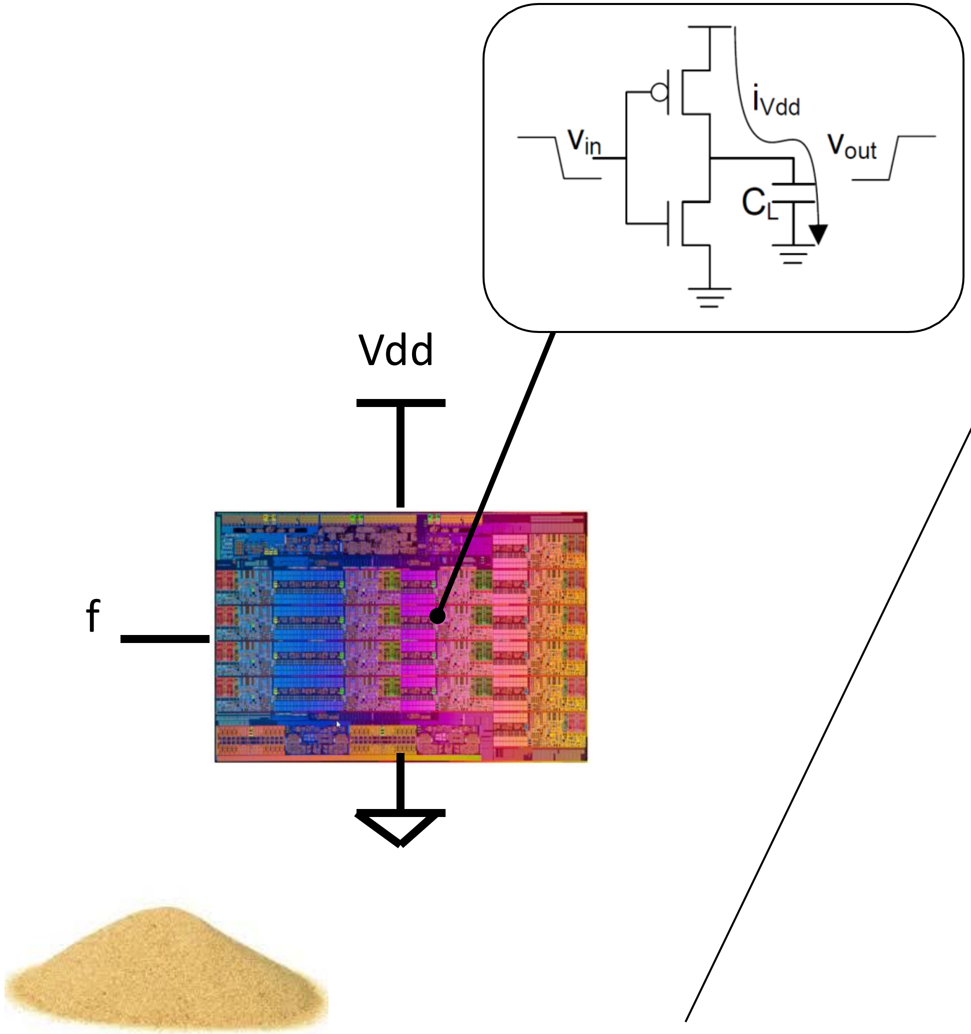
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts in seven acts

Prof Andrea Bartolini

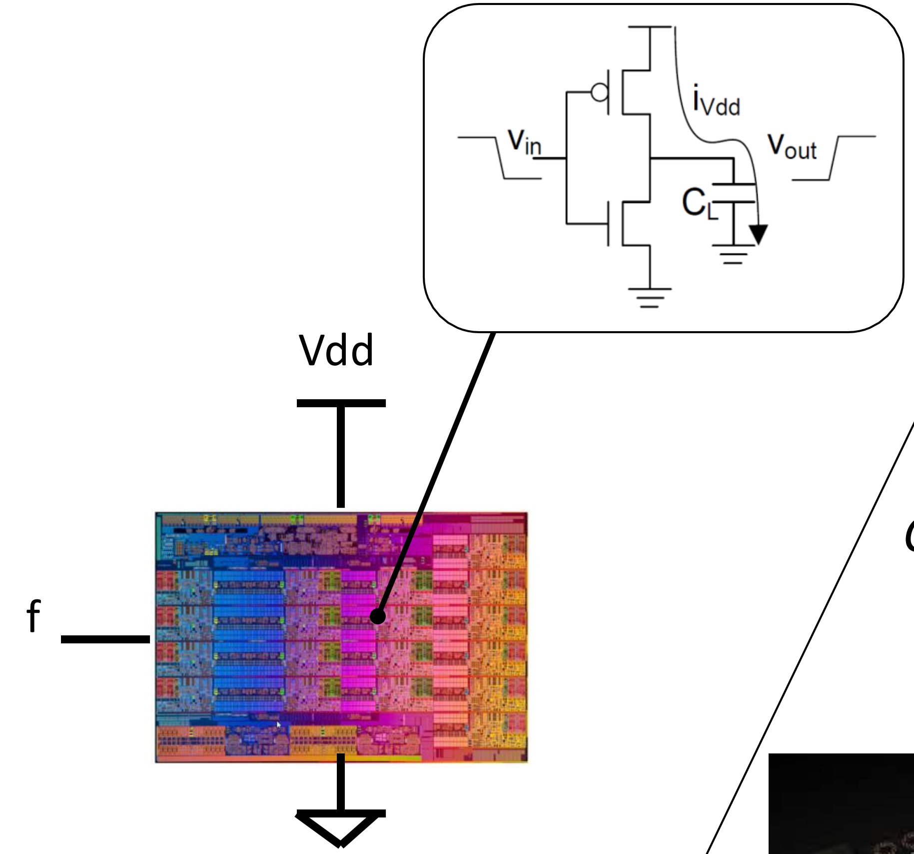
DEI, a.bartolini@unibo.it

From sand to megawatts – Act 1: Dynamic Power



$$P_{dynamic} = \overbrace{\underbrace{a}_{\text{activity factor}} \times \underbrace{C_L}_{\text{load capacitance}}}^{\text{“effective capacitance” } (C_{Effective})} \times \underbrace{V_{dd}^2}_{\text{supply voltage}} \times \underbrace{f}_{\text{clock frequency}}$$

From sand to megawatts – Act 1: Dynamic Power



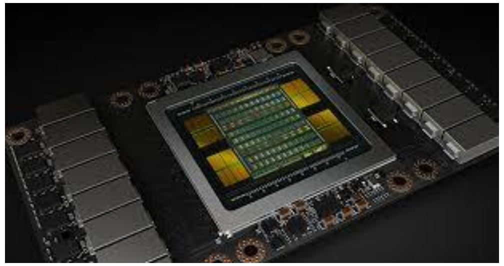
"effective capacitance" ($C_{Effective}$)

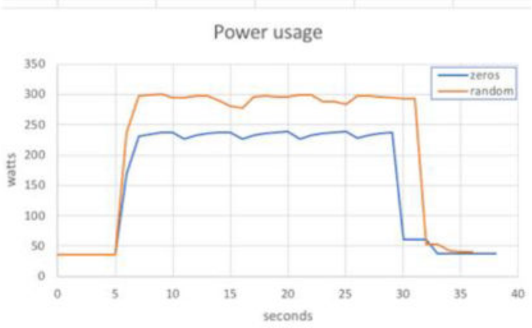
$$P_{dynamic} = \overbrace{a \times C_L}^{} \times V_{dd}^2 \times f$$

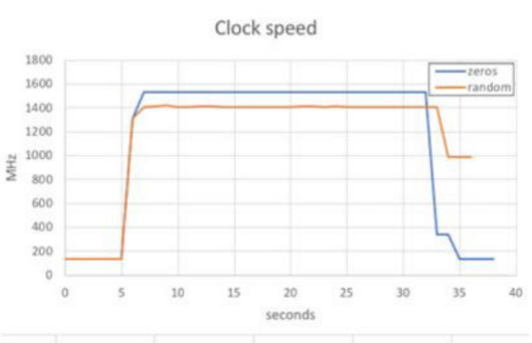
activity factor — load capacitance — supply voltage — clock frequency

Ceff in a power-limited architecture

- TC/HGEMM has surprising data-dependent performance: **125 TF** theoretical peak, **113 TF** achievable on zero-filled matrices, **105 TF** peak on random CCC matrices, **~95 TF** peak on matrices with fully random FP16 entries





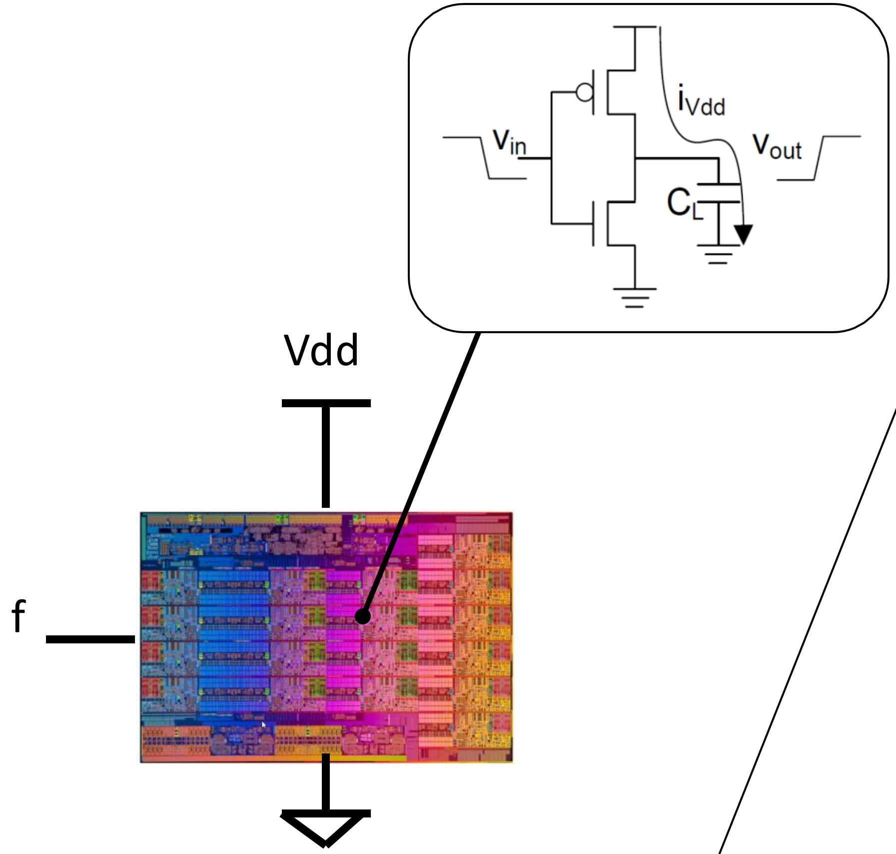


https://indico-jsc.fz-juelich.de/event/76/session/0/contribution/1/material/slides/0.pdf

Wayne Joubert - OpenPOWER ADG 2018

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts – Act 2: Static Power a.k.a Leakage



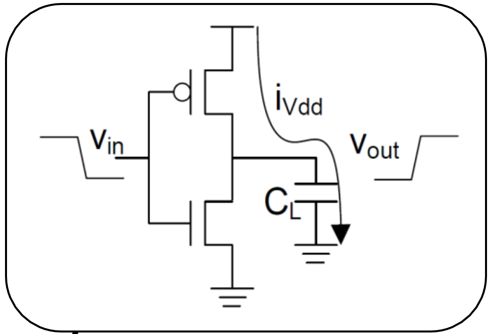
constants
temperature

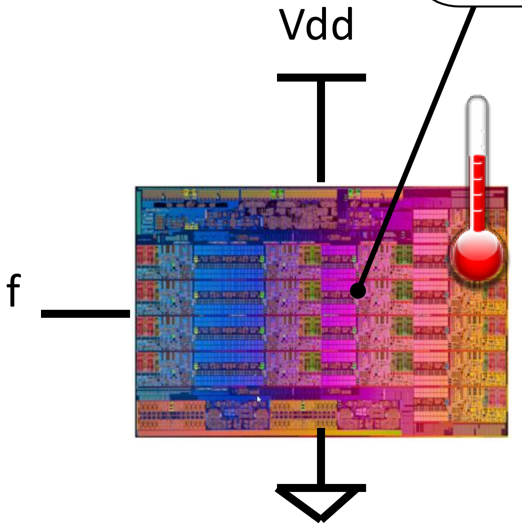
$$I_{leakage} = k_1 \times \left(1 - e^{-k_2 \times V_{ds}/T}\right) \times e^{k_3 \times \left(V_{gs} - V_{TH} - V_{off}\right)/T}$$

drain to source voltage
gate to source voltage
threshold voltage
empirical parameter

f

From sand to megawatts – Act 2: Static Power a.k.a Leakage





constants temperature

$$I_{leakage} = k_1 \times \left(1 - e^{-k_2 \times V_{ds}/T}\right) \times e^{k_3 \times \left(V_{gs} - V_{TH} - V_{off}\right)/T}$$

drain to source voltage | gate to source voltage | threshold voltage | empirical parameter

