BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

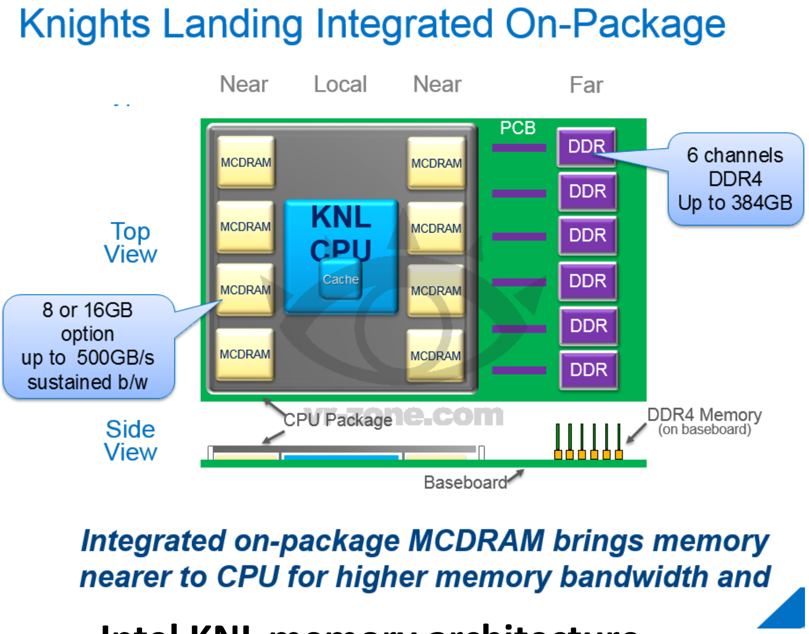
Toward Easing the Use of Optane DIMMs as Part of Heterogeneous Memory Systems

Marc Jordà, Harald Servat, and Antonio J. Peña

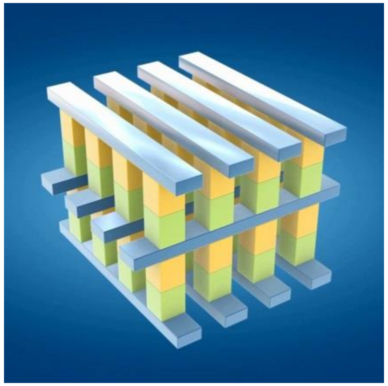
Sept. 25, 2019 IXPUG'19 Geneva, Switzerland

Motivation

- Heterogeneity in computing explored:
  - Heterogeneous processing ✓
  - Heterogeneous memory …
- Different memory technologies within cor
  - Scratchpad
    - Embedded processors
    - GPUs
  - High Bandwidth Memory (HBM)
    - Intel KNL
    - GPUs
  - (Byte-addressable) NVRAM
    - HP's "The Machine"
    - **Intel Optane**
- We expect more memory heterogeneity


Knights Landing Integrated On-Package
Near
Local
Near
Far
PCB
MCDRAM
KNL CPU
Cache
DDR
6 channels DDR4 Up to 384GB
Top View
8 or 16GB option up to 500GB/s sustained b/w
Side View
CPU Package
DDR4 Memory (on baseboard)
Baseboard
Integrated on-package MCDRAM brings memory nearer to CPU for higher memory bandwidth and

