www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Large-Memory Nodes for Energy Efficient High-Performance Computing

Darko Zivanovic, Milan Radulovic, Germán Llort, David Zaragoza, Janko Strassburg, Paul M. Carpenter, Petar Radojković, Eduard Ayguadé

The International Symposium on Memory Systems (MEMSYS16),
3-6 October 2016

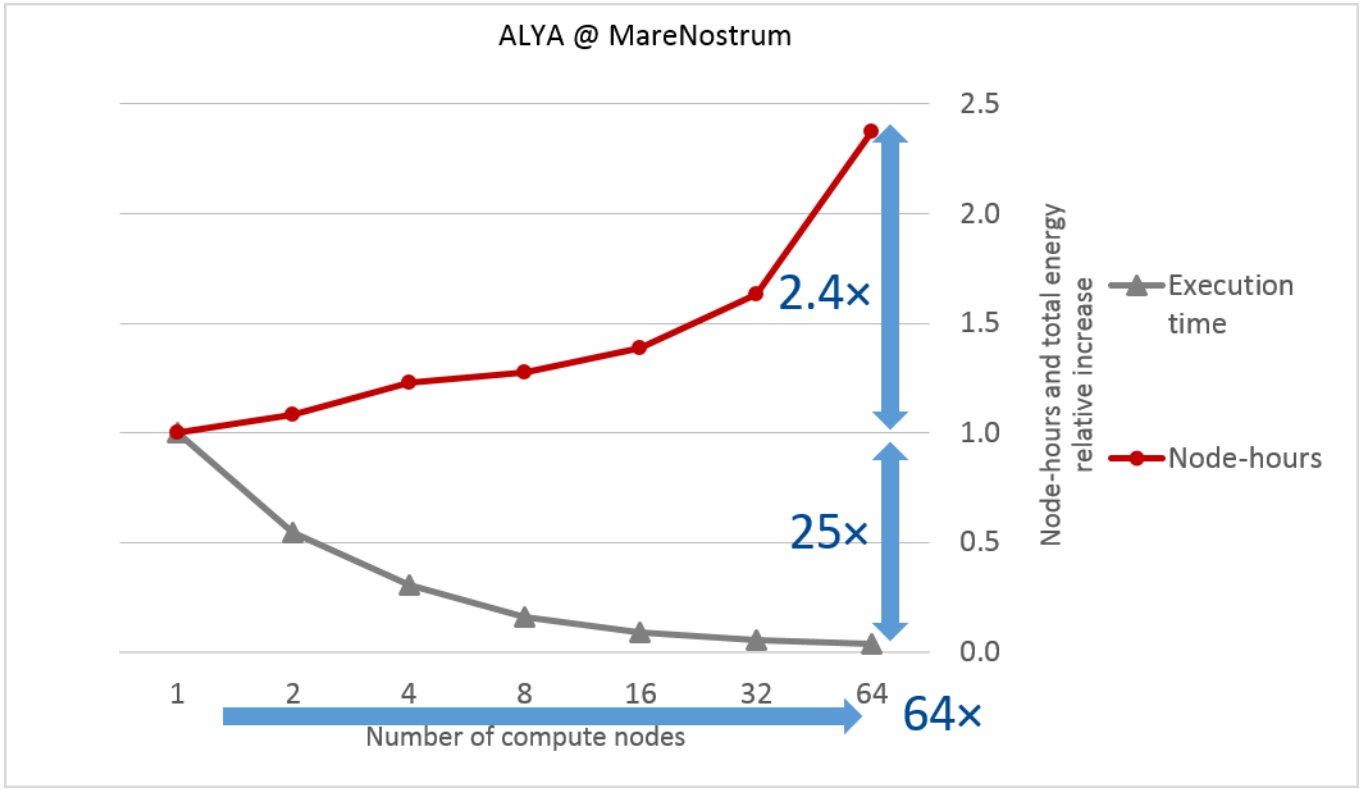
High-Performance Computing (HPC) --- today