Temperature in a power-limited architecture

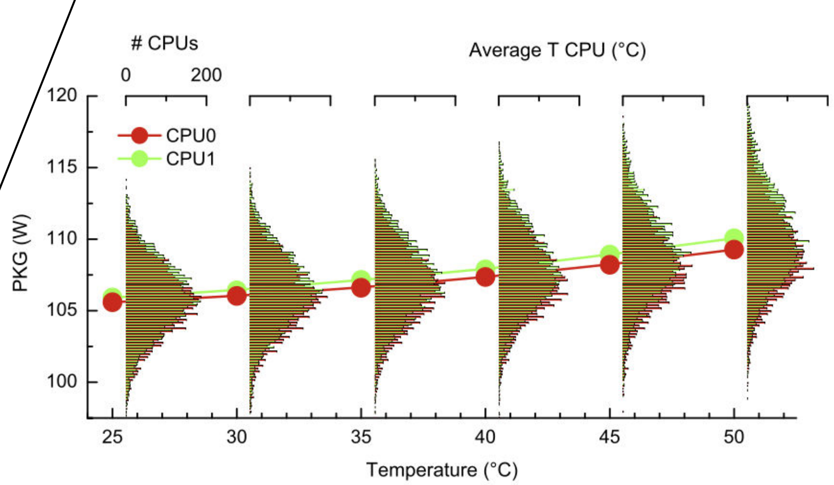


Fig. 7. Average power consumption distribution of SuperMUC Phase1 compute nodes (Intel Sandy Bridge-EP Xeon E5-2680 8C) running single node HPL (Turbo OFF) at different inlet water temperatures.

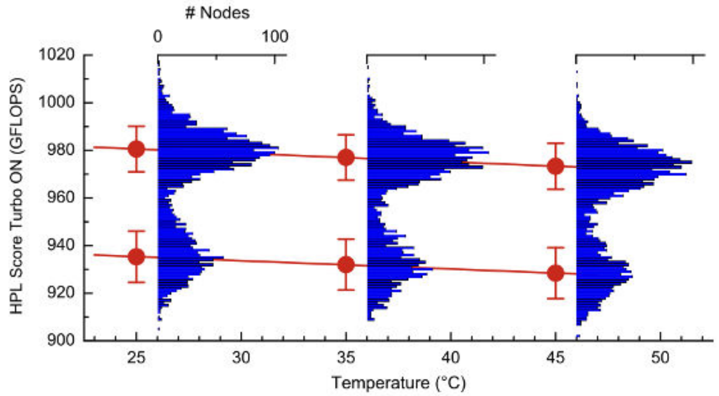
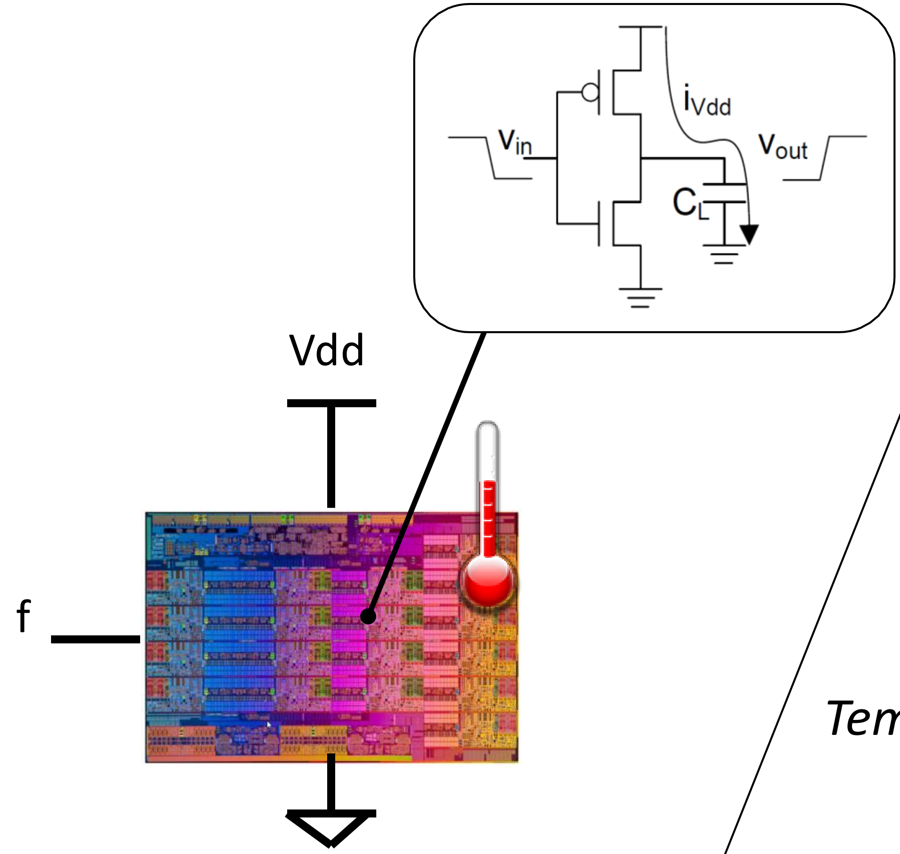


Fig. 5. Performance distribution of SuperMUC Phase2 compute nodes (Intel Haswell Xeon E5-2697 v3) at different inlet water temperatures.

Shoukourian et. al, Analysis of the efficiency characteristics of the first High-Temperature Direct Liquid Cooled Petascale supercomputer and its cooling infrastructure, JPDC 2017

From sand to megawatts – Act 3: Alpha-Power Thermal Inversion



Delay: $$D_p = \frac{C_{out} V_{dd}}{I_{ON}} = \frac{C_{out} V_{dd}}{\mu(T)[V_{dd} - V_{th}(T)]^{\alpha}}$$

Carrier Mobility: $$\mu(T) = \mu(T_0)\left(\frac{T_o}{T}\right)^m$$

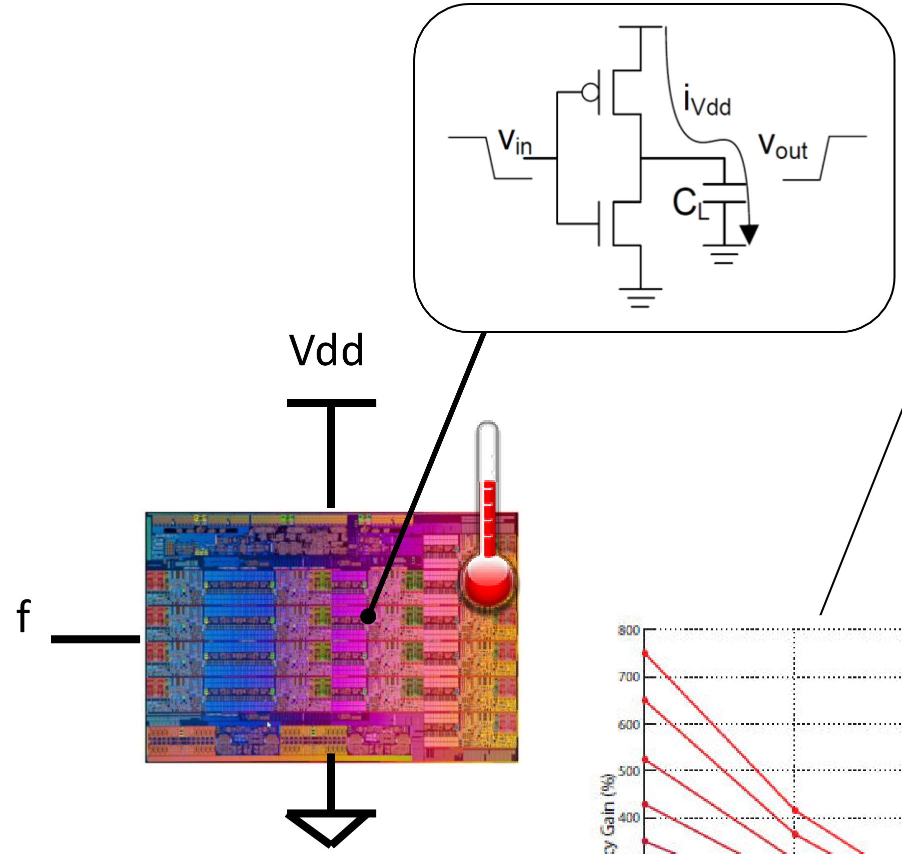
Threshold Voltage: $$V_{th} = V_{th}(T_0) - k(T - T_0)$$

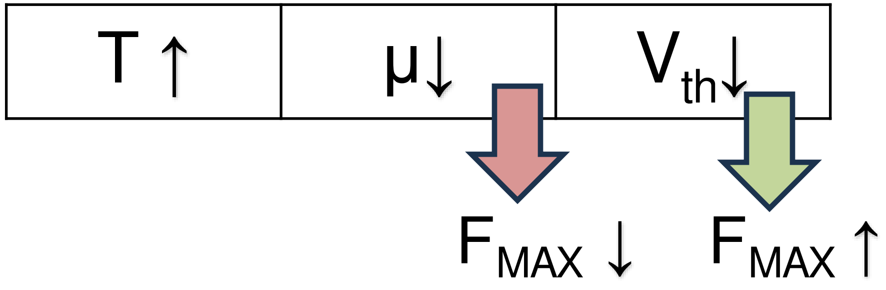
$T\uparrow$	$\mu\downarrow$	$V_{th}\downarrow$

Temperature in a power-limited architecture

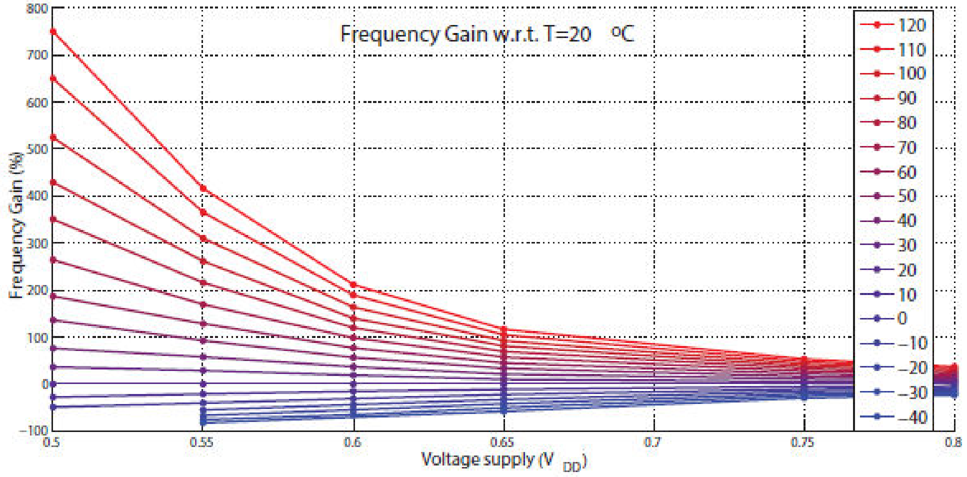
$F_{MAX}\downarrow$ $F_{MAX}\uparrow$

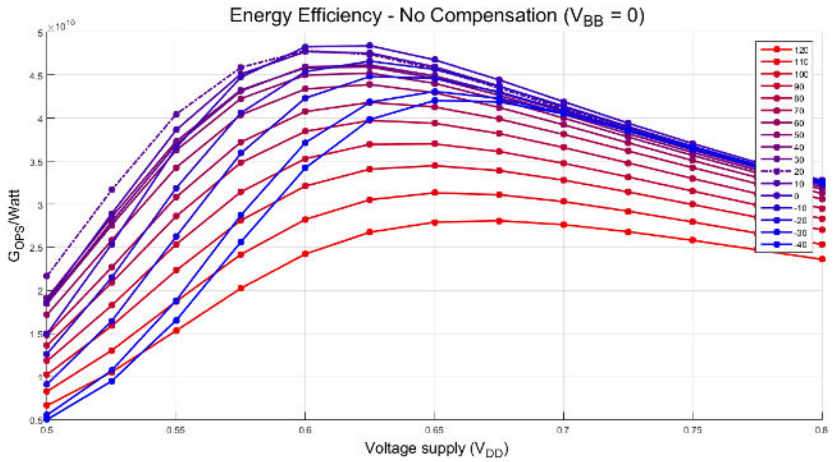
From sand to megawatts – Act 3: Alpha-Power Thermal Inversion