Intel KNL memory architecture

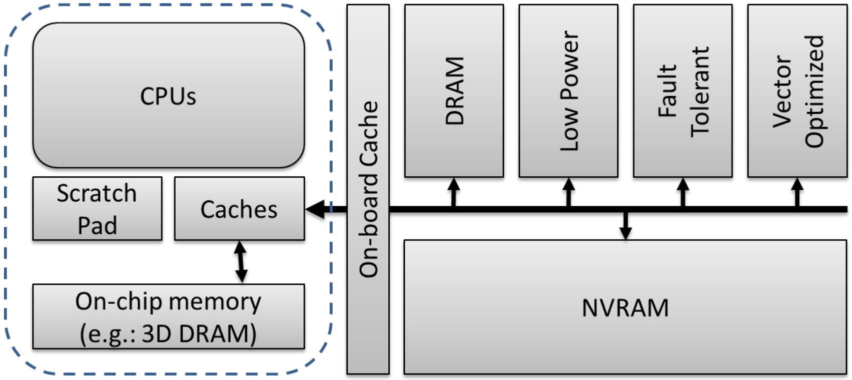


Intel 3D XPoint Technology

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

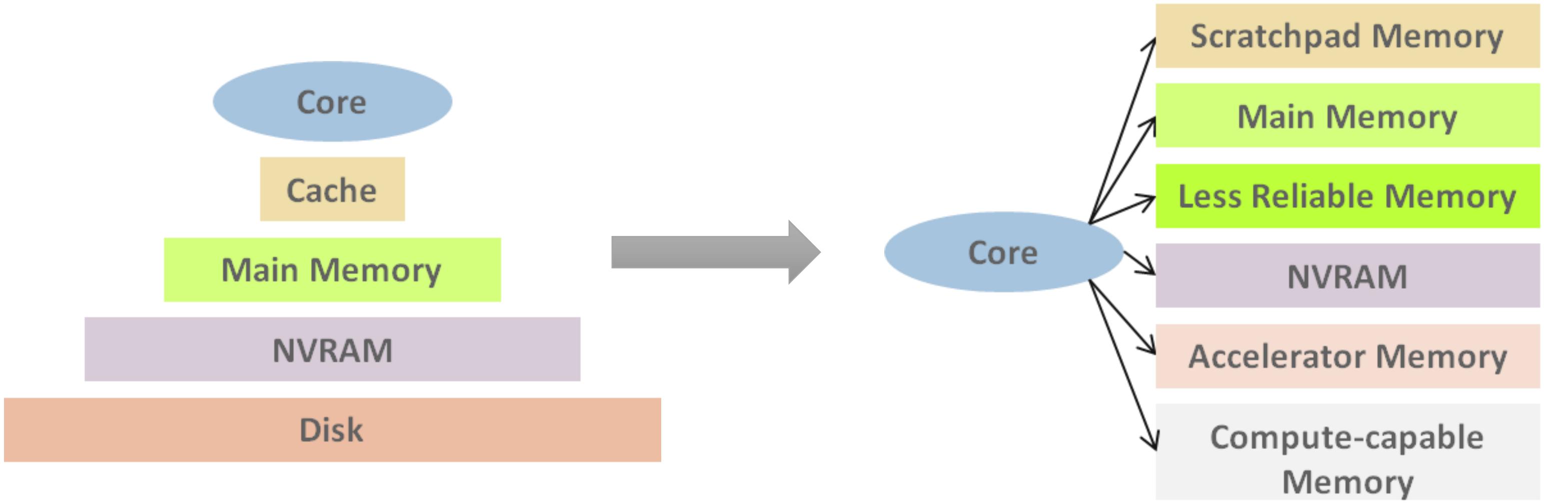
Motivation

- Different features:
  - Size, resilience, access patterns, energy, persistency…
- Examples:
  - Scratchpad:
    - Cachelike speeds, small sizes
  - Vector-specialized (e.g.: GDDR)
    - High bandwidth if cont. accesses
  - Low-power memory
    - Increased energy/speed ratio
  - ECC-enabled memory
    - Fault tolerance; speed & size ↑
  - I/O class (e.g.: NVRAM)
    - Large; reduced speeds & energy
    - Faster reading than writing



Motivation

- To efficiently exploit heterogeneous memory:
  - Bring them as first-class citizens
  - Move from hierarchical to explicitly managed

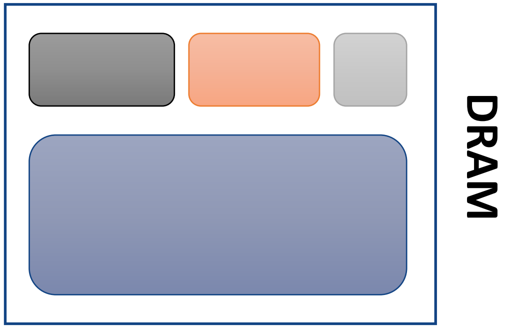


- Application's data distribution?
  - OS? Heuristics? On-the-fly monitoring? Hardware-assisted? Historic data? User hints?
  - Need ecosystem to assist users/developers: tools
    - Profilers, libraries, runtime systems

