WP1-3: Hw and Sw technologies for HPC and their use in SyeC applications

Eduard Ayguadé and Victor Viñals

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universidad Zaragoza

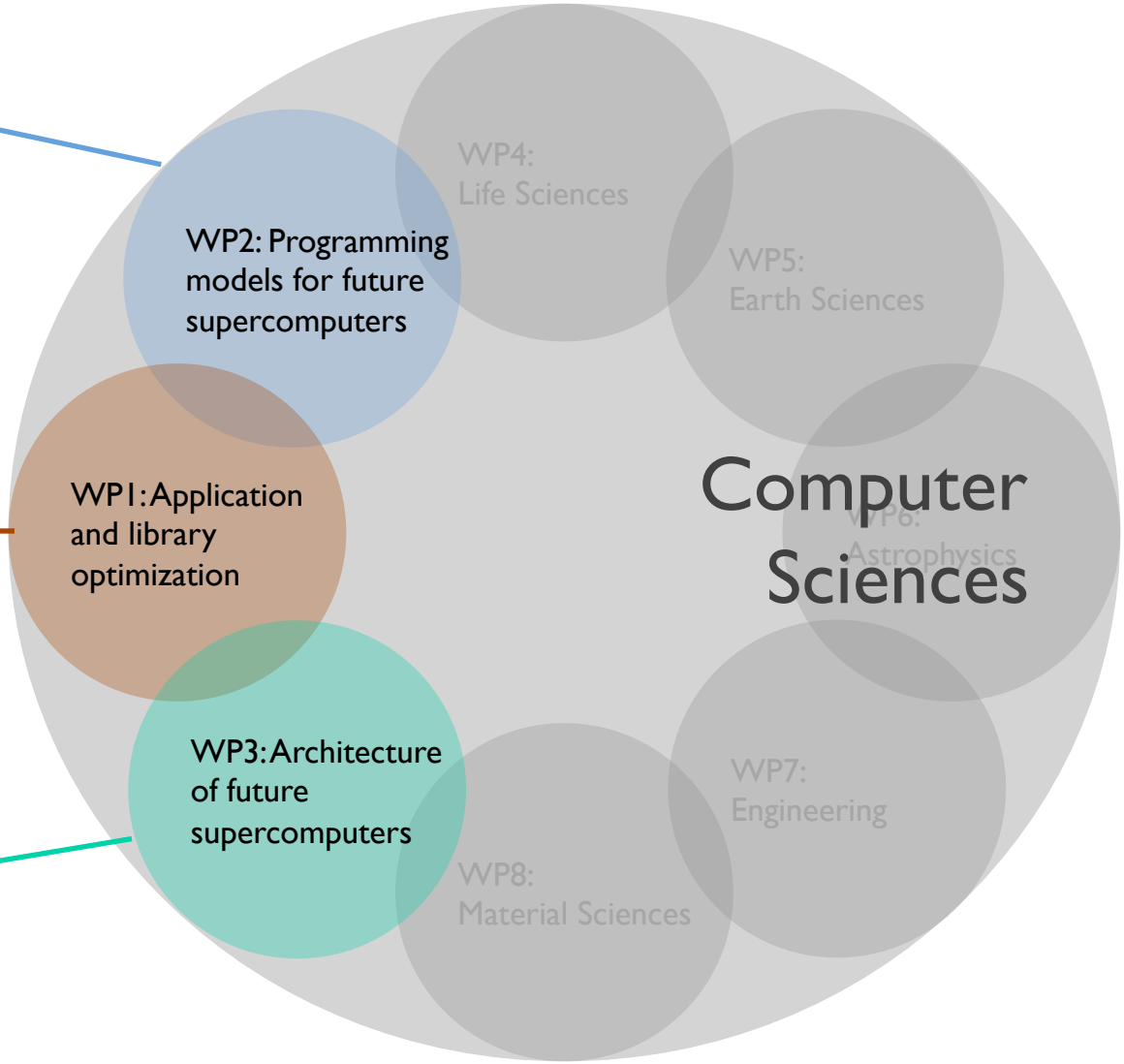
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
Consolider
PROGRAMA INGENIO 2010
SyeC
Supercomputación y eCiencia

WP1-3: project structure and participating groups (7)

- Arquitectura y Tecnología de Sistemas Informáticos (ArTeCS-UCM, **F. Tirado**)
- Arquitectura de Computadores (AC-UM, **E. López-Zapata**)
- Procesado Paralelo y Sistemas Distribuidos (PPDS-UAB/UdL, **A. Ripoll**)
- Computer Sciences (CS-BSC, **E. Ayguade**)

- Arquitectura y Tecnología de Sistemas Informáticos (ArTeCS-UCM, **F. Tirado**)
- Arquitectura de Computadores (AC-UM, **E. López-Zapata**)
- Procesado Paralelo y Sistemas Distribuidos (PPDS-UAB/UdL, **A. Ripoll**)
- Computer Applications in Science and Engineering(BSC-CASE, **J.M. Cela**)
- Computer Sciences (CS-BSC, **E. Ayguadé**)

- Arquitectura y Tecnología de Sistemas Informáticos (ArTeCS-UCM, **F. Tirado**)
- Arquitectura de Computadores (gAZ, **V. Viñals**)
- Arquitectura y Tecnología de Computadores (ATC-UC, **J. R. Beivide**)
- Arquitectura de Computadores (DAC-UPC, **M. Valero**)



SyeC
Supercomputación y eCiencia

WP1: activities and outcomes

From sequential optimization to parallelization

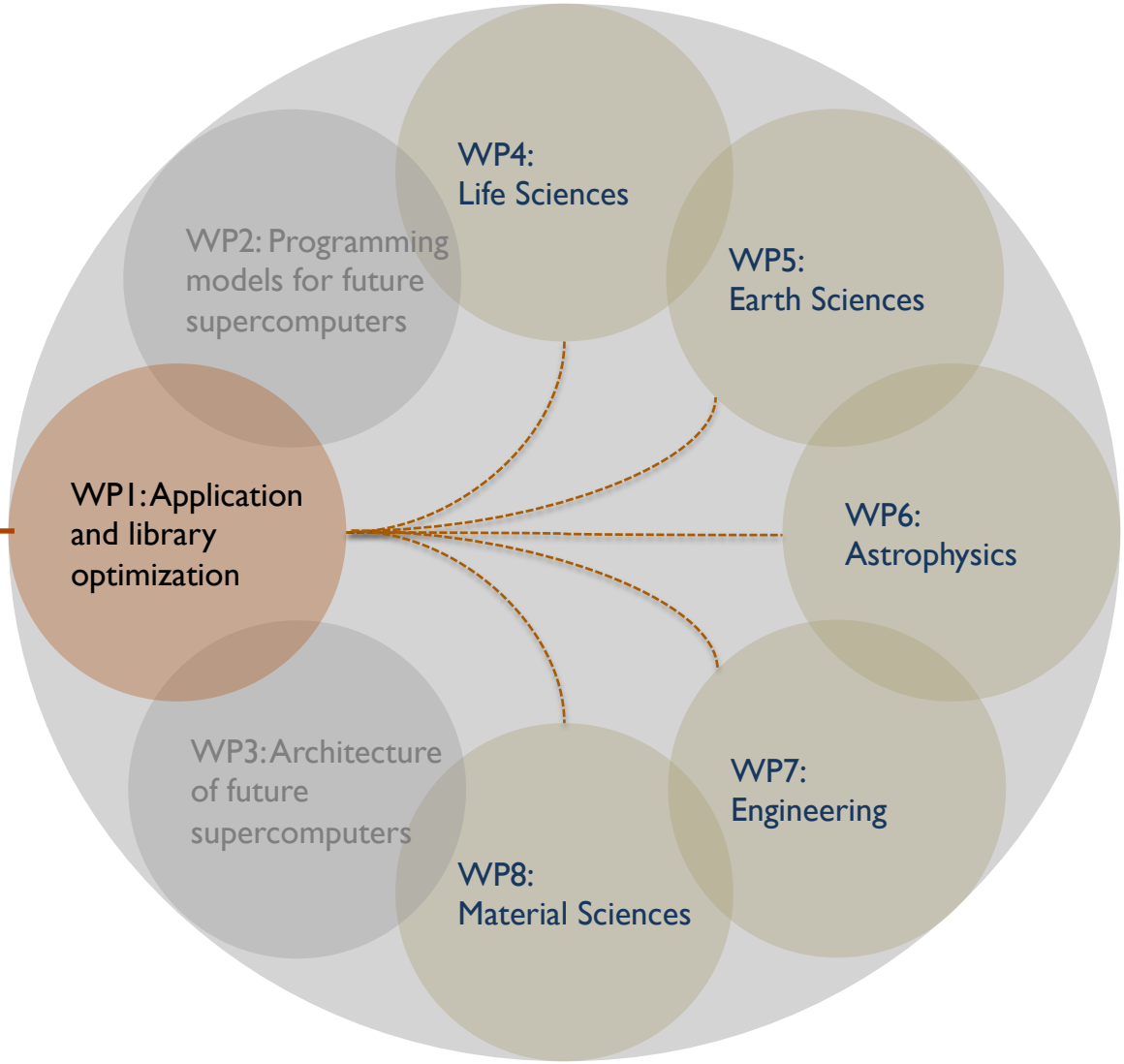
- Algorithmic improvements, load balancing
- Hybrid MPI/OpenMP parallelization
- Optimization for GPU and multicore

Collaborations:

- Gromacs: UB + CS-BSC
- T-coffee: CRG + PPDS-UAB/UdL
- GEM: CRG + gAZ
- NCAR-WACCM: UCM + ArTeCS-UCM
- Atmospheric: ES-BSC + CASE-BSC
- SPEV: UV + AC-UM
- MRGENESIS: UV + CS-BSC
- RATPENAT: UV + CASE-BSC
- pDEVA: UAM + CASE-BSC
- BLKTRI: CIEMAT + ArTeCS-UCM
- SIESTA: CSIC–ICN2 + CASE-BSC

Other applications

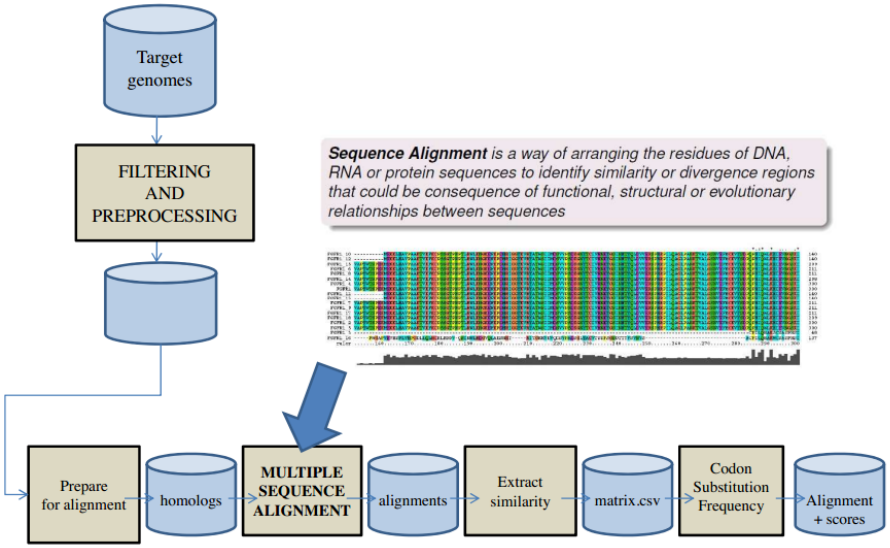
- Histogram, k-means, biomedicine, 3D Fast Wavelet, Generalized Hough Transform: AC-UM
- Image segmentation (gAZ)



SyeC
Supercomputación y eCiencia

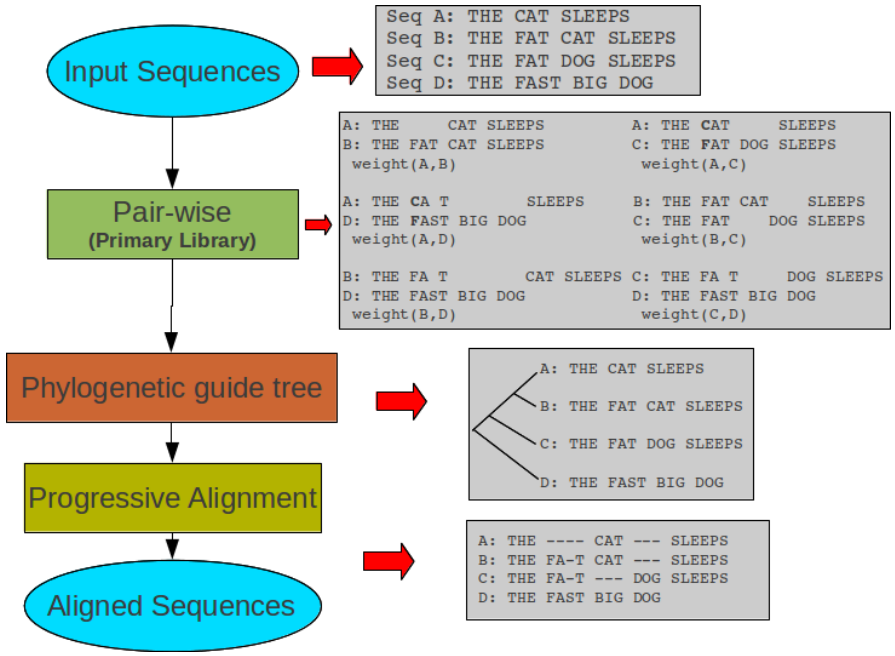
Multiple sequence alignment and T-coffee

- MSA common in many bio-computing workflows



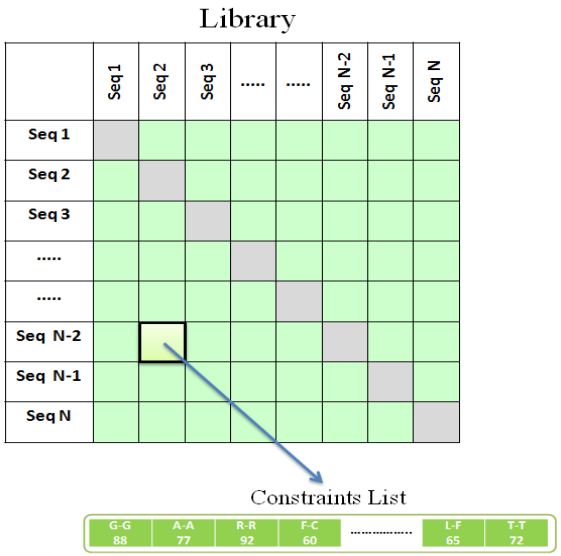
- Limitations: memory requirements and execution time (quadratic with number of sequences and length)

- T-coffee: reference MSA algorithm of Life Science community

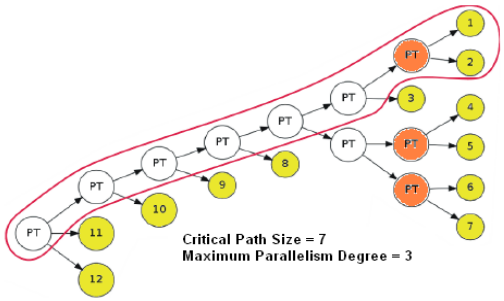


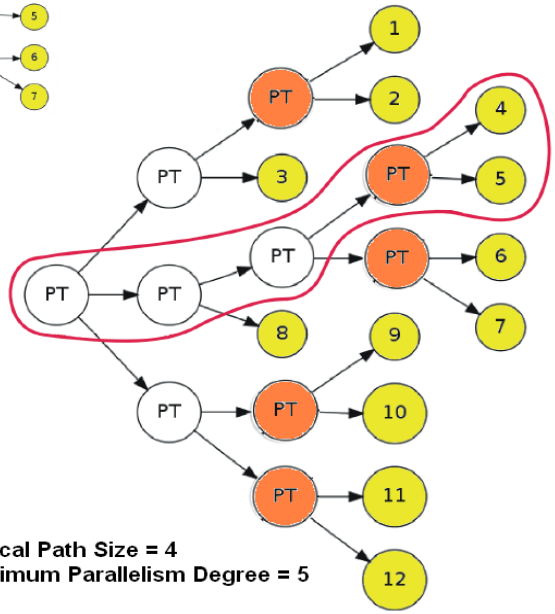
T-coffee: two main optimizations

- Improved **primary library generation**, discarding redundant/less representative information
  - Reduce memory requirements
  - Better execution time and increased scalability
  - Keep alignment quality



- Balanced **Guide Tree** (BGT)
  - Better parallelism degree and load balancing in the progressive alignment phase
  - Keep alignment quality

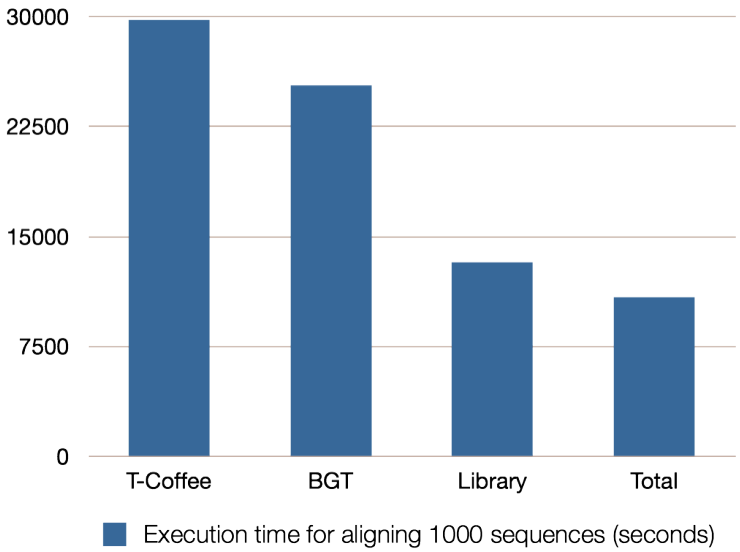


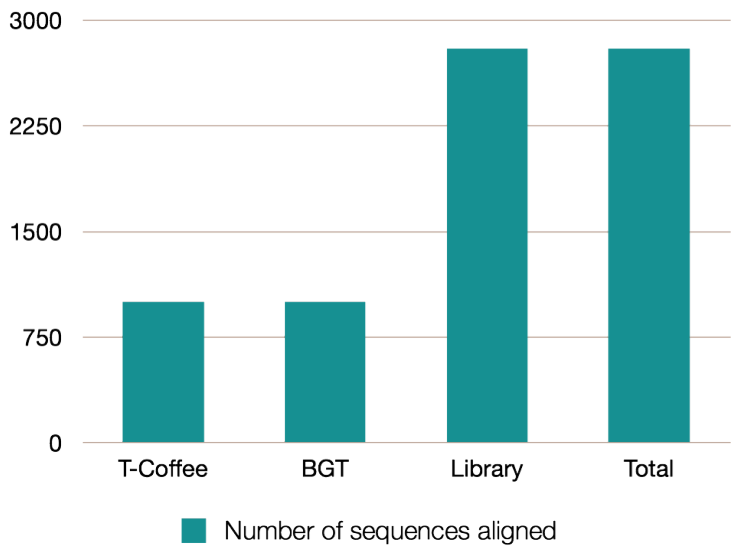


UAB Universitat Autònoma de Barcelona
Universitat de Lleida
CRG Centre de Regulació Genòmica