Temperature vs. peak performance vs. efficiency

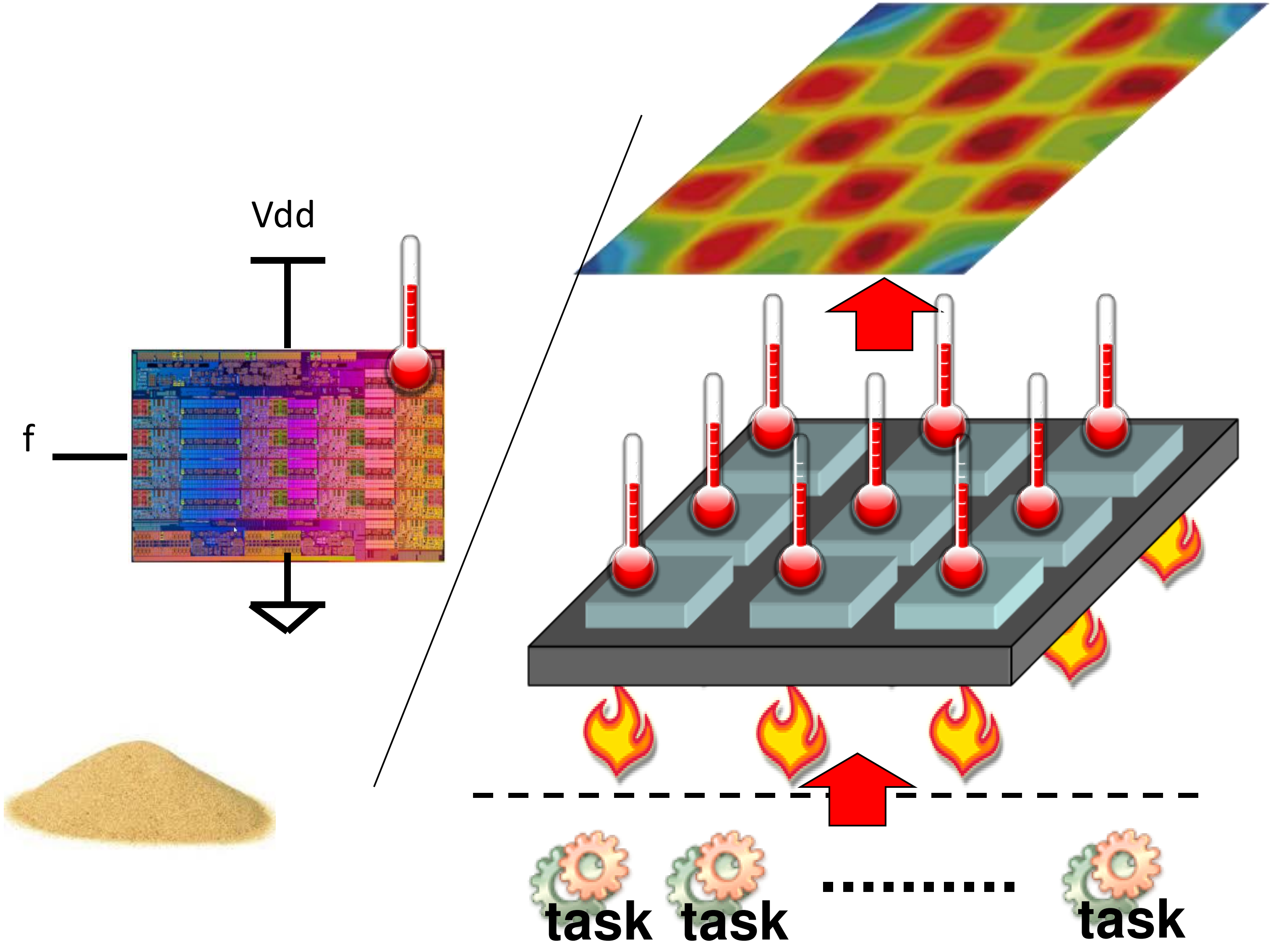




A 60 GOPS/W, −1.8 V to 0.9 V body bias ULP cluster in 28 nm UTBB FD-SOI technology,
D. Rossi,et al., Solid-State Electronics, 2016

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts – Act 4: Thermal Dissipation



$\text{Task}_j \rightarrow$ **Power model** $\xrightarrow{P_j}$ **Thermal model** $\xrightarrow{T_j}$

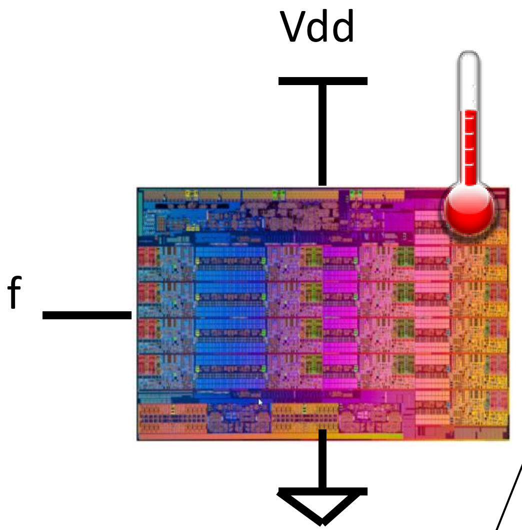
$$P = g(task, f)$$

Dynamic Model

$$ss\text{:}\ T[n+1] = AT[n] + BP[n]$$

$$tf\text{:}\ T[n+1] = \frac{B(z^{-1})}{A(z^{-1})} P[n]$$

From sand to megawatts – Act 4: Thermal Dissipation



Dynamic Model

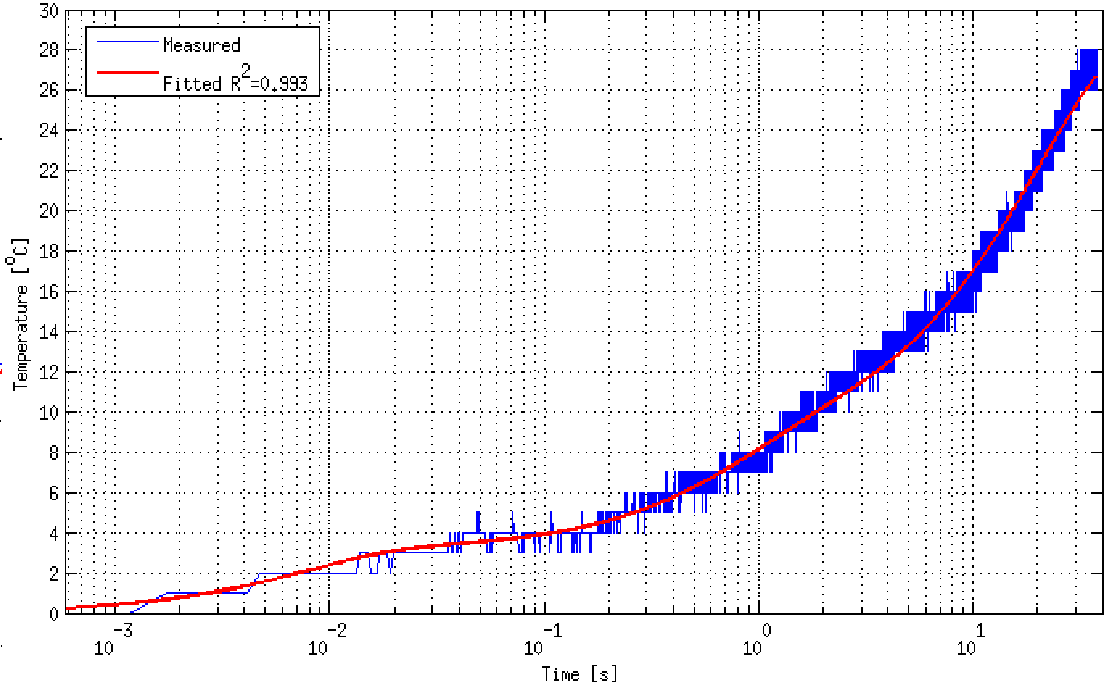
$$ss\colon T[n+1] = AT[n] + BP[n]$$

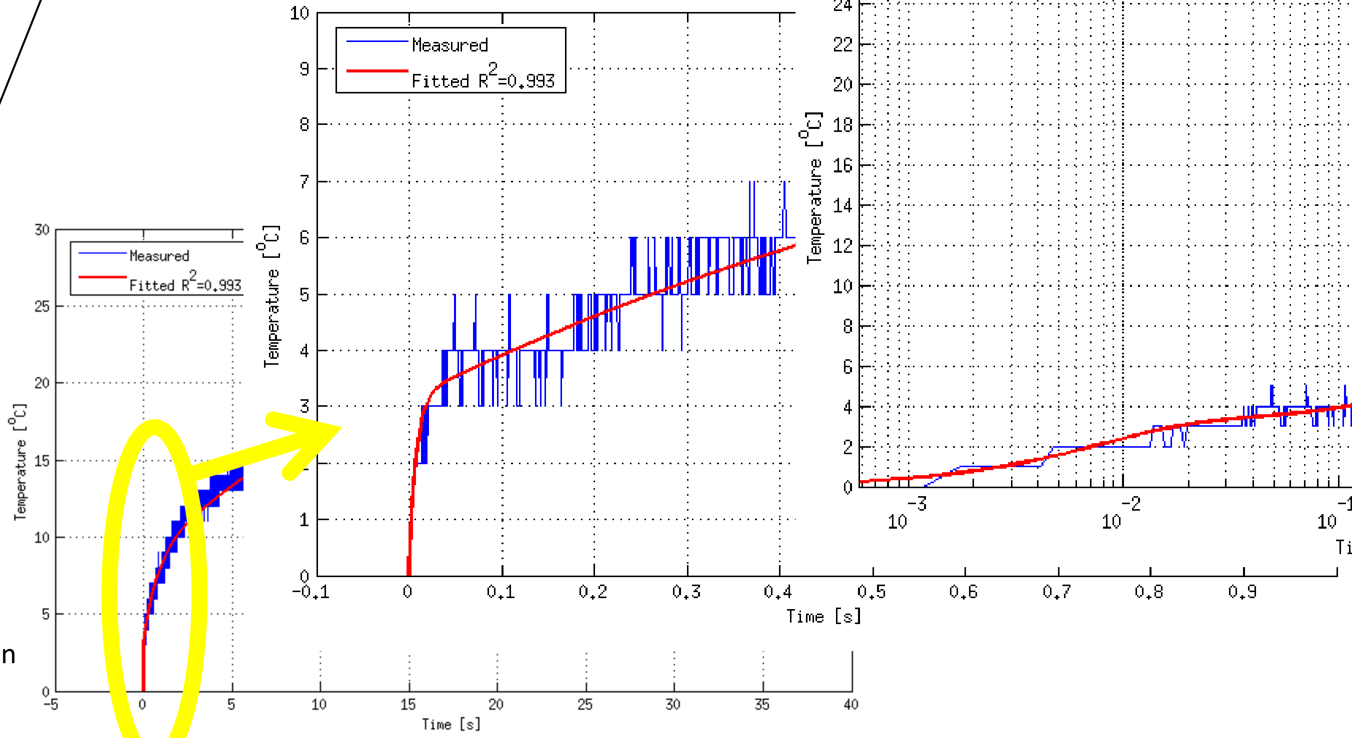
$$tf\colon T[n+1] = \frac{B(z^{-1})}{A(z^{-1})}P[n]$$

Thermal transient:
time constants
Model fitting

Tc@0.334ms	Time (s)
t1	0.0076
t2	0.7716
t3	20.0745

Is it temperature constant ?