- Traditional HPC = Performance (for a given cost)
- Current HPC = Performance for a given energy (for a given cost)

- Our work:
  1. Understand term “performance” in HPC
  2. Quantify tradeoff between “performance” and energy
  3. Memory sizing → changes the performance-energy tradeoff

High-Performance Computing (HPC)
Scale-out to reduce the execution time

- Not perfect – Amdahl's law, communication overheads
- Not for free – Increases the *cost* of the experiment
  - *node-hours* = #nodes × #hours (compute node = server)

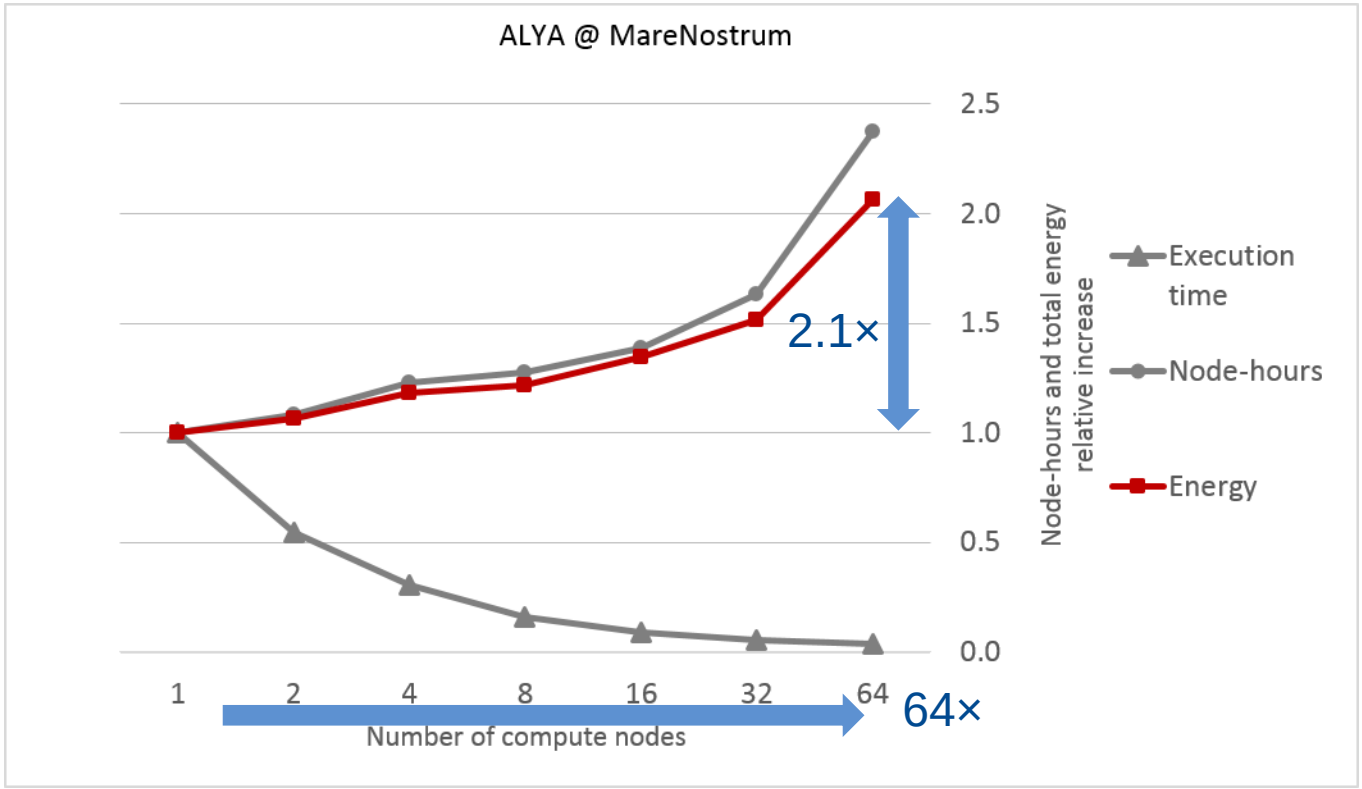


- 64× servers
- 25× faster
- 2.4× more expensive

ALYA – Computation mechanics - Unified European Application Benchmarks Suite (UEABS)
MN node: dual-socket 2 x 8-core Sandy-Bridge processor with 2GB of main memory per core (32 GB per node)

