New HPC architectures landscape and impact on code developments

Carlo Cavazzoni, Cineca

H2020 EUROPEAN CENTRE OF EXCELLENCE _ EU GRANT # 676598

Enabling Exascale Transition

- GOAL: “modernize” community codes and make them ready to exploit at best future exascale system for material science simulations (MAX Obj1.4)
- CHALLENGE: there is not yet a solution that fits all needs, and this is in common in all computational science domains
- STRATEGY: pragmatic approach based on building knowledge about exascale related problems and running proof of concepts to field test solutions, and finally deriving best practices that can consolidate in real solutions for the full applications and making their way through the public code release.
- OUTCOME: New code versions with development validated, libraries and module publicly available beyond MAX, extensive dissemination activities .
- IMPACT: modern codes, exploitation of today HPC systems, other applications fields as well as technology providers.

Changes in the road-map to Exa

Intel's Data Center Group GM Trish Damkroger describing the company's exascale strategy and other topics they are talking about at the SC17 conference, she offhandedly mentioned that the Knights Hill product is dead. More specifically she said that the chip will be replaced in favor of "a new platform and new microarchitecture specifically designed for exascale."

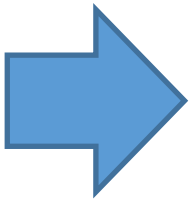
Exascale How serious the situation is?