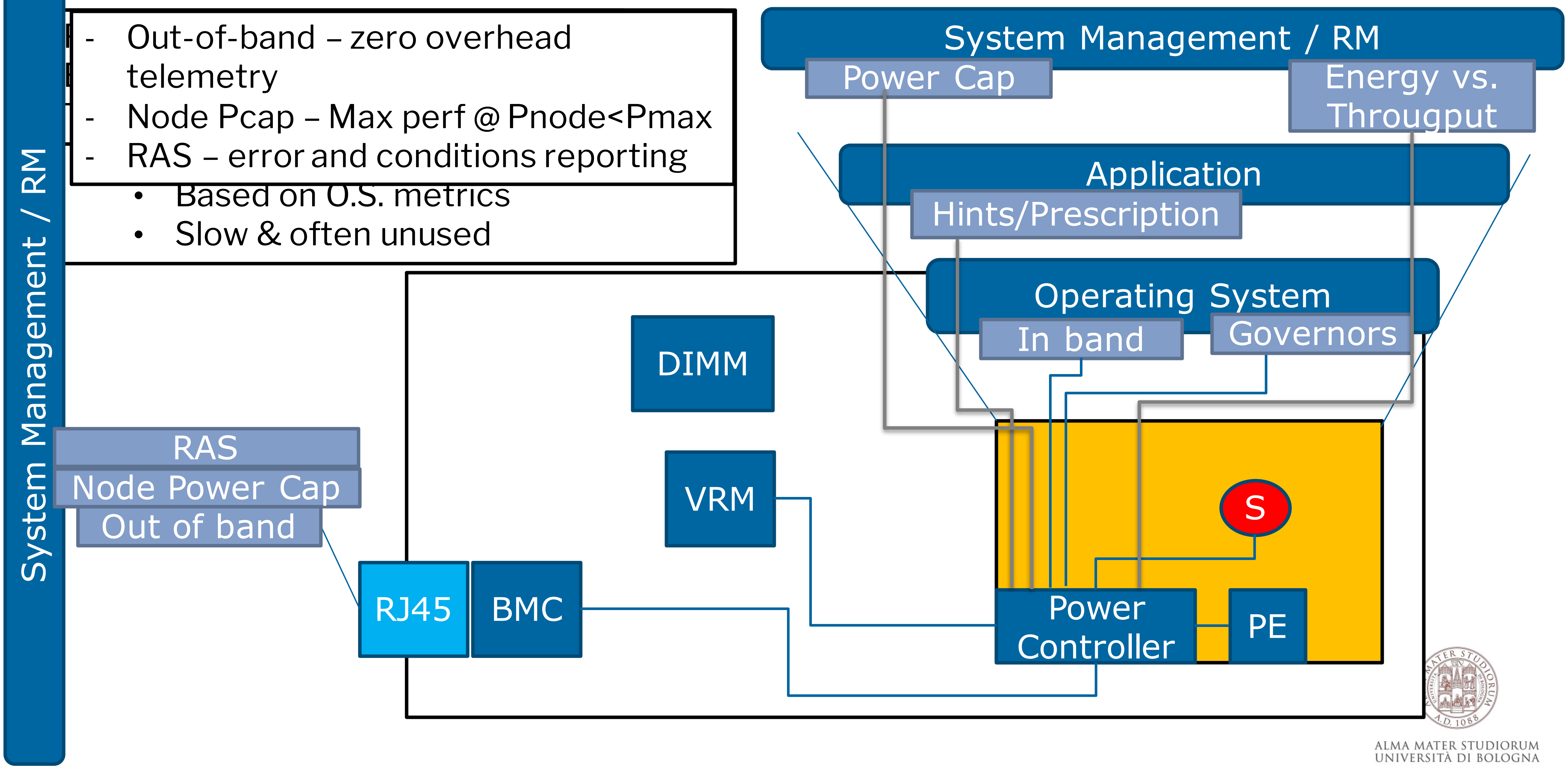




Beneventi et al., An Effective Gray-Box Identification Procedure for Multicore Thermal Modeling TC14

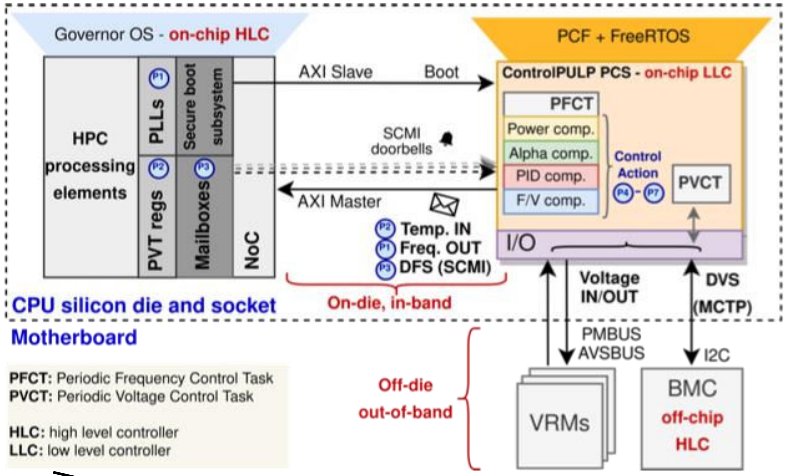
From sand to megawatts – Act 5: Compute Node power management

- Out-of-band – zero overhead telemetry
- Node Pcap – Max perf @ Pnode<Pmax
- RAS – error and conditions reporting
  - Based on O.S. metrics
  - Slow & often unused

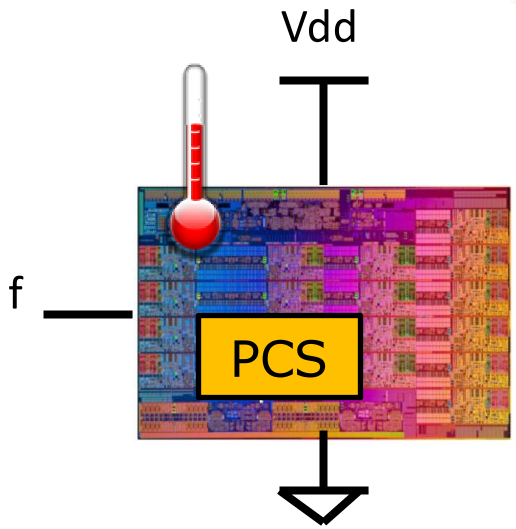


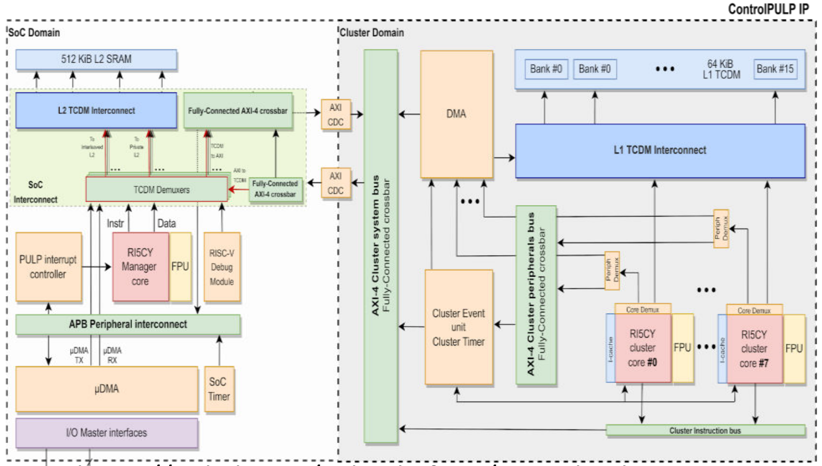
From sand to megawatts – Act 5: Compute Node power management

PCS co-design

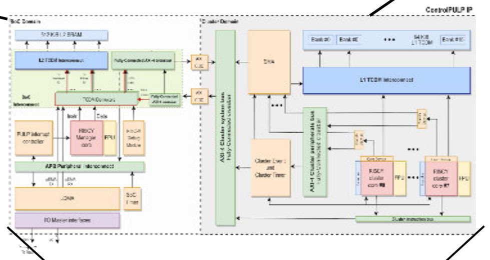


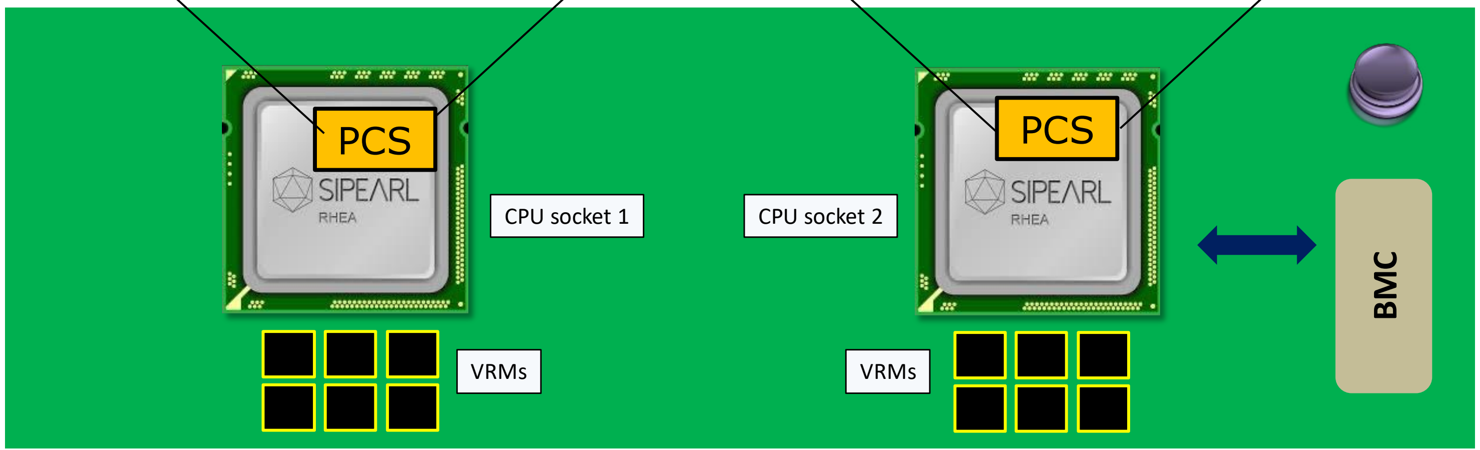
Ottaviano et al, ControlPULP: A RISC-V On-Chip Parallel Power Controller for Many-Core HPC Processors with FPGA-Based Hardware-In-The-Loop Power and Thermal Emulation, 2023





https://github.com/pulp-platform/control-pulp

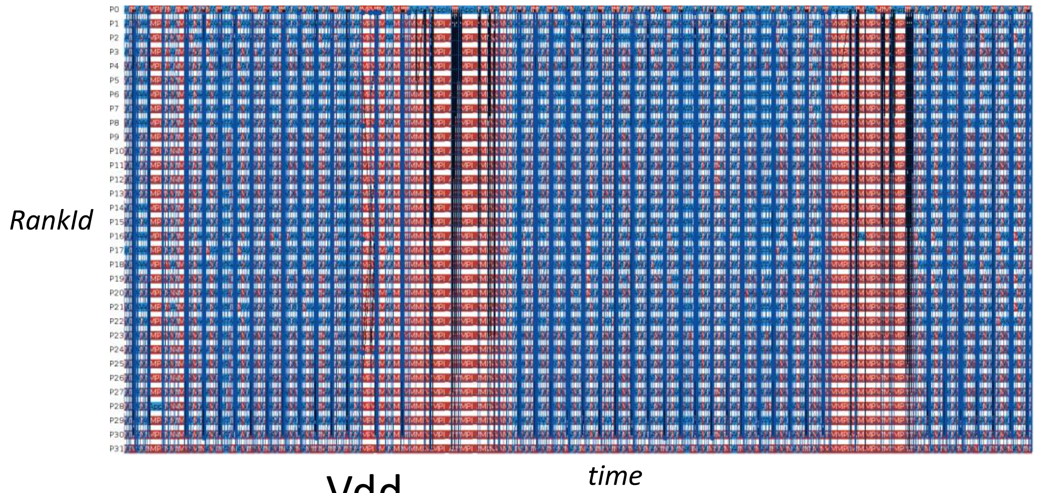




European Processor Initiative

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

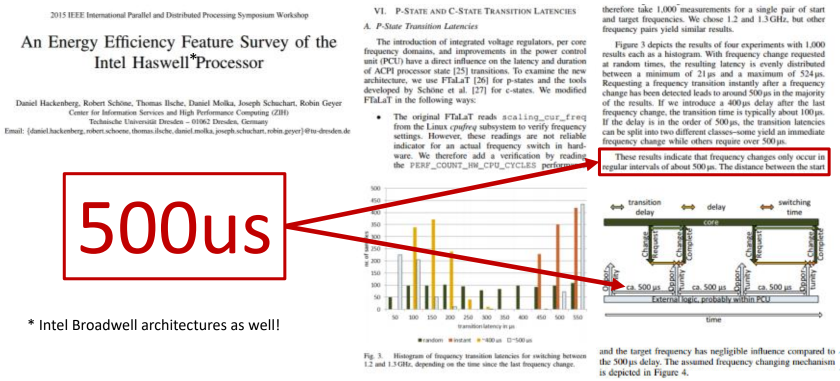
From sand to megawatts – Act 5: Compute Node power management



