Scaling performance In Power-Limited HPC Systems

Prof. Dr. Luca Benini

ERC Multitherman Lab
University of Bologna – Italy

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-ITET, Chair of Digital Dircuits
& Systems - Switzerland

erc European Research Council

European Commission | Horizon 2020 European Union funding for Research & Innovation

nano-tera.ch

FNSNF Fonds national suisse Schweizerischer Nationalfonds Fondo nazionale svizzero Swiss National Science Foundation

Outline

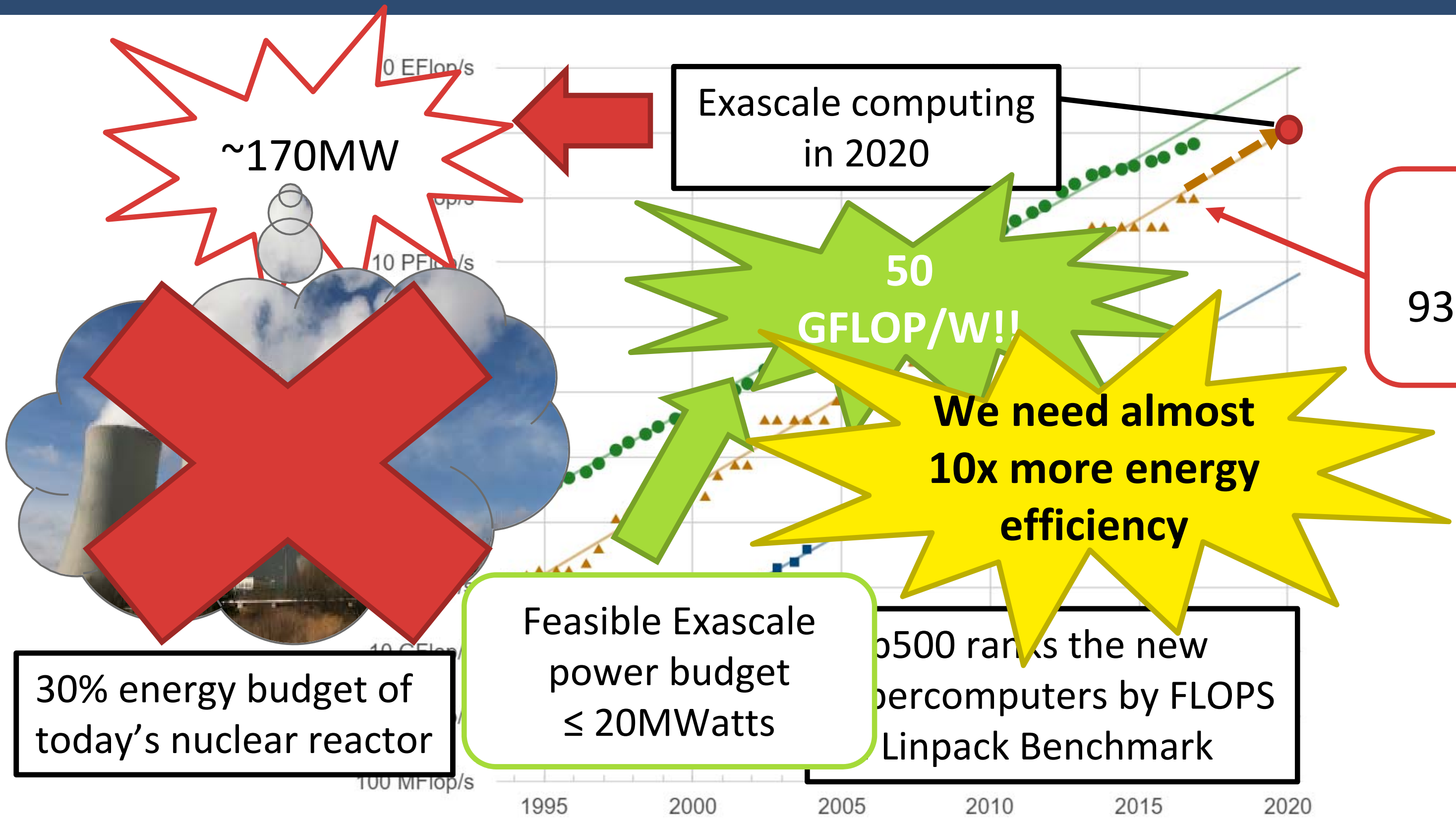
❑ Power and Thermal Walls in HPC

❑ Power and Thermal Management

❑ Energy-efficient Hardware

❑ Conclusion

Power Wall → Avg



~170MW

Exascale computing in 2020

Sunway TaihuLight
93 PF, 15.3 MW
6 GF/W

50 GFLOP/W!!

We need almost 10x more energy efficiency

30% energy budget of today's nuclear reactor

Feasible Exascale power budget ≤ 20MWatts

p500 ran s the new percomputers by FLOPS Linpack Benchmark

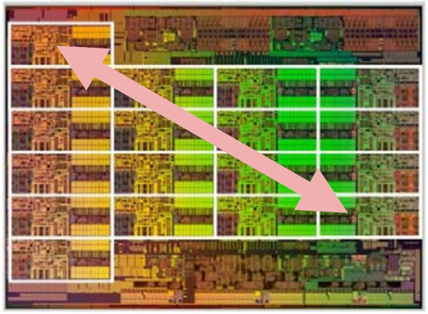
The second, Tianhe-2 (ex 1st) consumes 17.8 MW for "only" 33.2 PetaFLOPs, but…

24 MW

Cooling system matters!!!

Dynamic Power management (DPM)

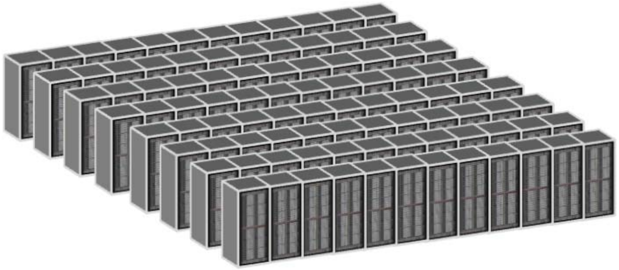
Thermal Wall → Max+



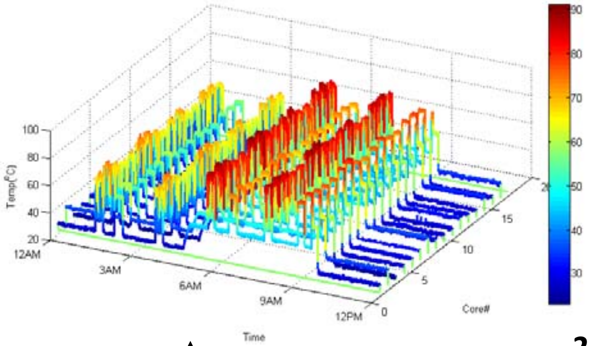
Intel Haswell – E5-2699 v3 (18 core)

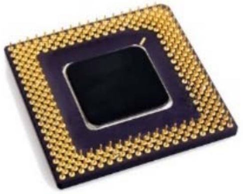
Up to **24°C** Temperature difference on DIE
More than **7°C** thermal heterogeneity under same workload

HPC System



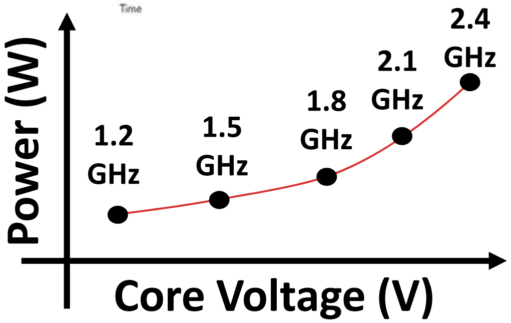
Thermal range: 69 C° – 101 C°





Power consumption: 40% - 66%

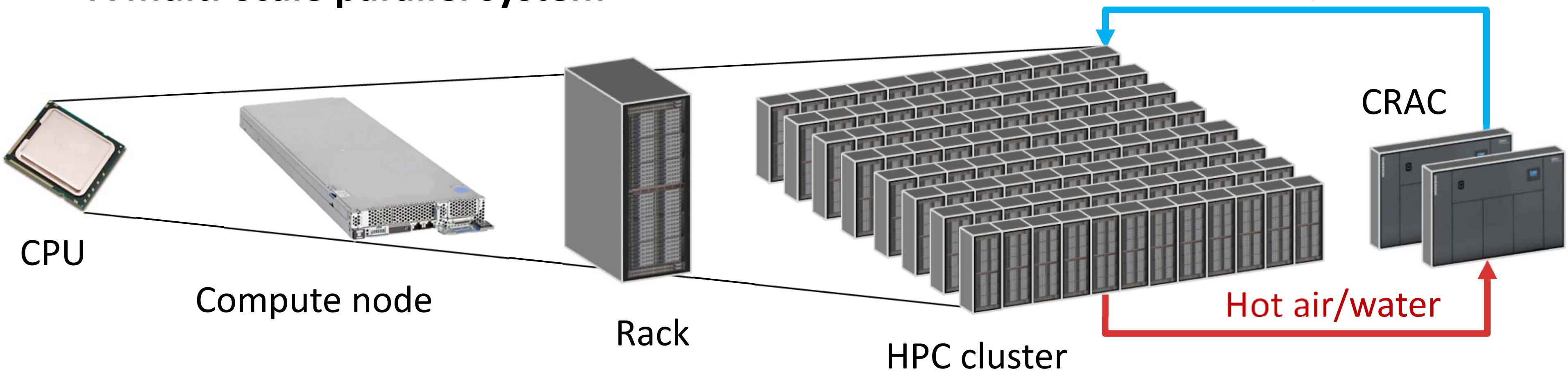
Per-core DVFS approach



Dynamic thermal management (DTM)