Peak Performance

10^18 Flops — Moore law

FPU Performance

10^9 Flops — Dennard law



Number of FPUs

10^9

10^5 FPUs in 10^4 servers

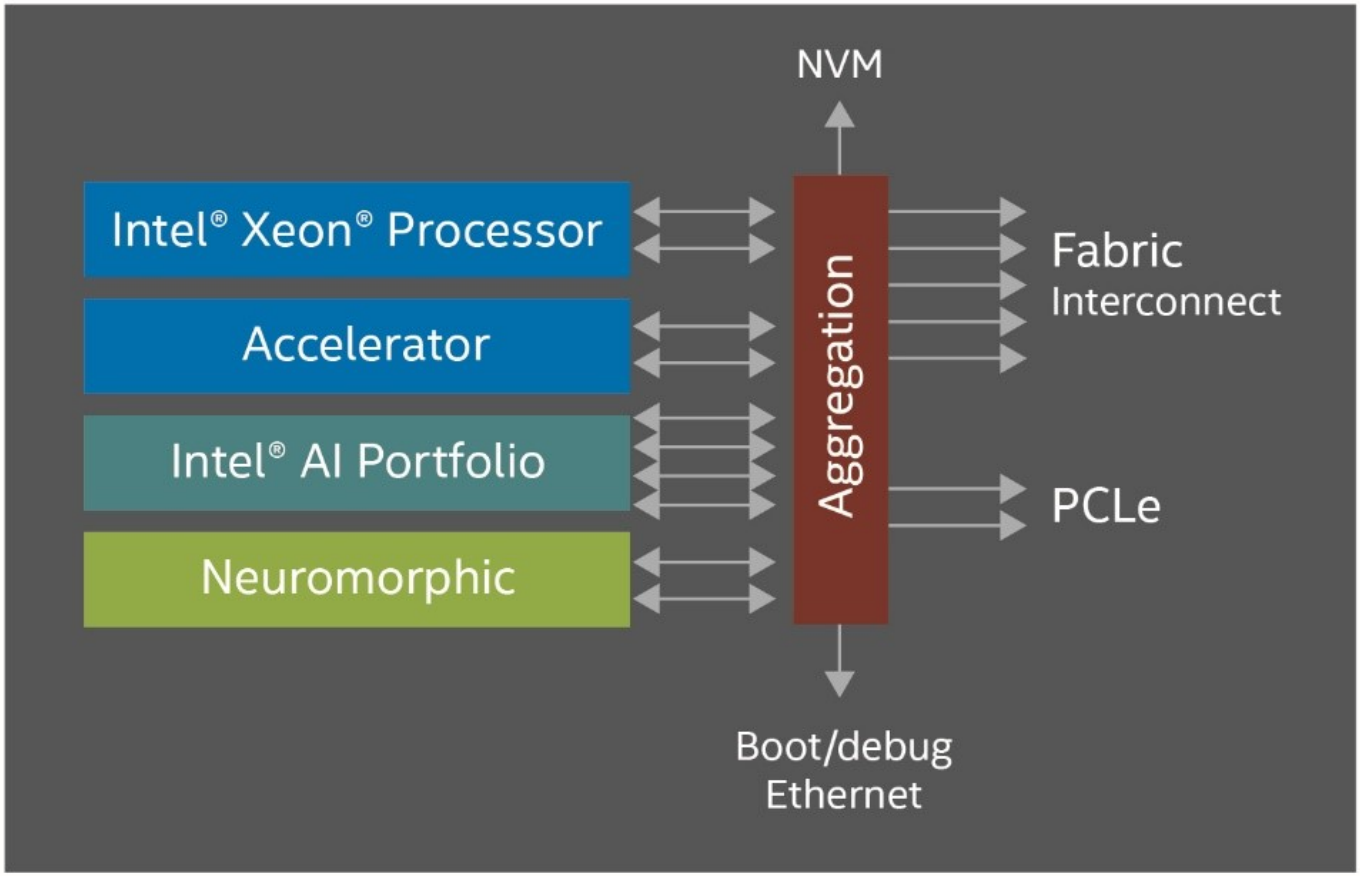
10^4 FPUs in 10^5 servers

Working hypothesis

Exascale Architectures

Heterogeneus

Exascale “node”, according to Intel

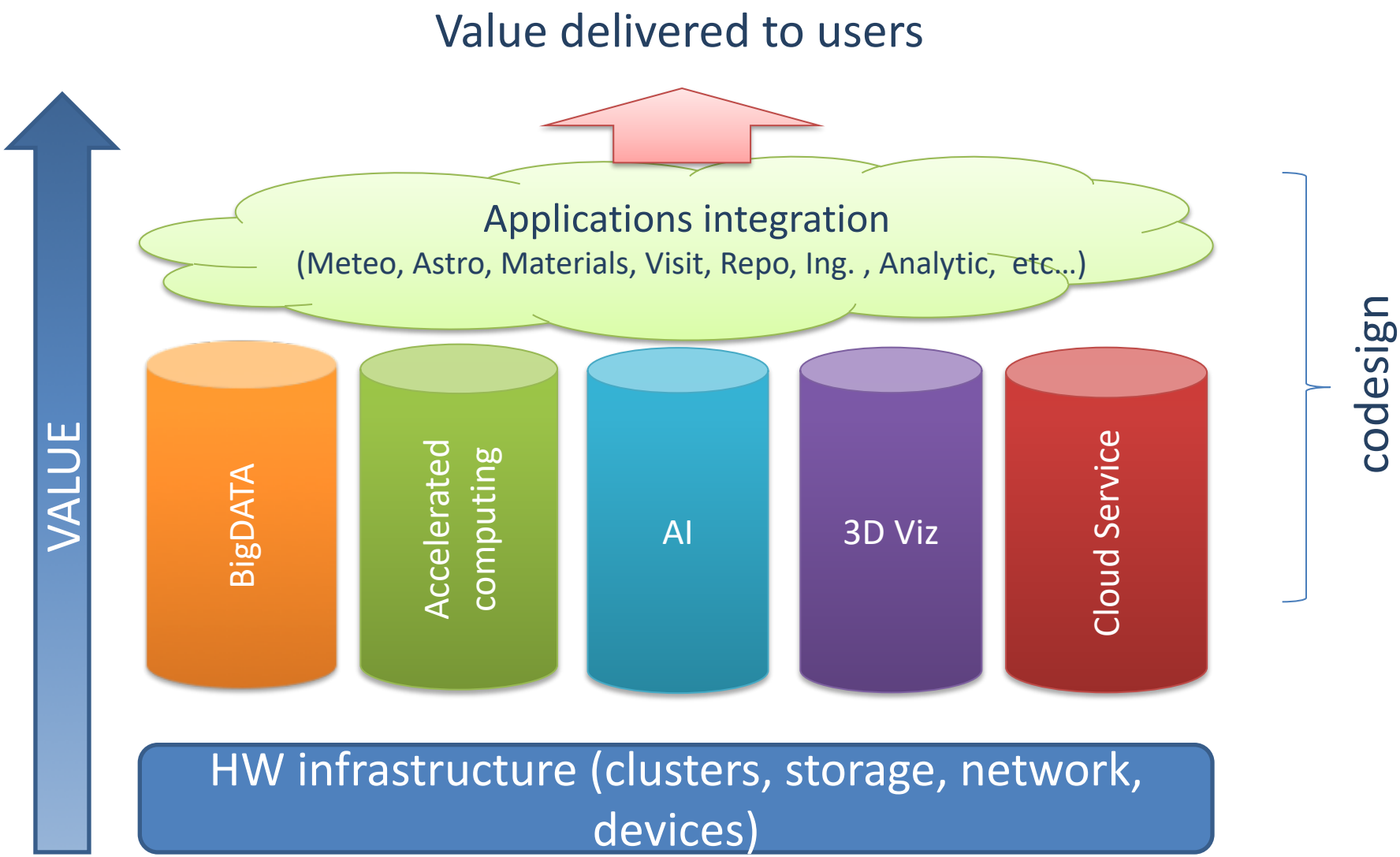


https://www.hpcwire.com/2018/01/25/hpc-ai-two-communities-future/

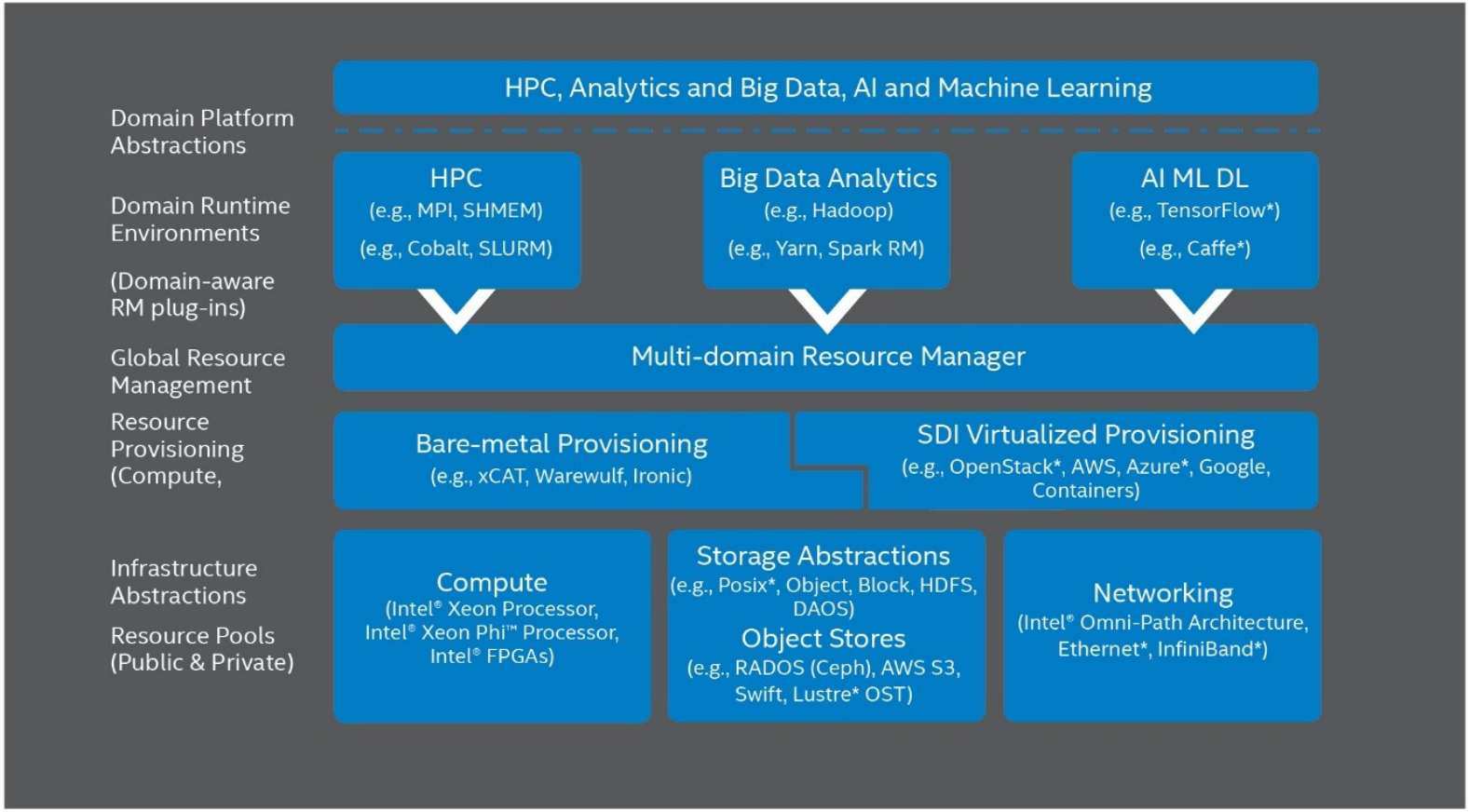
General Consideration

- Exascale is not (only) about scalability and Flops performance!