Heterogeneous Memory Systems

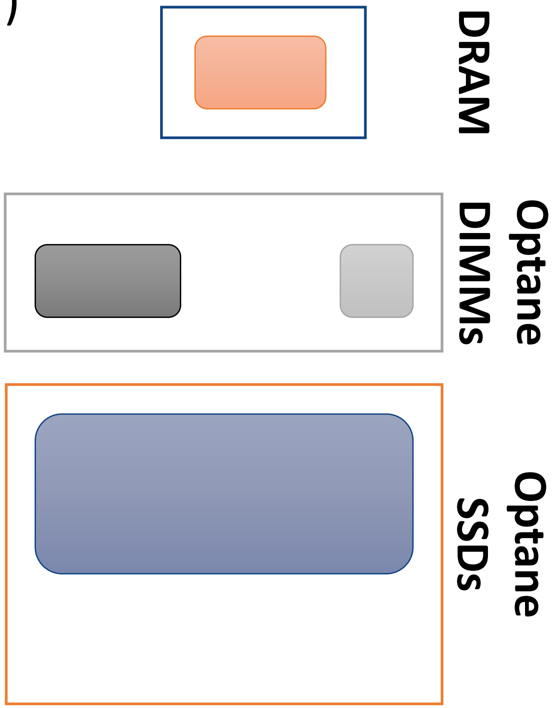
- KNL: DRAM + MCDRAM (↑ BW, ↑ Lat.) → R.I.P.
- Byte-addressable NVRAM (persistent)
  - Intel® Optane™ DC Persistent Memory (DIMM)
  - Intel® Optane™ SSD (NVMe)

- Goal: Assess optimal data distribution
  - Maximize performance
  - Minimize energy
  - …



Heterogeneous Memory Systems

- KNL: DRAM + MCDRAM (↑ BW, ↑ Lat.) → R.I.P.
- Byte-addressable NVRAM (persistent)
  - Intel® Optane™ DC Persistent Memory (DIMM)
  - Intel® Optane™ SSD (NVMe)



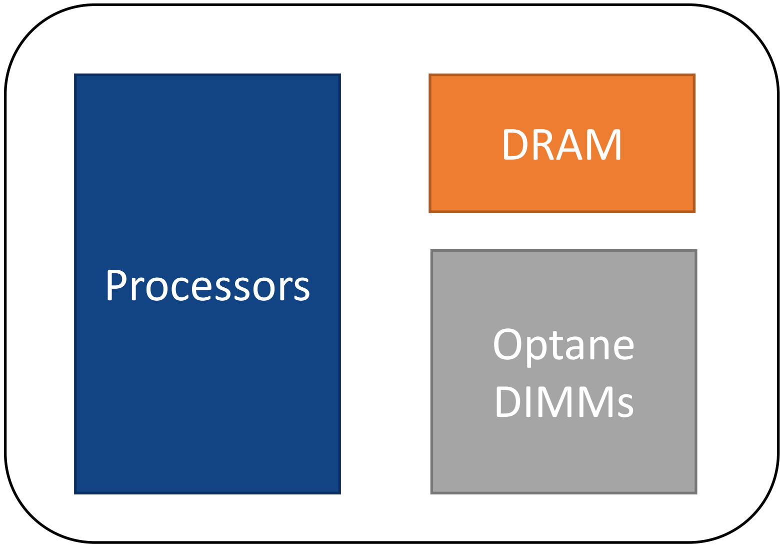
- Goal: Assess optimal data distribution
  - Maximize performance
  - Minimize energy
  - …

Heterogeneous Memory Systems

- Heterogeneous Memory Methodologies
  - Page level
    - Leverages OS's view
    - Can monitor hot vs. cold pages, # of allocations, total size, global status
    - Easy migrations
  - Object granularity (object: variable, static array, heap buffer, etc.)
    - Leverage object semantics
    - Usually same access pattern across entire object
    - User-friendly – user may hint / control

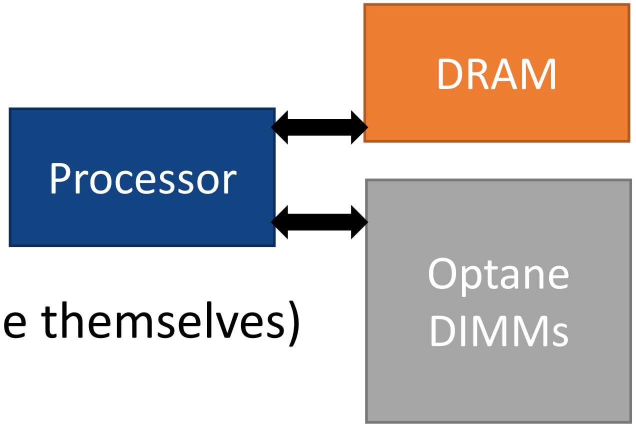
Optane DIMM Systems

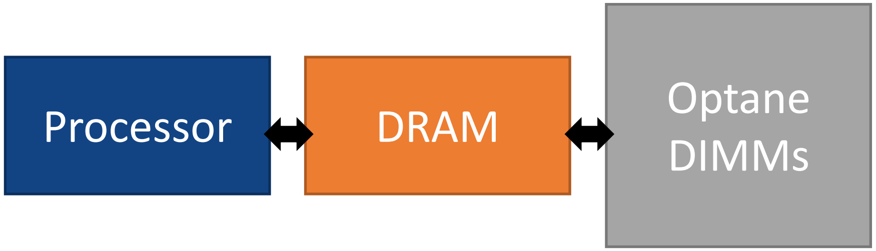
- Two levels of memory
  - Main memory (DRAM)
    - Processor has direct access to all of main memory
    - Regular DRAM latency/bandwidth
  - Intel® Optane™ DC Persistent Memory
    - Very high capacity + persistency
    - Higher latency, but much better than SSDs



Memory Modes

- App Direct Mode (Heter. Memory)
  - DRAM and Optane DIMMs are both available
  - More overall memory available
  - Software managed (applications need to handle themselves)





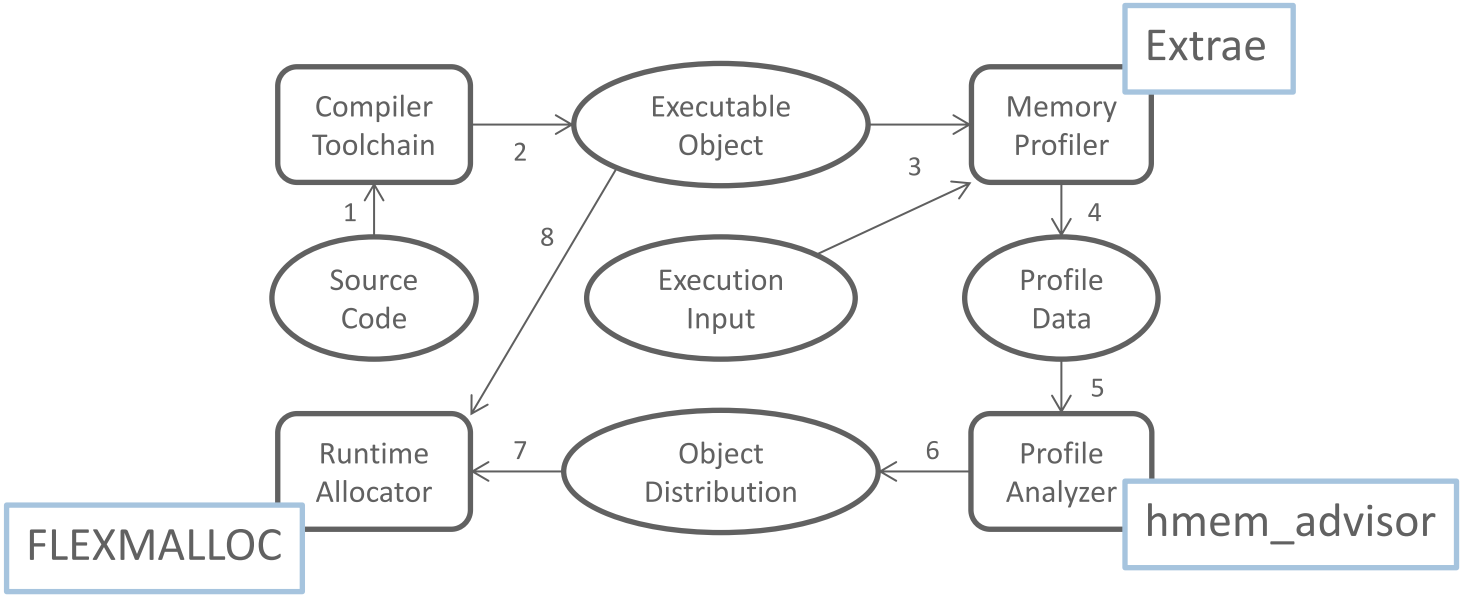
- 2LM Mode (Cache-mode)
  - DRAM as cache for Optane DIMMs
  - Only Optane DIMMs address space
  - Done in hardware (applications don’t need to be modified)

Methodology

- Object-differentiated data-oriented profiling + distribution algorithm (analysis):

  1. Profile to determine per-object last-level cache misses / avg. access time

  2. Assess the optimal distribution of the different objects among the memory subsystems
     - Minimize processor stall cycles



Evolved version of:

A. J. Peña and P. Balaji, "Toward the efficient use of multiple explicitly managed memory subsystems", IEEE Cluster 2014

Promising Early Results (KNL, loads only)

- Caveats:
  - Dynamic allocation (Lulesh)
    - Will require runtime vs. profiling
  - Lack of some HW counters
  - Stack frame allocation not managed by memkind
    - We can do some assembly to place these in different mems.

Speedup of Framework w.r.t. other approaches

Code	numactl –p 1 (MCDRAM*)	Cache Mode
miniFE	1.15x	1.27x
HPCG	1.49x	1.25x
Lulesh	1.22x	0.89x
BT	1.00x	1.00x
CGPOP	0.83x	0.85x
SNAP	0.90x	0.91x
MAXW-DGTD	1.04x	0.98x
GTC-P	1.34x	1.06x

MCDRAM*: allocate as much as it fits in HBM, FCFS

H. Servat, A. J. Peña, G. Llort, E. Mercadal, H. C. Hoppe, and J. Labarta. "Automating the application data placement in hybrid memory systems", in IEEE Cluster, Hawaii, USA, Sep. 2017.

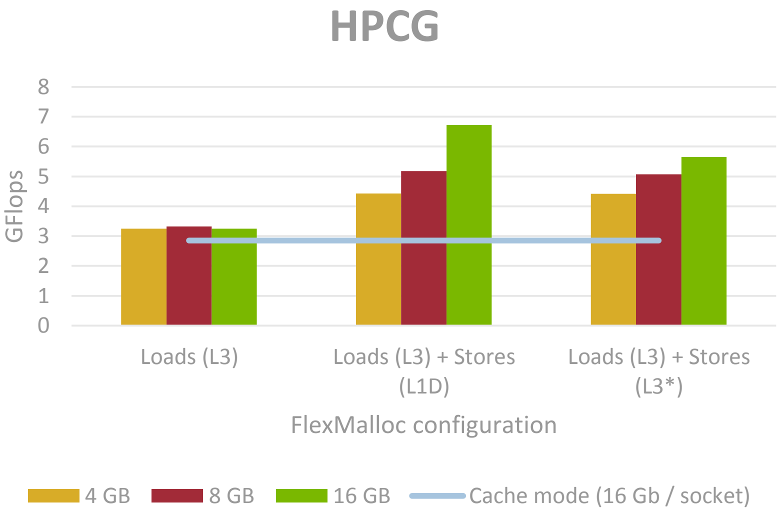
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

