E4
COMPUTER
ENGINEERING

E4

COMPUTER
ENGINEERING

E4 COMPUTER ENGINEERING

HW/SW co-design for Energy Efficiency and Performance: Charting the Path towards Exascale Computing

Food for Thought in Three Parts

HW/SW co-design for Energy Efficiency and Performance: Charting the Path towards Exascale Computing

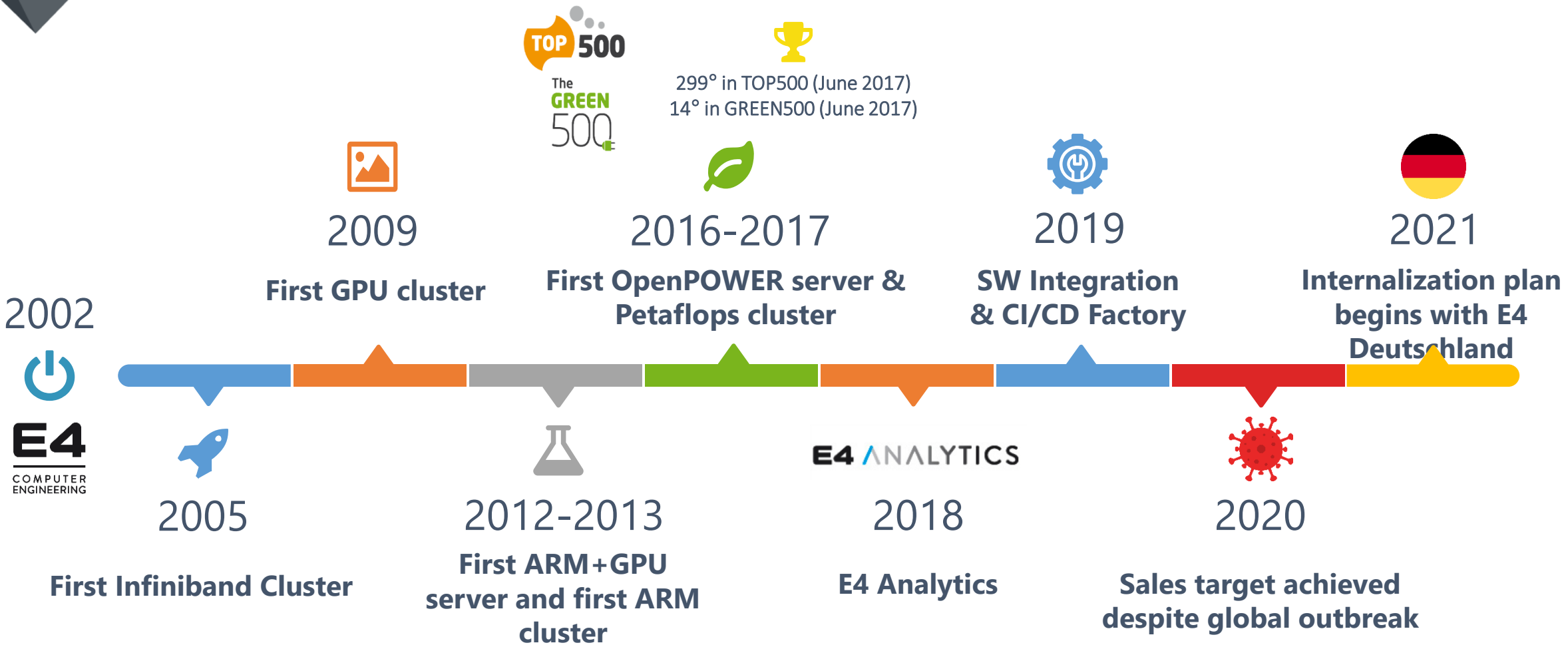
Part #1

Designing systems for Energy Efficiency and Performance

Fabrizio Magugliani
E4 Computer Engineering SpA
<Fabrizio.Magugliani@e4company.com>

Where we are coming from

WHEN PERFORMANCE MATTERS



Designing systems for Energy Efficiency and Performance

WHEN PERFORMANCE MATTERS

Customers are giving for granted performances and are demanding solutions having Energy Efficiency as an integral feature (‘plugged in’ since the beginning of design of the solution and not as a sort of afterthought of the design)

Building exascale-class systems requires (mandates..) a multi-disciplinary approach

3 stakeholders must collaborate:

- The supply side -> match the requirements of the customer with the proper technology (performance, components, cooling, smart schedulers)
- The system-level developers -> provide non-intrusive tools (system-level ‘intelligence’) for maximizing the energy efficiency of the solution by leveraging the HW components
- The end-user -> harmonize the solution with the requirements of the organization running the system in the datacenter

Designing systems for Energy Efficiency and Performance (The supply side)

E4 mantra: Build on proven technology, prepare/drive for next-gen technology, lead in innovation

5 “easy” steps:

- Start from the customers’ overall requirements (not ‘only’ the performance)
- Select the best technologies and the best technology providers
- Test, test, test
- Validate the solution with the technology providers and with the customer
- Deploy