Energy consumption

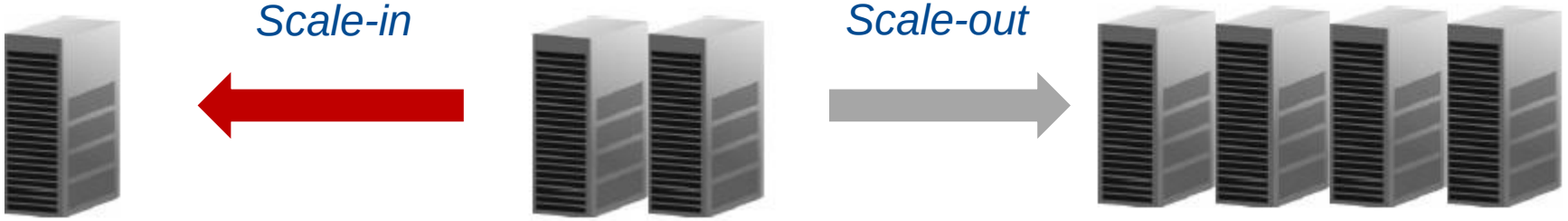
- Strongly correlated with *node-hours*
- Significant energy overheads:



- 2.1× for ALYA, 1→ 64 servers 16→1024 cores
- 60% on average for UEABS applications