HPC Architecture - Hardware

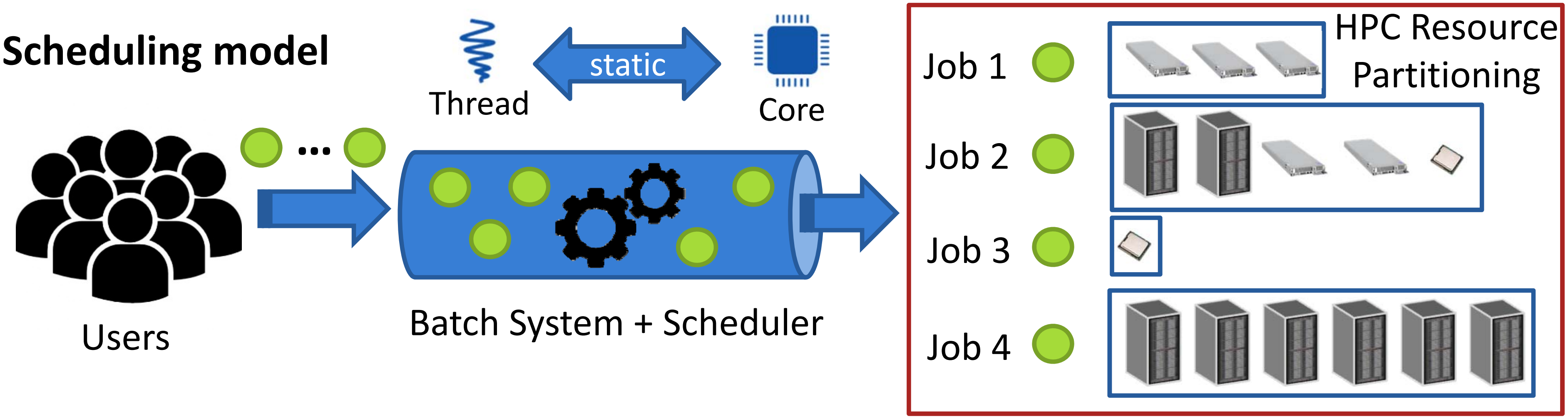
A multi-scale parallel system



DPM, DTM are Multi-scale Problems! ➡ Multitherman erc European Research Council

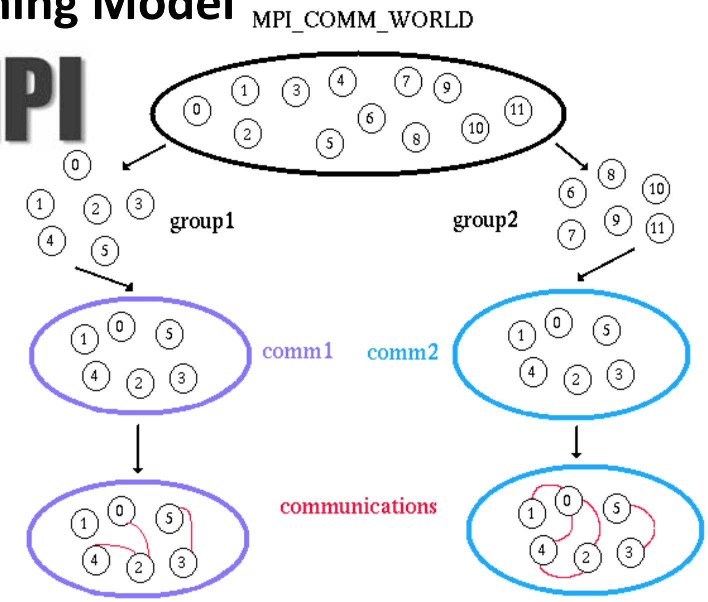
HPC Architecture - Software

Scheduling model



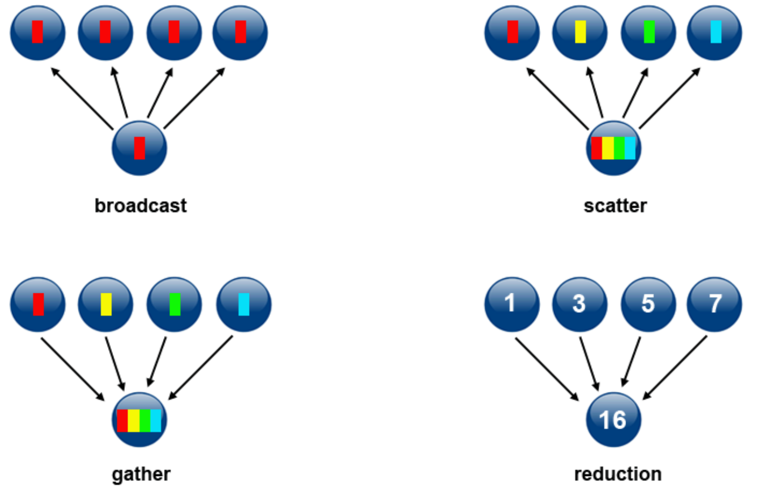
Programming Model

MPI



COMMUNICATIONS

one-to-one, one-to-many, many-to-one and many-to-many



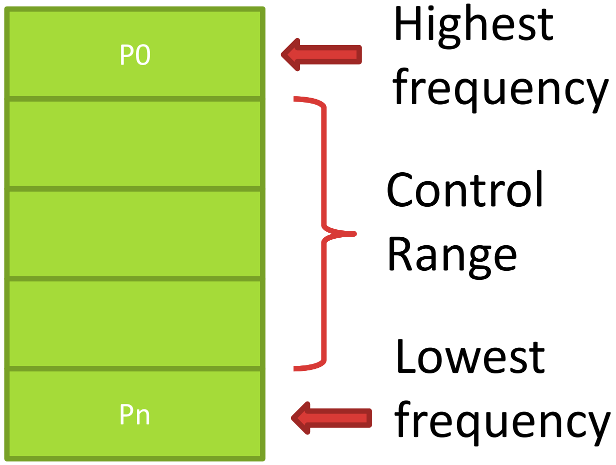
Programming & Scheduling model is essential!

Outline

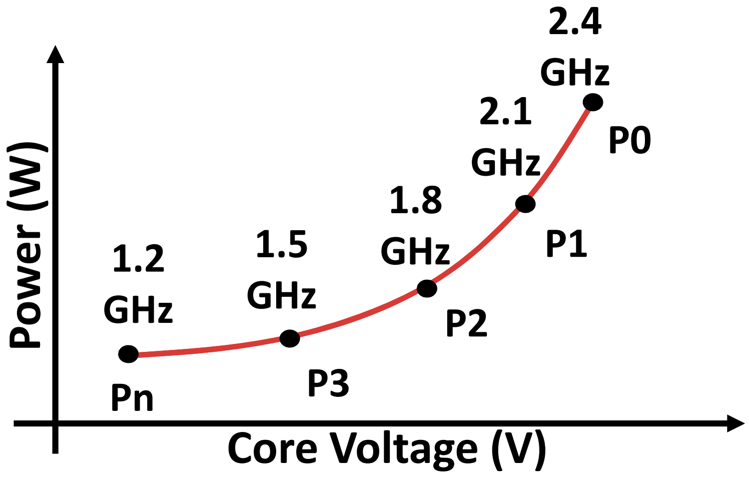
- □ Power and Thermal Walls in HPC
- □ Power and Thermal Management
- □ Energy-efficient Hardware
- □ Conclusion

HW Support for DPM, DTM

ACTIVE STATES
DVFS (P-State)



P-State: both a voltage and frequency level

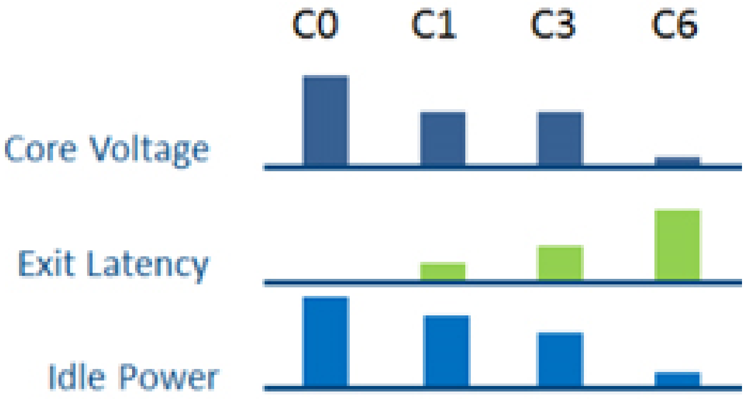


IDLE STATES
low power (C-State)

		Core C-States				
		C0	C1	C3	C6	
Package C-States	C0	🟩	🟩	🟩	🟩	Active State
	C1E	🟥	🟩	🟩	🟩	Lower P-State
	C2	🟥	🟥	🟩	🟩	Only L3 Snoop
	C3	🟥	🟥	🟩	🟩	Flush L3 - Off
	C6	🟥	🟥	🟥	🟩	Low Voltage
		Active State	Clock Gated	Flush L1,L2 Off	Power Gated	

🟩 Possible combination of core/package states

🟥 Impossible combination of core/package states



Intel provides a HW power controller called Running Average Power Limit (RAPL).

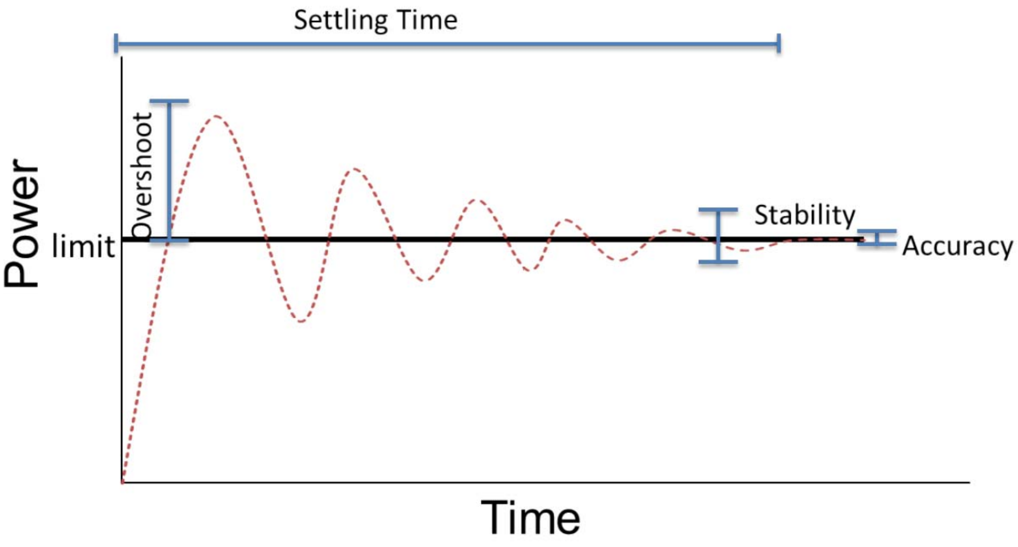
Power Management → Reactive

A significant exploration work on RAPL control:

- ❖ *Zhang, H., & Hoffman, H. (2015). "A Quantitative Evaluation of the RAPL Power Control System". Feedback Computing.*

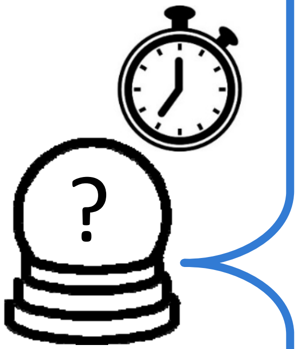
Quantify the behavior the control system in term of:

- ➢ **Stability**: freedom from oscillation
- ➢ **Accuracy**: convergence to the limit
- ➢ **Settling time**: duration until limit is reached
- ➢ **Maximum Overshoot**: the maximum difference between the power limit and the measured power



Power Management → HW Predictive

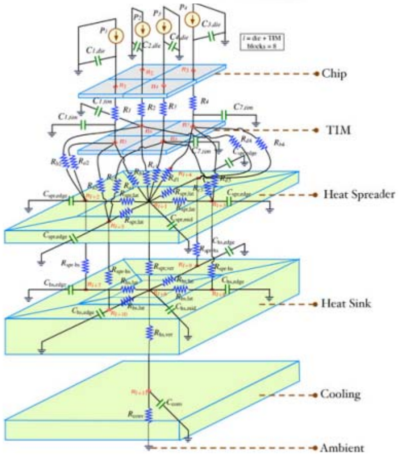
- **on-line optimization policies**
  - *A. Bartolini et al. "Thermal and Energy Managementof High-Performance Multicores: Distributed and Self-Calibrating Model-Predictive Controller.“ TPDS’13*



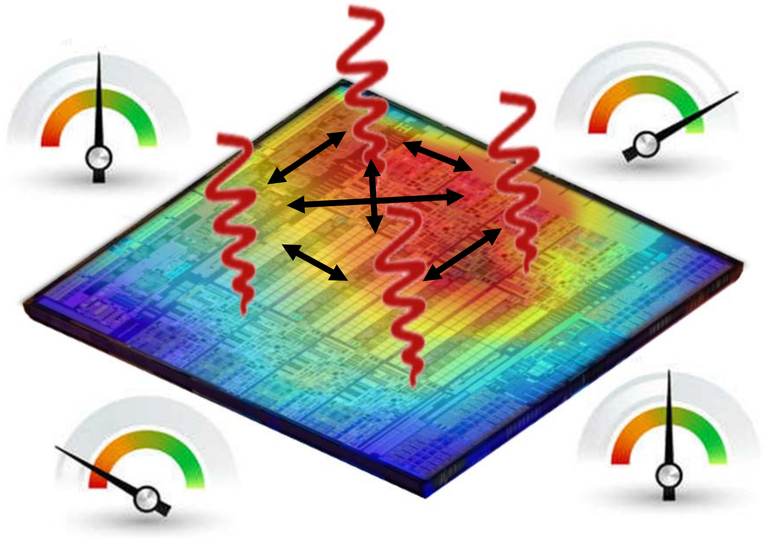
Temperature Prediction

AutoRegressive Moving Average (ARMA)

Thermal model based on RC approch



Implement proactive and reactive policies using DVFS selections and thread migrations



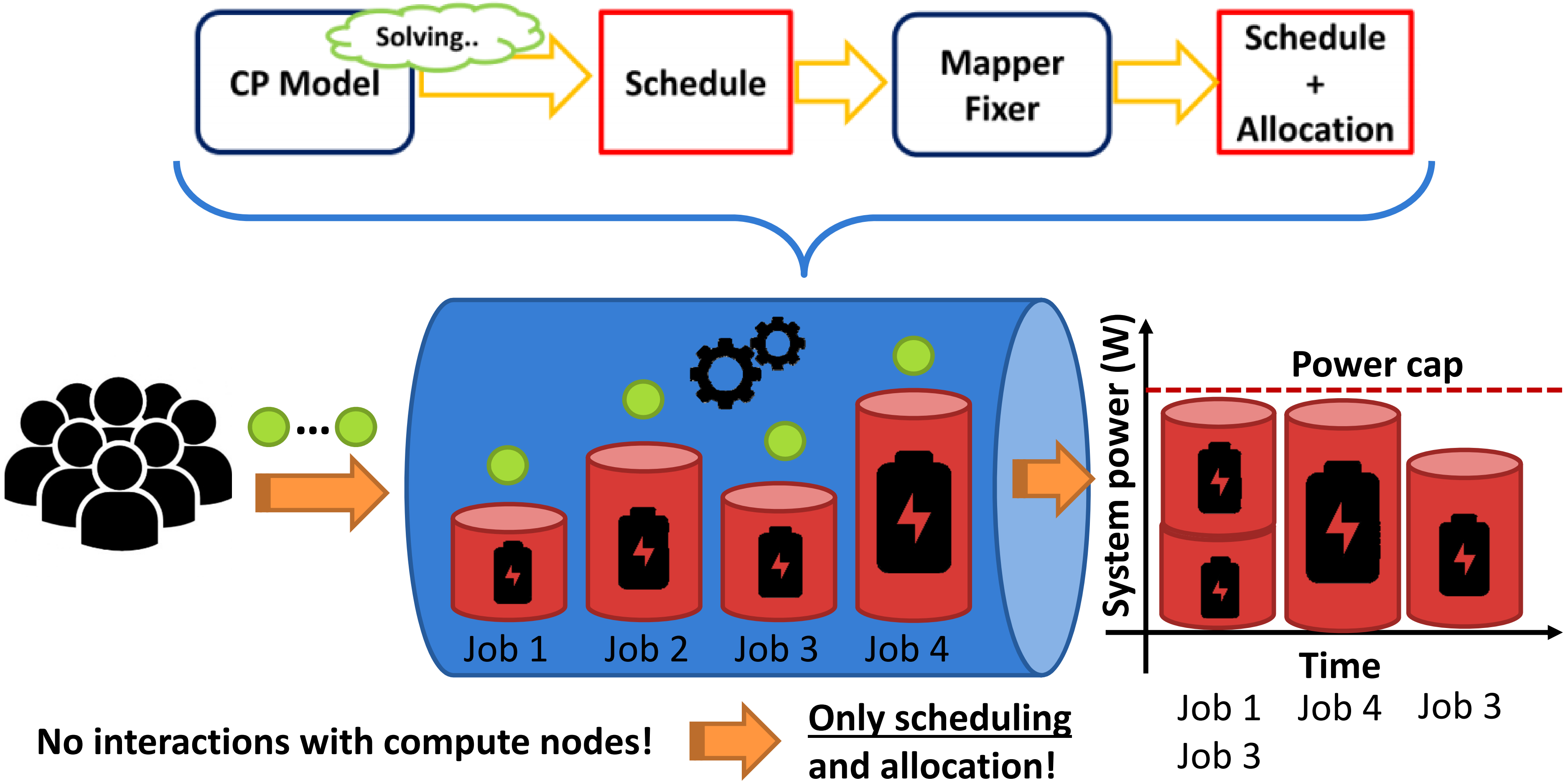
Scheduler based on convex optimization for DVFS selections and thread migrations

Online techniques are capable of sensing changes in the workload distribution and setting the processor controls accordingly.

