



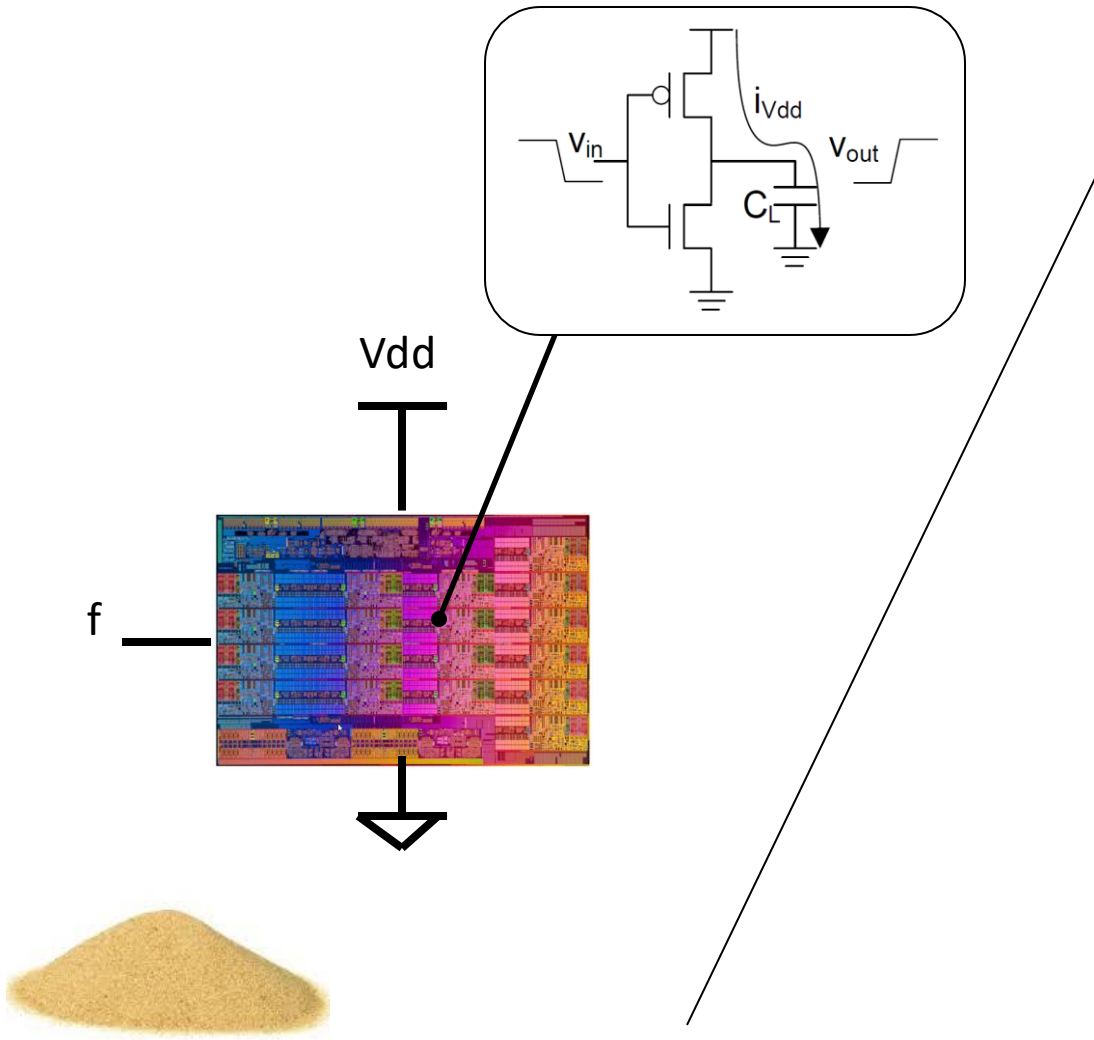
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

From sand to megawatts in seven acts

Prof Andrea Bartolini

DEI, a.bartolini@unibo.it

From sand to megawatts – Act 1: Dynamic Power

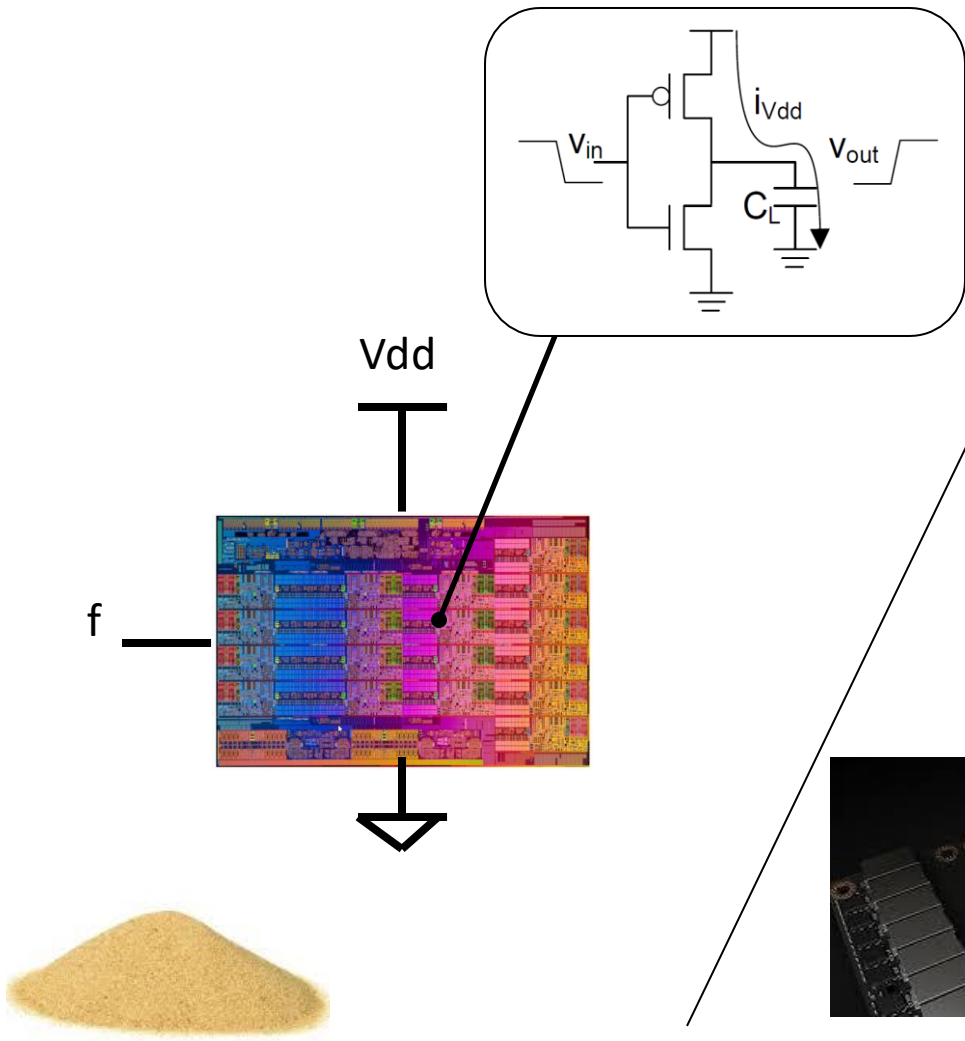


$$P_{dynamic} = \underbrace{a \times C_L}_{\text{"effective capacitance"} (C_{Effective})} \times V_{dd}^2 \times f$$

activity factor load capacitance supply voltage clock frequency



From sand to megawatts – Act 1: Dynamic Power

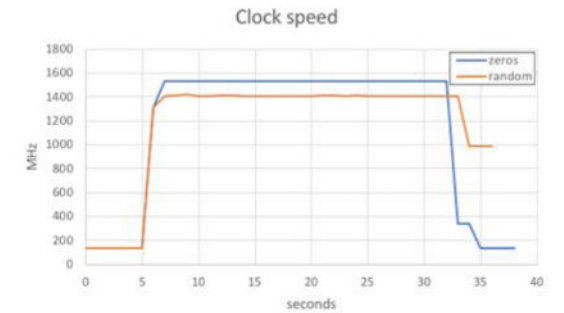
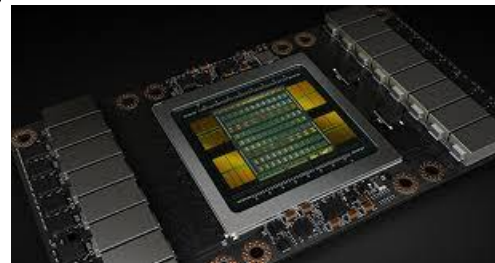


$$P_{dynamic} = \underbrace{a}_{\text{activity factor}} \times \underbrace{C_L}_{\text{load capacitance}} \times V_{dd}^2 \times \underbrace{f}_{\text{clock frequency}}$$

“effective capacitance” ($C_{Effective}$)

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

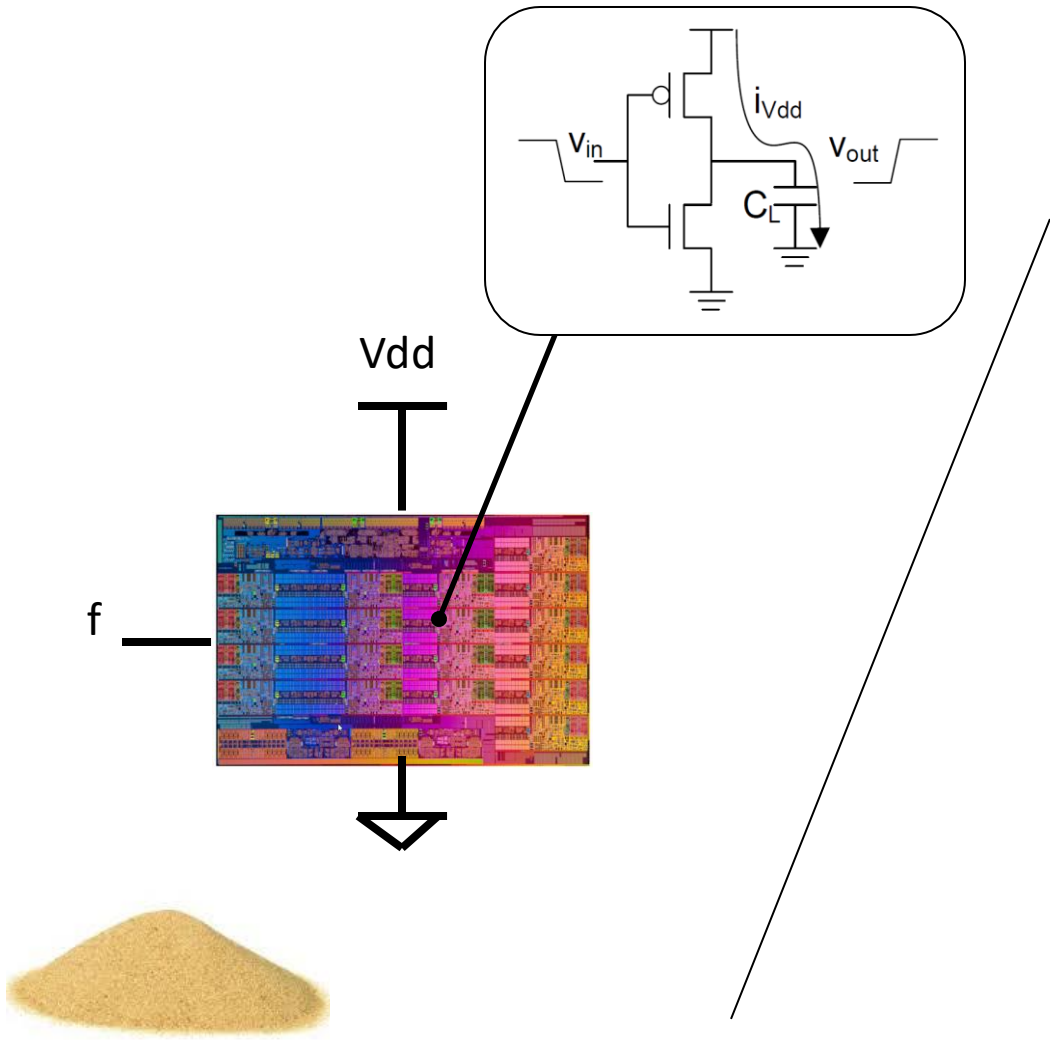


<https://indico-jsc.fz-juelich.de/event/76/session/0/contribution/1/material/slides/0.pdf>