Scale-out *vs.* Scale-in

Energy efficient HPC:
Scale-in

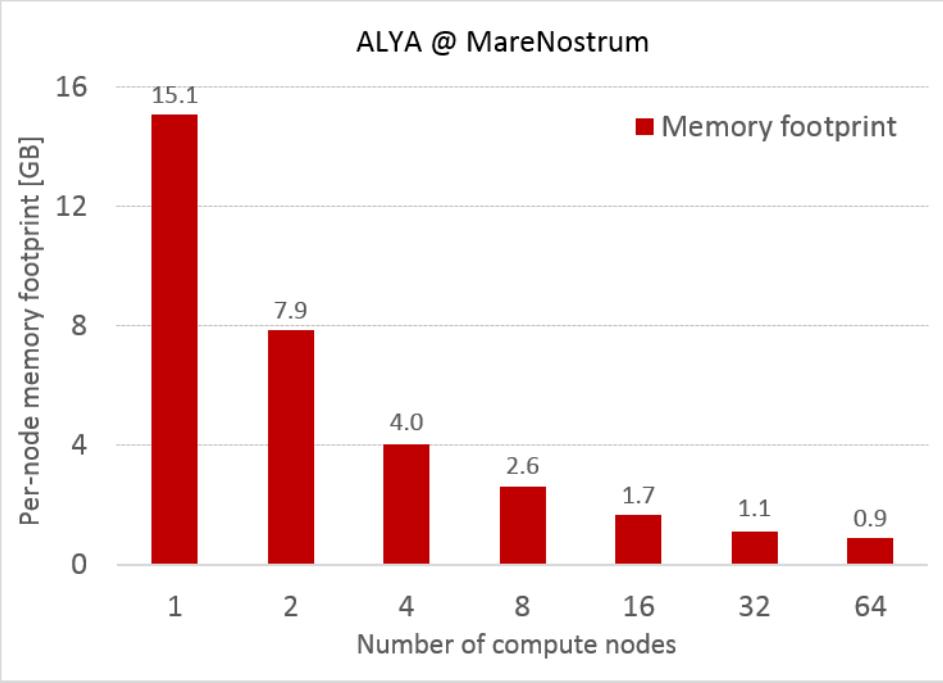


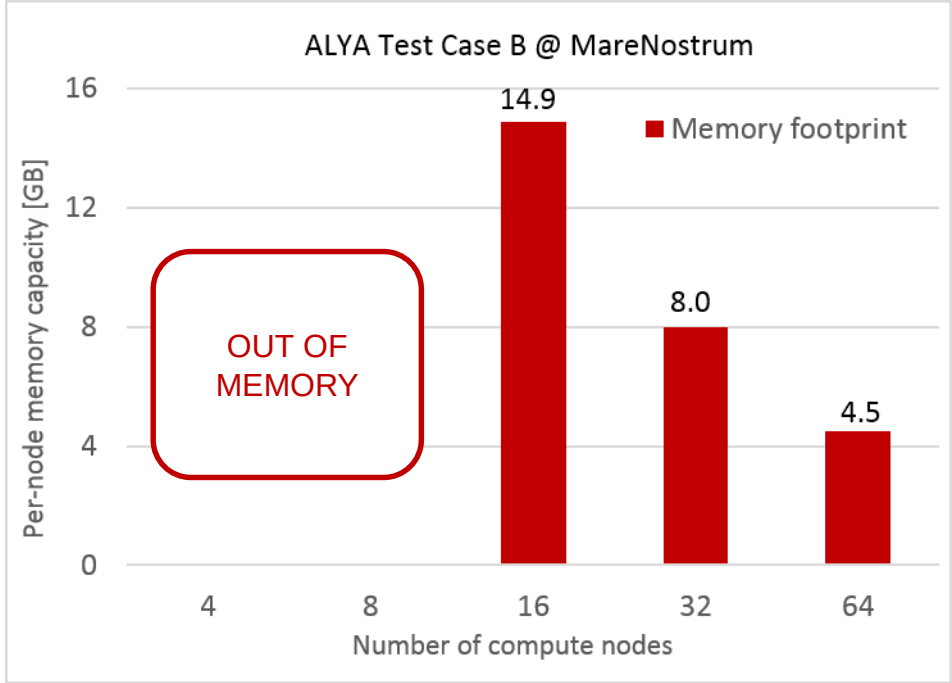
Traditional HPC:
Scale-out

60% energy overheads
on average

Scale-in → Hitting the memory capacity wall

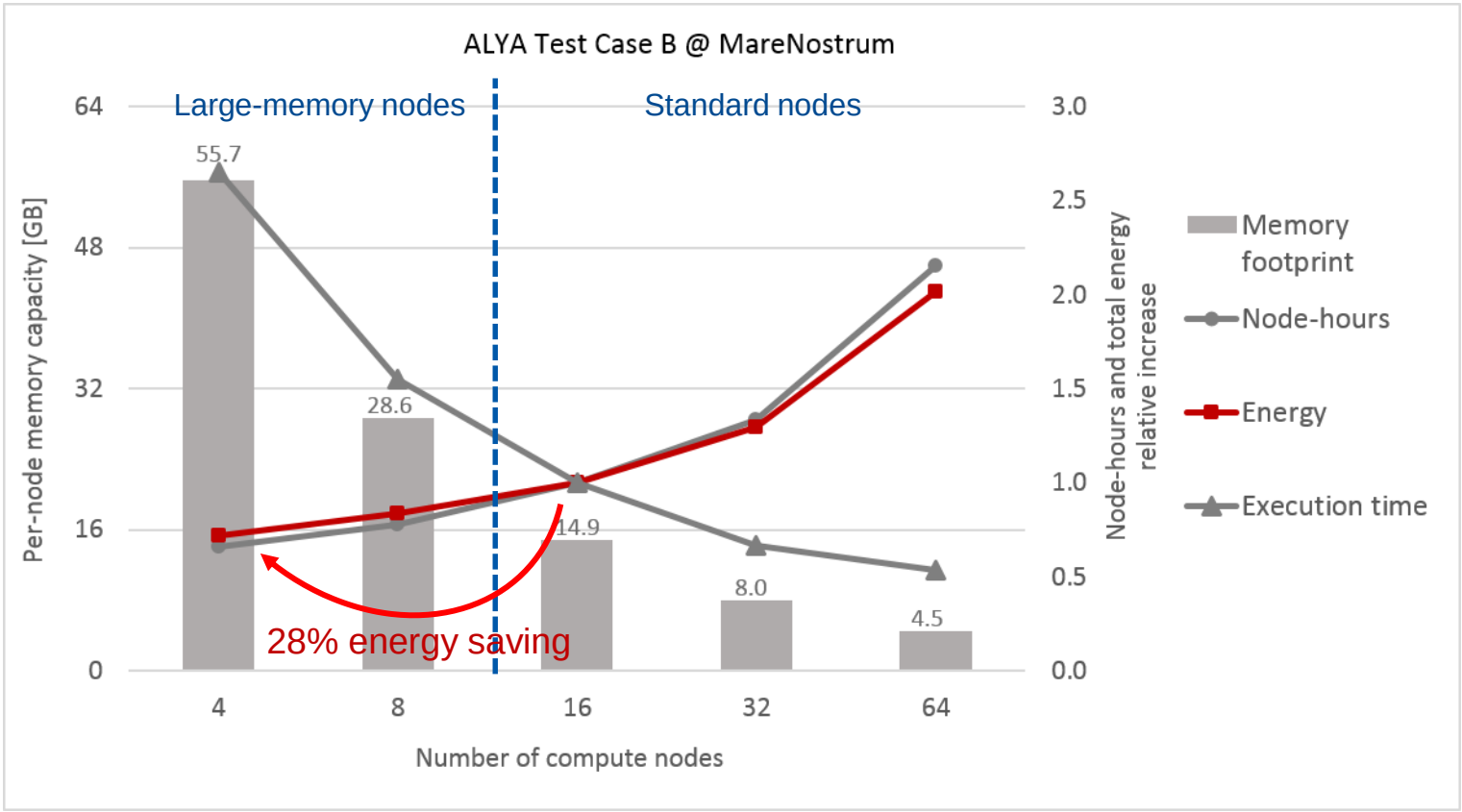
- Significant energy and node-hours savings
- Until we hit the memory capacity wall
  - → Higher per-node memory footprint
    - until the point when the dataset does not fit into available memory





What if we buy more memory?

- Large-memory MareNostrum nodes (32 → 128GB per node)
  - Additional energy savings