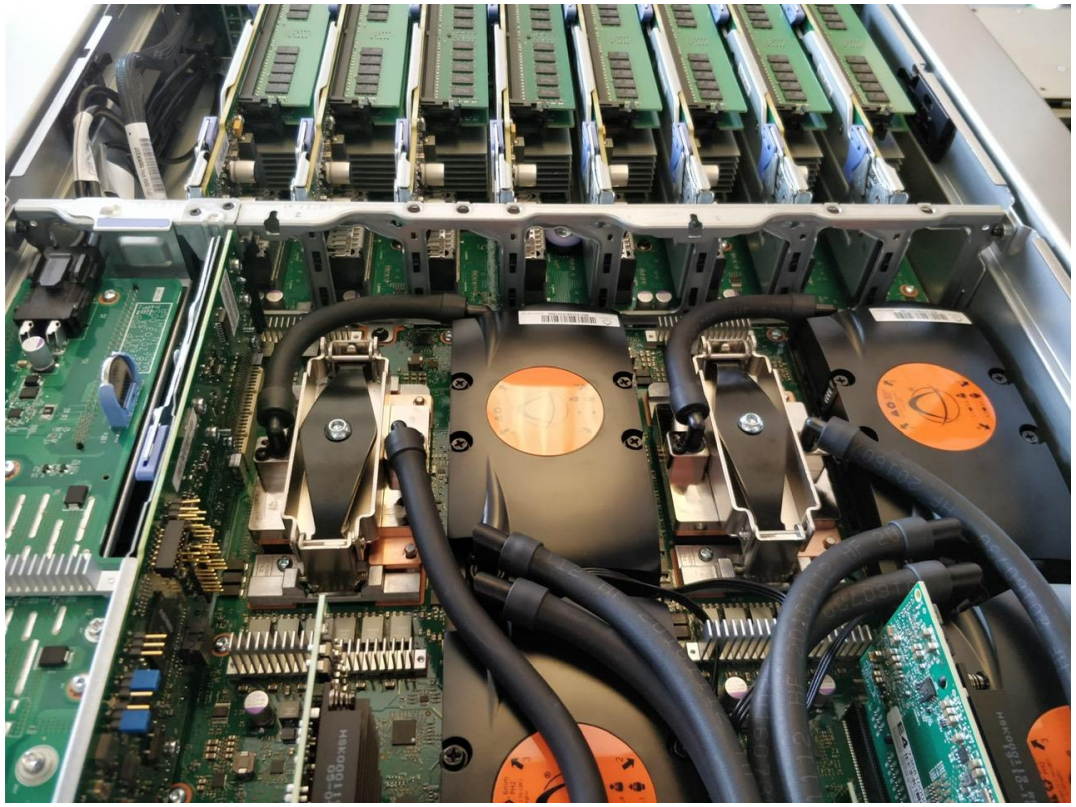
Build on proven technology: tight connections with silicon vendors

Prepare for next-gen technology:

Designing systems for Energy Efficiency and Performance (Cooling)

Build on proven technology: co-design with providers of liquid cooling technology

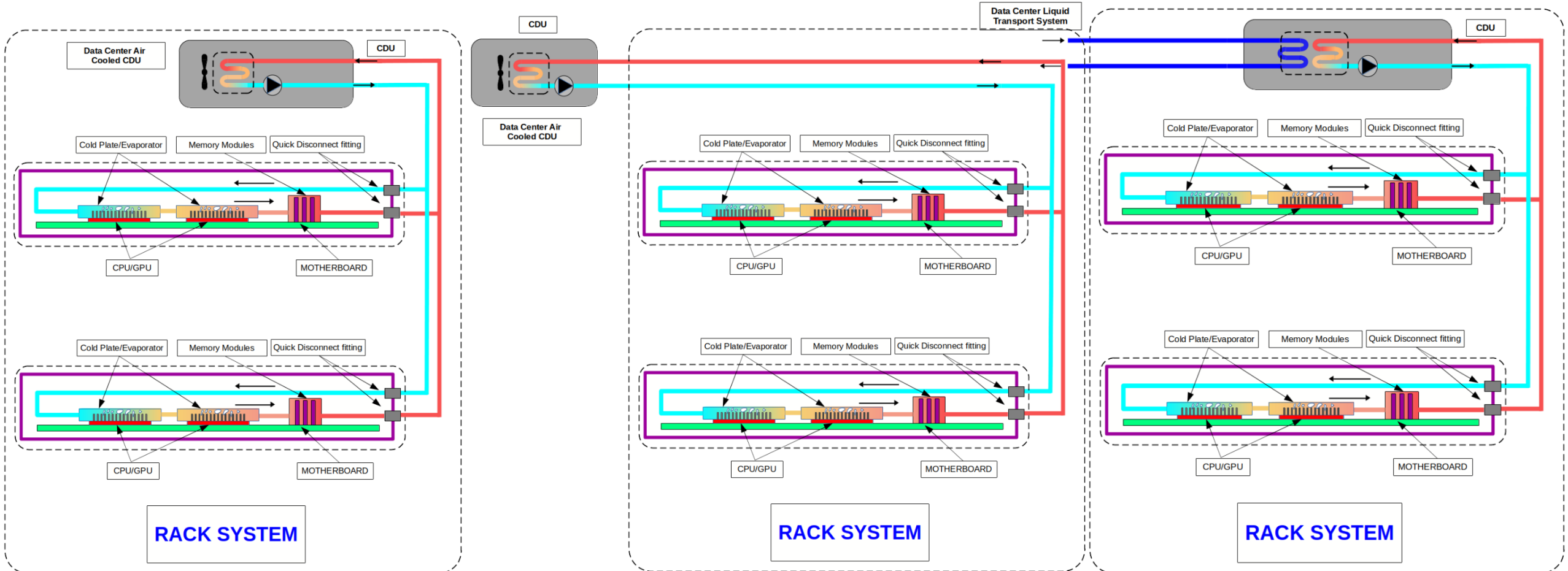
- DAVIDE (Development of an Added-Value Infrastructure Developed in Europe)
- PRACE-3IP PCP: Whole-System Design for Energy Efficient HPC



Designing systems for Energy Efficiency and Performance (Cooling)

Prepare for next-gen technology: collaborate with innovative start-up

- Co-designing the two-phase cooling equipment with InQuattro:
- TEXTAROSSA/EuroHPC calls 2019



Designing systems for Energy Efficiency and Performance
System-level ‘intelligence’

Build on proven technology: adapt the ‘standard’ tools

Prepare for next-gen technology: collaborate with innovative start-up and academia
Co-design with the owners of the technology: Università di Bologna (speaker in *Part #2*)

- ExaMon
- CountDown
- Tested on DAVIDE
- REGALE/EuroHCP call 2019

Designing systems for Energy Efficiency and Performance
Listening to the end-user