Power Management → SW predictive

❖ **Predictive models to estimate the power consumption**

- *Borghesi, A., Conficoni, C., Lombardi, M., & Bartolini, A. “MS3: a Mediterranean-Stile Job Scheduler for Supercomputers-do less when it's too hot!”. HPCS 2015*
- *Sîrbu, A., & Babaoglu, O. “Predicting system-level power for a hybrid supercomputer”. HPCS 2016*



No interactions with compute nodes! → **<u>Only scheduling and allocation!</u>**

Challenges

SW policies

👍 Application aware

👎 High overhead
Coarse granularity
(seconds)

HW mechanisms

👍 Low overhead
Fine granularity
(milliseconds)

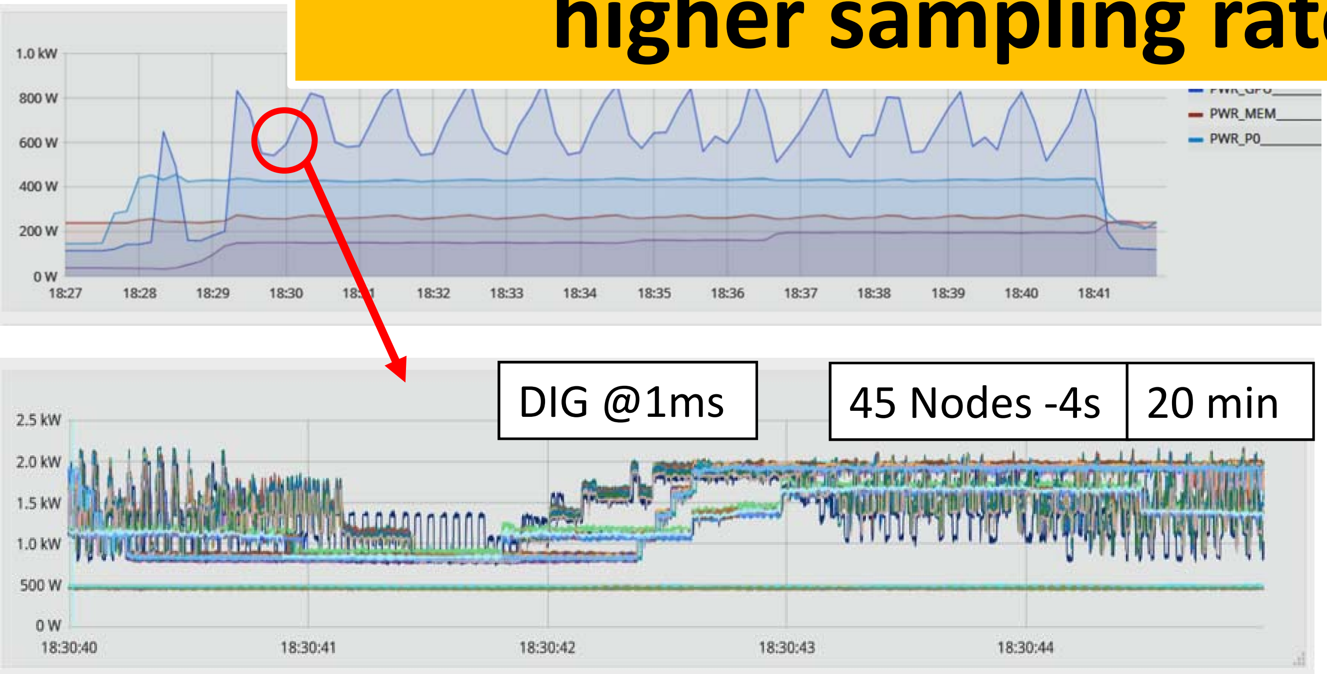
👎 No application awareness

1) Low-Overhead, accurate monitoring
2) Scalable data collections, analytics, decisions
3) Application awareness

Low Overhead, accurate Monitoring

High-resolution monitoring → more information available

How to analyze real-time with higher sampling rates?



Max. Ts = 1s

DIG @1ms

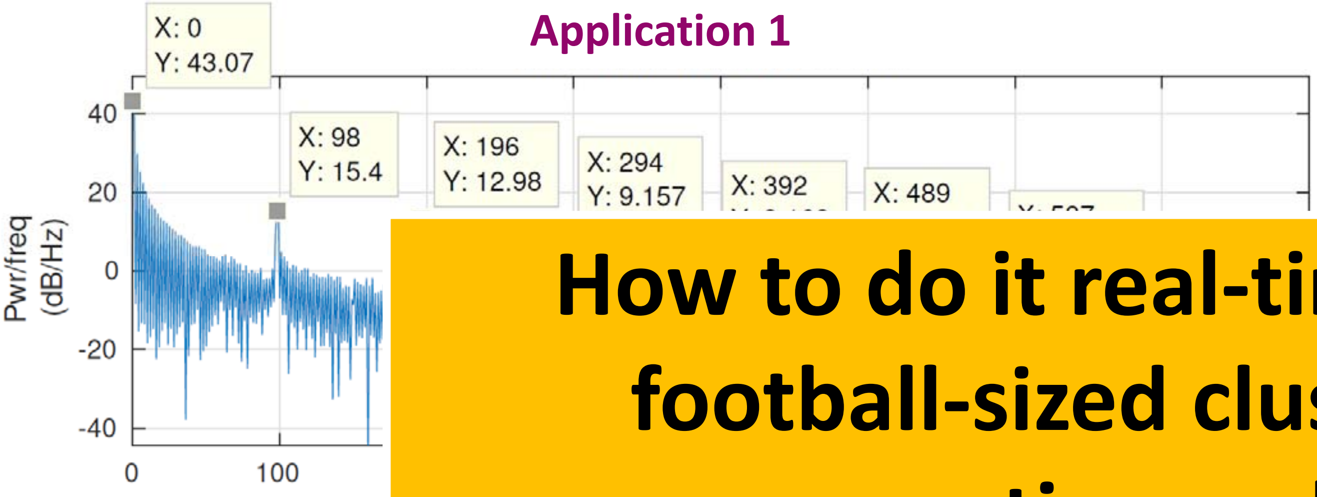
45 Nodes -4s

20 min

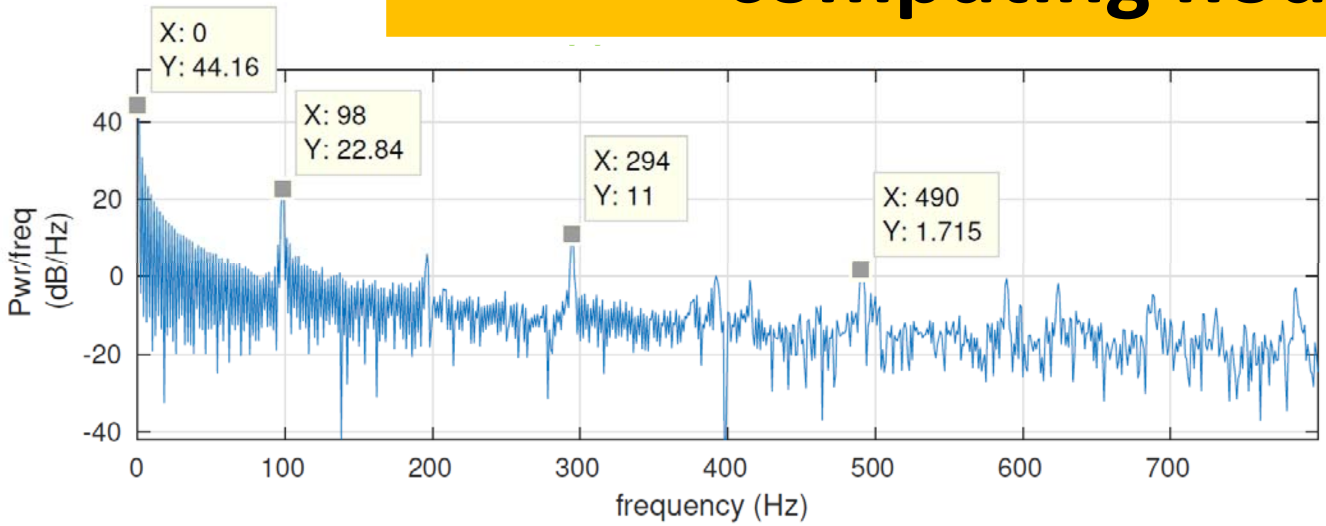
Low Overhead, accurate Monitoring

Real-time Frequency analysis on power supply and more…

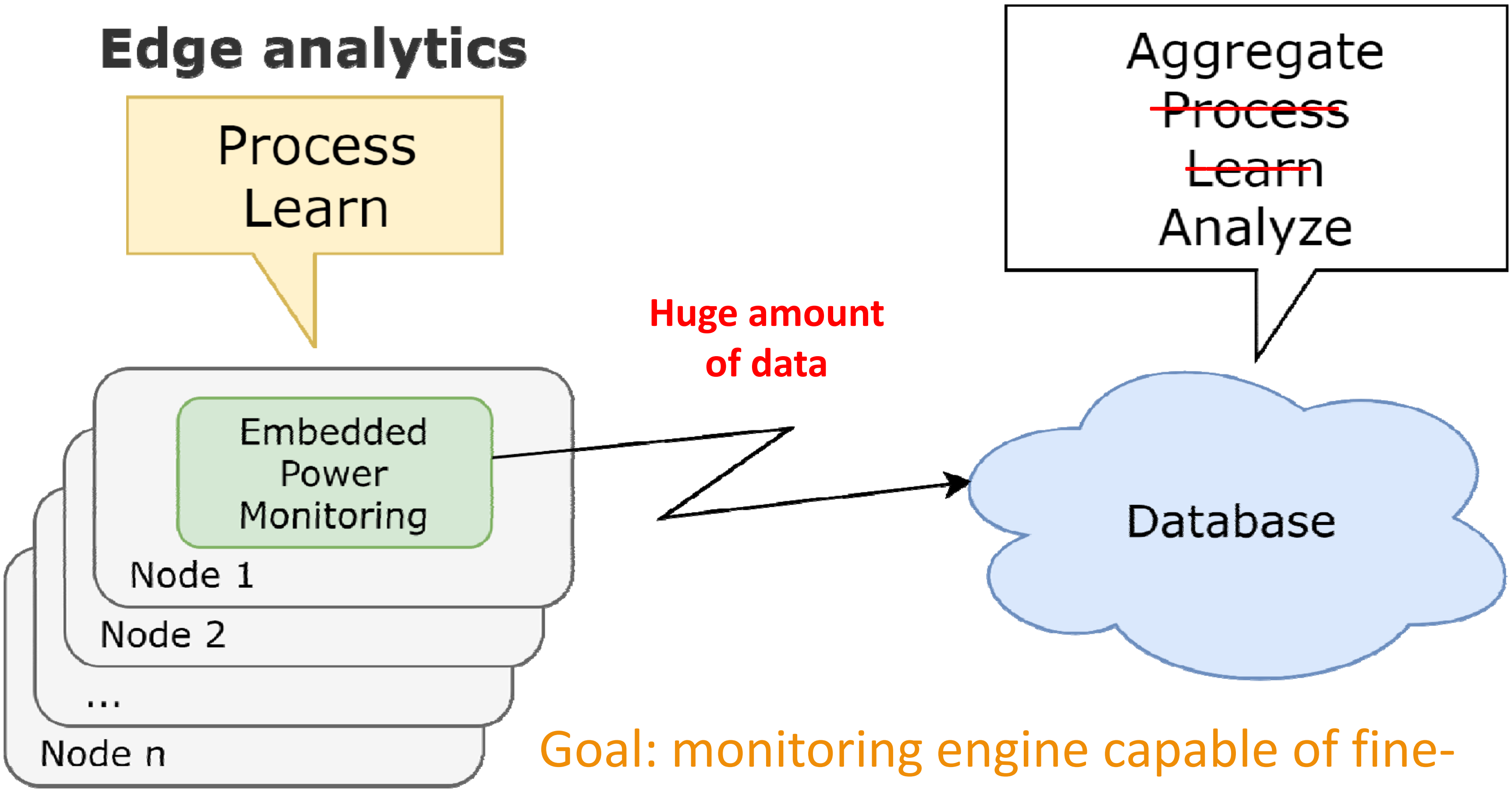
Application 1



How to do it real-time for a football-sized cluster of computing nodes?