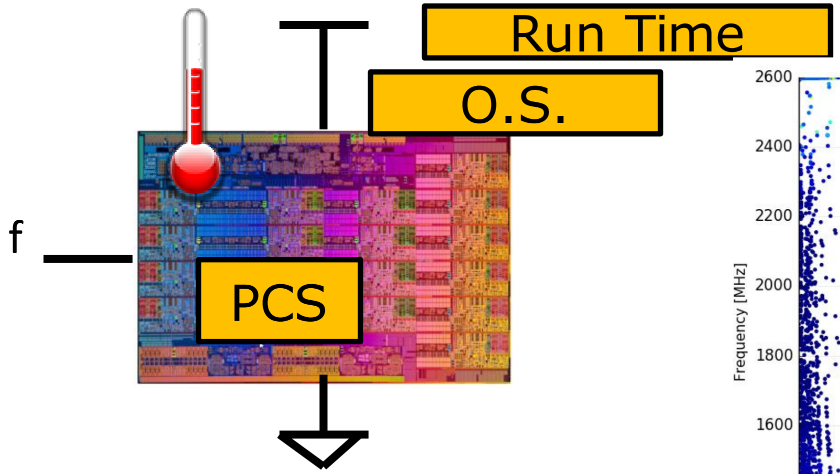
PCS & PM Runtime

Intel DVFS Power Manager

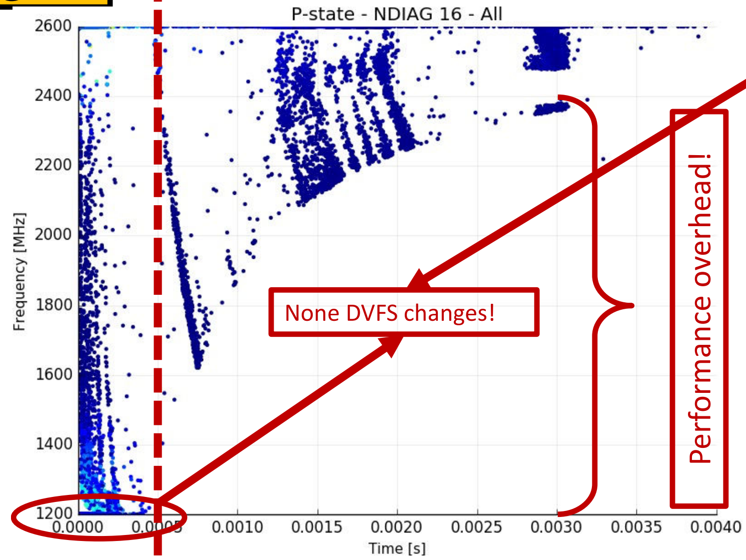
Today's HW power manager of Intel Architectures is quite slow in frequency variation!
Literatures studied this mechanism and, for reverse engineering, discovered a 500us latency!



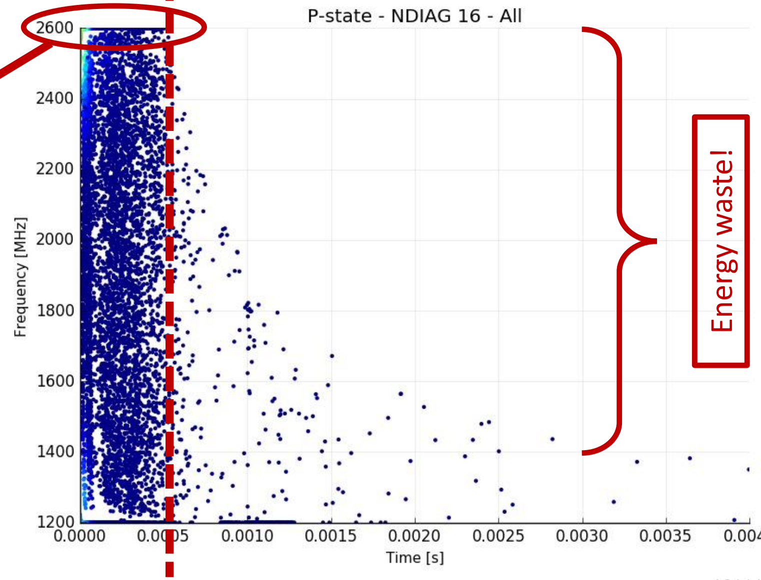
* Intel Broadwell architectures as well!



Application Phases



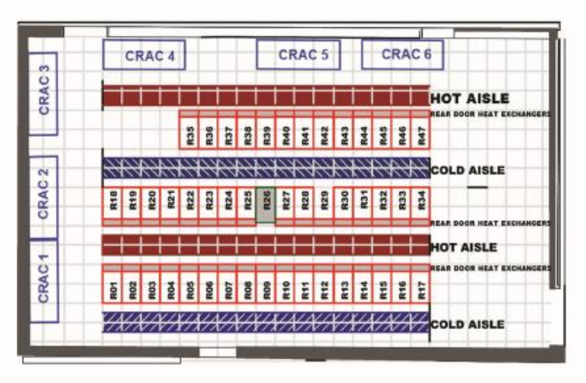
MPI Phases

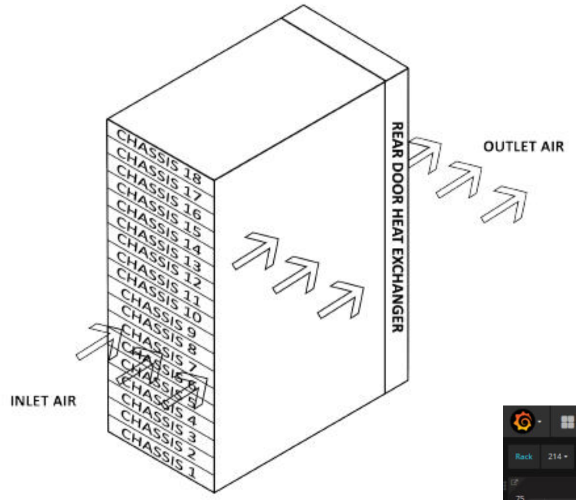


D. Cesarini et al. "COUNTDOWN: A Run-Time Library for Performance-Neutral Energy Saving in MPI Applications TC 2021

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

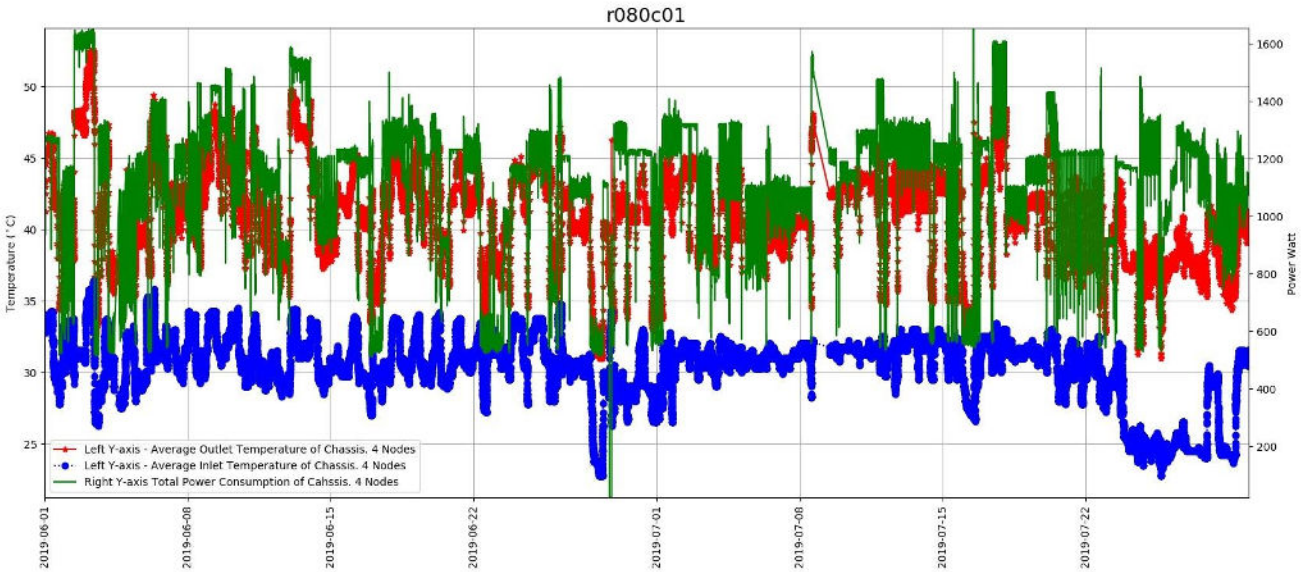
From sand to megawatts – Act 6: Room/DC Cooling Cost