  - In an exascale machine there will be 10^9 FPUs, bring data in and out will be the main challenge.
  - 10^4 nodes, but 10^5 FPUs inside the nodes!
- There is no silver bullet (so far)
  - heterogeneity is here to stay
  - deeper memory hierarchies

HPC and Verticals



AI Gara (Intel)



the same architecture will cover HPC, AI, and Data Analytics through configuration, which means there needs to be a consistent software story across these different hardware backends to address HPC plus AI workloads.

Exascale system

Dimension	2012 System	Exascale System Specifications
Nodes	Single nodes and DDR	Multiple aggregated nodes with HBM stacks
Performance	~20 PF peak	1.30-1.40 EF peak
Power	~10 MW	~ 20MW-40MW
Space	~4,000 sqft	~ 6,000sqft
Memory Cap.	1.5 PiB	6-12 PiB
Memory B/W	4 PB/s	100-200 PB/s
NVM I/O	NA	10-100 TB/s

Al Gara's vision for the unification of the "3 Pillars" of HPC currently underway.

"The convergence of AI, data analytics and traditional simulation will result in systems with broader capabilities and configurability as well as cross pollination."

Exascale… some guess

- From GPU to specialized core (tensor core)
- Specialized memory module HBM
- Specialized non volatile memory NVRAM