T-coffee: performance improvement

Execution time and number of sequences aligned





Biological accuracy of alignment

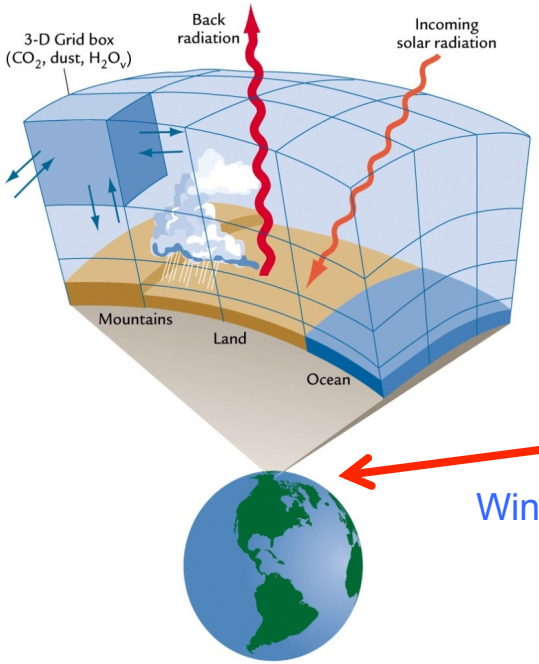
Similarity	standard T-Coffee quality	optimized T-Coffee quality
0-15	0,421	0,379
15-35	0,721	0,695
25-35	0,876	0,865
35-100	0,951	0,956
average	0,709	0,687

PREFAB MSA benchmark results

UAB Universitat Autònoma de Barcelona
Universitat de Lleida
CRG Centre de Regulació Genòmica

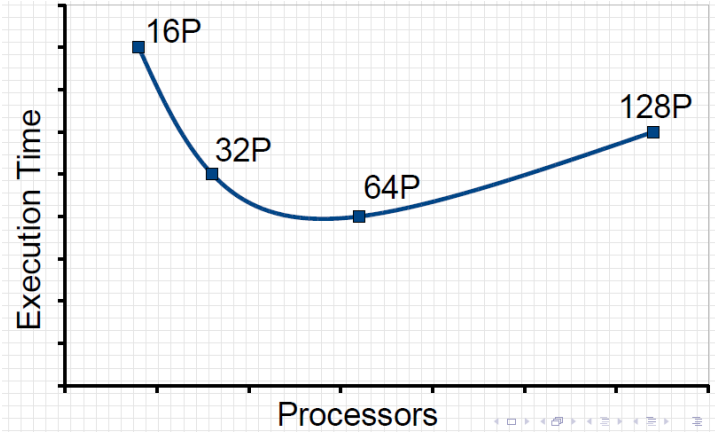
Whole-Atmosphere Community Climate Model (WACAMM)

- Main bottleneck in radiation code
  - Computational cost depends on seasonal cycles: temporal variations and spatial variations



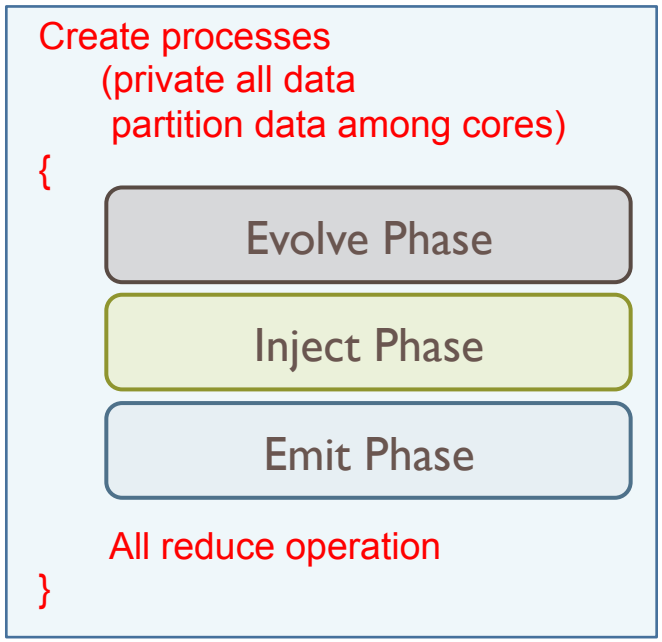
Winter season

  - Improving load balancing across processors using a cyclic data distribution of the earth grid

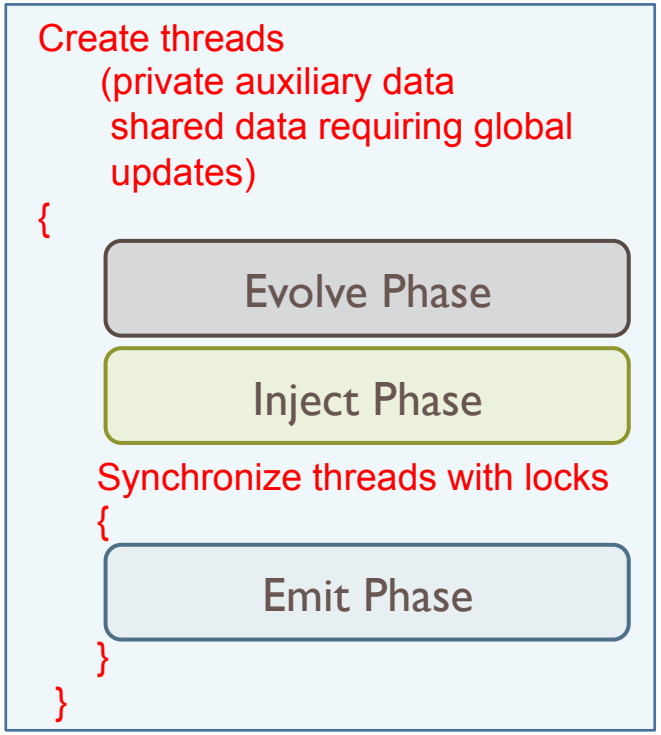


SPectral EVolution simulation (SPEV)

- Radio emissions from extragalactic jets → observed on Earth
- Data-sharing on multi-socket and multicore architectures
- Private data approach



- Shared data approach



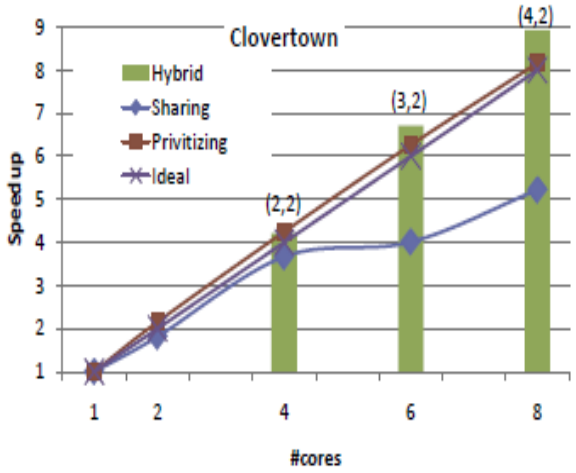
- Hybrid approach
  - Private data across chips (sockets)
  - Shared data inside chips

UNIVERSIDAD DE MÁLAGA

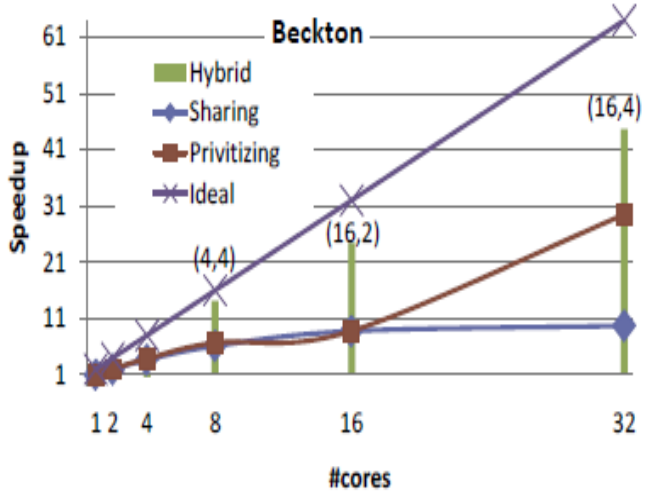
VNIVERSITAT DE VALÈNCIA

SPEV: performance results

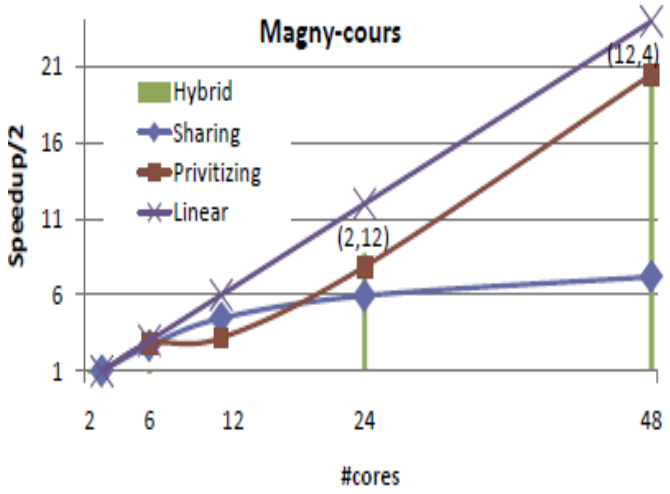
- Platforms
  - 2-socket 4-core Intel Clovertown
  - 4-socket 8-core Intel Beckton
  - 4-socket 12-core AMD Magny-Cours
- Programming models
  - OpenMP (threads and synchronization)
  - MPI (processes and reduction)