Build on proven technology: DAVIDE has been co-design with the end user/data center CINECA (speaker in *Part #3)*

Prepare for next-gen technology: tools and solution to integrate its solution within a data-center wide DCIM

HW/SW co-design for Energy Efficiency and Performance: Scalable and multiscale energy monitoring

Part #2

Developing non-intrusive (system-level 'intelligence') tools

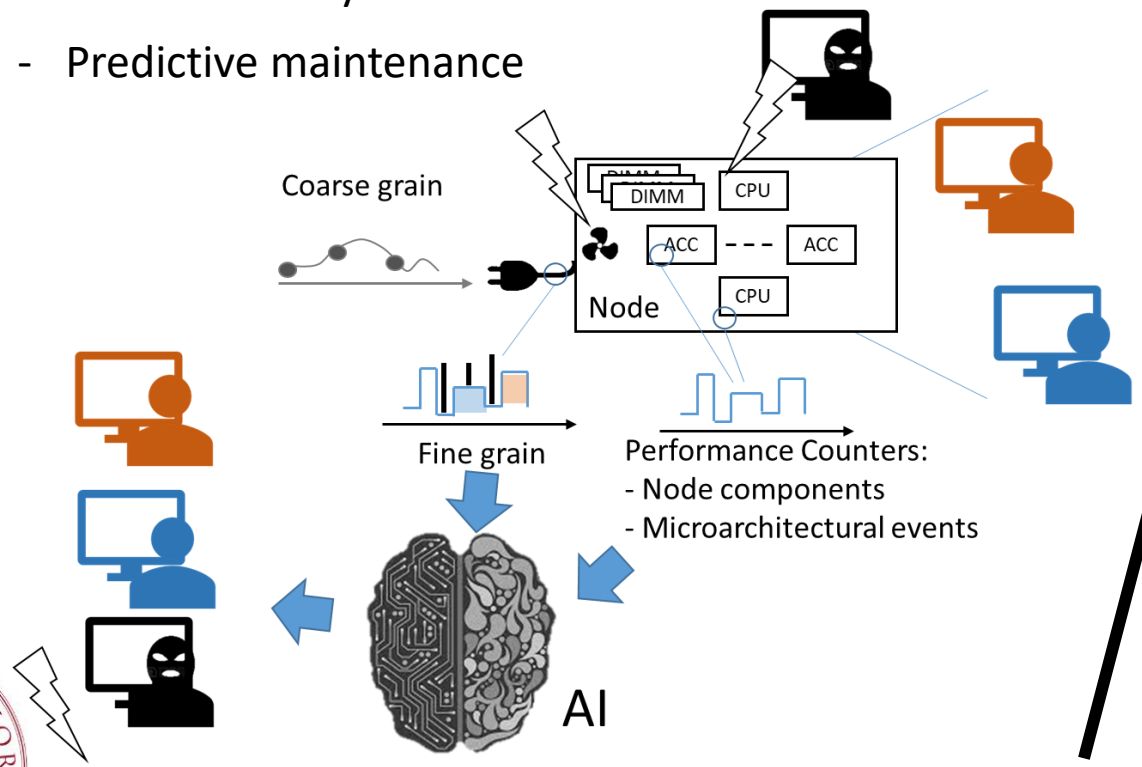
Benedetta Mazzoni
Università di Bologna
<benedetta.mazzoni3@unibo.it>

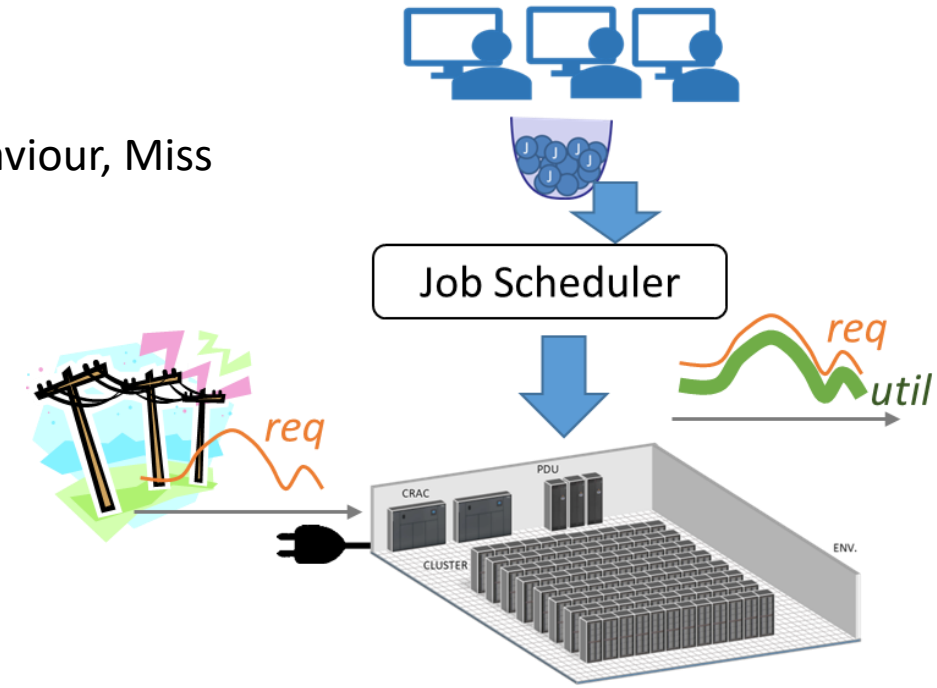
Holistic Monitoring

WHEN PERFORMANCE MATTERS

Fine Grain Power and Performance Measurements:

- Verify and classify node performance (In spec / out of spec behaviour, Miss configuration, Aging and wear out)
- Detect security hazards
- Predictive maintenance





