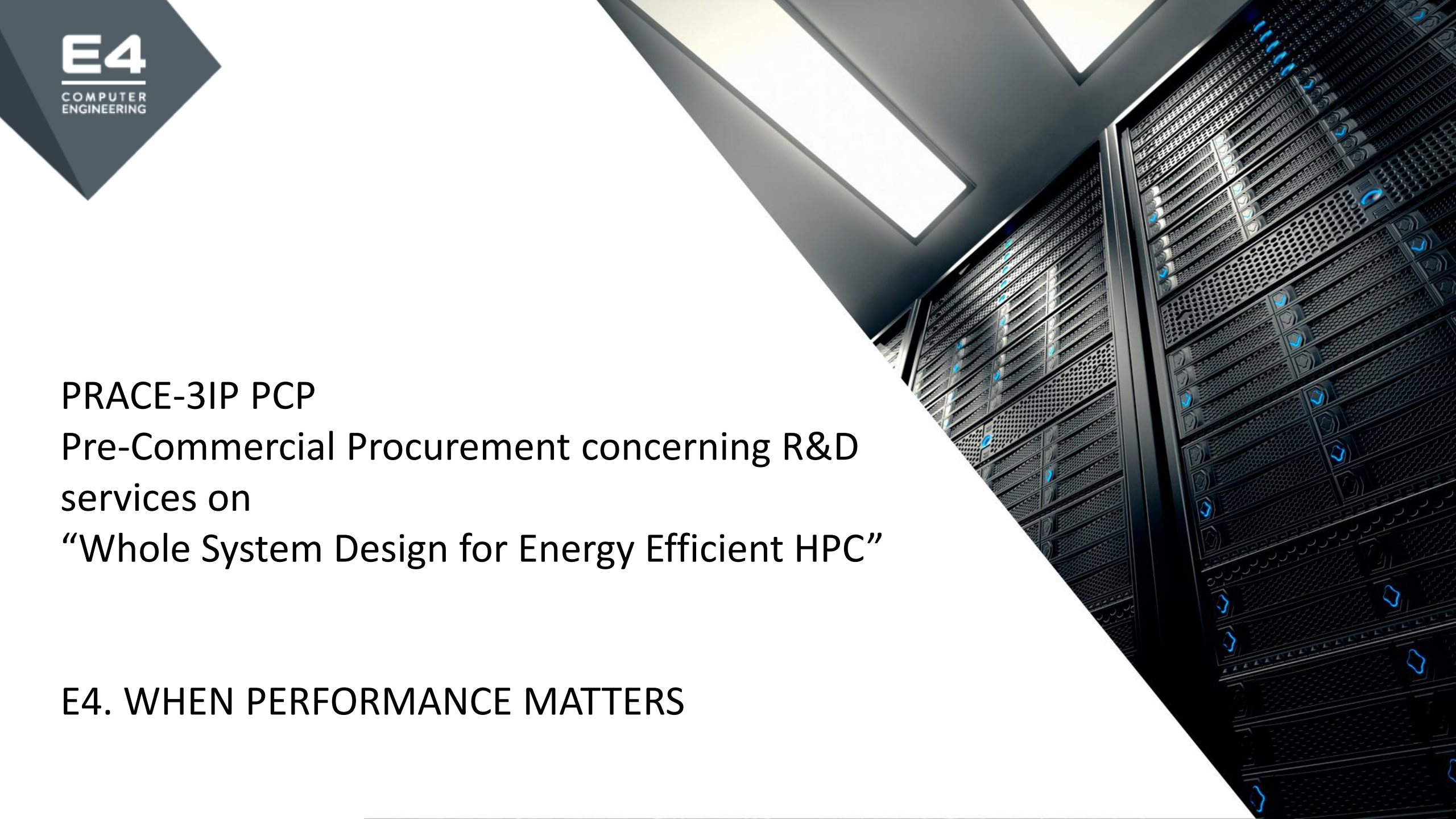


PRACE-3IP PCP
Pre-Commercial Procurement concerning R&D services on
“Whole System Design for Energy Efficient HPC”

E4. WHEN PERFORMANCE MATTERS

Successful R&I in Europe / Workshop on Innovation Procurement

The DNA of E4 is business-oriented innovation, combining technology excellence with enterprise-quality production-oriented systems

E4 Computer Engineering welcomed the opportunity to be engaged with the new procurement model of the PCP because this model gave E4 the opportunity to develop a chain of platforms incrementally showing the benefits of applying innovative technologies specifically developed for the project to off-the-shelf, commercially available platforms

Successful R&I in Europe / Workshop on Innovation Procurement

The continuous exchange of information among the PCP management and E4 has been key for success, because enabled E4 to better understand the goals of the project and develop strategies for meeting or exceeding the objectives of the project itself

Successful R&I in Europe / Workshop on Innovation Procurement

Because of having been selected for delivering a solution for the PRACE-3IP PCP Pre-Commercial Procurement concerning R&D services on “Whole System Design for Energy Efficient HPC”, E4 Computer Engineering

- Develped a new and innovative product
- Improved its competitiveness in the HPC market
- Developed in collaboration with European partners components and technologies that are now available on the market as standard products
- Enriched the European ecosystem of new product specifically targeted to reducing power consumption while not impacting performance

Successful R&I in Europe / Workshop on Innovation Procurement

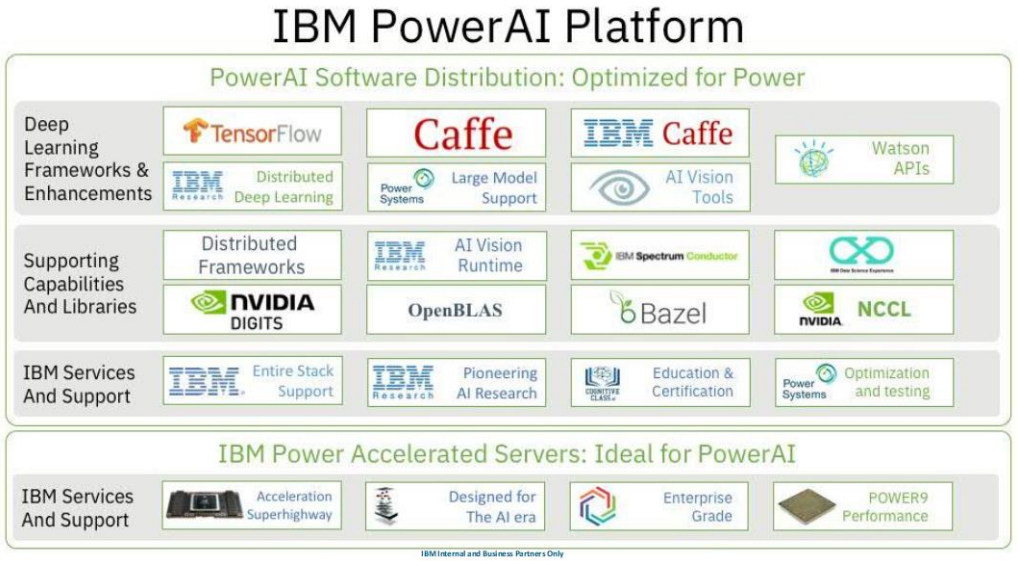
- As a results of the participation in the PCP program, E4 has a new product in its portfolio: OP206EC