(a)



(b)

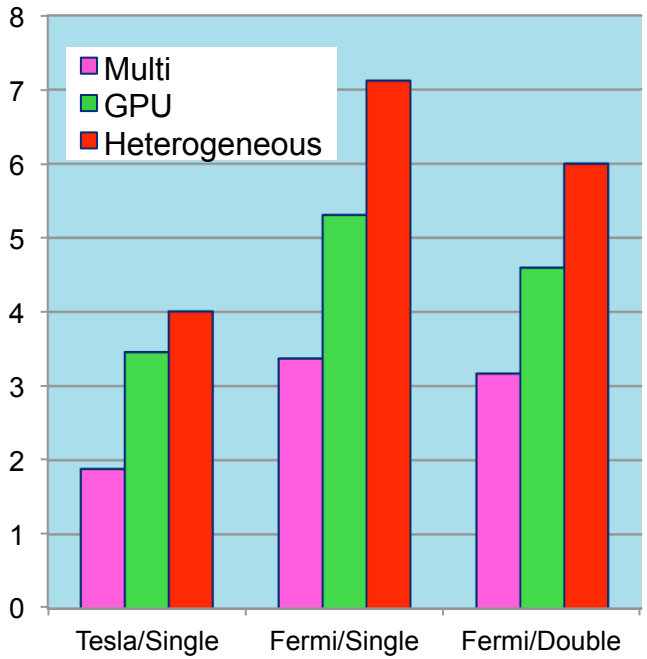


(c)

BLKTRI solver from FISH-PACK library

- Numerical simulation of incompressible fluid flows
  - The most time consuming part of incompressible unsteady Navier Stokes solver is related to the solution of a pressure Poisson equation
- Target architecture: heterogeneous multicore/GPU
  - Coarse-grain loop parallelism (OpenMP) in tri-diagonal problems for multicore
  - Fine-grain parallelism (CUDA) for GPU, with major algorithm changes
- Substantial speedup over the standard BLKTRI solver using a combination of the two strategies (overlapping)

- Tesla Platform (2009)
  - Xeon E5520
  - Tesla C1060
- Fermi Platform (2011)
  - Xeon E5645
  - Fermi C2050



WP2: activities and outcomes

Definition and implementation of programming models:

- **COMPSs** for the specification of workflows, support for web services
- **OmpSs** for asynchronous/dataflow programming in OpenMP, heterogeneity support (CUDA, OpenCL)
- **Chapel** (from Cray): implementation of the communication layer handling the transference of array regions in clusters
- Extensions to Intel **TBB** for the specification of wavefront computations
- **Source-to-source compilation (PPCG)**, from sequential to parallel code for modern GPU (collaboration with IMEC/INRIA)

Scheduling techniques for homogeneous and heterogeneous multicore architectures (accelerators, asymmetric multicores, …)

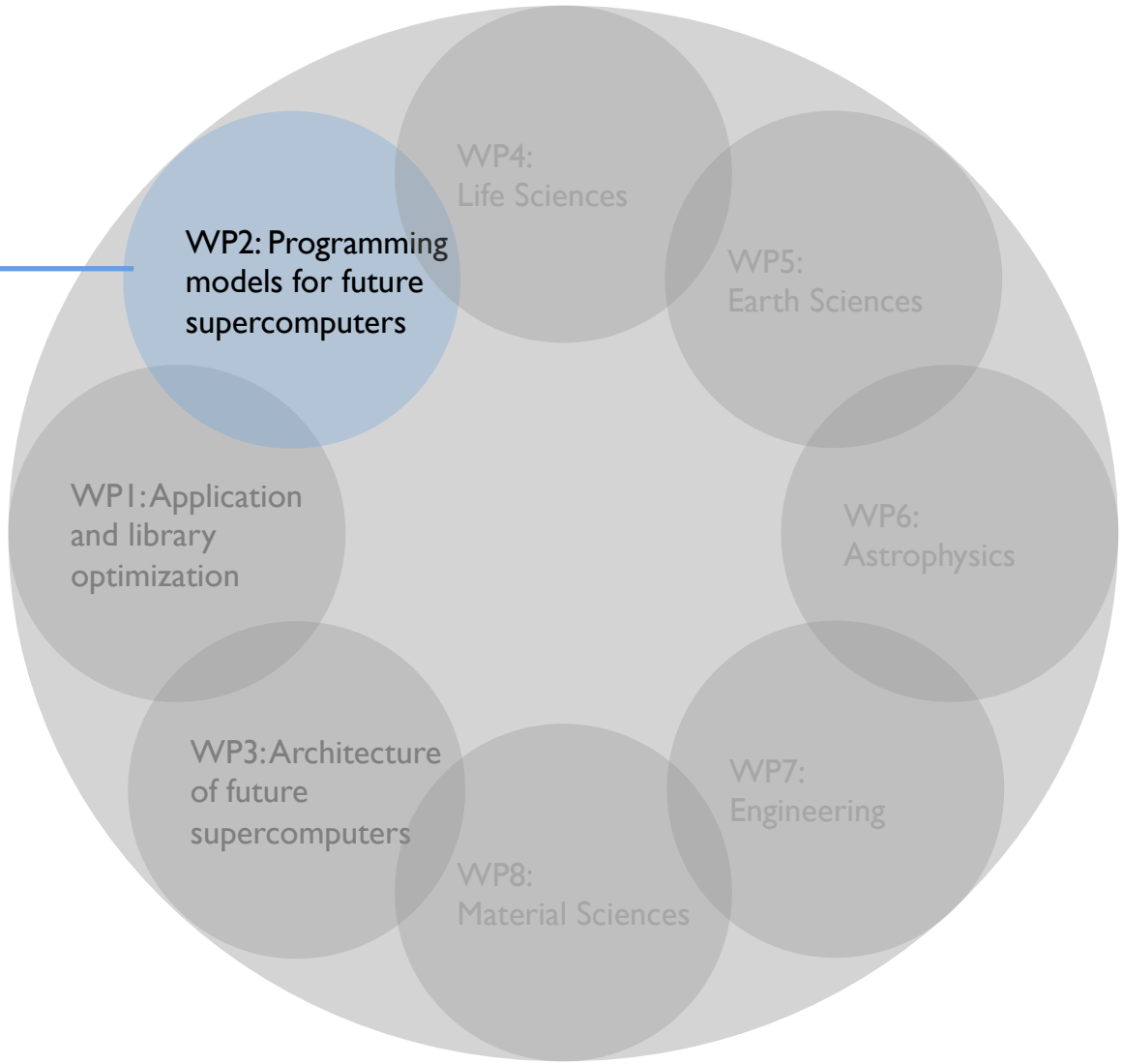
New data distribution policies and scheduling policies for **Hadoop/MapReduce**:

Optimization of application workflows

- COMPSs and MapReduce

Collaborations

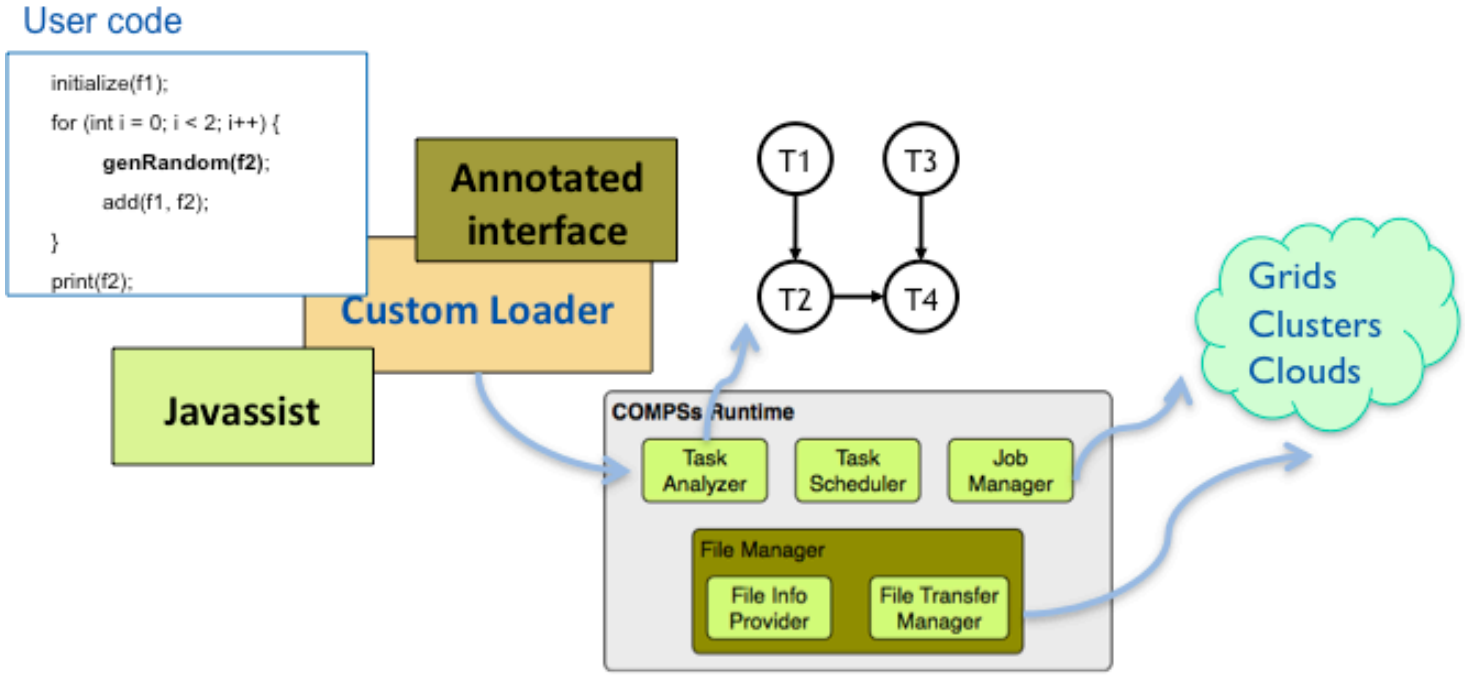
- Milky Way origin (GASS, GOG): UB + CS-BSC
- Protein dynamics, gene detection: IRB-UB + CS-BSC
- LNC-RNA: CRG + PPDS-UAB/UdL
- Ocean-atmosphere circulation model: ES-BSC + CS-BSC