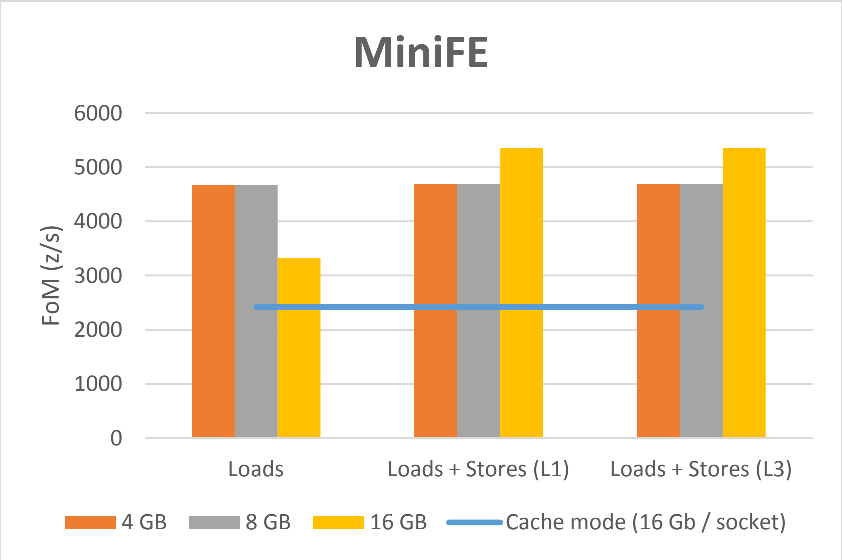
System description

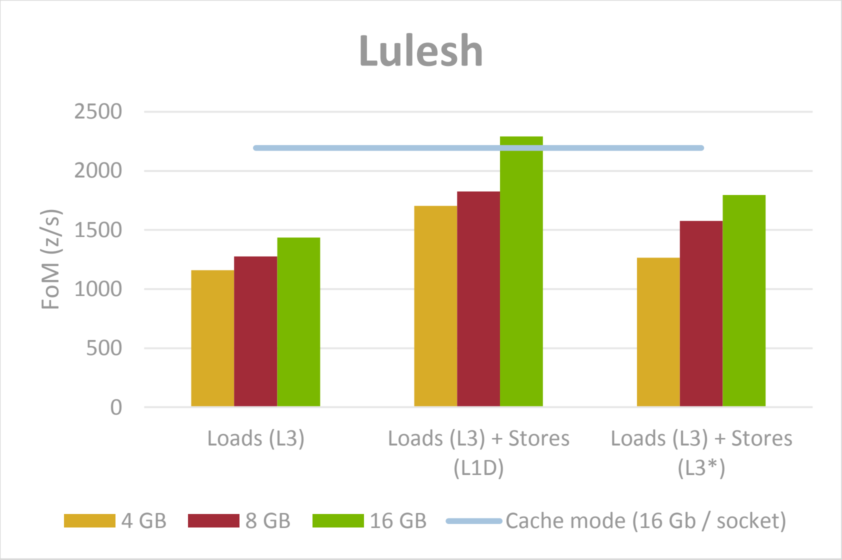
- Hardware
  - 2S – Intel Xeon Platinum 8260L CPU @ 2.30GHz (pre-qual), HT disabled
    - Only single socket executions
  - 2x 16 GB of DRAM (16 GB x socket)
    - 2 of 6 DIMM channels populated per socket – 1/3 of platform bandwidth
  - 12x 512 GB Optane™ DC Persistent Memory (3 TB x socket)
- Software stack
  - Fedora 27 (kernel 4.18.8-100.fc27.x86_64) – 2018ww40 BKC
  - Intel Compiler Suite 2019u3
  - Memkind checked out October’18

Some Results

- Quite some cases beat cache mode in fair comparison
  - E.g., MiniFE: ~100% improvement w.r.t. cache mode
    - Even ¼ RAM w.r.t cache mode
- In other cases we are within negligible performance
- And some other cases require runtime actions (next step yr 2)







Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Summary

- Heterogeneity is here for good and to stay
- Not only heterogeneous processing elements
  - Also memory and others
- Heterogeneous memory management APIs in production
  - Little help on deciding where to place data
- Research efforts on automatic/guided data distribution
- Some ongoing work ideas:
  - Runtime monitoring (migrations, reuse, get rid of previous profiling)
  - Seamless integration (no need for user intervention)
  - Improve profiling metrics
  - Integrate with other programming models (e.g., OpenMP)

Team Acknowledgements

- Muhammad Owais, former Jr. SW Engineer, BSC

- Marc Jordà, SW Engineer, BSC

- Antonio J. Peña, AccelCom group leader, BSC

- Jesús Labarta, CS Director, BSC

- Harald Servat, HPC SW Engineer, Intel

- Marie-Christine Sawley, Exascale Lab Director, Intel

Project Acknowledgements

- This work is done under the Intel-BSC Exascale Laboratory Statement of Work 5.1 on 3D Xpoint Memory Technology

- Antonio J. Peña has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie-Sklodowska Curie grant agreement No. 749516

MARIE CURIE ACTIONS

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Thank you

marc.jorda@bsc.es