System Power Capping

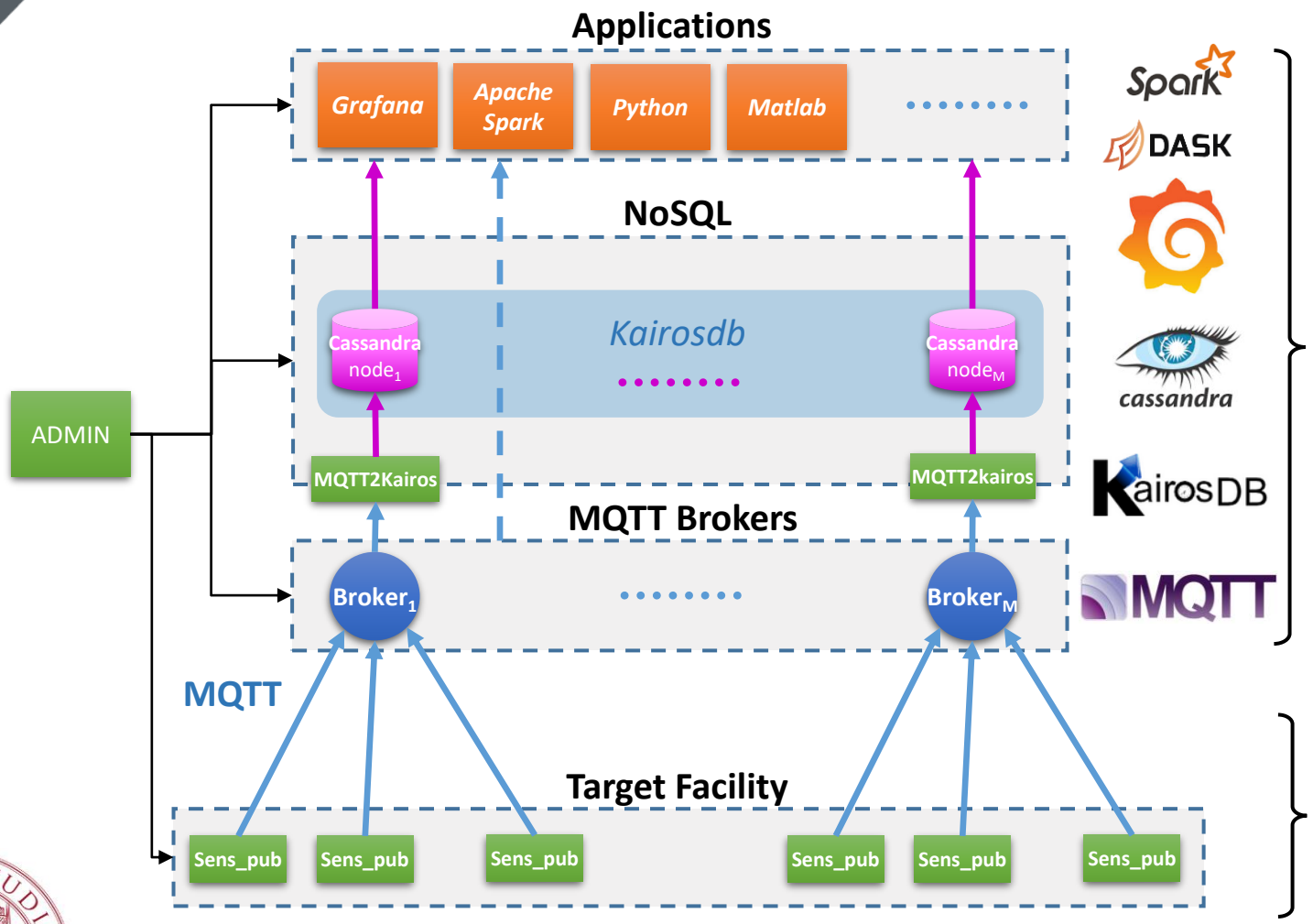
- New Installations, Grid SLA, Power Shortage, Natural Disasters
- Ensures operating power below a maximum power consumption level

A. Libri et al., “pAElla: Edge AI-Based Real-Time Malware Detection in Data Centers”, JIOT 2020

A. Borghesi et al, “A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems”, EAAI 2019

Scalable Monitoring Framework: ExaMon

WHEN PERFORMANCE MATTERS



EXAMON

Front-end

- **MQTT** Brokers
- Data **Visualization**
- **NoSQL** Storage
- **Big Data** Analytics

Back-end

- MQTT–enabled sensor collectors

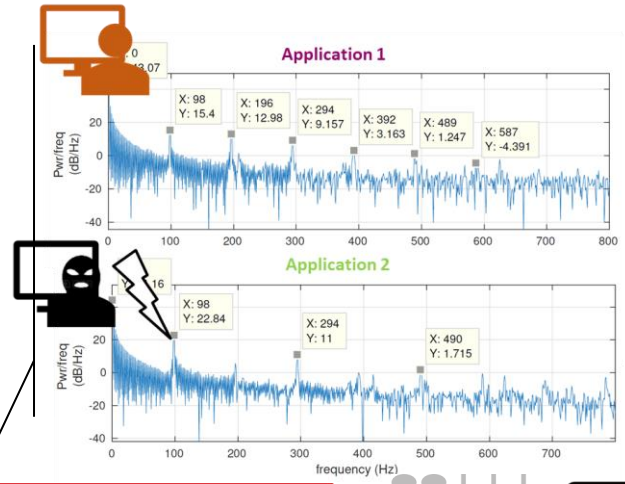
https://github.com/EEESlab/examon

F. Beneventi et al., “Continuous learning of HPC infrastructure models using big data analytics and in-memory processing tools”