- The product has been ranked
  299th in TOP500 (June 2017), 14th in GREEN500 (June 2017)
  440th in TOP500 (November 2017), 18th in GREEN500 (November 2017)

- Not bad….

SUCCESSFUL R&I IN EUROPE / WORKSHOP ON INNOVATION PROCUREMENT

- E4 Computer Engineering has installed and validated IBM PowerAI on the nodes and is now proposing an OpenPower-compliant platform for AI and HPDA



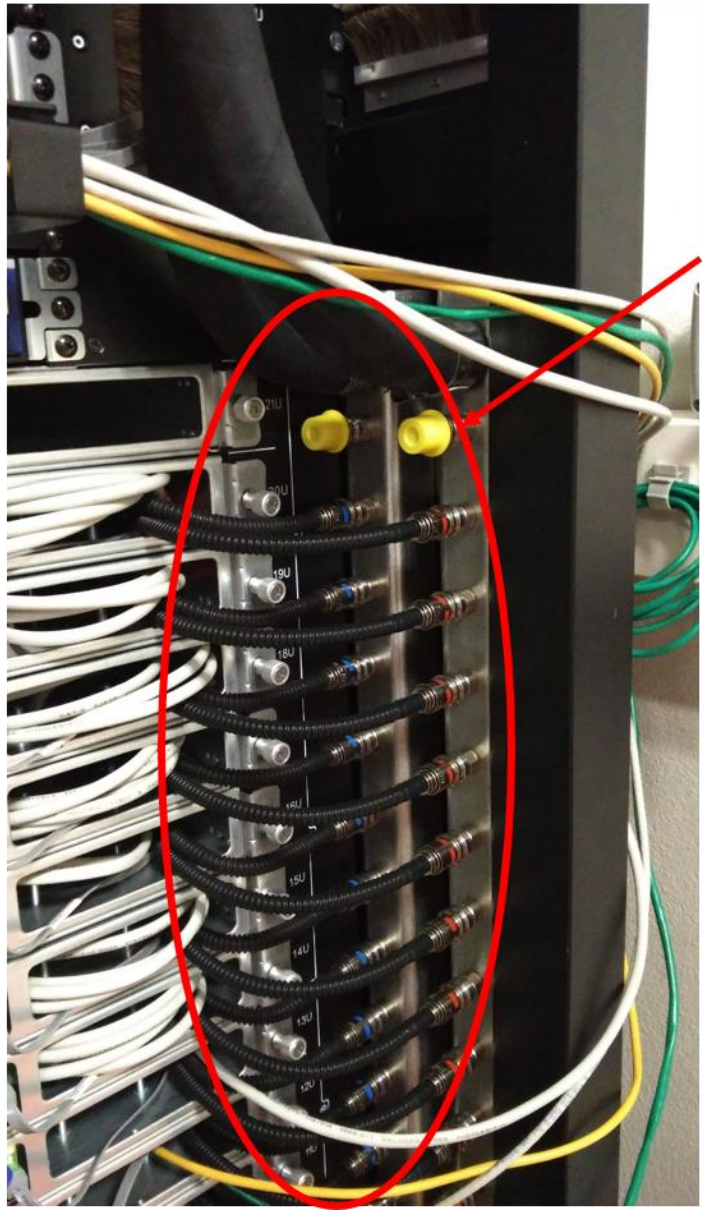
PCP Phase I and Phase II

- E4 Computer Engineering won the Phase I (solution design) and Phase II (prototype) of the PRACE-3IP-PCP tender presenting a project based on
  - ARMv8 SoC accelerated through
  - GPUs and
  - low latency network infiniband.

PCP Phase II

Back

Front



PCP Phase III

- E4 Computer Engineering won the Phase III of the 3IP-PCP tender presenting a project based on
  - IBM POWER8 accelerated through
  - NVIDIA P100 SXM2 GPUs and
  - Low latency network infiniband with
  - Water cooling and
  - Power monitoring and capping.

PCP PHASE III/Energy accounting and profiling

- To deliver a fine-grain and scalable energy monitoring and profiling we will leverage the Power Measurement Interposer (PMI) IP designed and verified in phase-II and the monitoring support already built in the POWER8 and NVIDIA compute system and motherboard.

- The PMI allows to offload the monitoring support to an external low-power SoC (the beaglebone black BBB) designed for internet-of-things IoT devices and it samples through dedicated voltage and current sensors the node power (energy) consumption with sampling frequency of 1KHz.

- These values are then exposed to the software infrastructure through a lightweight and scalable protocol (MQTT) both at coarse granularity (every second) and at the maximum sampling frequency.

PCP PHASE III/Energy accounting and profiling

- To extract a per-component power measurement our system feature a Board Management Controller (BMC) as well as an integrated On Chip Controller (OCC) which is a shadow core and it receives all the sensors information and handles the core power management.

- The OCC and BMC are connected to per component power sensors and can retrieve their power consumption and expose to the user through the Intelligent Platform Management Interface (IPMI) to the user.

PCP PHASE III/Energy accounting and profiling

- The monitoring system developed in phase-II will be extended to interact with the BMC and OCC, accessing per-component power measurements.

- It will provide power information to the accounting and profiler software. In each node all the measured information are available from the BBB and are exposed to the final user and system administrator through both a set of (i)APIs and tools(ii). The APIs will allow to access on demand to the power(energy) consumption at both 1s and 1ms sampling time. These request will be handled by the BBB present in each node.