Performance modelling

- Refactor code to better fit architectures with specialized HW

Paradigm and co-design

Identify latency and throughput
sub/module/class

App workflow

latency

throughput

comm

latency

throughput

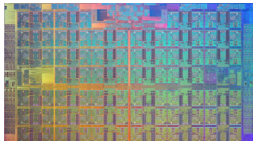
Re-factor

Latency code

Throughput code

Map and Overlap

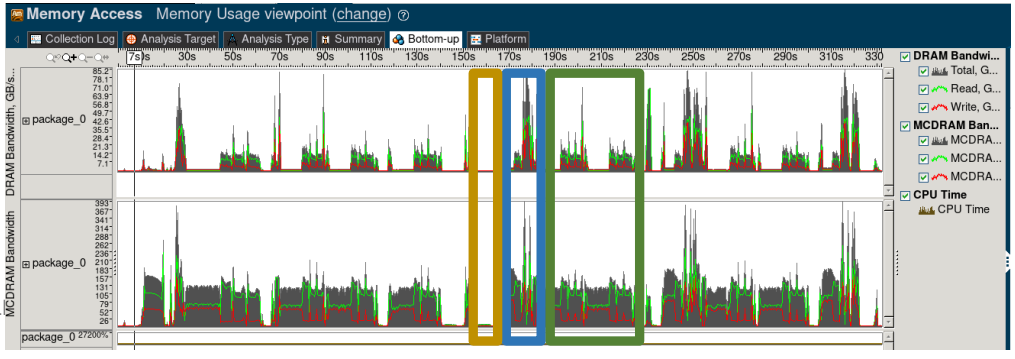
Map to HW



knl

host

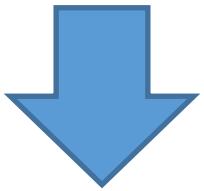