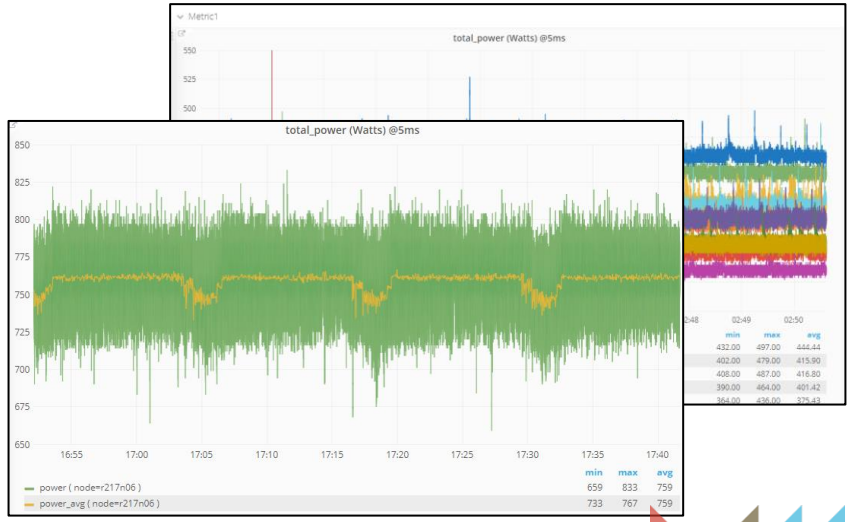
A. Bartolini et al, “The DAVIDE Big-Data-Powered Fine-Grain Power and Performance Monitoring Support”

High - Frequency Energy-monitoring

DiG



ETH Zurich / Univ. of Bologna
SoA out-of-band
High Resolution Power
and Performance Monitoring