SyeC
Supercomputación y eCiencia

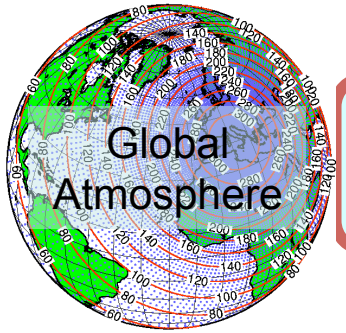
COMPSs programming model

- Easy programming, execution in any platform (cluster, Grid and Cloud)
- Task based programming model (automatic generation of data-dependent task graph)
  - Coarse grain tasks: methods and web services
  - Whole application can be exposed as a new web-service
- Cloud interoperability: commercial solutions (e.g. Azure and Amazon), open source (e.g. OpenNebula, EMOTIVE Cloud, OpenStack)



COMPSs: coupled models in Earth Sciences models

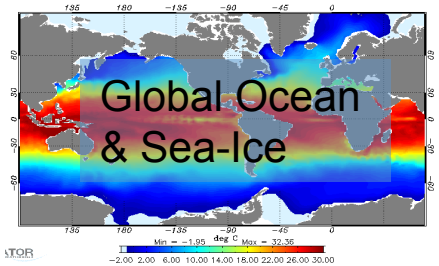
- Global coupled ocean-atmosphere general circulation model
  - Communication between atmospheric and ocean models done through the CMCC parallel version of OASIS3 coupler



ATMOSPHERE (dynamics, physics, prescribed gases and aerosols)

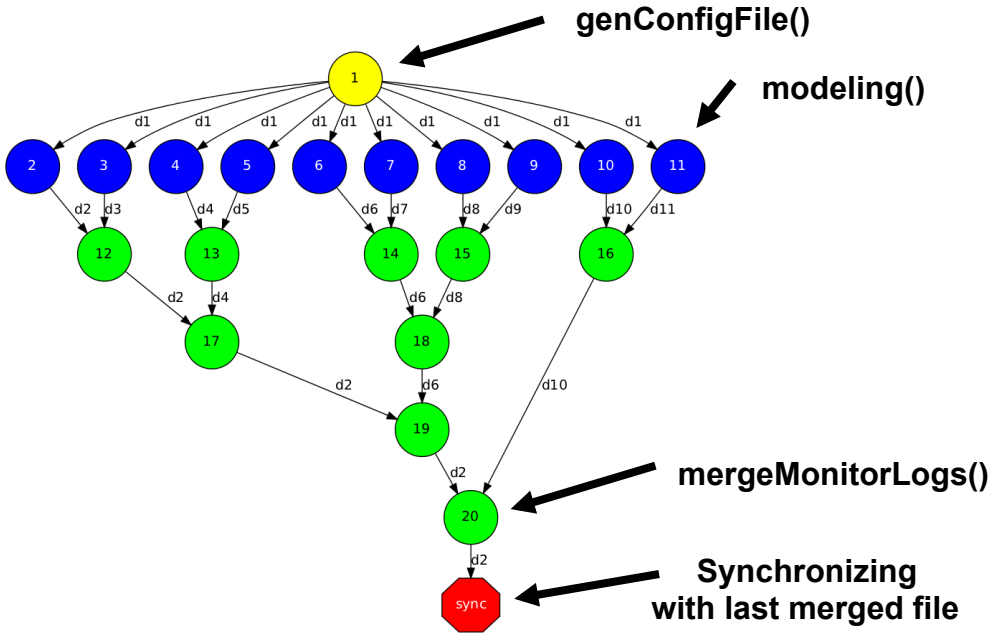
ECHAM5 T159 (~ 80 Km)

COUPLER Oasis 3



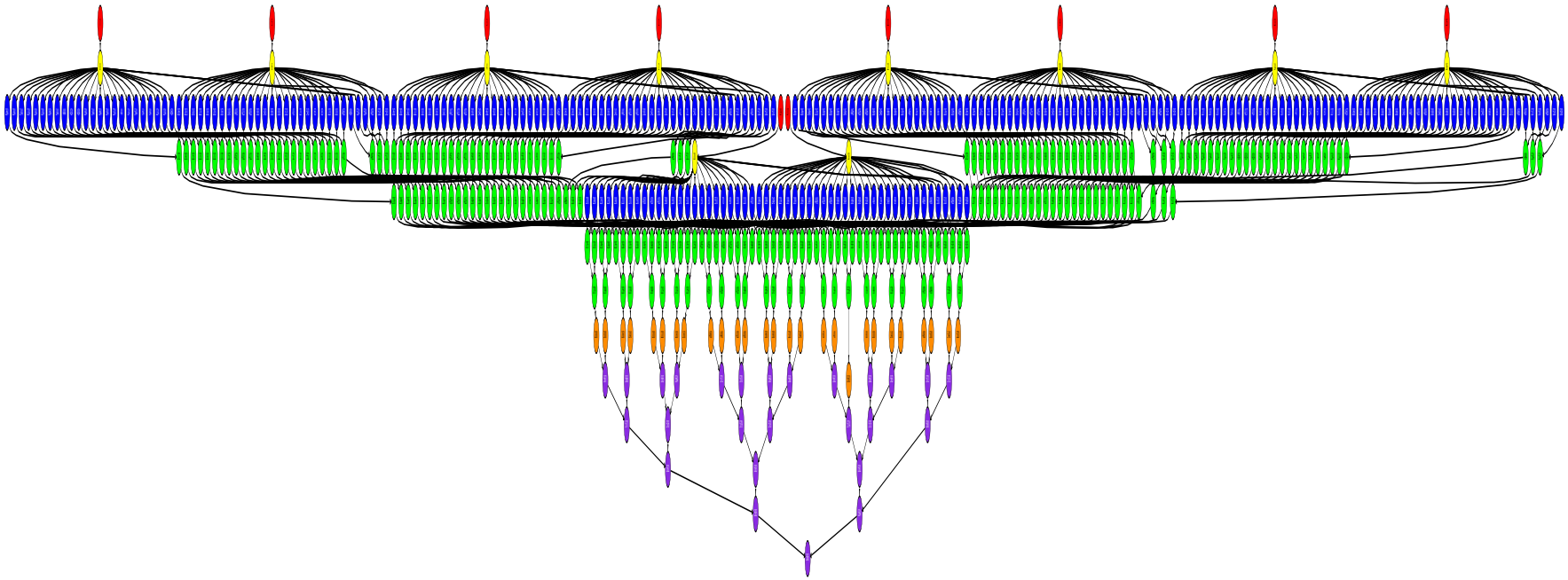
OCEAN:
OPA 8.2/ORCA2 (2º)

SEA-ICE: LIM



COMPSs: protein dynamics in Life Sciences workflow

- DISCRETE is a package devised to simulate the dynamics of proteins using the Discrete Molecular Dynamics (DMD) method
  - From a set of protein structures, perform a series of simulations to optimize 3 parameters: FDVW, FSOLV, EPS
  - For each structure, $N_{fdvw} \cdot M_{fsolv} \cdot L_{eps}$ simulations are done