Heterogeneus



One size do not fit all

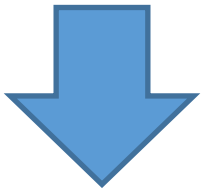
10^9 FPU to leverage

Best algo for 1FPU /= best algo for 10^9FPU



Implement the best algo for each scale

e.g. 2 FFT and data distribution in QE 6.2



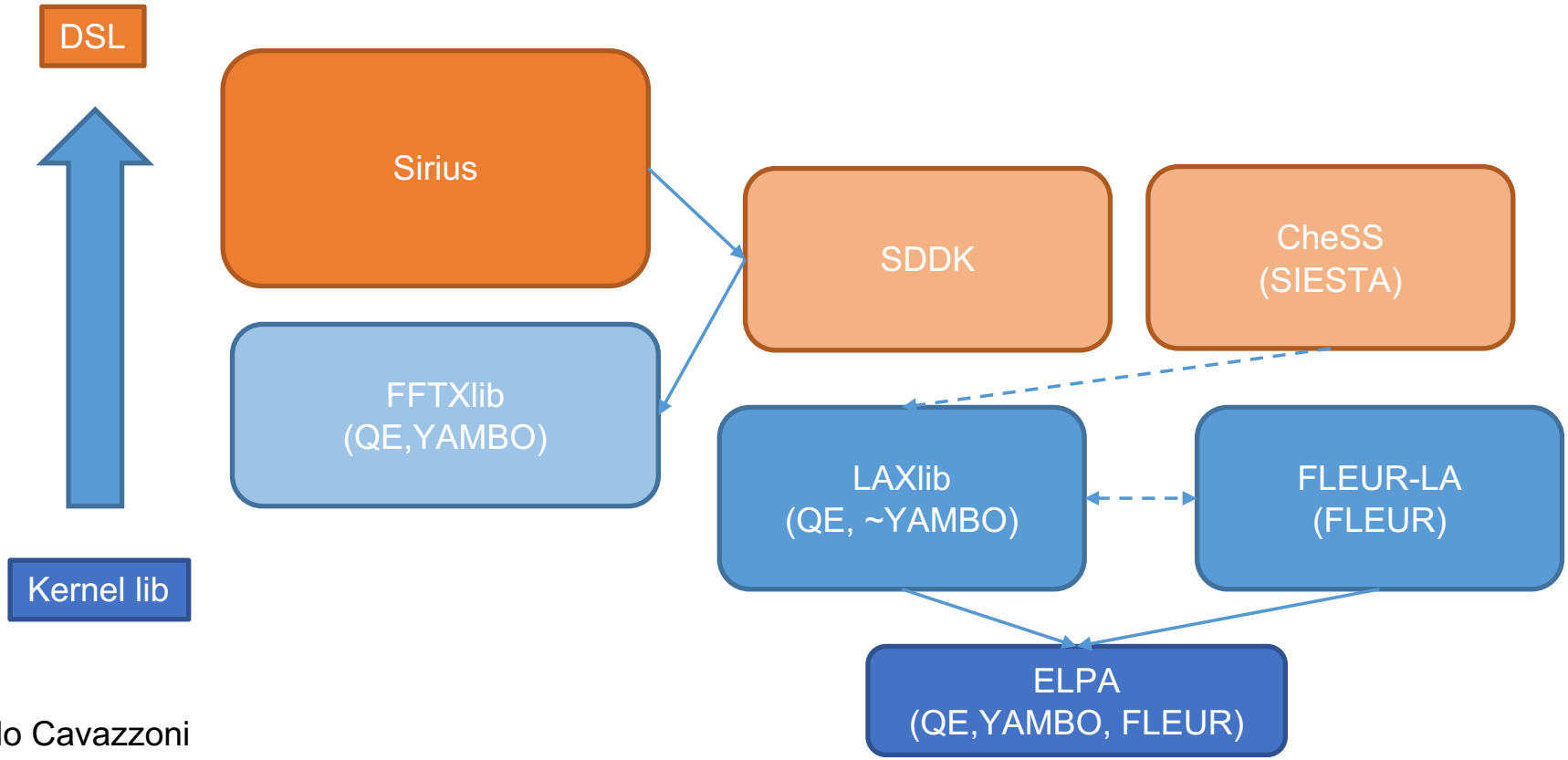
Autotuning

Choose the best at runtime

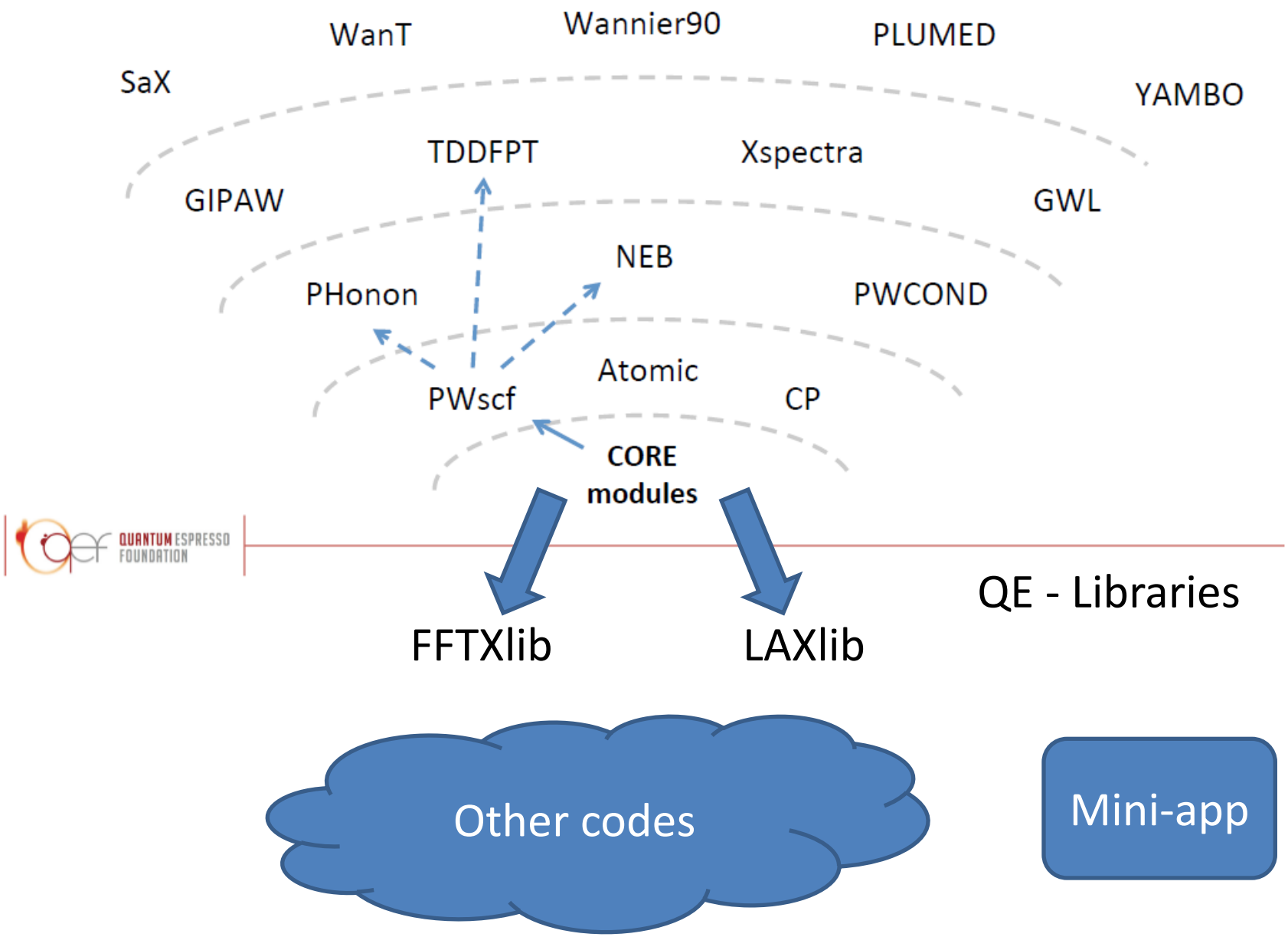
CoE what is my target architecture?