Solution – Dwarf In a Giant (DIG)



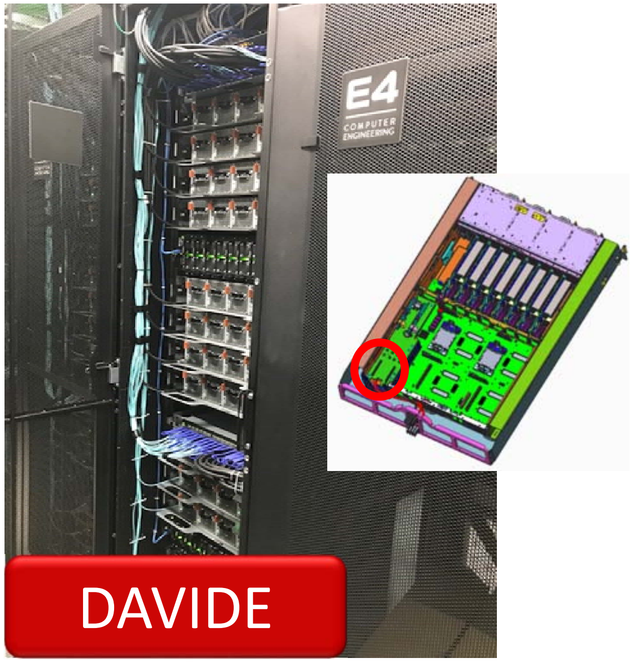
Goal: monitoring engine capable of fine-grained monitoring and spectral analysis distributed on a large-scale cluster

DIG in Real Life

Developing hardware extensions for fine-grained power monitoring: DIG deployed in production machines

- Intel Xeon E5 based
- Used for prototyping

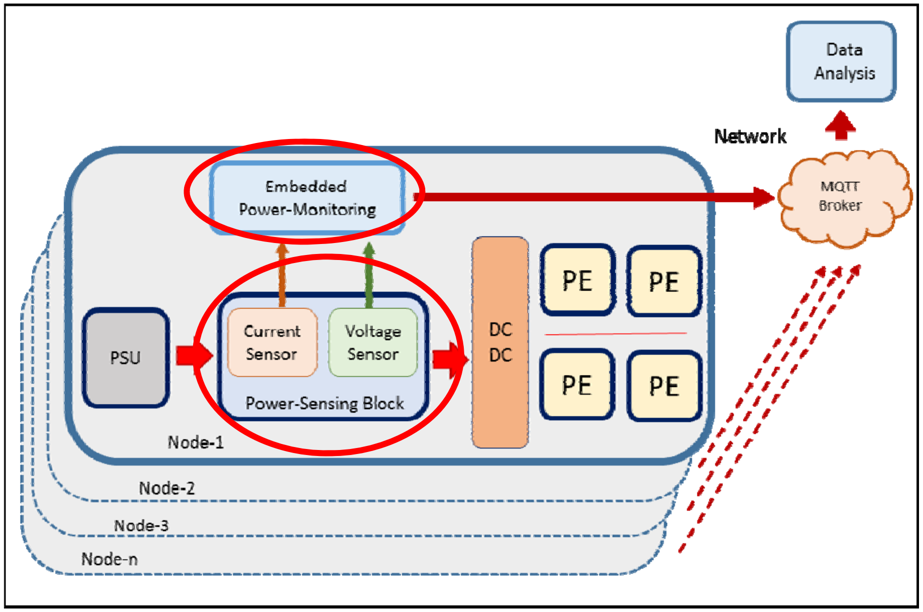
- ARM64 Cavium based
- Commercial system
- with E4 - PCP II



- IBM Power8 based
- Commercial system
- with E4 - PCP III
- 18th in Green500

DIG Architecture

High Resolution Out-of-band Power Monitoring



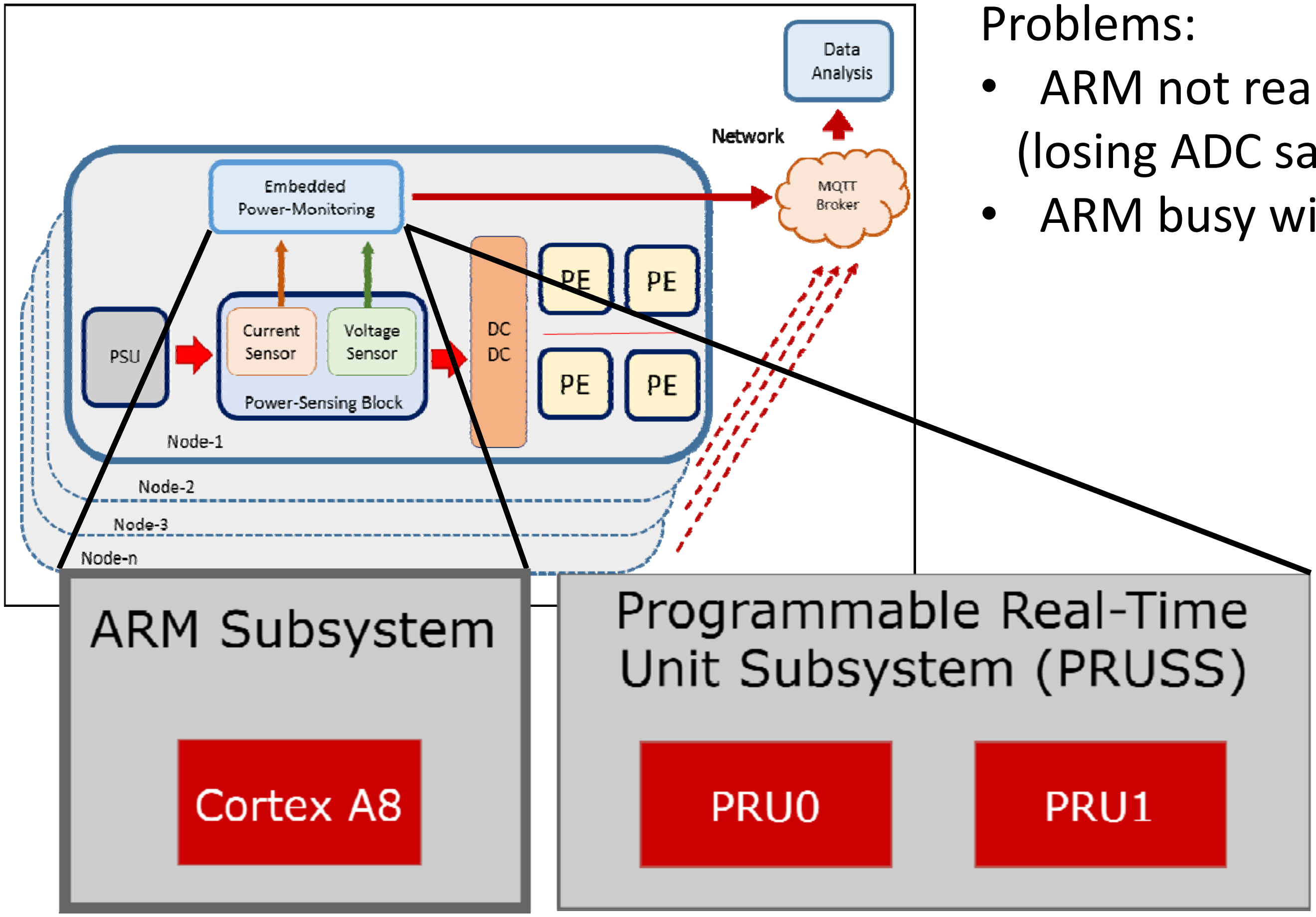
- Overall node power consumption
- Can support edge computing/learning
- Platform independent (Intel, IBM, ARM)
- Sub-Watt precision
- Sampling rate @50kS/s (T=20us)

State-of-the art systems (Bull-HDEEM and PowerInsight)

- Max. 1 ms sampling period
- Use data only offline

Hackenberg et al. "HDEEM: high definition energy efficiency monitoring”
Laros et al. "Powerinsight-a commodity power measurement capability."

Real-time Capabilities



Problems:

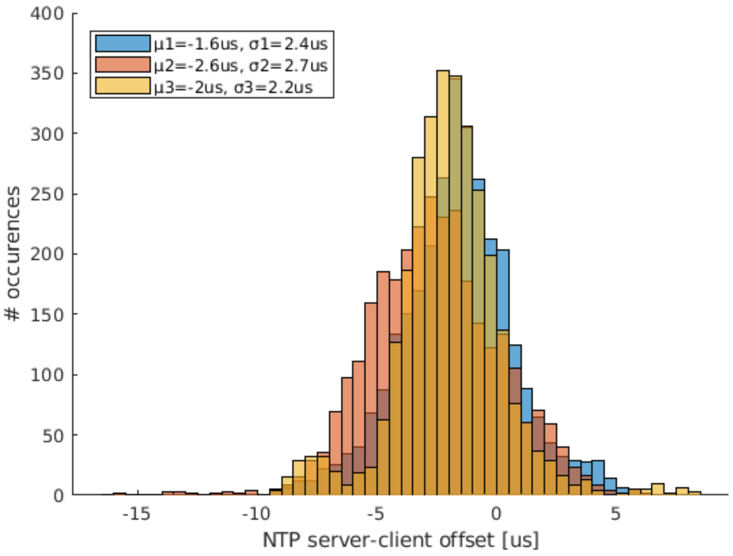
- ARM not real-time (losing ADC samples)
- ARM busy with flushing ADC

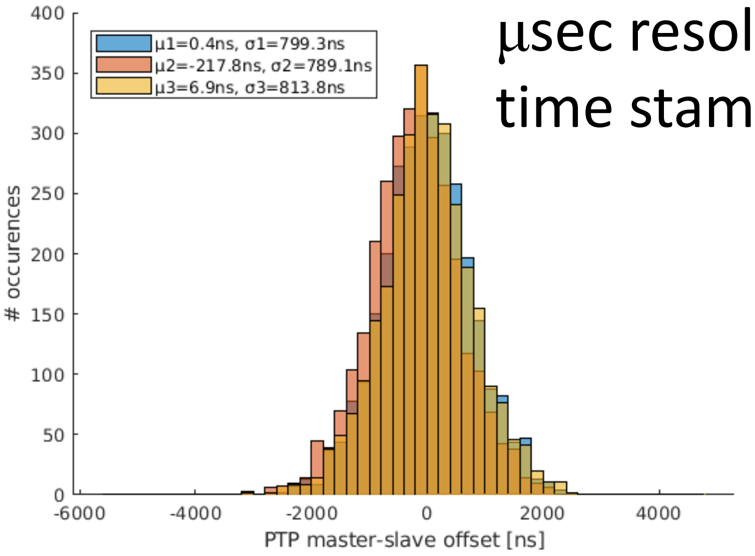
Goal:
Offload the processing to the PRUSS

DIG in production: E4’s D.A.V.I.D.E.

Possible tasks of the PRUs: Averaging @ 1ms, 1s → offline Computing, FFT → edge analysis

Framework	Fs_{max} [kHz]	CPU Overhead
DIG	50	~40%
DIG+PRU, edge analysis	400	<5%
DIG+PRU, offline	800	<5%
Bull-HDEEM	1	?
PowerInsight	1	?