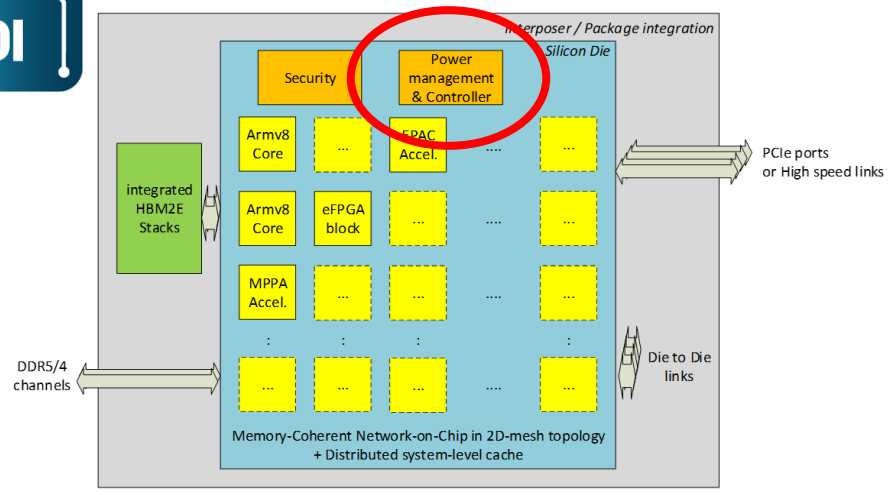
CINECA

COMPUTER ENGINEERING

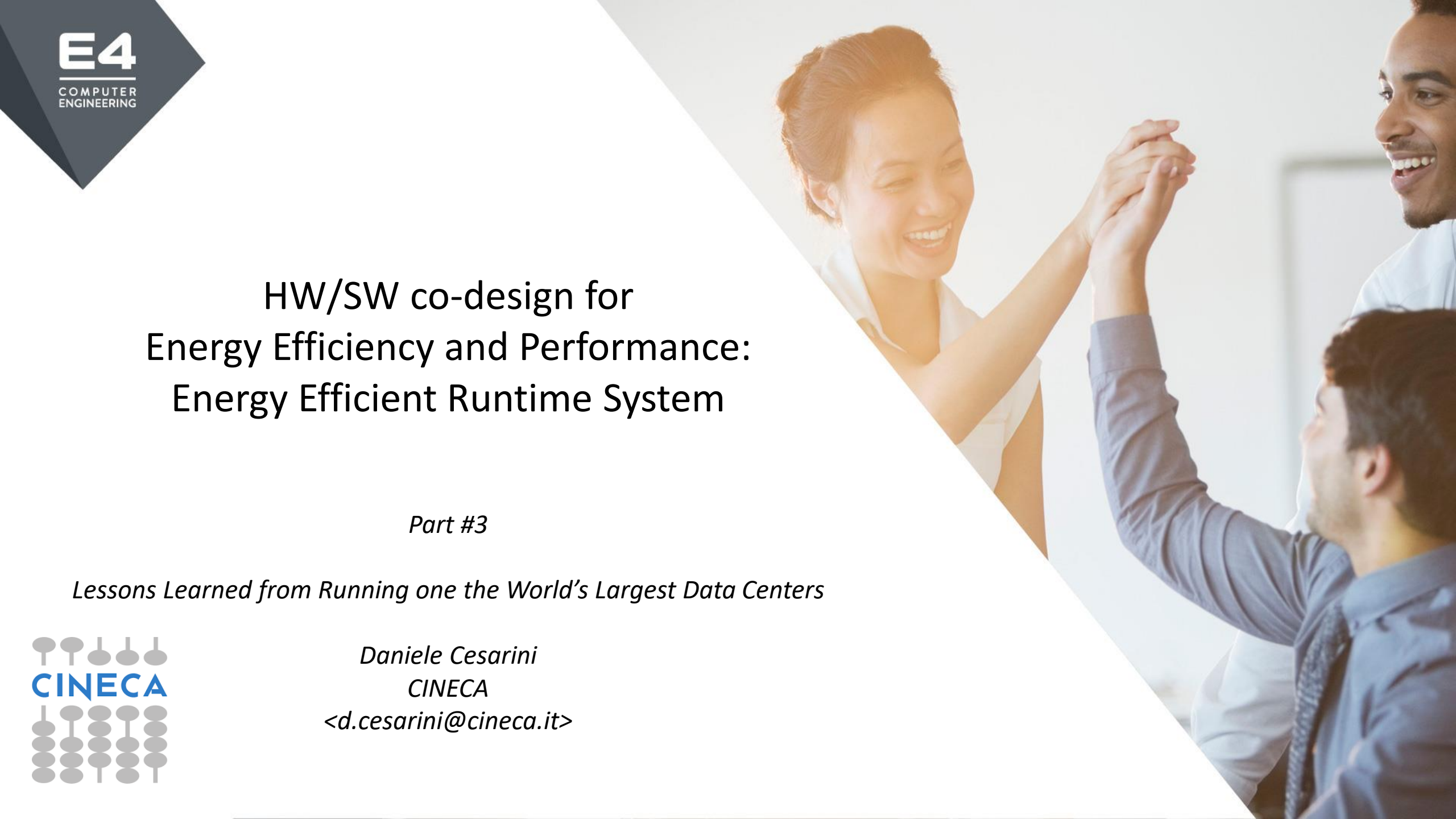
IBM

PPI4HPC

European Processor Initiative



Bambini et al. “An Open-Source Scalable Thermal and Power Controller for HPC Processors” ICCD2020
Bartolini et al. “A pulp-based parallel power controller for future exascale systems” ICECS 19



HW/SW co-design for Energy Efficiency and Performance: Energy Efficient Runtime System

Part #3

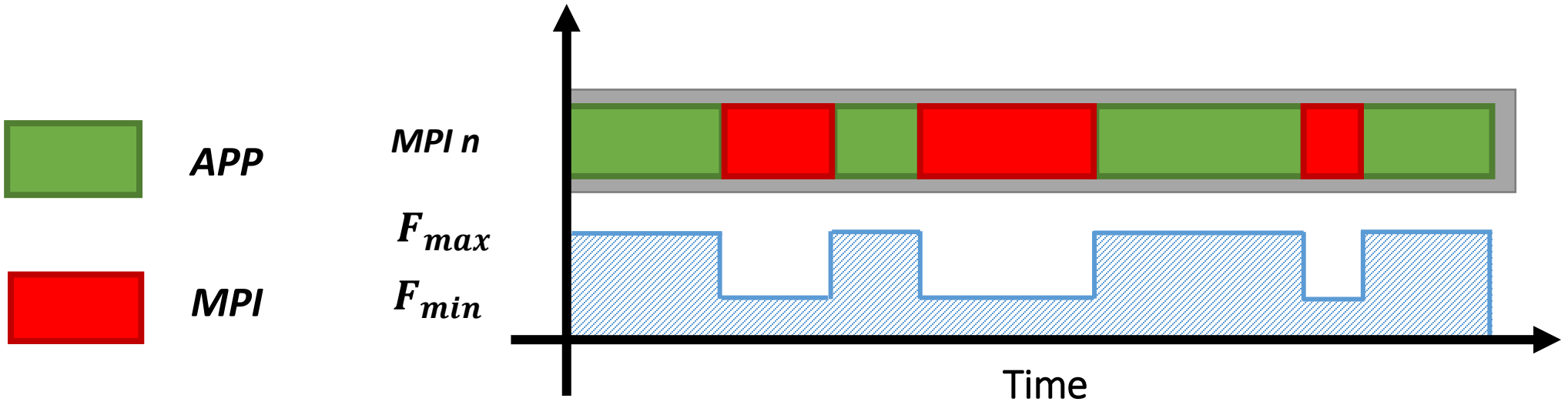
Lessons Learned from Running one the World's Largest Data Centers

Daniele Cesarini
CINECA
<d.cesarini@cineca.it>

Energy Efficient Runtime System

WHEN PERFORMANCE MATTERS

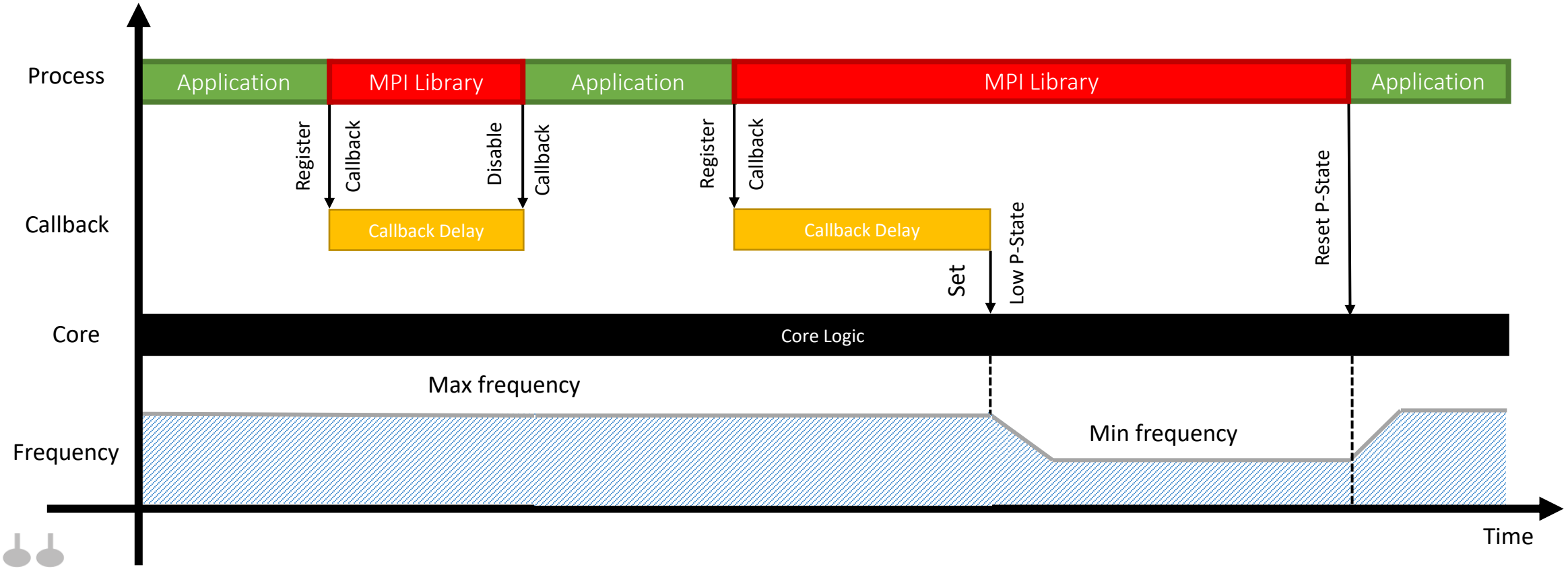
COUNTDOWN is a tool to identify and automatically reduce the power consumption of the CPU cores during communication phases of an MPI-based application.