How could I cope with GPUs, many-cores, FPGAs?
I like homogeneous architecture! Why should I care about heterogeneous?

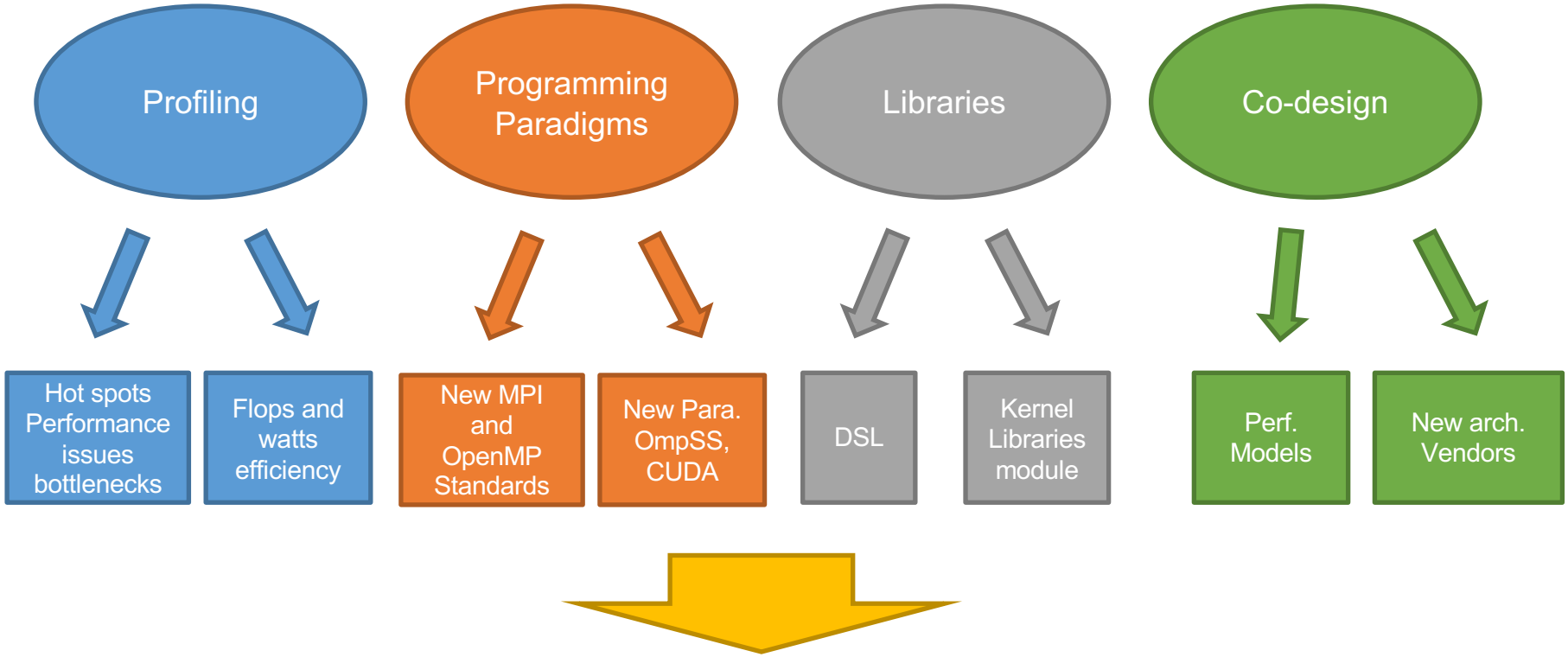
Heterogeneity is here to stay!!! → **Separation Of Concern** → DSL, Kernel libraries Modularization, API Encapsulation