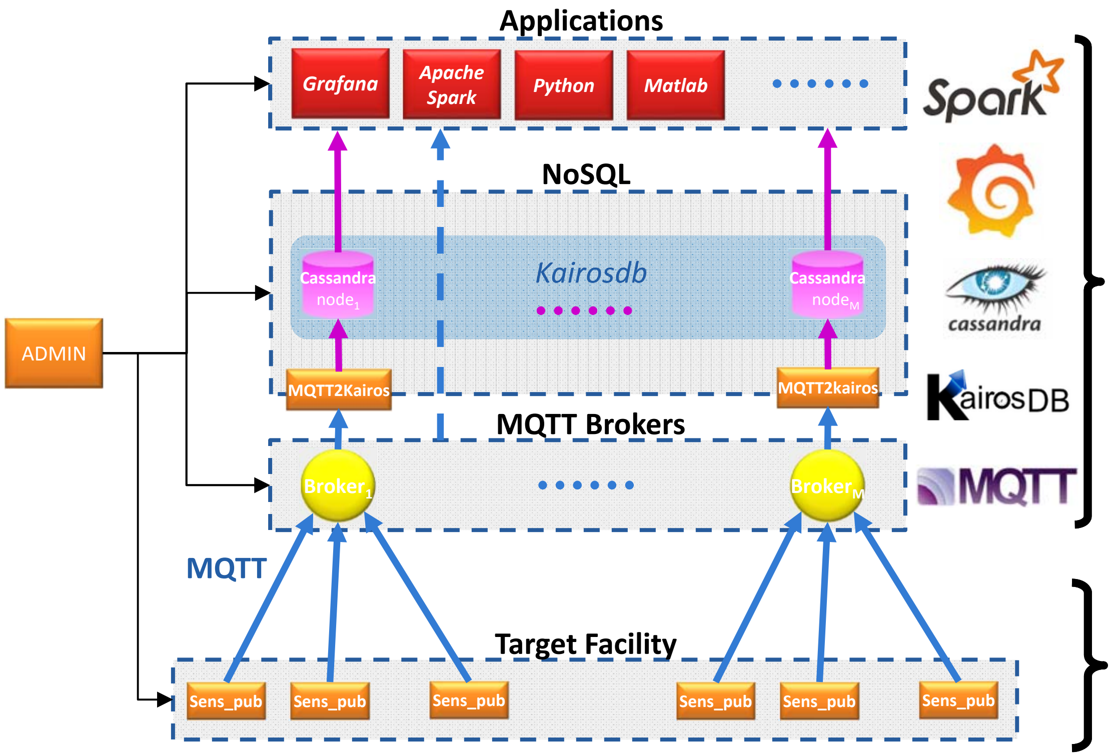




μsec resolved time stamps

Scalable Data Collection, Analytics

EXAMON



Front-end

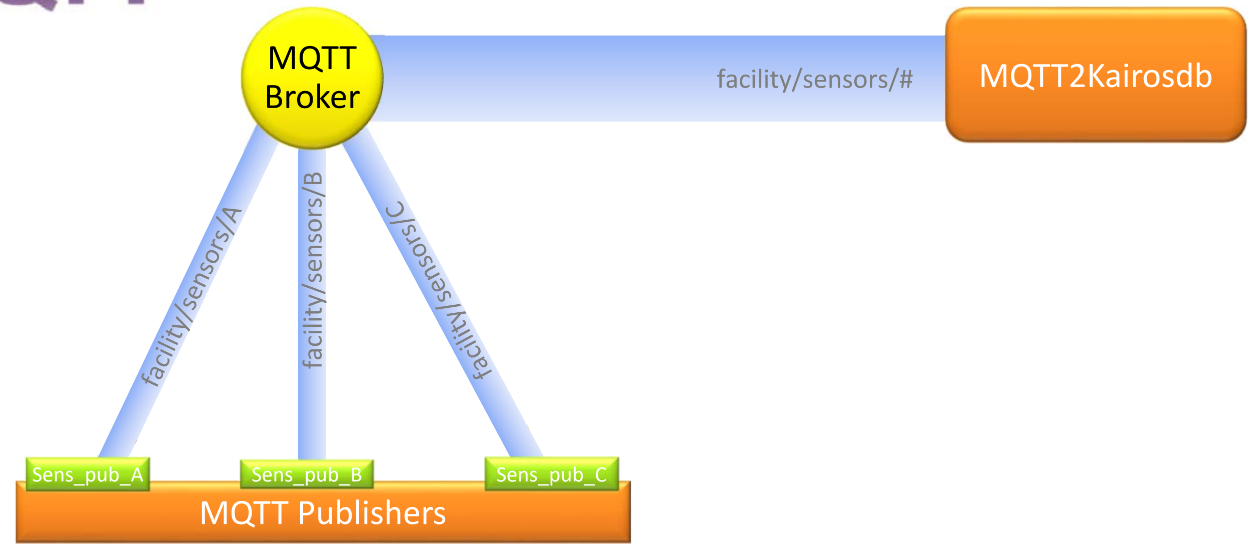
- **MQTT** Brokers
- Data **Visualization**
- **NoSQL** Storage
- **Big Data** Analytics

Back-end

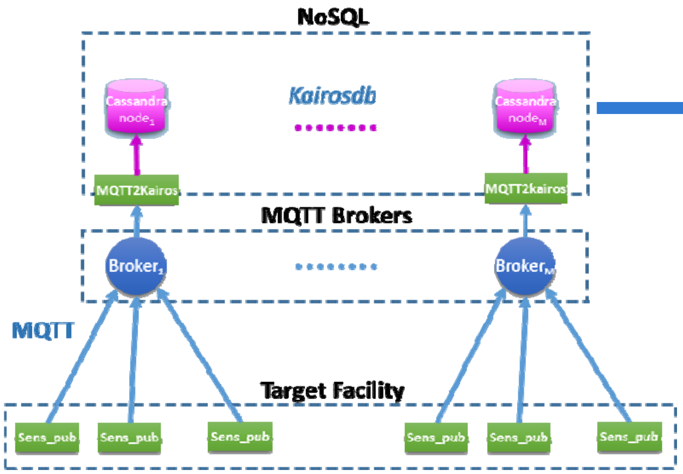
- MQTT–enabled sensor collectors

MQTT to NoSQL Storage: MQTT2Kairosdb

● = {Value;Timestamp}

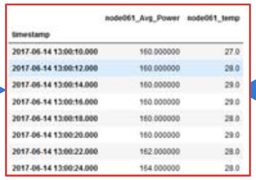


Examon Analytics: Batch & Streaming



examon-client
(REST)

Pandas
dataframe



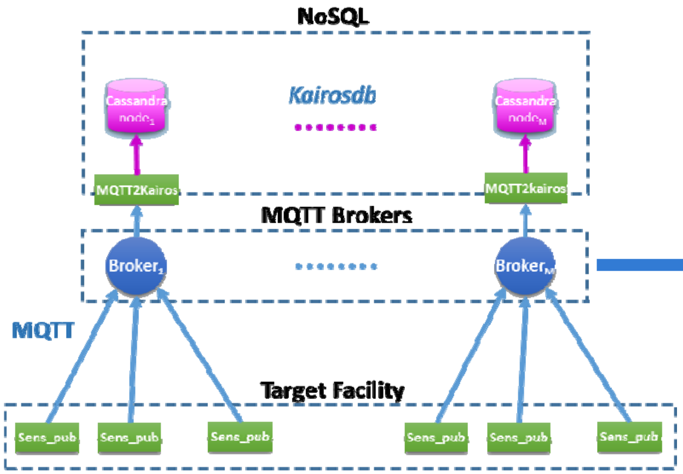
	node061_Avg_Power	node061_temp
timestamp		
2017-06-14 13:00:10.000	160.000000	27.0
2017-06-14 13:00:12.000	160.000000	28.0
2017-06-14 13:00:14.000	160.000000	29.0
2017-06-14 13:00:16.000	160.000000	29.0
2017-06-14 13:00:18.000	160.000000	28.0
2017-06-14 13:00:20.000	160.000000	29.0
2017-06-14 13:00:22.000	162.000000	28.0
2017-06-14 13:00:24.000	164.000000	28.0

Spark
(Batch)

PYTORCH

TensorFlow

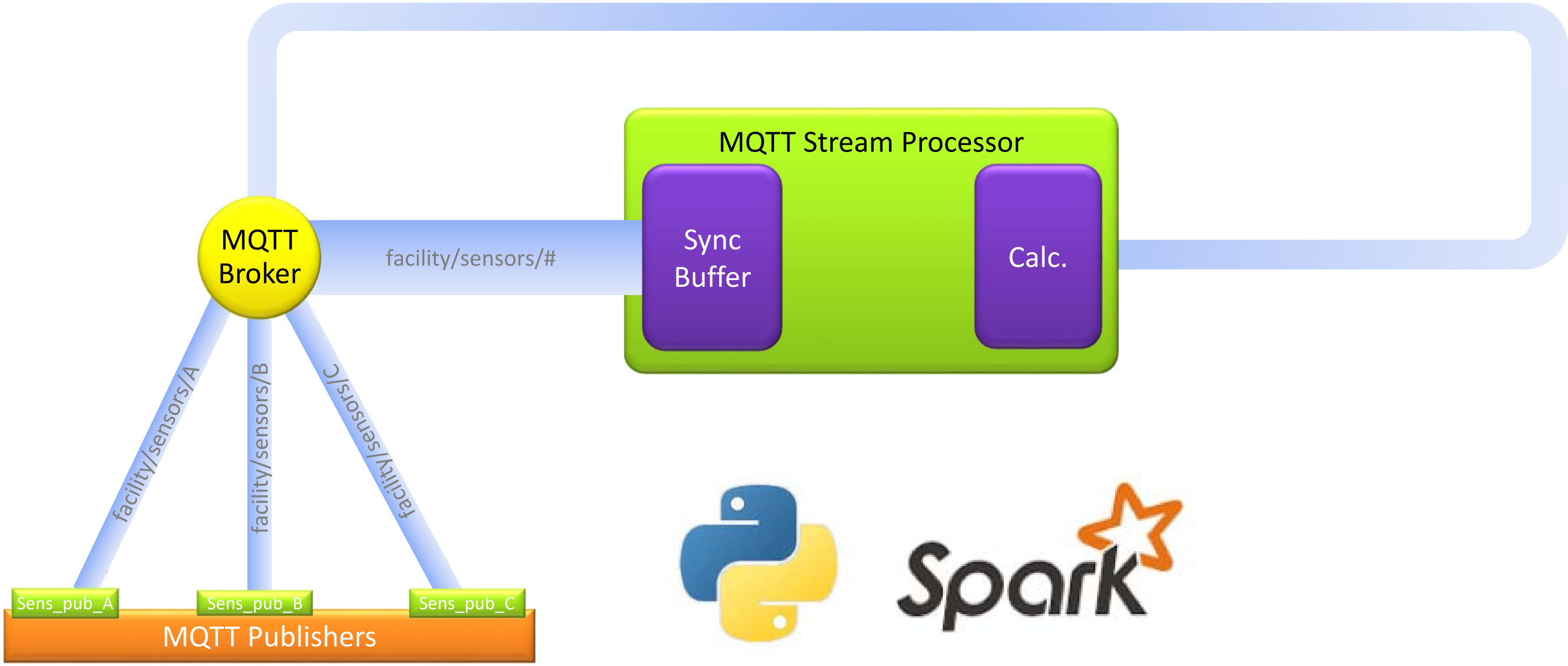
⋮



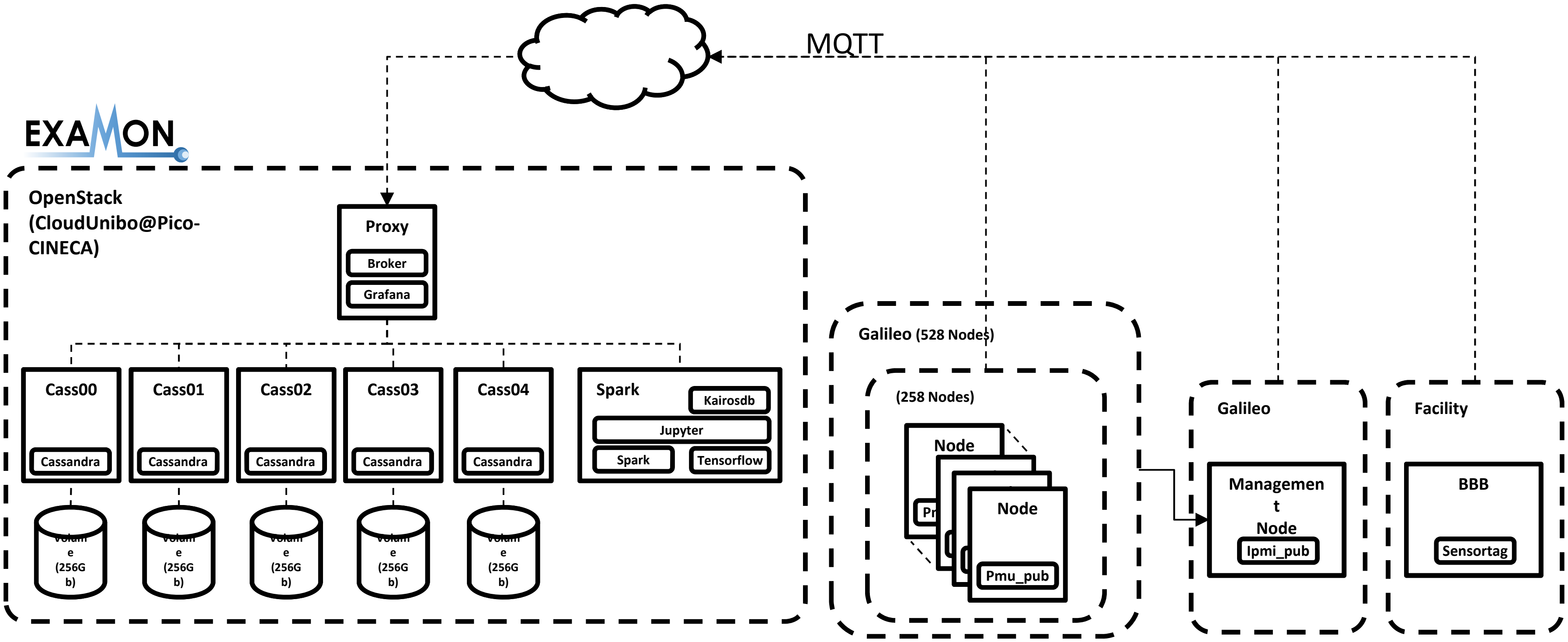
Bahir-mqtt
(Spark connector)

Spark
(Streaming)

Streaming Analytics: virtual sensors!



Examon in production: CINECA’s GALILEO



Data Ingestion Rate	~67K	Metrics/s
DB Bandwidth	~98	Mbit/s
DB Size	~1000	GB/week
DB Write Latency	20	us
DB Read Latency	4800	us

Tier1 system 0.5-1TB every week
Tier0 *estimated* 10TB per 3.5 Days

➡

Stream analytics & distributed processing are a necessity

Application Aware En2Sol Minimization

HARDWARE

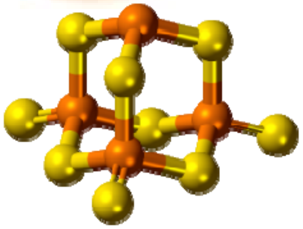
Galileo: Tier-1 HPC system based on an IBM NeXtScale cluster

- **Cluster**: 516 nodes (14 rack)
- **Node**: Dual socket Intel Haswell E5-2630 v3 CPUs with 8 cores at 2.4 GHz (85W TDP), DDR3 RAM 128 GB
- **Power consumption**: 360 KW
- **OS**: SMP CentOS Linux version 7.0
- **Top500**: Ranked at 281th

Compute node

SOFTWARE

Quantum ESPRESSO is an integrated suite of HPC codes for electronic-structure calculations and materials modelling at the nanoscale.

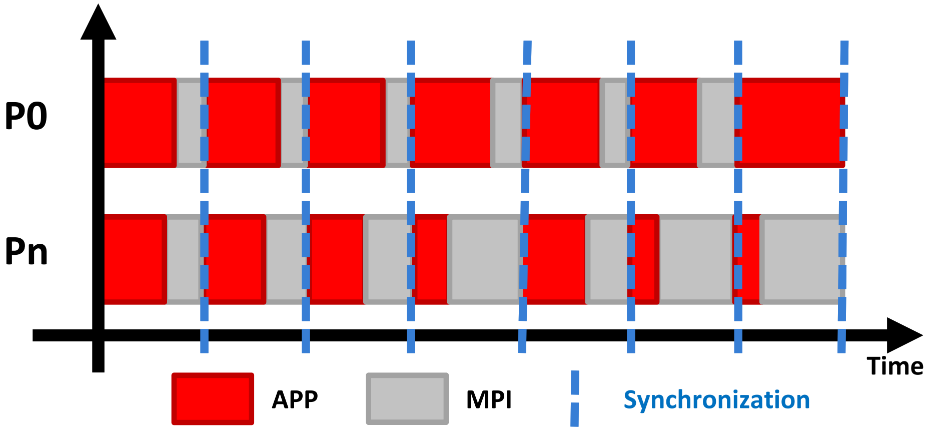


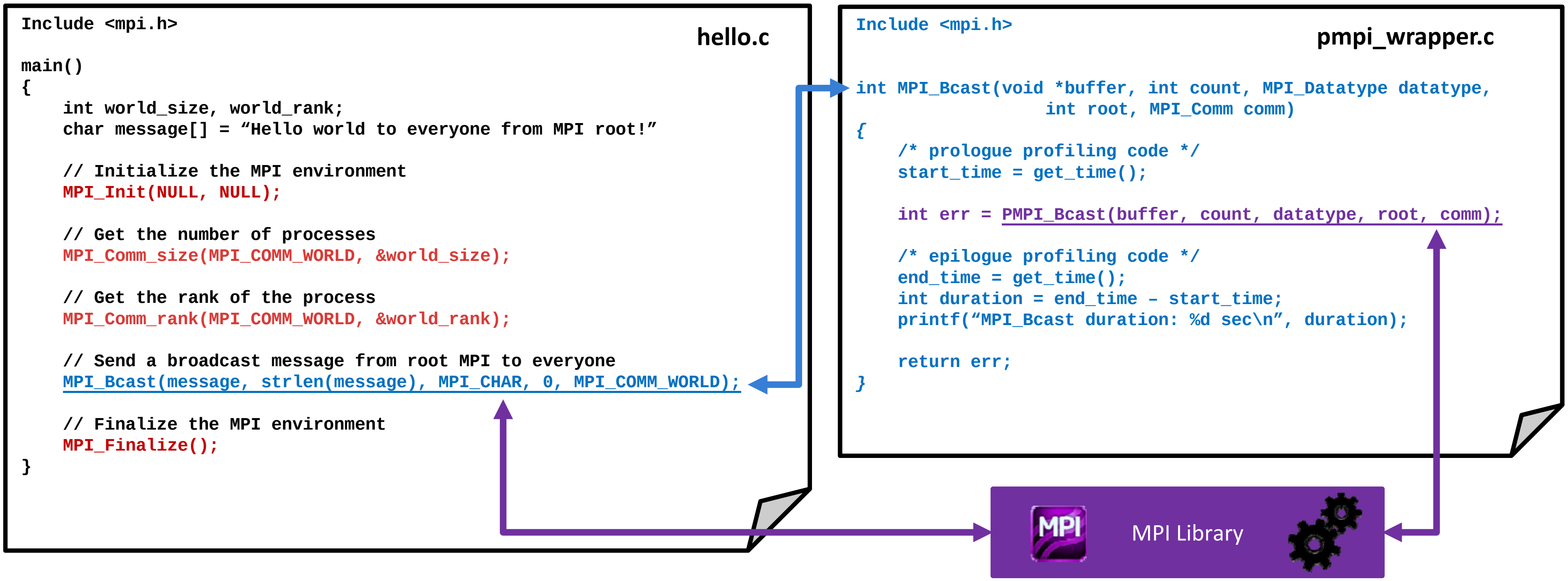
Car-Parrinello

Kernels

PMPI

MPI profiling interface Augment each standard MPI function with profiling collection functionality





hello.c

```
Include <mpi.h>

main()
{
    int world_size, world_rank;
    char message[] = “Hello world to everyone from MPI root!”

    // Initialize the MPI environment
    MPI_Init(NULL, NULL);

    // Get the number of processes
    MPI_Comm_size(MPI_COMM_WORLD, &world_size);

    // Get the rank of the process
    MPI_Comm_rank(MPI_COMM_WORLD, &world_rank);

    // Send a broadcast message from root MPI to everyone
    MPI_Bcast(message, strlen(message), MPI_CHAR, 0, MPI_COMM_WORLD);

    // Finalize the MPI environment
    MPI_Finalize();
}
```

pmpi_wrapper.c

```
Include <mpi.h>

int MPI_Bcast(void *buffer, int count, MPI_Datatype datatype,
                  int root, MPI_Comm comm)
{
    /* prologue profiling code */
    start_time = get_time();

    int err = PMPI_Bcast(buffer, count, datatype, root, comm);

    /* epilogue profiling code */
    end_time = get_time();
    int duration = end_time – start_time;
    printf(“MPI_Bcast duration: %d sec\n”, duration);

    return err;
}
```

PMPI Runtime

Our PMPI implementation has the following features:

- **Number MPI calls**: 50 MPI functions wrapped (all the QE's MPI calls)
- **Timing**: record TSC for timing (time clock accuracy)
- **Network data**: record all data sent and received from the process
- **Fixed perf counters**: monitor 3 fixed performance counters using low overhead *rdpmc()* instruction
  - Fixed 1: Number of instructions retired
  - Fixed 2: Clock at the nominal frequency at every active cycle
  - Fixed 3: Clock coordinated at frequency of the core at every active cycle
- **PMC perf counters**: monitor 8 configurable performance counters using low overhead *rdpmc()* instruction

Time Overhead: 0,59%

Memory Overhead?

It is related to:

- Number of MPI processes
- Application time
- Number of MPI calls

Example: 16 MPI processes, 7.40 min of application time and 3,5 Mln of MPI calls

Memory overhead: ≈250 MB

Average timing error wrt Intel Trace Analyzer: 0.45%

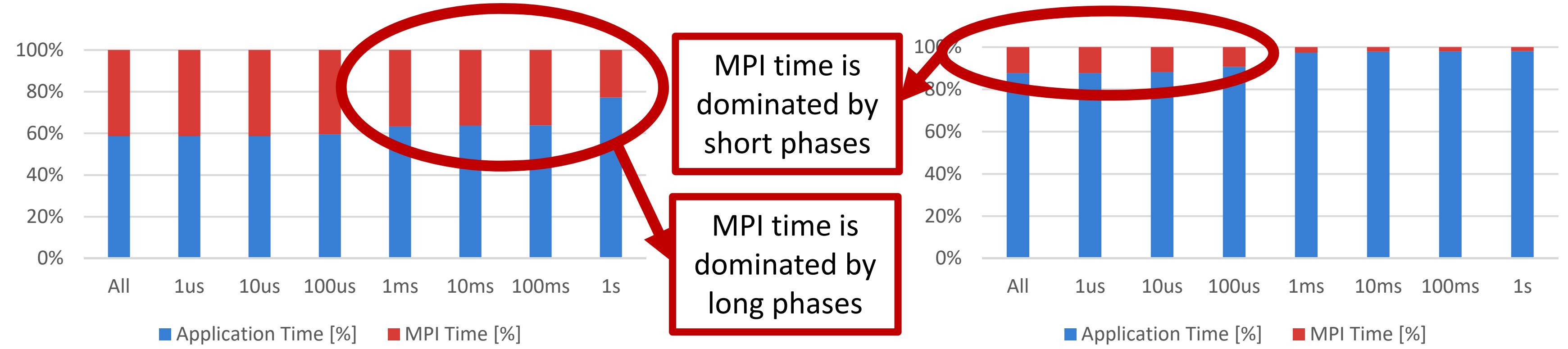
APP time vs MPI Time

Ndiag 1 **Workload MPI root**: 10.25% **AVG workload (no root)**: 5.98%

Linear algebra is computed only by the root MPI → unbalanced workload

Ndiag 16 **Workload MPI root**: 6.59% **AVG workload (no root)**: 6.23%

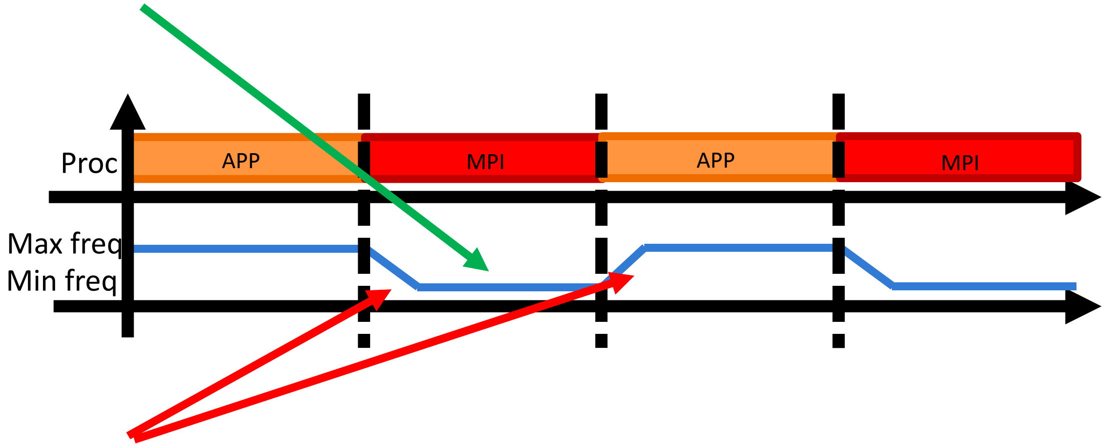
Linear algebra is computed by all MPI processes → balanced workload



MPI time is dominated by short phases

MPI time is dominated by long phases

Idea: use DVFS to slow down cores during MPI-phases

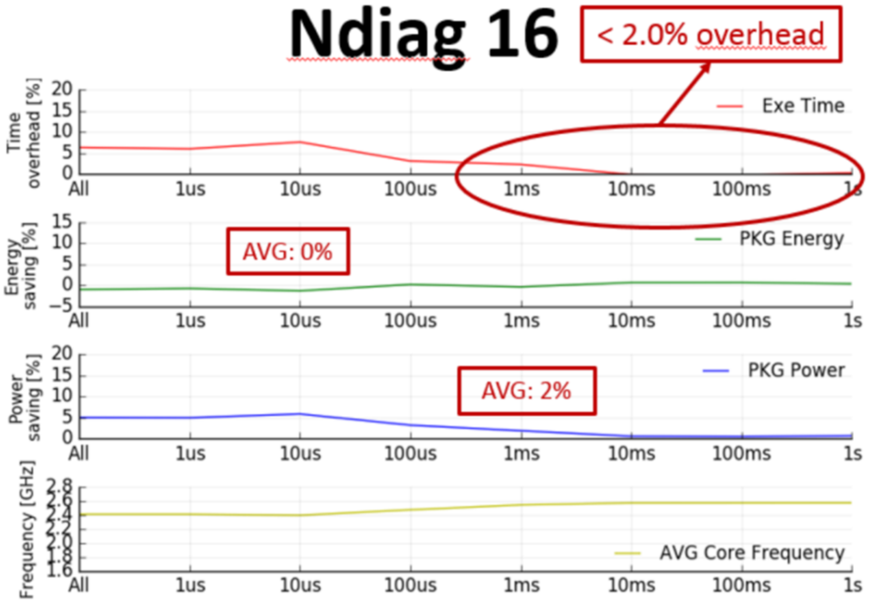


Challenge: Account for DVFS inertia, and appl. slowdown

PMPI-based E2Sol minimization

If QE has significant percentage of MPI time with MPI phases longer than 500us

➡ **PMPI needed to gauge and exploit (PMPI + PM) power saving opportunity**

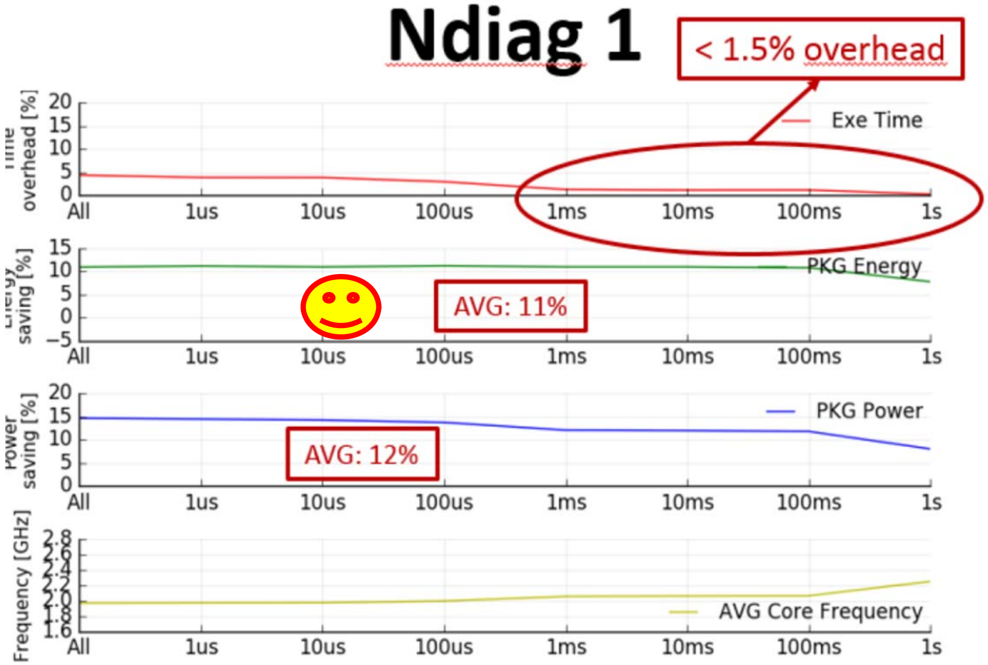


Ndiag 16

< 2.0% overhead

AVG: 0%

AVG: 2%



Ndiag 1

< 1.5% overhead

AVG: 11%

AVG: 12%

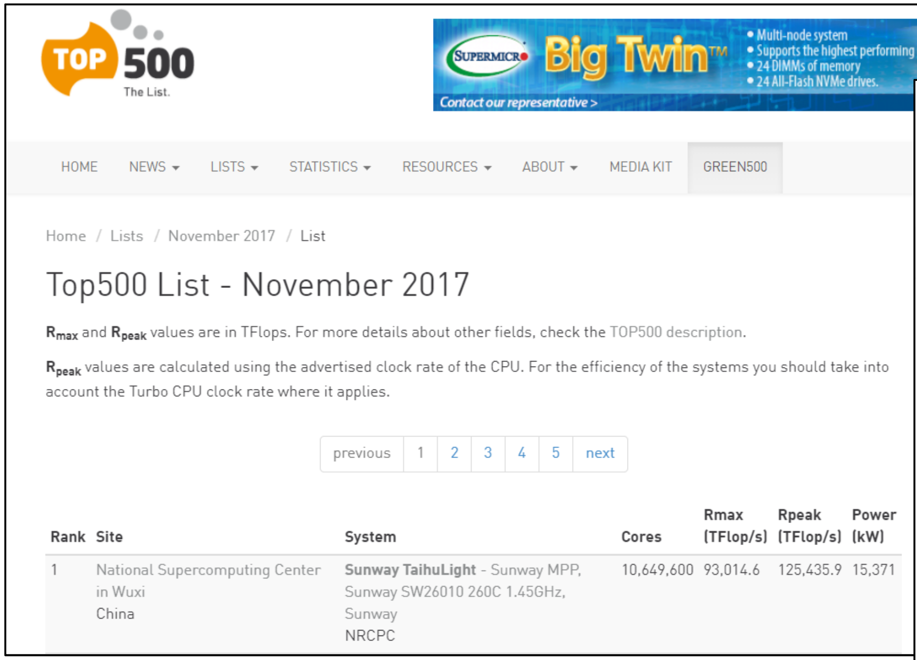
Unbalanced benchmark on a single node (negligible MPI communication time)

Up to 11% of energy and 12% of power saved with no impact on performance

Outline

- ❑ Power and Thermal Walls in HPC
- ❑ Power and Thermal Management
- ❑ Energy-efficient Hardware
- ❑ Conclusion

The era of Eterogeneous Architecture



Home / Lists / November 2017 / List

Top500 List - November 2017

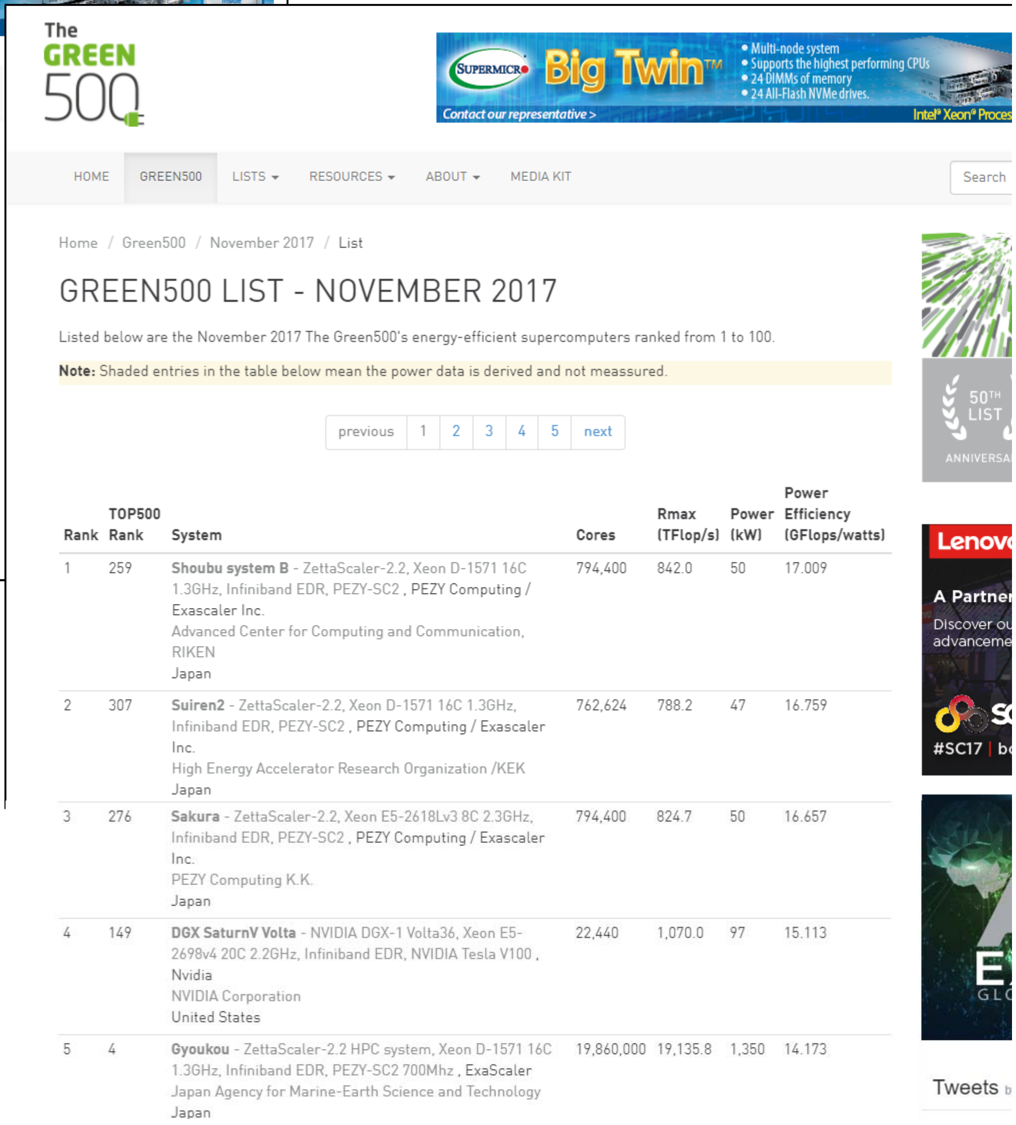
R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

R_{peak} values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

previous | 1 | 2 | 3 | 4 | 5 | next

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371

Massive presence of accelerators in TOP500



Home / Green500 / November 2017 / List

GREEN500 LIST - NOVEMBER 2017

Listed below are the November 2017 The Green500's energy-efficient supercomputers ranked from 1 to 100.

Note: Shaded entries in the table below mean the power data is derived and not meassured.

previous | 1 | 2 | 3 | 4 | 5 | next

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	259	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	842.0	50	17.009
2	307	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	788.2	47	16.759
3	276	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	149	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	1,350	14.173

Absolute dominance in GREEN500

Recipe for Energy-efficient Acceleration

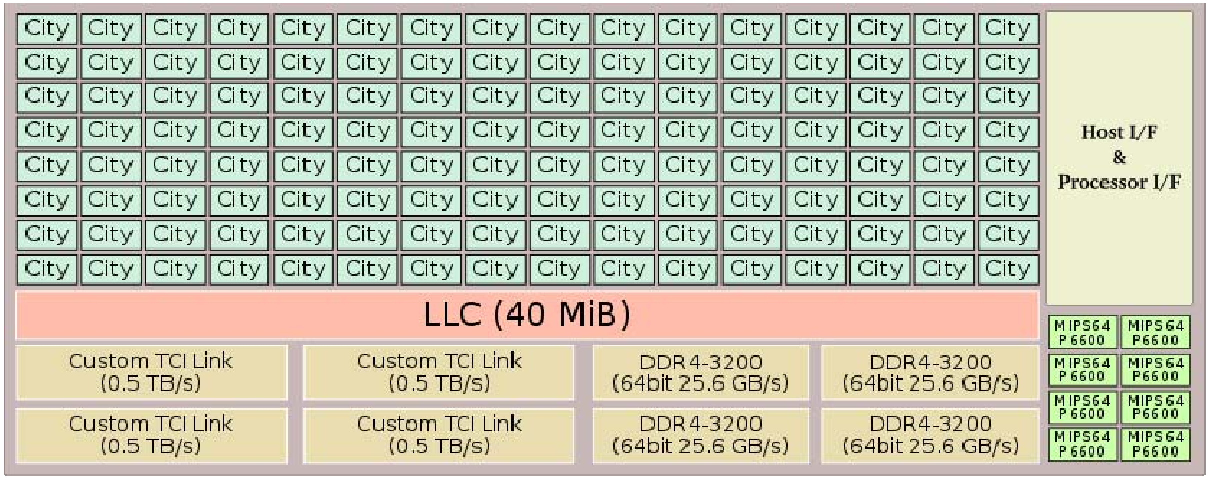
- Many (thousands) “simple” cores, managing FP units and special-function units for key workload patterns (stencil, tensor units) → maximize FP/mm2
- Non-coherent caches and lots of “non-cache” memory (registers for multithreading, scratchpad) → maximize “useful” Bit/mm2 for on-chip
- Large memory bandwidth based on tightly coupled-memory (HBM) → maximize GBps/mm2 for off-chip
- Low Operating voltage and moderate operating frequency → keep W/mm2 under control
- From 2D to 3D (now 2.5D)

Is there room for differentiation, or are GP-GPUs the only answer?

Pezy-SC2 (top 1-2-3 GREEN500 Nov17)

Pezy-SC highlights:

- Technology (16nm TSMC) - 54% power reduction
- Advance and integrated power delivery – 30% power reduction
- Low voltage operation (0.7v) – 16% power reduction
- Low performance host processor – 15% power reduction


City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
City City City City City City City City City City City City City City City City
Host I/F & Processor I/F
LLC (40 MiB)
MIPS64 P6600
MIPS64 P6600
MIPS64 P6600
MIPS64 P6600
MIPS64 P6600
MIPS64 P6600
MIPS64 P6600
MIPS64 P6600
Custom TCI Link (0.5 TB/s)
Custom TCI Link (0.5 TB/s)
DDR4-3200 (64bit 25.6 GB/s)
DDR4-3200 (64bit 25.6 GB/s)
Custom TCI Link (0.5 TB/s)
Custom TCI Link (0.5 TB/s)
DDR4-3200 (64bit 25.6 GB/s)
DDR4-3200 (64bit 25.6 GB/s)

Combines low-power design, simple (no legacy!) instruction set, advanced power management

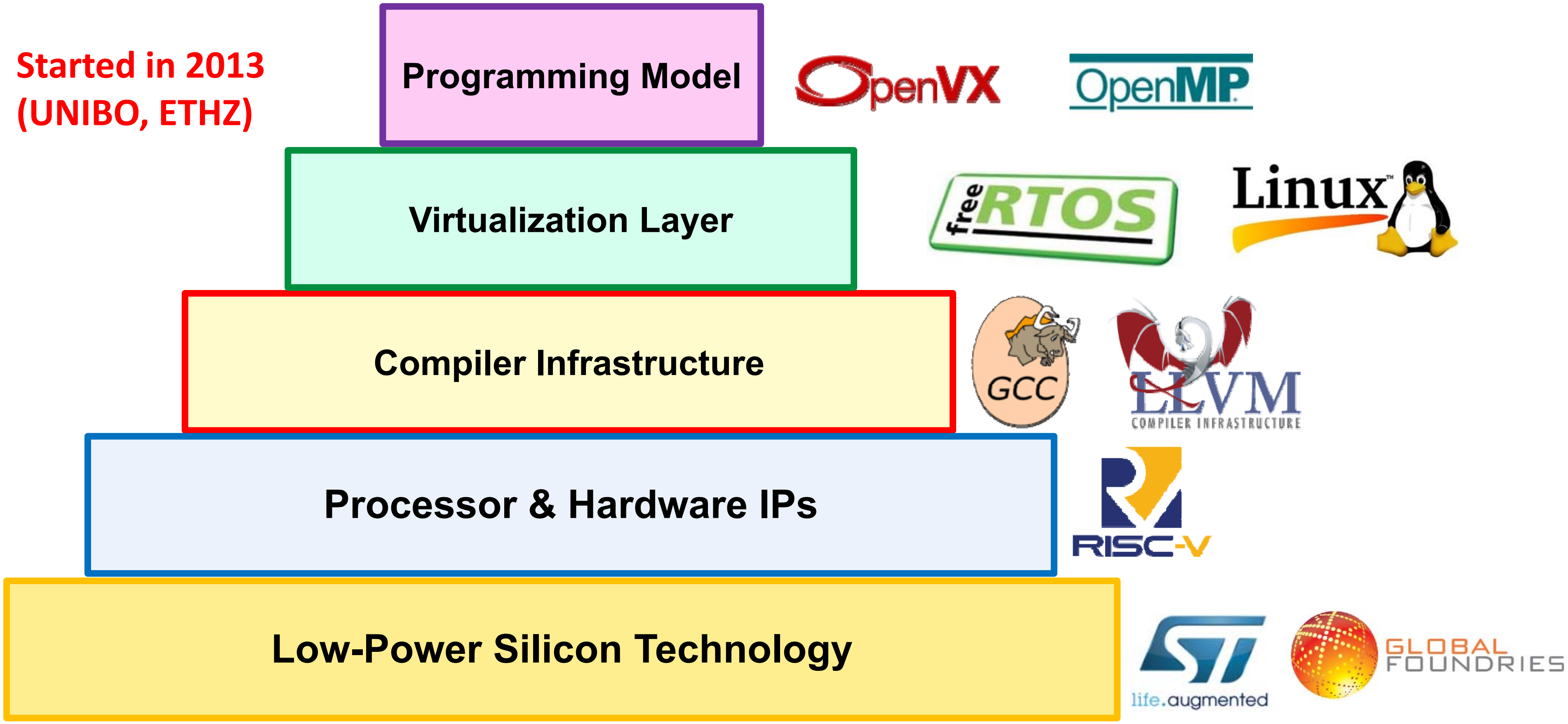
Opportunity for (EU) HPC: open ISA

RISC-V **open** RISC ISA developed by UCB and supported now by the RISC-V foundation (riscv.org), with 70+ members (including, NVIDIA, IBM, QUALCOMM, MICRON, SAMSUNG, GOOGLE…)

- Reasonable, streamlined ISA → distills many years of research, conceived for efficiency not for legacy support
- Safe-to-use free ISA → freedom to operate (see RISC-V genealogy project), freedom to change/evolve/specialize, no licensing costs
- Wide community effort already on-going on tools, verification… → leverage this to jumpstart and compensate for our initial inertia
- Rapidly gaining traction in many application domains (IoT, big data) → large “dual-use” markets opportunity
- **Spec covers 64bit, vector ISA (on-going), 128bit (planned)**
- **HPC-profile RISC-V startups already active (esperanto.ai)**

PULP: An Open Source Parallel Computing Platform

Started in 2013
(UNIBO, ETHZ)



Programming Model

Virtualization Layer

Compiler Infrastructure

Processor & Hardware IPs

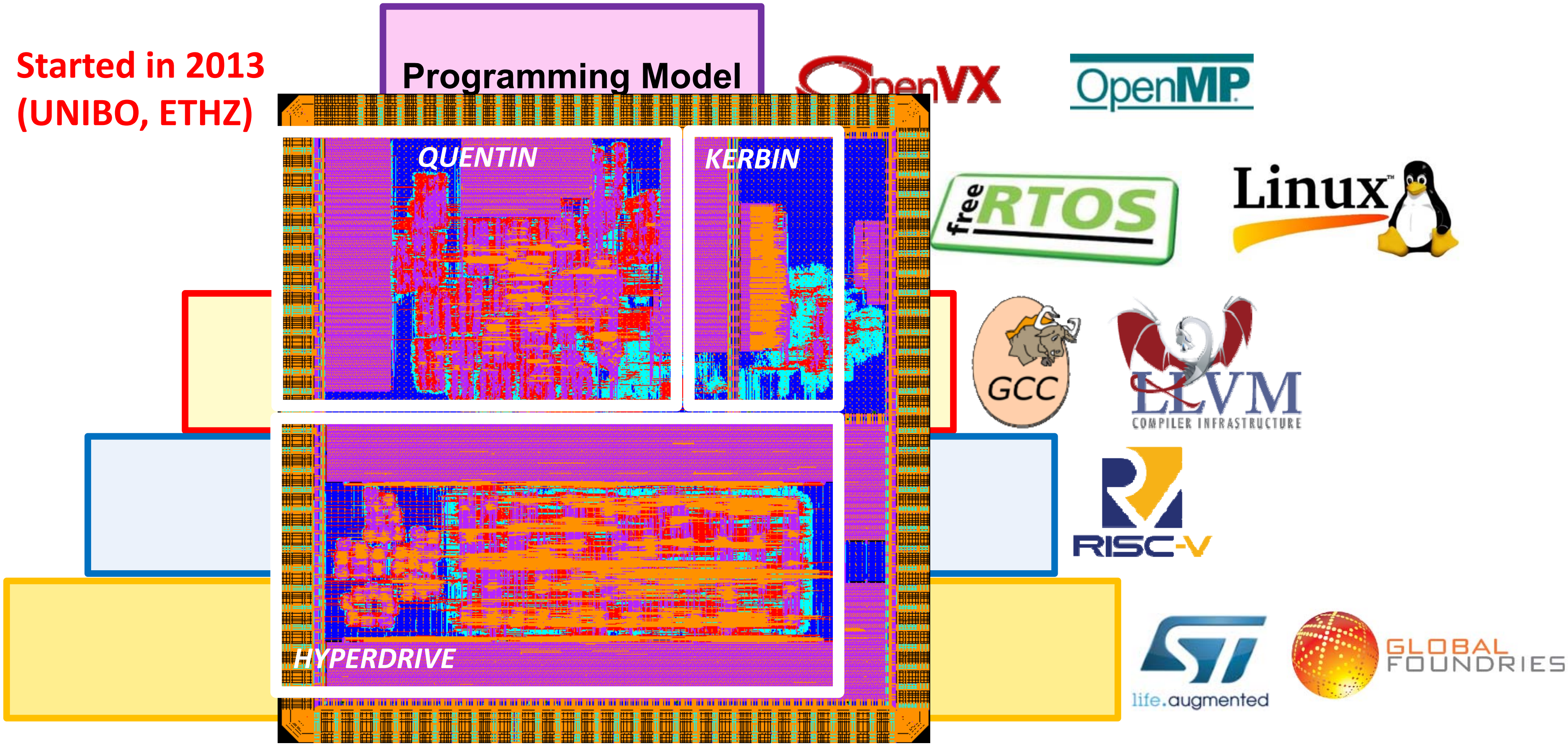
Low-Power Silicon Technology

PULP Hardware and Software released under Solderpad License

Used by tens of companies and universities, taped out in 14nm FINFET, 22FDX,…
64bit core "Ariane" + Platform to be launched in Q1 2018 (taped out in 22FDX)

PULP: An Open Source Parallel Computing Platform

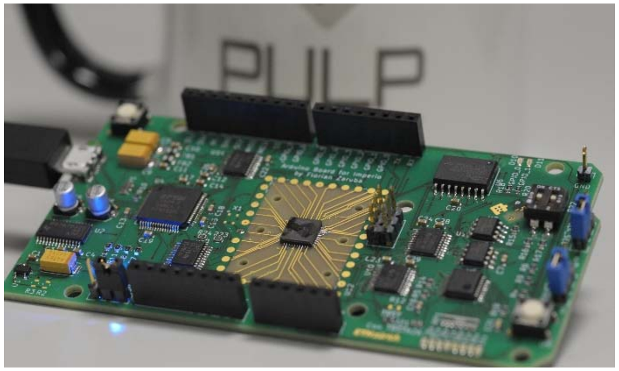
Started in 2013
(UNIBO, ETHZ)

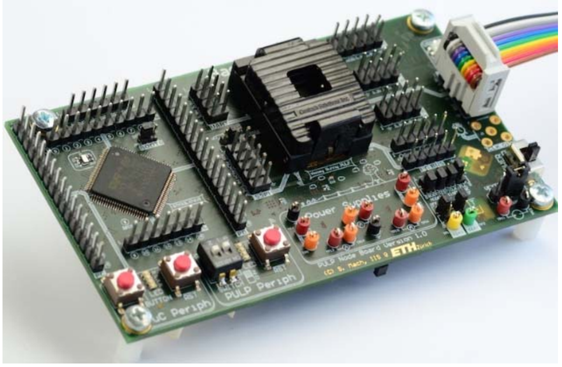


PULP Hardware and Software released under Solderpad License

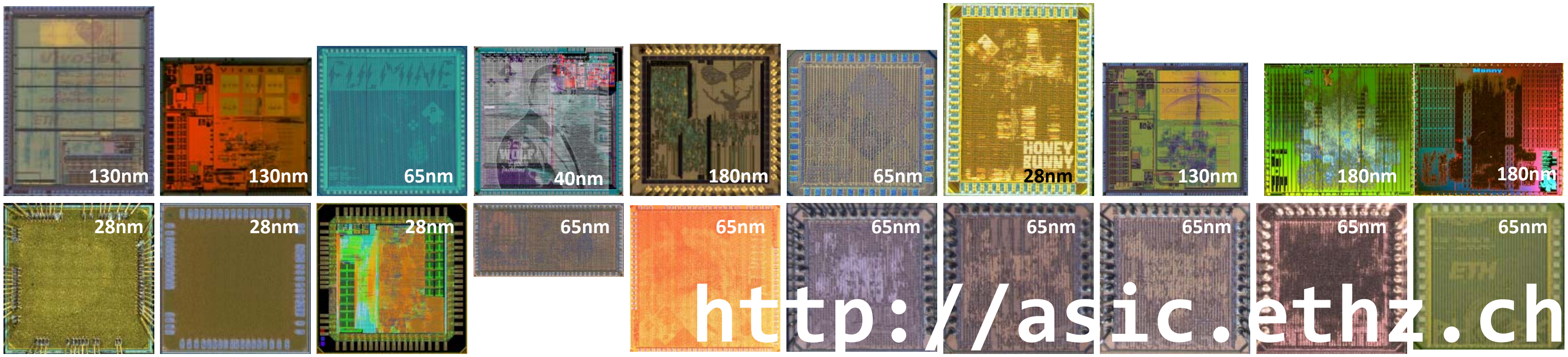
Used by tens of companies and universities, taped out in 14nm FINFET, 22FDX,…
64bit core "Ariane" + Platform to be launched in Q1 2018 (taped out in 22FDX)

Thanks for your attention!





www.pulp-platform.org



http://asic.ethz.ch

The fun is just beginning…