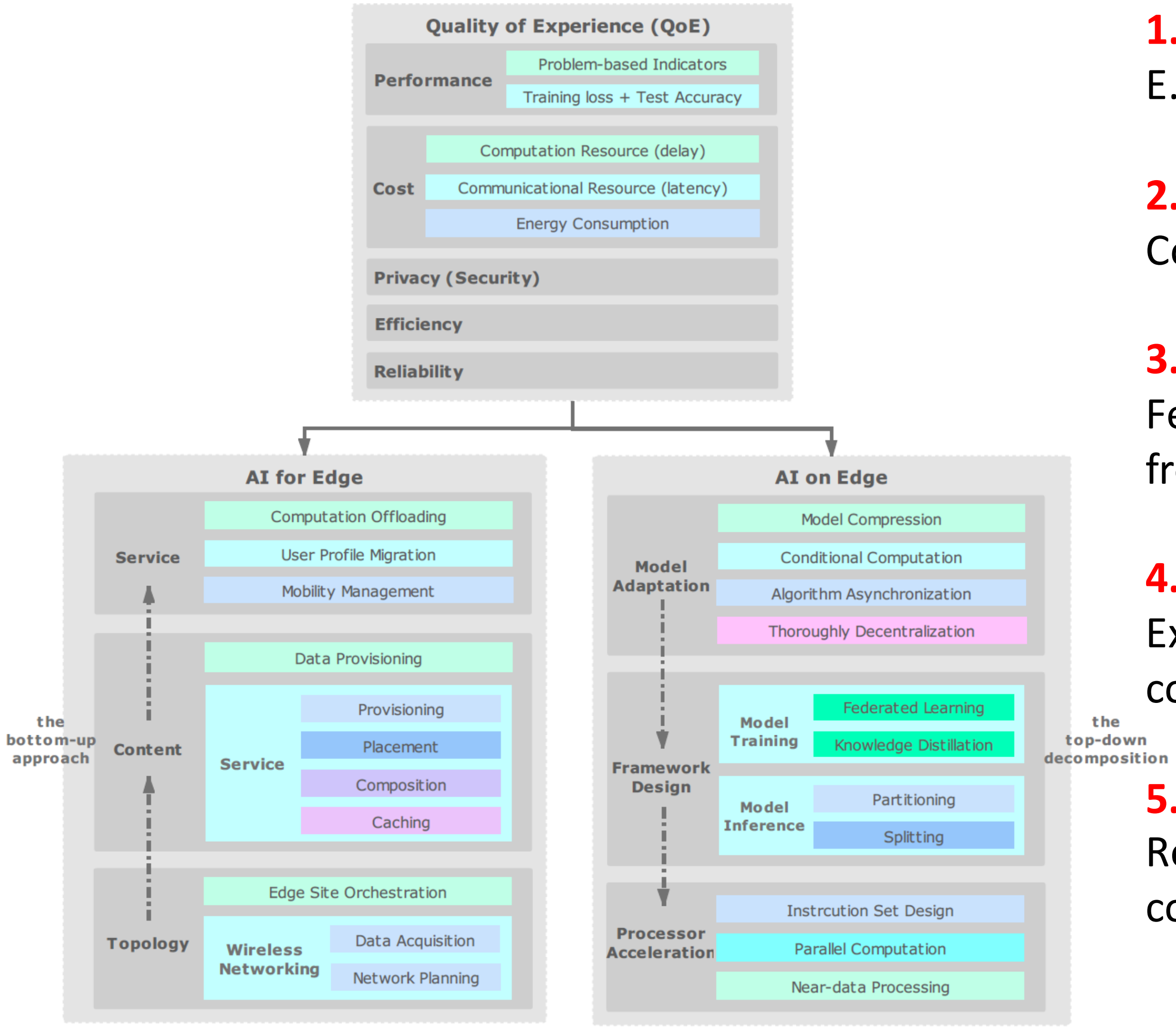
- How to obtain enough data?

