ETH zürich

ALMA MATER STUDIORUM
UNIVERSITA DI BOLOGNA

A Review of SoA HPC power monitoring systems & future trend on fine-grain data analytics

EU H2020 FETHPC
project ANTAREX
(g.a. 671623)

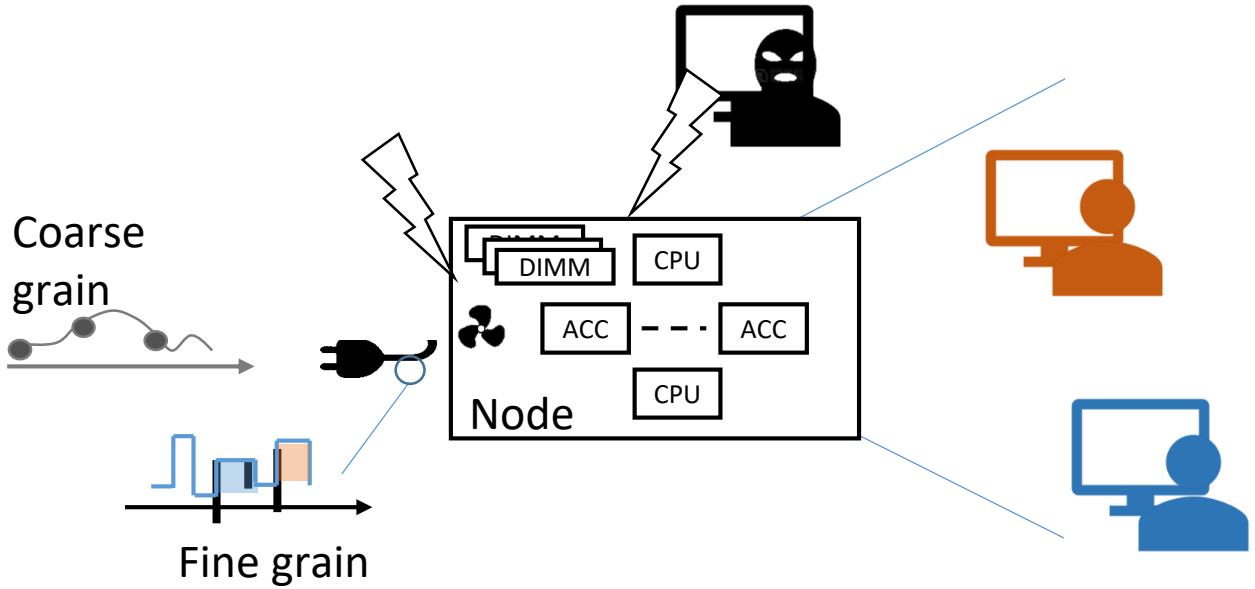
EE in HPC @CINECA
Ostrava, IT4I
30-Jan-2020

Andrea Bartolini
(slides from Antonio Libri)

Outline

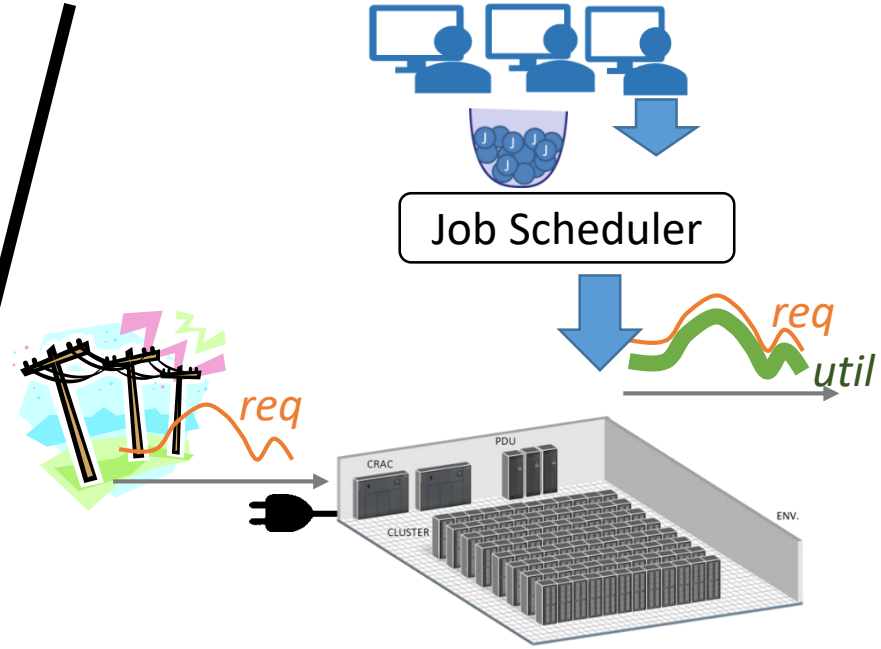
- Motivation – why reliable and high-res power/energy mon?
- SoA power monitoring systems
- Future trend on fine-grain measurements: data analytics

Several Challenges for HPC and Data-centers



Fine-Grain Power and Performance Measurements:

- Verify and classify node performance (& anomalies)
  - In spec / out of spec behaviour
  - Miss configuration
  - Aging and wear out
- Predictive maintenance



System Power Capping (**reliable** energy measurements)

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

Top500/Green500 Power Meas methodology [1]

- Document in [1] reports requirements to rank a HPC system in Top500/Green500

Requirements	Level 1 (Min Quality)	Level 2	Level 3 (Best Quality)

Granularity	1S/s	1S/s	• Continuously integrated energy
			• V & I sampled **at least**
			  • @5kS/s for AC
			  • @120S/s for DC

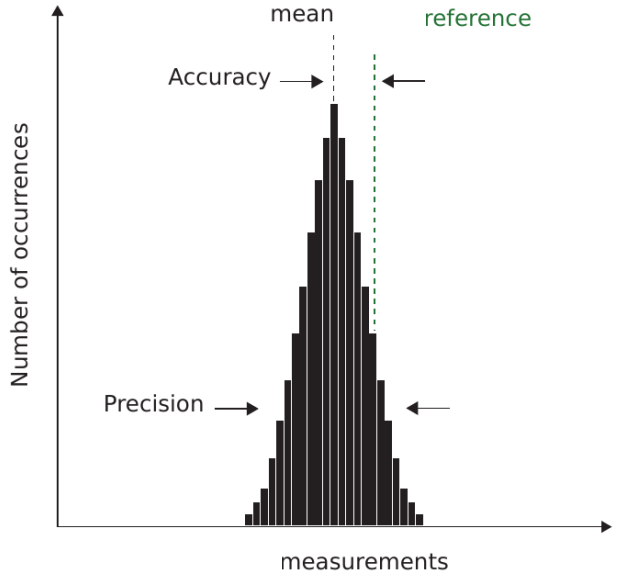
Precision (1σ - relative error)	5%	2%	**Below 1%**
Meas **Synch** b/w different meters	Below sampling period (e.g., NTP)	Below sampling period (e.g., NTP)	**Below sampling period** (e.g., NTP)

[1] EEHPC WG, "*Energy Efficient High Performance Computing Power Measurement Methodology", v2.0 RC 1.0*

Some (CORRECT) terminology

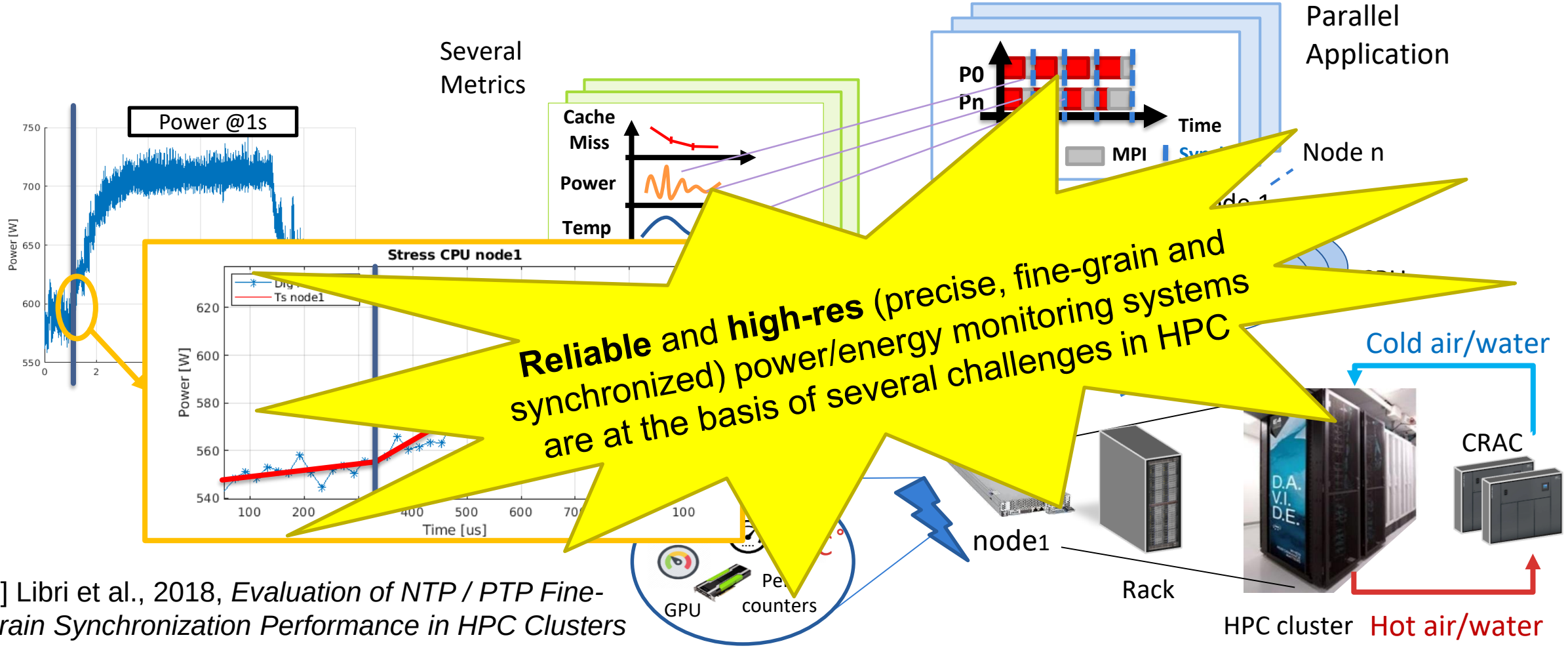
- Often the terms accuracy and precision are wrongly interchanged
- However
  - Accuracy → mean (can be fixed by calibration)
  - Precision → Std Dev (no fix)

Note: This definitions are used for any set of measurements (e.g., **power** meas, but also **time synch** measurements)



Accuracy and precision of a set of measurements.

Fine-grain Sync w. App Phases on HPC Clusters [1]



[1] Libri et al., 2018, *Evaluation of NTP / PTP Fine-Grain Synchronization Performance in HPC Clusters*

Outline

- Motivation – why reliable and high-res power/energy mon?
- SoA power monitoring systems
- Future trend on fine-grain measurements: data analytics

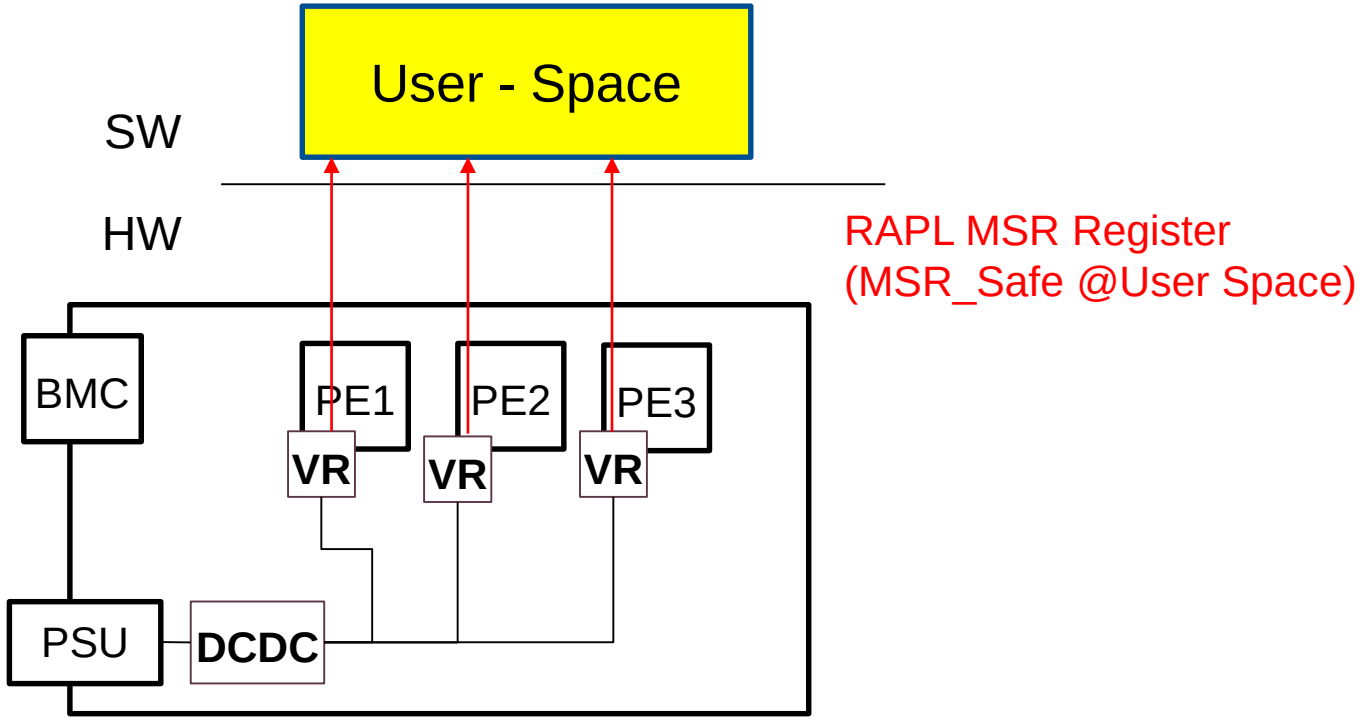
SoA HPC Monitoring Systems

- Current solutions allow to collect measurements **in-band** and **out-of-band** (no overhead on PE) via
  - built-in tools (e.g., IPMI, Amester, RAPL → hw perf counters)
  - custom sensors (e.g., HDEEM, HAEC → fine grain power meas)

SoA In-band HPC Power Monitoring systems

State-of-the-art:
1. **Intel RAPL [1]**

- Sampling time up to 1ms
- Reading energy via RAPL MSR registers
- Synchronization via NTP/PTP (vendor dependent)
- Precision is vendors dependent
- Scalable

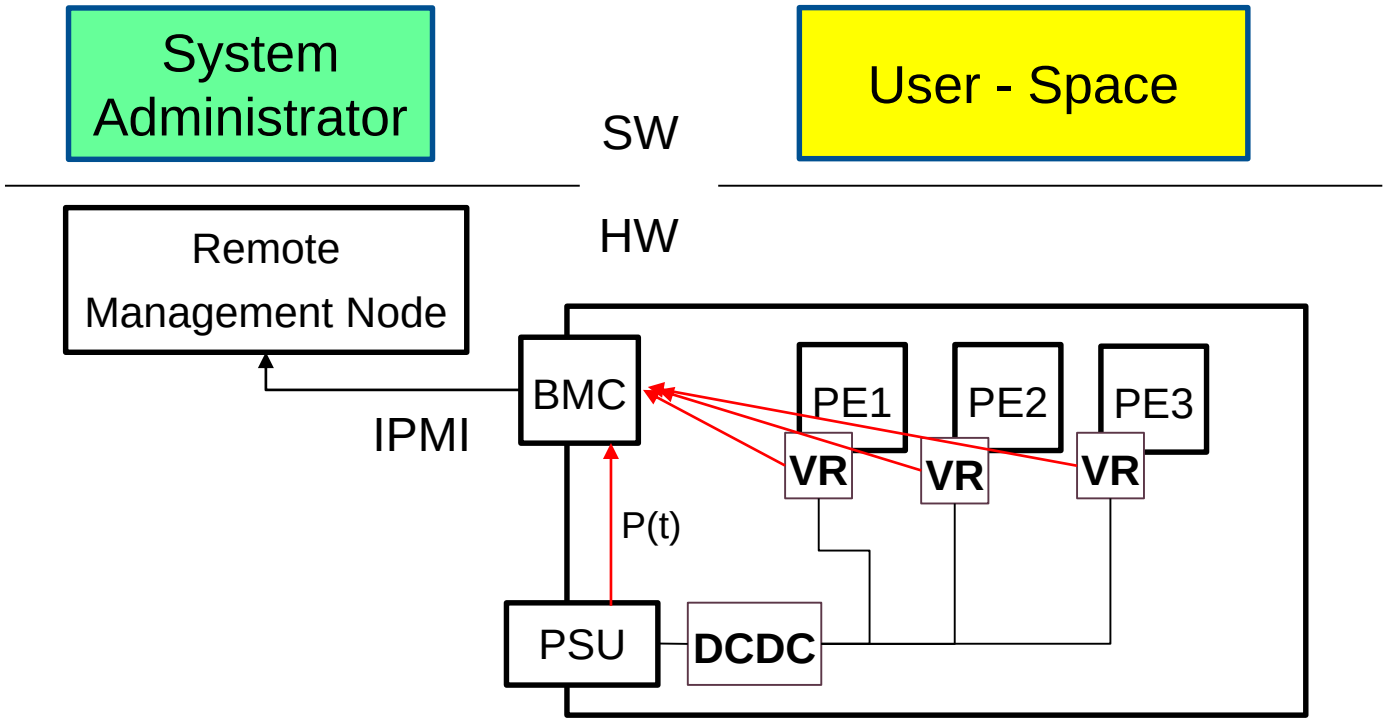


[1] M. Hahnel et al., "*Measuring Energy Consumption for short Code Paths Using RAPL*"

SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. **BMC – IPMI [1]**

- Slow sampling time (seconds)
- Unreliable time stamping
- Instantaneous power measurement – no energy / aliasing
- Precision is vendors dependent
- Scalable

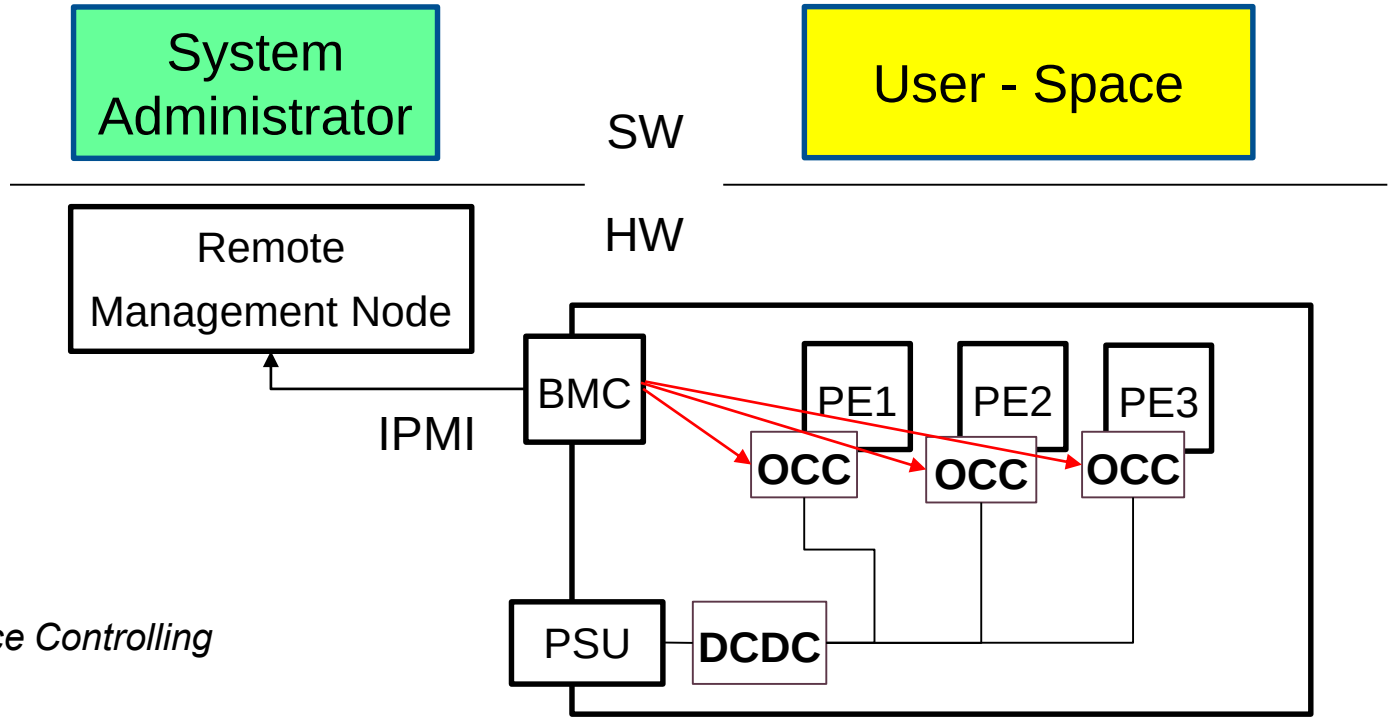


[1] IPMI spec v2.0 rev1.1, April 2015

SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. BMC – IPMI
2. **IBM Amester [1]**

- BMC reads Power measurements via OCC (within PE)
- Also in-band version in new Power9
- Time resolution up to 250µs (8kB buffers)
- Synchronization via NTP/PTP (vendor dependent)
- Precision is vendors dependent
- Scalable

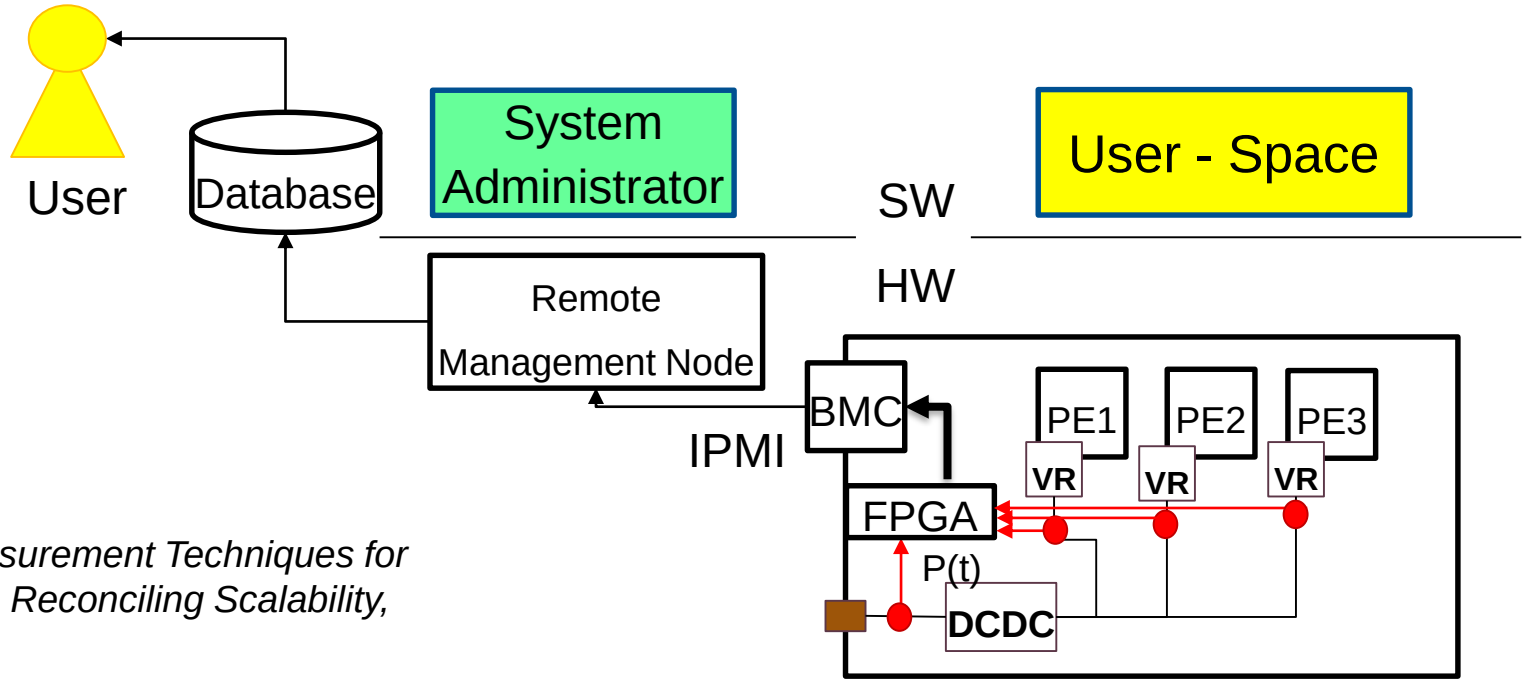


[1] T. Rosedahl et al., *"Power/Performance Controlling Techniques in OpenPOWER"*

SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. BMC – IPMI
2. IBM Amester
3. **HDEEM [1]**

- Time resolution up to 1ms (VR on CPU, DDR) - 125 µs (plug)
- Precision of 2% and 3%, respectively
- Time synchronization up to ms via NTP
- Scalable

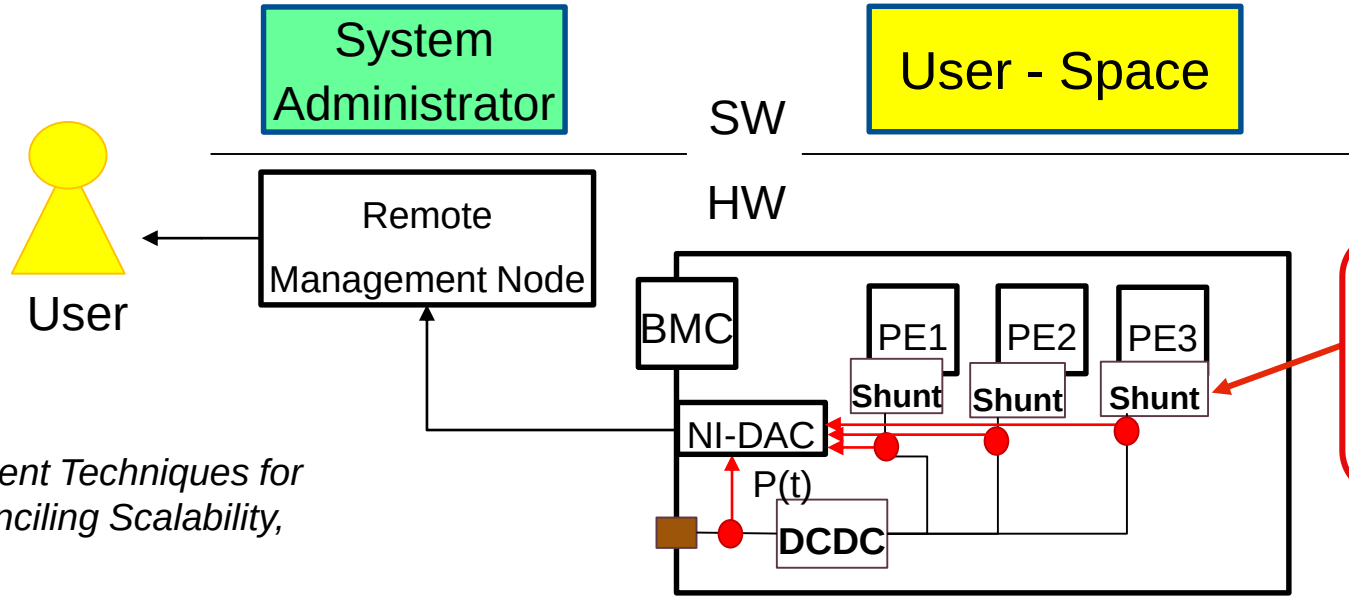


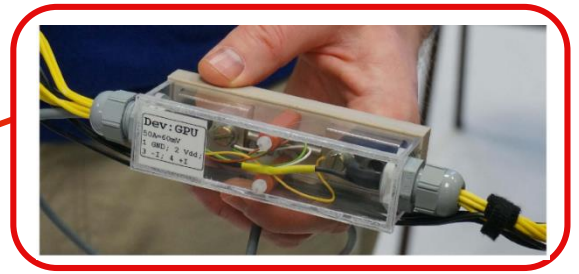
[1] Ilsche et al., "*Power Measurement Techniques for Energy-Efficient Computing: Reconciling Scalability, Resolution and Accuracy*"

SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. BMC – IPMI
2. IBM Amester
3. HDEEM
4. **HAEC [1]**

- 2 NI-DAC, one @7 kS/s (T=143µs - VR on CPU, DDR), one @500kS/s (T=2µs – power plug)
- Precision below 2%
- Current monitoring with Shunt Resistors (tested also HE Sensor)
- Time synchronization up to ms via NTP
- Not Scalable (single node only)





Shunt Resistor

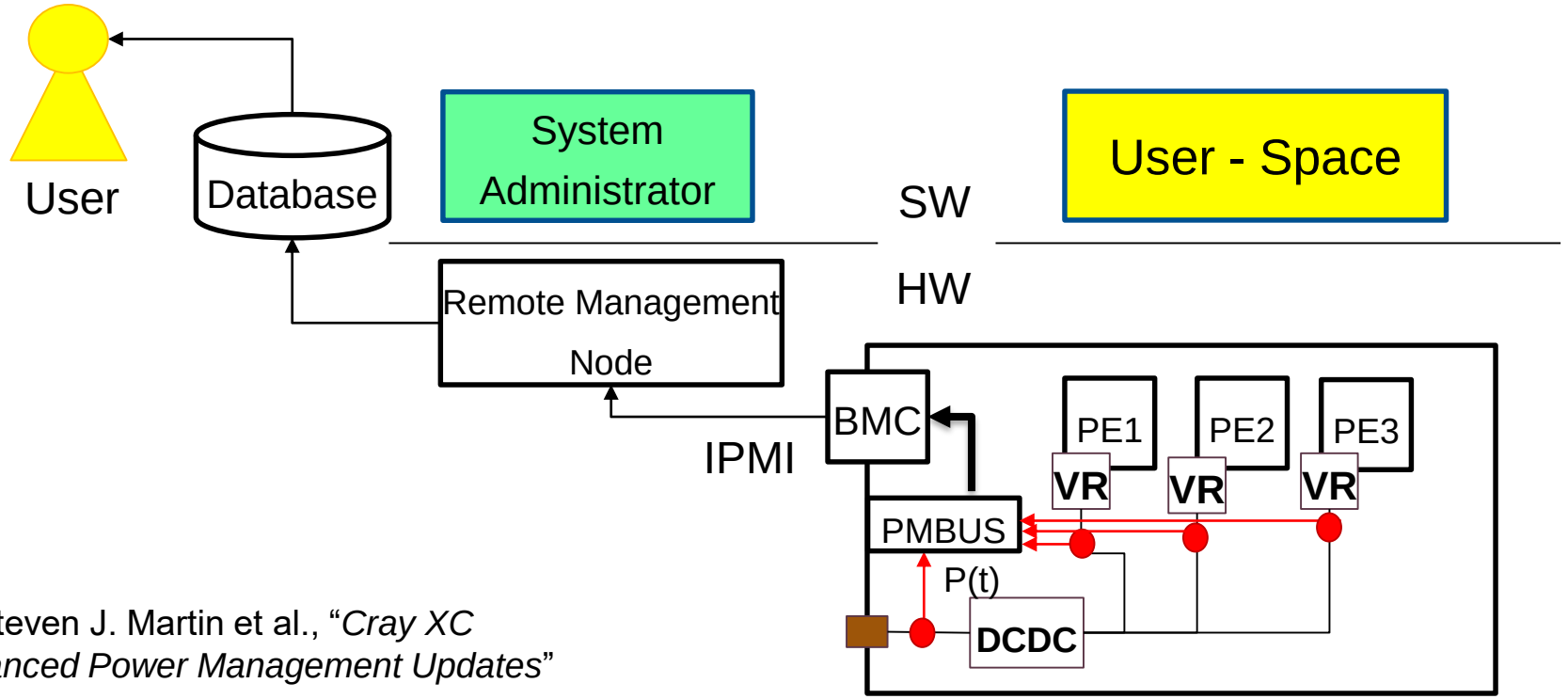
[1] Ilsche et al., "*Power Measurement Techniques for Energy-Efficient Computing: Reconciling Scalability, Resolution and Accuracy*"

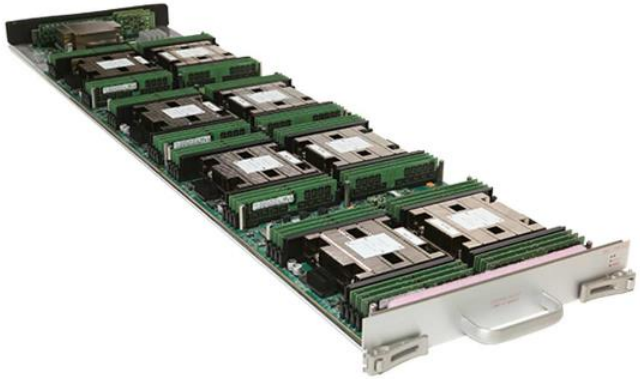
SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. BMC – IPMI
2. IBM Amester
3. HDEEM
4. HAEC
5. **CRAY XC APM [1]**

→

- Measurements via PMBus
- Time res up to 1s (Node Power) and 10s (CPU & Memory Power)
- Precision ±2.5% (by datasheet)
- Time synchronization up to ms via NTP
- Scalable





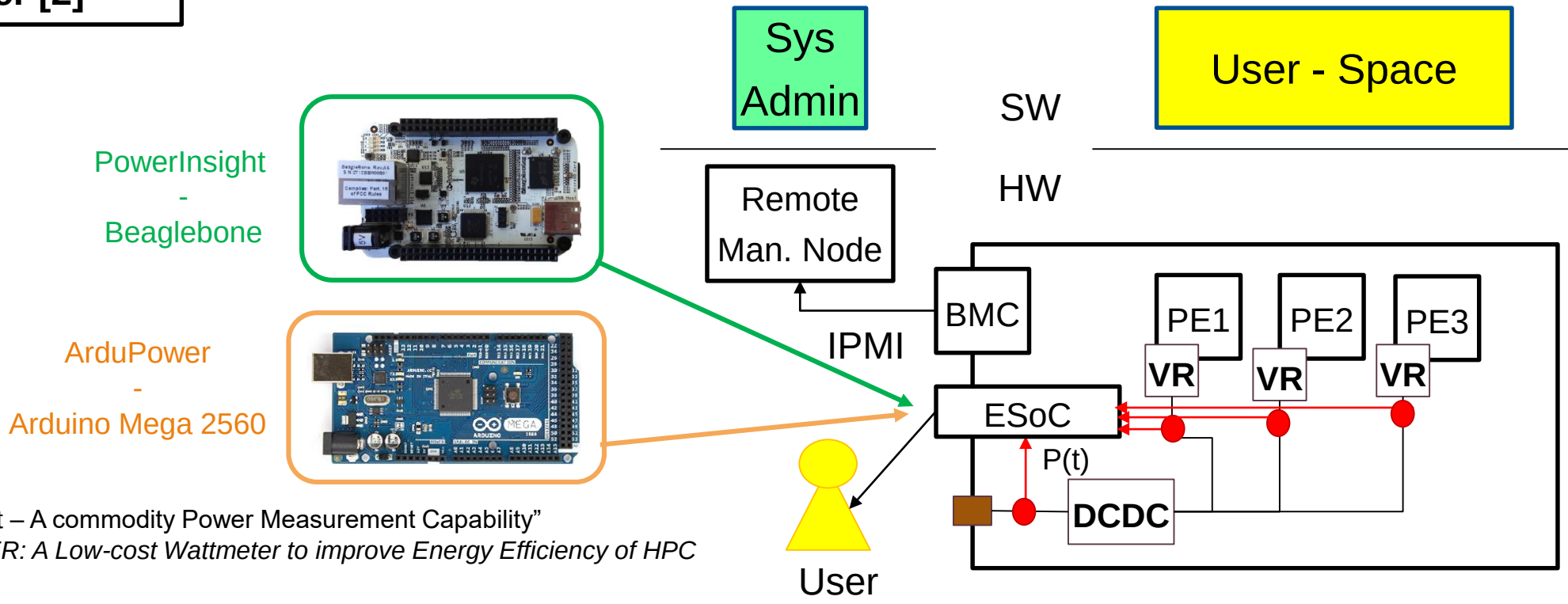
Cray XC50 Scalable Processors Blade

[1] Steven J. Martin et al., "*Cray XC Advanced Power Management Updates*"

SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. BMC – IPMI
2. IBM Amester
3. HDEEM
4. HAEC
5. CRAY XC APM
6. **PowerInsight [1] / Ardupower [2]**

- Open, low cost embedded SoC (PI → Beaglebone; AP → Arduino)
- Current monitoring with HE Sensor
- Time res up to 1ms on PowerInsight; ~2ms on ArduPower
- Precision 1.8% PowerInsight; not reported on ArduPower
- Time synchronization up to ms via NTP
- Scalable



[1] J. L. Laros et al., "PowerInsight – A commodity Power Measurement Capability"
[2] M. F. Dolz et al., "*ARDUPOWER: A Low-cost Wattmeter to improve Energy Efficiency of HPC Applications*"

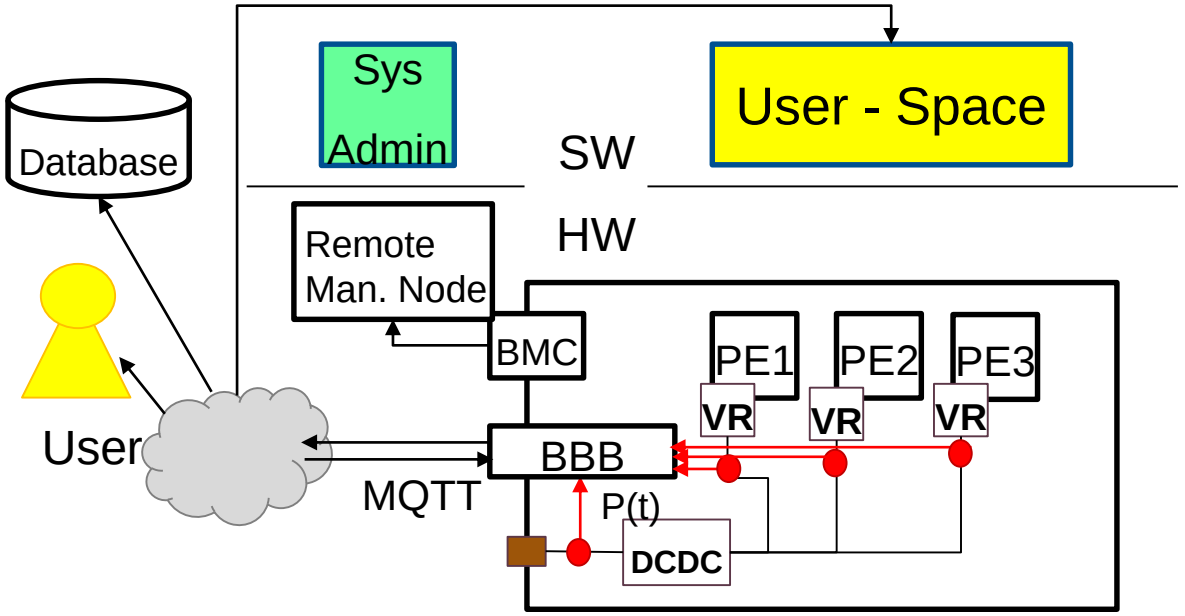
SoA Out-of-band HPC Power Monitoring systems

State-of-the-art:
1. BMC – IPMI
2. IBM Amester
3. HDEEM
4. HAEC
5. CRAY XC APM
6. PowerInsight / Ardupower
7. **DiG [1]**

- Open, Low-Cost Embeeded (BeagleboneBlack - ARM-A8)
- Tested both HE Sens / Shunt Res on several arch (Intel, ARM, IBM)
- Time res up to 20 µs @plug and reading VR at different rates depending on arch
- Precision below 1%
- Time synchronization up to µs via PTP
- Scalable: Big data communication protocol (MQTT) + Real-time edge analytics of the fine-grain measurements

Beaglebone Black (BBB)

- 1GHz ARM Cortex-A8
- 12bit 8-ch SAR ADC
- PTP HW enabled

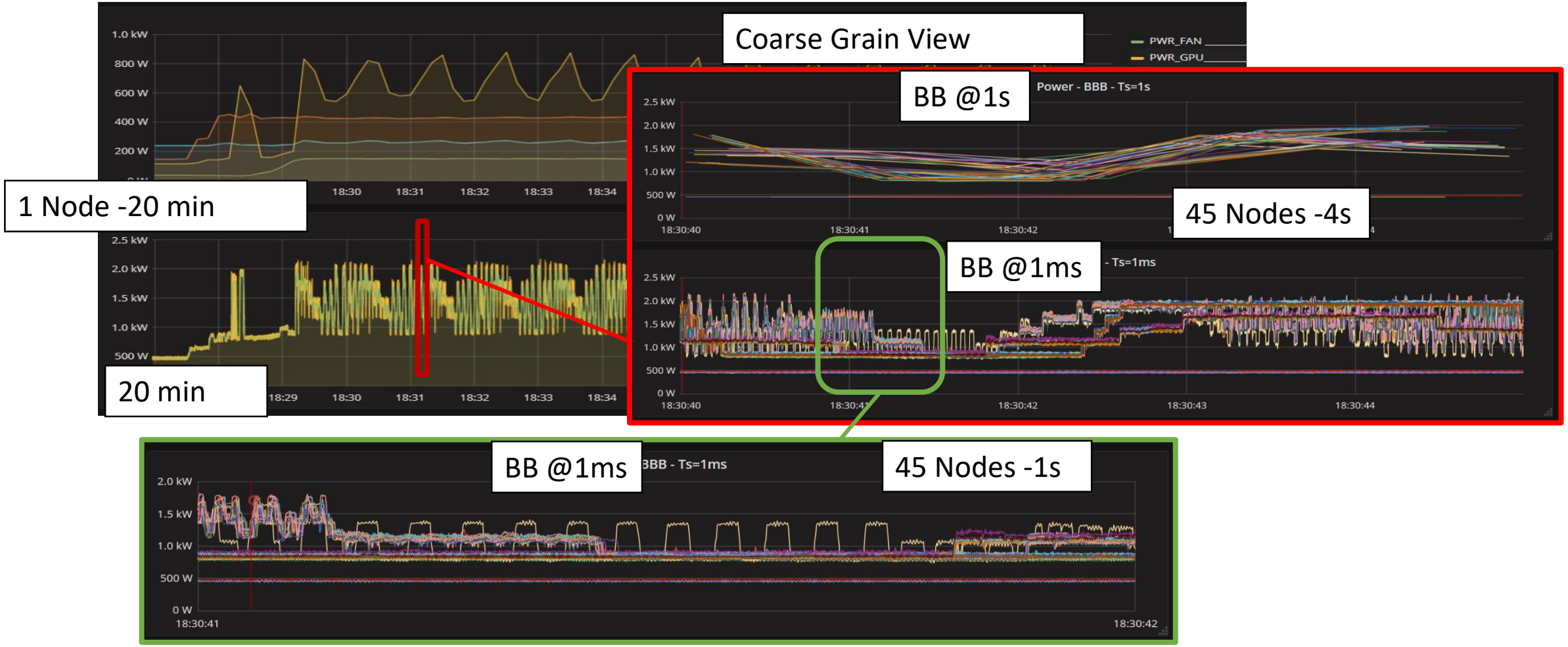


[1] Libri et al., "*DiG: Enabling Out-of-Band Scalable High-Resolution for Data-Center Analytics, Automation and Control*"

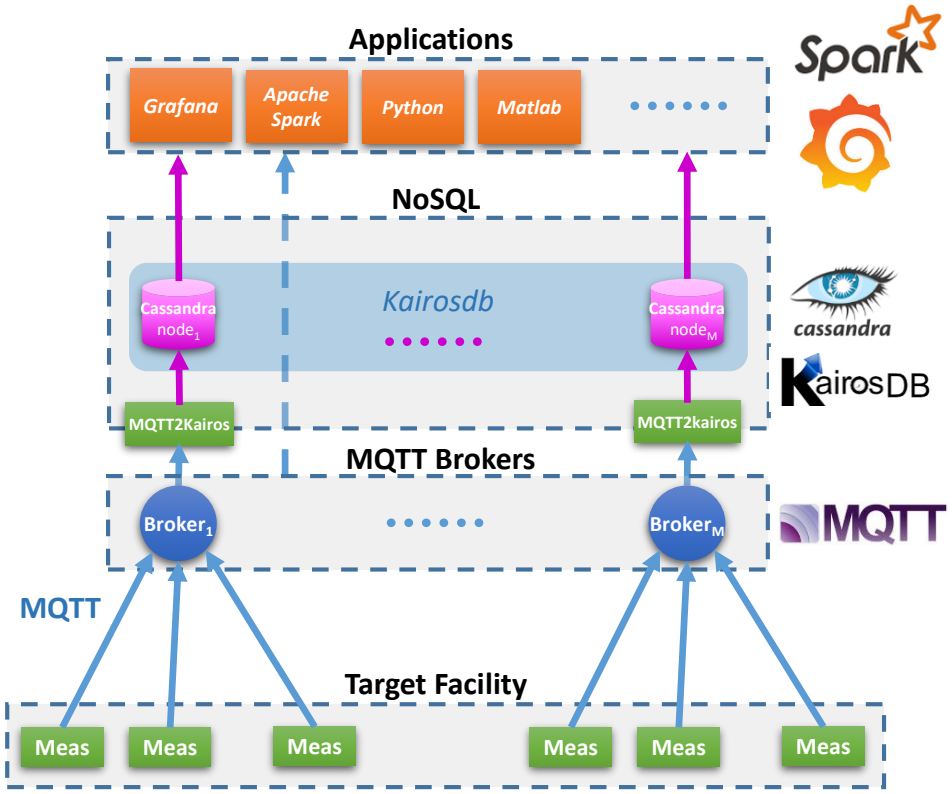
Outline

- Motivation – why reliable and high-res power/energy mon?
- SoA power monitoring systems
- Future trend on fine-grain measurements: data analytics

Example of Fine-Grain Power Meas w. DiG



Cluster Analytics – Data Collection SW Stack [1,2]



Front-end

- Exploit Cassandra (NoSQL DB)
- Data Visualization (Grafana) and Cluster Level ML (Spark, both RT & batch mode)

Back-end

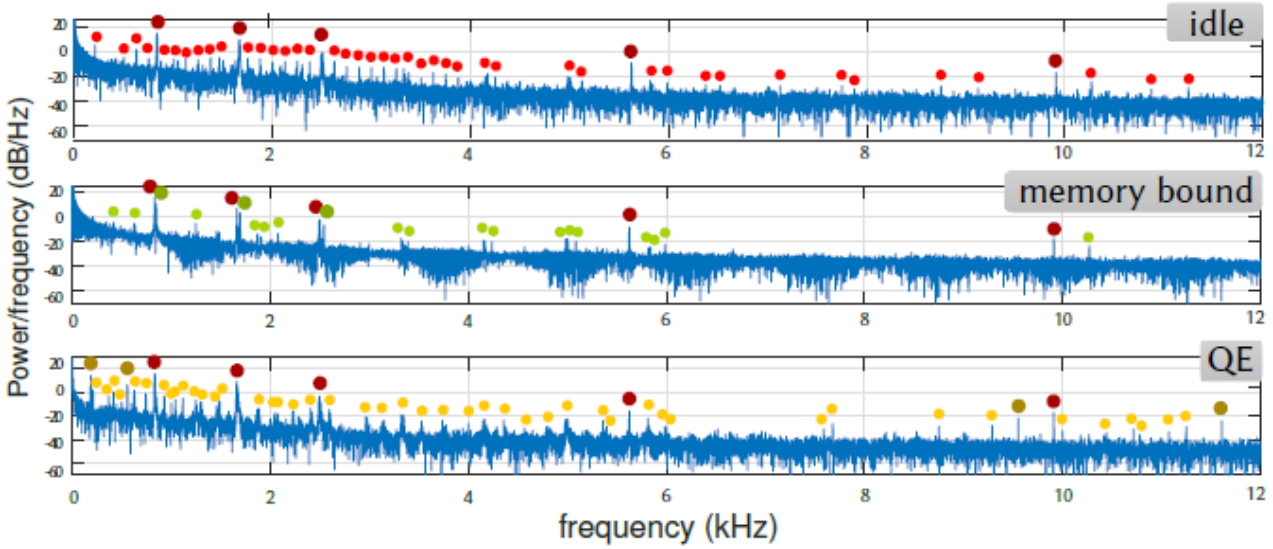
- Send Pow and Perf measurements for cluster-level analytics

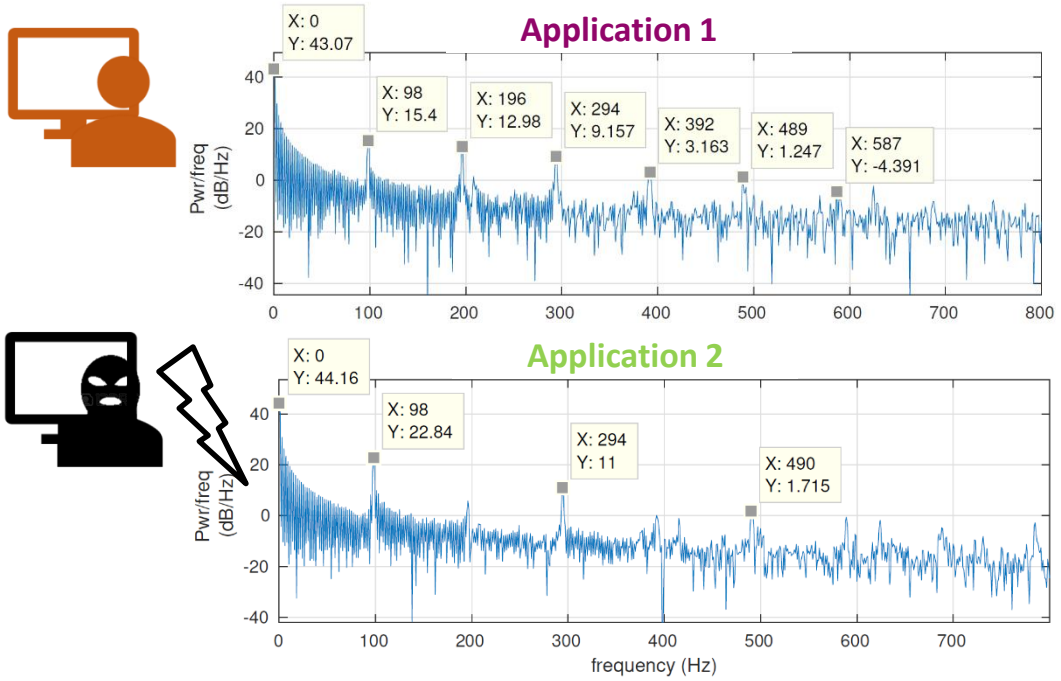
[1] https://github.com/EEESlab/examon

[2] F. Beneventi et al., "Continuous learning of HPC infrastructure models using big data analytics and in-memory processing tools"

DiG: High-Res Power Monitoring [1,2]

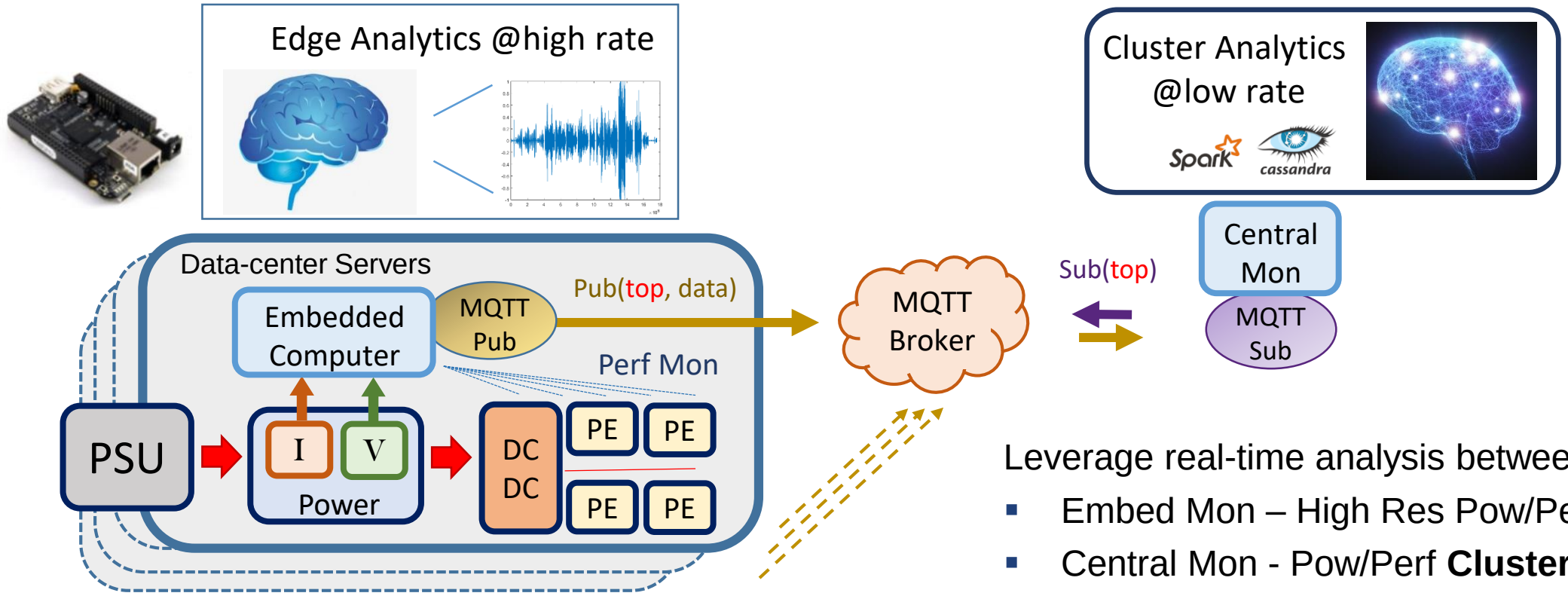
- Fourier on high-resolution power measurements as example of feature extraction technique for time series





[1] Libri et al., 2018, *DiG: Enabling Out-of-Band Scalable High-Resolution Monitoring for Data-Center Analytics, Automation and Control*
[2] Borghesi et al., 2018, "*Online Anomaly Detection in HPC systems*"

Our Envision on Future Fine-Grain Analytics – The DiG Approach [1,2]



Leverage real-time analysis between:

- Embed Mon – High Res Pow/Perf **Edge Analytics**
- Central Mon - Pow/Perf **Cluster Analytics**

[1] Libri et al., 2018, *DiG: Enabling Out-of-Band Scalable High-Resolution Monitoring for Data-Center Analytics, Automation and Control*
[2] Borghesi et al., 2018, "*Online Anomaly Detection in HPC systems*"

Take home messages

- Several challenges on HPC systems require **reliable** and **high-resolution** measurements (precise, fine-grained & synchronized)
- Several SoA methods → **in-band** and **out-of-band** methods which use built-in and custom sensors
- Fine-grain measurements can reveal **precious information** that can be used to profiling applications and system behavior (e.g., detection of anomalies)
- Leverage the huge amount of data of fine-grain measurements between **edge** and **cluster level** analytics

ETH zürich

Thanks for your interest

Contact:

- Antonio Libri, a.libri@iis.ee.ethz.ch

Acknowledge:

- *Andrea Bartolini, Francesco Beneventi, Andrea Borghesi and Luca Benini*

EU H2020 FETHPC project ANTAREX (g.a. 671623)

Backup Slides

DiG: SW Overhead, Scalability & ML Infer

- **SW overhead**: DiG CPU usage of mon daemons < 46 % (soon 0% thanks to co-processor offloading)
- **Synch:** via PTP up to µs (below sampling period 20µs)
- **MQTT Scalability**: tested on 512 nodes of GALILEO (CINECA) → suitable for large-scale systems
- **Real-Time ML Inference** preliminary benchmarks:
  - RT feature extraction via FFT on high-resolution measurements in a time window of 40ms w. around 7% of DiG CPU usage
  - RT ML inference via TF w. Resnet of 16 layers and chan {16;16; 32; 64} respecting real-time constraint of 40ms (FFT time window)