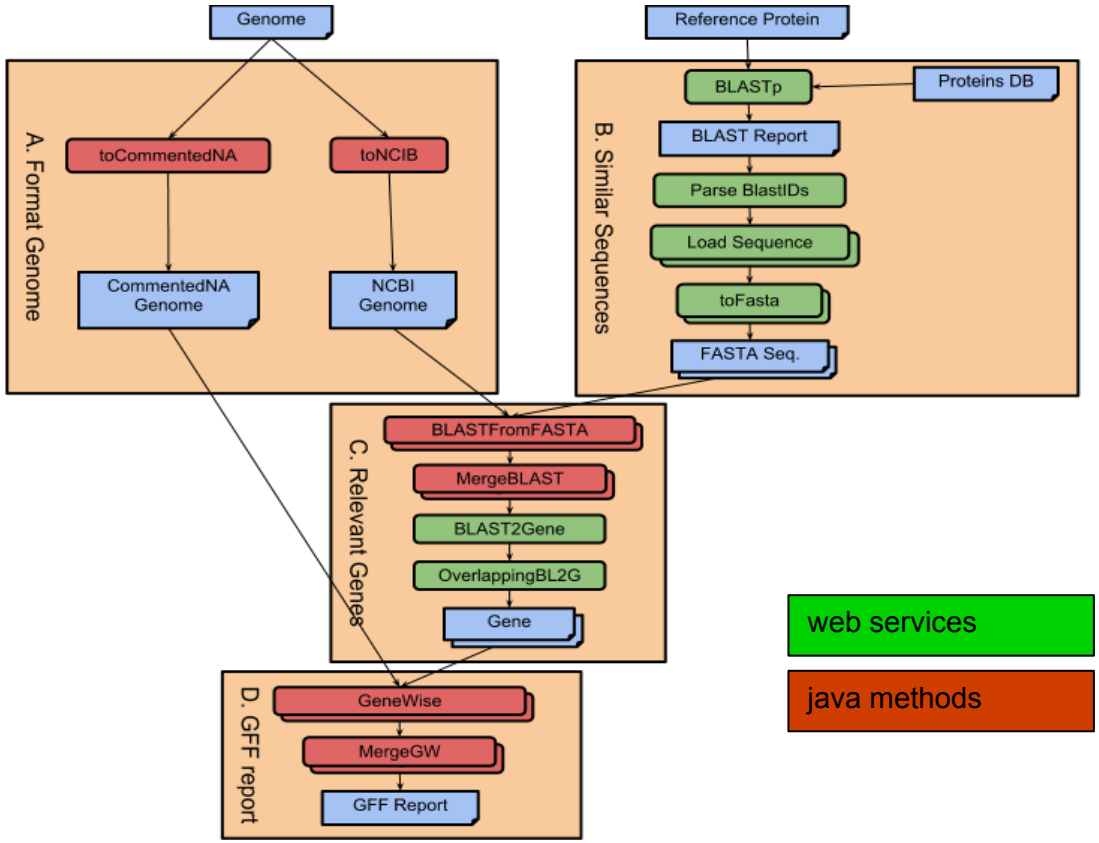


IRB BARCELONA INSTITUTE FOR RESEARCH IN BIOMEDICINE

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

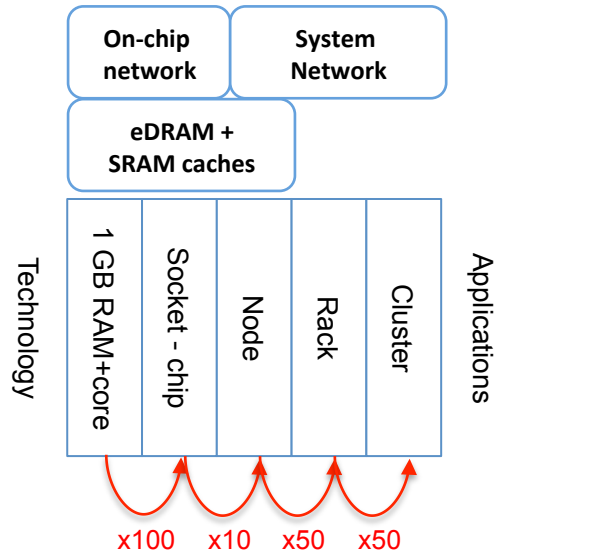
COMPSs: gene detection in Life Sciences workflow

- Automatic identification of genes which causes a disease
- Combine web services with computations
- Implemented as a new composite service



WP3: activities and motivation

Memory hierarchy and interconnection hierarchy



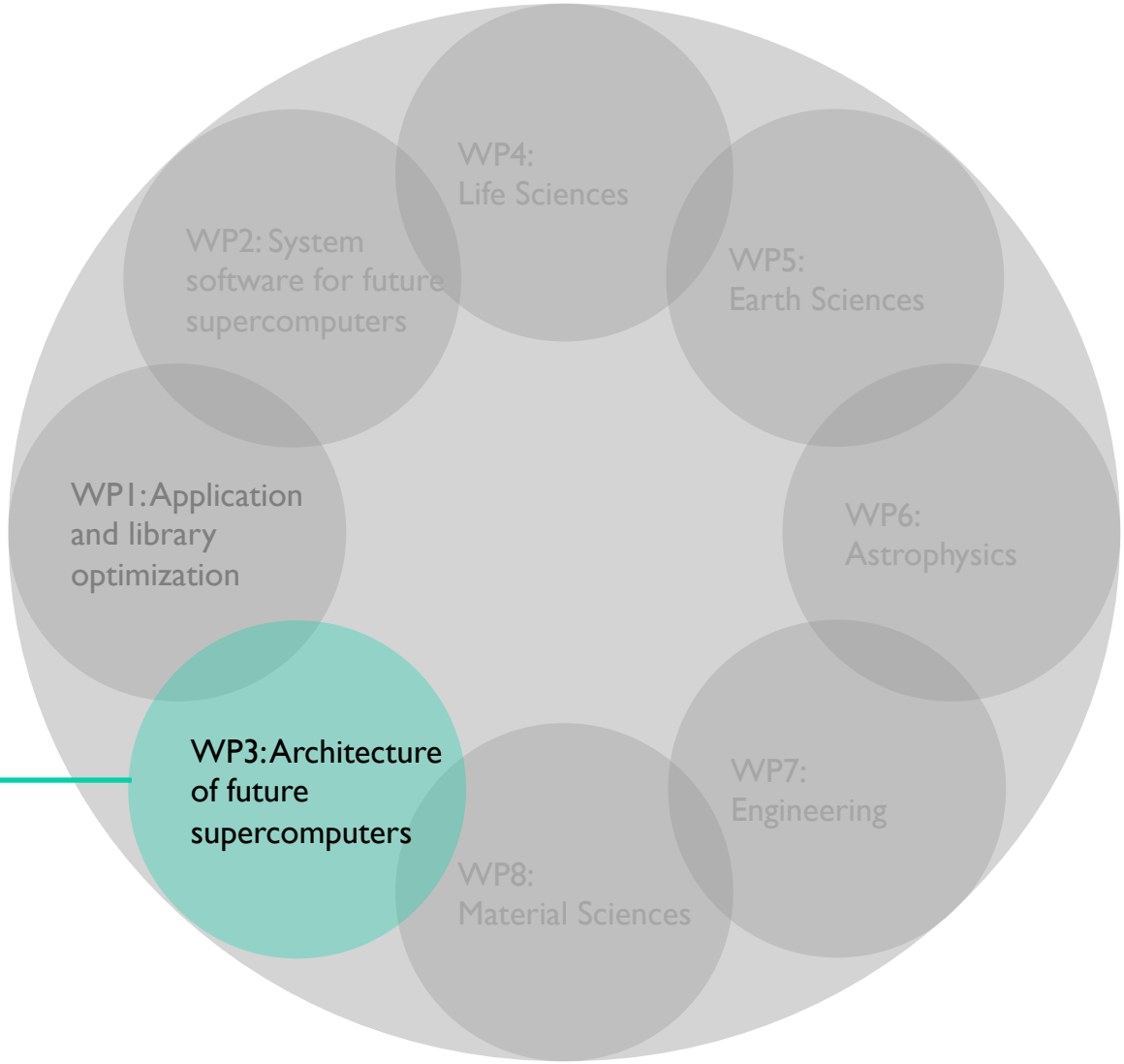
R = 2.5 Mcores x 40 GFLOPS/core
x 0.8 efficiency = **80 PFLOPS**

M = 2.5 PBytes

P < 10 MW

x4 Titan

Qualitative changes needed in several layers in order to reach Exascale levels of performance



