Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

A Novel Tool for Data Placement in Multi-Level Memory Hierarchies

Antonio J. Peña

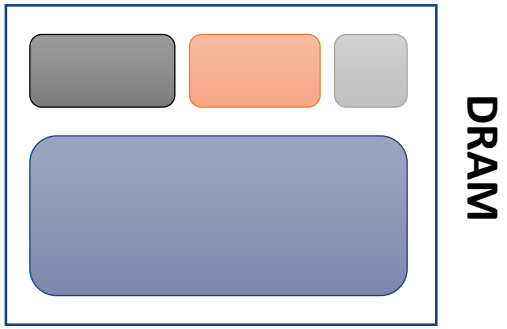
Sr. Researcher, Team Lead

Multi-Level Memory and Storage for HPC, Data Analytics & AI – ISC19 BoF

June 18, 2019

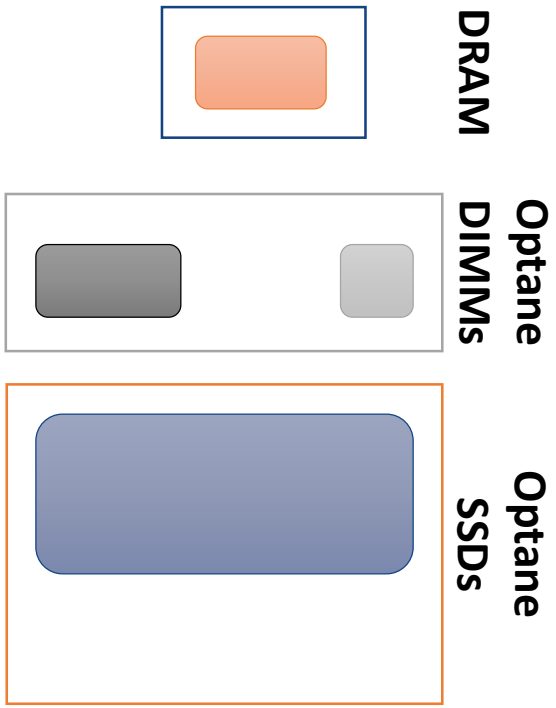
Heterogeneous Memory Systems

- KNL: DRAM + MCDRAM (↑ BW, ↑ Lat.) → R.I.P.
- Byte-addressable NVRAM (persistent)
  - Intel's 3D XPoint
    - Optane DIMMs
    - Optane SSDs
- Goal: Assess optimal data distribution
  - Maximize performance
  - Minimize energy
  - …



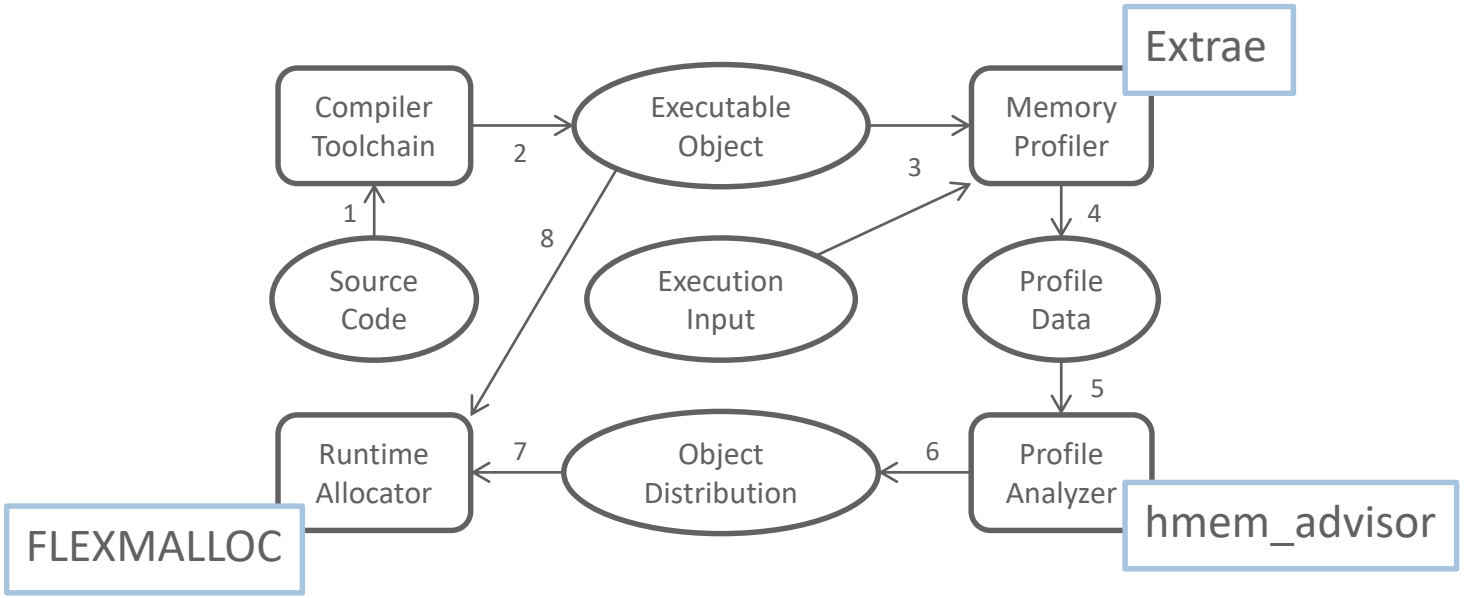
Heterogeneous Memory Systems

- KNL: DRAM + MCDRAM (↑ BW, ↑ Lat.) → R.I.P.
- Byte-addressable NVRAM (persistent)
  - Intel's 3D XPoint
    - Optane DIMMs
    - Optane SSDs
- Goal: Assess optimal data distribution
  - Maximize performance
  - Minimize energy
  - …



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.

