A journey of cloud FPGAs on converged data centres

—

Dr. Dionysios Diamantopoulos
Research Staff Member
Cloud FPGAs & Tape Group
Cloud & AI Systems Research Department
IBM Research Europe

did@zurich.ibm.com

Hipeac Workshop 20-22 June 2022, Budapest, Hungary
EVEREST: Design and Programming High-performance, distributed, reconfigurable and heterogeneous platforms for extreme-scale analytics
https://www.hipeac.net/2022/budapest/

This work is partially funded by the EU Horizon 2020 Programme under grant agreement No 957269 (EVEREST).

IBM

IBM Legal Disclaimer

This content was provided for informational purposes only. The opinions and insights discussed are those of the presenter and guests and do not necessarily represent those of the IBM Corporation.

Nothing contained in these materials or the products discussed is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers, or altering the terms and conditions of any agreement you have with IBM.

The information presented is not intended to imply that any actions taken by you will result in any specific result or benefit and should not be relied on in making a purchasing decision. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal contact of any party.

All product plans, directions and intent are subject to change or withdrawal without notice. References to IBM products, programs or services do not imply that they will be available in all countries in which IBM operates. IBM, the IBM logo, and other IBM products and services are trademarks of the International Business Machines Corporation, in the United States, other countries or both. Other company, product, or services names may be trademarks or services marks of others.

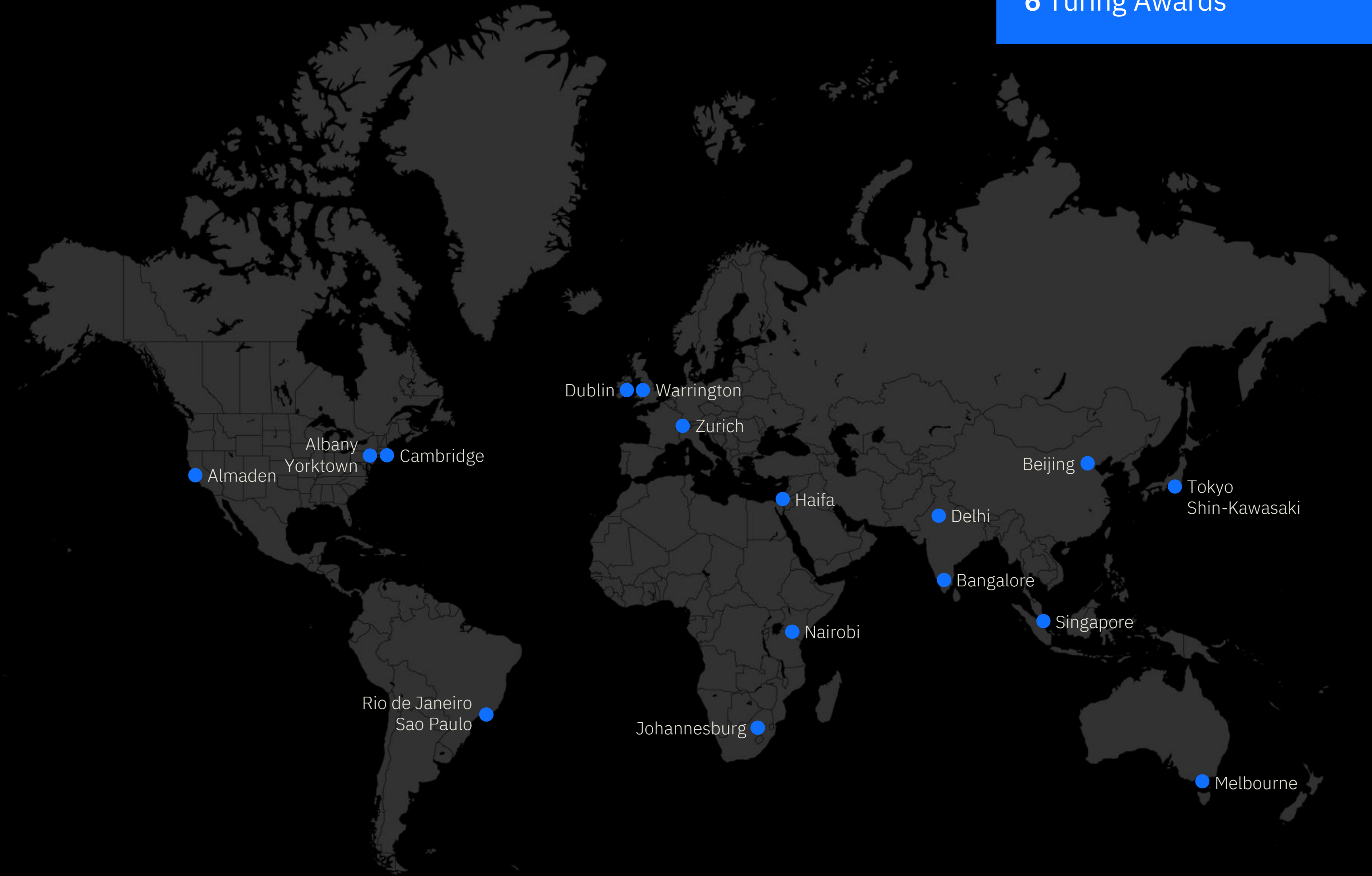
For copyright and trademark information go to: http://www.ibm.com/legal/us/en/copytrade.shtml

3000
Researchers

19
Locations

6
Continents

6 Nobel Laureates
10 Medals of Technology
5 National Medals of Science
6 Turing Awards



IBM Research – Zurich

Established in 1956
45+ different nationalities
Open Collaboration:

- Horizon2020: 50+ funded projects and 500+ partners

Two Nobel Prizes:

- 1986: Nobel Prize in Physics for the invention of the scanning tunneling microscope by Heinrich Rohrer and Gerd K. Binnig
- 1987: Nobel Prize in Physics for the discovery of high-temperature superconductivity by K. Alex Müller and J. Georg Bednorz

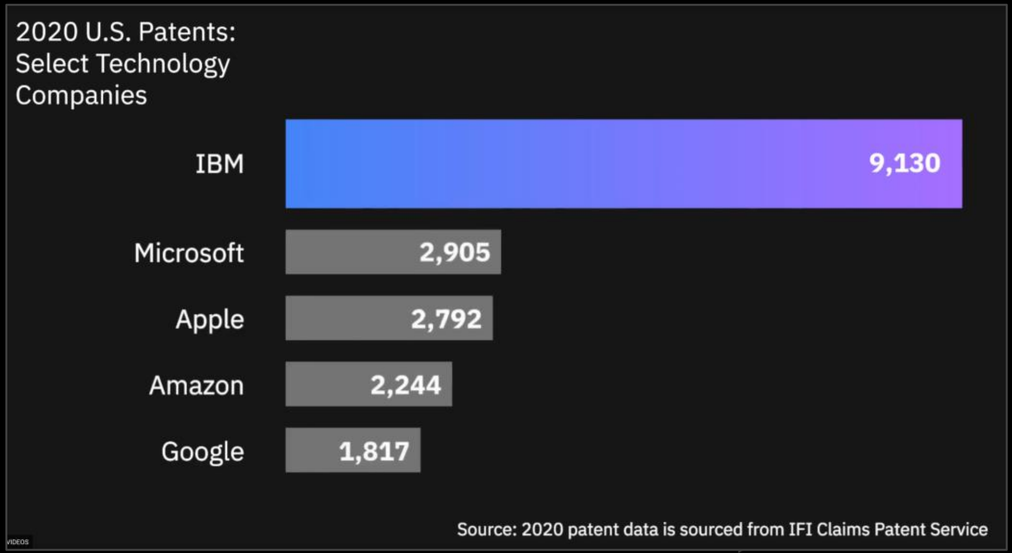
European Physical Society Historic Site

Binnig and Rohrer Nanotechnology Centre (Public Private Partnership with ETH Zürich and EMPA)

7 European Research Council Grants

We’re Inventing What’s Next in:

Hybrid Cloud
AI
Quantum
Science



IBM’s innovation: Topping the US patent list for 28 years running

https://www.ibm.com/blogs/research/2021/01/ibm-patent-leadership-2020/

From automated teller machine (ATM), speech recognition technology, DRAM to a novel way to search multilingual documents using NLP, 2300 AI patents !

Agenda

Context & Competitive Landscape

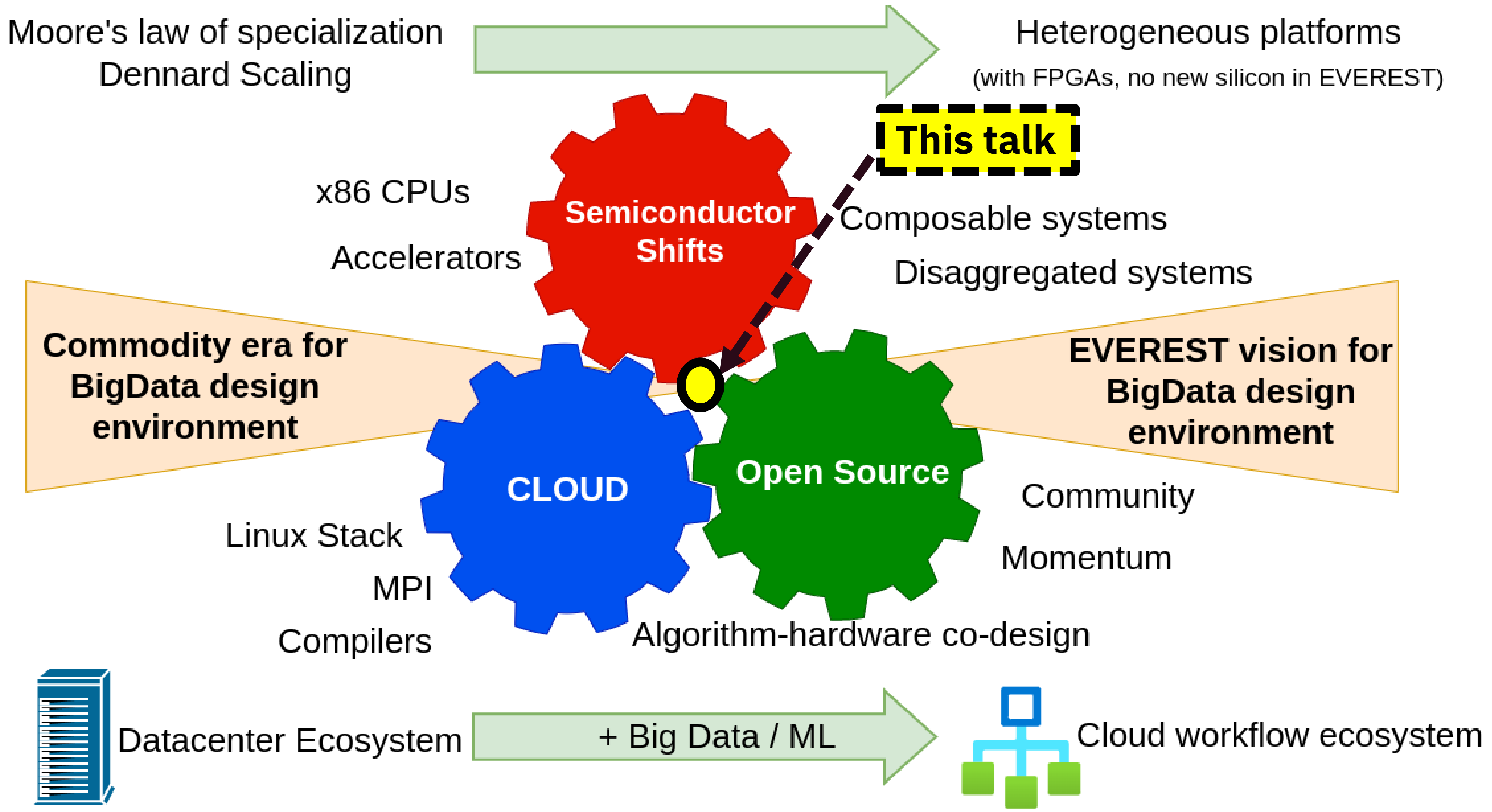
EVEREST use cases

cloudFPGA

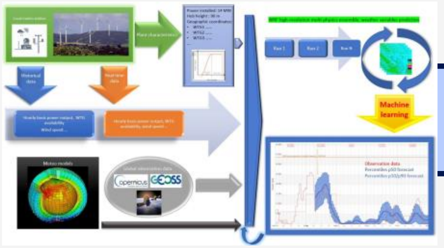
Converged Infra

Example on Traffic Simulation

Context & Competitive Landscape for EVEREST



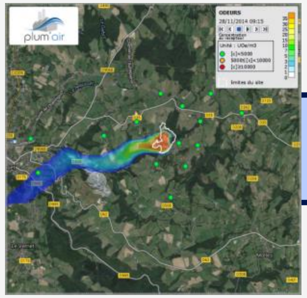
EVEREST Use Cases



Renewable energy production prediction

★ Improve **quality of the predictions**

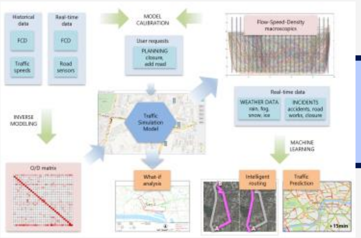
Weather prediction modelling (WRF)



Air-quality monitoring of industrial sites

★ Improve the **response time of predictions**

★ Accelerate kernels to execute more tests

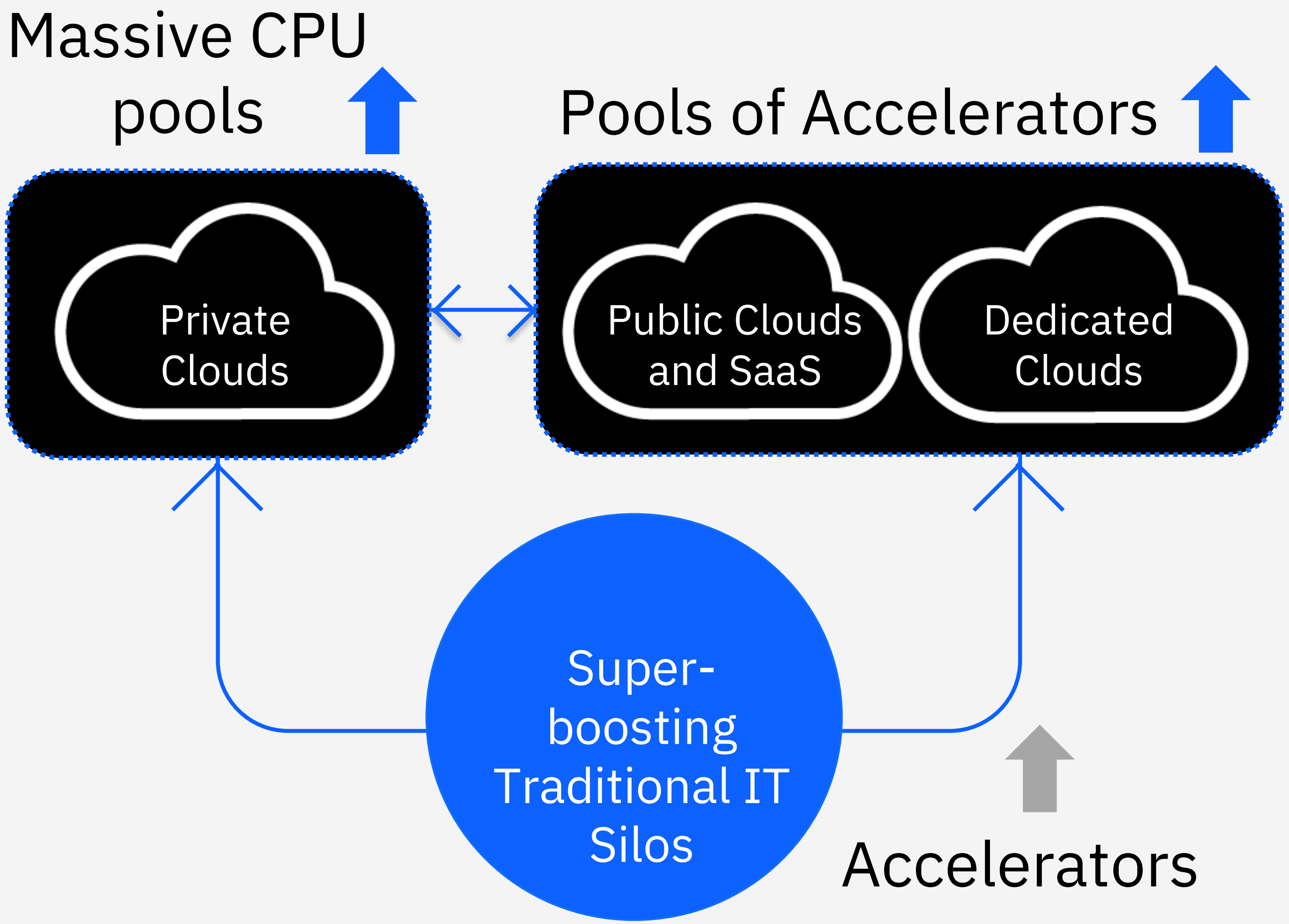


Traffic modeling for intelligent transportation

★ Improve the **overall performance of traffic simulation**

Accelerated computationally-intensive kernels + **Machine-learning** kernels

A Converged computing journey



Super-boosting is not enough, breaking silos is imperative …
Converged infra to the rescue

A journey of cloud FPGAs on converged data centres, D. Diamantopoulos, Hipeac 2022, Budapest, Hungary, 22 June 2022

An optimal strategy will address key challenges:

50% Reduce by 50% the development costs

10x Reduce programming efforts by one order of magnitude

Unleashes the full potential

30% energy cost

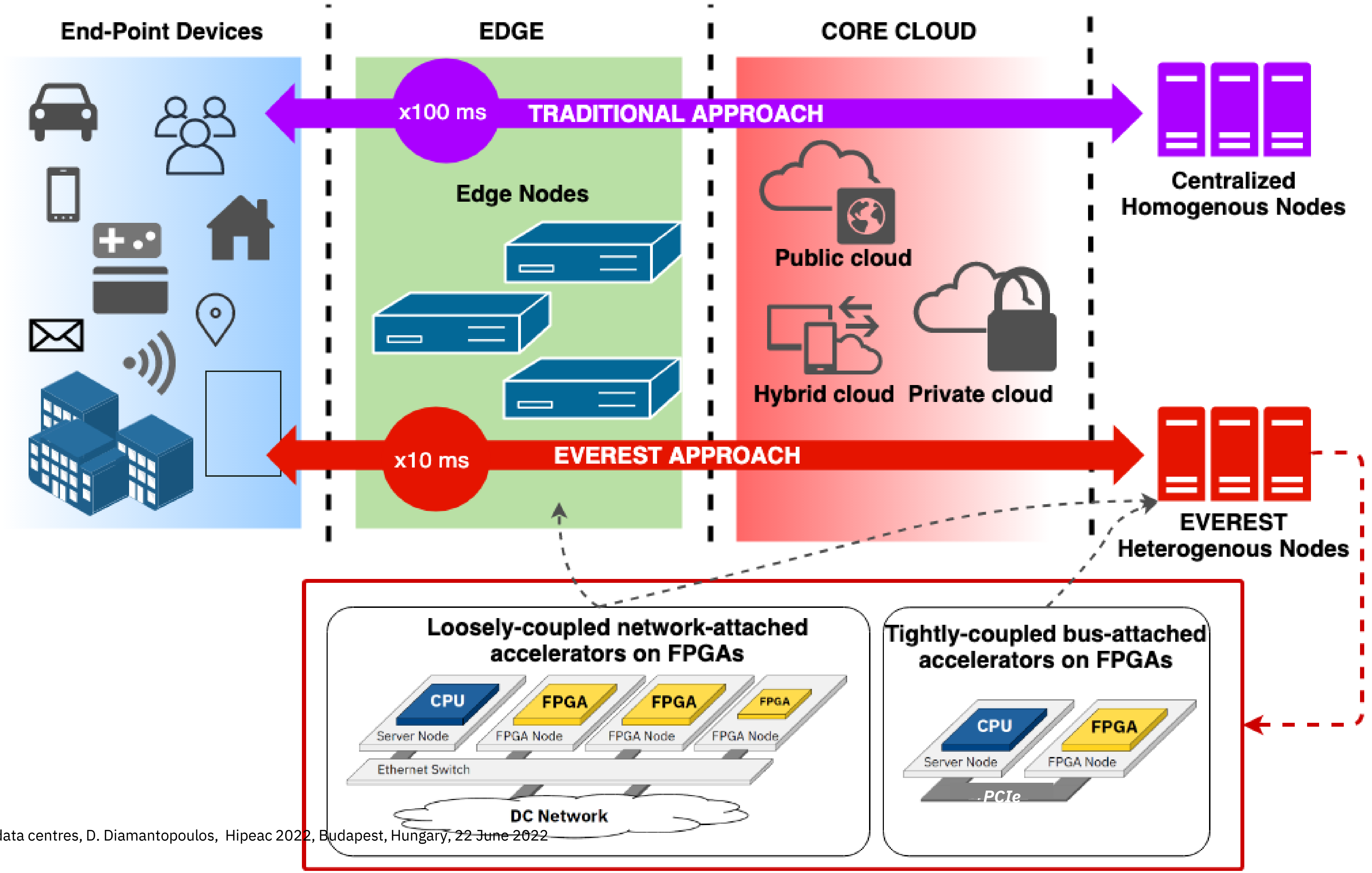
10x the performance of simulations for renewable energies prediction

2x the response time of the air-quality predictions

3x the overall performance of traffic model framework

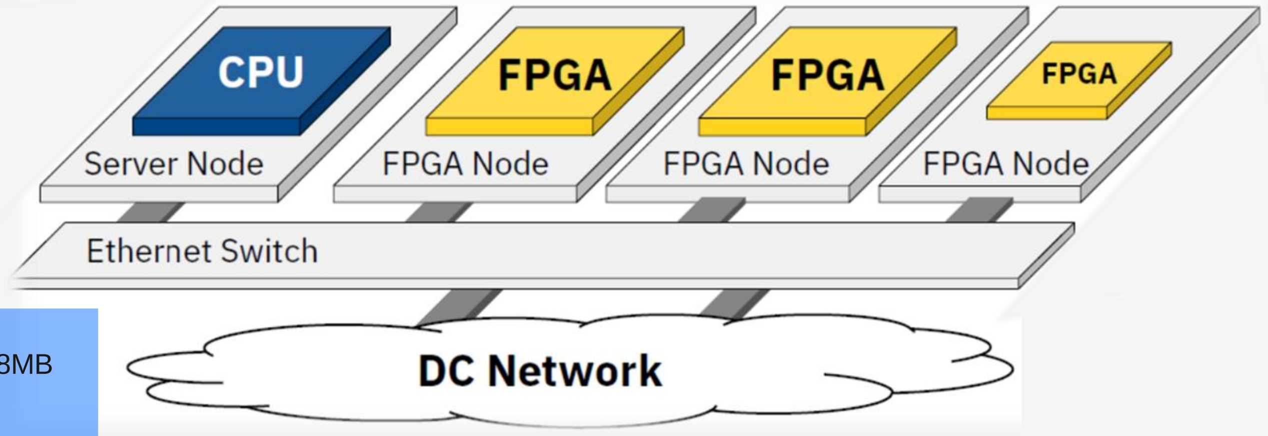
Computing continuum to enable cloud-to-edge integration

Cloud & bus-attached FPGAs at H2020 EVEREST Project

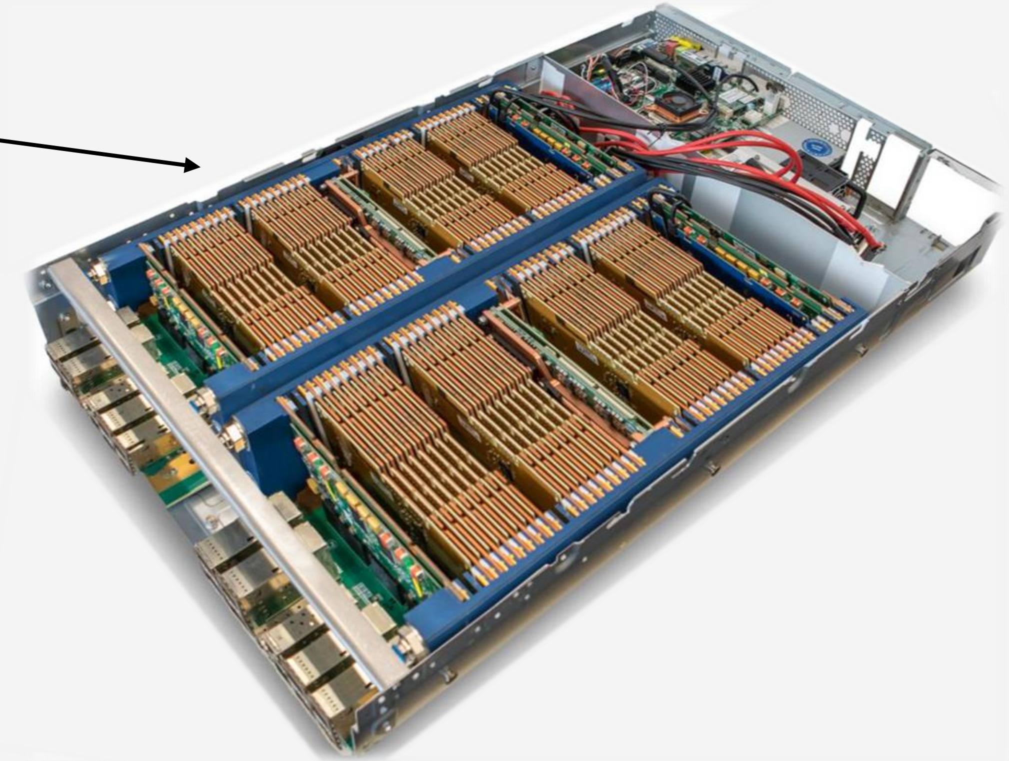


There's no AI without IA

cloudFPGA concept



IP Address: 10.10.1.9
DRAM: 8GB, BRAM: 38MB
CLBs:660.000,
DSPs: 2760



Highlights

- dense
  → chassis w/ 64 compute units
  → ~1000 FPGAs / rack
- integration of 1st level switch
  → full cross-sectional BW
  → low cost (cables / rack space)
- energy efficient
  → no SW/FW overhead
  → no CPU overhead
  → (hot) water cooling
- self-hosted / network-attached
  → bare-metal support
  → scalabl

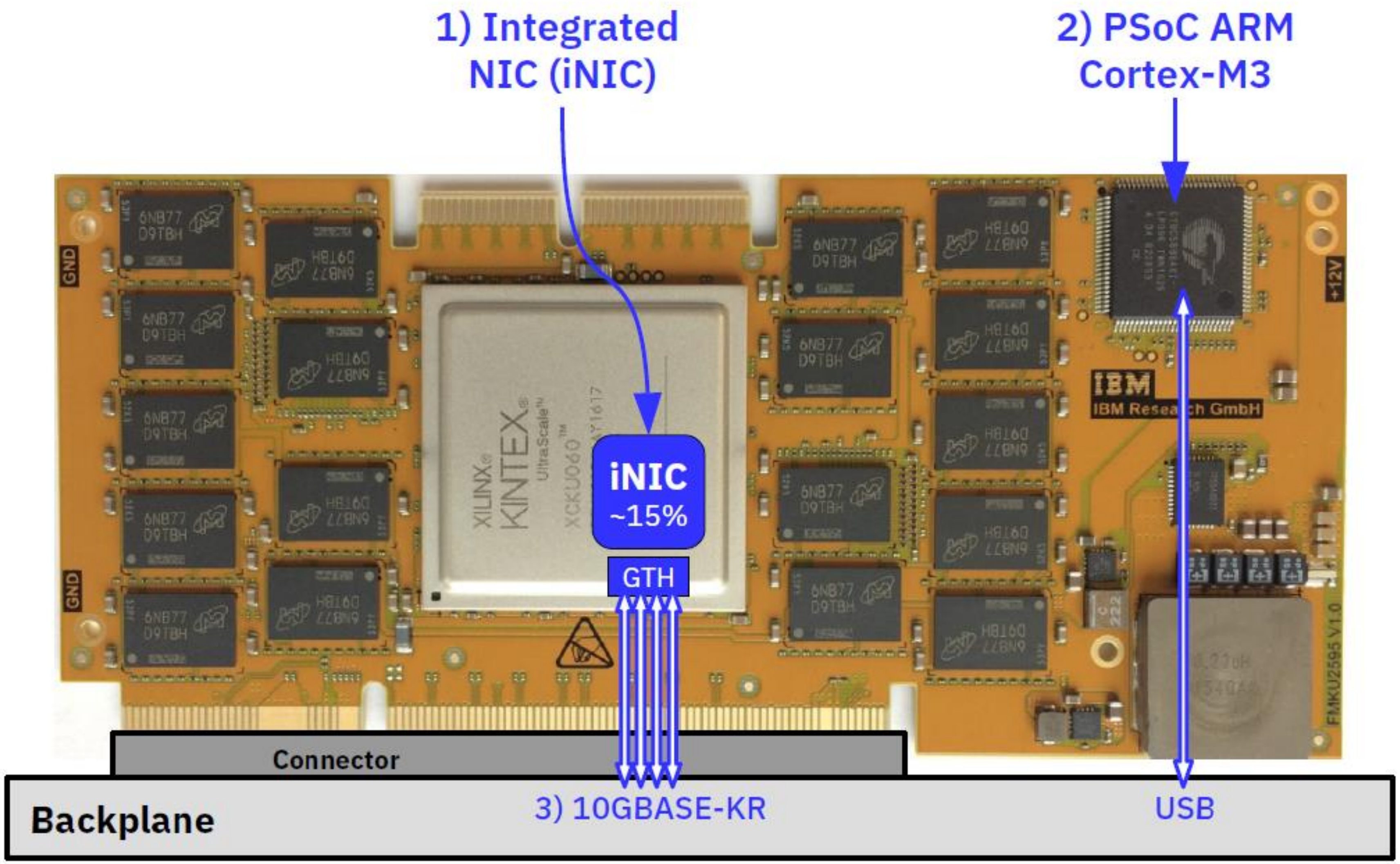
The FPGA becomes the node !

IP Address: 10.10.1.50
DRAM: 32GB, Cores: 4

Goal → Deploy FPGAs at large scale in data centers

 1-10s of thousands per DC

Standalone network-attached FPGA



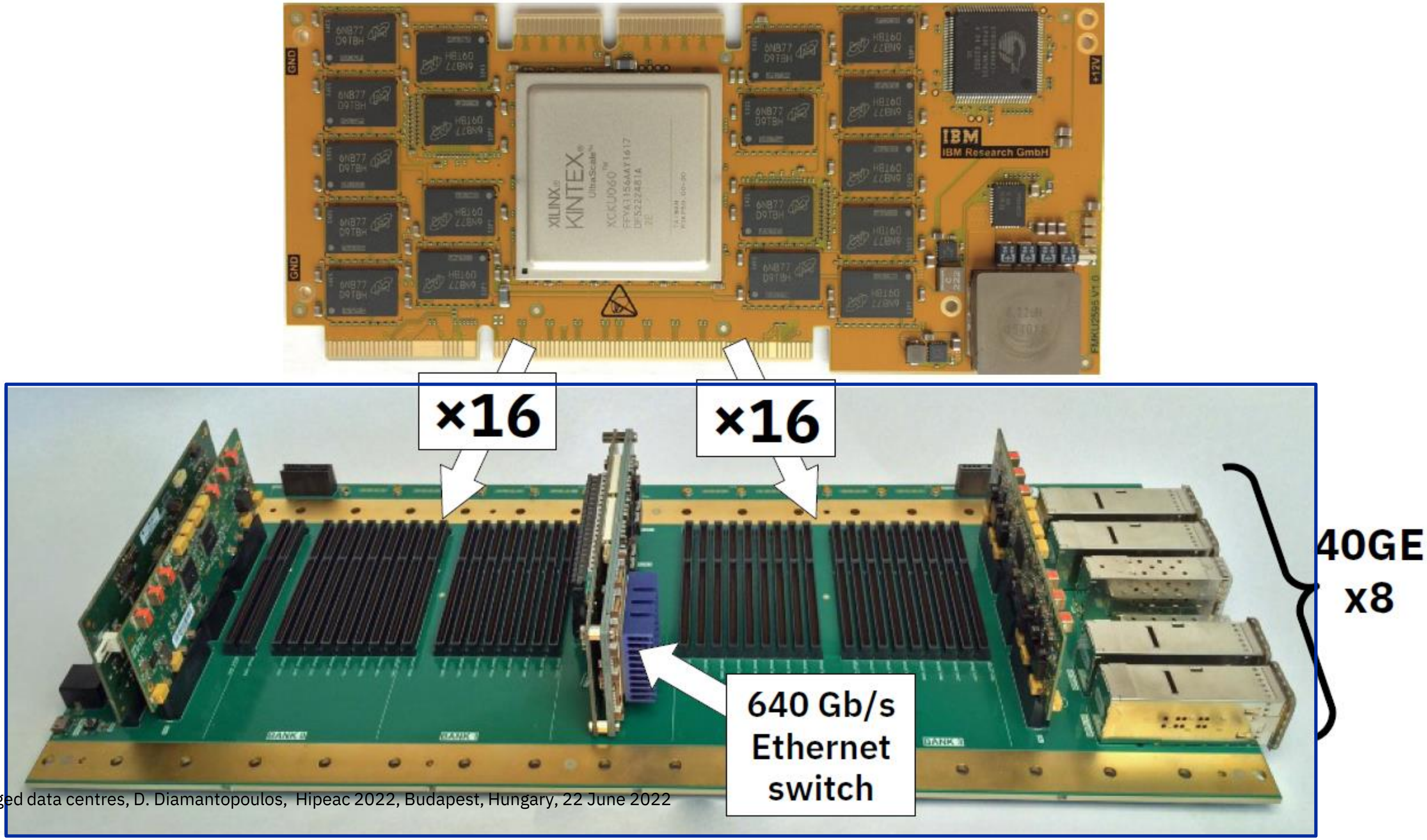
1. Replace PCIe I/F with integrated NIC (iNIC).
2. Turn FPGA card into a self-contained appliance.
3. Replace transceivers w/ backplane connectivity.

One carrier sled = 32 FPGA modules

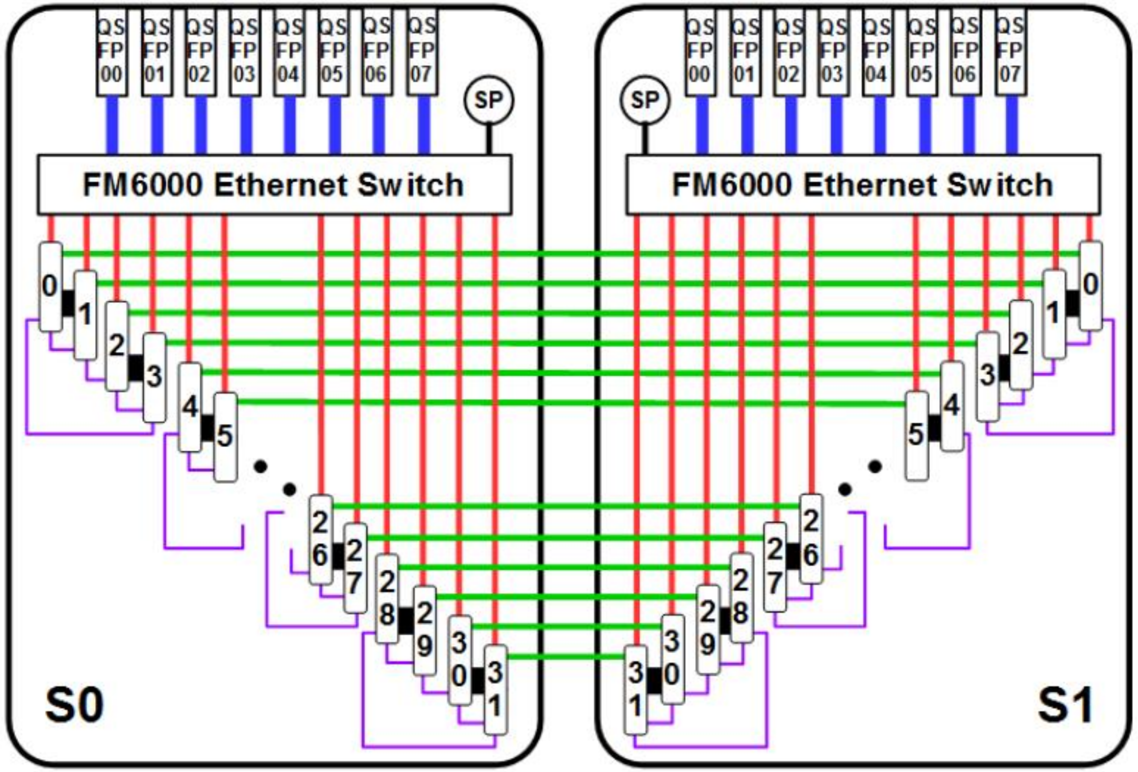
1. Our first FPGA module uses a Xilinx Kintex Ultrascale KU060
   - A mid-range FPGA with high performance/price and low wattage

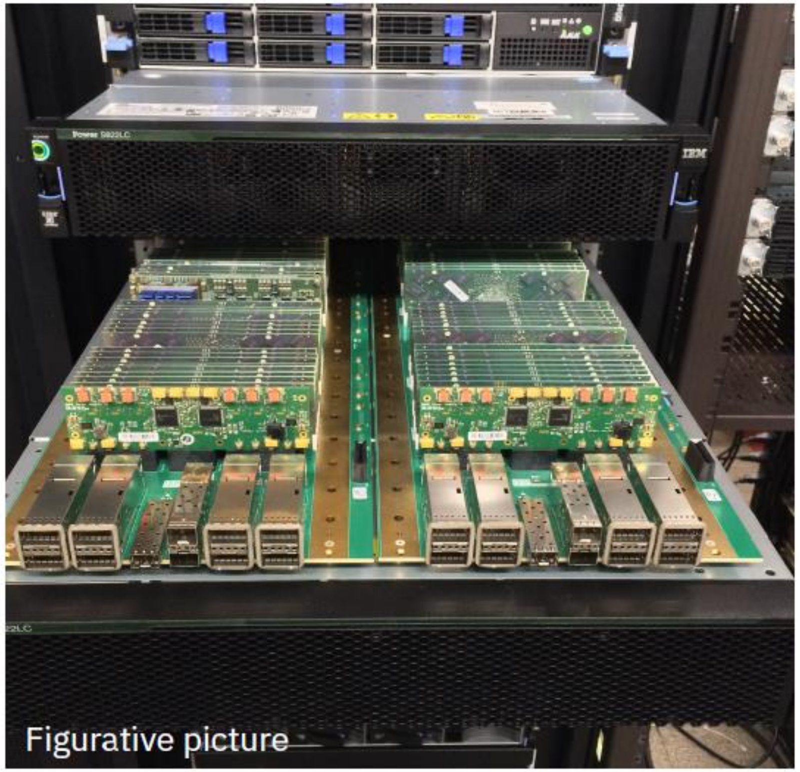
One carrier sled = 32 FPGA modules

1. Our first FPGA module uses a Xilinx Kintex Ultrascale KU060
   - A mid-range FPGA with high performance/price and low wattage



Two carrier sleds per chassis = 64 FPGAs



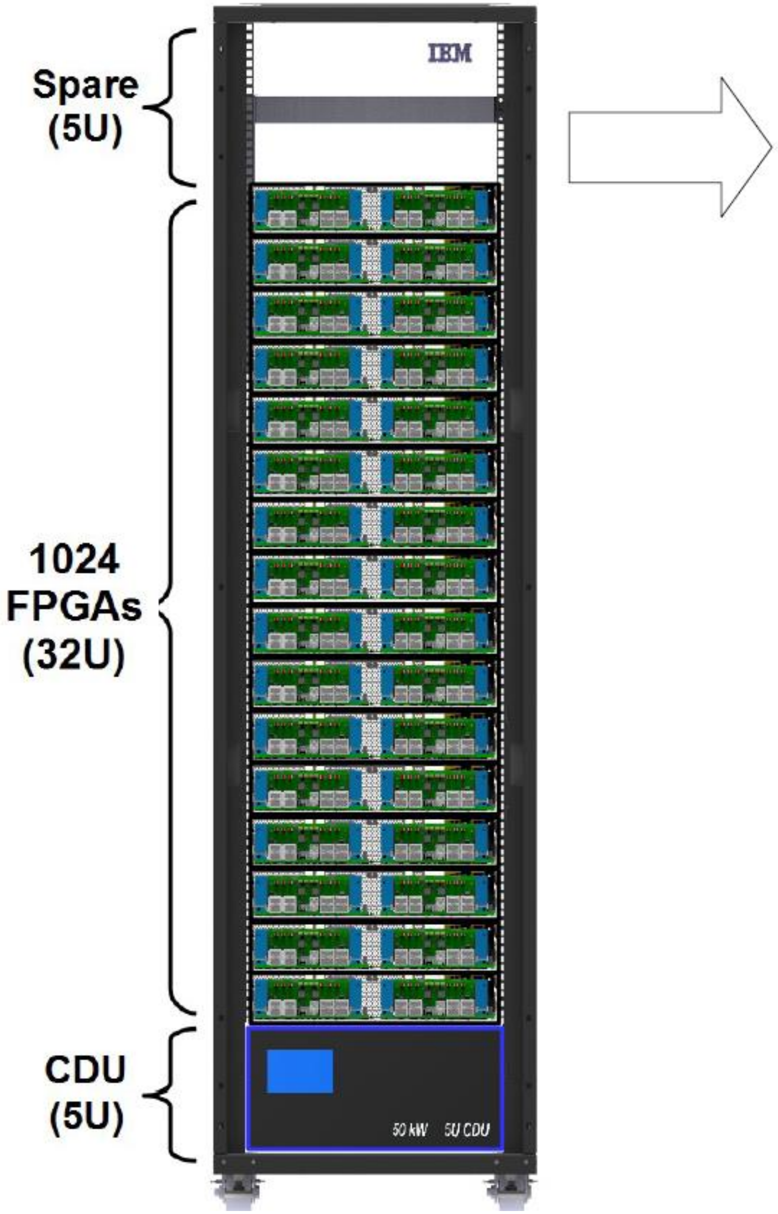


Figurative picture

```
Legend (per slice):
  [==]  x8 40GbE up links                    (320 Gb/s)
  [--] x32 10GbE FPGA-to-Switch links (320 Gb/s)
  [--] x32 10GbE redundant links
  [--] x32 10GbE FPGA-to-FPGA links
  [██] x16 PCIe x8 Gen3
   SP   x1 Service Processor
```

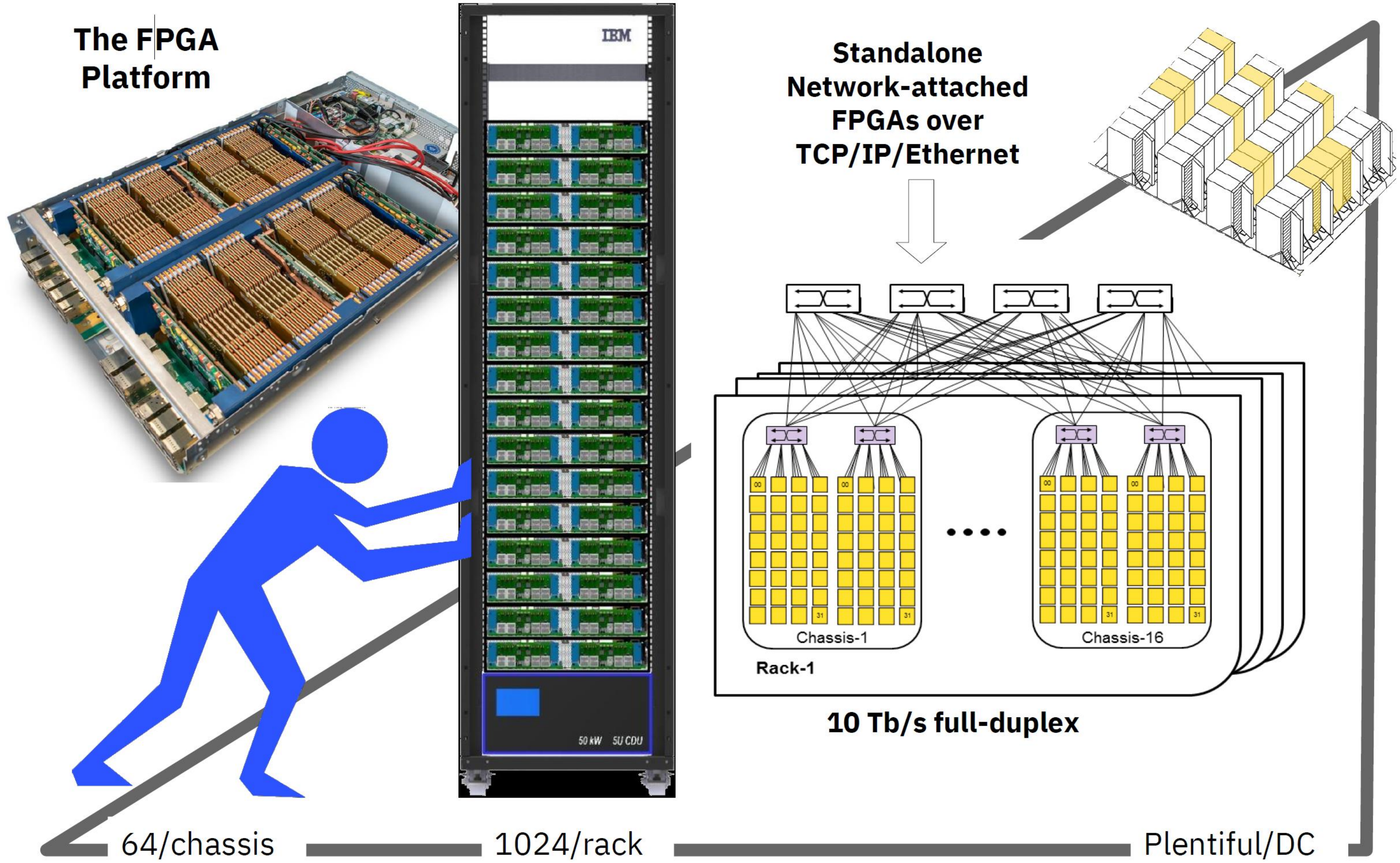
Balanced (i.e. no over-subscription) between the north and south links of the Ethernet switch

Sixteen chassis per rack = 1024 FPGAs

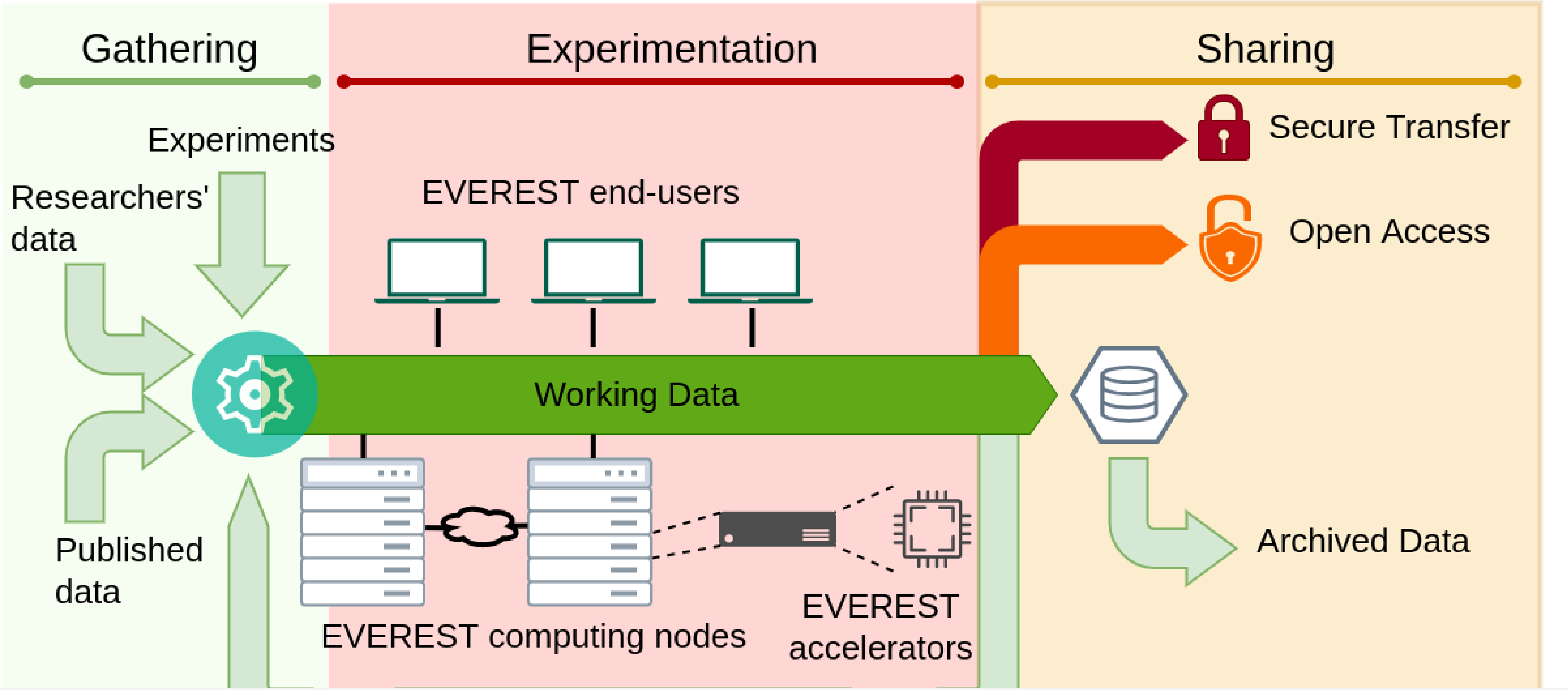


1024 FPGAs → 2.8M DSPs,
2×10^{15} Fixed-Point Multiply-Accumulates/s
10 Tb/s bi-sec. Bw – 16 TB DDR4 – 40 kW max.

cloudFPGA


The FPGA Platform
IBM
50 kW 5U CDU
Standalone Network-attached FPGAs over TCP/IP/Ethernet
Chassis-1
Chassis-16
Rack-1
10 Tb/s full-duplex
64/chassis
1024/rack
Plentiful/DC

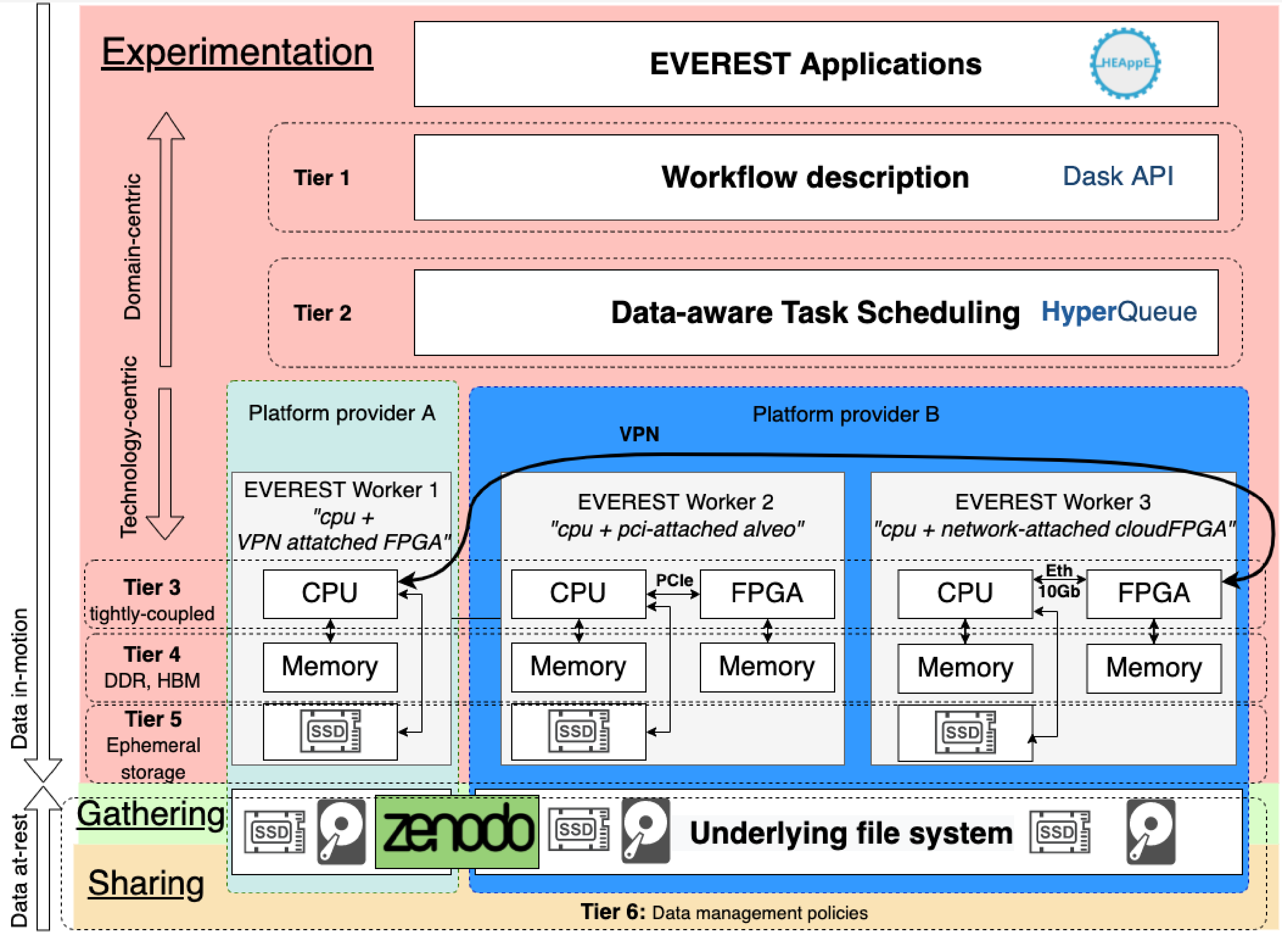
EVEREST Data lifetime



Key Components

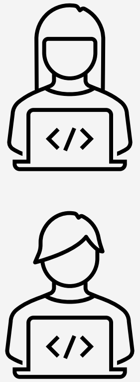
Building the stack…

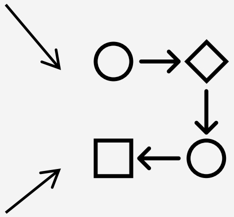
Converged FPGAs



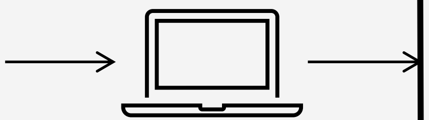
Use case: Traffic simulation on converged EVEREST platform

EVEREST users





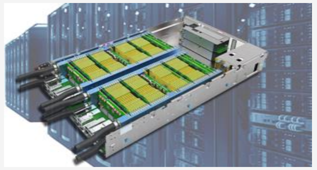
Traffic simulation workflow

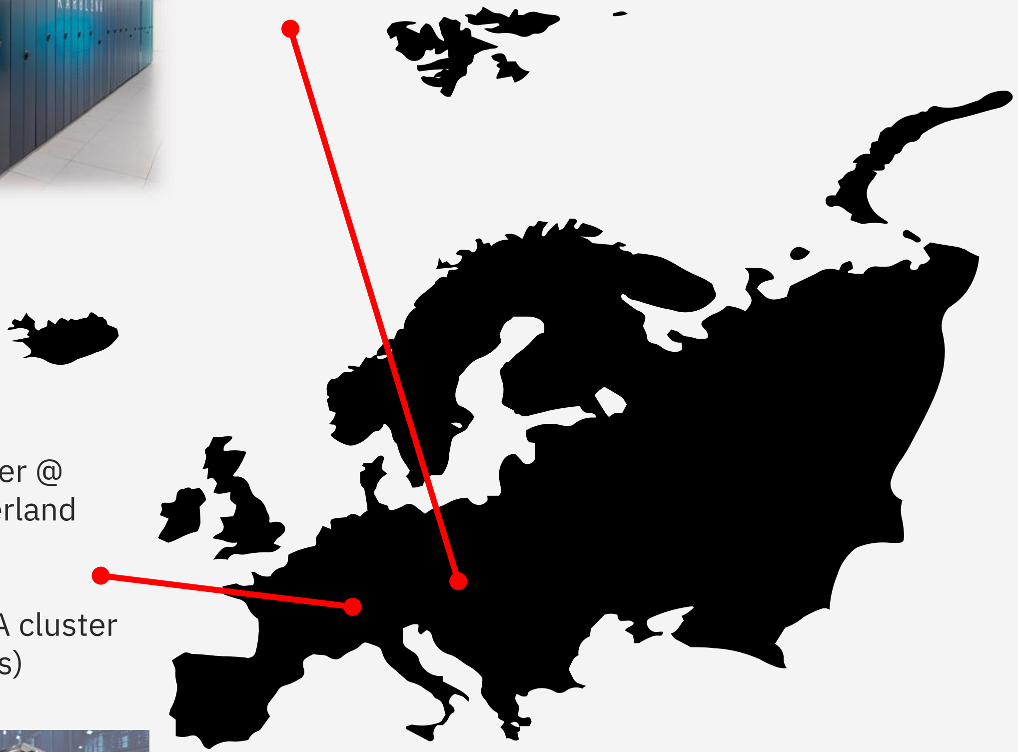


https://portal.lexis.tech/

IT4I Datacenter @ Ostrava, Czechia
- x86 nodes
- Lexis portal

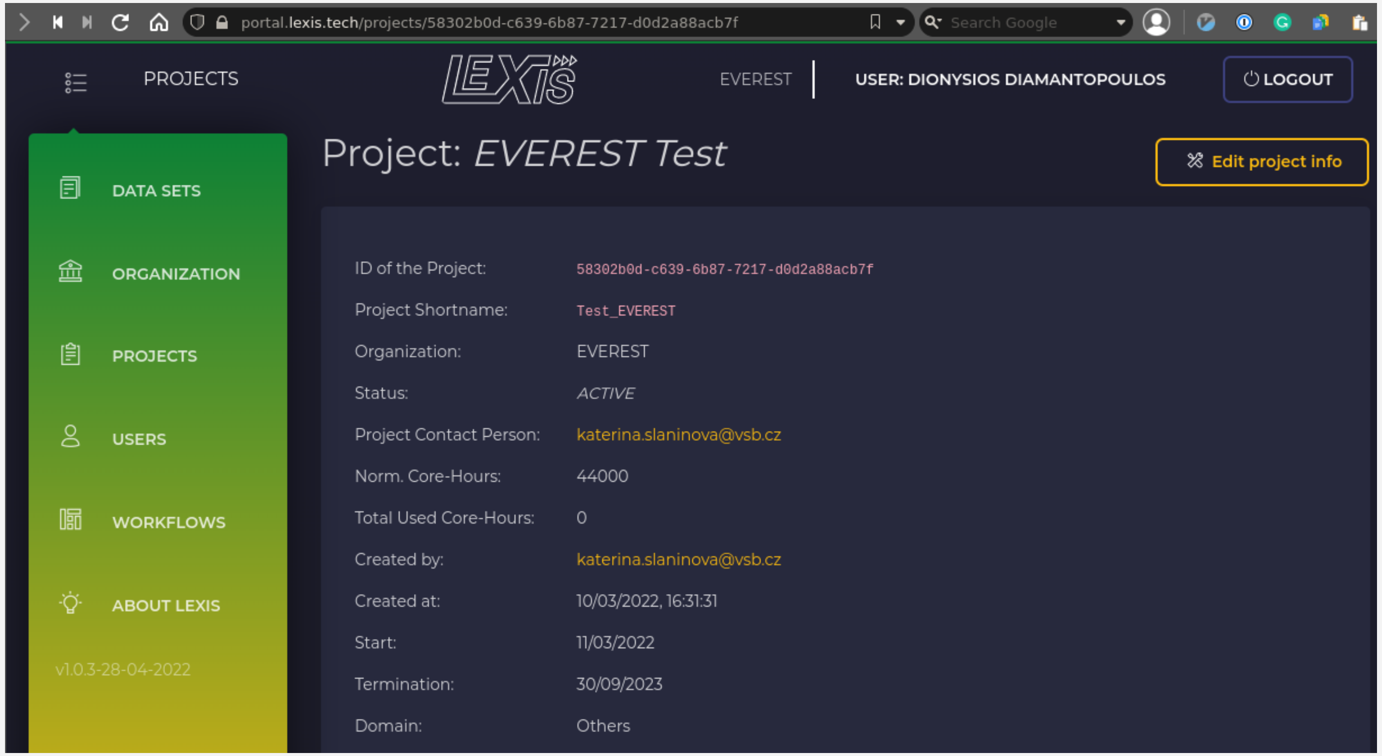
IBM Datacenter @ Zurich, Switzerland
- x86 nodes
- cloudFPGA
- Alveo-FPGA cluster (in-progress)



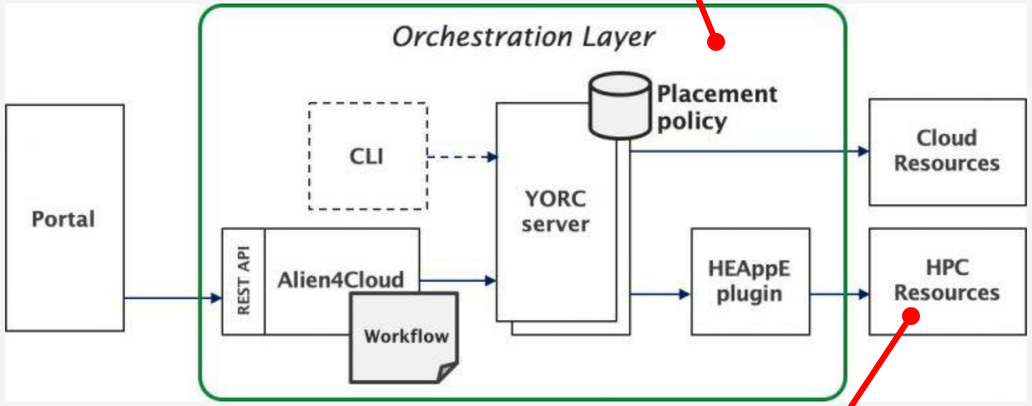


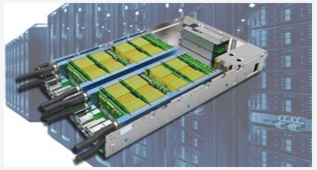
Use case: Traffic simulation on converged EVEREST platform

EVEREST project within LEXIS portal



IT4I Datacenter @ Ostrava, Czechia
- x86 nodes
- LEXIS portal

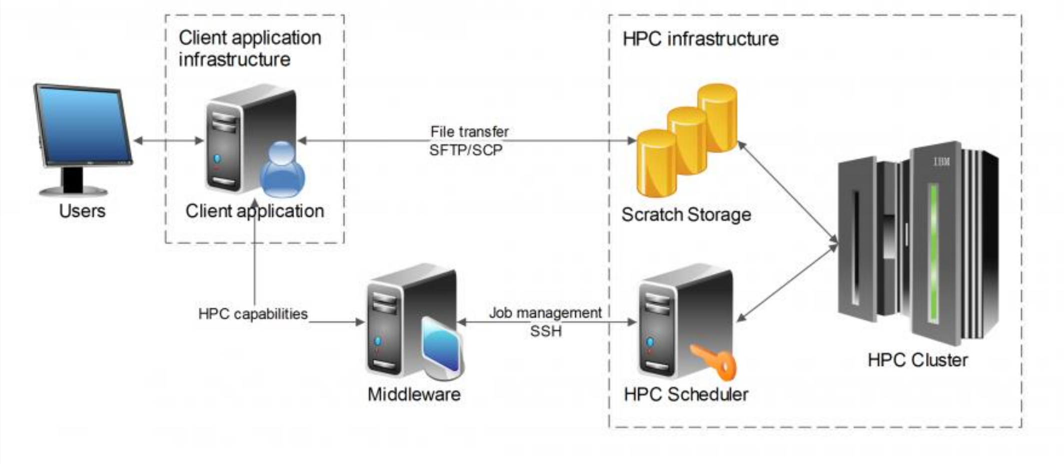


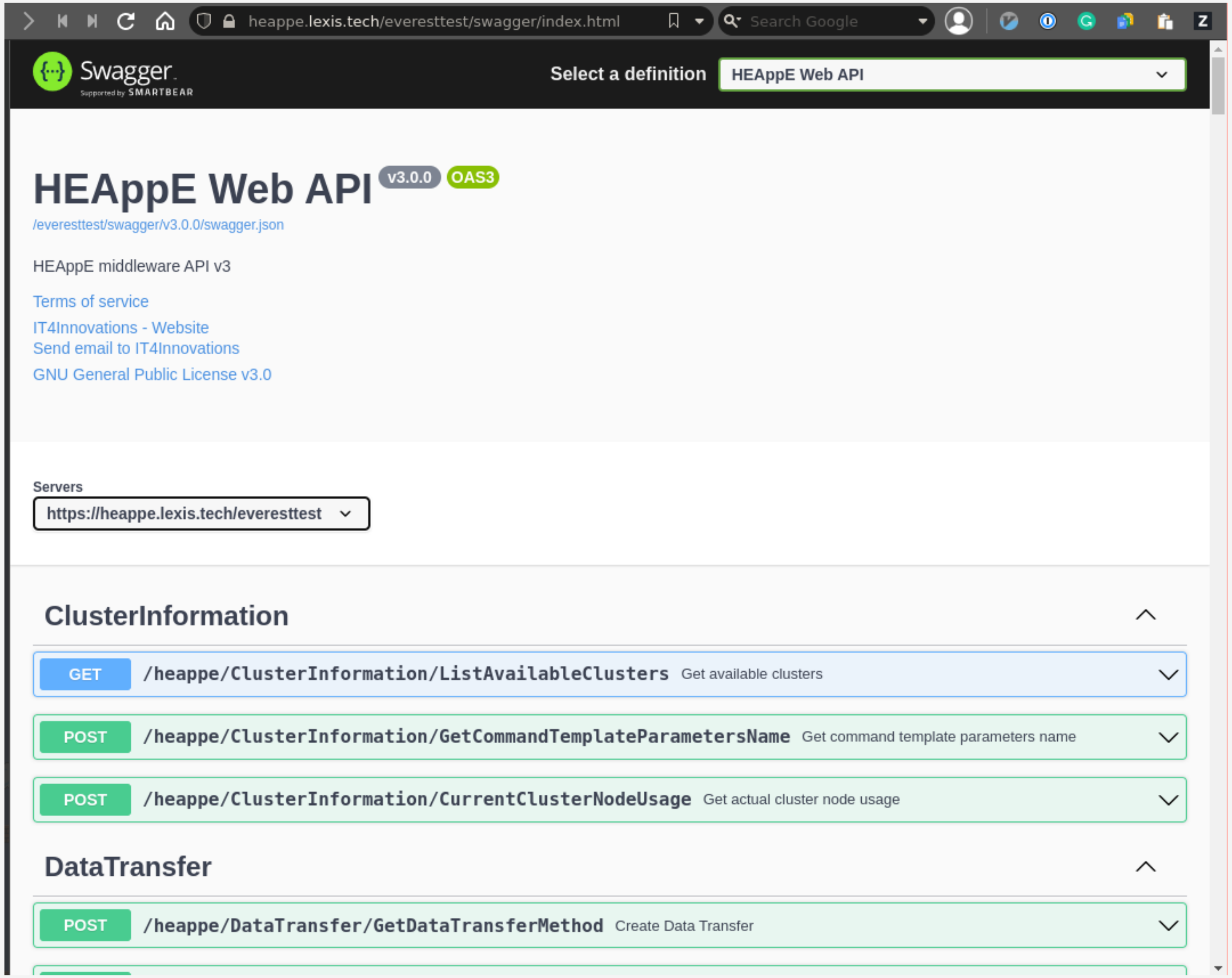


IBM Datacenter @ Zurich, Switzerland
- x86 nodes
- cloudFPGA
- Alveo-FPGA cluster (in-progress)

Use case: Traffic simulation on converged EVEREST platform

EVEREST HEAppE API

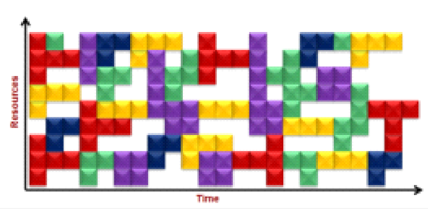




HPC requires more than just Cloud infrastructure

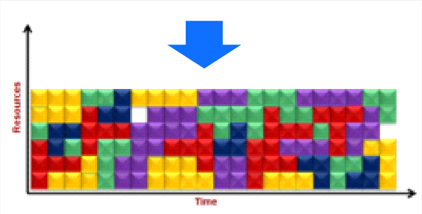
Without scheduling, workloads are dispatched in a haphazard fashion leading to sub-optimal resource use and longer execution times

Without Scheduling

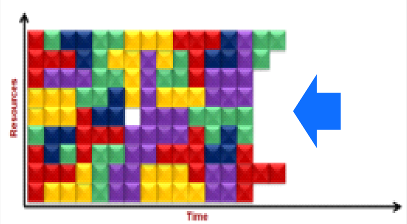


With sophisticated scheduling, we use resources more efficiently and can optimize for high-utilization, better performance, or both

Fewer Resources



Faster Results



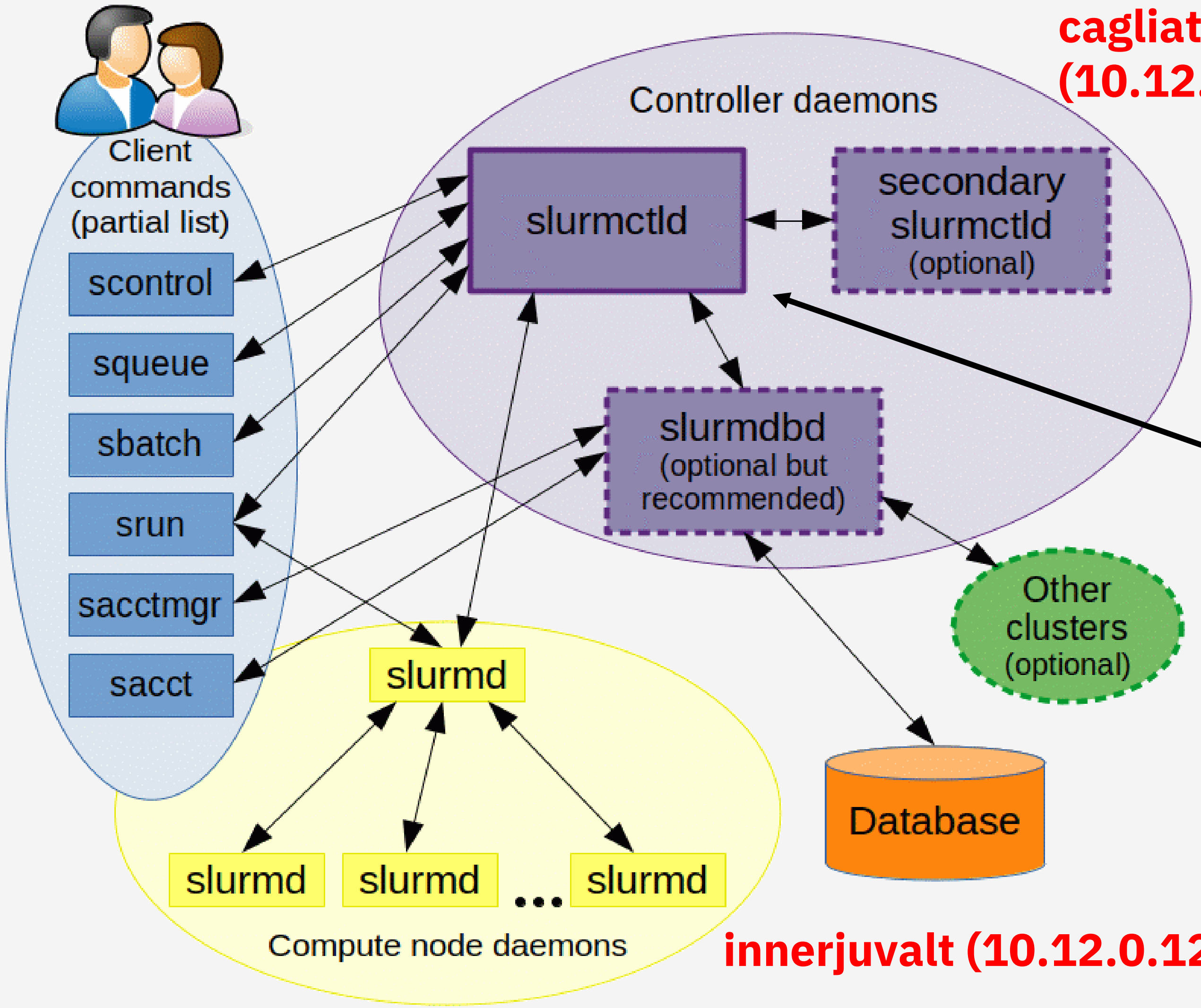
Use case: Traffic simulation on converged EVEREST platform