COUNTDOWN does not impact on the application tasks but only on the communication phases!

COUNTDOWN

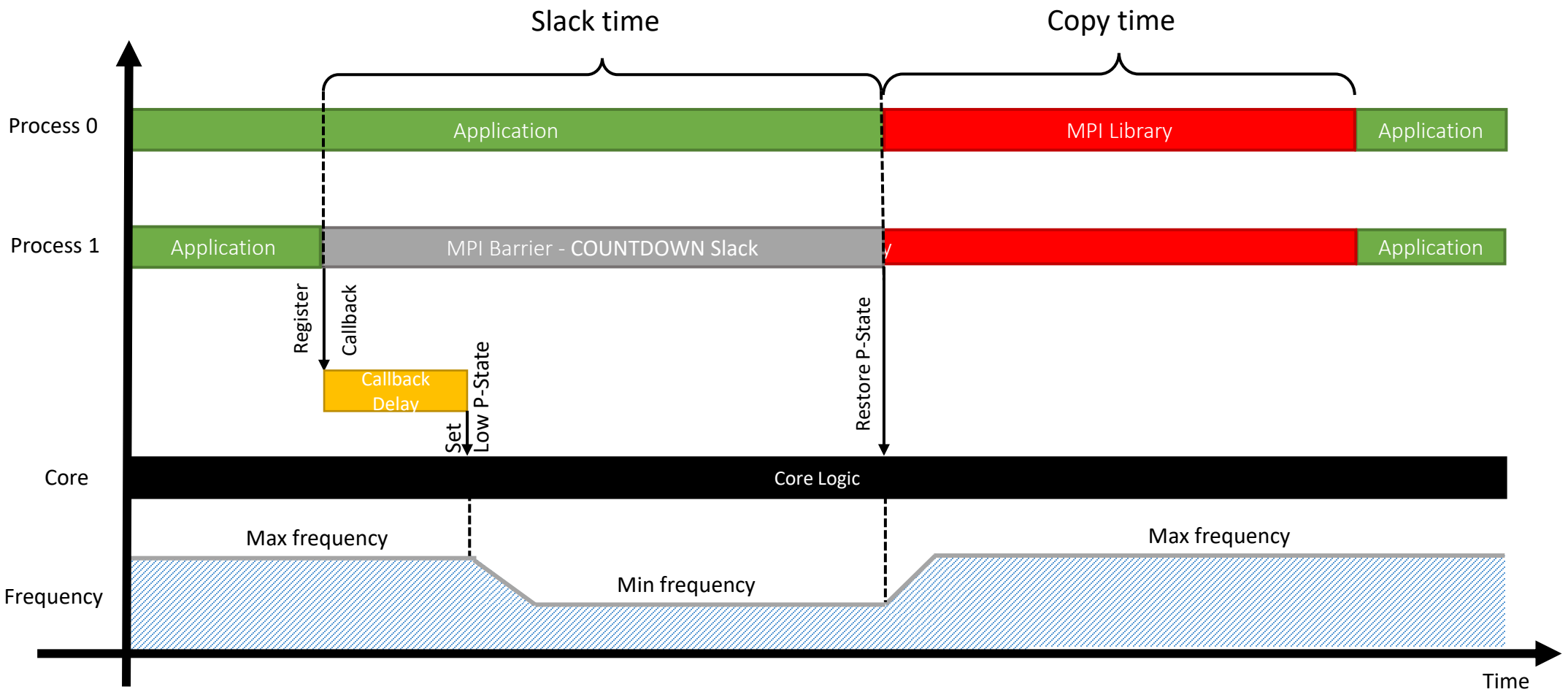
WHEN PERFORMANCE MATTERS



Not always beneficial to reduce the core's frequency during the entire MPI regions -> overhead!