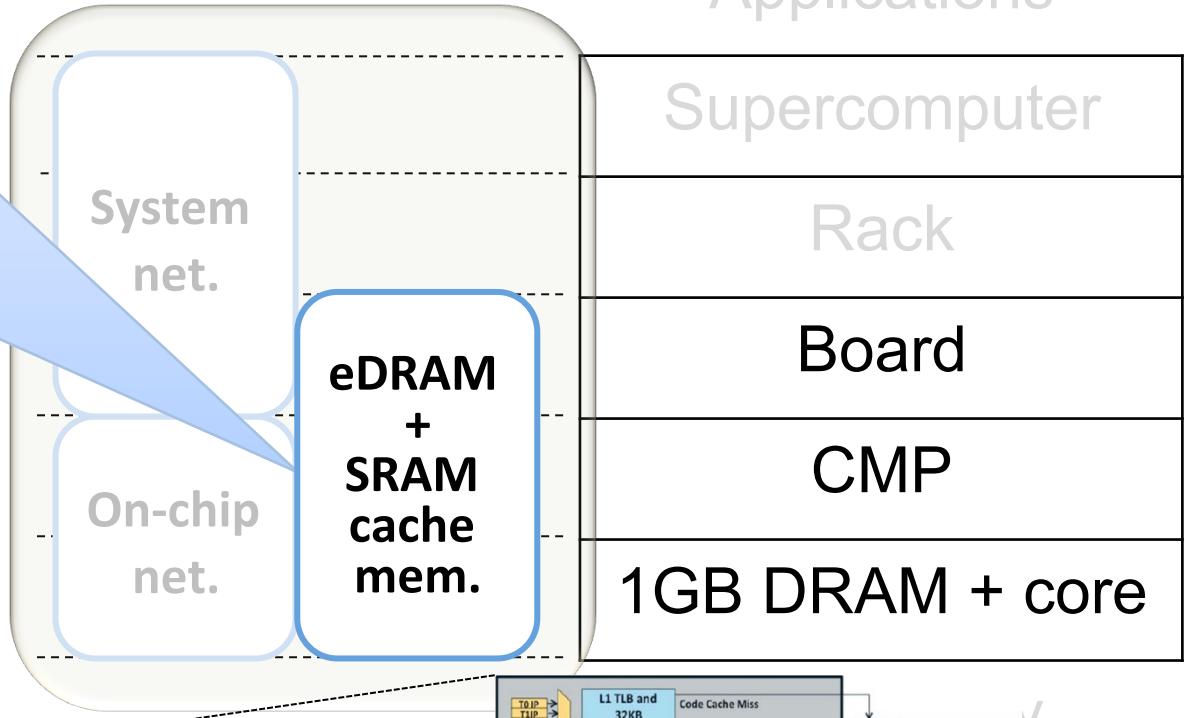
UC UNIVERSIDAD DE CANTABRIA

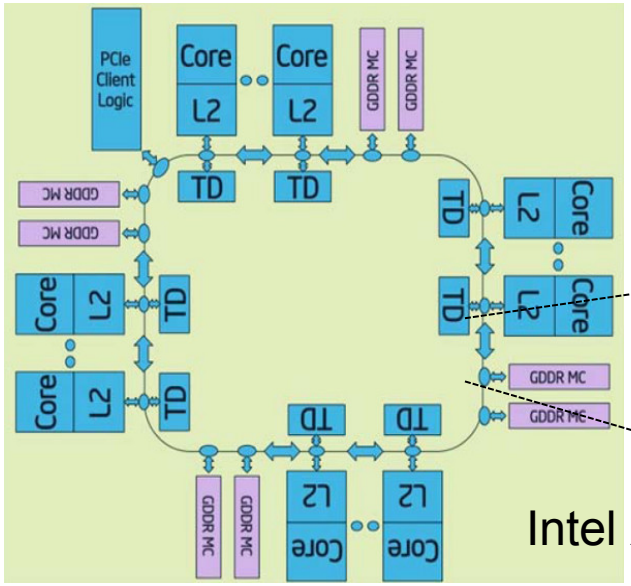
Universidad Zaragoza 1542

Cache hierarchy

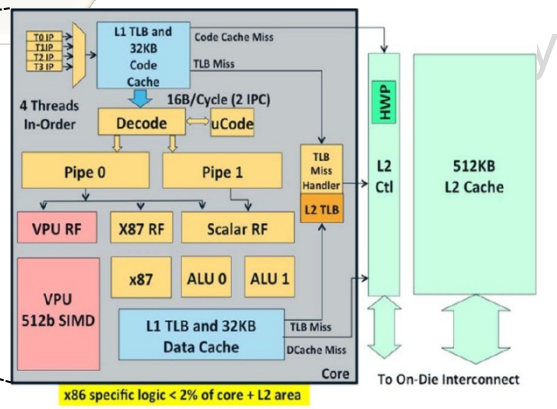
- First-level data caches
  - L-NUCAs, LP-NUCAs
- Adaptive data prefetching in last-level caches
- Reuse-based replacement and storage management
- Filtered coherence schemes
- Synchronization speeding-up

The Exaflop computer: qualitative changes needed in several layers





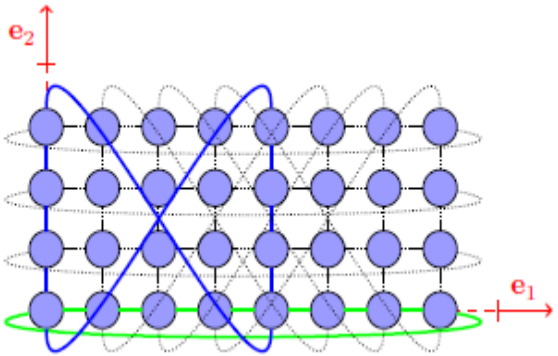
Intel Xeon Phi, 60 cores

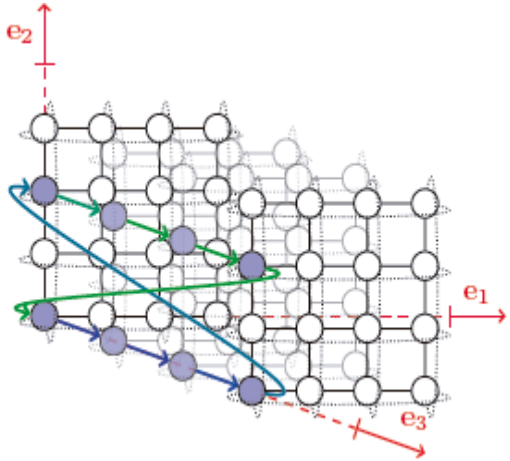


Network foundations

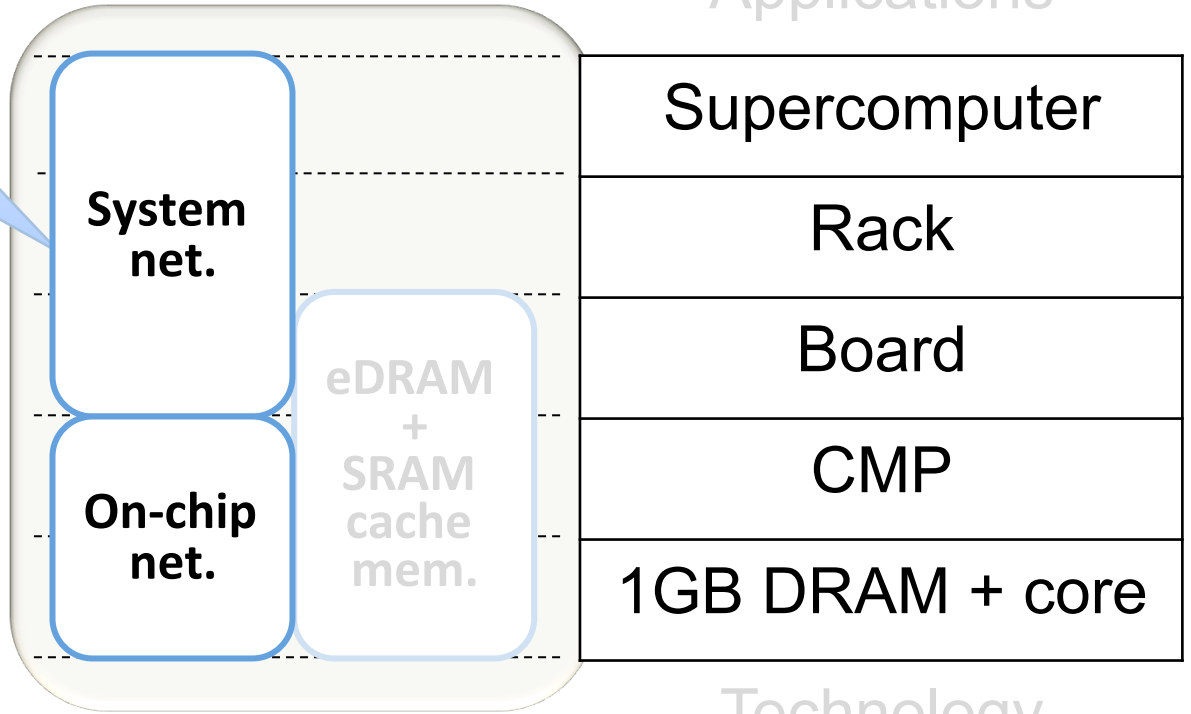
Theory: applying topological results to network design

- low-degree networks
- cayley graphs
- gaussian integers
- error-correcting codes





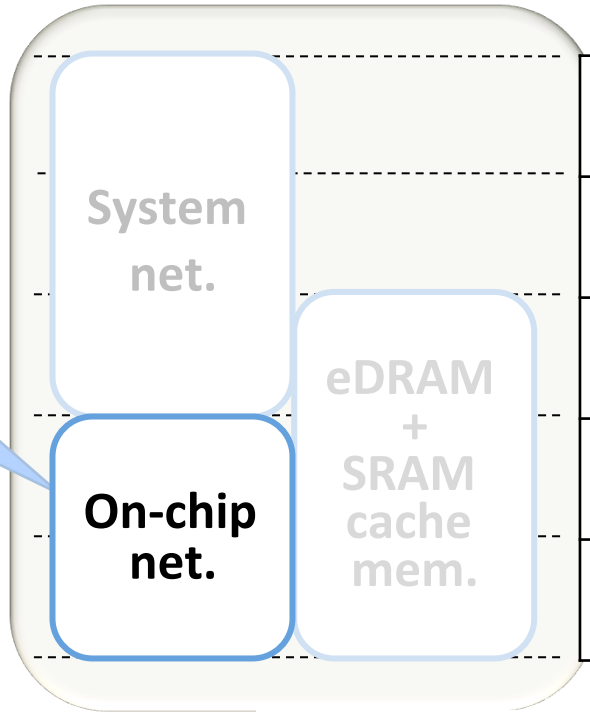
The Exaflop computer: qualitative changes needed in several layers



Networks on Chip

- Design driven by
  - realistic traffic (coherence and DRAM messages)
  - 2D feasibility
- King lattices
- Concentrated topologies
- Starting photonic networks