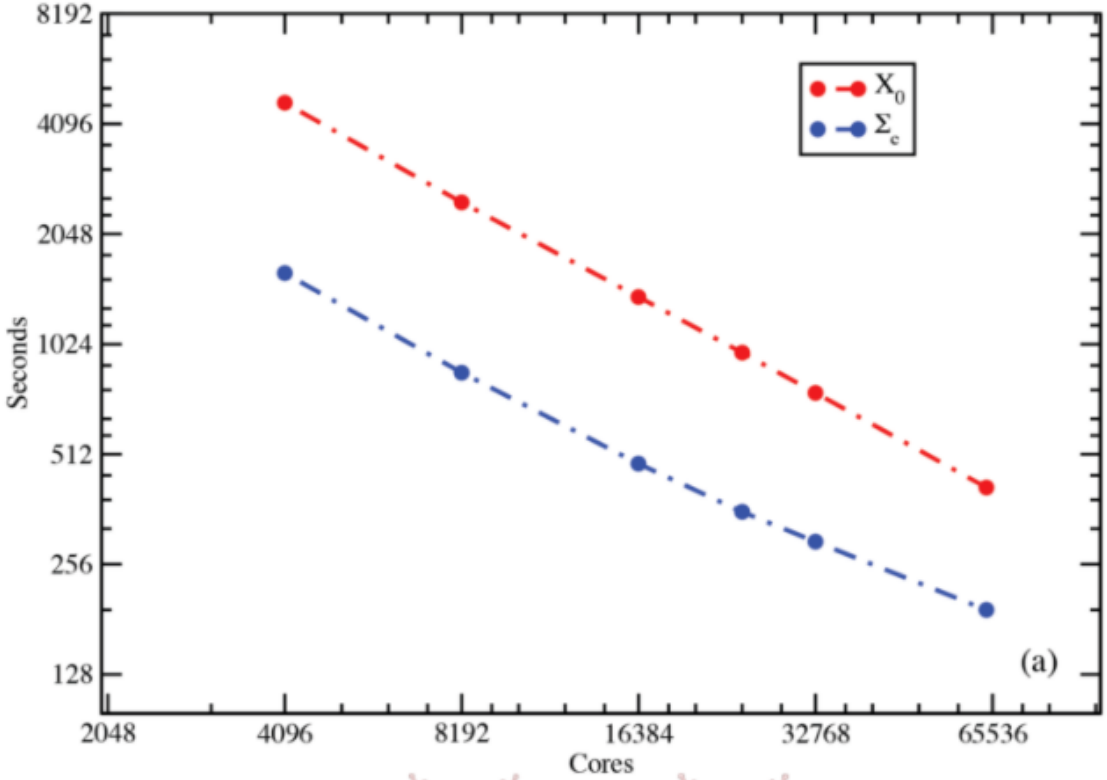
Beyond Modularization

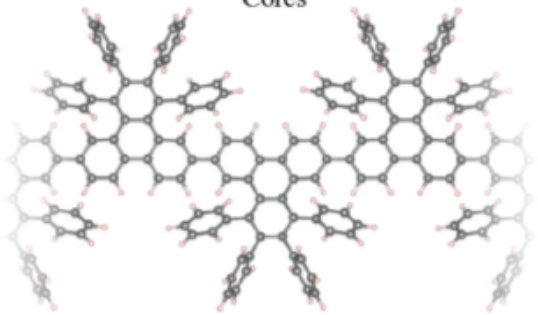


MaX Activities



Scaling-out YAMBO



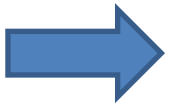


A single GW calculation has run on 1000 Intel Knights Landing (KNL) nodes of the new Tier-0 MARCONI KNL partition, corresponding to 68000 cores and ~ 3 pFlop/second.
The simulation, related to the growth of complex graphene nanoribbons on a metal surface, is part of an active research project combining computational spectroscopy with cutting edge experimental data from teams in Austria, Italy, and Switzerland.
Simulations were performed exploiting computational resources granted by PRACE (via call 14).

http://www.max-centre.eu/2017/04/19/a-new-scalability-record-in-a-materials-science-application/

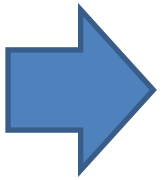
Planning for Exascale

Performance model, co-design, POC



code re-factoring

today	Pre-exascale	exascale
Yambo @ 3Pflops	Yambo @ 10-15Pflops	Yambo @ 50Pflops
Socket perf 3-5TFlops	Socket perf 10-15TFlops	Socket perf 20-40TFlops



Thank you!