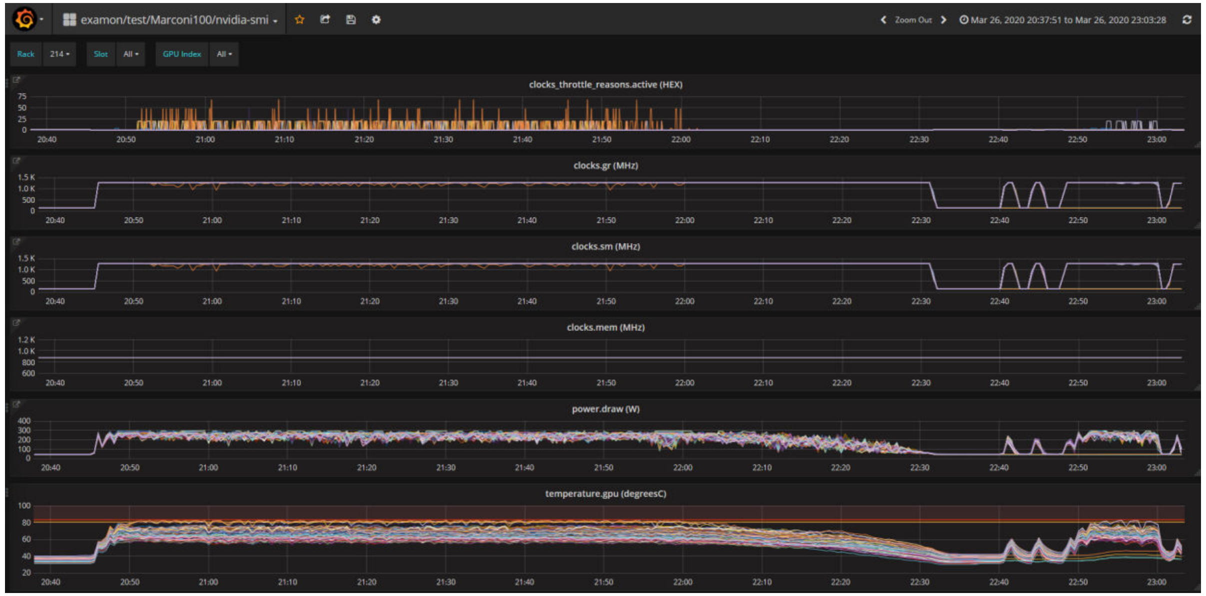




Power Usage Efficiency (PUE) $\triangleq \frac{P_{IT}+P_{COOLING}}{P_{IT}}$

PUE is always greater than 1 w. values ranging from: 1.0X to 1.40

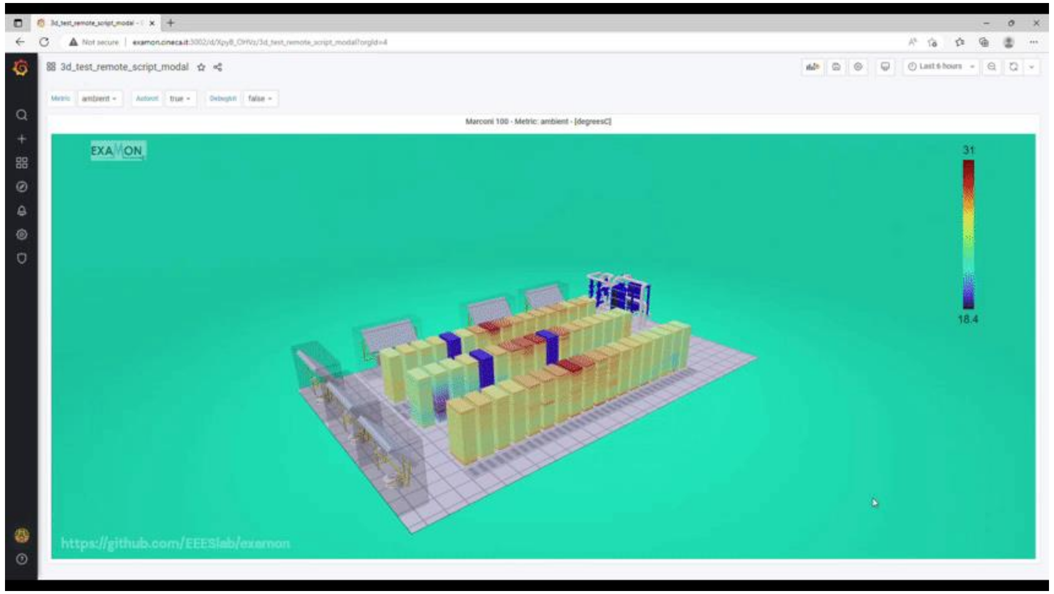




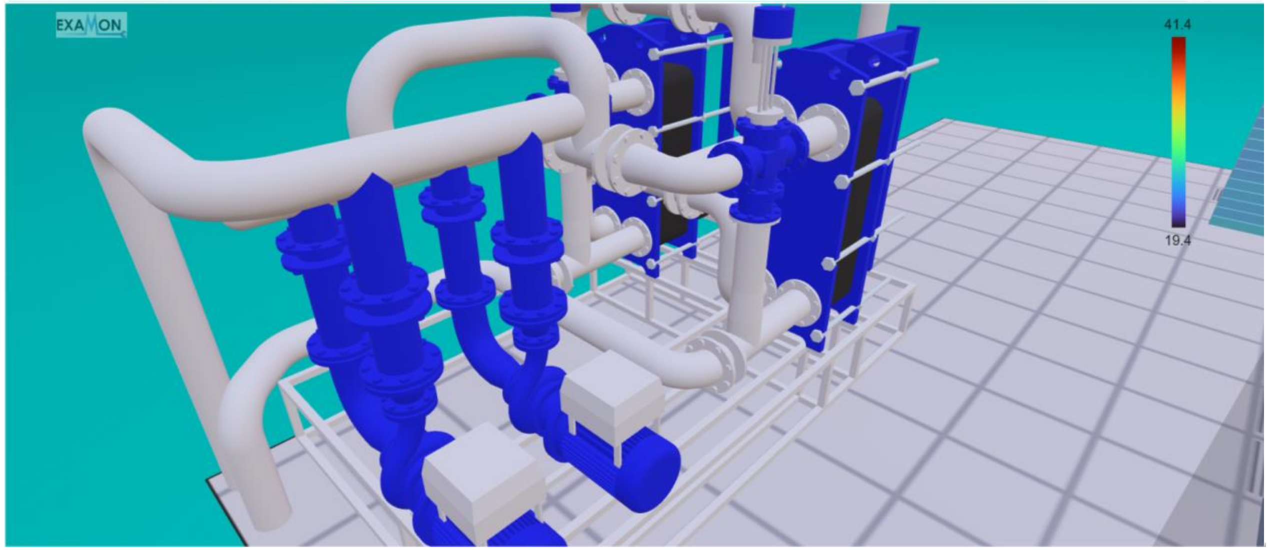
Bartolini et al, Paving the way toward energy-aware and automated datacentre ICPP19
Bartolini et al, Examon-x: a predictive maintenance framework for automatic monitoring in industrial iot systems JIOT21

From sand to megawatts – Act 7: Data-driven Large-scale Optimization

Using 3D visualization tool linked to the real time data provided by ExaMon can bring several benefits



- Visualization and Analysis
  - Helps identify and understand events and behaviors in relation to the location of objects.
  - Enables **XR** (VR/AR/MR) applications

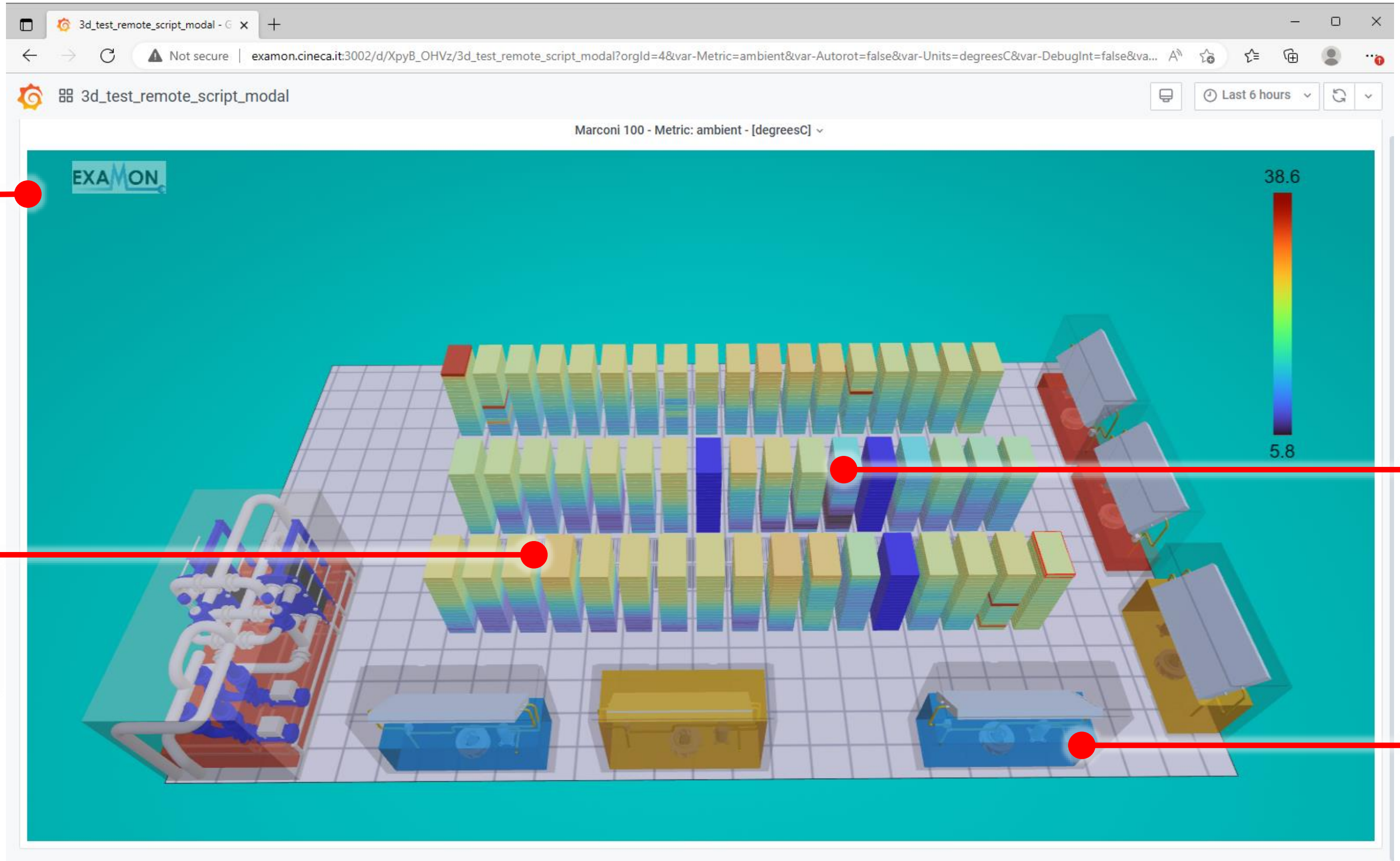


- Improved collaboration
  - Visualizing data and issues in a common and familiar visual representation enables better decision-making through improved communication and collaboration.

From sand to megawatts – Act 7: Data-driven Large-scale Optimization

PoC#1: Data center room power and thermal analysis

The overall view of the **heat generation and dissipation process** allows for a qualitative and immediate evaluation of the current cooling strategy.



Hot spots, thermal imbalances, and cooling efficiency can be quickly investigated to improve the overall system.

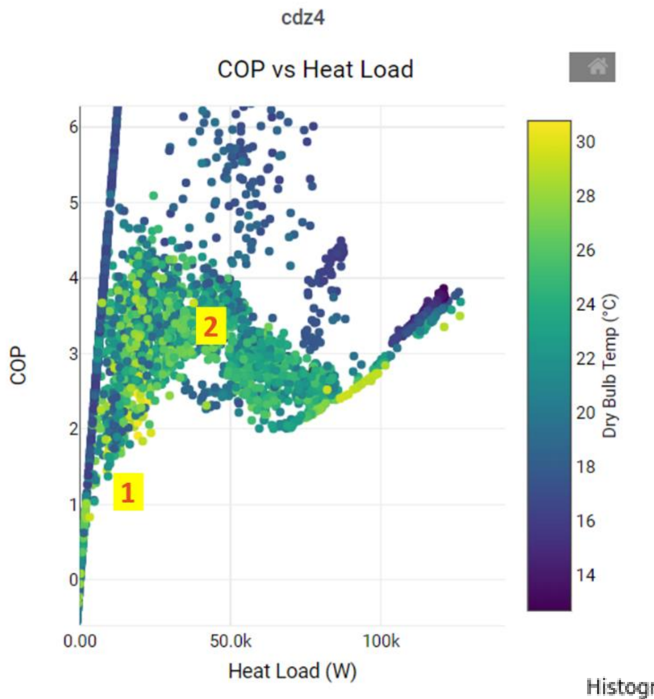
Node **power** consumption and **inlet temperatures** can be mapped and aggregated by node or by rack (average, total, max, ...).

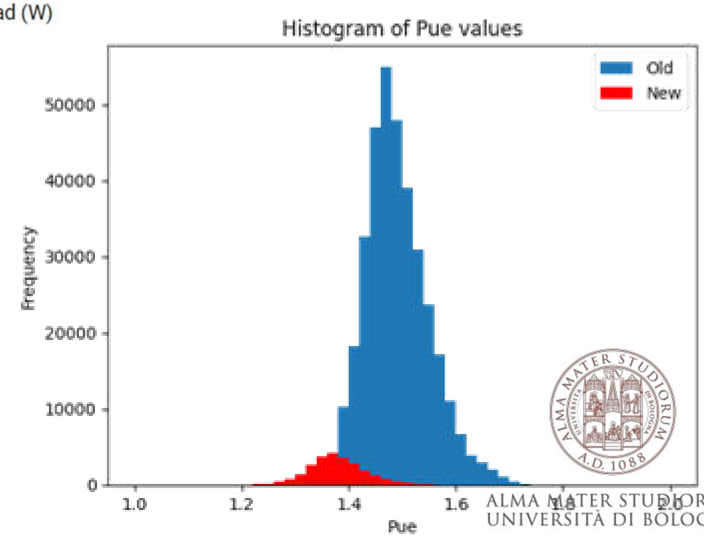
3D **widgets** can display complex metrics such as CRAC and chiller **cooling load** (bar height) and **efficiency** (color).

