UMEÅ UNIVERSITY

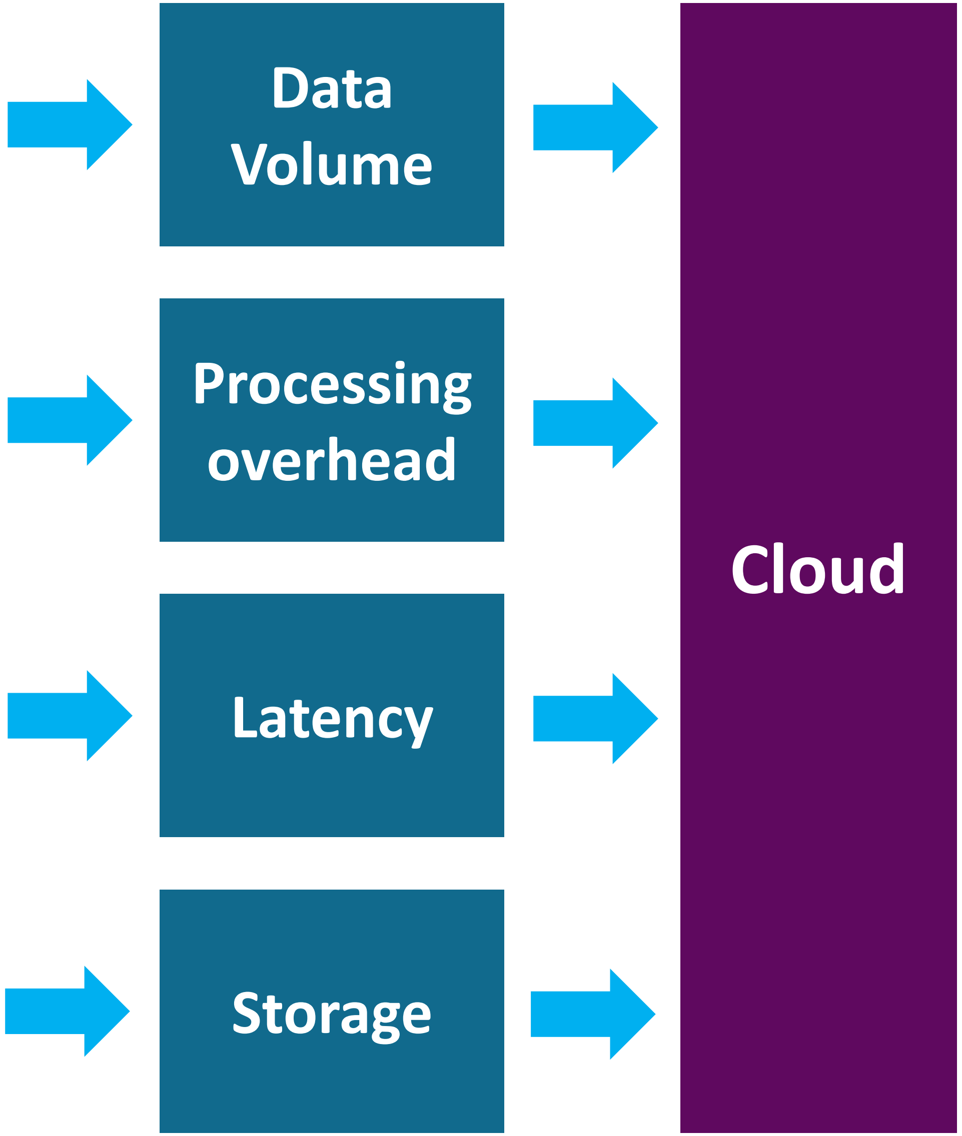
MODELLING AND MANAGING THE FUTURE CLOUD-EDGE CONTINUUM

PAUL TOWNEND

UMEÅ UNIVERSITY, SWEDEN

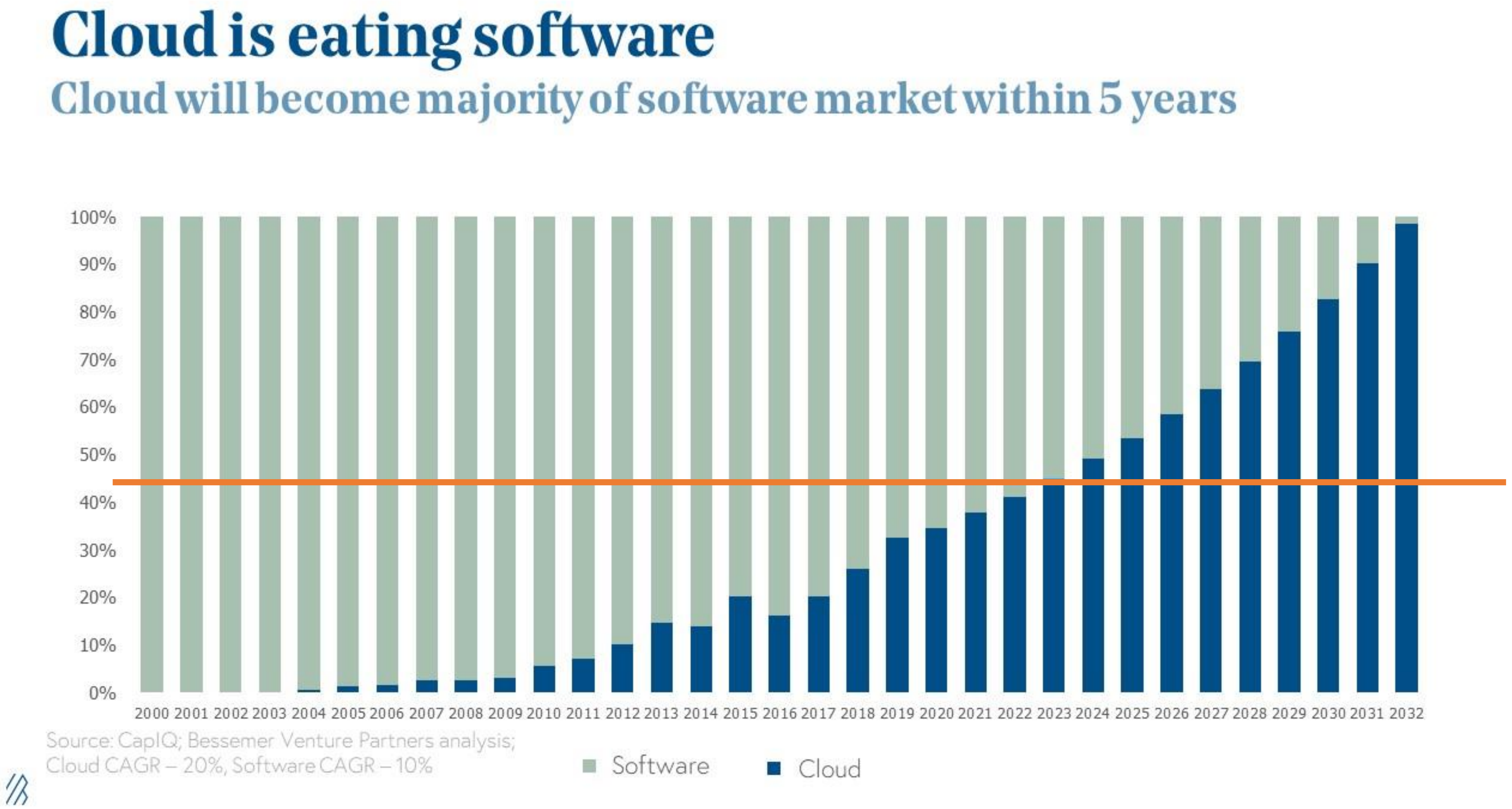
MODERN DISTRIBUTED SYSTEMS AND DATA

UMEÅ UNIVERSITY



OUCH

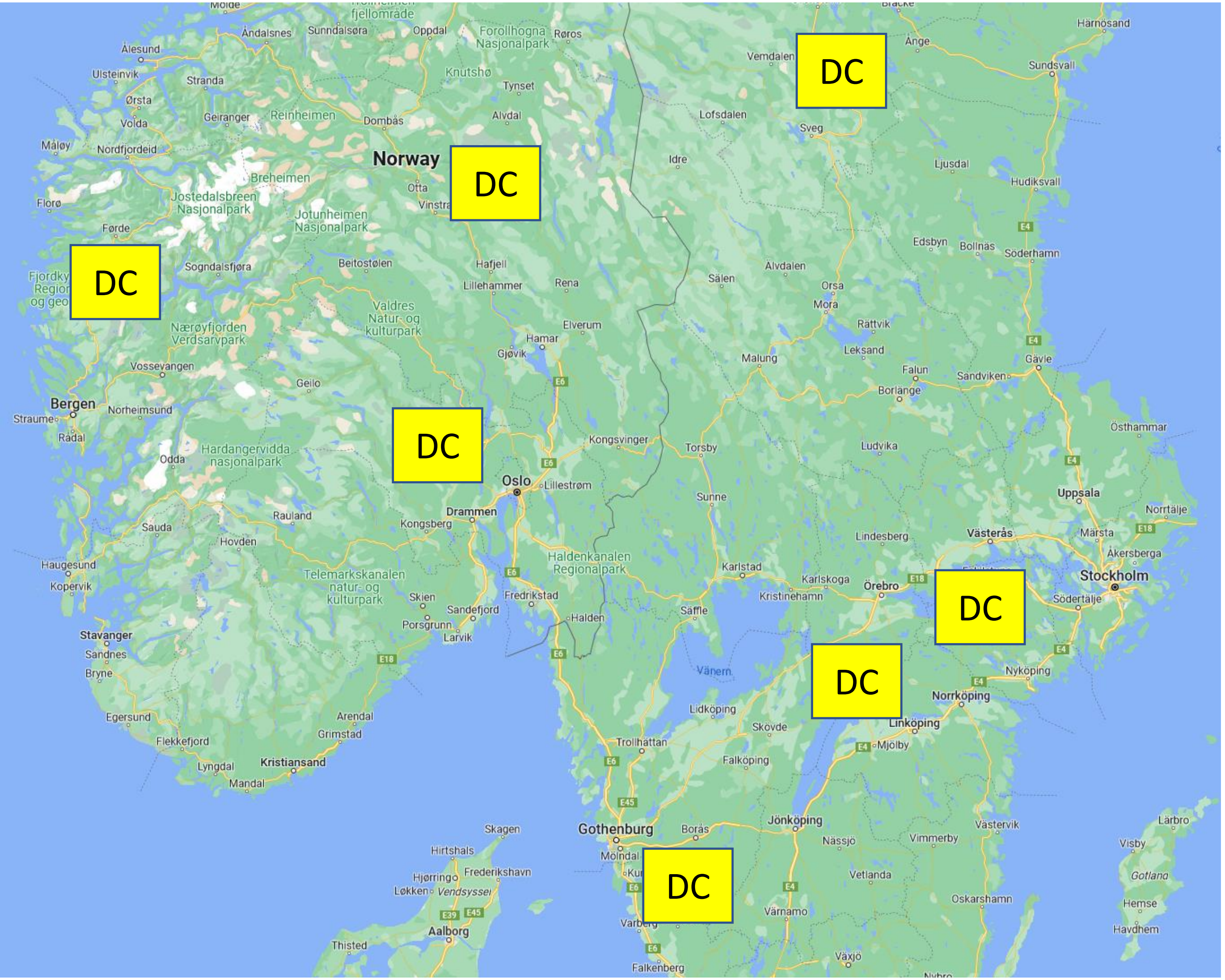
UMEÅ UNIVERSITY



SO WHAT IS CLOUD?

"TRADITIONAL CLOUD COMPUTING"

UMEÅ UNIVERSITY



Users connect to large data centres

(e.g. Facebook, Google, etc.)

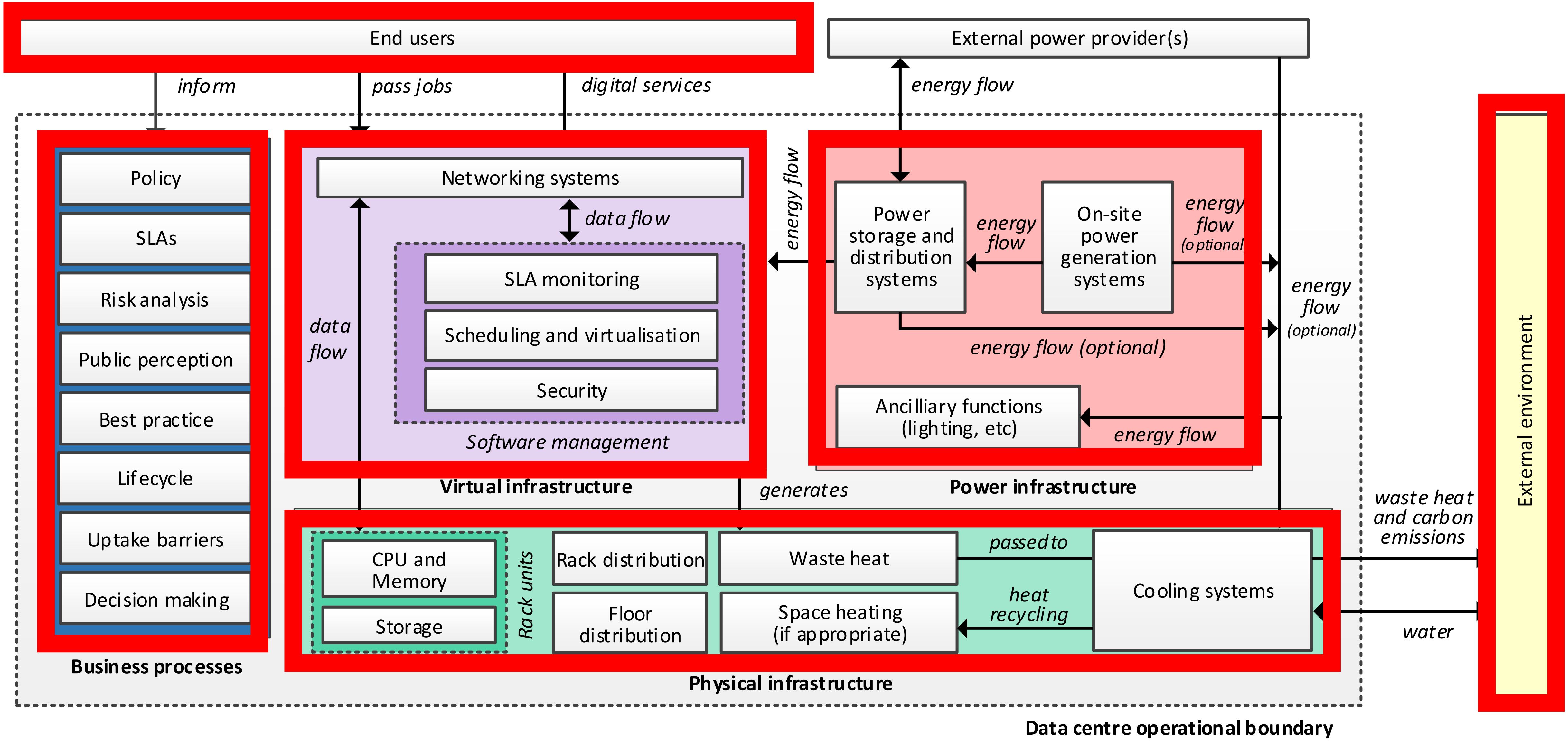
These are powerful resources

Lots of servers, storage, bandwidth, etc.

Expensive, slow to build, huge impact on power grids and the environment

DATA CENTRES AS COMPLEX SYSTEMS-OF-SYSTEMS

UMEÅ UNIVERSITY



GOOGLE DATA CENTRE, HAMINA, FINLAND

RECENT

The Biggest Problem in AI? Lying Chatbots
MAY 30, 2023

Biden's Former Tech Adviser on What Washington Is Missing About AI
MAY 30, 2023

HPE and Ampere Take Aim at Intel With Vision of Arm-Based Open RAN Server
MAY 25, 2023

Amazon's Answer to ChatGPT Seen as Incomplete
MAY 24, 2023

Are Data Centers Taking Over Oregon's

COMPANIES > GOOGLE (ALPHABET)

Google Using Sea Water to Cool Finland Project

Google will use cool sea water in the cooling system for its new data center in Hamina, Finland, which may be the first sea-cooled data center. The initiative continues Google's focus on data center efficiency and sustainability.

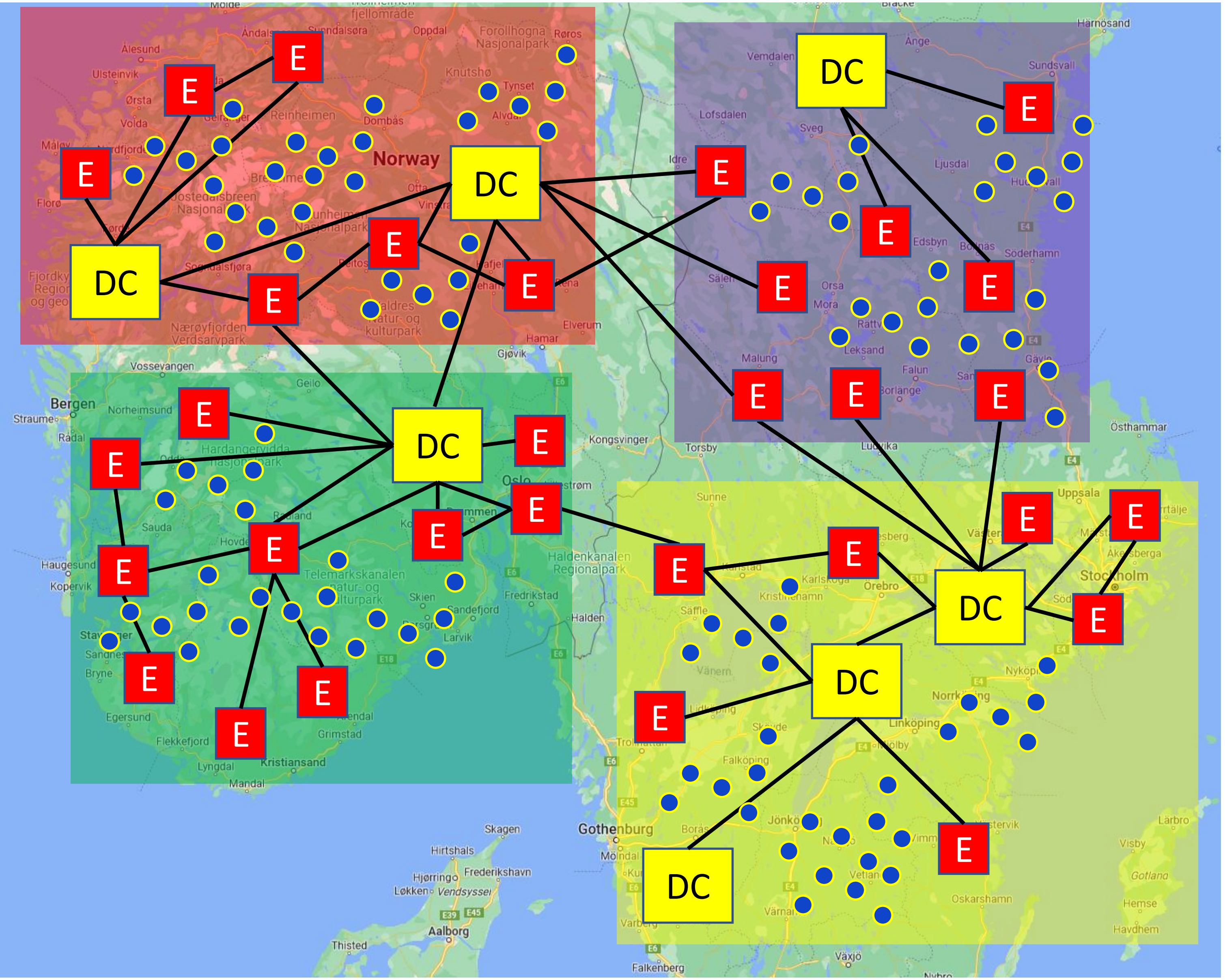
Rich Miller | Sep 15, 2010

Google will use cool sea water in the cooling system for its new data center in Hamina,

The Cloud-Edge Continuum and its characteristics

THE CLOUD-EDGE CONTINUUM

UMEÅ UNIVERSITY



50 billion connected devices by 2025

This would overwhelm centralised DCs

Proliferation of "intermediate" edge DCs

Complex federations of devices

Regional considerations

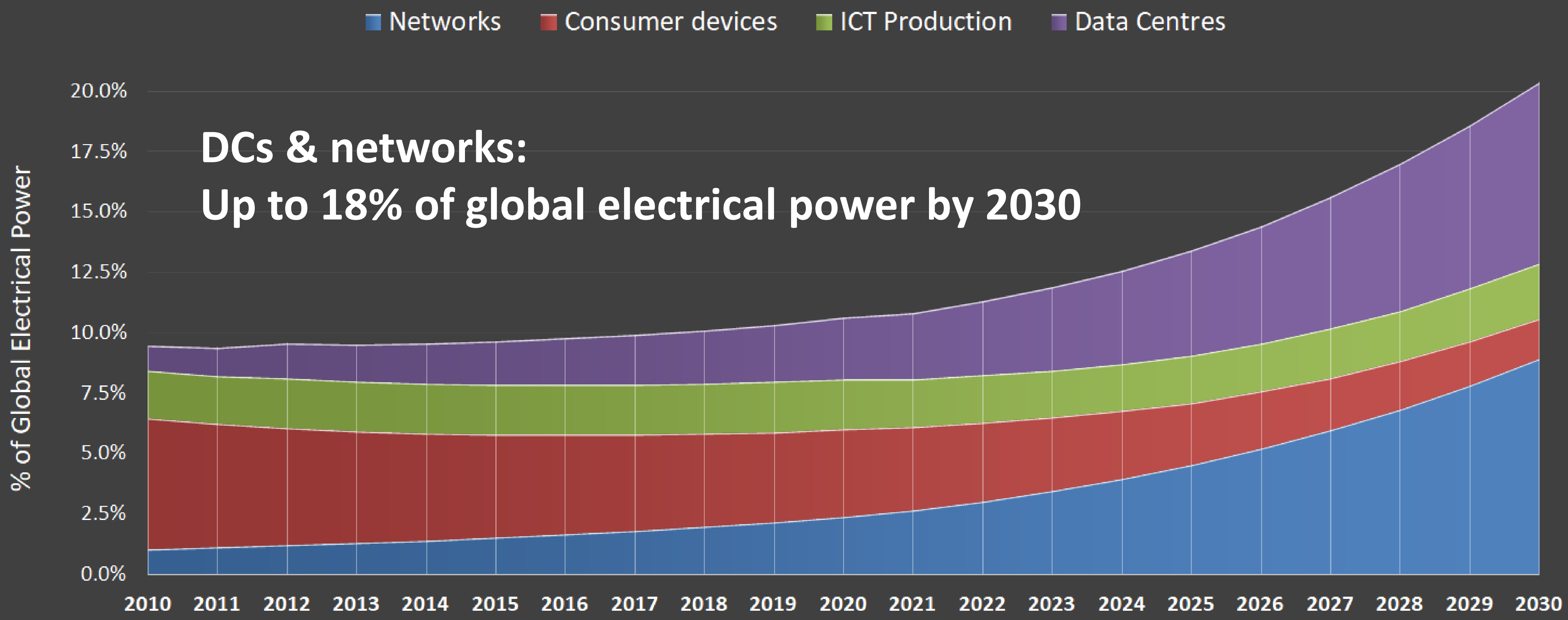
CONTINUUM CHARACTERISTICS

- Multiple disparate providers
- Platform heterogeneity
- Resource constrained devices
- Infrastructural dynamicity
- Secure orchestration over public networks

Massive complexity

High energy consumption

DATA CENTRE AND NETWORK ENERGY CONSUMPTION



DCs & networks:
Up to 18% of global electrical power by 2030

A. Anders, T. Edler, "On global electricity usage of communication technology: trends to 2030.", Challenges 6, no. 1 (2015): 117-157

Managing complexity:

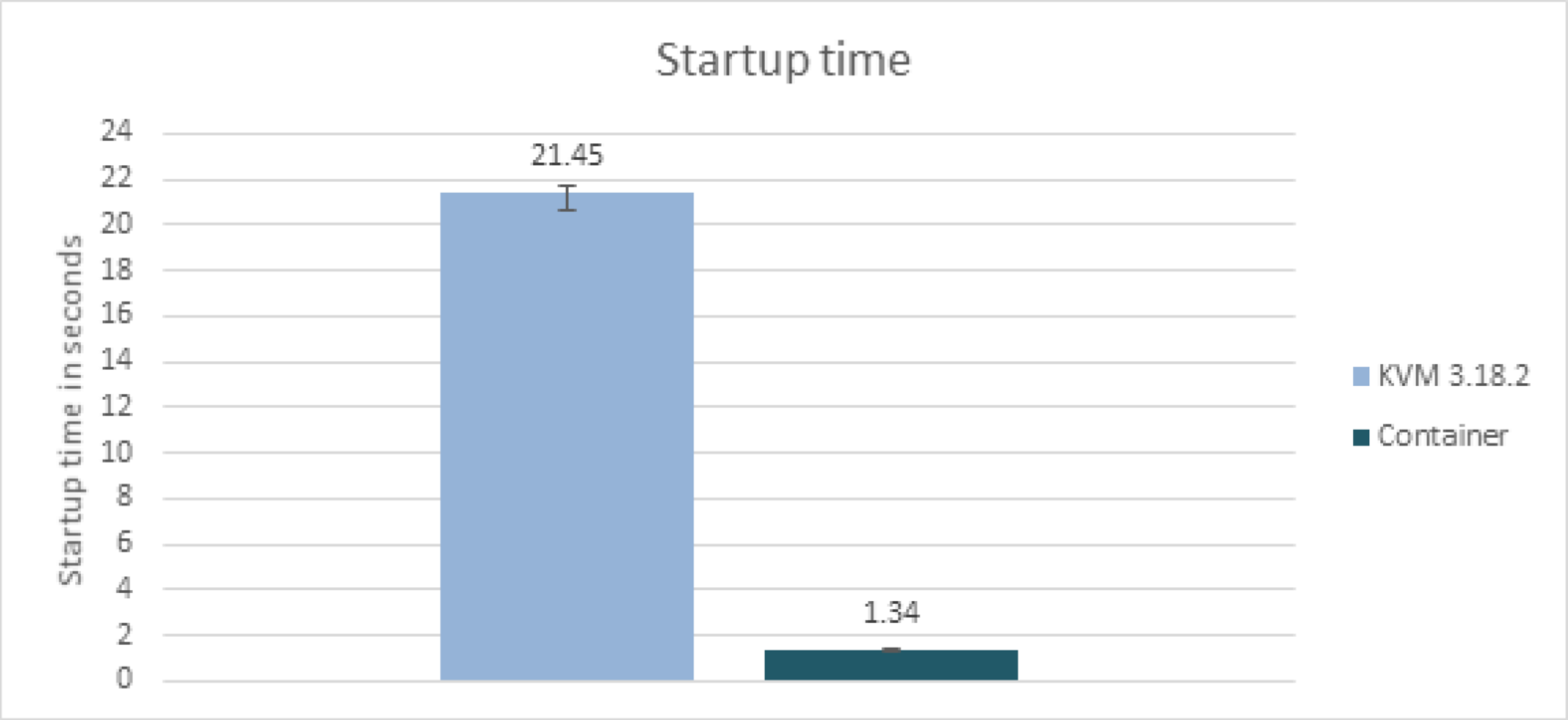
Containers and Serverless computing

CONTAINERS AND VIRTUAL MACHINES

PERFORMANCE COMPARISON

UMEÅ UNIVERSITY

Average Startup Time (Seconds) for a KVM Linux Virtual Machine and a Container Over Five Measurements



VIRTUALISED NETWORK INFRASTRUCTURE

We have an infrastructure to instantiate, run, and manage containers almost instantly

Why not virtualise network functions?

Network address translation
Firewalls
Routing services
Etc.

Run and manage the network on cheap commodity servers

The challenge then becomes:

running and managing the network on cheap commodity servers

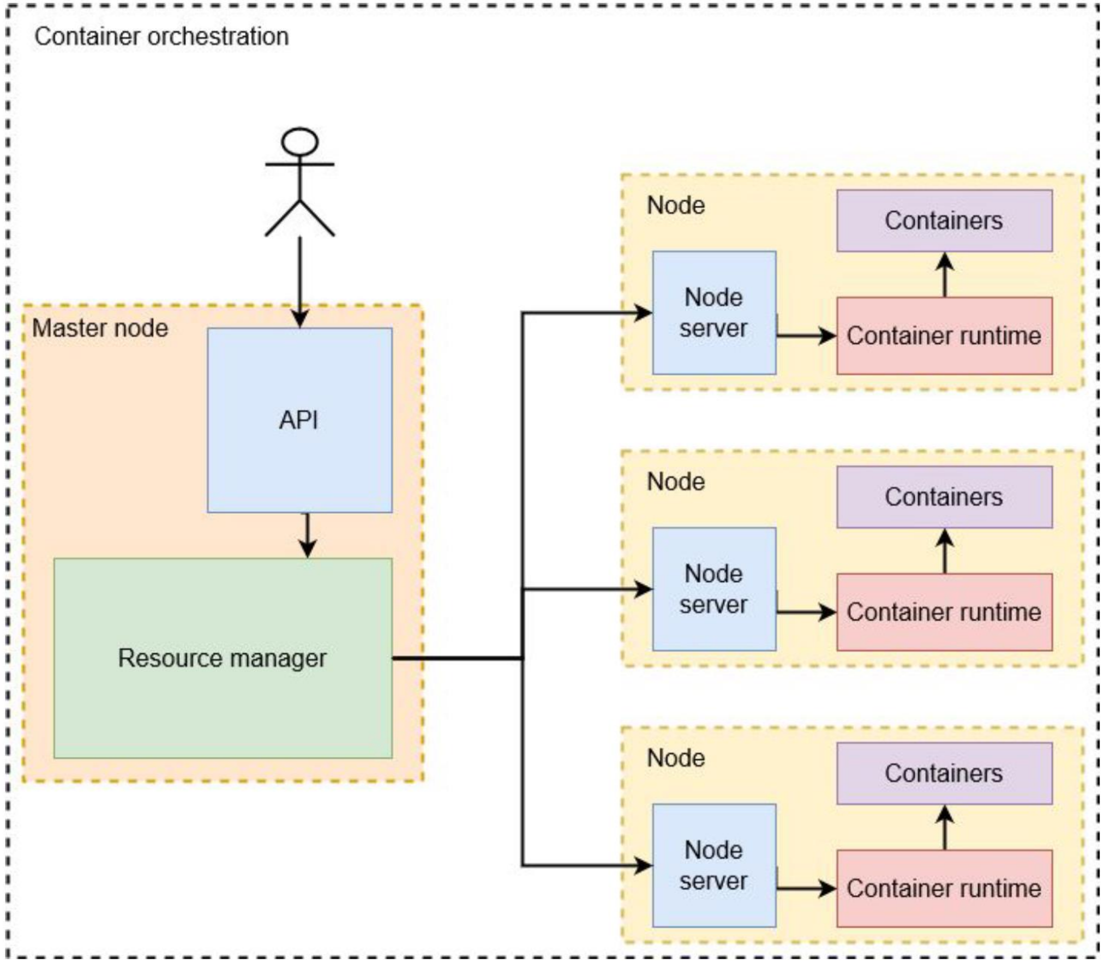
CONTAINER ORCHESTRATION

UMEÅ UNIVERSITY

Container orchestrators are crucial for deploying, managing, and monitoring container systems

Container engines deploy container images, running container runtimes

Container orchestrators manage the runtimes and the live system as a whole



CONTAINER ORCHESTRATION PROVIDERS

UMEÅ UNIVERSITY

Azure Kubernetes Service (AKS)

IBM Cloud Kubernetes Service

Google Kubernetes Engine

Amazon ECS

KUBERNETES

UMEÅ UNIVERSITY

By far the leading container orchestration platform in the world

Portable

Extensible

Open-Source

Huge ecosystem

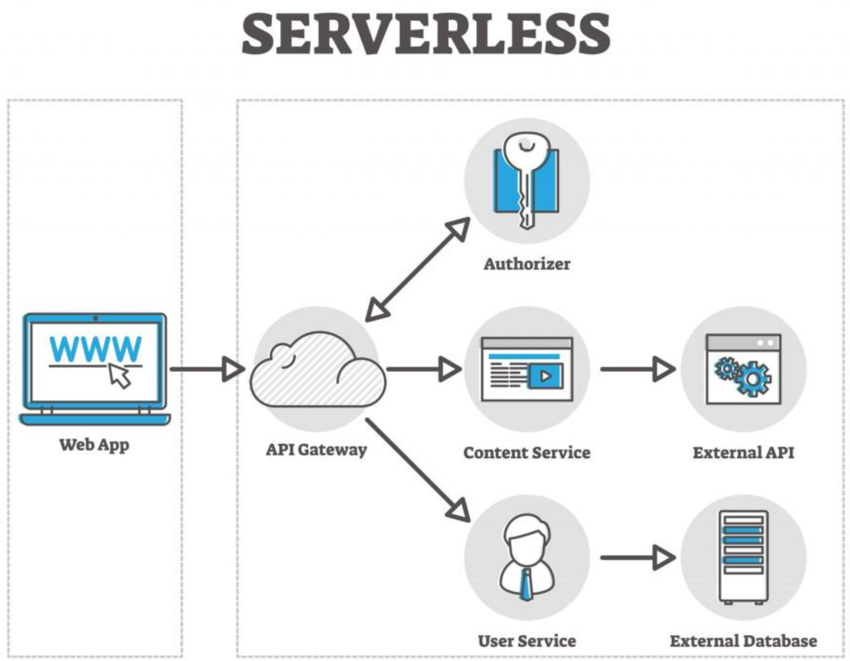
KubeCon 2022

17,000 attendees

3,000 companies

CAN SERVERLESS HELP?

UMEÅ UNIVERSITY



Users / developers:

Application functionality composed of invokable services (FaaS)

Cloud providers:

Auto-provision, deploy, and scale the services based on range of criteria

Abstract infrastructure from applications

How does this work in a Cloud-Edge continuum, given the characteristics discussed?

Serverless at the Edge: Research challenges

SERVERLESS CHALLENGES IN THE CONTINUUM

Cognitive Cloud-Edge management:
disaggregated, highly distributed, SDNs, etc.

Deployment and migration of functions:
Overlays to abstract heterogeneity etc.

Seamless user access and programming models:
automatic offload with specific requirements

Secure and trusted execution:
Fine-grained control in multi-provider networks

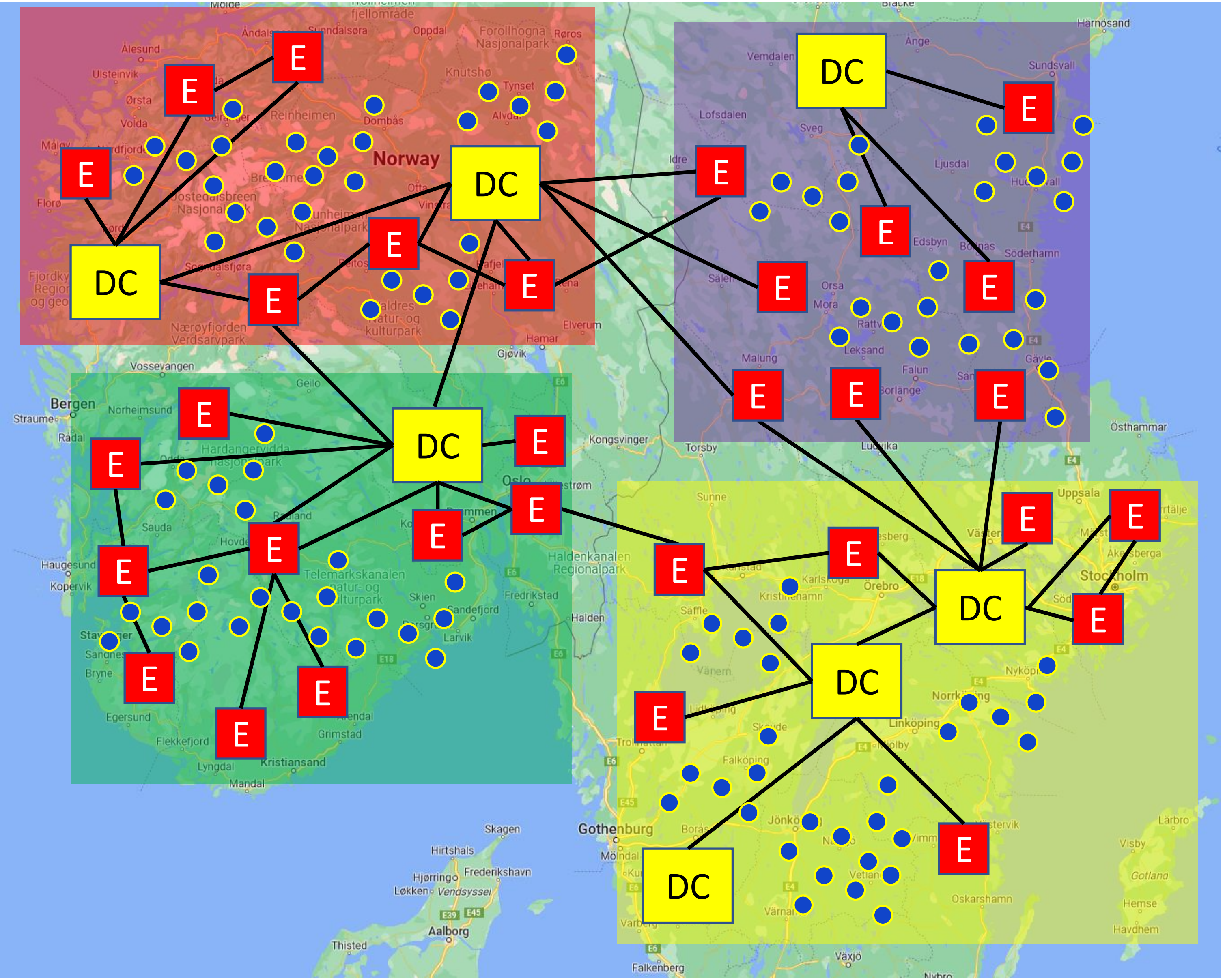
New serverless functionalities:
Long-term execution, location-awareness, etc.

Energy efficiency and adaptation to green energy:
Placement, renewables, models, etc.

AI models and optimisation techniques:
How to accurately manage and predict without incurring excessive latency or overhead

RECAP OF CLOUD-EDGE

UMEÅ UNIVERSITY



50 billion connected devices by 2025

This would overwhelm centralised DCs

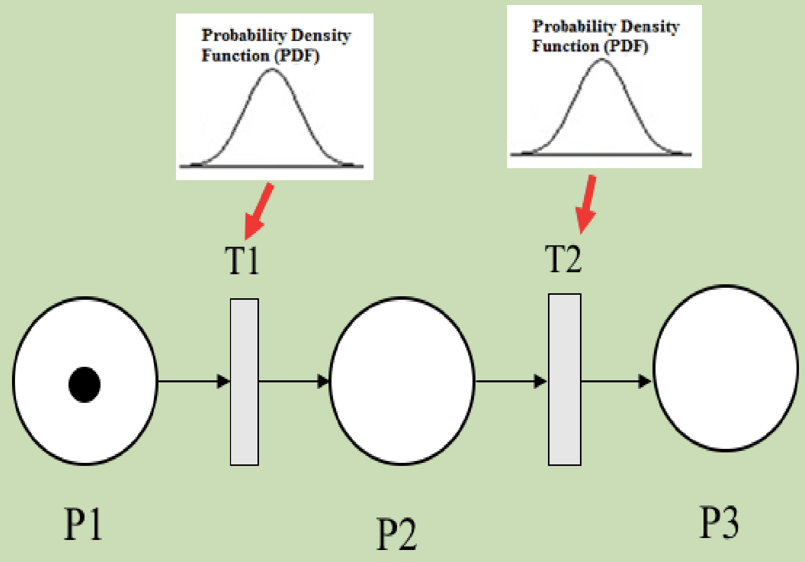
Proliferation of “intermediate” edge DCs

Complex federations of devices

Regional considerations

MODELLING THE CONTINUUM

UMEÅ UNIVERSITY



A Simple Generalized Stochastic Petri Net Model

Conceptual models:

Stochastic process algebra, Discrete event simulation, Queueing theory, Approximation theory, Game theory, etc.

Need of modelling solutions:

Understanding the behavior, performance, and resource orchestration in cloud-edge systems.

Few formal models of federated Cloud-Edge systems exist

None adequately represent and integrate energy and network considerations

MODELLING CHALLENGES

UMEÅ UNIVERSITY

How to model the system

Stochastic features, network, pricing, energy distribution, policies, etc.

How to maintain energy-perf trade-offs

Best practices, Conflicting SLOs, local & global optima, etc.

How to combine multiple models

Model types, granularities, scaling, interactions

How to model energy-driven systems

Power grids capacities, renewables, service levels

How to model different system regions

Optimization, heterogeneity, monitoring, scheduling

How to develop validation models

Scalability scenarios, Iterative testing, mobility, topology, network behavior, energy, etc.

SovereignEdge.COGNIT

Sovereign EDGE.EU

European Open Source for Europe's Next-Gen Edge Cloud

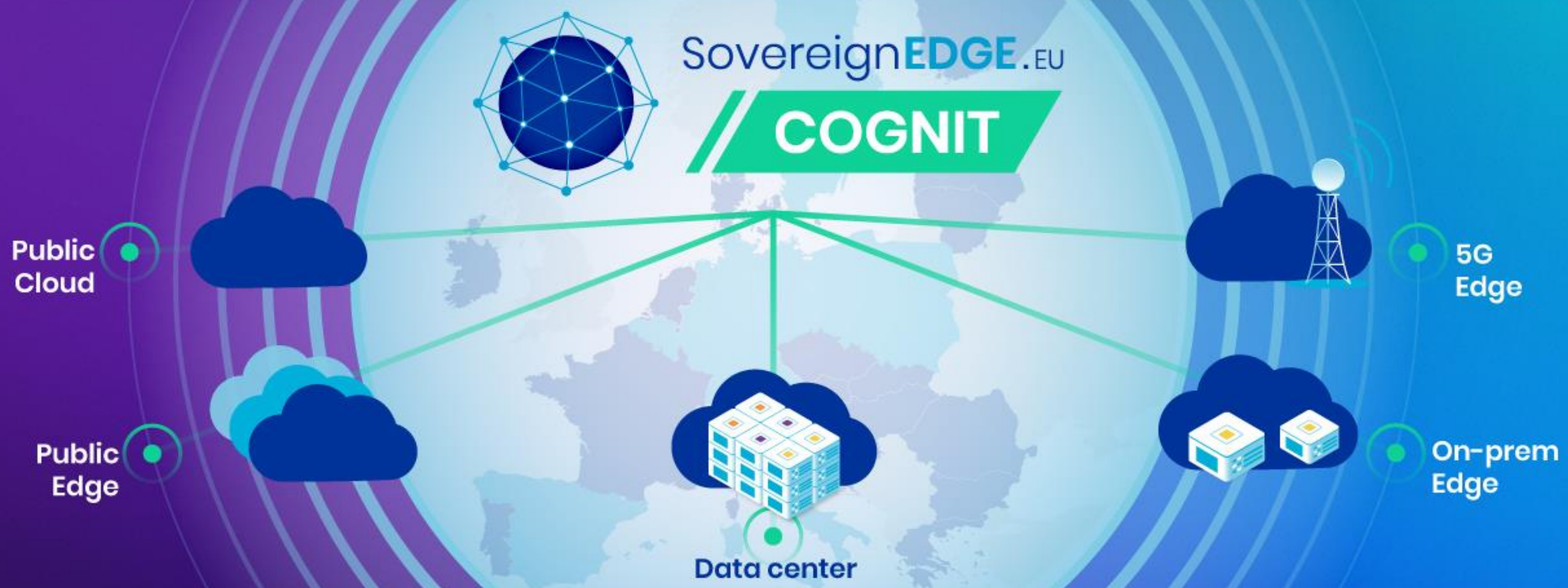
Building a Sovereign Edge Cloud Stack for the Digital Decade

Technologies | Webinar

https://sovereignedge.eu/

A Cognitive Serverless Framework for the Cloud-Edge Continuum

Topic: **HORIZON-CL4-2022-DATA-01-02** (Cognitive Cloud) · Execution Dates: **2023 - 2025**



OpenNebula · UMEÅ UNIVERSITET · ikerlan · cetic · RISE Research Institutes of Sweden · SUSE · acisa · Nature 4.0 · Phoenix Systems · ATENDE INDUSTRIES

PROJECT PARTICIPANTS

UMEÅ UNIVERSITY

Umea University, Sweden
Paul Townend
Monowar Bhuyan
P-O Ostberg
Erik Elmroth

Ikerlan, Spain
Idoia de la Iglesia
Marco González
Iván Valdés
Aritz Brosa
Martxel Lasa
Goiuri Peralta
Samuel Pérez

Open Nebula, Spain
Alberto P. Martí
Constantino Vázquez
Marco Mancini
Ignacio M. Llorente
Michael Abdou

ACISA, Spain
Joan Iglesias
Antonio Lalaguna
Behnam Ojaghi

SUSE, Germany
Torsten Hallmann
Holger Pfister

Nature 4.0, Italy
Riccardo Valentini
Francesco Renzi
Micaela Onorati

RISE, Sweden
Thomas Timoudas
Daniel Olsson
Johan Kristiansson
Shuai Zhu

Atende, Poland
Dominik Bocheński
Tomasz Piasecki
Grzegorz Gil

CETIC, Belgium
Nikolaos Matskanis
Philippe Massonet
Sébastien Dupont
Malik Bouhou

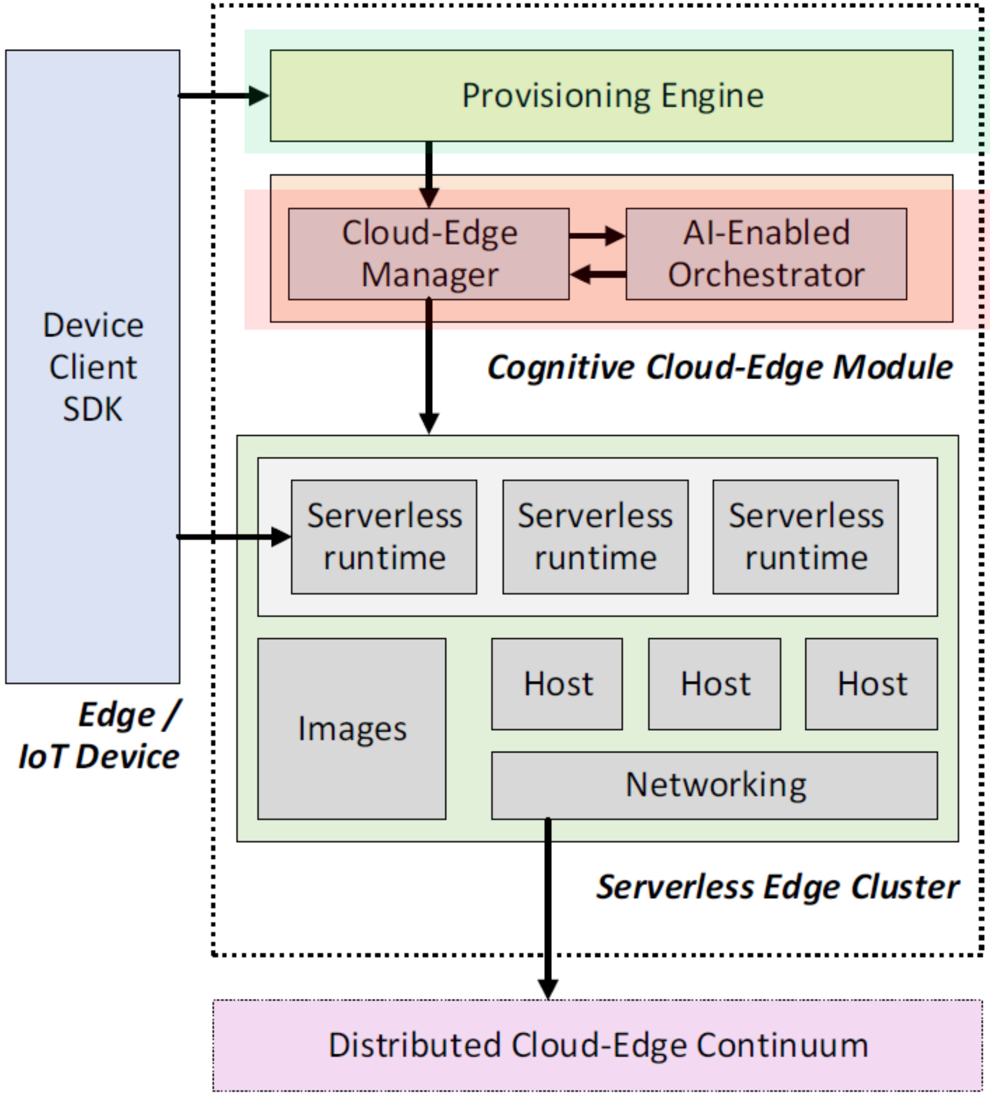
Phoenix Systems, Poland
Kaja Swat
Tomasz Korniluk
Marek Białowąs
Gerard Świderski
Rafał Jurkiewicz

10+ organisations
40 researchers

COGNIT architecture and use cases

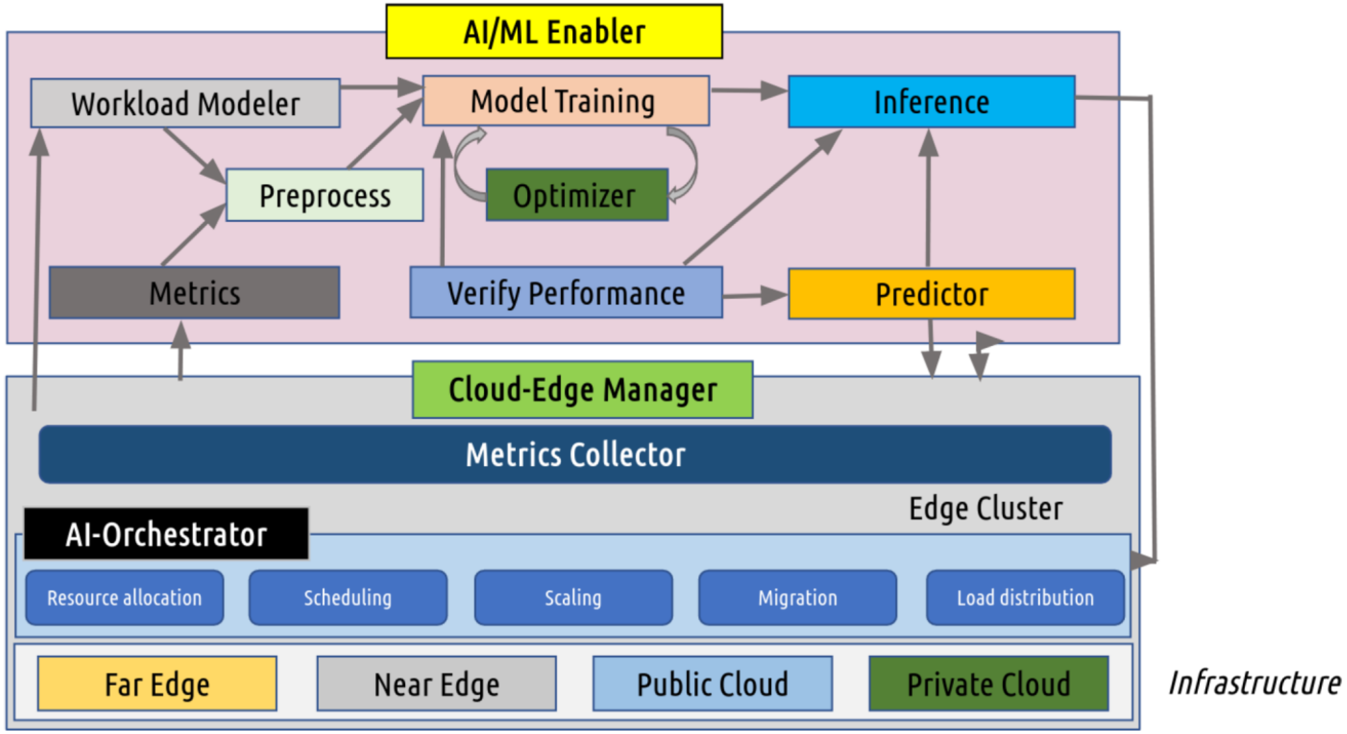
MAIN ARCHITECTURE

UMEÅ UNIVERSITY



Stateless component responsible for managing the life cycle of the Serverless Runtimes

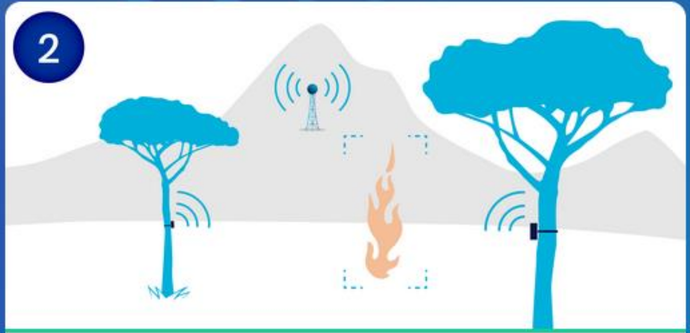
AI reasoning. First version will look at workload placement based on renewable power source availability



FOUR USE CASES

Smart Cities

Coordinated by ACISA



Wildfire Detection

Coordinated by Nature 4.0

Energy

Coordinated by Phoenix Systems & Atende Industries

Cybersecurity

Coordinated by CETIC and SUSE

COGNIT TESTBED

UMEÅ UNIVERSITY

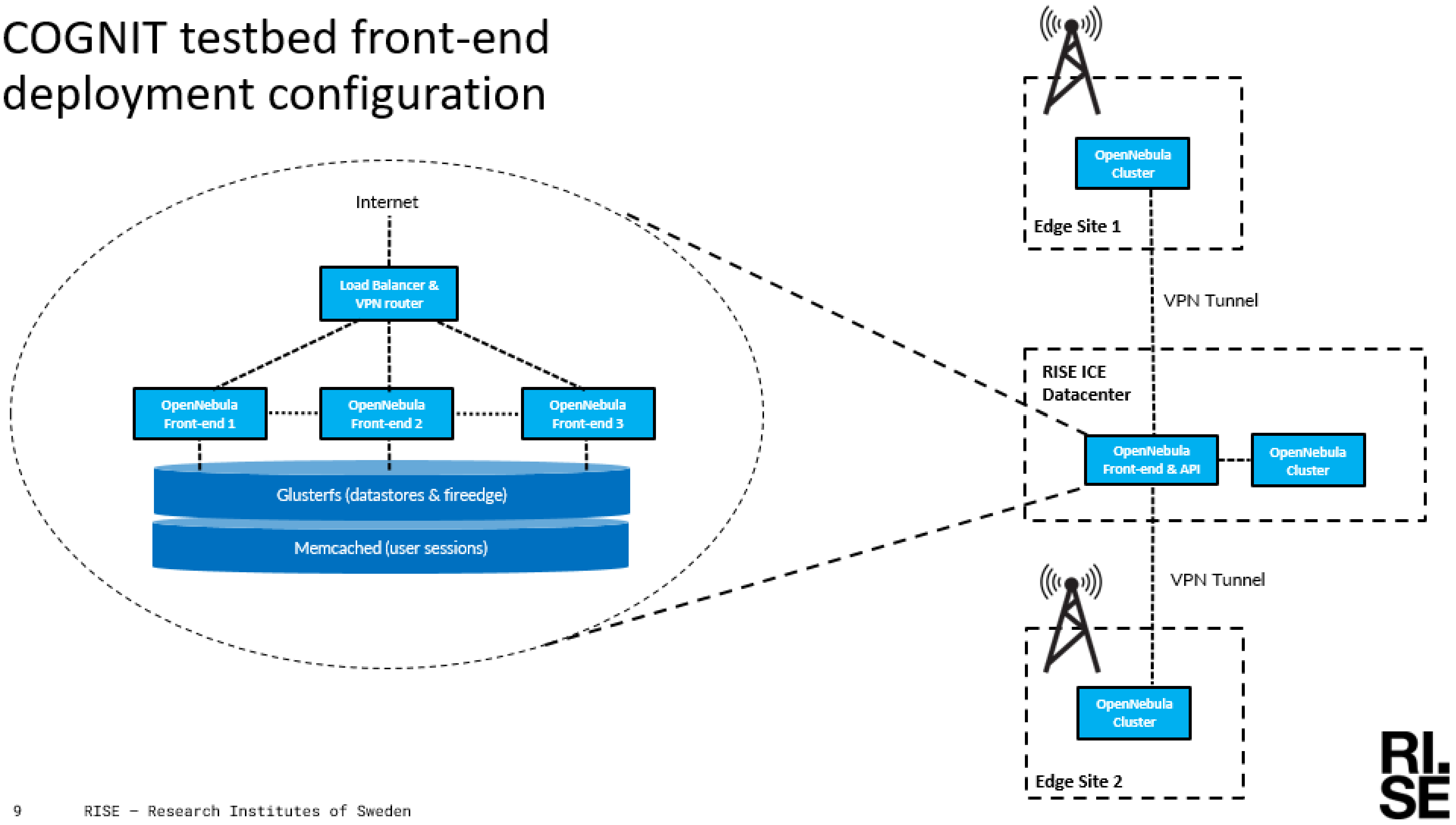
Commercial and Research Data Center

DCD Best Data Center Initiative 2017

IEEE Scale Award 2017

Building 2000+ Node Container Facility

COGNIT testbed front-end deployment configuration



An open source reference implementation for serverless Continuum computing

First version released in **September 2023**, more advanced version in **March 2024**

All versions of COGNIT will be tested in physical + virtual testbeds and use cases

Happy to collaborate, incorporate interesting new technologies, etc.

Going forward

AT UMEÅ

Energy-aware Autonomous management

Intelligently allocate resources: e.g. target renewable energy, etc.

Horizon Europe SovereignEdge (2023-2025)

Efficiently monitor, predict, and audit at massive scale

How to adaptively do this to avoid huge overhead?

WASP WARA-Ops (2023-2025)

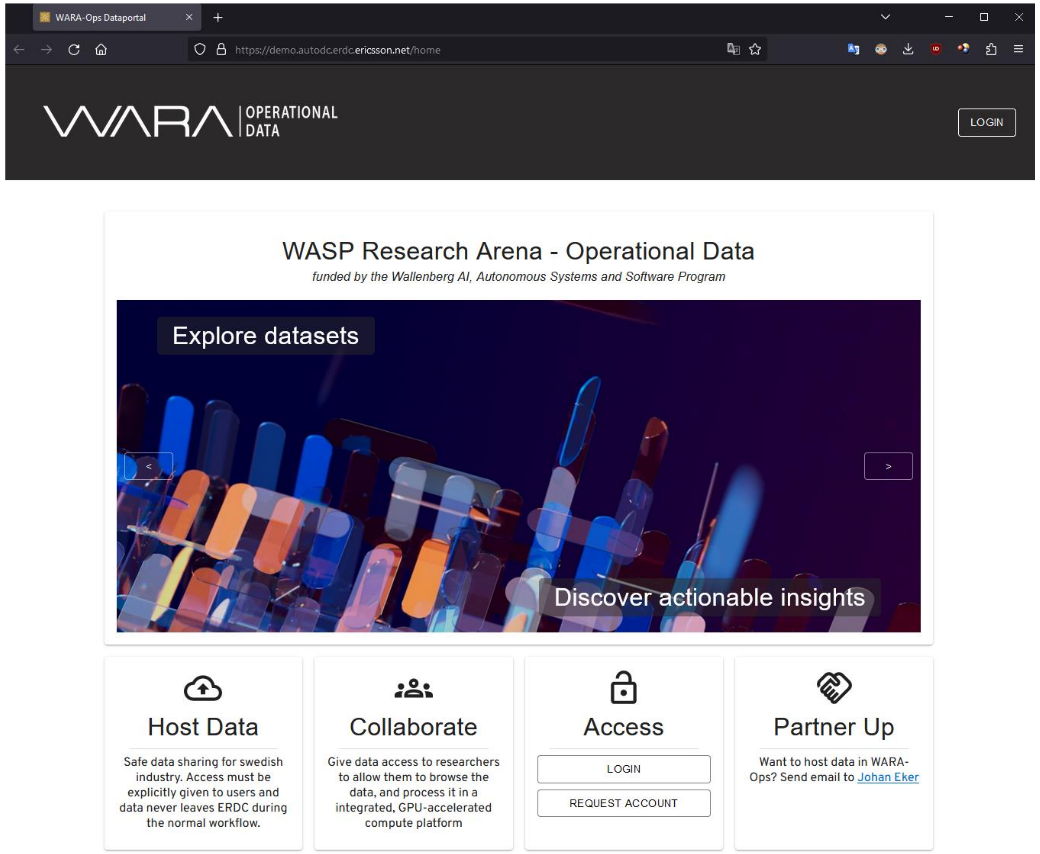
A formal model for energy-aware Cloud-Edge Systems

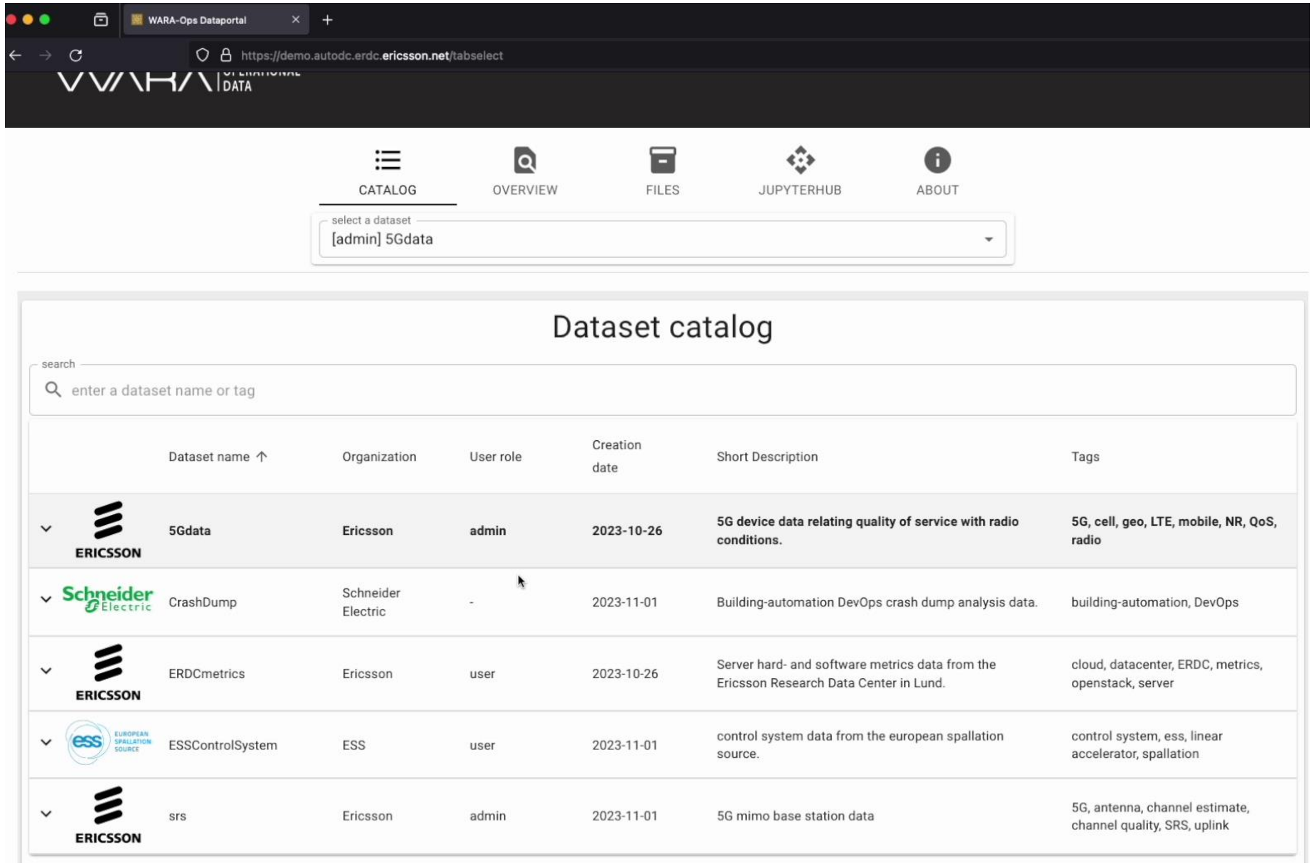
Integrate energy providers, pricing, renewables, etc. into existing Cloud-Edge models

WASP Academic PhD (2023-2026)

UMEÅ UNIVERSITY

WARA | OPERATIONAL DATA





MEDIA

HOME > NEWS >

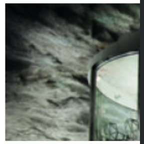
Bahn
powe
data

Small nuc
Bond styl

March 07, 2023

Swedish data
center, accord

The company,
together plans
Stockholm wh
households, al



Nyheterna | Tipsa TV4 | Nyhetsmorgon | Efter fem | Kalla fakta

Molntjänsterna slukar el – nu ska forskarna lösa problemet

Molntjänsterna slukar el – nu ska forskarna lösa problemet

Får tillgång till telekomföretagets serverhall • Ska öka effektiviteten

Uppdaterad: 3 februari, 2022 Publicerad: 3 februari, 2022

Vi blir allt mer digitala och molntjänster drar mycket el: enligt beräkningar som EU-kommissionen gjort kommer nästan en tiondel av elproduktion i världen gå åt till våra internetaktiviteter år 2030. Nu ska forskare vid Umeå universitet försöka hitta metoder som ökar effektiviteten och får ut mer av elen som förbrukas.

– När du tittar på en video på Youtube spelas den egentligen upp någon annan stans, på ett datacenter på någon plats i Europa, säger Paul Townend, universitetslektor i datavetenskap vid Umeå universitet.

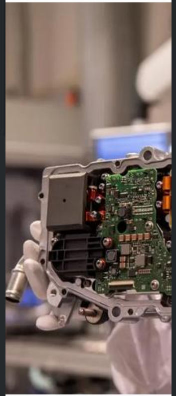
Får tillgång till serverhall

En stor serverhall drar ungefär lika mycket el som 80 000 villor. För att ta reda på hur man ska kunna energieffektivisera kommer forskarna få tillgång till telekomföretaget Ericssons serverhall i Lund.

S > INNOVATION > AI

reen IT Is
ption For

in Duranton Contributor
about tech, deep tech, gr
gence

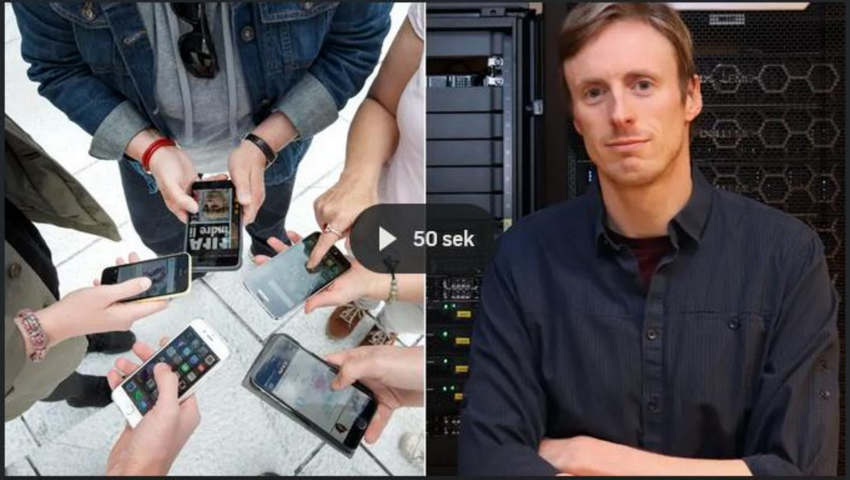


MBERG, GERMANY - MARCH
... [+] GETTY IMAGES

come to the wild worl
al infrastructure of o
unts of electricity and
ake.

svt NYHETER | Nyheter | Lokalt | Sport | SVT Play | Barn

/ VÄSTERBOTTEN



"Vi hoppas kunna minska energiåtgången med upp till 25 procent" – se Umeåforskaren Paul Townend i klippet. Foto: TT/Victoria Skeidsvoll/Umeå universitet

Molntjänster slukar el – Umeåforskare: "Bråttom att effektivisera dessa"

UPPDATERAD 8 FEBRUARI 2022 PUBLICERAD 8 FEBRUARI 2022

Världens datacenter förbrukar allt mer el, så pass att de inom åtta år ser ut att kräva nio procent av all elförsörjning. Nu hoppas forskare i Umeå hitta en lösning – med hjälp av artificiell intelligens. Se mer i klippet.

s TikTok data
for
mmunition

plans to increase
demands of nearby

apacity to assist Ukraine has been
by TikTok. Photograph: Geir

rers has said efforts to meet
een stymied by a new TikTok
e region close to its biggest

SOME CHALLENGES WE ARE CONSIDERING

How do we model and integrate energy and network into Cloud-Edge?

How do we monitor at massive-scale without being overwhelmed with data?

How do we deal with conflicting demands between DCs and energy providers?

How do we optimise and negotiate in near real-time?

How do we store so much information for later audit?

Where does ML and other AI fit into this?

6G CONSIDERATIONS

What are the key components we need to introduce into our models?

How to integrate network models into our existing Cloud-Edge models?

How to integrate (lightweight) AI/ML for 6G?

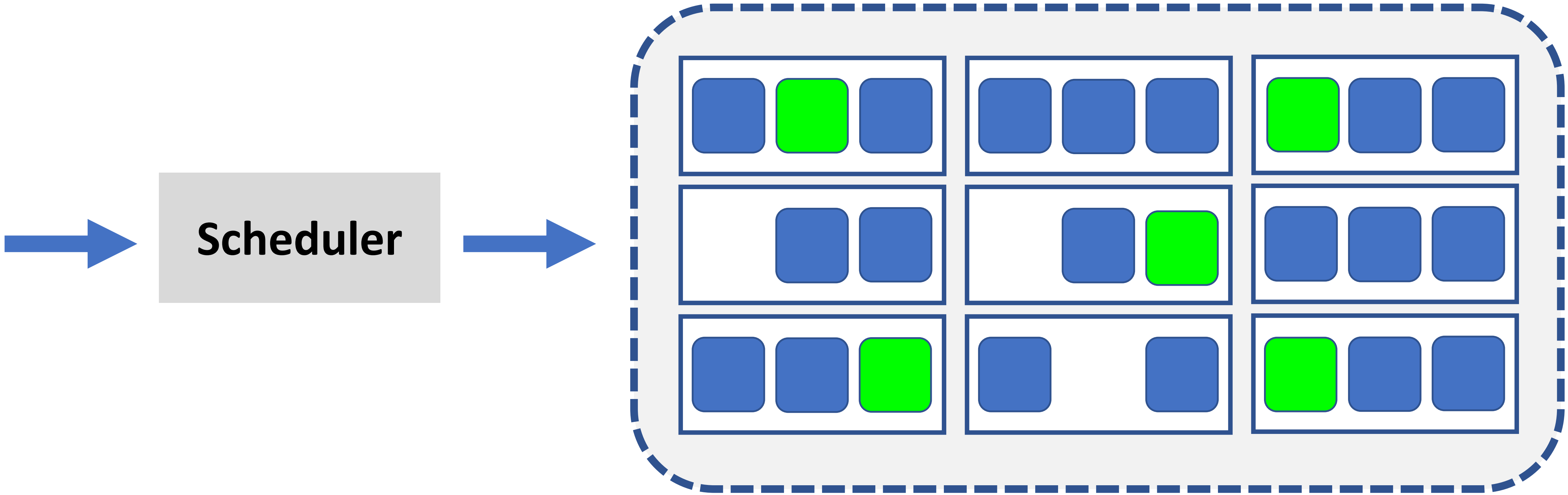
Rapid anomaly detection

Behavioural prediction

What orchestration technology is appropriate?

An example of container orchestration for energy

Virtualise resources – and **schedule workloads** in a more effective manner



WHAT CAN WE DO WITH SMARTER SCHEDULING?

UMEÅ UNIVERSITY

Over-allocation	Interference	Optimise hardware
Allocate more work on the same nodes	Avoid contention between co-located workloads	Allocate work until nodes are at “optimum” efficiency
Use less machines	Reduce power, improve performance	Reduce power, improve performance

CASE STUDY

UMEÅ UNIVERSITY

RISE SICS, NORTHERN SWEDEN

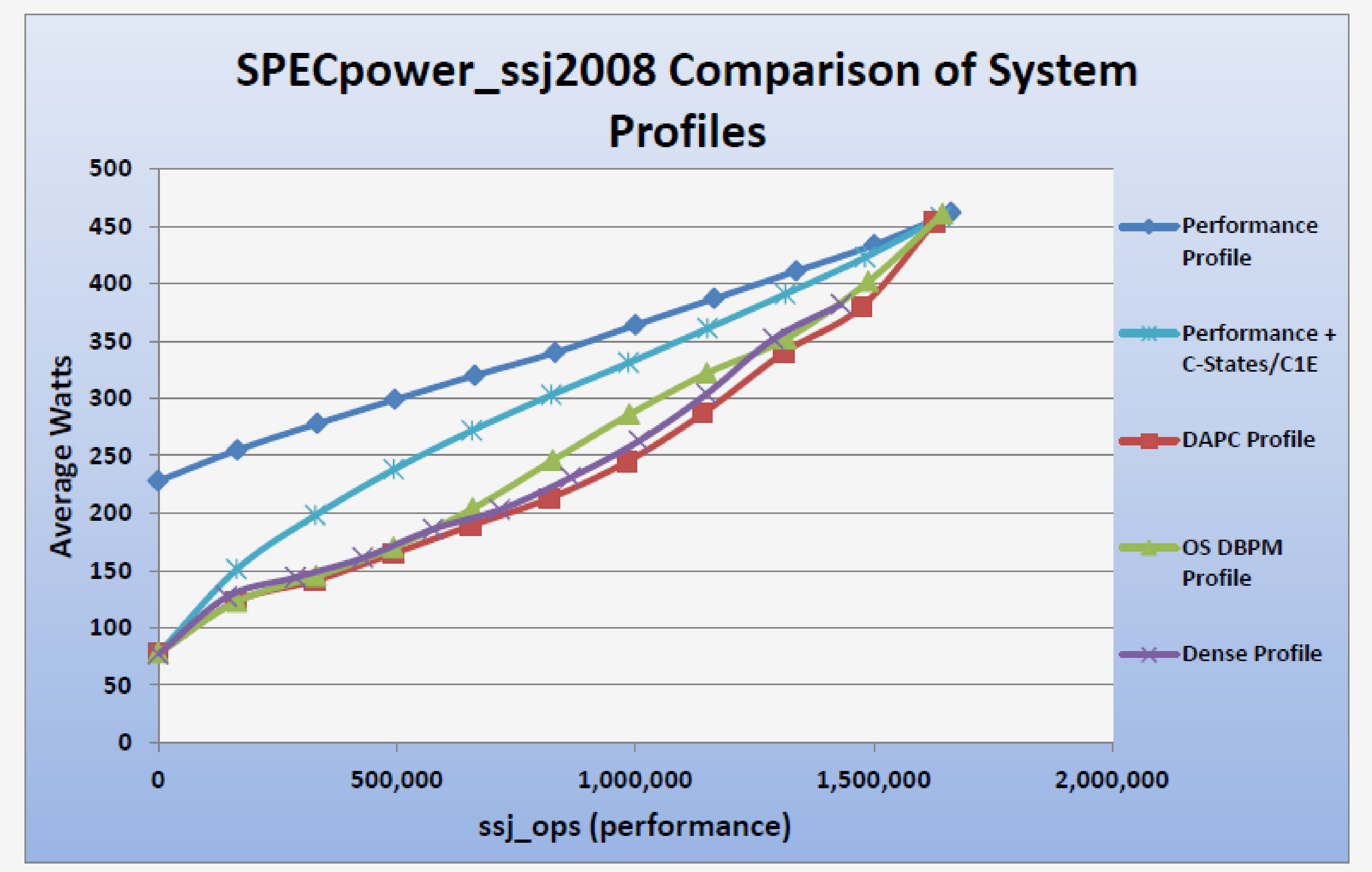
Commercial and Research Data Center

DCD Best Data Center Initiative 2017

IEEE Scale Award 2017

Building 2000+ Node Container Facility

EXAMPLE SPECPOWER BENCHMARK

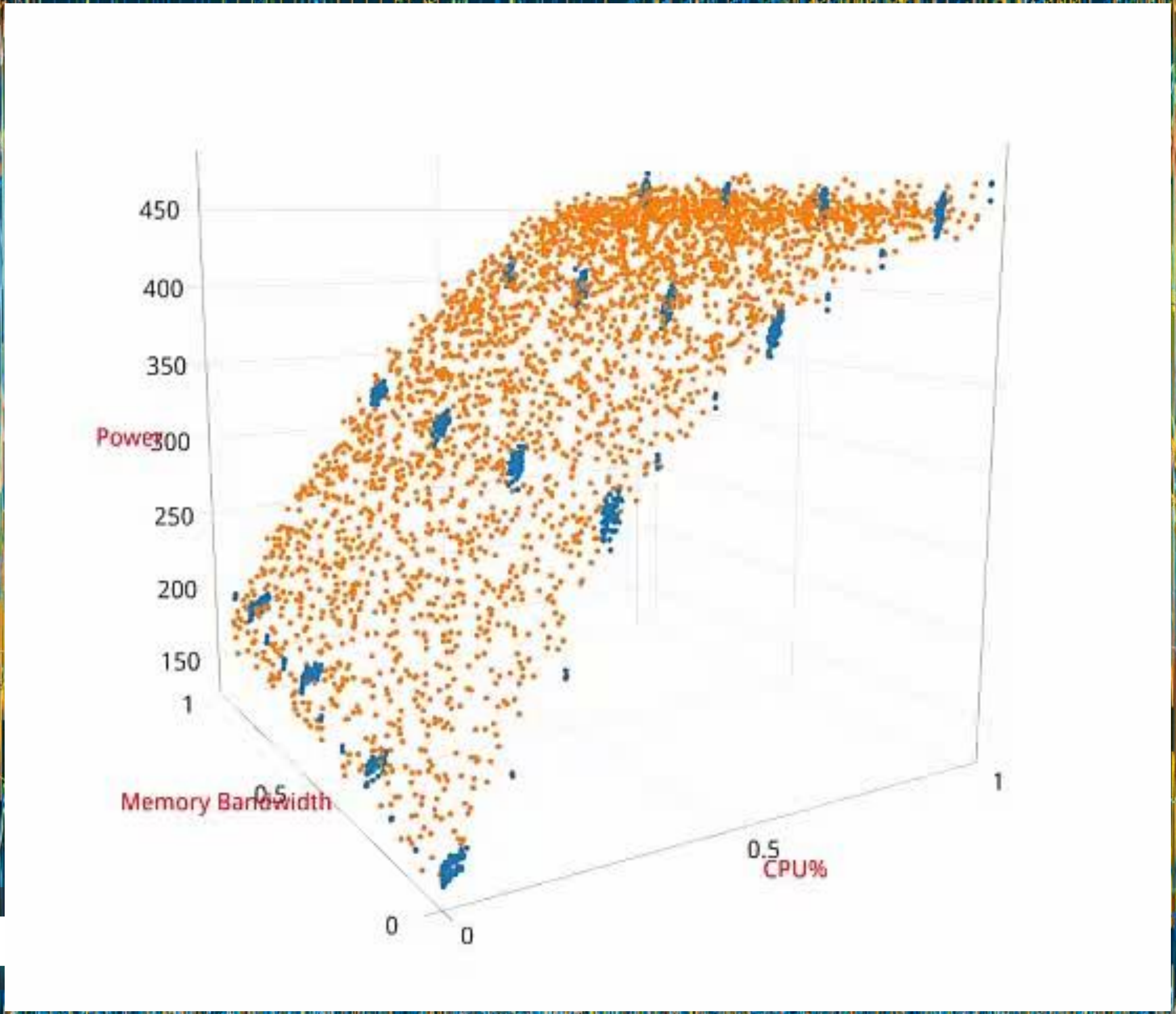


WIND TUNNEL BASED SERVER MODELLING

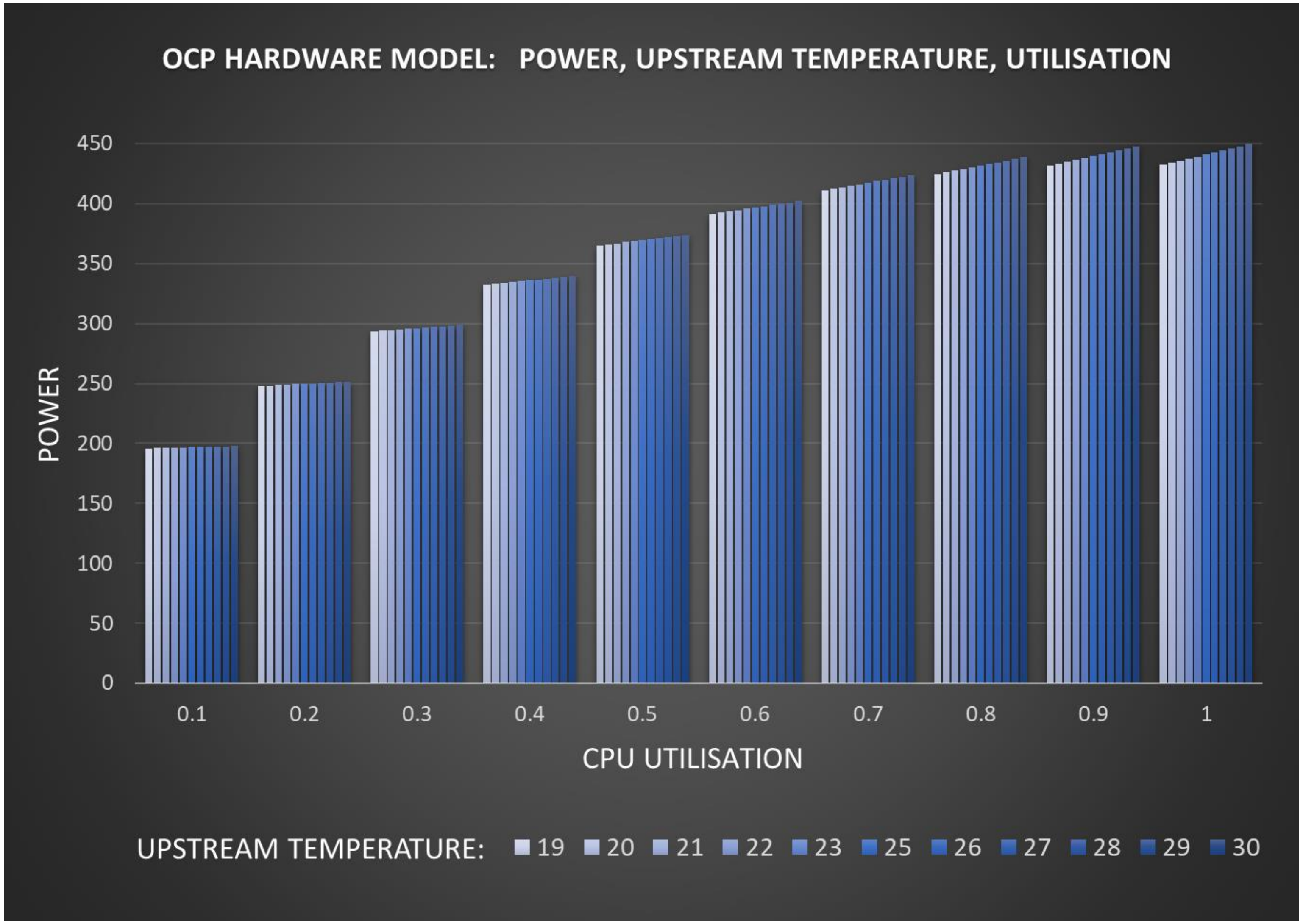
SERVER MODELLING

Actual readings (in blue)
Predicted values (orange)

ND T



HOW ENVIRONMENT AFFECTS POWER

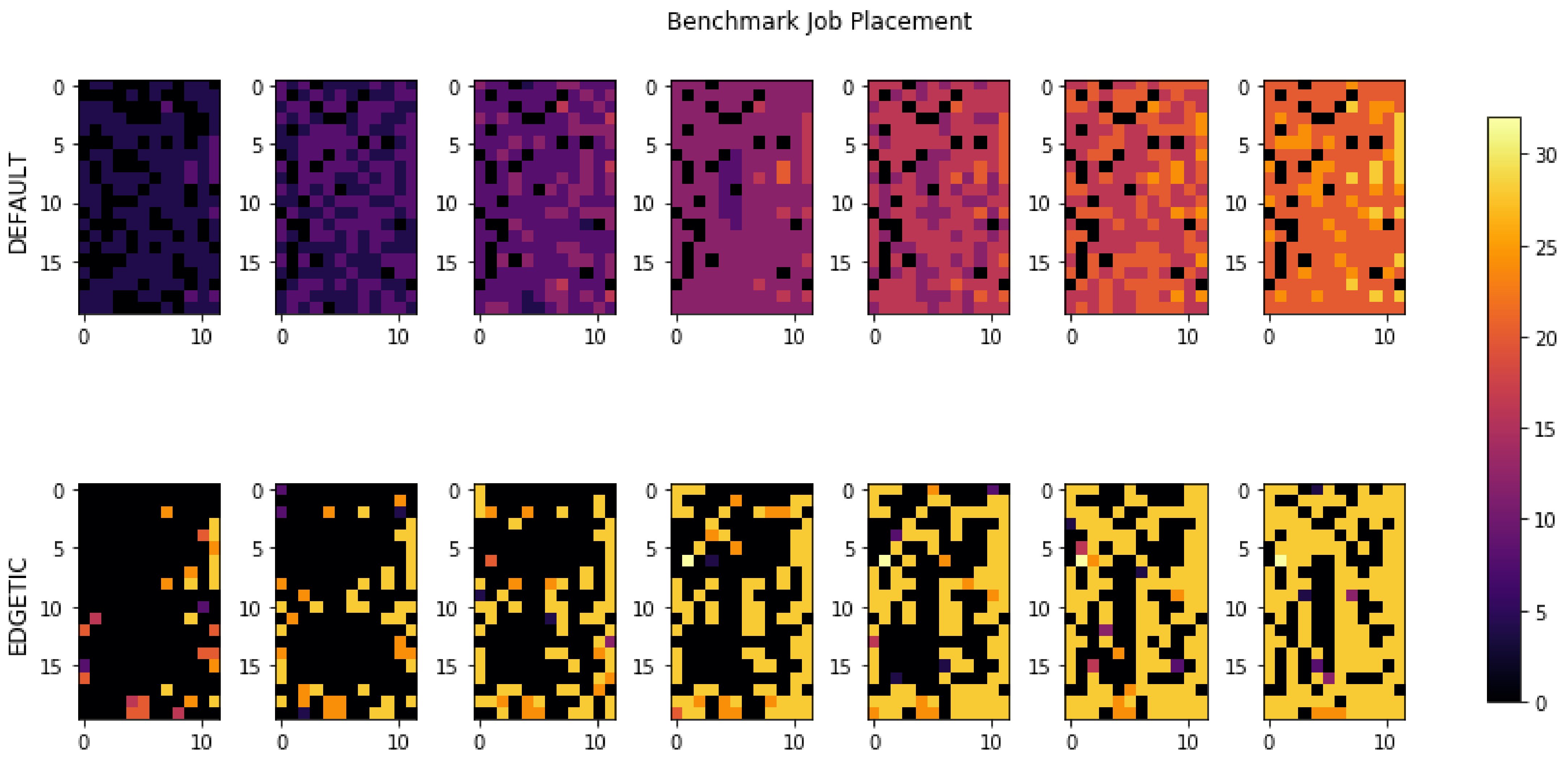


INITIAL EXPERIMENT

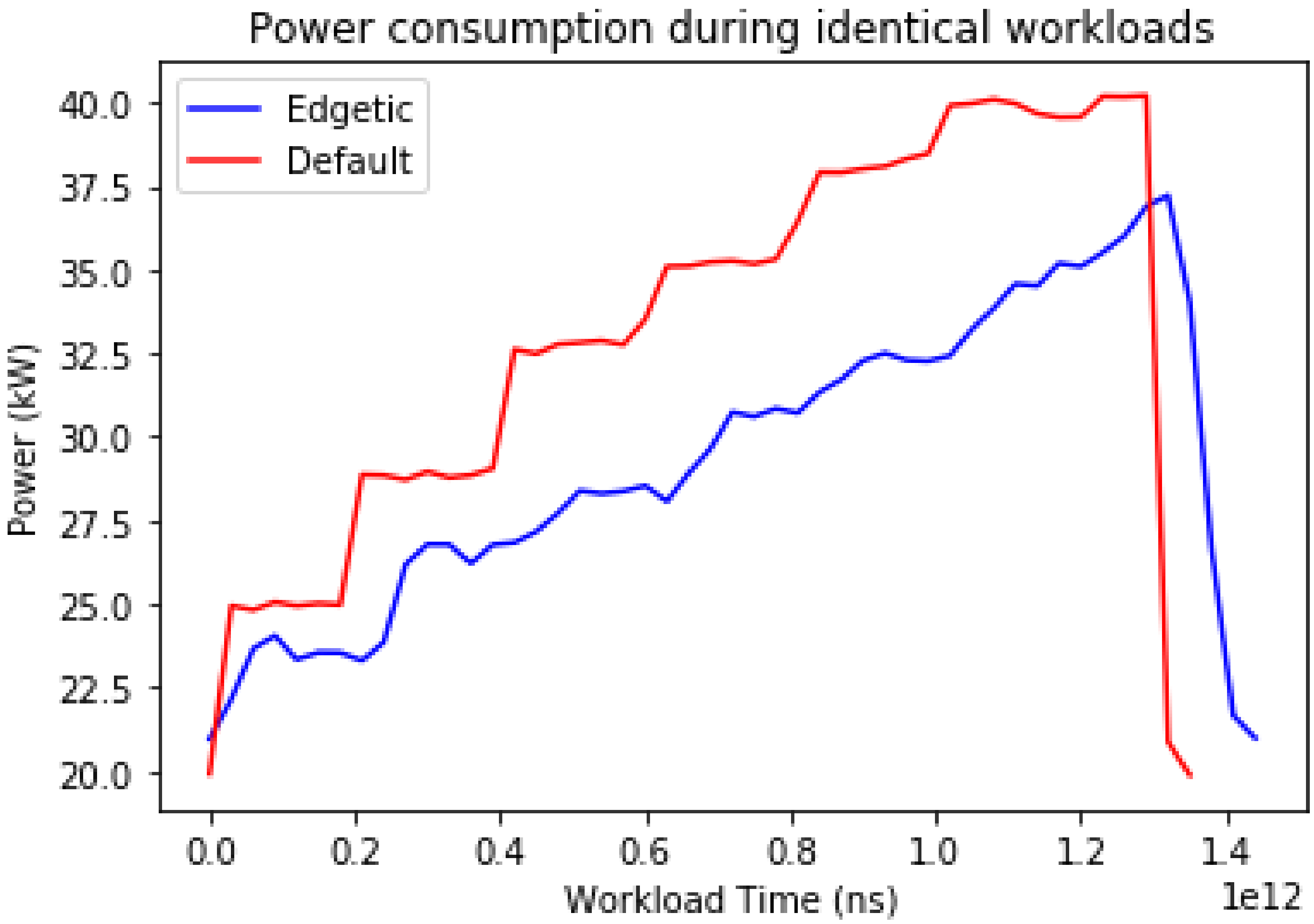
1. Set the BIOS on RISE machines
2. Create 215 node Kubernetes cluster
3. Deploy energy-aware scheduler
4. Submit tests to Workload Generator
5. Deploy replayable workloads
6. "Fill up" cluster to a given level of utilisation
7. Measure power consumption
8. Measure performance of each task
9. Record results
10. Let cluster cool then move to next test

UTILISATION AT DIFFERENT TIME PERIODS

UMEÅ UNIVERSITY



RESULTS


Power consumption during identical workloads
Edgetic
Default
40.0
37.5
35.0
32.5
30.0
27.5
25.0
22.5
20.0
Power (kW)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e12
Workload Time (ns)

RESULTS IN SUMMARY

215 nodes

OCP Hardware

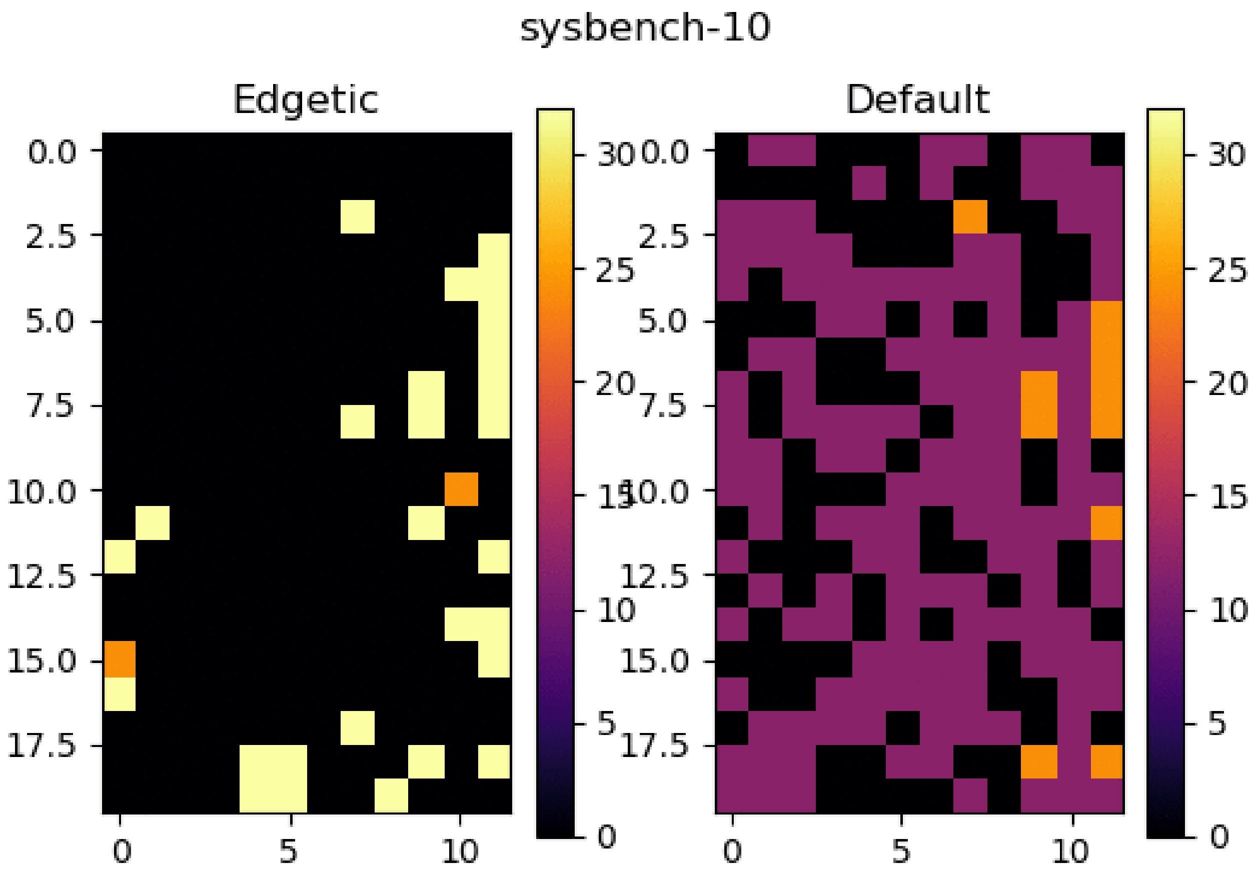
Variety of workloads

Kubernetes containers

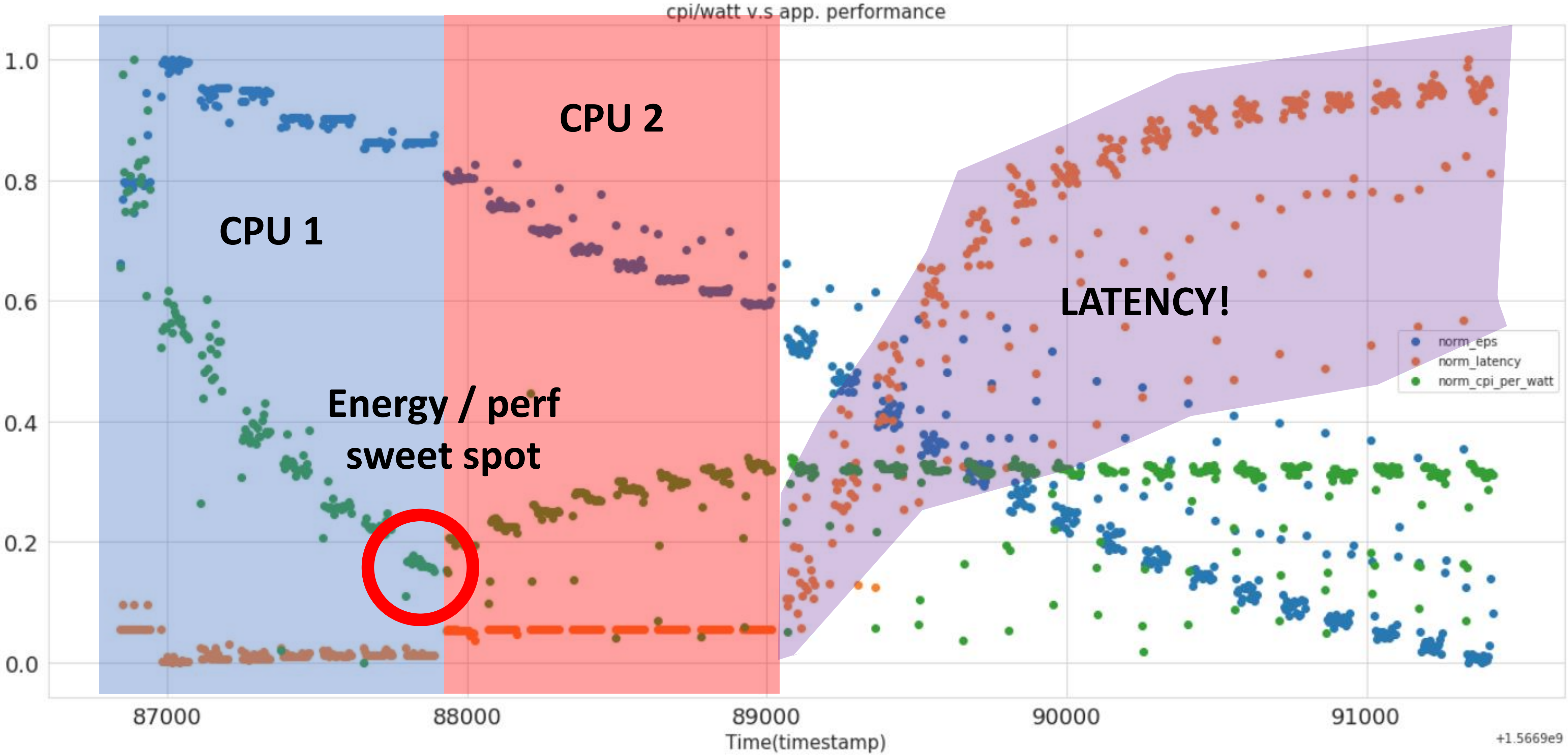
No prior workload knowledge

10-20% power savings

12 terabytes of telemetry data



Overhead of Scheduler: 10ms per incoming workload


cpi/watt v.s app. performance
CPU 1
CPU 2
LATENCY!
Energy / perf
sweet spot
1.0
0.8
0.6
0.4
0.2
0.0
87000
88000
89000
90000
91000
Time(timestamp)
+1.5669e9
norm_eps
norm_latency
norm_cpi_per_watt

THANKS - FEEL FREE TO GET IN TOUCH!

paul.townend @ umu.se

OpenNebula

SUSE

RI. SE Research Institutes of Sweden

Phoenix-RTOS

cmcc

UNIVERSIDAD COMPLUTENSE MADRID

acisa

cetic
Your Connection to ICT Research

ikerlan

Natur 4.0
Sensing Biodiversity

UMEÅ UNIVERSITY

ATENDE INDUSTRIES