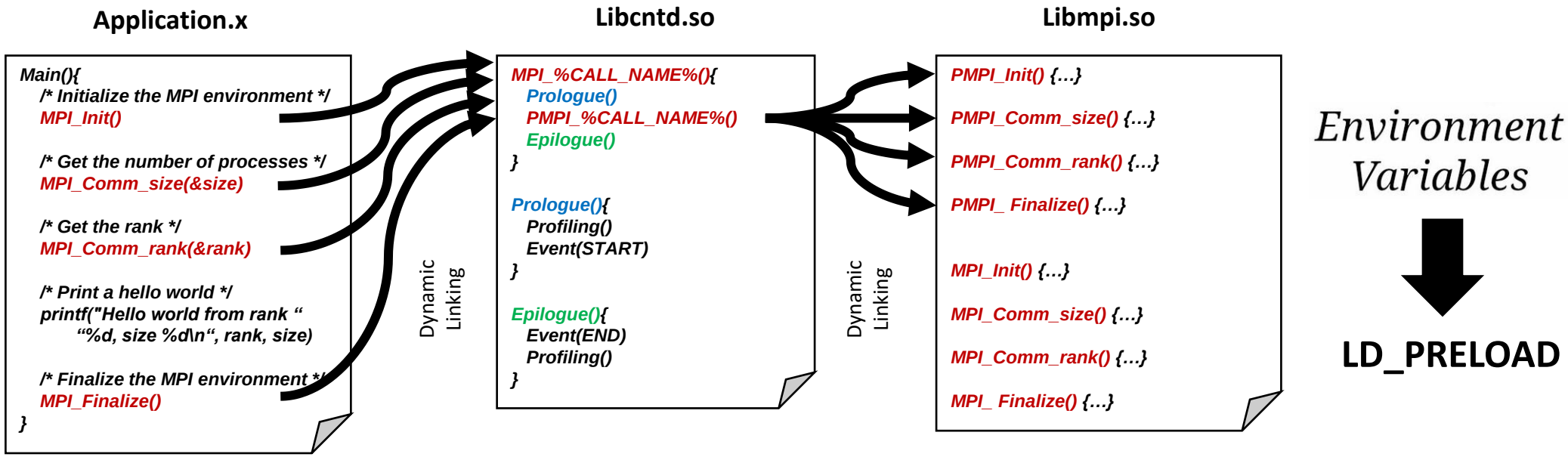
COUNTDOWN Slack

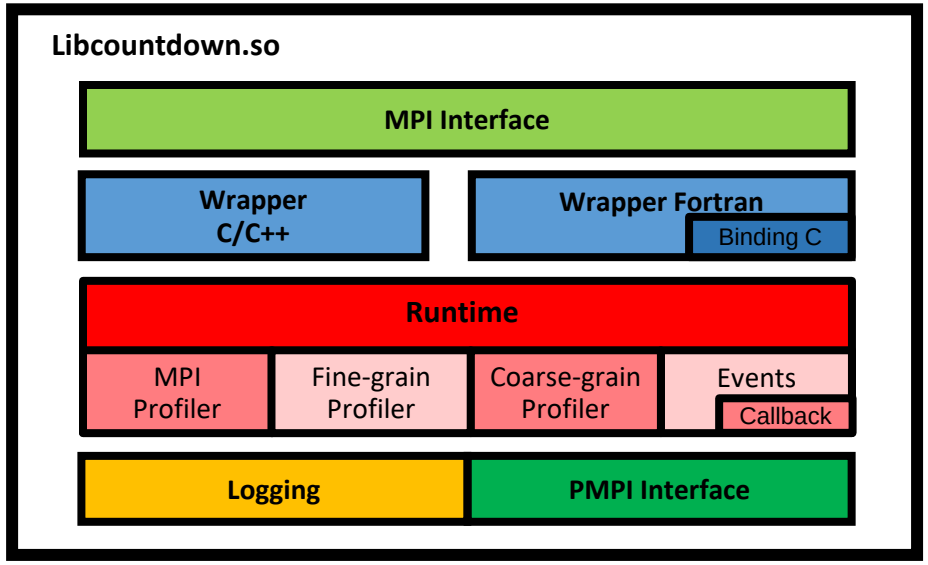
WHEN PERFORMANCE MATTERS



COUNTDOWN Slack can split slack time to copy time!

Framework





Does not require modification of source code nor rebuilding process!

Example:

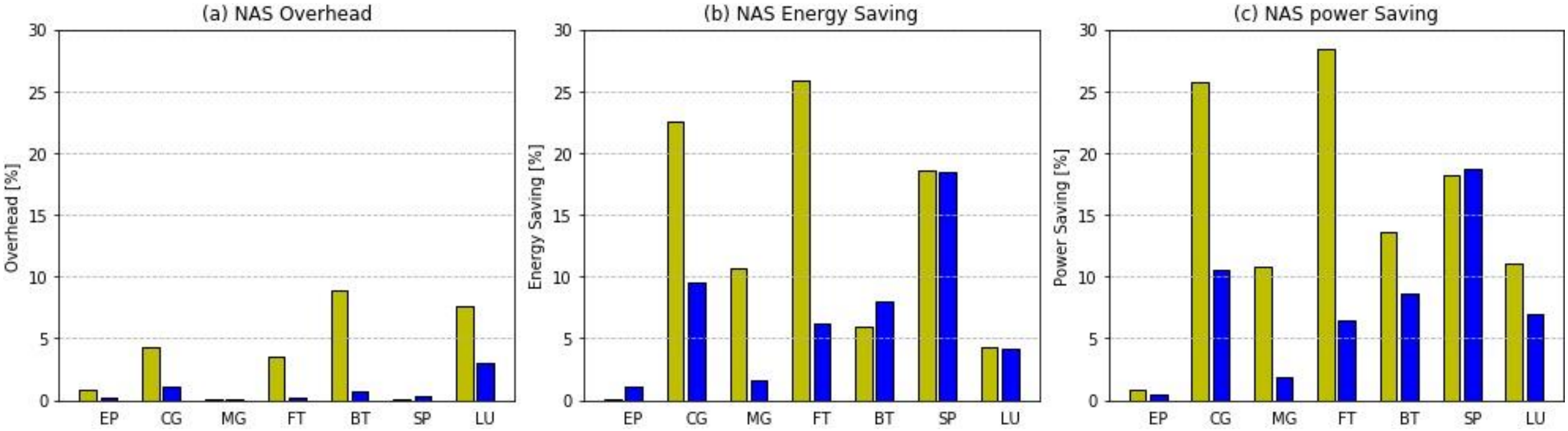
$>: mpirun cp.x -input SiO2.in

$>: mpirun -genv LD_PRELOAD=/path/libcntd.so cp.x -input SiO2.in

COUNTDOWN Slack - Exploration Results

NAS Benchmarks are a small set of programs designed to help evaluate the performance of parallel supercomputers.

Computing Resources: 1024 Cores (29 nodes)



Avg Energy/Power Saving: 9.96% - 10.73%

Max Energy/Power Saving: 18.56% - 19.04%

Avg Overhead: 0.79%

Worst-case Overhead: 3.02%

Github: http://github.com/EEESlab/countdown

Examon + COUNTDOWN Slack

Conclusions

Co-design is the ‘only’ way to go to achieve Energy-Efficient systems without sacrificing performance

Addressing one component at-a-time won’t result in maximizing the overall, system-wide and ecosystem-wide efficiency

An ecosystem-wide, synergic approach is mandatory

Our vision

WHEN PERFORMANCE MATTERS

https://ecs-org.eu/working-groups/transcontinuum-initiative

Thanks, from all of us

E4company.com

UniBo.it

cineca.it