Wayne Joubert - OpenPOWER ADG 2018



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



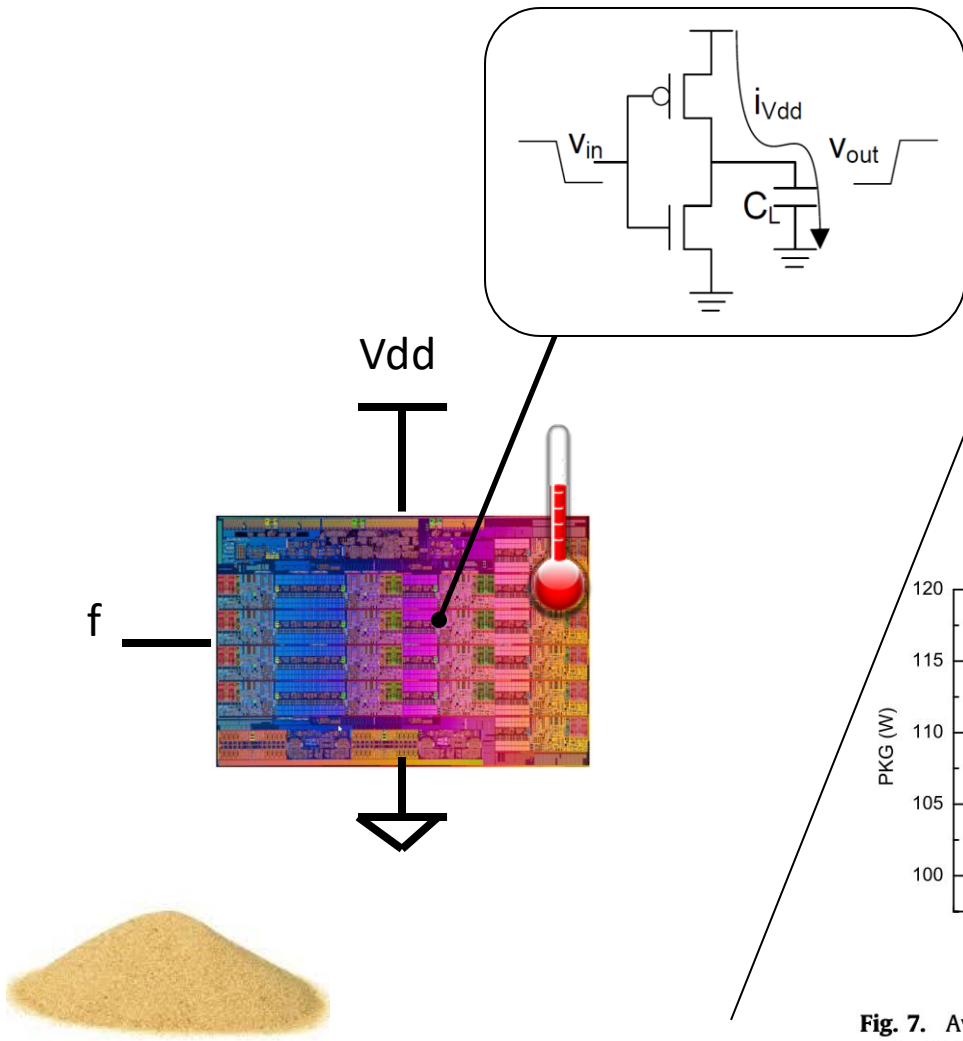
$$I_{leakage} = k_1 \times \left(1 - e^{-k_2 \times V_{ds} / T}\right) \times e^{k_3 \times (V_{gs} - V_{TH} - V_{off}) / T}$$

constants → k_1
 temperature → T
 drain to source voltage → V_{ds}
 gate to source voltage → V_{gs}
 threshold voltage → V_{TH}
 empirical parameter → V_{off}

f



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



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

Annotations for the equation:

- k_1 : constants
- k_2 : constants
- k_3 : constants
- V_{ds} : drain to source voltage
- V_{gs} : gate to source voltage
- V_{TH} : threshold voltage
- V_{off} : empirical parameter
- T : temperature

Temperature in a power-limited architecture

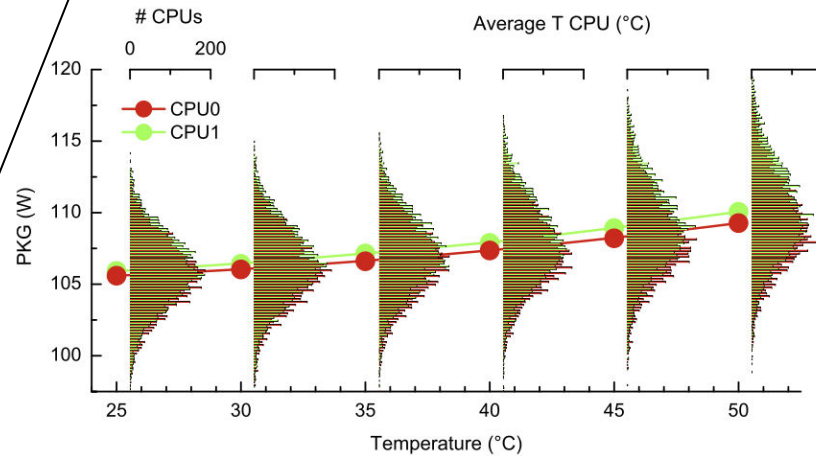


Fig. 7. Average power consumption distribution of SuperMUC Phase1 compute nodes (Intel Sandy Bridge-EP Xeon E5-2680 8C) running single node HPL (Turbo OFF) at different inlet water temperatures.

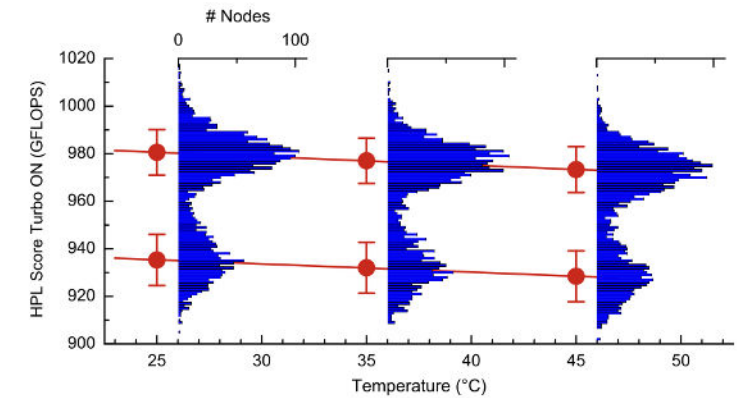
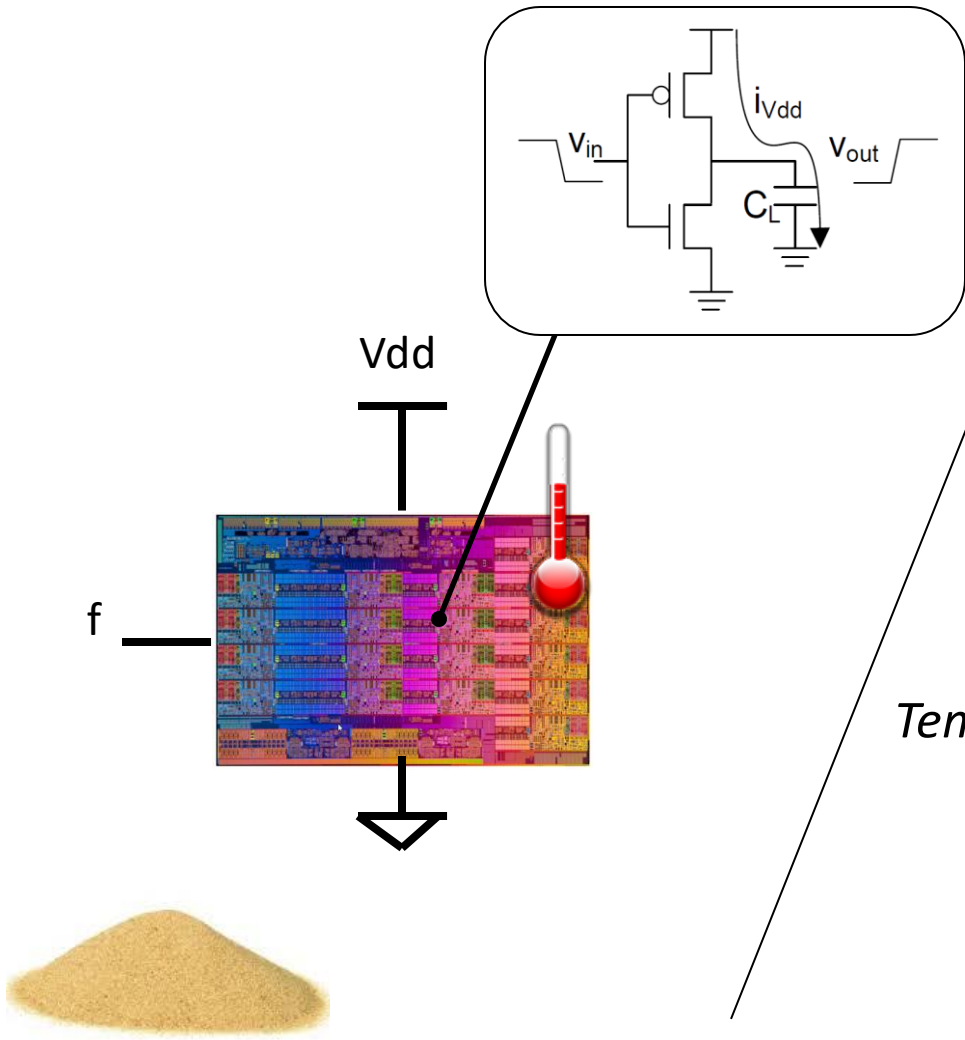


Fig. 5. Performance distribution of SuperMUC Phase2 compute nodes (Intel Haswell Xeon E5-2697 v3) at different inlet water temperatures.

