



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



Perspectives on achieving Exascale: View from Europe

Jesús Labarta

BSC

ModSim 2017

Seattle, August 11th 2017

What does it mean to achieve exascale?

- Exa what ?
 - Scale definition ?
 - € ?
 - Timeline?
 - vision ?

- View from Europe
 - EU ?
 - Personal ?

Disclaimer

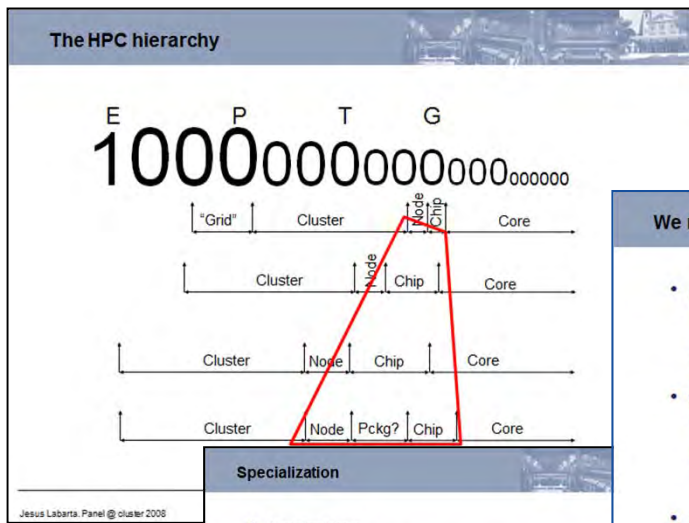
Limited understanding on some of the topics

Yes, I am European → one view from Europe

Personal (controversial?) opinions

I have been here before

- Will the First Exascale Machine be a Commodity Cluster, or Something Else? (Panel @ CLUSTER 2008 (Tsukuba))
 - Nice to look at slides again ☺
 - Tempted to reuse them !!!



Specialization

- Specialized blocks
 - Efficiency for target application, unused by other applications
 - Has always been there: ie: Functional units, SIMDs,...
 - But at coarse granularity: explosion of potential functions.
 - Will they be used? What is a general purpose special processor
- Specialized device lifecycle
 - Potential to be first in achieving some "impressive" result.
 - if successful → push towards other areas and general purpose. Good
 - ie. GPUs, BGL, Cell
 - if successful → general purpose evolves (fights, absorbs, adopts) towards it
- Specialized application area
 - What if an area requires apps. of different characteristics?
 - What about others areas?

We need ...

- Asynchronism, data flow
 - MPI and OpenMP too synchronous
 - Collectives/barriers multiply effects of microscopic load imbalance, OS noise,....
 - Need ways to execute asynchronously within process, across processes.
 - Need to work on algorithms!!!
- Deal with the Memory Wall
 - Hw:
 - Technology at rescue? 3D, Si-photonics
 - Architecture/run time
 - Latency tolerance. Back to vectors? Aren't tasks an extension of vectors?
 - Bandwidth: locality aware scheduling. Across distant tasks.
- Malleability
 - Capability to adapt the parallelism structure of the application to available resources
 - Dynamic load balance within large node.
- Coordination between levels
 - Power to the run time ↔ resource management
 - Controlled sharing of resources (difference between power and force)

Exaflops

- Exawhat? Whatflops? 1000000000000000000
- Numbers make me nervous, overwhelmed, computersick,...
- Can I say something about exaflops ... and not speak about power?
- Don't know how the first exaflop computer will be. The second will probably be very different.
- Fortunately it is still very far in the future. We actually have problems for two orders of magnitude less.
- Unfortunately it may arrive before we have time to digest current revolution.
- Let me just express an unstructured set of thoughts that may be related to it.

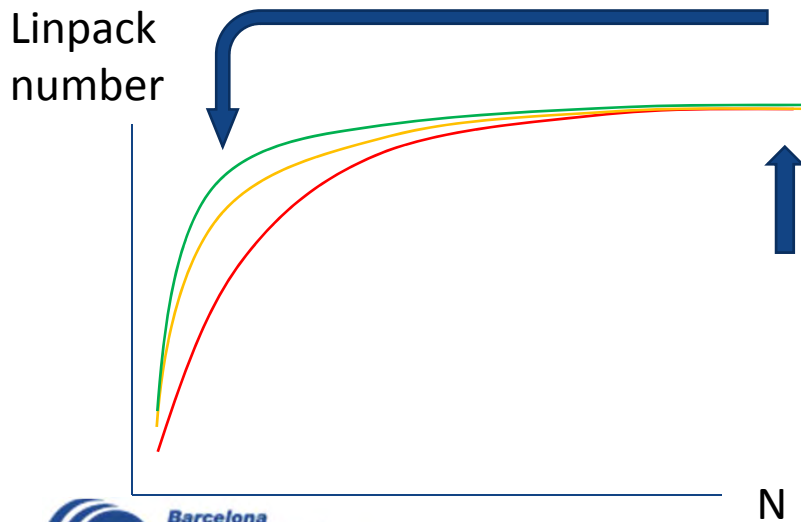
- Hierarchical structure
 - ~10000 nodes x ~10 chips/node x (~100 cores/chip x ~100 GF/core + ~20 cores/chip x fast general purpose + 4 task graph handling engines)
 - <30000 processes x 400 threads/process → 12M threads

How many racks?
How much power?
Can we feed this?
Can we generate this for a large number of apps?
Is this manageable?



Exascale definition ?

- Would certainly accept ECP definition
 - can demonstrate part of the 50x in today's machines !!!!
 - Wide class of apps !!!
- Vindicate Linpack !! (Just a controversial statement !! ??)
 - What is wrong with Linpack?

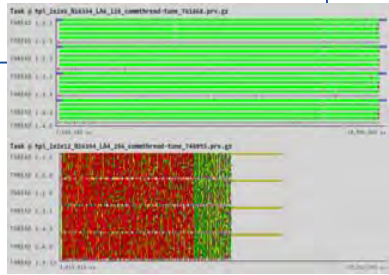


Speed / acceleration ?



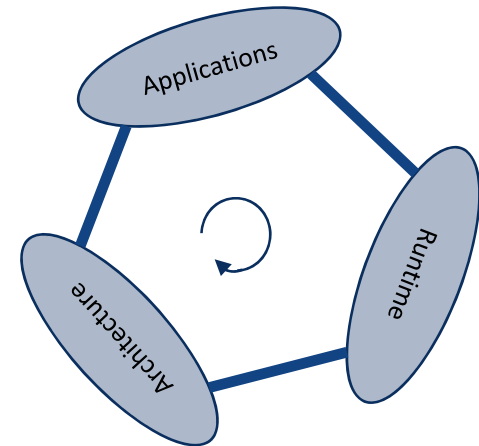
GF/W → GF/J ?

Opportunity for different approaches, different traversals of the computation space



Vision context

- Ensure programmers survive “exascale”
- Maximize take-up → standards
- Co-design
 - Best place to address issues ?
 - Economic, elegant, forward looking place
 - Generic, fundamental characteristics
 - Differentiation between design and dimensioning



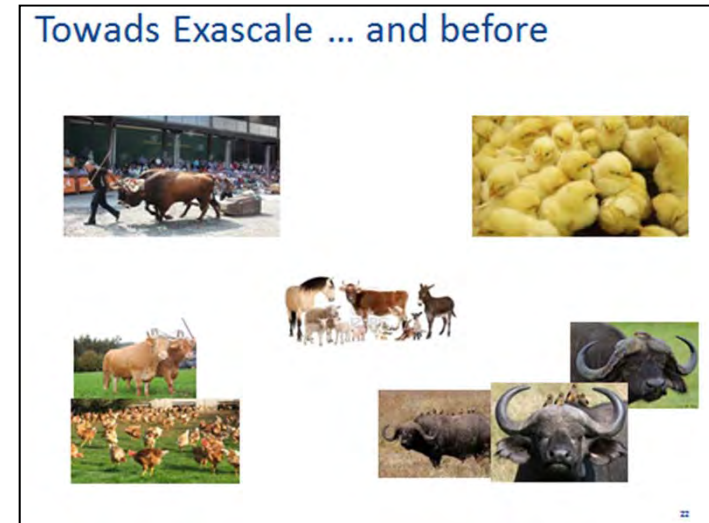
Vision Fundamentals

- Efficiency focus
 - Latency → throughput oriented computing (High % of peak)
 - Lookahead and asynchrony/overlap
 - Malleability and dynamic resource exploitation
 - Performance depends on amount of resources, not so much on other configuration decisions
- Hierarchy
 - Nesting of program structure and resources
 - Top – down
- Integrated concurrency and locality management
 - Same mechanism to specify both issues
 - Try to avoid global synchronization
- Need deep insight on actual system performance
- Incremental programming and productivity is critical

*At all levels: applications, architecture,
system software,...*

A few more thoughts

- “limited” number of control flows
 - few words, lots of work
- High throughput devices
 - Long Vectors
 - Optimize memory throughput
- Accelerator devices (optimized energy efficiency)
 - LONG Vector generic accelerator
 - Specific high semantics “functional units”
- Heterogeneity
 - “Symbiotic”
 - **Homogenized**
- Intelligent runtimes & Runtime Aware Architectures
 - Handle overlap and locality
 - Latency toleration
 - improve B/F
 - Reduce overhead
 - Hierarchical granularity (coarse → fine)



Strong political push for HPC in Europe



“Our goal is for Europe to become one of the top 3 world leaders in high-performance computing by 2020”.

European Commission President
Jean-Claude Juncker (27 October 2015)

EuroHPC



France

Germany

Italy

Luxembourg

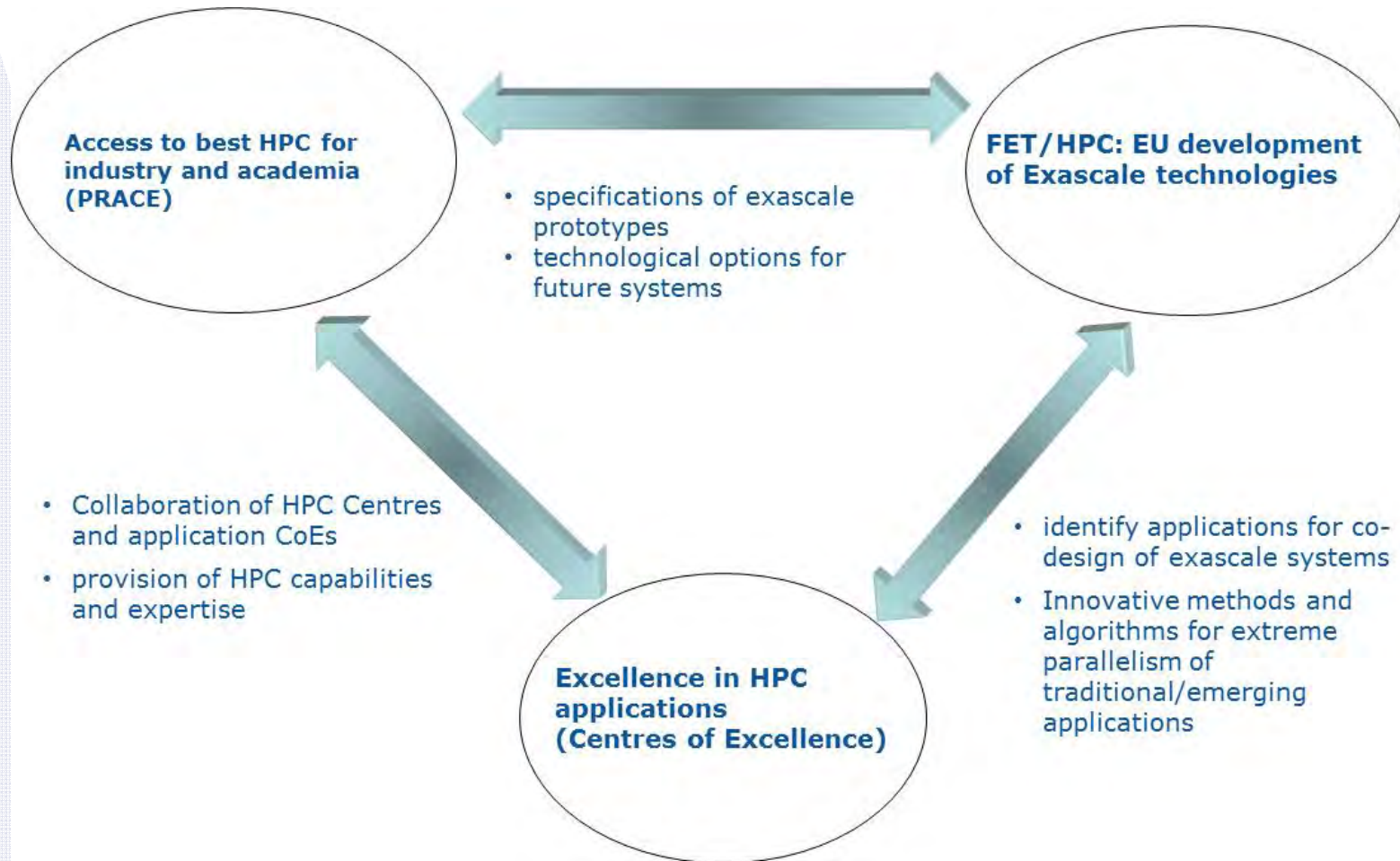
Netherlands

Portugal

Spain

Agree to work towards the establishment of a cooperation framework for acquiring and deploying an integrated exascale supercomputing infrastructure that will be available across the EU for scientific communities as well as public and private partners, no matter where supercomputers are located.

The European HPC policy





Distributed Supercomputing Infrastructure

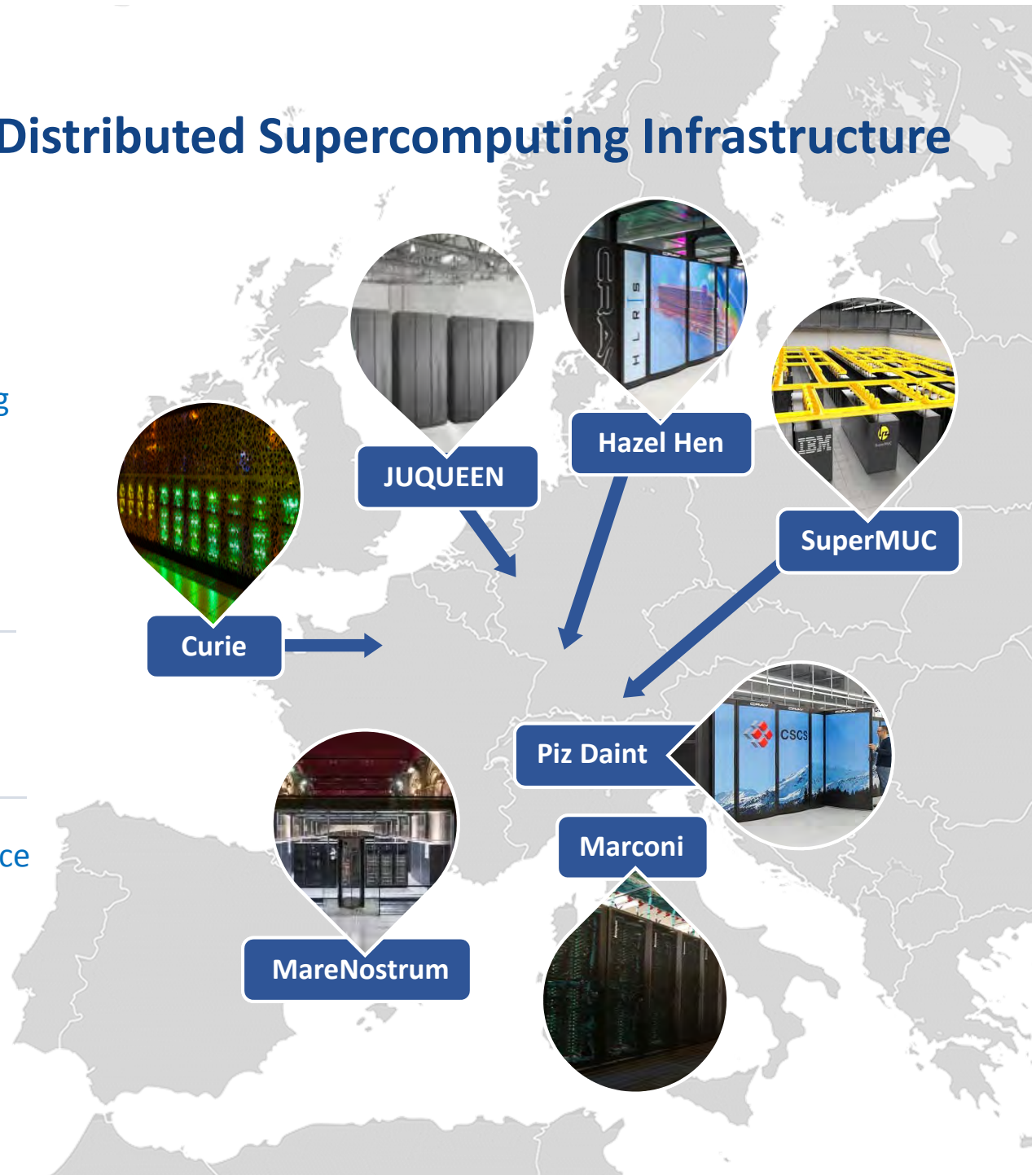
25 member states, including

5 Hosting Members

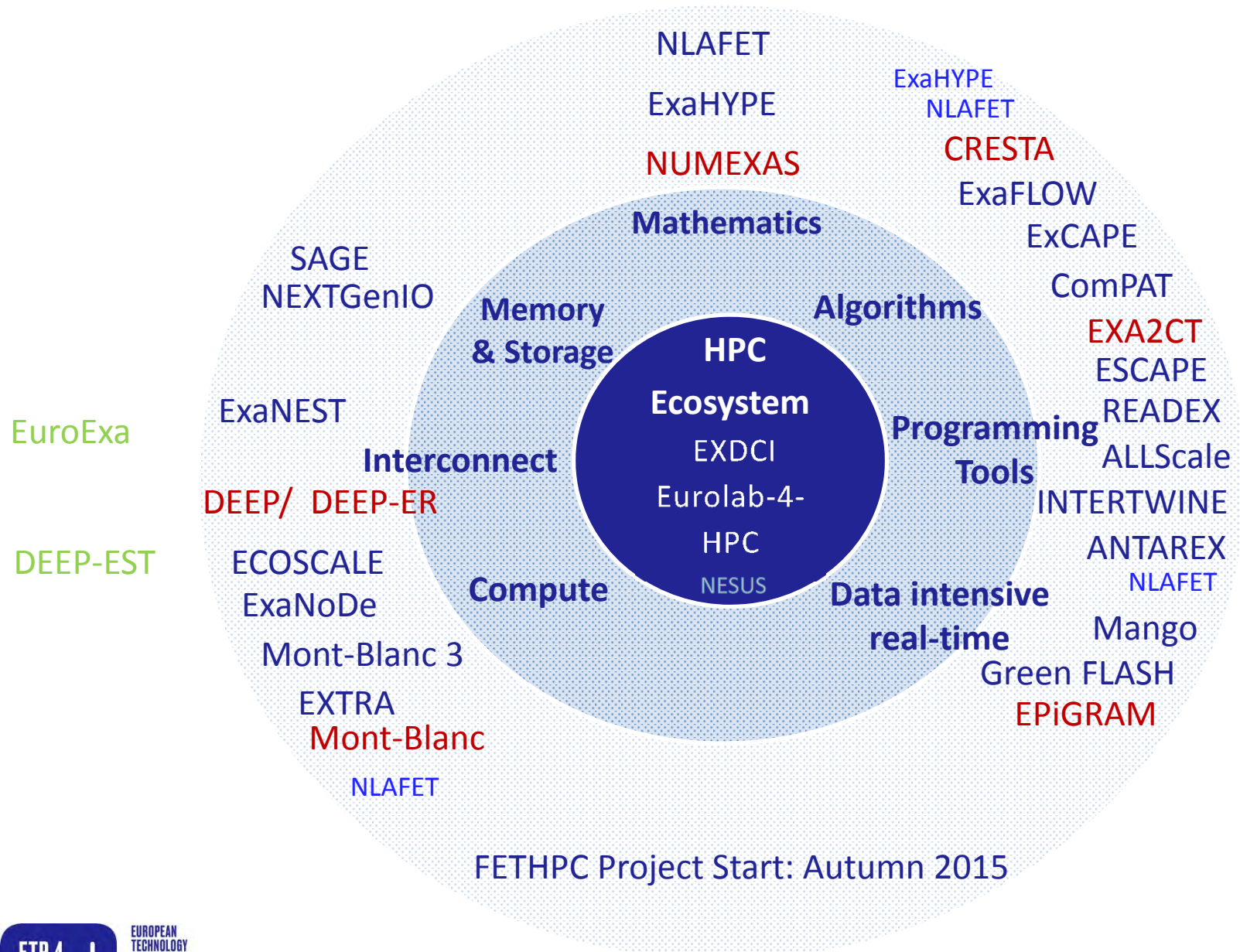
(Switzerland, France, Germany, Italy and Spain)

465 scientific projects enabled

40 Pflop/s of peak performance on 7 world-class systems



THE CURRENT EUROPEAN HPC TECHNOLOGY LANDSCAPE



Centers of Excellence

Materials



NOVEL MATERIALS DISCOVERY

Climate

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER
AND CLIMATE IN EUROPE

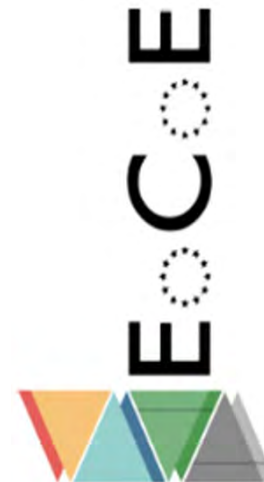


Bio



bioexcel

Energy



Global systems



Coe GSS

Centre of excellence



Performance analysis and programming models

Some apps. / frameworks

Materials

Quantum Espresso
Siesta
aiIDA
Gaussian, VASP,
Gromacs, NAMD
GPAW, CASINO

Yambo
Fleur

Aiida

Climate

OpenIFS

NEMO

OASIS 3 MCT
XIOS
Cylc

ICON
EC-EARTH
MPI_ESM2

XIOS
CYLC

Bio

Gromacs
HADDOCK
CPMD

Chaste
HemeLB
Alya
Palabos
AceMD
OpenSim
Vizualization

Galaxy, Taverna,
OpenPHACTS
and KNIME

Energy

MetalWall
Gysela
Alya

Global systems

Pandora

(Repast HPC)
Self-developed
graph-based
simulation tool
(no name so far)

Hadoop
Spark

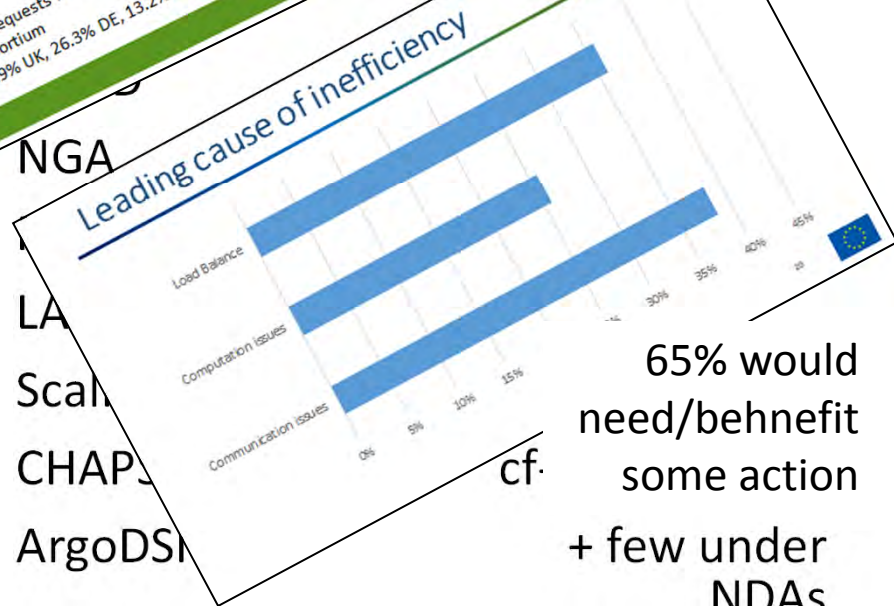
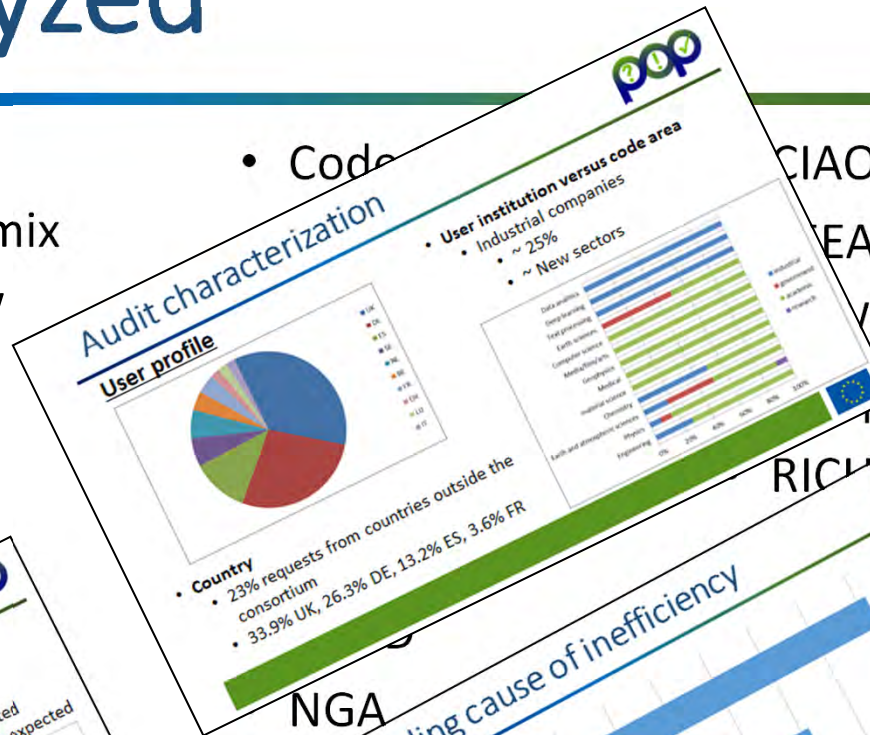
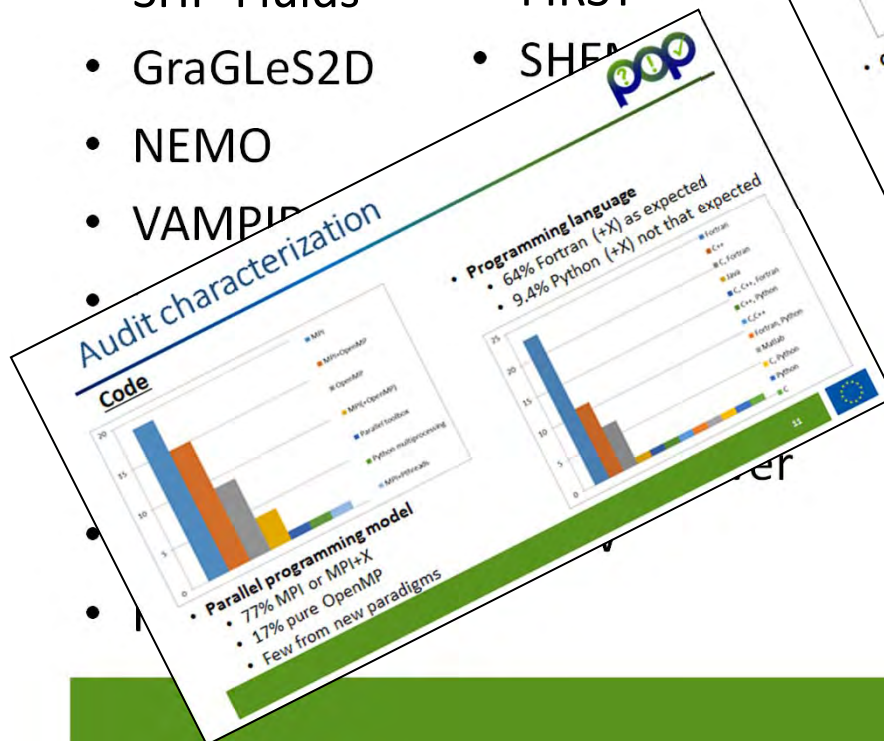
BSC tools, Score-P, Scalasca, Vampir, ...

Performance analysis and programming models

Codes analyzed



- DPM
- Quantum Espresso
- DROPS
- Ateles
- SHP-Fluids
- GraGLEs2D
- NEMO
- VAMPIR
- Baleen
- Modynamix
- ParFlow
- GITM
- BPMF
- FIRST
- SHEM



- Code
- CIAO
- IEA
- have
- plus
- RIC
- NGA
- LA
- Scal
- CHAPS
- ArgoDS



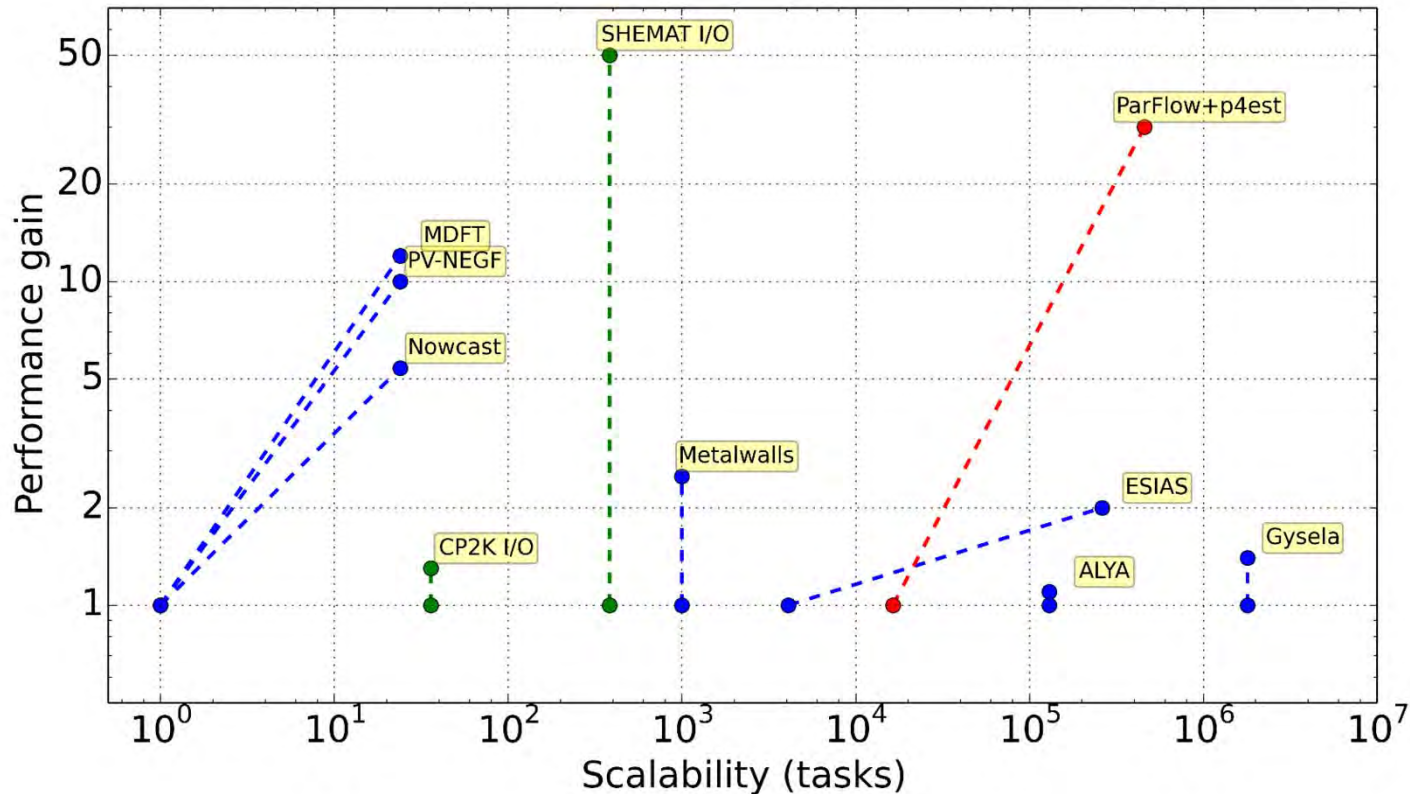
Codes analyzed



- DPM
 - Quantum Espresso
 - DROPS
 - Ateles
 - SHP-Fluids
 - GraGLEs2D
 - NEMO
 - VAMPIRE
 - psOpen
 - GYSELA
 - AIMS
 - OpenNN
 - FDS
 - Baleen
 - Mdynamix
 - ParFlow
 - GITM
 - BPMF
 - FIRST
 - SHEMAT
 - GS2
 - ADF
 - DFTB
 - ICON
 - dwarf2-ellipticsolver
 - EPW
 - Code Saturne
 - ONETEP
 - Ms2
 - SIESTA
 - Oasys GSA
 - SOWFA
 - BAND
 - NGA
 - Fidimag
 - LAMMPS
 - ScalFMM
 - CHAPSIM K.W.
 - ArgoDSM
 - CIAO
 - FFEA
 - k-Wave
 - DSHplus
 - RICH
 - COOLFluid
 - Ondes3D
 - ATK
 - Molcas
 - GBMol_DD
 - Kratos
 - cf-python
- + few under NDAs

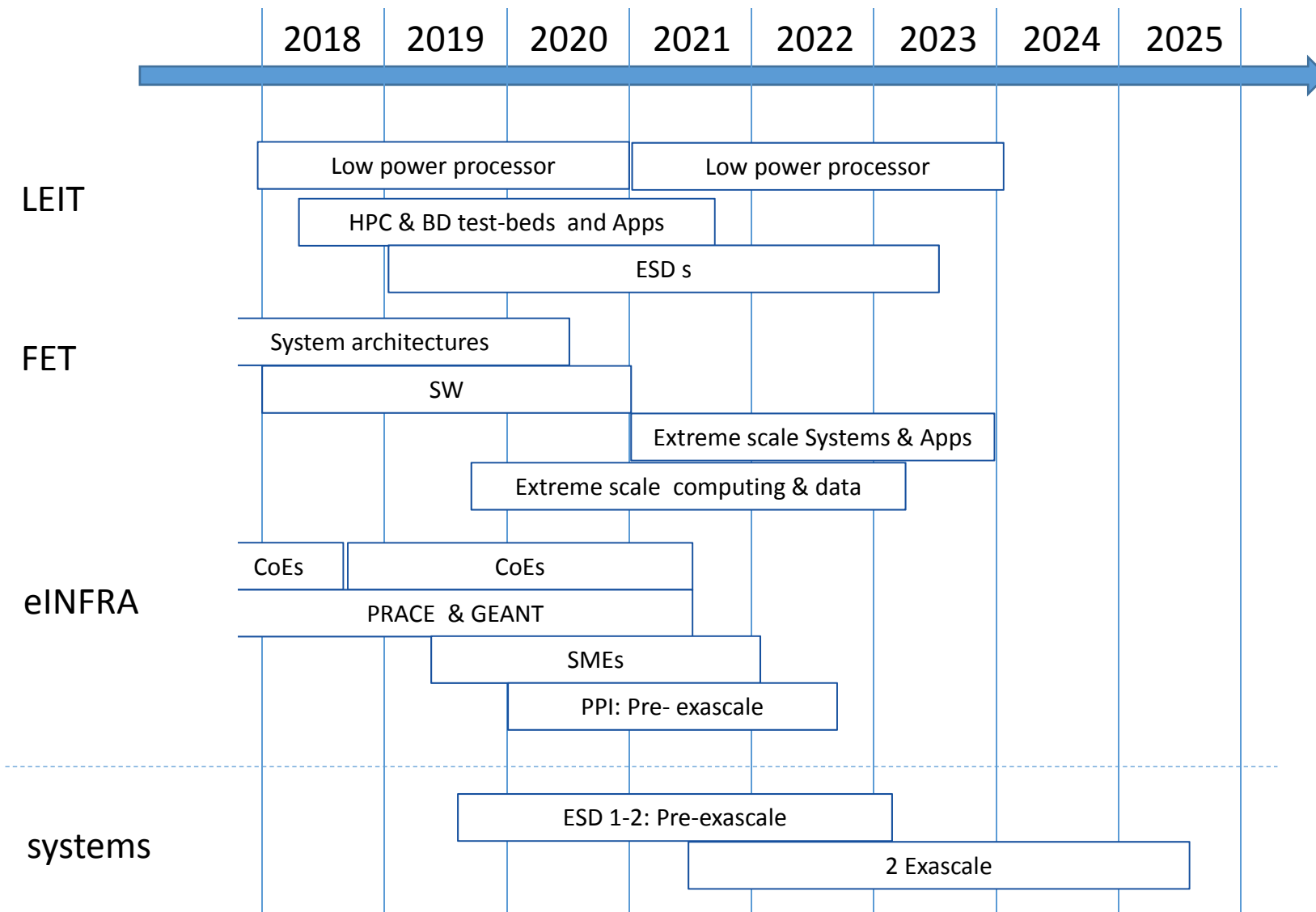


EoCoE application progress



Improvements made to selected applications in the EoCoE consortium: optimisation and parallelisation (blue), parallel I/O library (green), kernel refactoring (red). The performance gain indicates the time-to-solution improvement since the code was first examined - typically at one of the joint EoCoE-POP benchmarking workshops.

HPC in Programs/calls



Disclaimer: Indicative
 Blame Jesus Labarta for inaccuracies

A personal opinion ...

- Activities, experience, interest, ...
 - Some good experiences
 - With involvement of EU research institutions ...and worldwide vendors
- ... but lack of integration
 - Lacking a technical manager/catalyst (a la Paul Messina)
 - National politics
- Vision ?
 - Is there a EU technical vision ?
 - Many individual proposals/research plans
 - SRA: Topics of interest, but not list of actors and responsibilities.
 - Same words everywhere, differences are in the details
 - Like asynchronous task based models
- International cooperation
 - Seems to be difficult at institutional level ?

Some controversial pils ?

(Impression I miss something/think different from the world)

(psychologists are expensive)

(do not take without doctor supervision)

- Order before overhead
 - Why SO MUCH worry about individual runtime operation overhead ?
 - Exascale != atto granularity ?
 - Avoid stalling the work instantiation engine !!!
- Where is the UNDO button?
 - Refactorings done in the past for the sake of optimizing performance that would better not have been done (mallocs, nonblocking calls,...)
- Important to provide “Hope for Lazy programmers “
 - Fighting Amdahl’s law == lookahead + nesting
- The 80-20 rule for Energy efficiency
 - Isn’t it mostly about technology ?
- (Is there really an I/O bottleneck?)
 - Of couse systems I/O often under dimensioned ☹, but ...
 - Human brain capability vs human I/O bandwidth
 - Examples of Optimization of MPI apps for I/O vs port on top of distributed KVS

What does it mean to achieve exascale?

- The challenge

“Let programmers survive the exascale”
- The revolution: a change in the mentality of programmers

“From the latency to the throughput age”



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



Thank you

Jesus.labarta@bsc.es

07/07/2017