From sand to megawatts – Act 7: Data-driven Large-scale Optimization

PoC#2: @CINECA: Marconi 100 PUE optimization

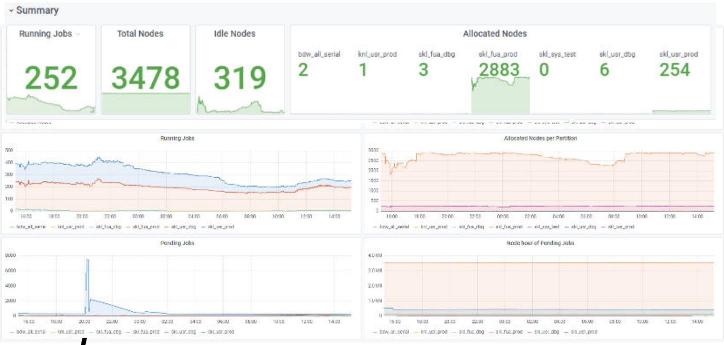
- Results obtained :
  - By analyzing the efficiency curves obtained using historical data, it was possible to determine the optimal operating point of the devices as a function of load, temperature and humidity.
  - Thanks to the immediate feedback provided by the dashboards, the operators were able to set the individual set points of the devices optimally.
  - During the trial period, we were able to achieve a **PUE reduction of approximately 8%** when compared to historical data measured under the same environmental and operating conditions.





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts – Act 7: Data-driven Large-scale Optimization w. AI !!

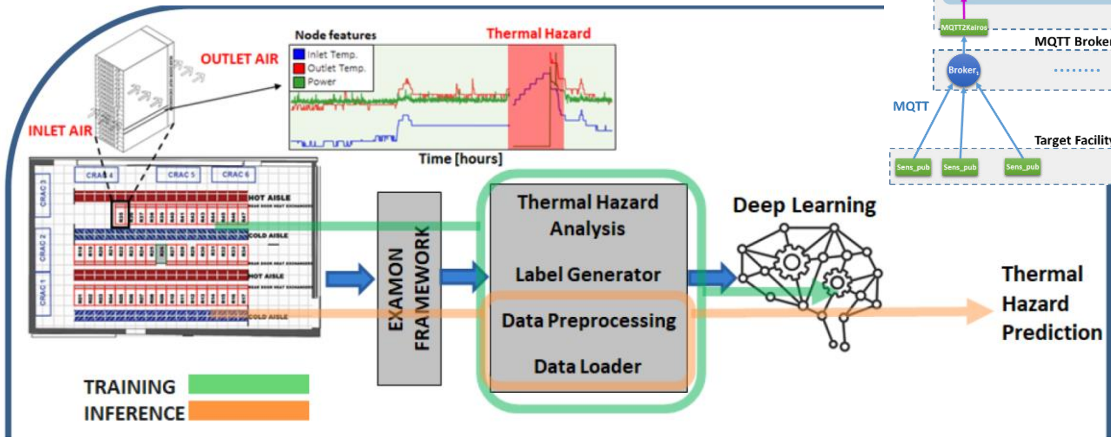


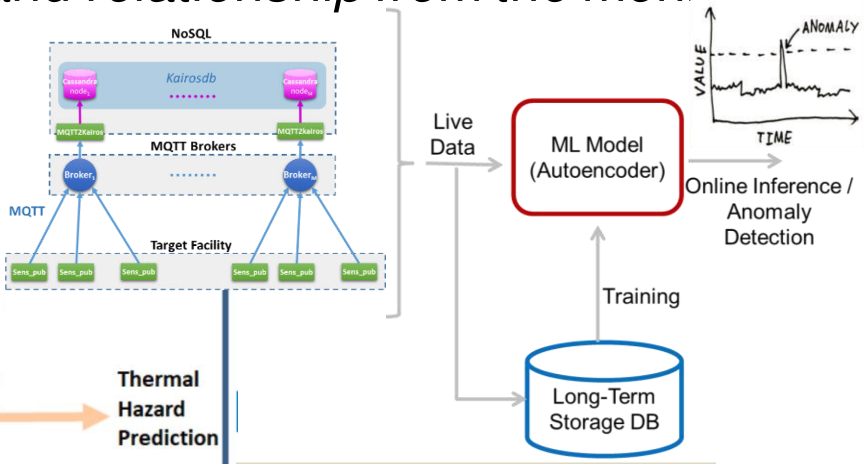
ExaMon@2021:

- Deployed on CINECA Datacentre since 2015
- Monitoring Operation, Facility, ICT and Users: >1M sensors, DB: 7TB online, 12GBs/Day, 21KSa/s
- Flexible dashboard for User Support, Admin and Facility managers

ExaMon + AI → **Anomaly Detection & Anticipation**!

Idea: use DL to extract normal behaviour and relationship from the monitored sensors.





Detect thermal hazards and cooling shortage

Detect node's miss-configurations & anomalies

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

From sand to megawatts – Act 7: Data-driven Large-scale Optimization w. AI !!

ExaData – open dataset – just released

scientific **data**

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > scientific data > data descriptors > article

Data Descriptor | Open Access | Published: 18 May 2023

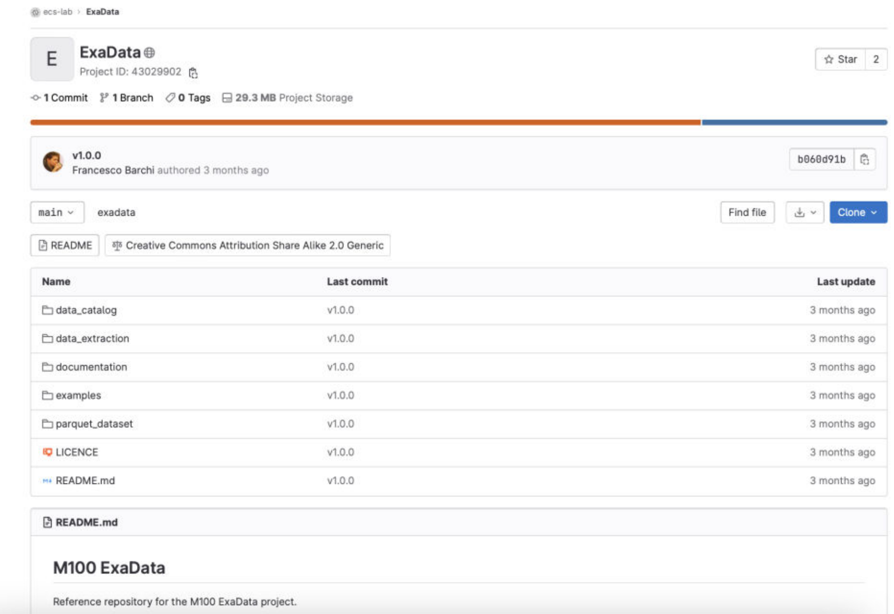
M100 ExaData: a data collection campaign on the CINECA's Marconi100 Tier-0 supercomputer

Andrea Borghesi, Carmine Di Santi, Martin Molan, Mohsen Seyedkazemi Ardebili, Alessio Mauri, Massimiliano Guarrasi, Daniela Galetti, Mirko Cestari, Francesco Barchi, Luca Benini, Francesco Beneventi & Andrea Bartolini

Scientific Data **10**, Article number: 288 (2023) | Cite this article

1 Altmetric | Metrics

https://www.nature.com/articles/s41597-023-02174-3



https://gitlab.com/ecs-lab/exadata

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts – Act 7: Data-driven Large-scale Optimization w. AI !!

[Antici et al, PM100: A Job Power Consumption Dataset of a Large-scale Production HPC System](#) SC23

ExaData description

- 31 months of data
- 573 metrics, 980+ nodes, approx. 50 TB uncompressed
- Vertiv, Schneider, IPMI, Ganglia, Logics, Weather, Nagios, SLURM, Job table
- Hardware data, system monitoring data, external information
- Different sampling granularities (from seconds) to minutes
- Zenodo + Nature Dataset

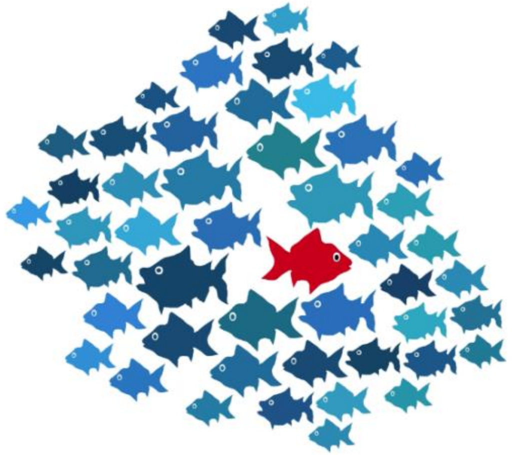
Plugin	#Metrics	#Plugin-specific columns	Description
Vertiv	25	1	Mainly collects data from the air-conditioning units (CDZ) located in room F (Marconi 100) of Cineca. The plugin uses the RESTful API interface available on the individual devices to extract the most interesting metrics.
Schneider	164	1	Dedicated data collector designed to acquire data from an industrial PLC by accessing its HMI module (from Schneider Electric). The PLC controls the valves and pumps of the liquid cooling circuit (RDHx) of Marconi 100. It consists of two (redundant) twin systems controllable by two identical HMI panels, Q101 and Q102.The ExaMon plugin extracts and stores all the metrics available on both panels.
IPMI	104	1	Collects all the sensor data provided by the OOB management interface (BMC) of cluster nodes.
Ganglia	177	1	Connects to the Ganglia server (gmond), collects and translates the data payload (XML) to the ExaMon data model.
Logics	37	2	Data collection system already installed at Cineca. It is specialized for collecting power consumption data from equipment in the different rooms, typically using multimeters that communicate via Modbus protocol. The ExaMon plugin dedicated to collecting this data interfaces to the Logics database (RDBMS) via its REST API. NOTE: Since the translation process is fully automated, the same inconsistencies present in the original db may result in the ExaMon database: e.g., metric names in the Italian language, units of measure as metric name, etc.
Weather	10	0	Collects all the weather data related to the Cineca facility location (Casalecchio di Reno) using an online open weather service (https://openweathermap.org).
Nagios	1	5	Interfaces with a Nagios extension developed by CINECA called "Hnagios", collects and translates the data payload to the ExaMon data model.
SLURM	54	4	Collects aggregated data from the SLURM server; these information is gathered through ad hoc scripts created by CINECA system administrators.
Job table	1	89	Collects information regarding the jobs executed on the cluster (and store in the SLURM database); the information collected are those provided by users at submission time.

https://www.nature.com/articles/s41597-023-02174-3

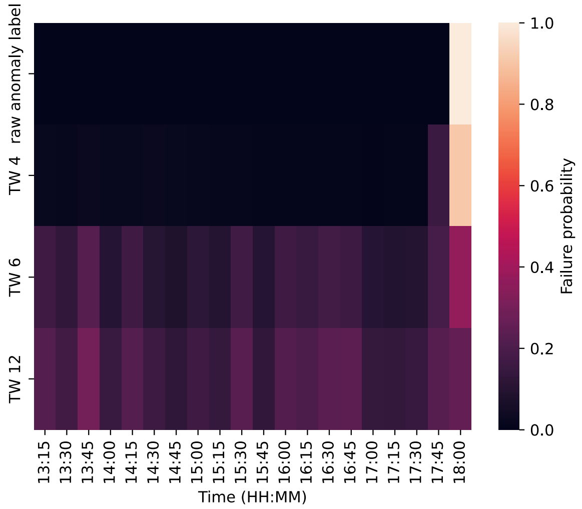
https://gitlab.com/ecs-lab/exadata

From sand to megawatts – Act 7: Data-driven Large-scale Optimization w. AI !!