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



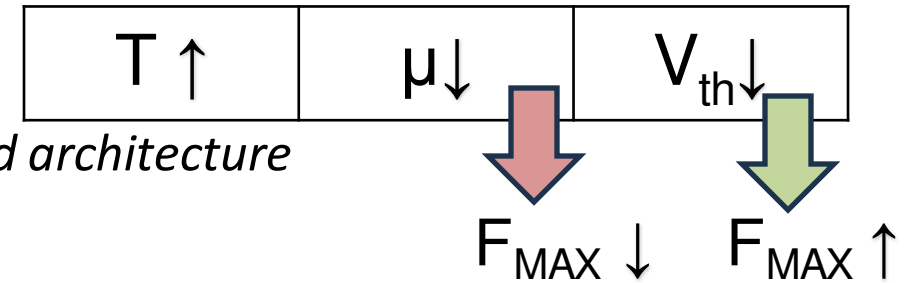
Delay:
$$D_p = \frac{C_{out} V_{dd}}{I_{ON}} = \frac{C_{out} V_{dd}}{\mu(T)[V_{dd} - V_{th}(T)]^\alpha}$$

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

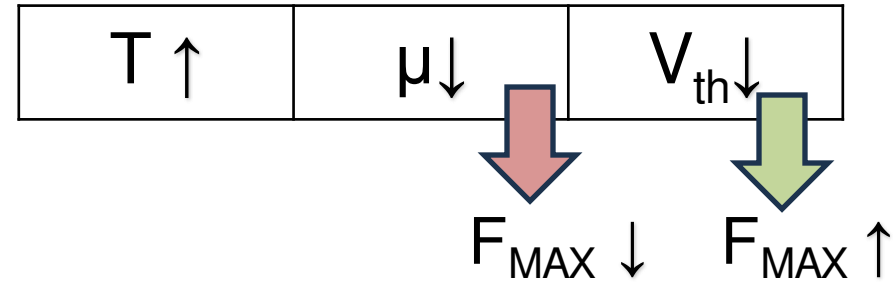
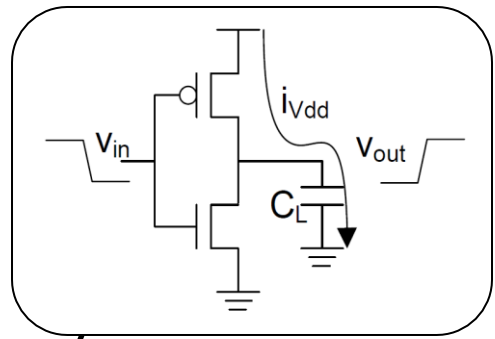
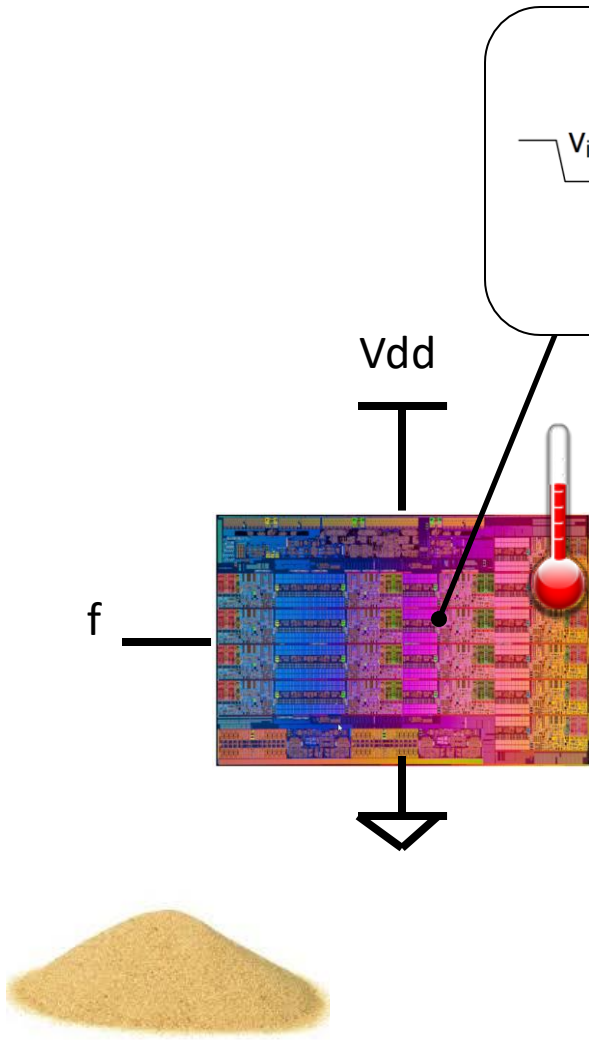
Threshold Voltage:
$$V_{th} = V_{th}(T_0) - k(T - T_0)$$



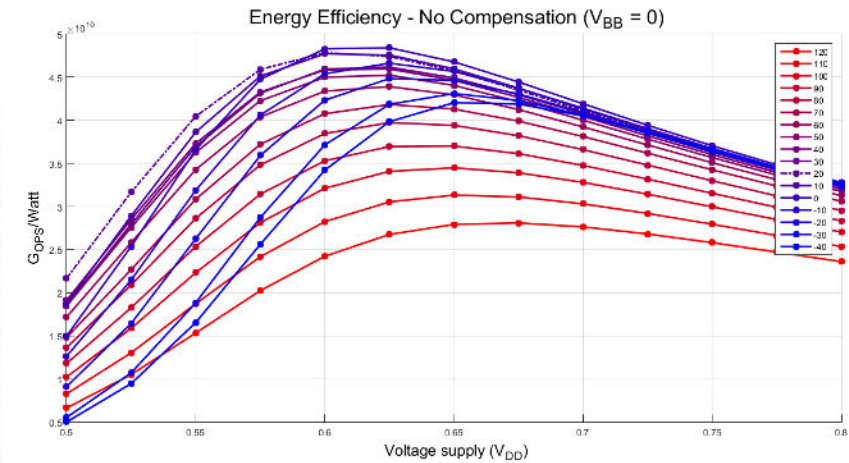
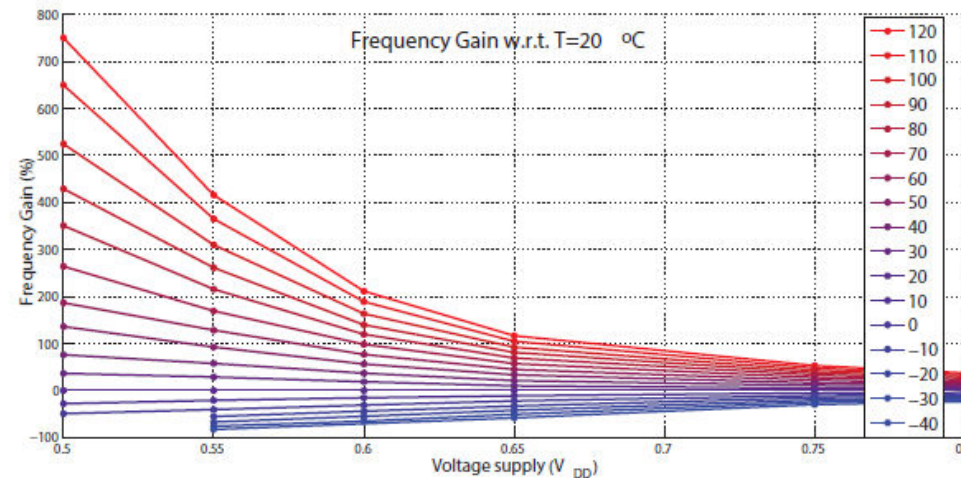
Temperature in a power-limited architecture



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



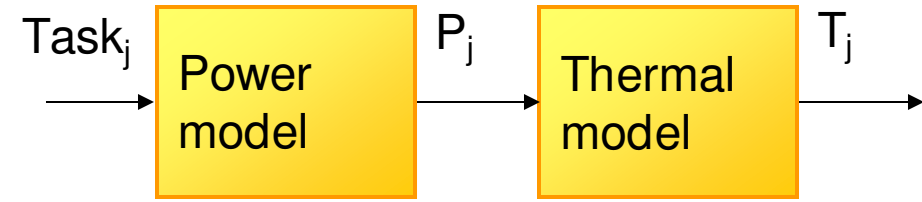
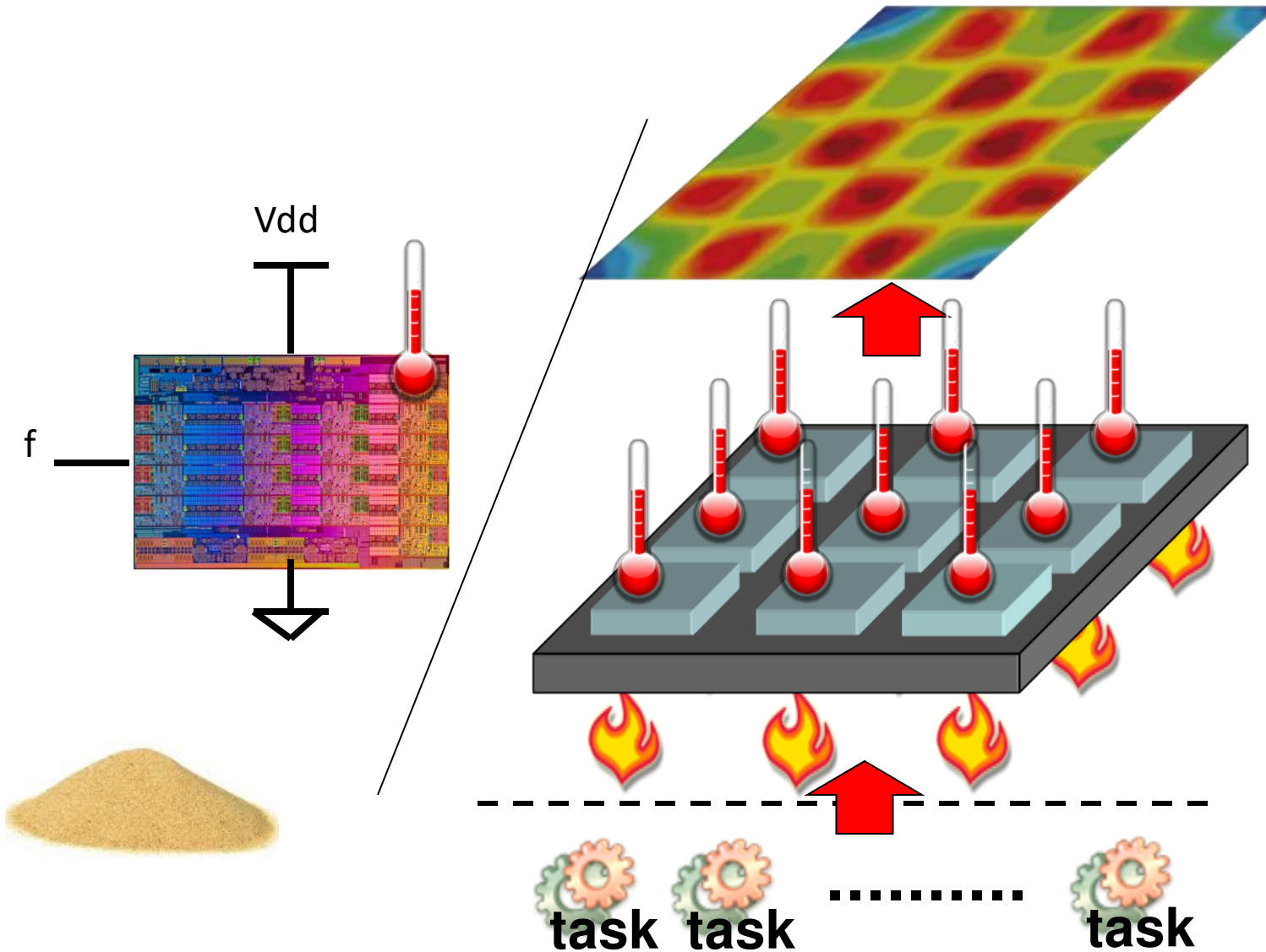
Temperature vs. peak performance vs. efficiency



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



From sand to megawatts – Act 4: Thermal Dissipation



$$P = g(\text{task}, f)$$

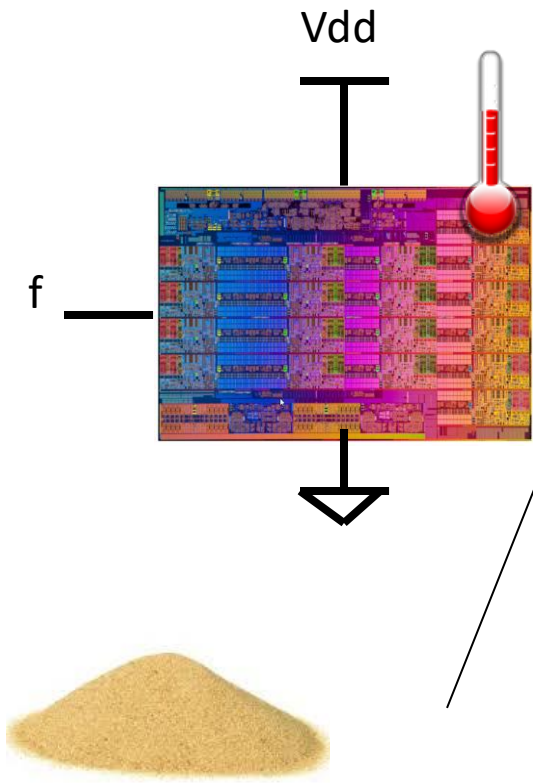
Dynamic Model

$$ss: T[n + 1] = AT[n] + BP[n]$$

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



From sand to megawatts – Act 4: Thermal Dissipation



Dynamic Model

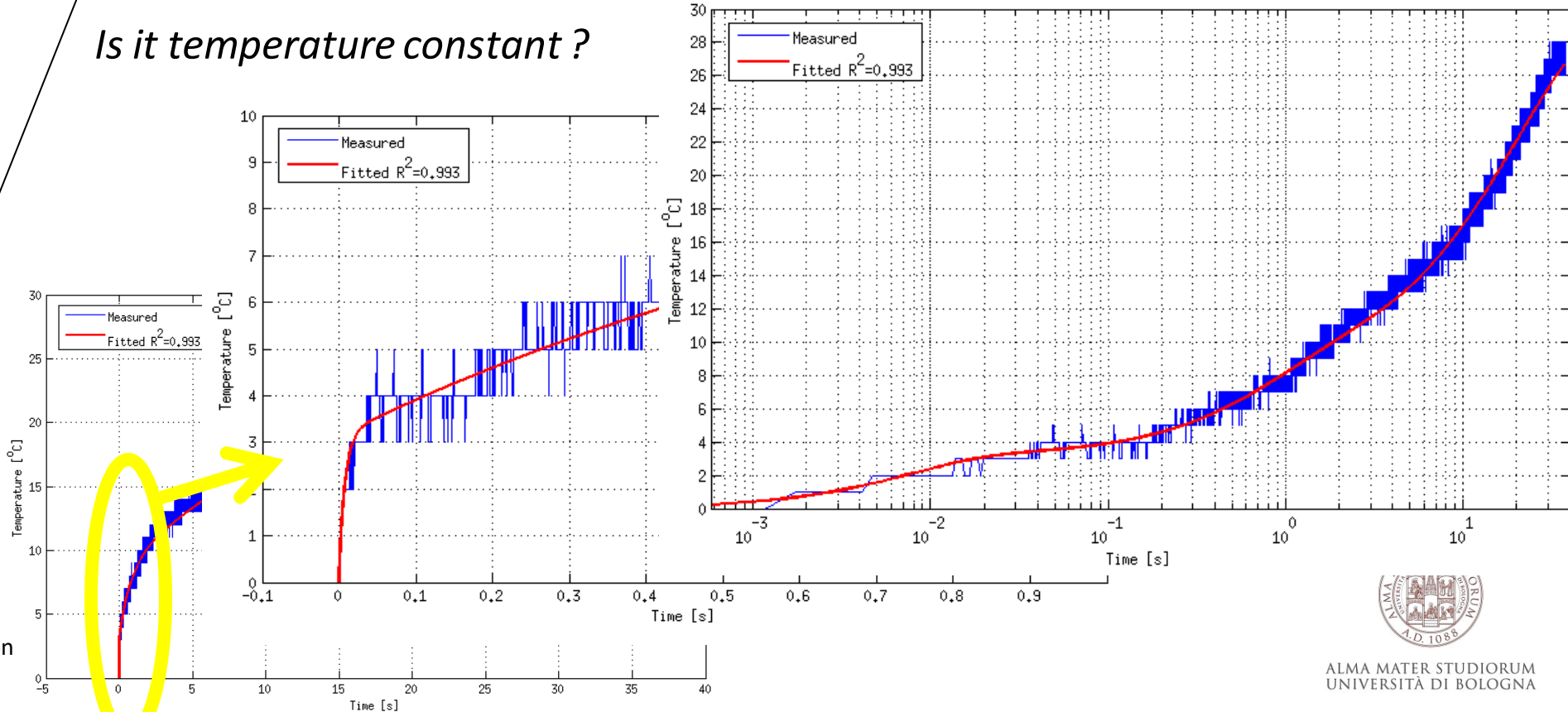
$$ss: T[n + 1] = AT[n] + BP[n]$$

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

Thermal transient:
time constants
Model fitting

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

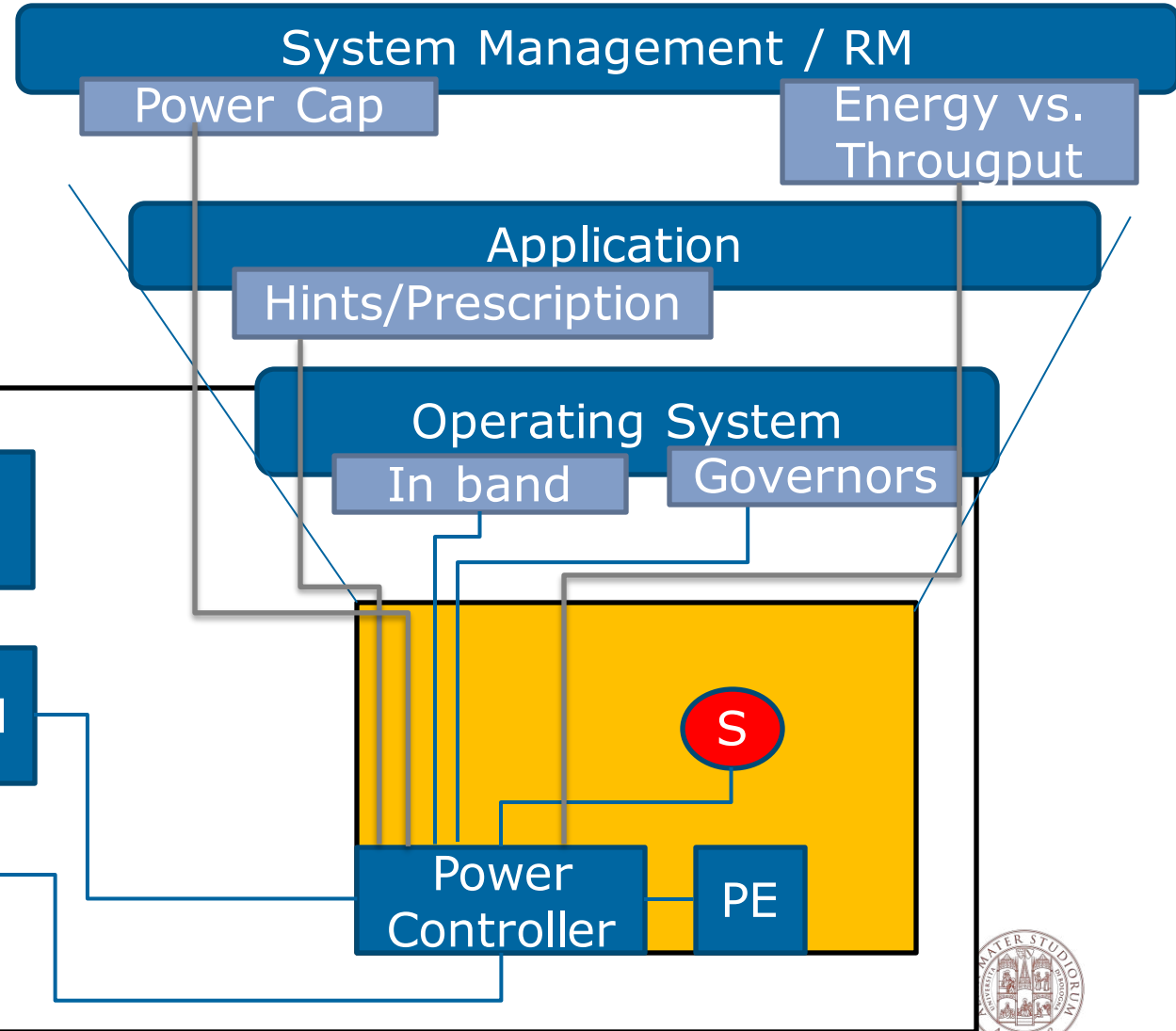
Is it temperature constant?



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

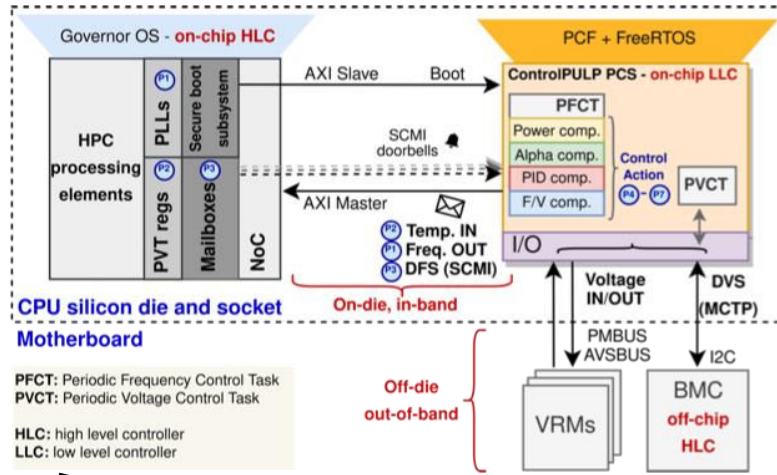
System Management / RM

- Out-of-band – zero overhead telemetry
- Node Pcap – Max perf @ $P_{node} < P_{max}$
- RAS – error and conditions reporting
 - Based on U.S. metrics
 - Slow & often unused



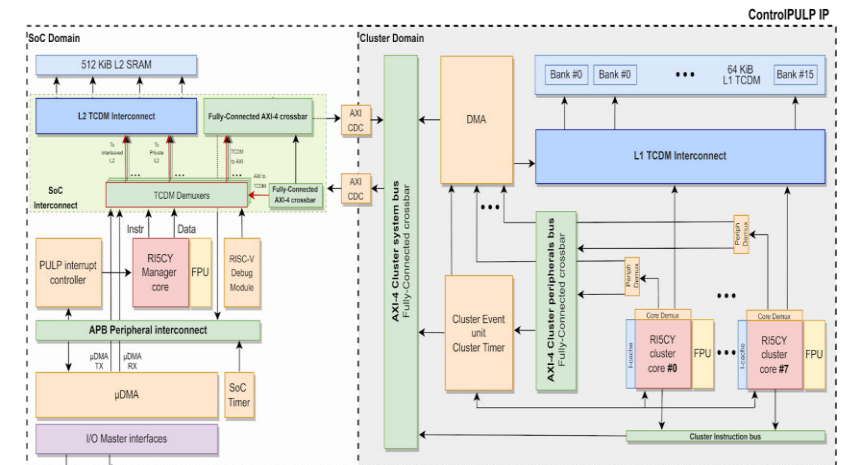
From sand to megawatts – Act 5: Compute Node power management