EVEREST user

SLURM setup at IBM:

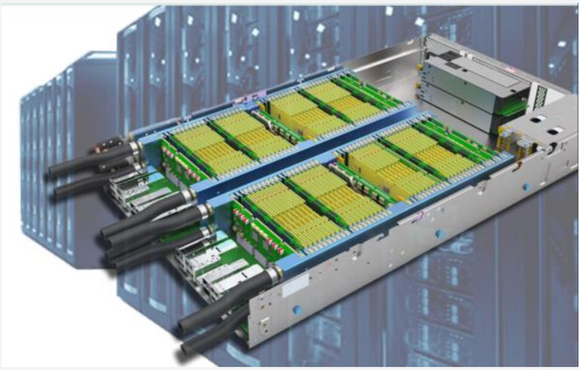
- SLURM controller & daemon up and running
- Connection to cloudFPGA
- NFS shared storage
- Exposure of a task to workflows out of IBM VPN (HEAppE, SOCKS5)

cagliatscha (10.12.0.11)



innerjuvalt (10.12.0.12)

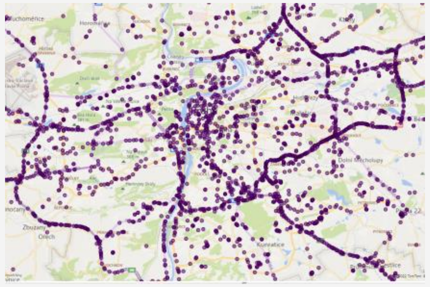
cloudFPGA



Use case: Traffic simulation on converged EVEREST platform

Floating car data (FCD)

Sygic

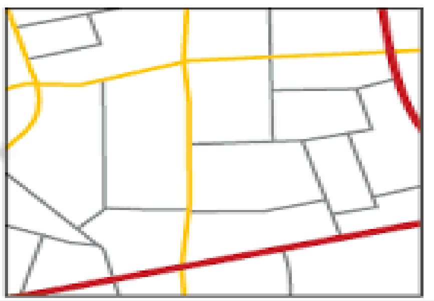


Origin-Destination matrix (ODM)

Telekom

Road network

Sygic



Traffic computation eco-system

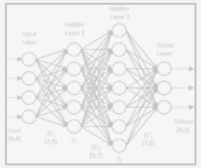
Map matching processor

Traffic simulator

Prediction model training

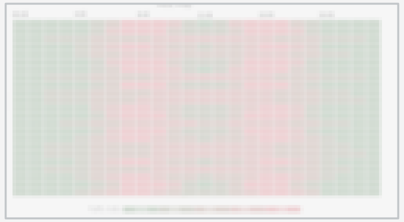
Prediction service

Intelligent routing

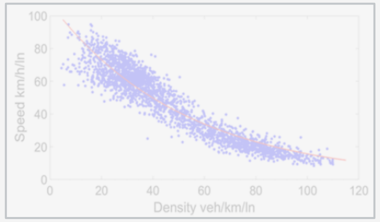


Training data
Prediction model

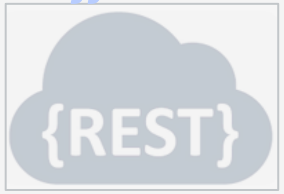
Speed profiles



3D traffic model



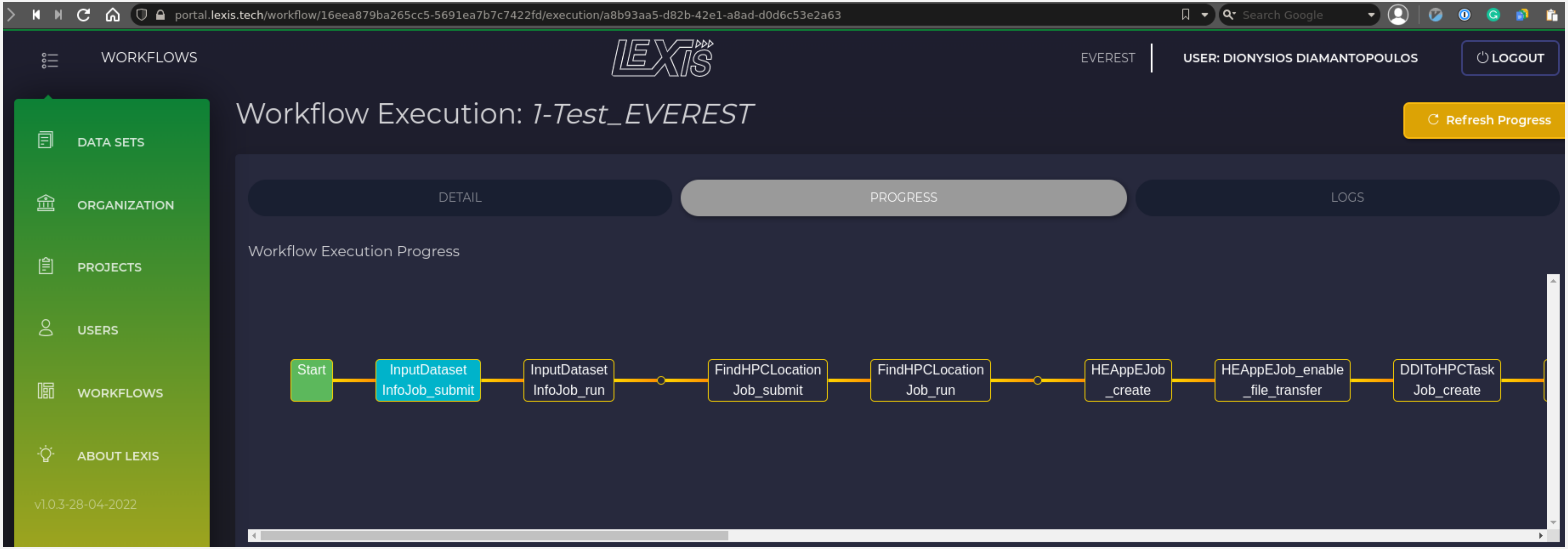
Traffic services



Analyst
Planner
Driver

Use case: Traffic simulation on converged EVEREST platform

Create an EVEREST workflow for traffic simulation inside LEXIS

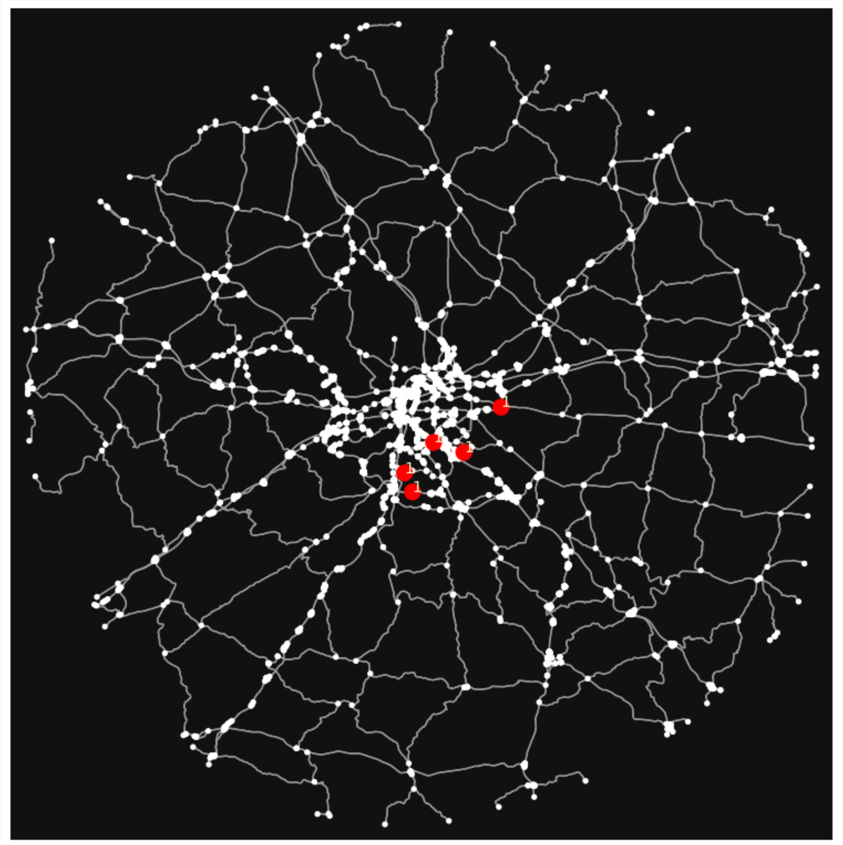


Use case: Traffic simulation on converged EVEREST platform

Traffic simulation initiated by LEXIS workflow is being executed at IBM SLURM computing node.

```
--------------------------------------------------------------------------
No OpenFabrics connection schemes reported that they were able to be
used on a specific port.  As such, the openib BTL (OpenFabrics
support) will be disabled for this port.

  Local host:           cagliatscha
  Local device:         mlx5_0
  Local port:           1
  CPCs attempted:       udcm
--------------------------------------------------------------------------
Loading data ... INPUT-od_matrix_10.parquet
Map file path: ./data/custom_8a777d0ce5460e8b94cbfc8b14dc8b3b.graphml
Loading network for 'custom_8a777d0ce5460e8b94cbfc8b14dc8b3b' from local map.
StepInfo(step=0, active=10)
StepInfo(step=1, active=2)
StepInfo(step=2, active=2)
StepInfo(step=3, active=9)
StepInfo(step=4, active=3)
StepInfo(step=5, active=7)
StepInfo(step=6, active=5)
StepInfo(step=7, active=2)
StepInfo(step=8, active=6)
StepInfo(step=9, active=5)
StepInfo(step=10, active=2)
StepInfo(step=11, active=6)
StepInfo(step=12, active=5)
StepInfo(step=13, active=6)
StepInfo(step=14, active=1)
StepInfo(step=15, active=6)
StepInfo(step=16, active=7)
StepInfo(step=17, active=3)
StepInfo(step=18, active=7)
StepInfo(step=19, active=1)
StepInfo(step=20, active=4)
```



Conclusion for EVEREST Converged infra:

Simplified cluster management

Design your own cluster with cloud FPGAs

Security & isolation

Onboard apps within HPC workflows

Native open-source experience

Integrated operational tools

Köszönöm ! - Thank you!

IBM

This work is partially funded by the EU Horizon 2020 Programme under grant agreement No 957269 (EVEREST).