Large, hyper-distributed and open systems. How to know the system is accurate?

- And many more… How to deal with IID data? How to tackle uncertainty?

Research Roadmap – Quality of Experience

Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence, *IEEE Internet of Things Journal*, Vol.7, Issue 8, pp. 7457-7469



1. Performance
E.g., the ratio of computation offloading

2. Cost
Computation | Communication | Energy consumption costs

3. Privacy & Security
Federated learning, i.e., aggregating local machines models from distributed edge devices

4. Efficiency
Excellent performance with low overhead, e.g., model compression, conditional computation

5. Reliability
Relates to model upload and download and wireless network congestion

AI *for* Edge

1. Topology

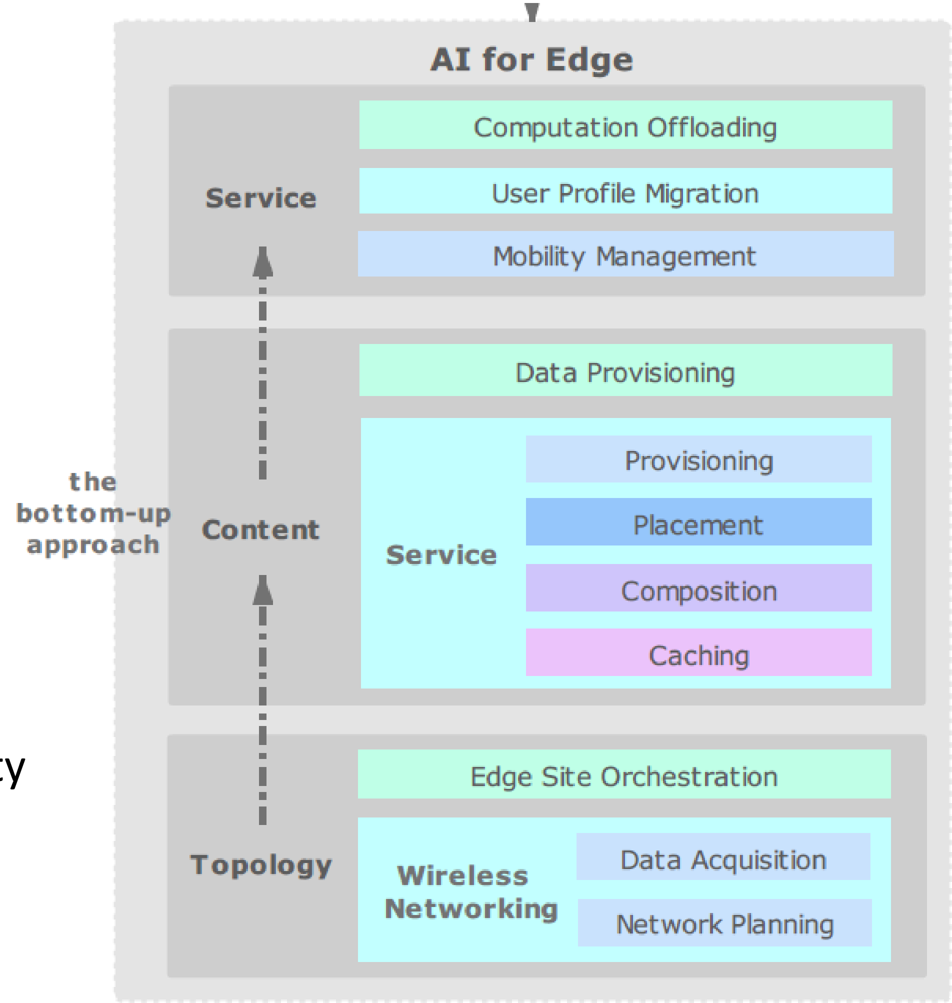
- Edge orchestration and coordination with small base stations
- Unmanned Aerial Vehicles (UAVs) and access points

2. Content

Lightweight service frameworks for QoS-aware services, e.g., on mobile devices

3. Service

Computation offloading, User profile migration and mobility management



Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence, *IEEE Internet of Things Journal*, Volume 7, Issue 8, pp. 7457-7469

Grand Challenges – AI *for* Edge

- **Model Establishment – restraining the optimization model**
  - Stochastic Gradient Descent (SGD)
  - MBGD (Mini-Batch Gradient Descent)

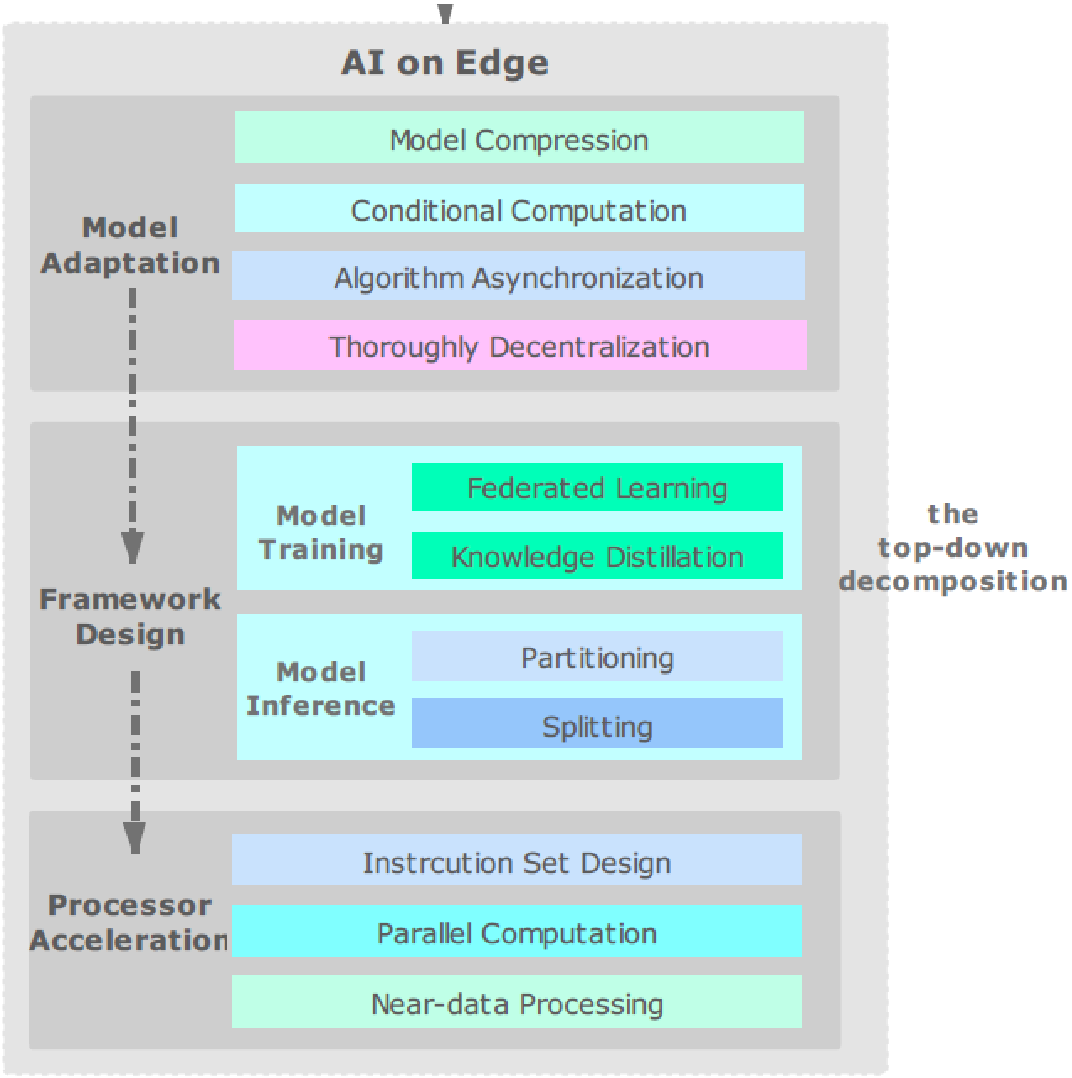
- **Algorithm Development**
  - Selection of *which* edge device should be responsible for deployment and execution in an online manner
  - SOTA formulates combinatorial and NP-hard optimization problems with high computational complexity

- **Trade-off between optimality and efficiency**
  - Consider resource constraint devices

Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence, *IEEE Internet of Things Journal*, Volume 7, Issue 8, pp. 7457-7469

AI *on* Edge

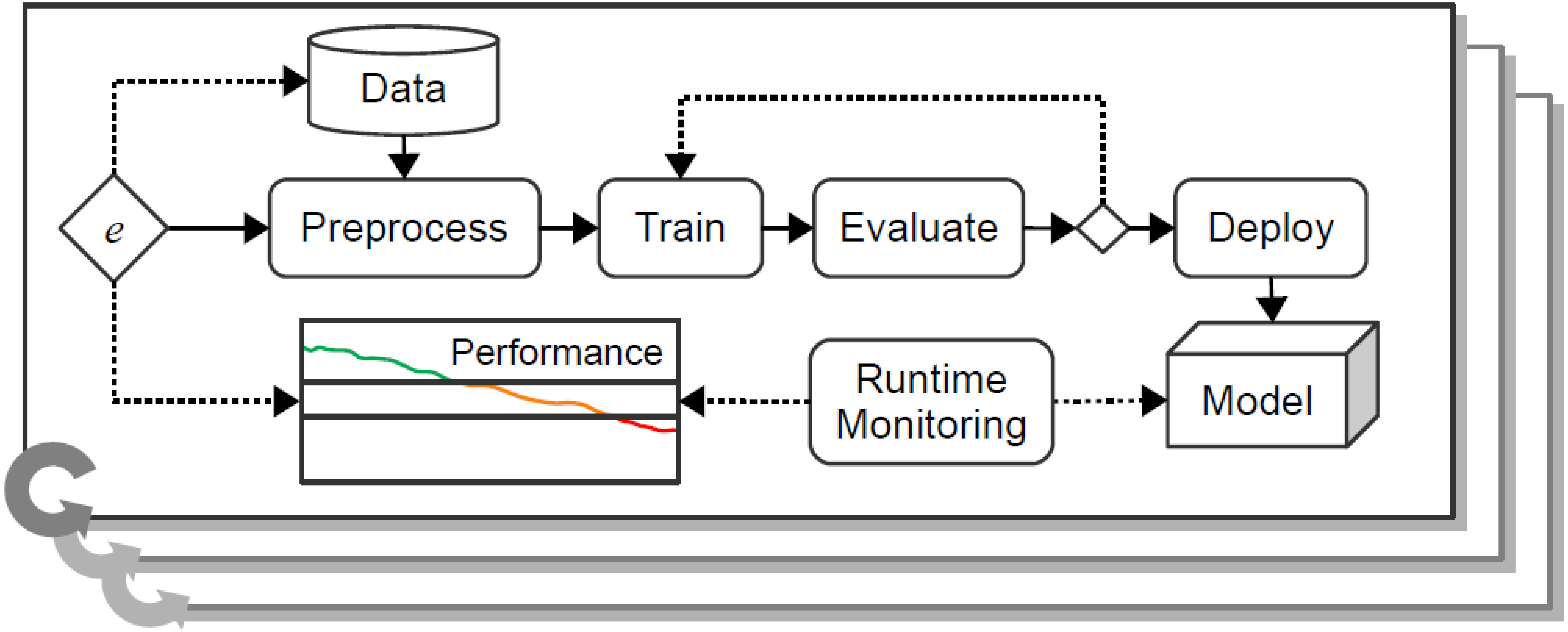
- **Data Availability**
  - Challenge of lack of availability and usability of raw training data for model training and inference
  - Bias of raw data from various end user/mobile devices
- **Model Selection**
  - SOTA requires selection of need-to-be trained AI models has challenges
  - Threshold of learning accuracy and scale of AI models for quick deployment and delivery
  - Selection of probe training frameworks and accelerator architectures under limited resources
- **Coordination Mechanisms**
  - Coordination between heterogeneous edge devices, cloud, and various middlewares and APIs



Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence, *IEEE Internet of Things Journal*, Volume 7, Issue 8, pp. 7457-7469

Managing the AI Lifecycle

AI lifecycle pipeline with a rule-based trigger *e* that monitors available data and runtime performance data to form an automated retraining loop



AI Operations Workflows – Edge to Cloud

	Data characteristics	Model characteristics	Enabling technologies	Example use cases
C2C	- Training data is centralized - Massive data sets	- Models are large - Huge number of inferencing requests need to be load balanced	- Scalable learning infrastructure [39] - Data warehousing	- Image search - Recommender systems
C2E	- Training data is centralized - Inferencing data may be sensitive	- Inferencing may need to happen in near-real time - Large number of model deployments - Models run on specialized hardware	- Model compression [42] - Latency/accuracy tradeoff [43] - Distributed inferencing [44] - Transfer learning [45]	- Surveillance systems - Self driving cars - Fieldwork assistants
E2C	- Training data is distributed - Training data may be sensitive	- Models can be centralized - Huge number of inferencing requests need to be load balanced	- Decentralized/federated learning [41]	- Volunteer computing - Novel Smart City use cases
E2E	- Training data is distributed - Training and inferencing data may be sensitive	- Inferencing may need to be near-real time	- Decentralized/federated learning - Distributed inferencing	- Industrial IoT (e.g., predictive maintenance) - Privacy-aware personal assistants - Novel IoT use cases

Rausch, T., Dustdar, S. (2019). Edge Intelligence: The Convergence of Humans, Things, and AI. In *IEEE International Conference on Cloud Engineering (IC2E) 24-27 June 2019.*

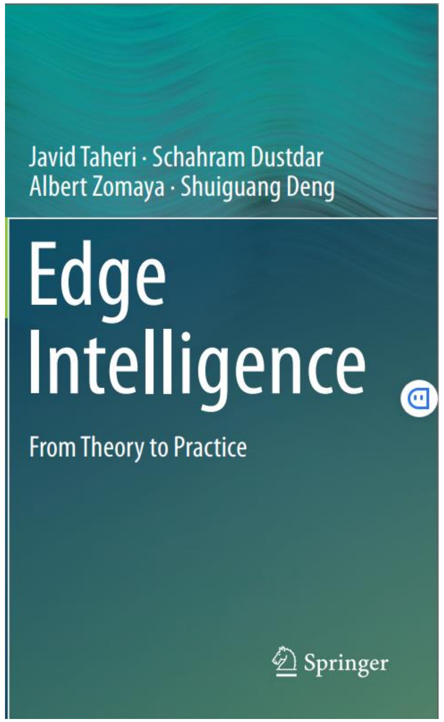
Conclusions

1. Leverage the “Distributed Computing Continuum“ from IoT->Edge->Fog->Cloud

2. Need for an Edge Intelligence AI Fabric and a “clear“ distributed systems ecosystems understanding

3. Differentiate between AI *for* Edge and AI *on* Edge. Both bring their distinct research challenges

Thanks for your attention



Prof. Schahram Dustdar

IEEE Fellow | EAI Fellow | I2CIC Fellow | AAIA Fellow

Member *Academia Europaea*

President of the AAIA (Asia-Pacific Artificial Intelligence Association)

ACM Distinguished Scientist | ACM Distinguished Speaker

TCI Distinguished Service Award by the IEEE Technical Committee on the Internet (TCI)

IEEE TCSVC Outstanding Leadership Award in Services Computing

IEEE TCSC Award for Excellence in Scalable Computing

IBM Faculty award

Distributed Systems Group
TU Wien, Austria **dsg.tuwien.ac.at**