The Exaflop computer: qualitative changes needed in several layers



Applications

Supercomputer

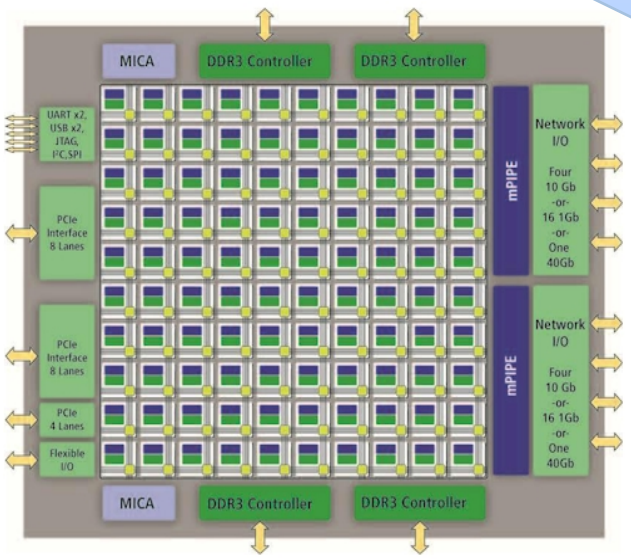
Rack

Board

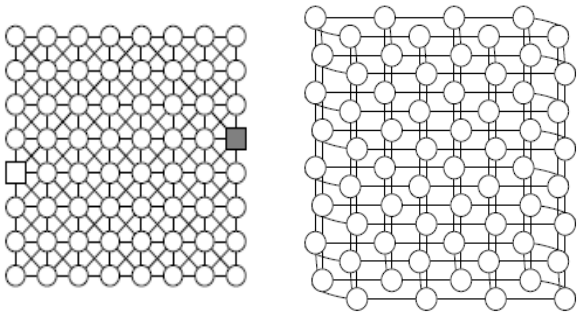
CMP

1GB DRAM + core

Technology



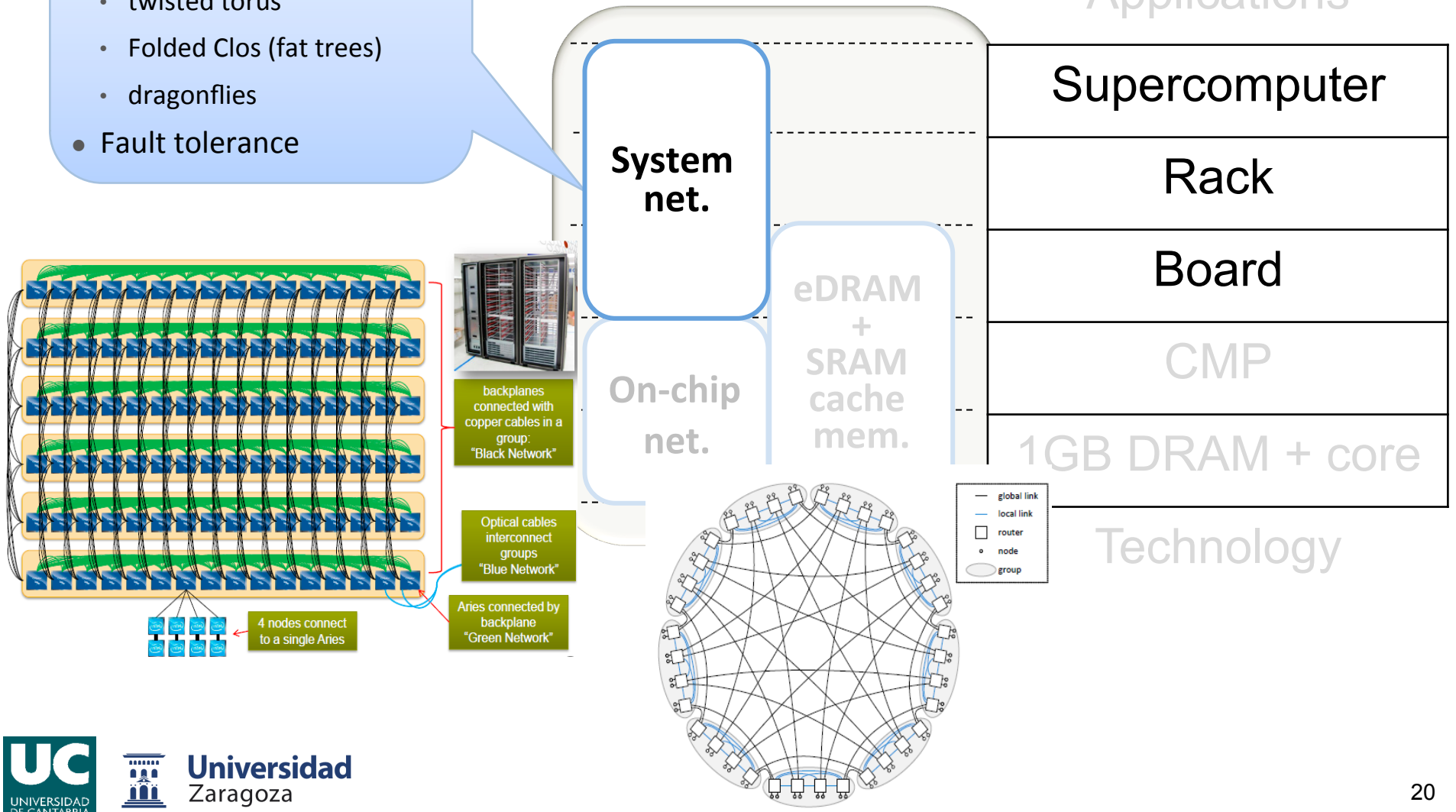
On-chip 2D mesh in 100-core Tilera



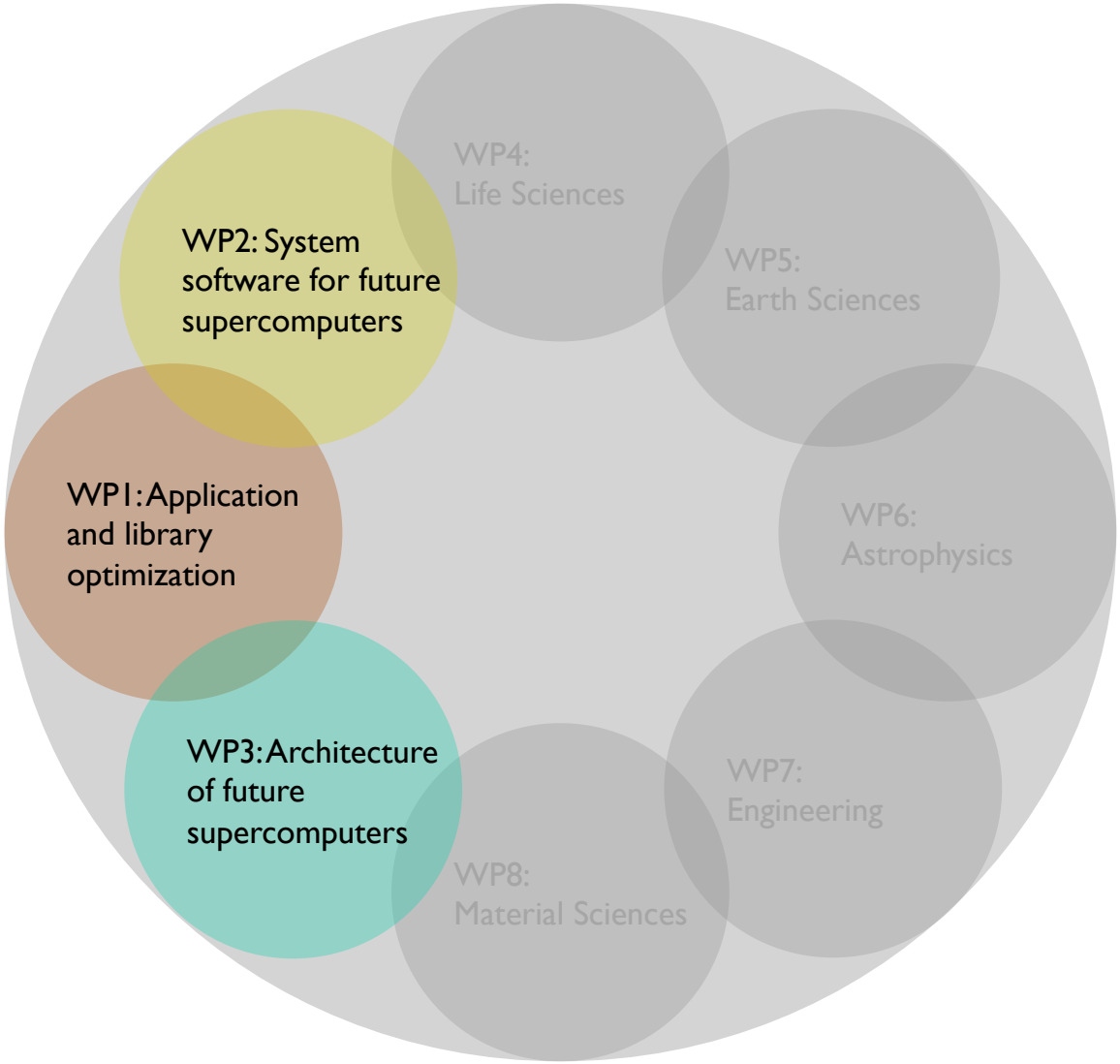
System networks

- New router and network architectures
  - twisted torus
  - Folded Clos (fat trees)
  - dragonflies
- Fault tolerance

The Exaflop computer: qualitative changes needed in several layers



WP3: conclusions



Conclusions

- Collaboration between Computer Science and Application groups resulted in important improvements
  - Needed to adapt to novel hw technologies, sometimes difficult to predict their impact in actual applications
- Impact in the evolution of programming models
  - Novel programming models: OmpSs
  - Evolution of current ones: TBB (Thread Building Blocks)
  - Efficient implementation of current ones: data movement for Chapel, automatic compilation for CUDA
  - We missed the possibility of using some of them in applications from other groups in the Consolider project (only COMPSs for workflows)
- Architectural design of future supercomputers
  - Intra-chip and intra-node interconnect and memory hierarchy
  - Inter-node interconnection network
  - We also missed the opportunity of driving architectural simulations using real applications from the Consolider project