PCP PHASE III/Energy accounting and profiling

- In addition to these information, an accounting and profiling tool will be integrated as a plug-in.

- The Energy Accounting (EA) plugin already developed in phase-II will allow to feed this information to the job scheduler software and enables energy accounting.

- The Energy Profiling (EP) plugin will instead provide a scalable storage buffer for fine grain energy traces suitable to be correlated in between the different nodes and with architectural events and applications events. Time synchronization in between the different nodes and monitoring devices (BBB) will be ensured by network time protocols which we have evaluated in the phase-II and allows time stamp accuracy in the microsecond range.

D.A.V.I.D.E.
SUPERCOMPUTER
(**D**evelopment of an
Added
Value
Infrastructure
Designed in
Europe)

D.A.V.I.D.E.
SUPERCOMPUTER
(**D**evelopment of an
Added
Value
Infrastructure
Designed in
Europe)

First & Only OpenPOWER PFLOPS-class Cluster in TOP500 and GREEN500
299th in TOP500 (June 2017)
14th in GREEN500 (June 2017)
440th in TOP500 (November 2017)
18th in GREEN500 (November 2017)

D.A.V.I.D.E. SUPERCOMPUTER

(Development of an Added Value Infrastructure Designed in Europe)

OCP form-factor compute node based on IBM Minsky

4x NVIDIA Tesla P100 HSMX2

2 x IBM POWER8 with NVLink

2xIB EDR

BusBar

E4/Università di Bologna POWER MANAGEMENT COMPONENTS

Liquid cooling

Design of an Energy Aware peta-flops Class High Performance Cluster Based on Power Architecture

Wissam Abu Ahmad[1], Andrea Bartolini[2,3], Francesco Beneventi[2], Luca Benini[2,3], Andrea Borghesi[2], Marco Cicala[1], Privato Forestieri[1], Cosimo Gianfreda[1], Daniele Gregori[1], Antonio Libri[3], Filippo Spiga[4,5], Simone Tinti[1]

[1] E4 Computer Engineering, Scandiano (RE), Italy.

[2] DISI, DEI, University of Bologna, Bologna, Italy.

[3] Department of Information Technology and Electrical Engineering, ETH, Zurich, Switzerland.

[4] Quantum ESPRESSO Foundation, UK

[5] University of Cambridge, Cambridge, UK

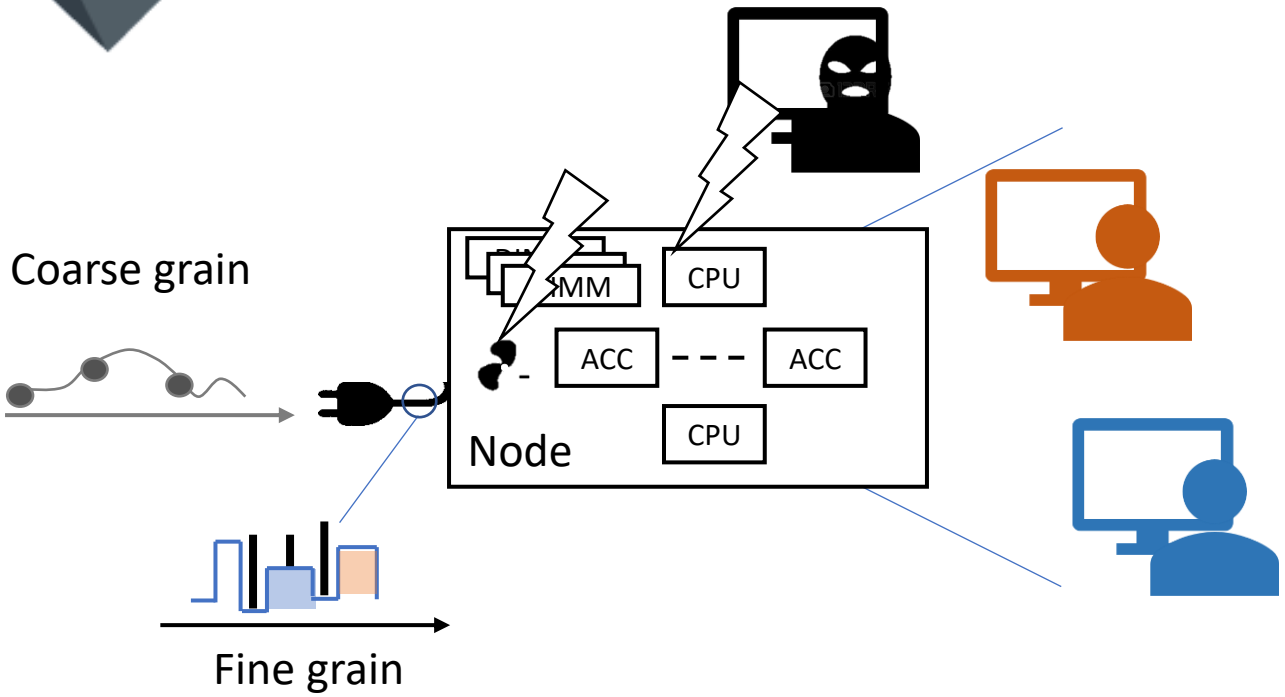
wissam.abuahmad@e4company.com, a.bartolini@unibo.it, francesco.beneventi@unibo.it, luca.benini@unibo.it, andrea.borghesi@unibo.it, marco.cicala@e4company.com, tino.forestieri@e4company.com, cosimo.gianfreda@e4company.com, daniele.gregori@e4company.com, a.libri@iis.ee.ethz.ch, filippo.spiga@quantum-espresso.org, simone.tinti@e4company.com

***Abstract*—In this paper we present D.A.V.I.D.E. (Development for an Added Value Infrastructure Designed in Europe), an innovative and energy efficient High Performance Computing cluster designed by E4 Computer Engineering for PRACE (Partnership for Advanced Computing in Europe). D.A.V.I.D.E. is built using best-in-class components (IBM's POWER8-NVLink CPUs, NVIDIA TESLA P100 GPUs, Mellanox InfiniBand EDR 100 Gb/s networking) plus custom hardware and an innovative system middleware software.**

has caused an increment of the total power consumption. This reached a practical limit with Tianhe-2 (the former most powerful supercomputer, 1st from 06/2013 to 11/2015 Top500 lists), with 17.8 MW of IT power consumption for 33.8 PFlops. The current most powerful supercomputer TaihuLight reaches 93 PFlops with a power envelope of only 15.4 MW. This was possible thanks to an energy efficiency increment of 3x w.r.t. Tianhe-2. The fact that the improvement in energy

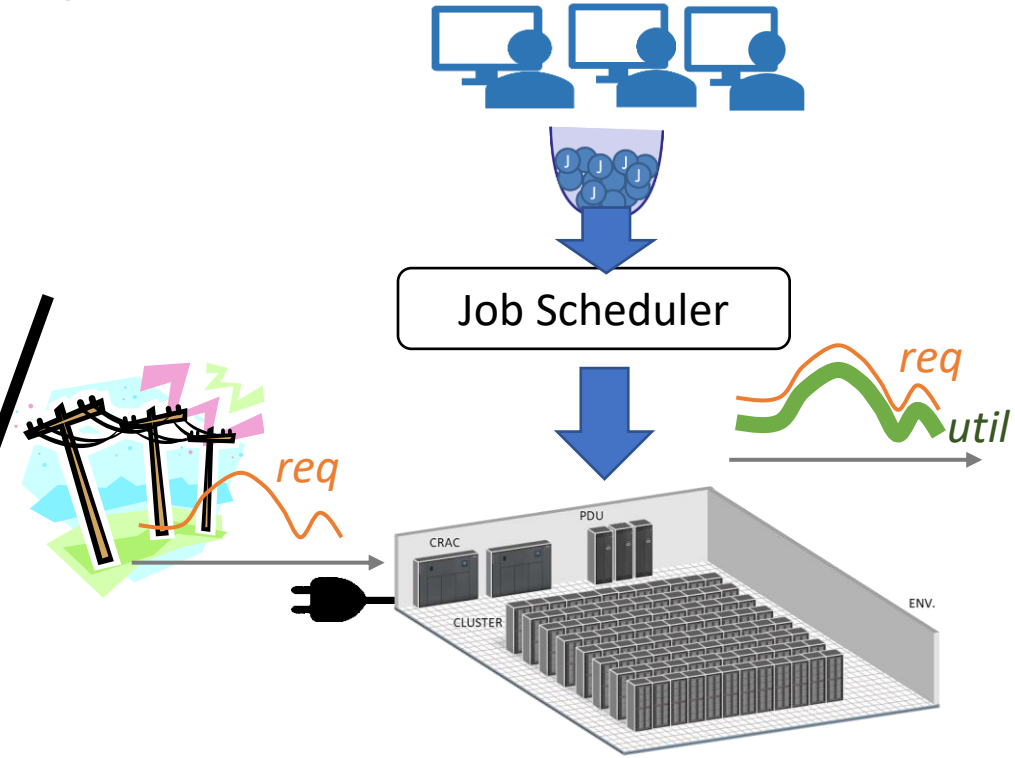
Submitted to IEEE Conference

Target Use Cases



Fine Grain Power and Performance Measurements:

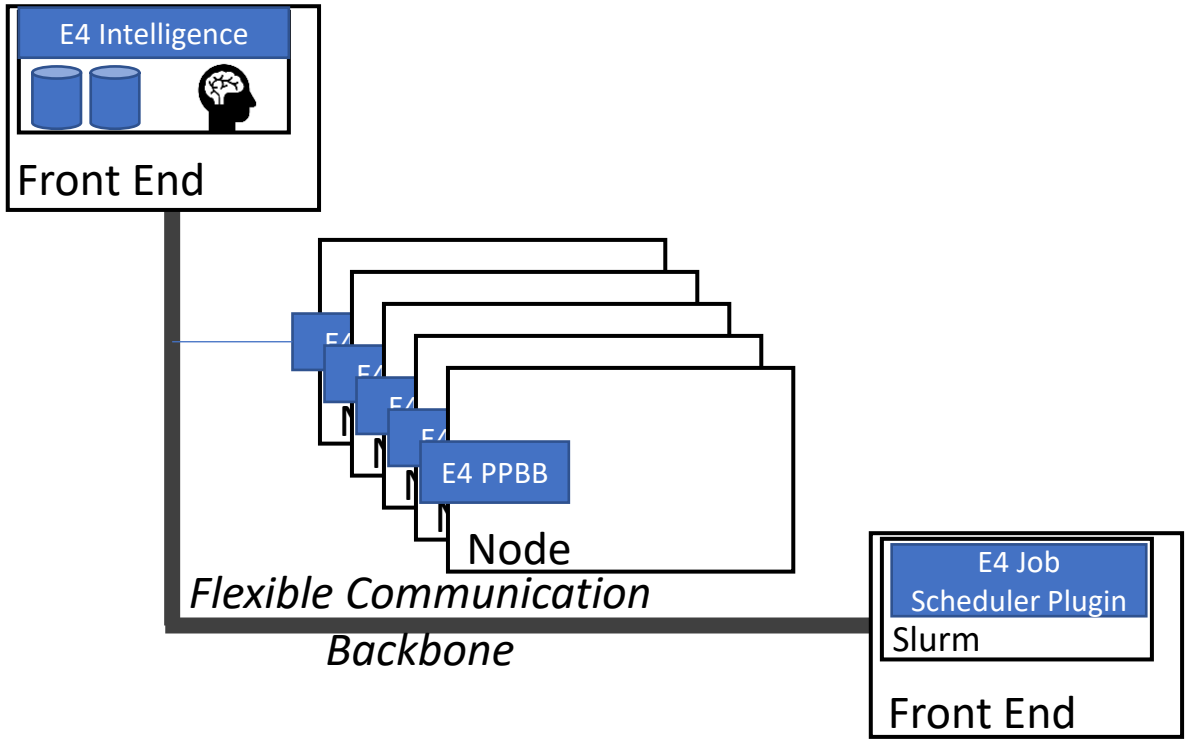
- Verify and classify node performance
  - In spec / out of spec behavior
  - Aging and wareout
- Predictive maintenance
- Per user - Energy / Performance – accounting



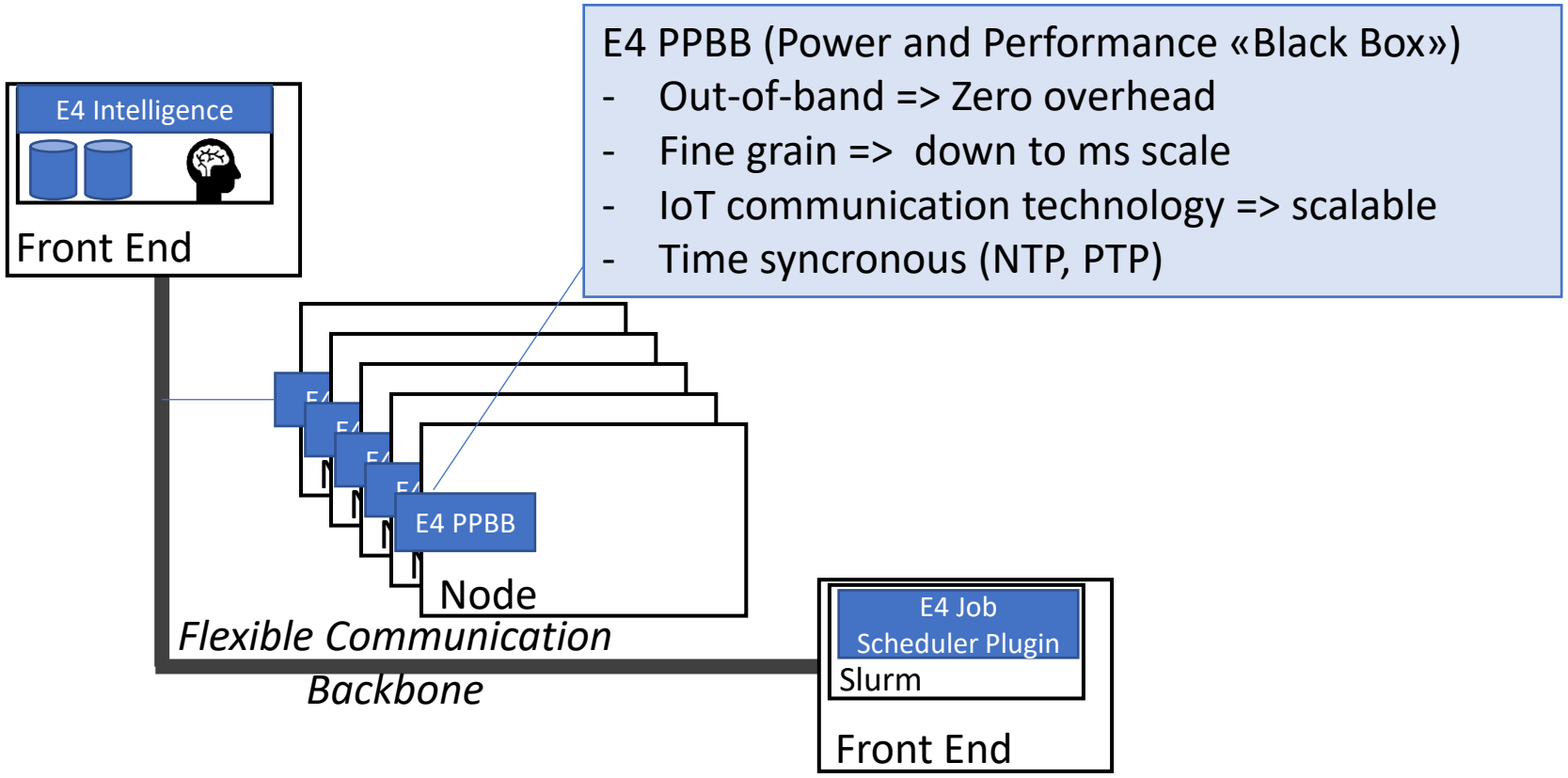
System Power Capping

- New Installation, Grid SLA, Power Shortage, Environmental disaster (i.e. Japan, Huricane,. .)
- Ensure operating power below a maximum power consumption level

E4 Solution Key IPs (co-designed with University of Bologna)



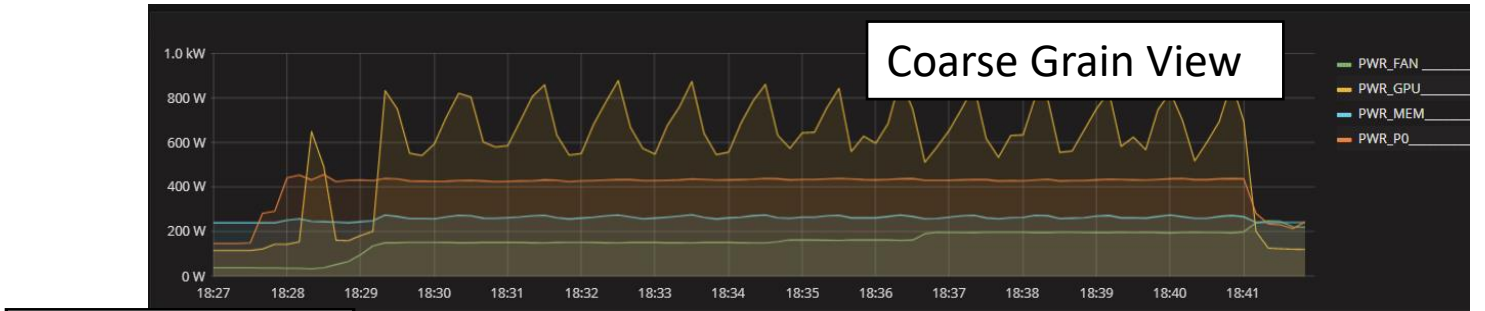
E4 Solution Key IPs (co-designed with University of Bologna)



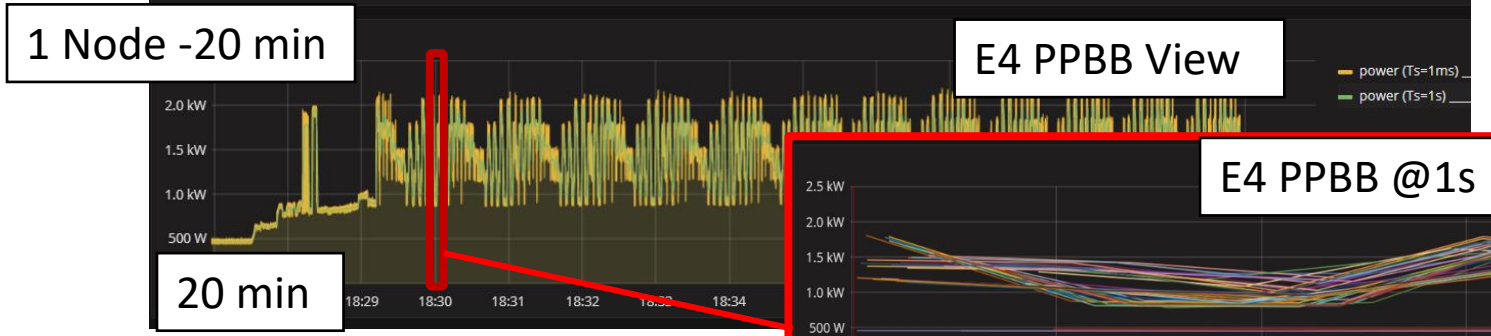
E4 Solution Key IPs #1 (co-designed with University of Bologna)

E4 PPBB

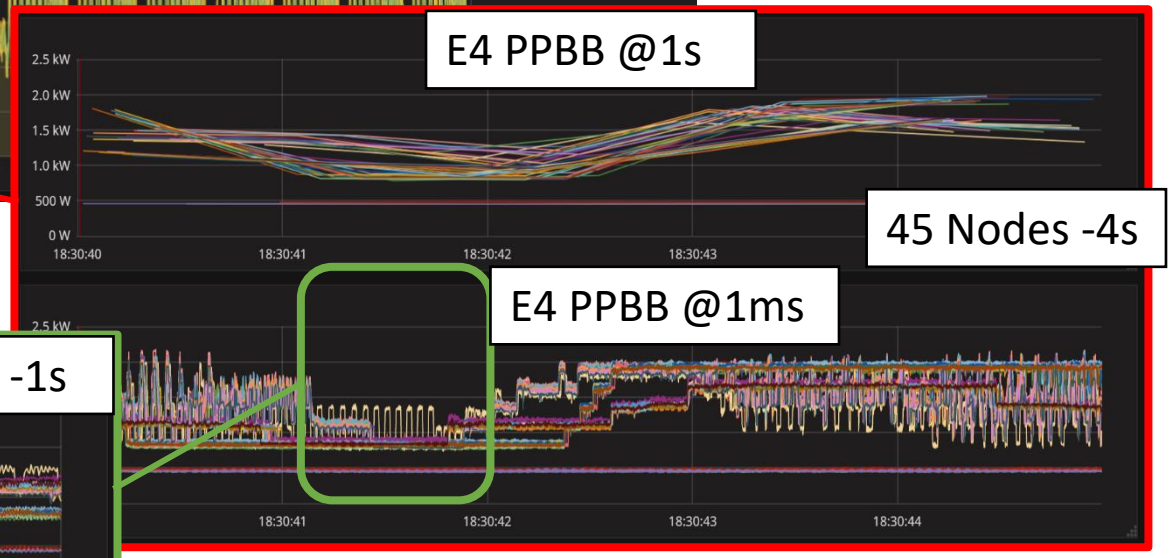
Node

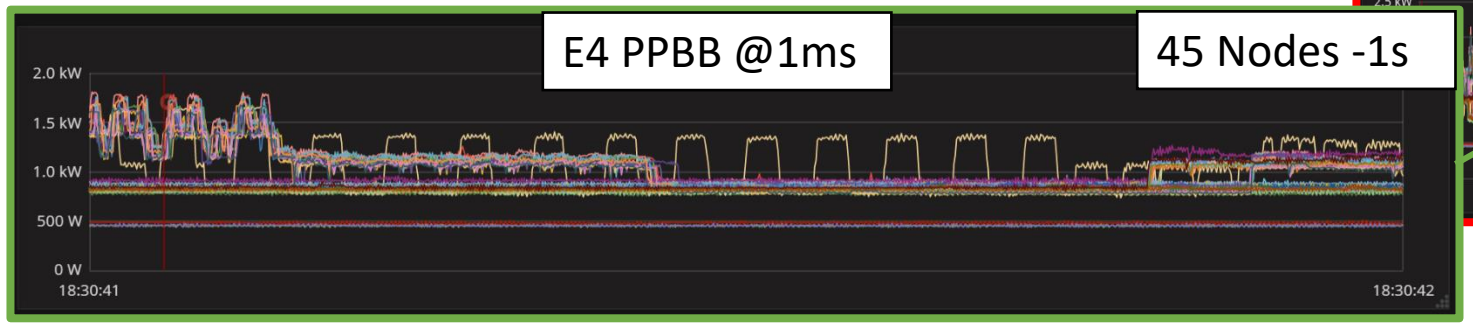


1 Node -20 min



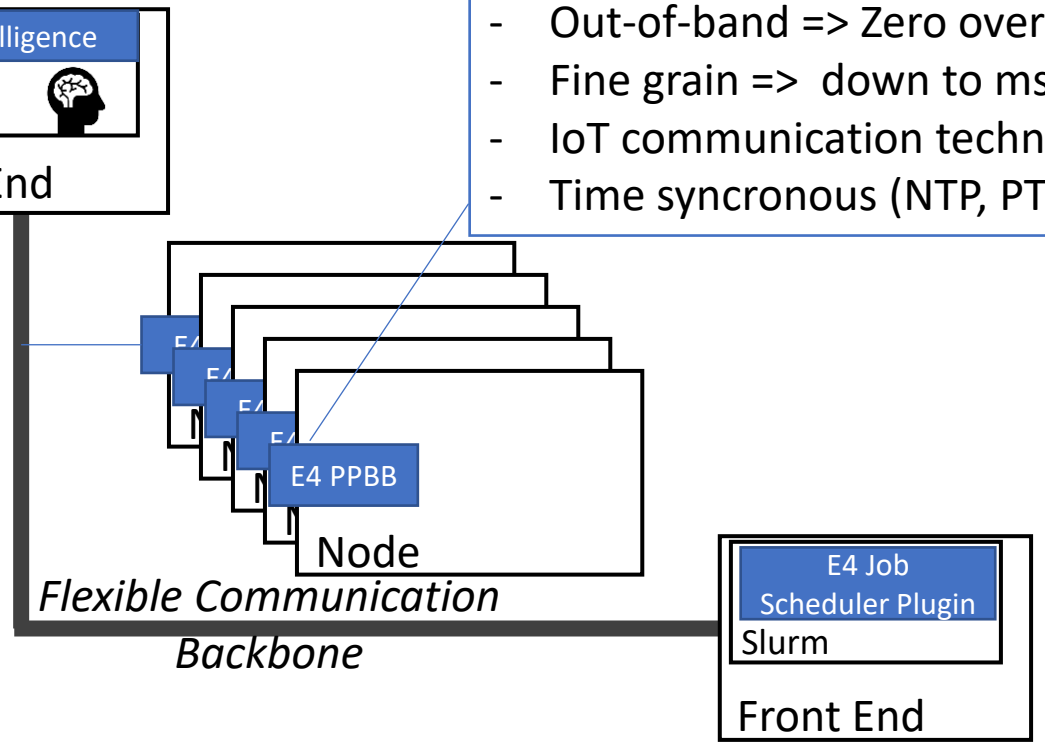
20 min





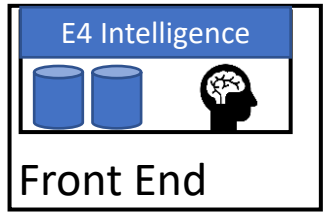
E4 Solution Key IPs (co-designed with University of Bologna)

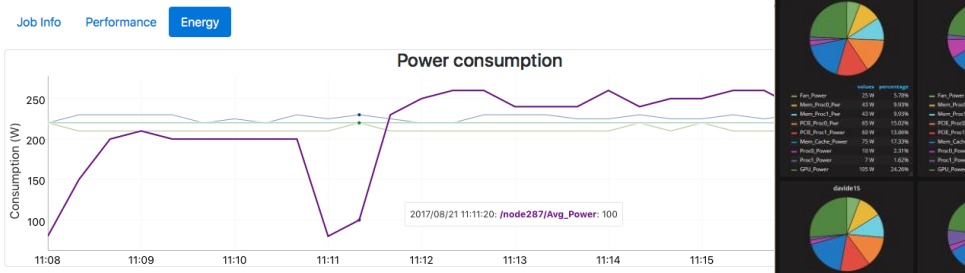
E4 PPBB (Power and Performance «Black Box»)
- Out-of-band => Zero overhead
- Fine grain => down to ms scale
- IoT communication technology => scalable
- Time syncronous (NTP, PTP)

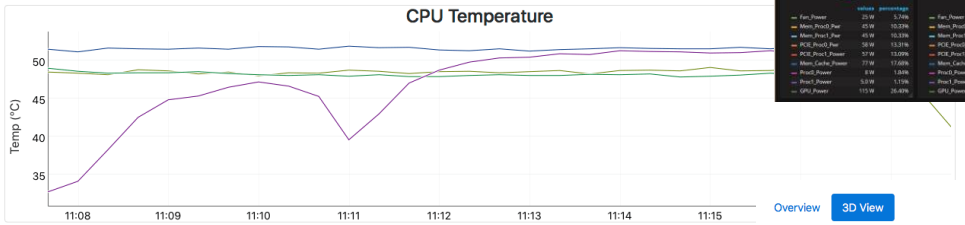


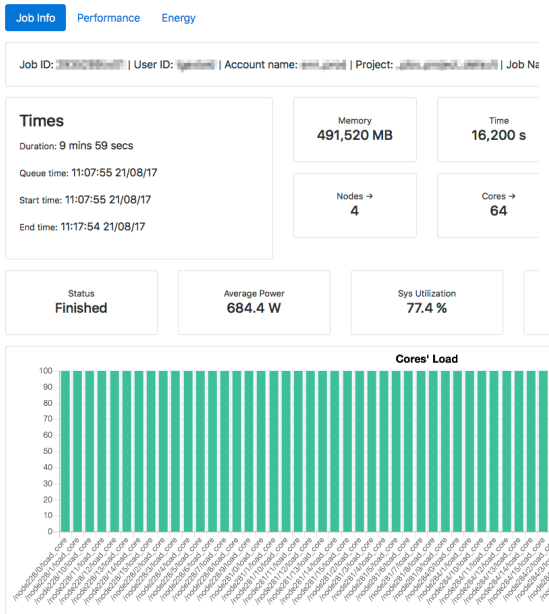
E4 Intelligence
- Capable of aggregating Job, Power, Performance information in real-time and at fine-granularity
- Based on opensource Big Data SW
- Store, Process, Visualize and Analyse Monitored Data
  - Historical Buffer (i.e. 2Weeks)
  - Per Job Perpetual
- Web frontend

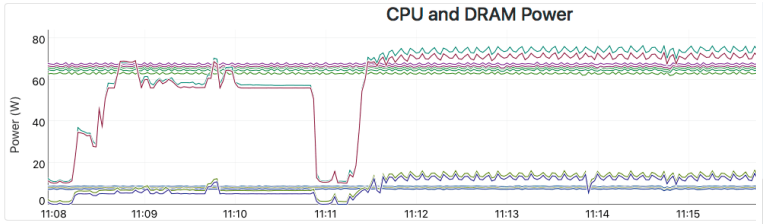
E4 Solution Key IPs #2 (co-designed with University of Bologna)



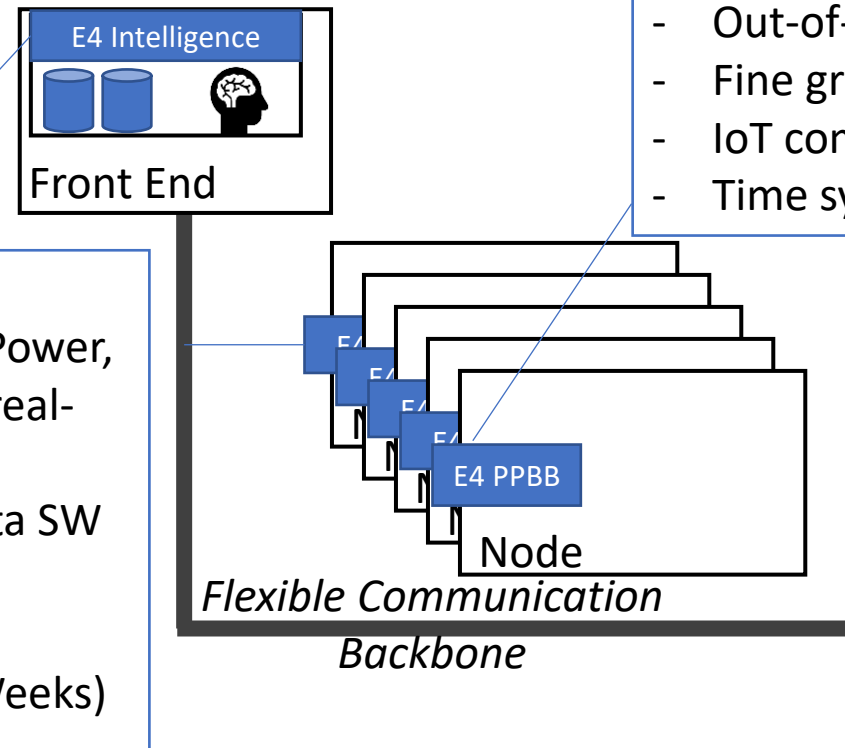








E4 Solution Key IPs (co-designed with University of Bologna)

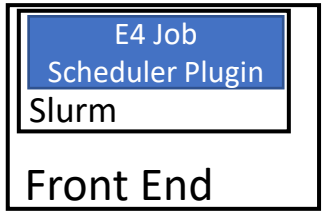


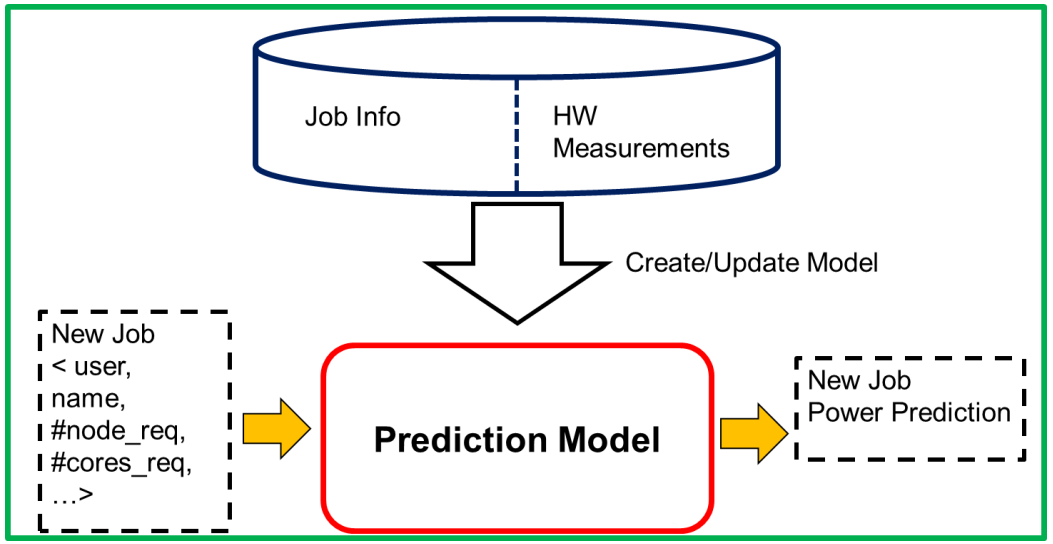
E4 PPBB (Power and Performance «Black Box»)
- Out-of-band => Zero overhead
- Fine grain => down to ms scale
- IoT communication technology => scalable
- Time syncronous (NTP, PTP)

E4 Intelligence
- Capable of aggregating Job, Power, Performance information in real-time and at fine-granularity
- Based on opensource Big Data SW
- Store, Process, Visualize and Analyse Monitored Data
  - Historical Buffer (i.e. 2Weeks)
  - Per Job Perpetual
- Web frontend

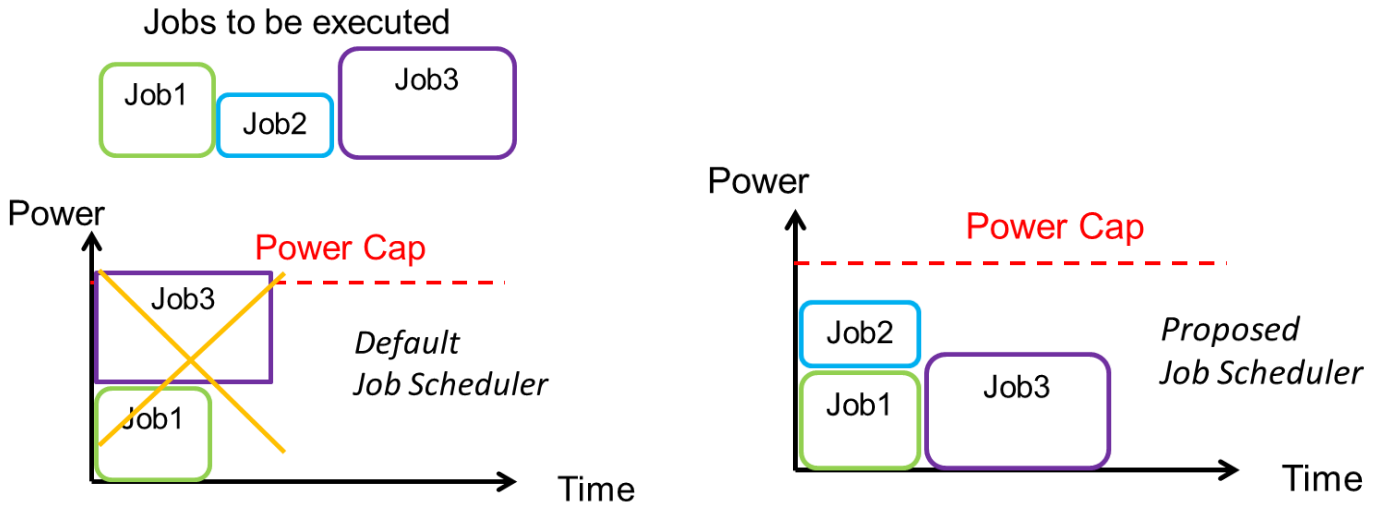
E4 Job Scheduler Plugin
- Monitors Job submission and scheduling
- Estimate Job Power Consumption
- Control Power Consumption

E4 Solution Key IPs #3 (co-designed with University of Bologna)





1. Machine Learning models to predict the power consumption of HPC applications
2. Slurm Custom Extensions to schedule jobs based on their power
3. Run-time monitoring and power management
   - Frequency scaling/RAPL-like mechanism



Successful R&I in Europe / Workshop on Innovation Procurement

PCP-Participation - Advantages for a Company like E4

- Cooperation of E4 with the University of Bologna
- Delivery of solution design, prototye and testing separated into three PCP-Phases
- Keeping IP rights
- Commercialisation of the project results
- European impact

Thank you very much for your attention!

Fabrizio Magugliani
Strategic Planning and Business Development
E4 Computer Engineering SpA, Italy