Datacenter Automation (Anomaly Detection & Anticipation)



- Detect anomalies/faults in a HPC system
- Hundreds/thousands of possible sources:
  – HW components that malfunction, breakages, misconfigurations, intruders, etc.
- Strong incentive to automatize the detection process
  - Downtime are *very* expensive
  - It's better to identify a problem as soon as possible

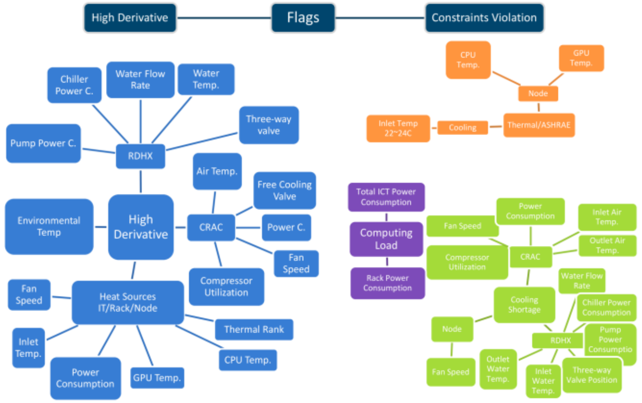


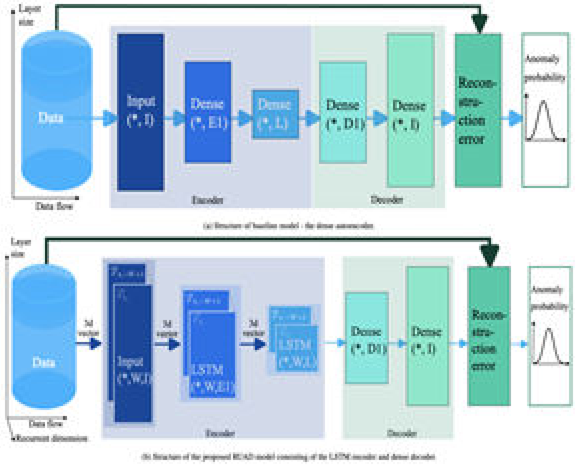
- **Solution: DL models that can distinguish anomalies from normal situations**

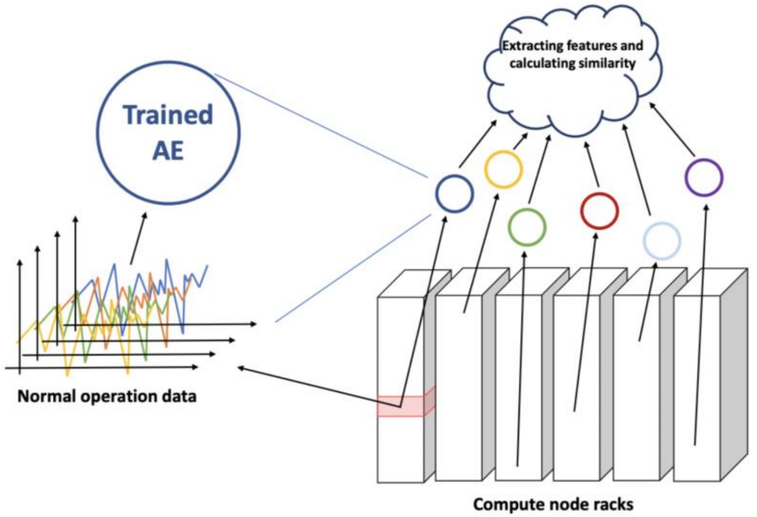
Borghesi et al., “Anomaly Detection using Autoencoders in High Performance Computing Systems”, AAAI'19
Borghesi et al, “Online anomaly detection in hpc systems”, AICAS’1
Borghesi et al., "A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems", EAAI 2019
Molan et al, RUAD: Unsupervised anomaly detection in HPC systems, FGCS23
Molan et al. Graph Neural Networks for Anomaly Anticipation in HPC Systems ICPE23

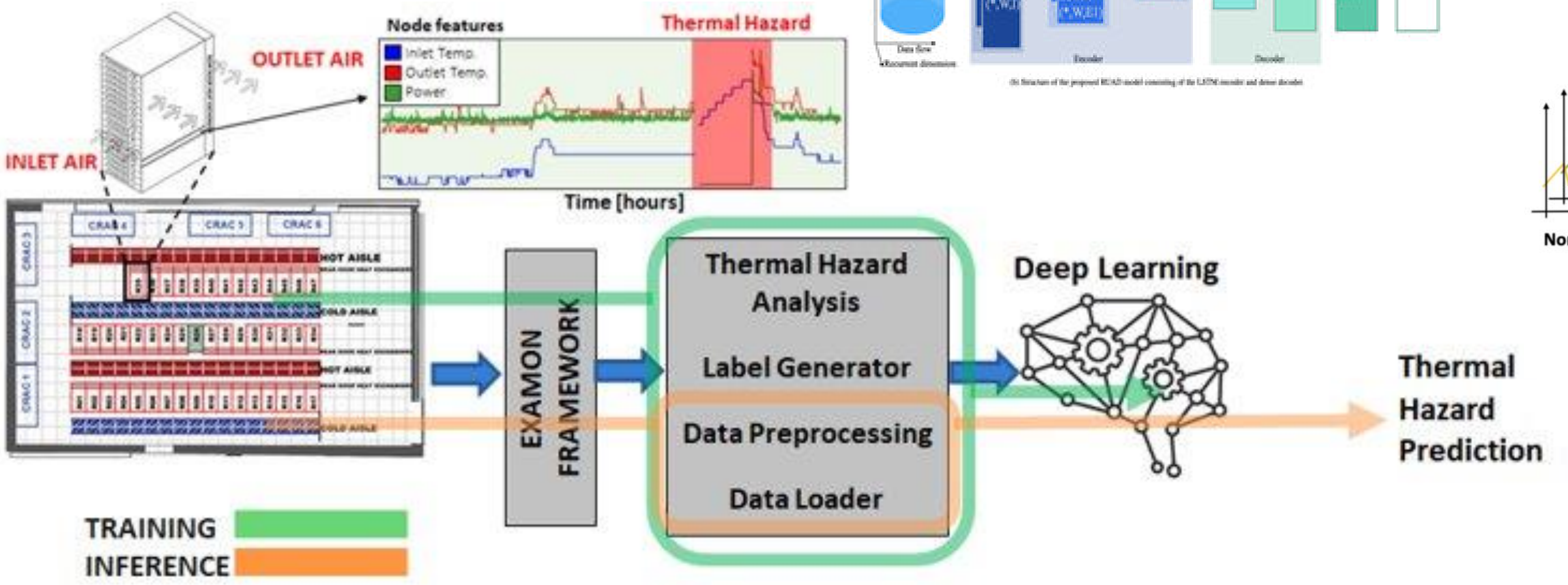
Example – Anomaly Detection per node-based or full room based

RUAD: Unsupervised anomaly detection in HPC systems, FGCS23
Rule-Based Thermal Anomaly Detection for Tier-0 HPC Systems ISC22
Examon-x: a predictive maintenance framework for automatic monitoring in industrial iot systems JIOT21
Integrated energy-aware management of supercomputer hybrid cooling systems TII06
Graph Neural Networks for Anomaly Anticipation in HPC Systems ICPE23







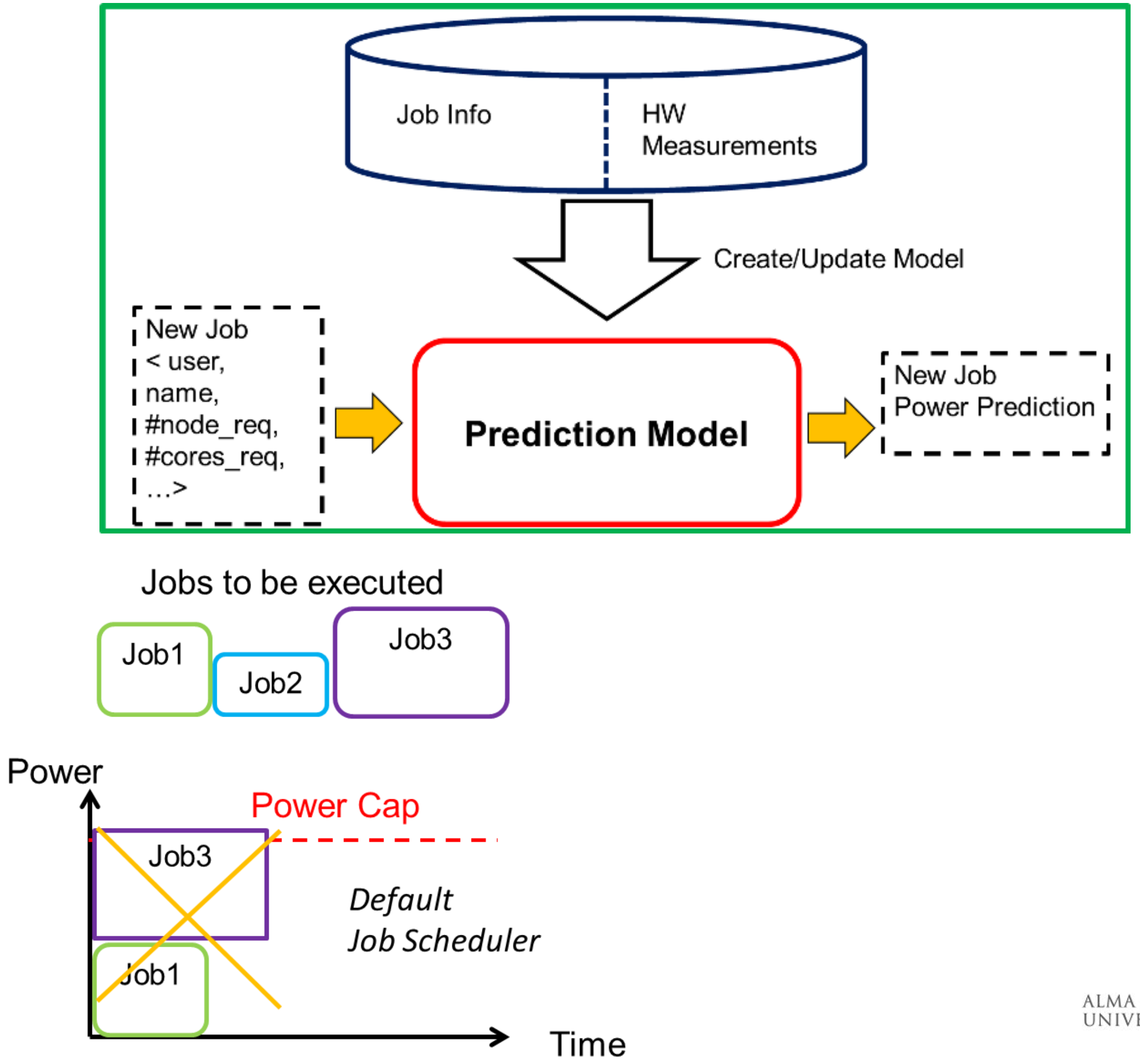


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts – Act 7: Data-driven Large-scale Optimization w. AI !!

Job power prediction

1. Machine Learning models to predict the power consumption of HPC applications
2. Slurm Custom Extensions to schedule jobs based on their power
3. Interacts with power management



Aknowledgment

European Processor Initiative

EUPEX European Pilot for Exascale

REGALE Open Architecture for Exascale Supercomputers

GRAPH MASSIVIZER

The european-project-initiative has received funding from the European High Performance Computing Joint Undertaking (JU) under Framework Partnership Agreement No 800928 and Specific Grant Agreement No 101036168 (EPI SGA2). The JU receives support from the European Union's Horizon 2020 research and innovation programme and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland.

The EUPEX project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No.101034126. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Switzerland, Germany, France, Greece, Sweden, Croatia and Turkey.

This REGALE-project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956560. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Greece, Germany, France, Spain, Austria, Italy.

The **Spoke Future HPC** of the Italian **National Center of HPC, Big Data e Quantum Computing** is funded by the **National Recovery and Resilience Plan (NRRP).**

CINECA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA