Leveraging Cloud-native tools as the glue to seamlessly converge different computing worlds

ICCS contribution to H2020 project: EVOLVE

1st Open Annual Workshop on Future ICT

Achilleas Tzenetopoulos

25/5/2022

Supervisor:

Dimitrios Soudris

Team Members:

Sotirios Xydis, George Lentaris, Dimosthenis Masouros, Konstantina Koliogeorgi, Konstantinos Iliakis, Dimitrios Danopoulos, Ioannis Stamoulias, Ioannis Oroutzoglou, Aggelos Ferikoglou, Aimilios Leftheriotis

Outline

1. Related Research & Projects @ Microlab

2. EVOLVE H2020: Integrating HPC, Cloud and Big Data worlds

3. AI@EDGE H2020: AI Platform for Edge Computing in Beyond 5G Networks

Microprocessors and Digital Systems Laboratory

- School of Electrical and Computer Engineering @ National Technical University of Athens
- 21 people (2 Professors, 5 Postdoc, 14 Ph.D. students), est. 1985

Microprocessors and Digital Systems Laboratory

- School of Electrical and Computer Engineering @ National Technical University of Athens
- 21 people (2 Professors, 5 Postdoc, 14 Ph.D. students), est. 1985

Main Areas of Expertise

- High-performance computing
- Cloud, Big-Data and Datacenter technologies
- Heterogeneous platforms
- Embedded and/or reconfigurable processing (with UCs, multi-/many cores, GPU, DSP, FPGA)
- Digital circuit design (VHDL, Verilog, HLS)
- HW/SW co-design and acceleration

H2020 projects

VINEYARD (H2020 2016-2018, P.Coordin.)
heterogeneous platforms (FPGA+GPU+CPU) for datacenter services, acceleration, architecture

AEGLE (H2020, 2014-2018, T.Coordin.)
big data, acceleration of health-care analytics on cloud-based platforms (genomics, diabetes, ICU)

EXA2PRO (H2020, 2018-2020, P.Coordin.)
programming models for HPC and (pre-)exascale apps (quant. comp., materials for CO2 capture,…)

SDK4ED (H2020, 2018-2020)
SW tools for embedded platforms, tradeoff analysis of energy efficiency & security & SW maintenance

5GPHOS (H2020, 2017-2020)
telecom fiber-wireless links, ARoF with low-level DSP acceleration on complex FPGAs (RFSoC)

EVOLVE (H2020 2018-2020)
downstream testbed (HPC+bigdata), Sentinel-2 data
EO pilot apps (e.g., agri, maritime, surveillance)

AI@EDGE (H2020 2021-2023)
A secure and reusable Artificial Intelligence platform for Edge computing in beyond 5G Networks

FabSpace 2.0 (H2020 2016-2018)
open-innovation network for geodata-based applications (service-oriented platform for EO)

ESA Funded: HPCB, SPARTAN/SEXTANT/COMPASS, HIPNOS, QUEENS(1/2/3), radiation testing of COTS FPGAs

Horizon 2020
European Union Funding
for Research & Innovation

esa
European Space Agency

Market Landscape

Cloud market has been doubled since 2018

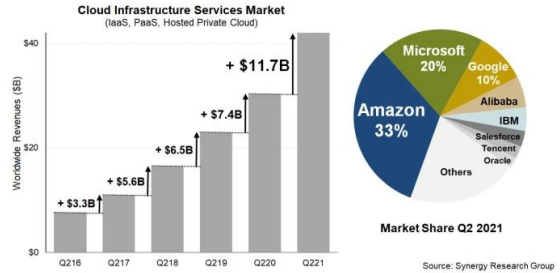
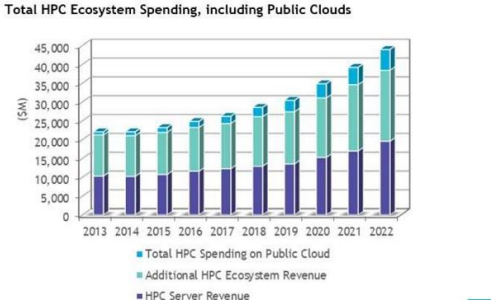


Image Credits: *Synergy Research*

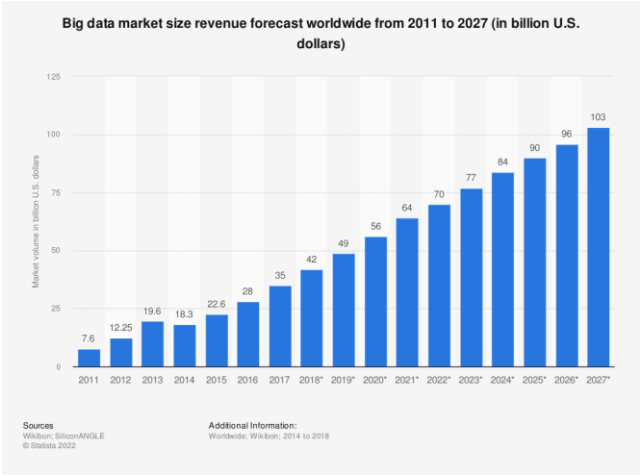
HPC growth has been healthy but more modest: **+50%**

The Total HPC Market Including Public Cloud Spending

- TOTAL HPC spending grew from $22B in 2013 to $26B in 2017, and is projected to reach $44B in 2022



Big-Data grow on the same pace: **+50%**



EVOLVE H2020 - Overview

- **EVOLVE** is a European Innovation Action funded by the European Union's Horizon 2020 Research and Innovation programme.
- The project is composed by **19 specialized partners** from **11 European countries.**
- **Budget:** ~15M
- **Duration:** 2018-2021

evolve
Leading the Big Data Revolution

EVOLVE H2020 - Overview

EVOLVE H2020 aims to build a large-scale testbed by integrating technology from three areas:

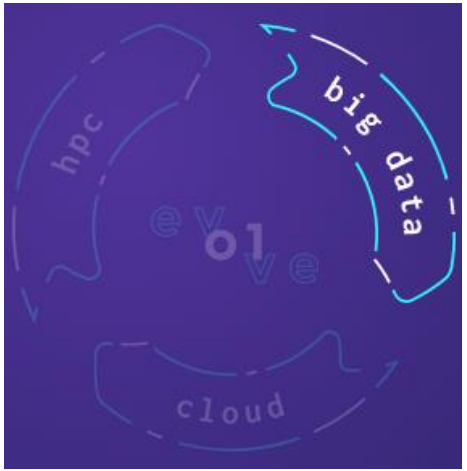
- **High Performance Computing (HPC)**
  By providing an advanced compute platform with HPC features and systems software

EVOLVE H2020 - Overview

EVOLVE H2020 aims to build a large-scale testbed by integrating technology from three areas:

- High Performance Computing (HPC)
  By providing an advanced compute platform with HPC features and systems software

- **Big Data**
  By providing a versatile big-data processing stack for end-to-end workflows



EVOLVE H2020 - Overview

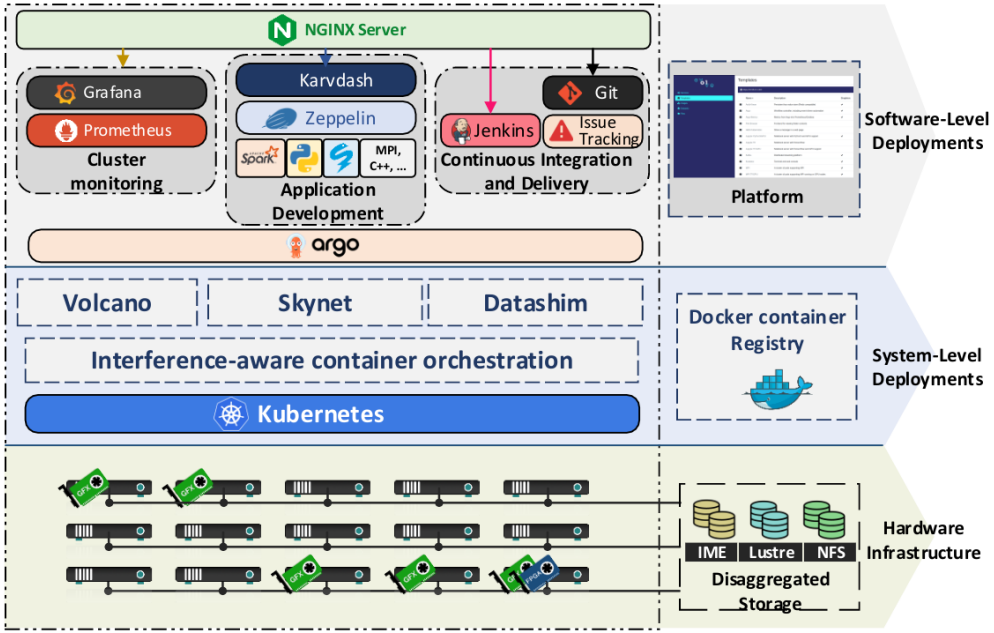
EVOLVE H2020 aims to build a large-scale testbed by integrating technology from three areas:

- High Performance Computing (HPC)
  By providing an advanced compute platform with HPC features and systems software

- Big Data
  By providing a versatile big-data processing stack for end-to-end workflows

- **Cloud**
  By providing ease of deployment, access, and use in a shared manner, while addressing data protection

Advanced Compute Platform



Pilots & Domains

Automated Driving

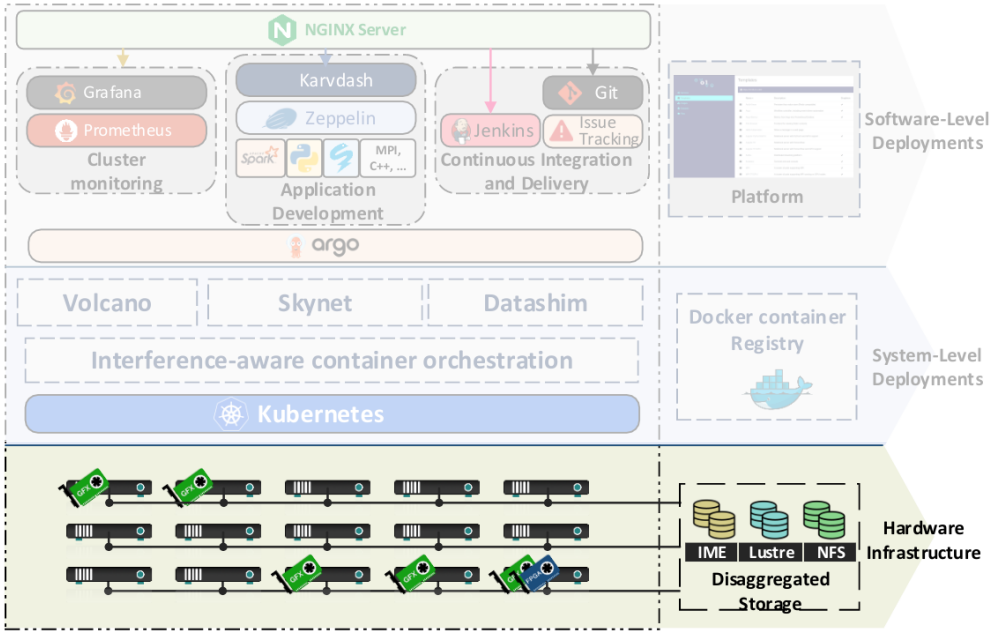
Bus transportation

Maritime surveillanve

Sentinel-2 Satellite images

Automotive service

Advanced Compute Platform



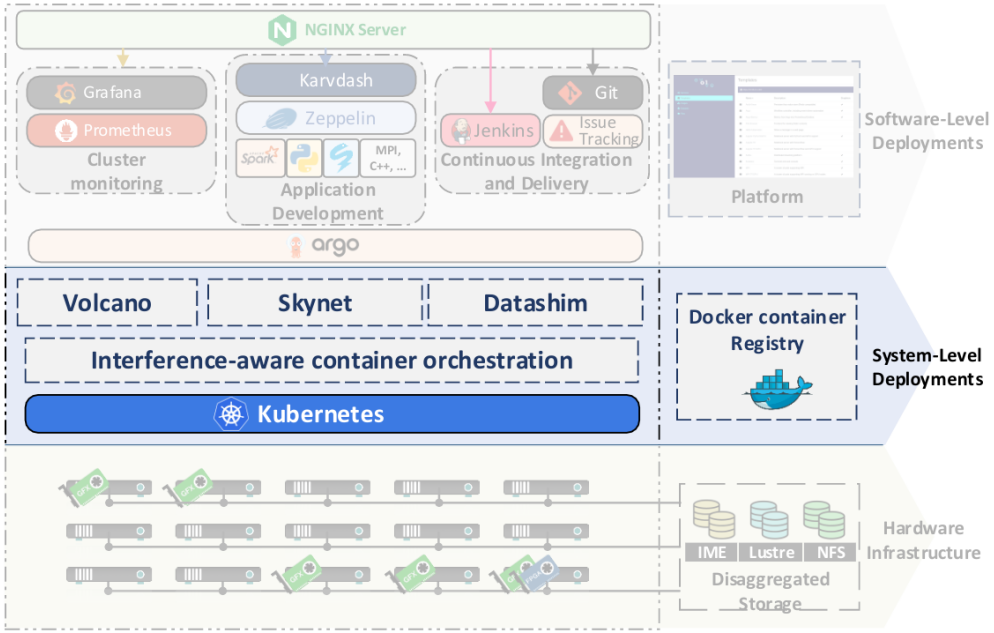
Heterogeneous, HPC enabled HW stack:

Compute: CPUs, GPUs, FPGAs

Storage: NFS, IME & Lustre

Network: Ethernet, Infiniband (56Gb/s)

Advanced Compute Platform



Interoperability & Performance:

Performance-driven plugins: Volcano, iKube, Skynet

Data-driven plugins: Unified Storage Layer (Datashim & H3)

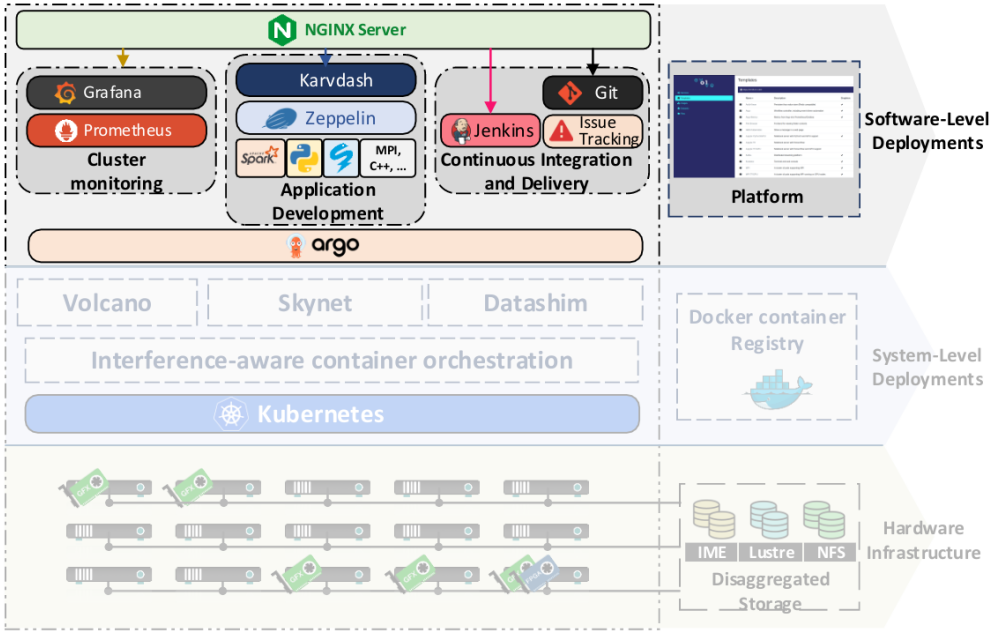
Heterogeneous, HPC enabled HW stack:

Compute: CPUs, GPUs, FPGAs

Storage: NFS, IME & Lustre

Network: Ethernet, Infiniband (56Gb/s)

Advanced Compute Platform



Seamless platform integration and workflow deployment:

Karvdash: Interface for accessing EVOLVE's technologies (Kubernetes, USL, Prometheus, Spark, Grafana, Zeppelin)

SparkLE: Spark Autotuning framework

Interoperability & Performance:

Performance-driven plugins: Volvano, iKube, Skynet

Data-driven plugins: Unified Storage Layer (Datashim & H3)

Heterogeneous, HPC enabled HW stack:

Compute: CPUs, GPUs, FPGAs

Storage: NFS, IME & Lustre

Network: Ethernet, Infiniband (56Gb/s)

EVOLVE H2020 – ICCS Activities

❑ Tools Integration

- ✓ **Container Orchestration**: Kubernetes & Docker
- ✓ **CI/CD**: Git & Jenkins
- ✓ Spark (CPU & GPU)
- ✓ **Monitoring**: (Prometheus & Grafana)

❑ Use-Cases Integration

(with the provided technologies)

❑ System/Software-level components

(performance-driven)

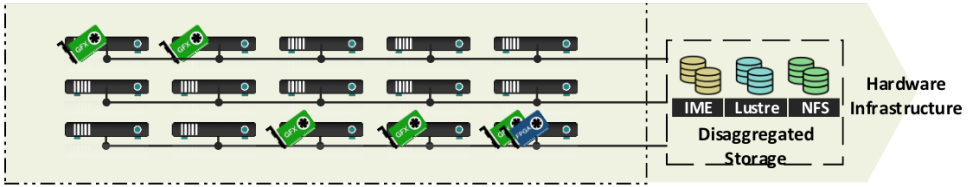
- ✓ Spark Autotuning (SparkLE)
- ✓ Interference-aware container orchestration (**iKube**)

❑ GPU & FPGA Acceleration

(of containerized applications)

EVOLVE H2020 – ICCS Activities

- To test the acceleration capabilities, we have developed in-house versions of VGG16 and ResNet50 **inference models** for our target devices
- The use of GPUs and FPGAs yeld speedups that range from **2x** up to **11x**.

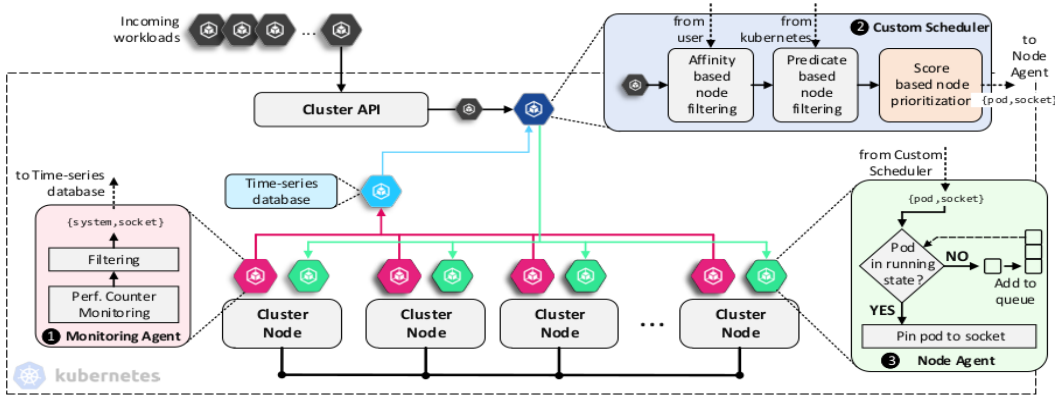


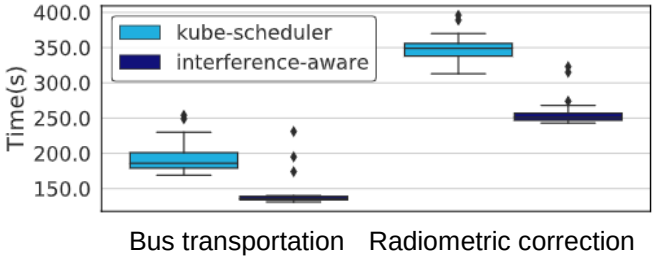
Platform	Execution time (ms)	
	VGG16	ResNet50
TF Intel Xeon Platinum 8153	1067.26	1162.79
TF Nvidia K20Xm	1095.19	410.25
TF Nvidia P40	608.35	187.71
TF Nvidia V100	318.24	143.83
OpenCL Stratix10	103.1	101.21