PCS
co-design

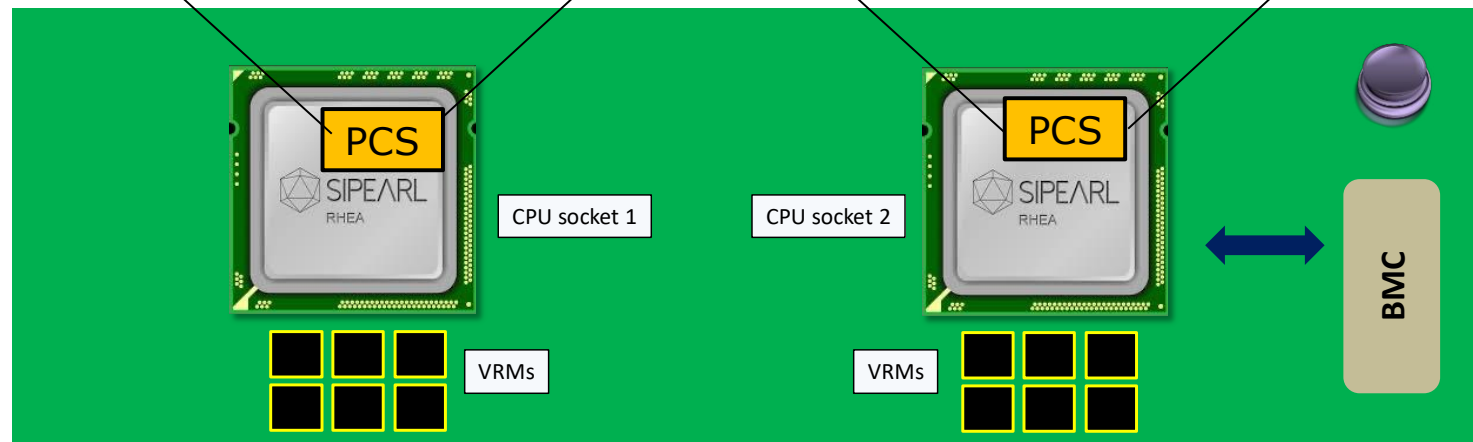
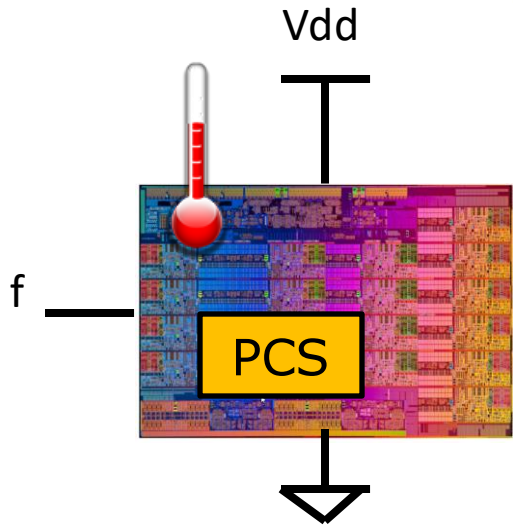


PFCT: Periodic Frequency Control Task
PVCT: Periodic Voltage Control Task
HLC: high level controller
LLC: low level controller

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

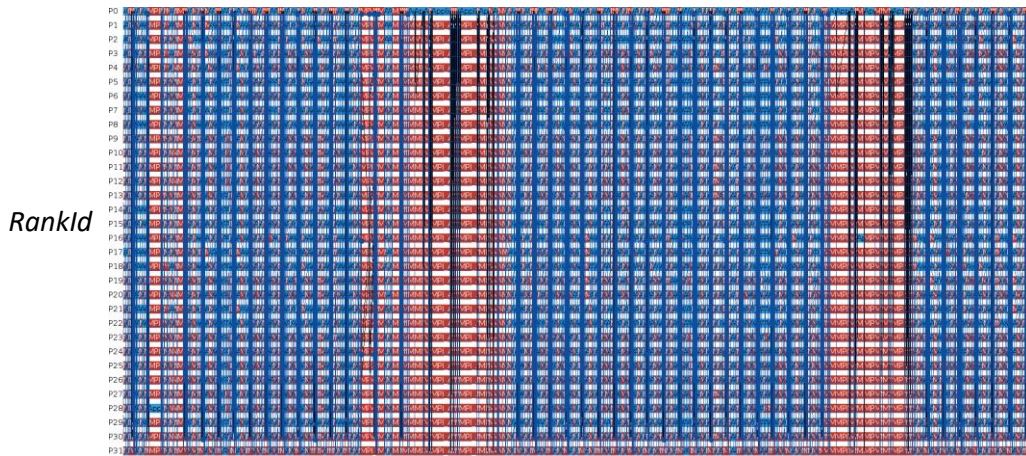


<https://github.com/pulp-platform/control-pulp>



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!

2015 IEEE International Parallel and Distributed Processing Symposium Workshop
An Energy Efficiency Feature Survey of the Intel Haswell Processor
Daniel Hackenberg, Robert Schöne, Thomas Ische, Daniel Molka, Joseph Schuchart, Robin Geyer
Center for Information Services and High Performance Computing (ZIH)
Technische Universität Dresden, 01062 Dresden, Germany
Email: {daniel.hackenberg, robert.schoene, thomas.ische, daniel.molka, joseph.schuchart, robin.geyer}@tu-dresden.de

VI. P-STATE AND C-STATE TRANSITION LATENCIES

A. P-State Transition Latencies

The introduction of integrated voltage regulators, per core frequency domains, and improvements in the power control unit (PCU) have a direct influence on the latency and duration of ACPI processor state [25] transitions. To examine the new architecture, we use FTLAT [26] for p-states and the tool developed by Schöne et al. [27] for c-states. We modified FTLAT in the following ways:

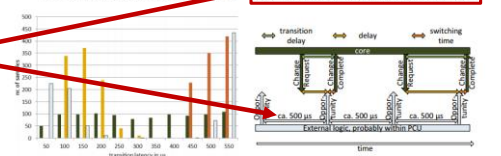
- The original FTLAT reads `scaling_cur_freq` from the Linux `cpufreq` subsystem to verify frequency settings. However, these readings are not reliable indicator for an actual frequency switch in hardware. We therefore add a verification by reading the `PERF_COUNT_HW_CPU_CYCLES` performance

therefore take 1,000 measurements for a single pair of start and target frequencies. We chose 1.2 and 1.3 GHz, but other frequency pairs yield similar results.

Figure 3 depicts the results of four experiments with 1,000 results each as a histogram. The resulting latency is evenly distributed between a minimum of 21 μ s and a maximum of 524 μ s. Requesting a frequency transition instantly after a frequency change has been detected leads to around 500 μ s in the majority of the results. If we introduce a 400 μ s delay after the last frequency change, the transition time is typically about 100 μ s. If the delay is in the order of 800 μ s, the transition latencies can be split into two different classes—some yield an immediate frequency change while others require over 500 μ s.

These results indicate that frequency changes only occur in regular intervals of about 500 μ s. The distance between the start

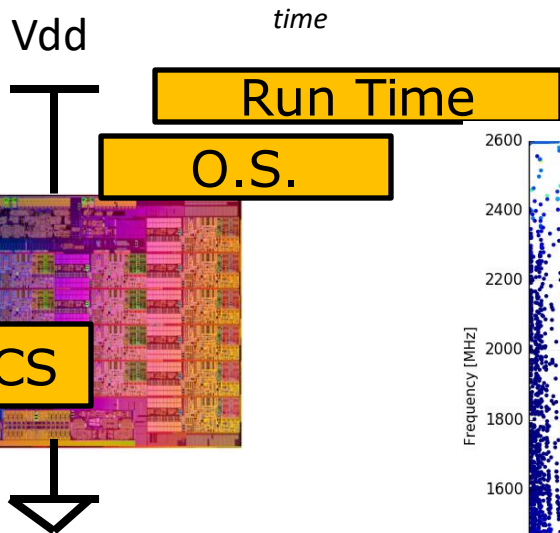
500us



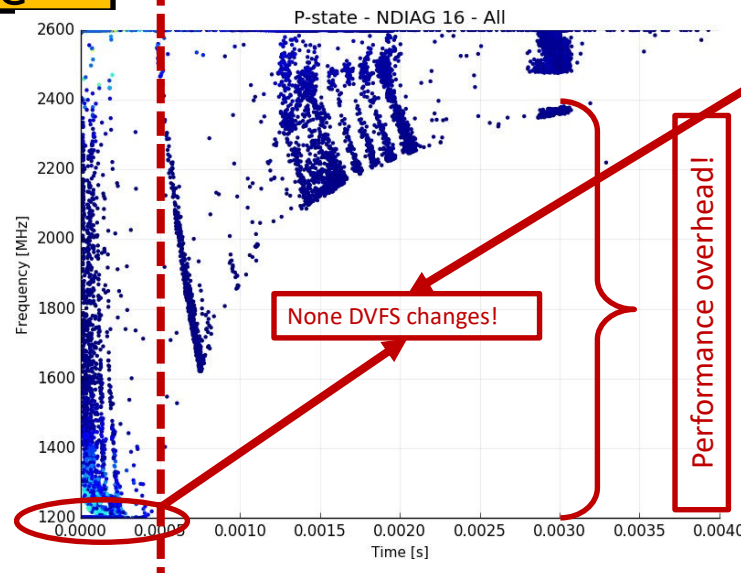
* Intel Broadwell architectures as well!

Fig. 3. Histogram of frequency transition latencies for switching between 1.2 and 1.3 GHz, depending on the time since the last frequency change.

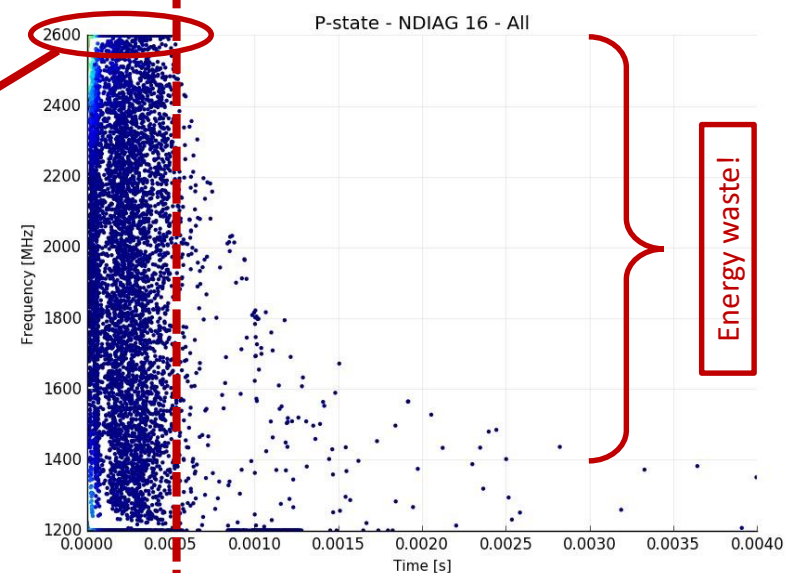
and the target frequency has negligible influence compared to the 500 μ s delay. The assumed frequency changing mechanism is depicted in Figure 4.



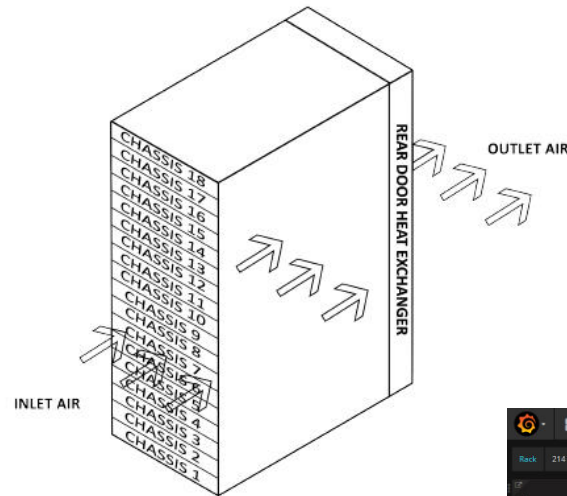
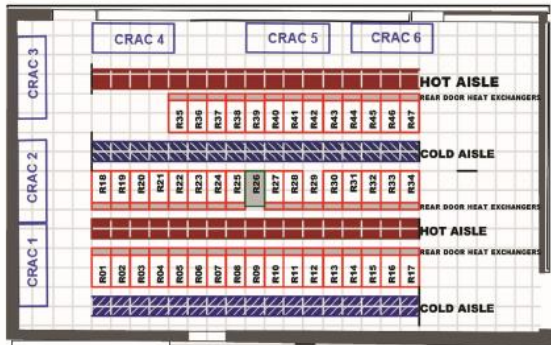
Application Phases



MPI Phases

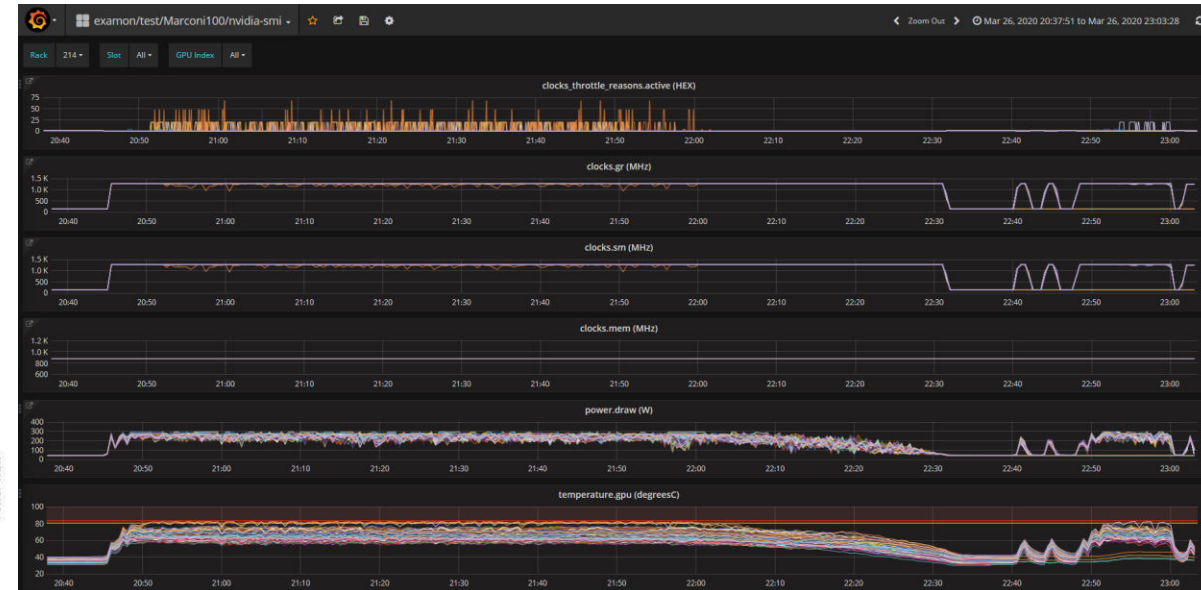
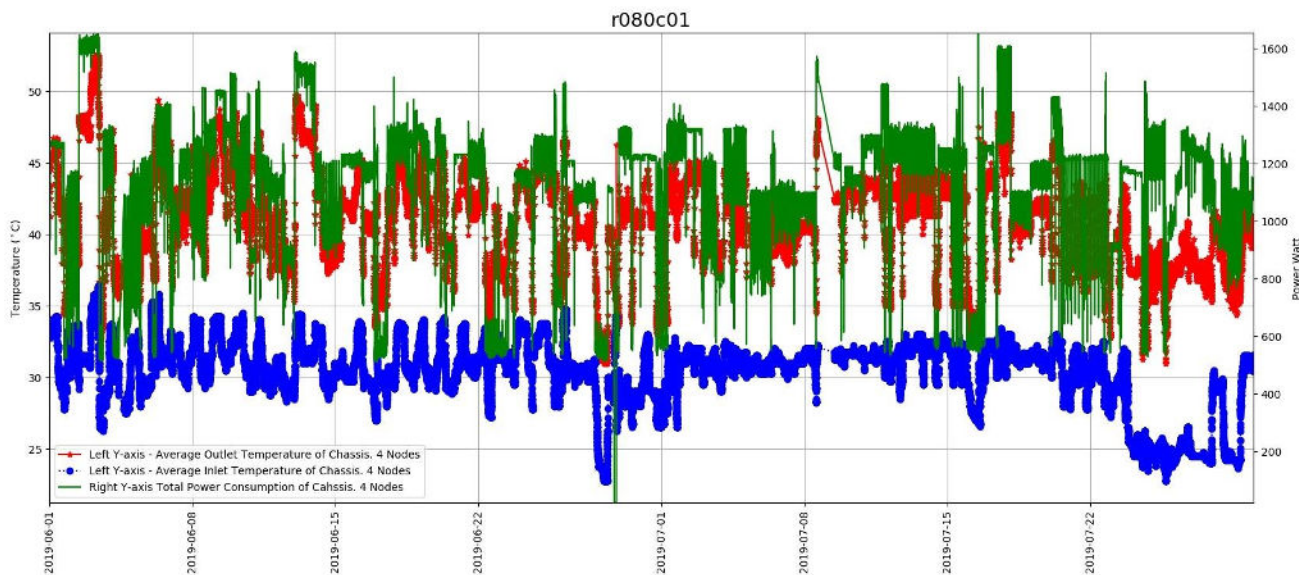


From sand to megawatts – Act 6: Room/DC Cooling Cost



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

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

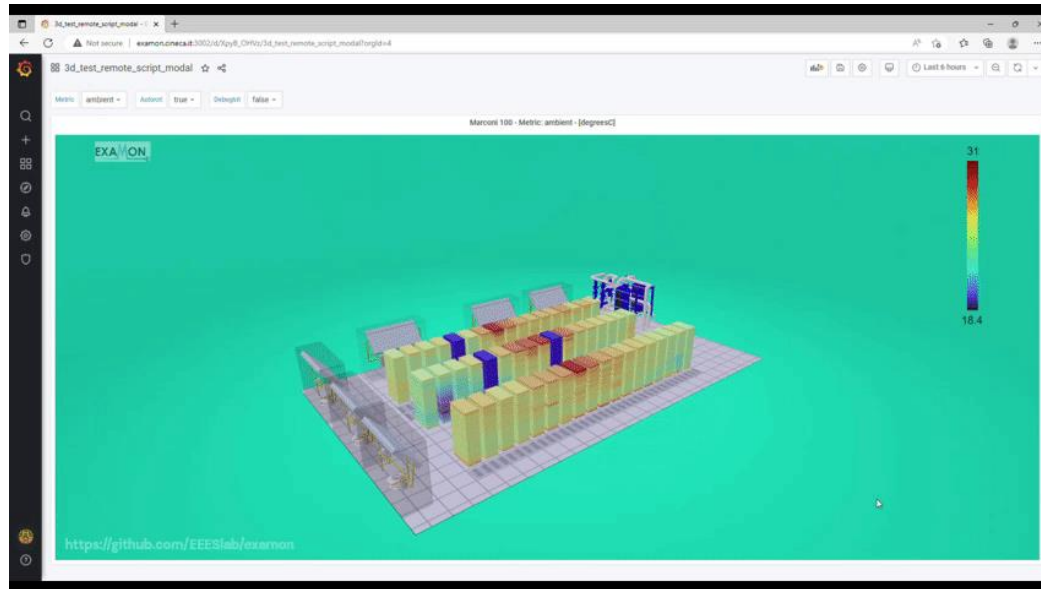


From sand to megawatts – Act 7: Data-driven Large-scale Optimization

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



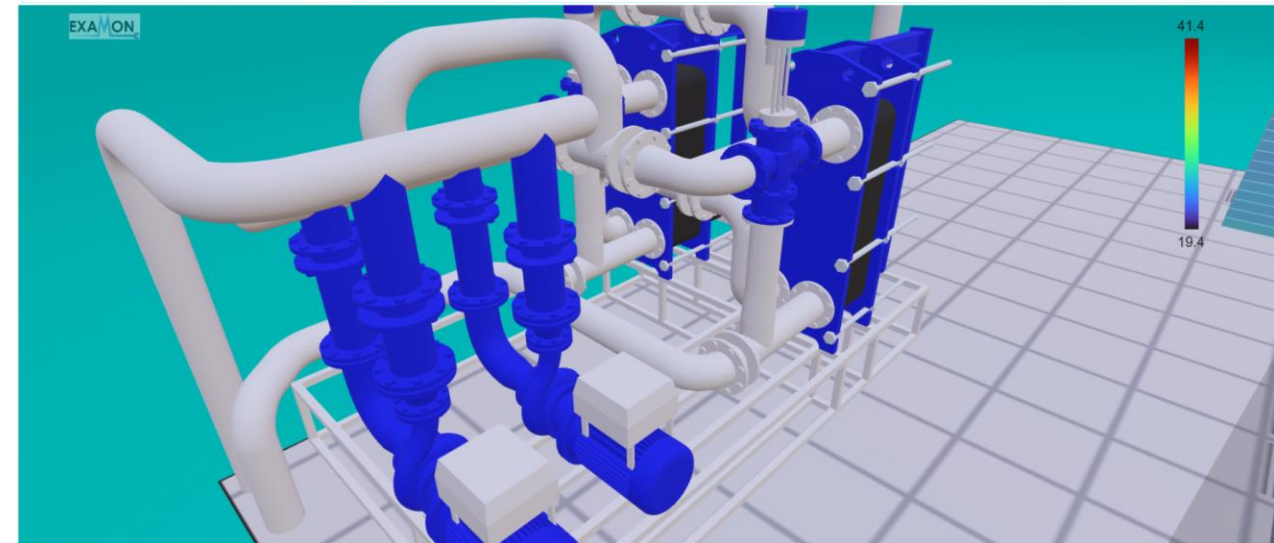
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



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

- Visualization and Analysis

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

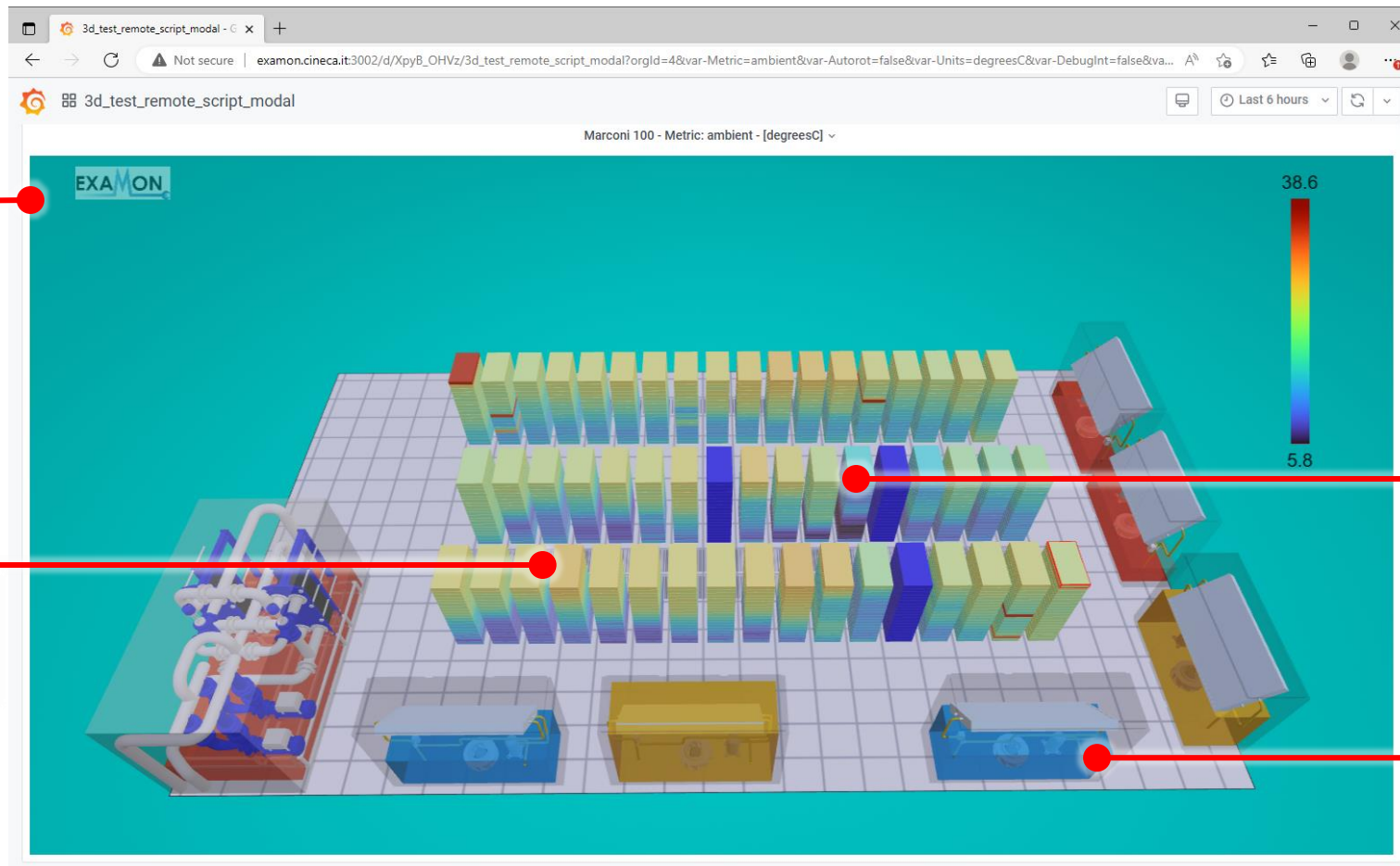


From sand to megawatts – Act 7: Data-driven Large-scale Optimization

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.

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



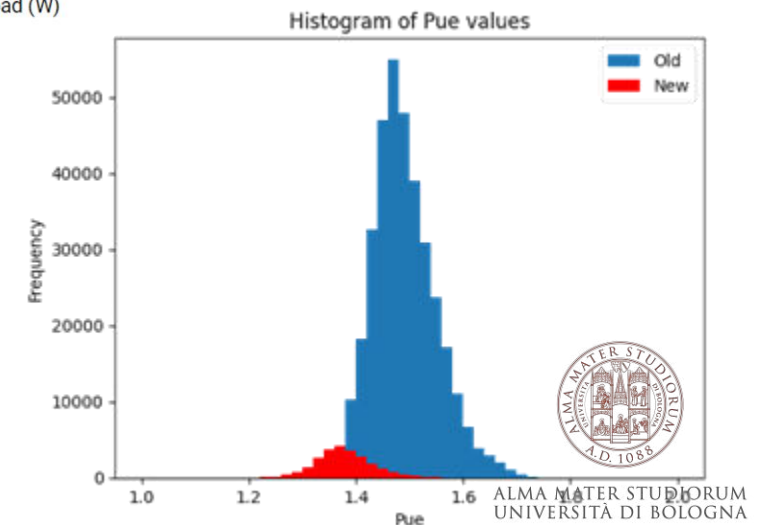
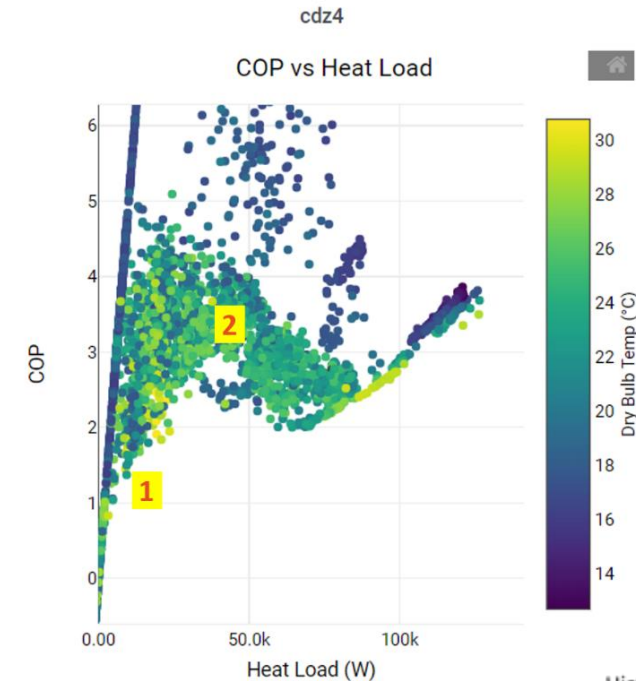
Hot spots, thermal imbalances, and cooling efficiency can be quickly investigated to improve the overall system.

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.



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

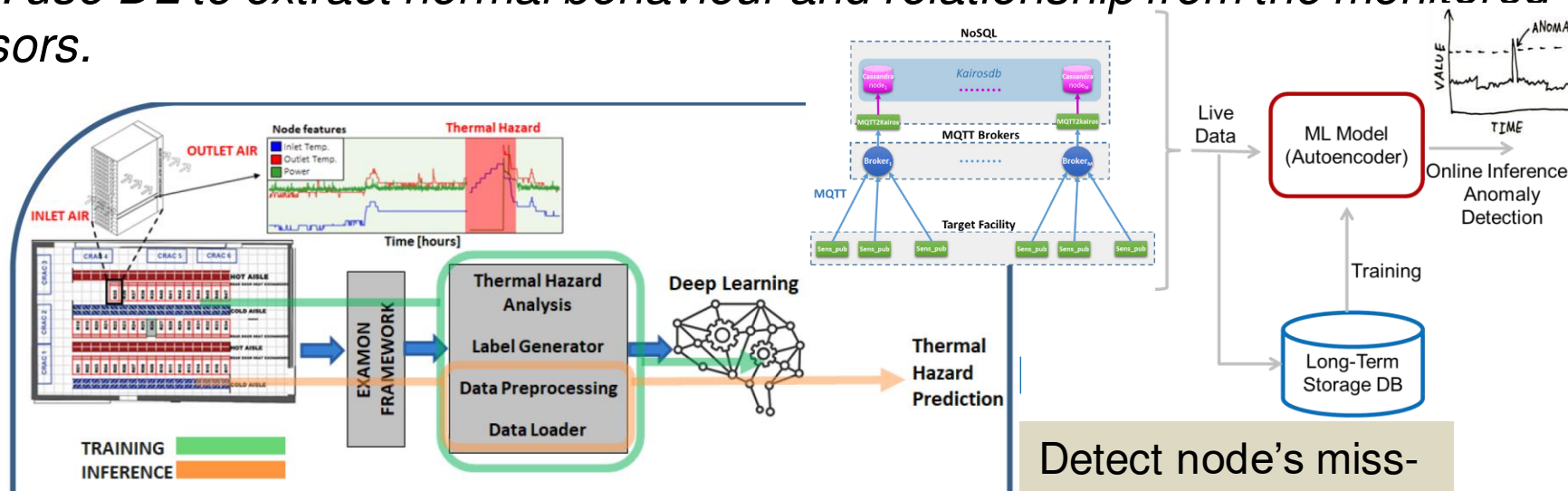
ExaMon@2021:

- 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



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>

Upload Communities

Celebrating our 10th anniversary! Send us your birthday greeting here. 🎉

January 31, 2023

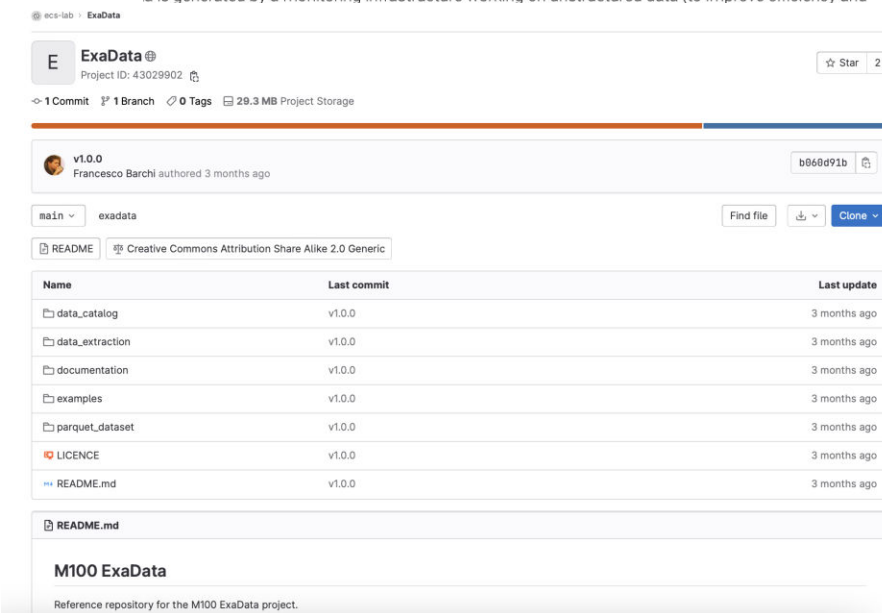
[Dataset](#) [Open Access](#)

M100 dataset 6: 22-03

 [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](#)

This entry is a part of a larger data set collected from the most recent Tier-0 supercomputer hosted at CINECA (<https://www.hpc.cineca.it/hardware/marconi100>). The data covers the entirety of the system, ranging from nodes (980+ computing nodes) internal information such as core loads, temperatures, frequencies, memory usage, CPU power consumption, fan speed, GPU usage details, etc., to the system-wide information, including cooling infrastructure, the air conditioning system, the power supply units, workload manager statistics, system status alerts, and weather forecast. Hundreds of metrics measured on each computing node, in addition to hundreds of other metrics gathered and monitored along all system components. This dataset is stored as a collection of Zenodo entries; this particular entry corresponds to the period: 22-03. It is provided as a partitioned Parquet dataset, with this partitioning hierarchy: year_month ("YY-MM"), plugin, and is distributed as tarball files, each corresponding to one month of data (first-level partitioning,

and is generated by a monitoring infrastructure working on unstructured data (to improve efficiency and



ecs-lab · ExaData

ExaData
Project ID: 43029802

1 Commit · 1 Branch · 0 Tags · 29.3 MB Project Storage

v1.0.0
Francesco Barchi authored 3 months ago

main · exadata

README Creative Commons Attribution Share Alike 2.0 Generic

Name	Last commit	Last update
data_catalog	v1.0.0	3 months ago
data_extraction	v1.0.0	3 months ago
documentation	v1.0.0	3 months ago
examples	v1.0.0	3 months ago
parquet_dataset	v1.0.0	3 months ago
LICENCE	v1.0.0	3 months ago
README.md	v1.0.0	3 months ago

README.md

M100 ExaData

Reference repository for the M100 ExaData project.