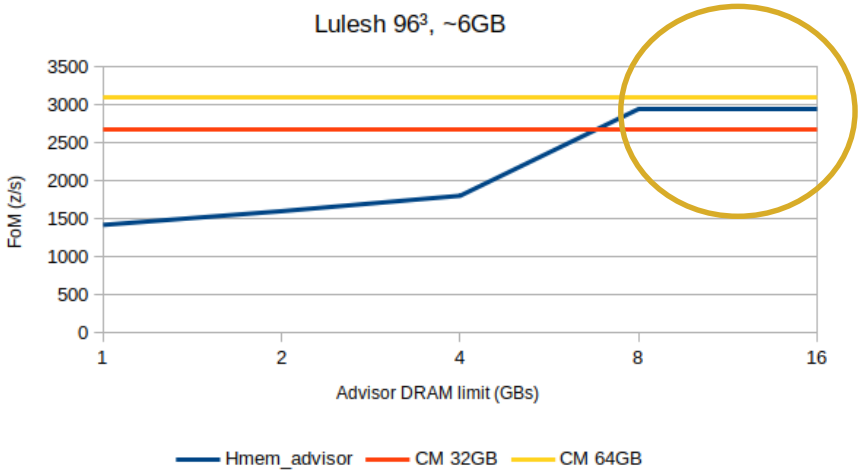
Speedup of Framework w.r.t. other approaches

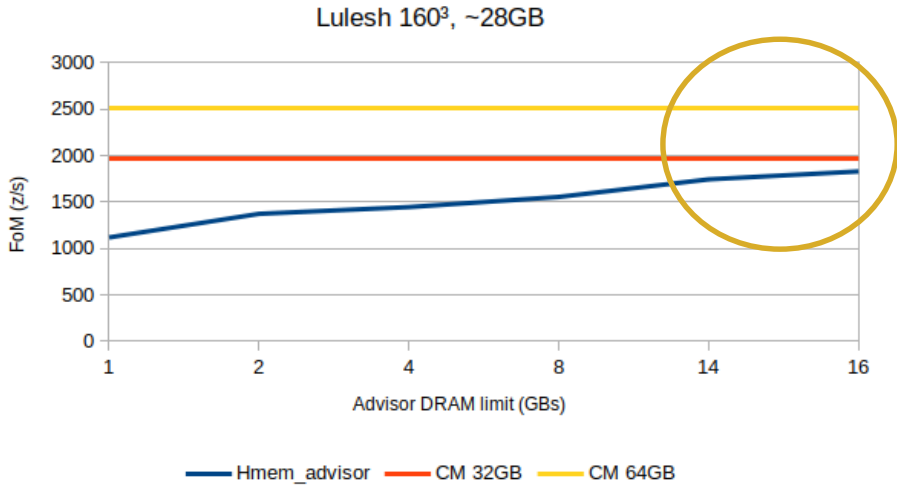
Code	numactl –p 1 (MCDRAM*)	Cache Mode
miniFE	1.15x	1.27x
HPCG	1.49x	1.25x
Lulesh	1.22x	0.89x
BT	1.00x	1.00x
CGPOP	0.83x	0.85x
SNAP	0.90x	0.91x
MAXW-DGTD	1.04x	0.98x
GTC-P	1.34x	1.06x

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

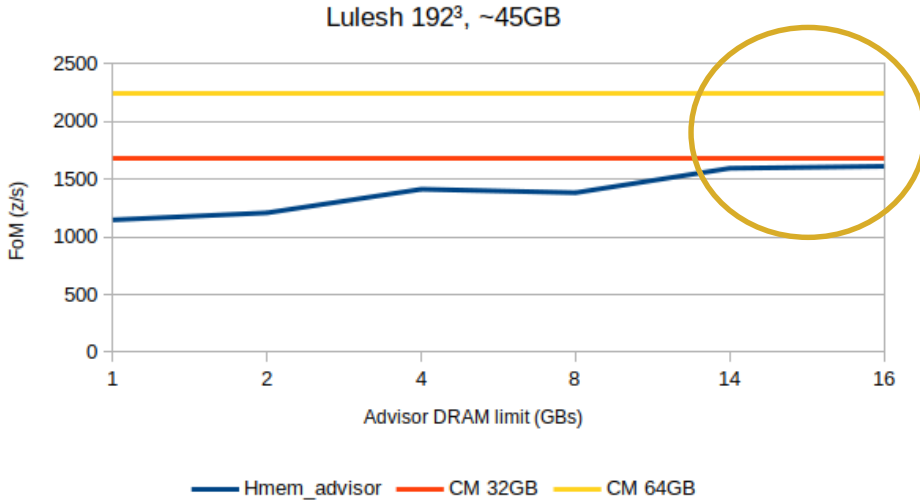
H. Servat, A. J. Peña, G. Llort, E. Mercadal, H. C. Hoppe, and J. Labarta. “Automating the application data placement in hybrid memory systems”, in IEEE Cluster, Hawaii, USA, Sep. 2017.

Early Optane Results





- Executions are pinned to one socket, actual cache size of cache-mode (CM) is half of the total DRAM
- Advisor placement for 16GB is close to CM with 16GB cache for 1603 and 1923 datasets, but not better
- Issue with other apps: executions with large datasets are using more DRAM than expected
- hmem_advisor distributions seem to comply with the DRAM limit
- We are investigating the reasons



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, Jr. SW Engineer, BSC

- Marc Jordà, SW Engineer, 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

- The speaker's research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie-Sklodowska Curie grant agreement No. 749516

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

antonio.pena@bsc.es