EVOLVE H2020 – ICCS Activities

iKube
Interference-aware container orchestration in Kubernetes

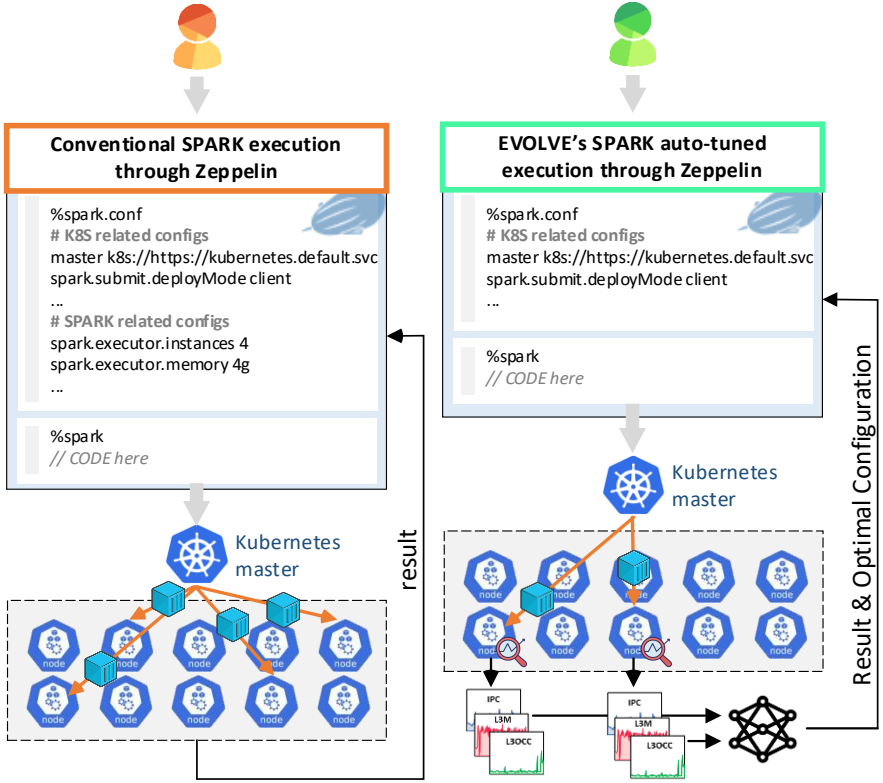
- Extended the default Kubernetes scheduler
- Utilize **hardware counters**
- Identify shared resource **interference**
- Socket-level **container placement**



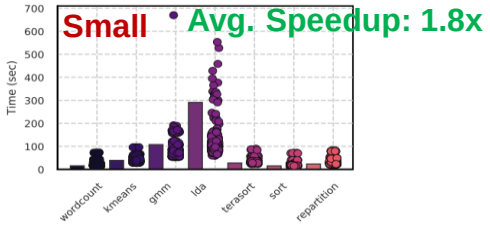


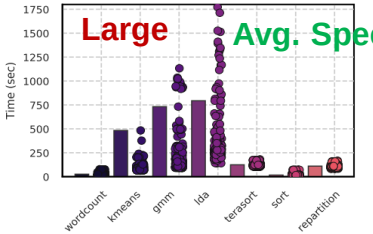
- Container **placement** on Kubernetes cluster under severe **shared resource interference**
- Achieved **26% lower median** of workloads' distribution on average

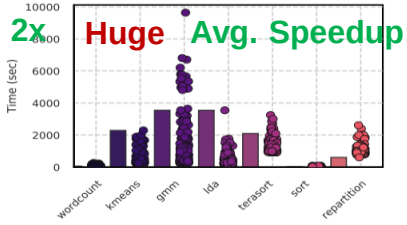
EVOLVE H2020 – ICCS Activities



- Spark auto-tuning (with HiBench)
  - Full set of Spark parameters (>150)
  - 7 benchmarks for 3 dataset sizes
  - Accurate prediction: 82% to 98%
  - Improved Spark execution time by **1.8x – 3x**