<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 !!

ExaData description

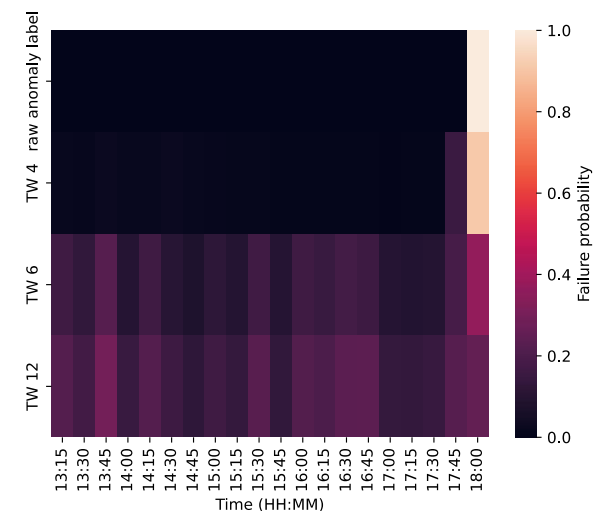
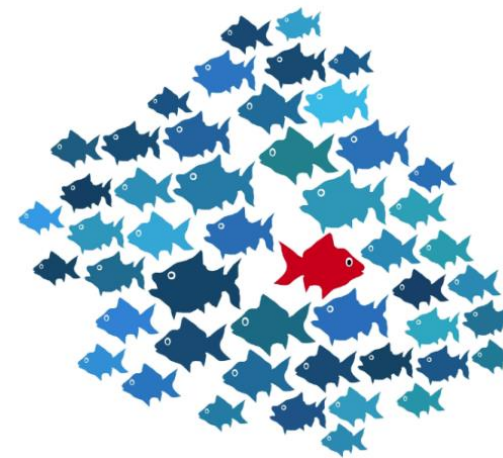
- 31 months of data
- 573 metrics, 980+ nodes, approx. 50 TB uncompressed
- Vertiv, Schneider, IPMI, Ganglia, Logics, Weather, Nagios, SLURM, Job table
- Hardware data, system monitoring data, external information
- Different sampling granularities (from seconds) to minutes
- Zenodo + Nature Dataset

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.

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

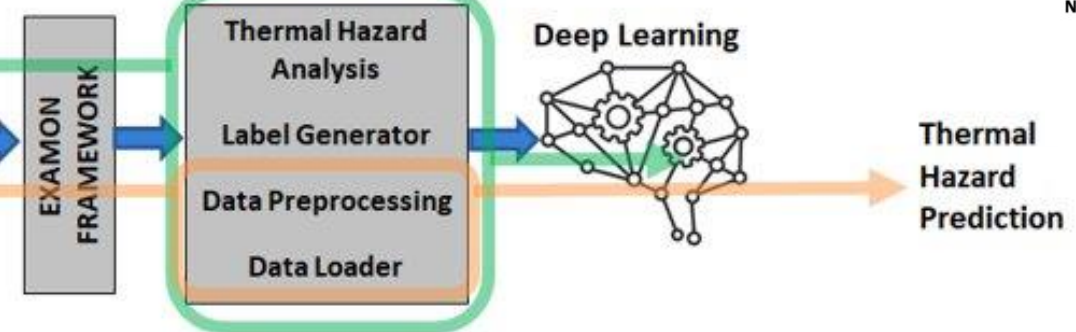
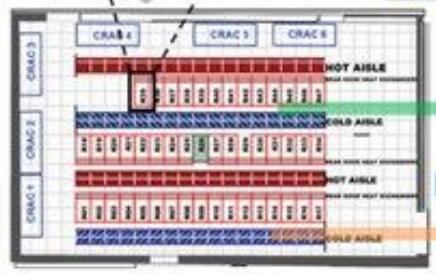
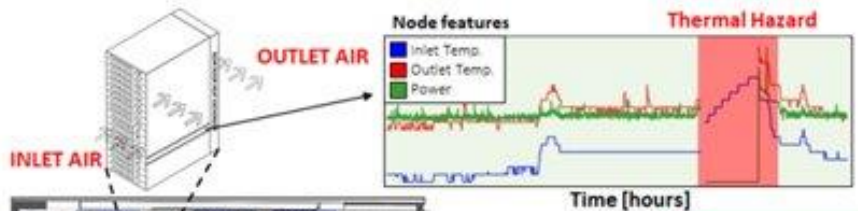
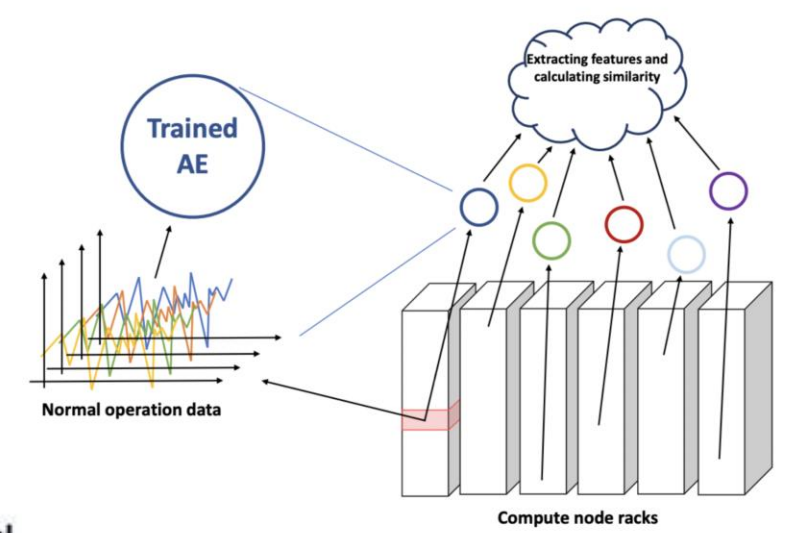
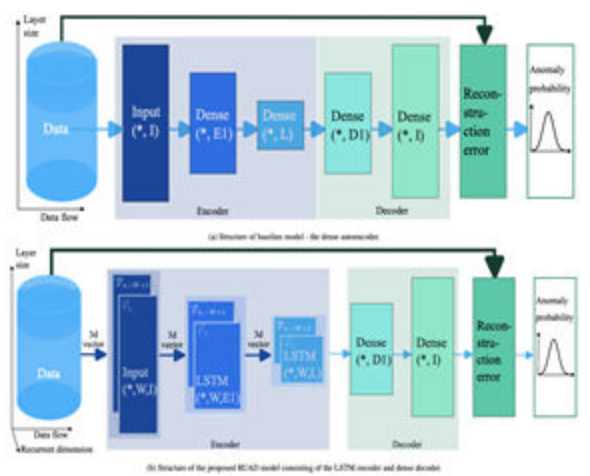
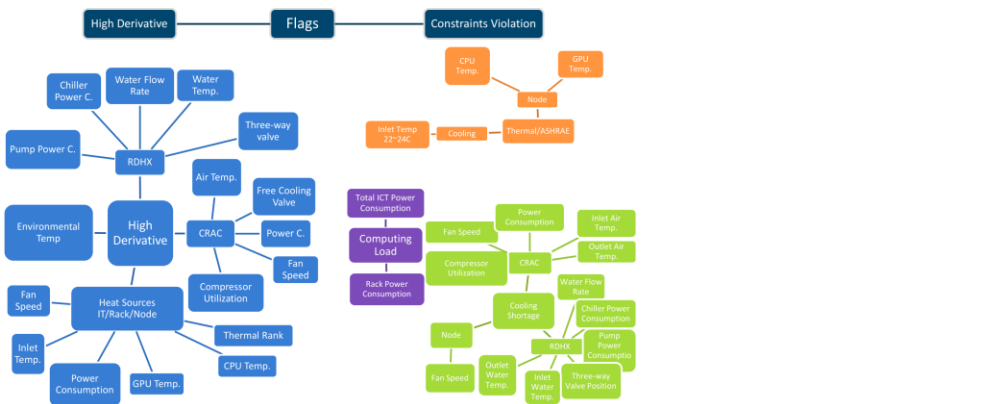
RUAD: Unsupervised anomaly detection in HPC systems, FGCS23

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 TI06

Graph Neural Networks for Anomaly Anticipation in HPC Systems ICPE23



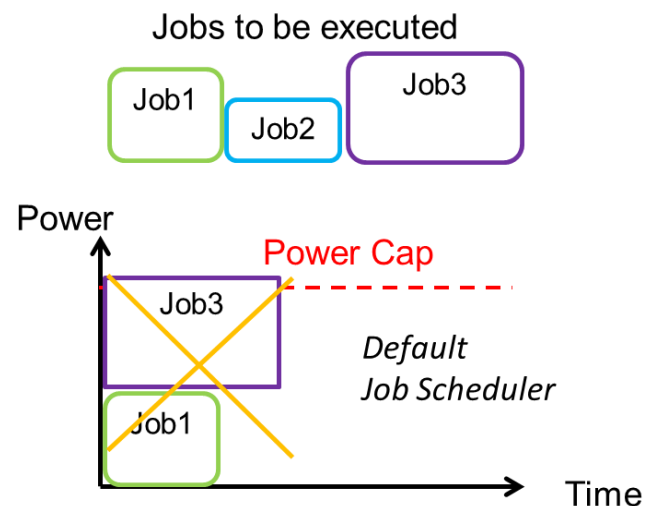
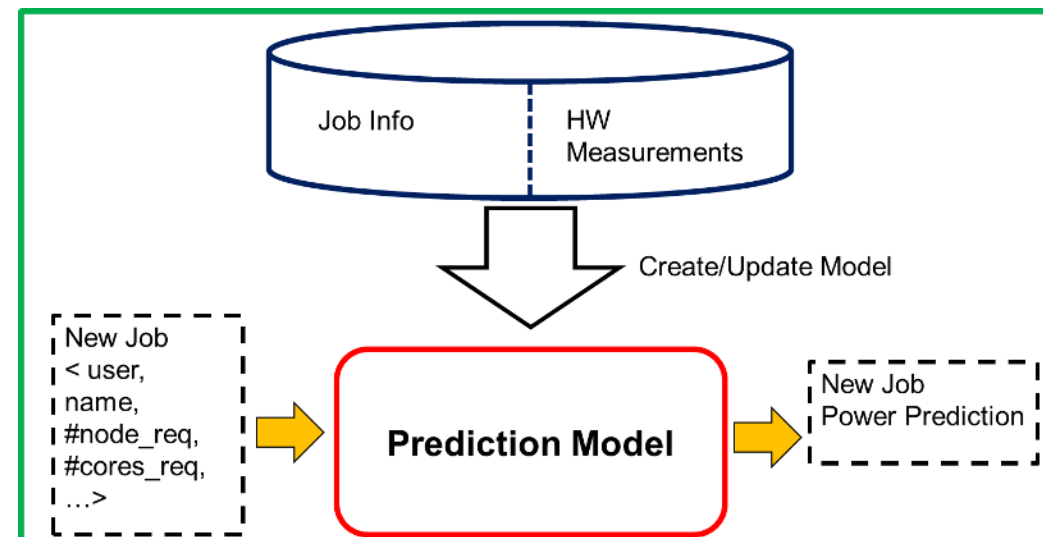
TRAINING █
INFERENCE █



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



Acknowledgment



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)**.