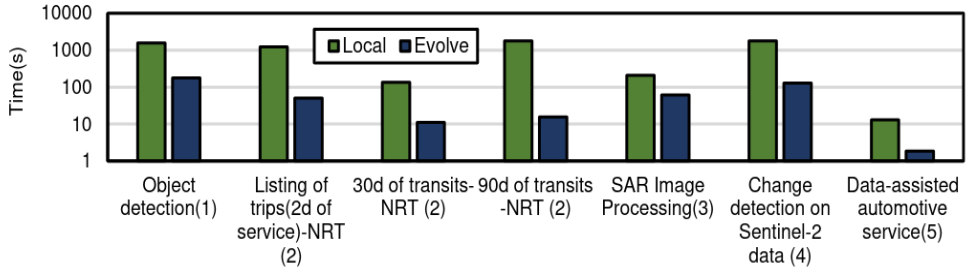




EVOLVE H2020 – Evaluation

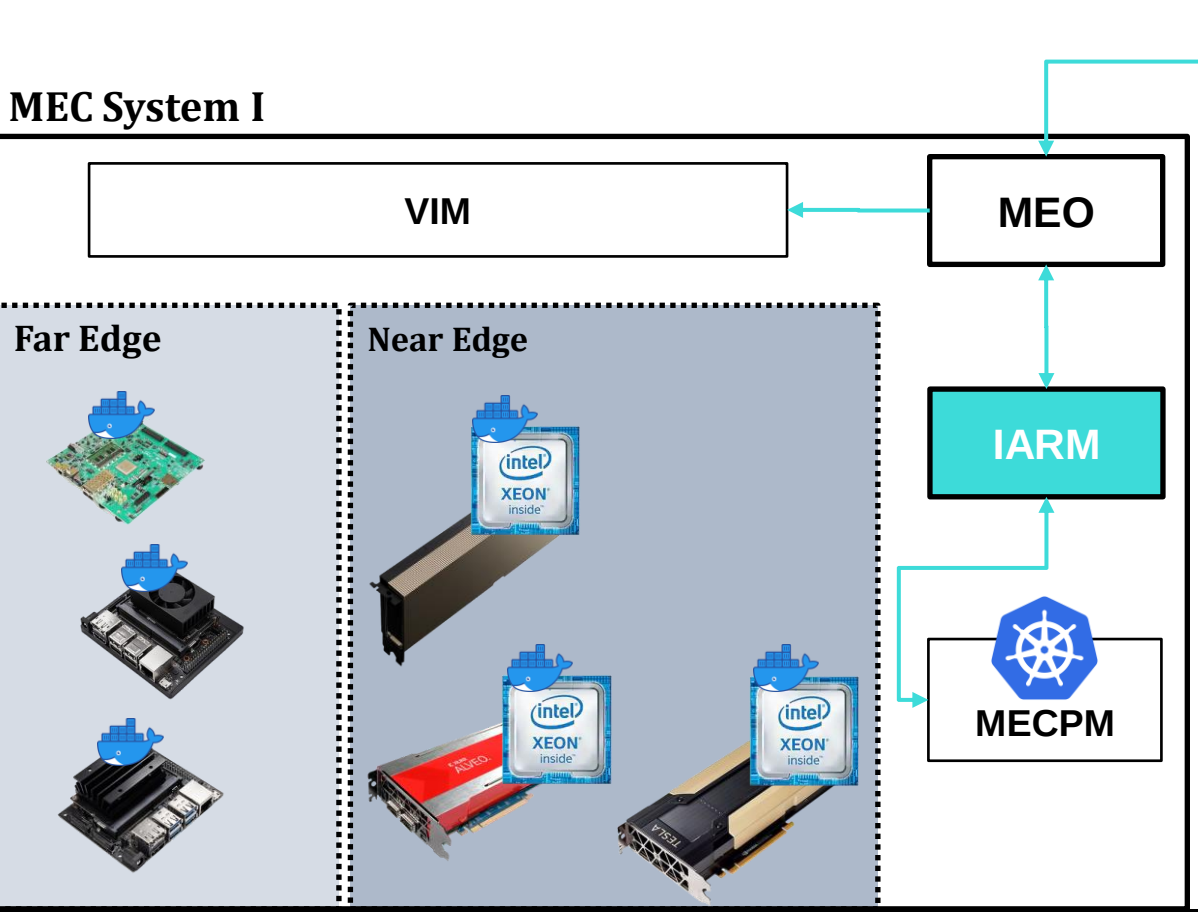
Achieved an average **latency improvement** of **26.3x** related to the legacy deployments

1. **Automated Driving - 8.73x** speedup
2. **Bus transportation** service using observation and historic operational data – **24.7x**, **12x** and **114x** improvement on 30 and 90 days transits
3. **Maritime surveillanve** – SAR Image processing latency improved by **3.45x**
4. **Radiometric correction** and change detection on satellite images – **13.84x**
5. Data-assisted **automotive service** development – **7x**



Legacy and EVOLVE platform comparison on various use-cases

AI@EDGE: acceleration in AI Platform for Edge Computing in Beyond 5G Networks



- AI Functions development to **accelerate** and **increase efficiency** at the near and far edge.
- Integrate with Kubernetes Virtual Machines that include:
  - Intel Xeon CPUs
  - NVIDIA GPUs
  - Xilinx Alveo FPGAs
- Provide **seamless** deployment of AI functions on any device (without user's intervention) based on user's objectives

Thank you for your attention

Q & A

Achilleas Tzenetopoulos
Ph.D. Student @ Microlab ECE NTUA
atzenetopoulos@microlab.ntua.gr