    - 28% for ALYA
    - 36% on average



Are we cheating?

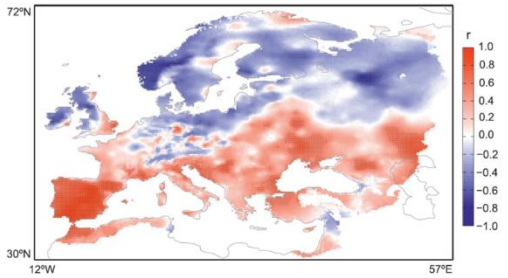
- Execution time

- Weather forecast for tomorrow
  - 64 servers: 5h
  - 1 server:
    - 60% lower node-hours
    - 50% energy savings
    - 5 days
    - But I need if before the 8pm news!

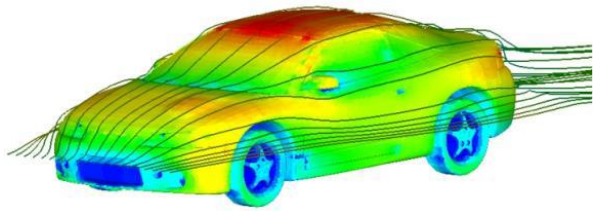
Not all HPC is a weather forecast

Two traditional HPC categories are distinguished:*

[1] Capability computing refers to the use of a large and high-performing computing infrastructure to solve a single, highly complex problem in the *shortest possible time*



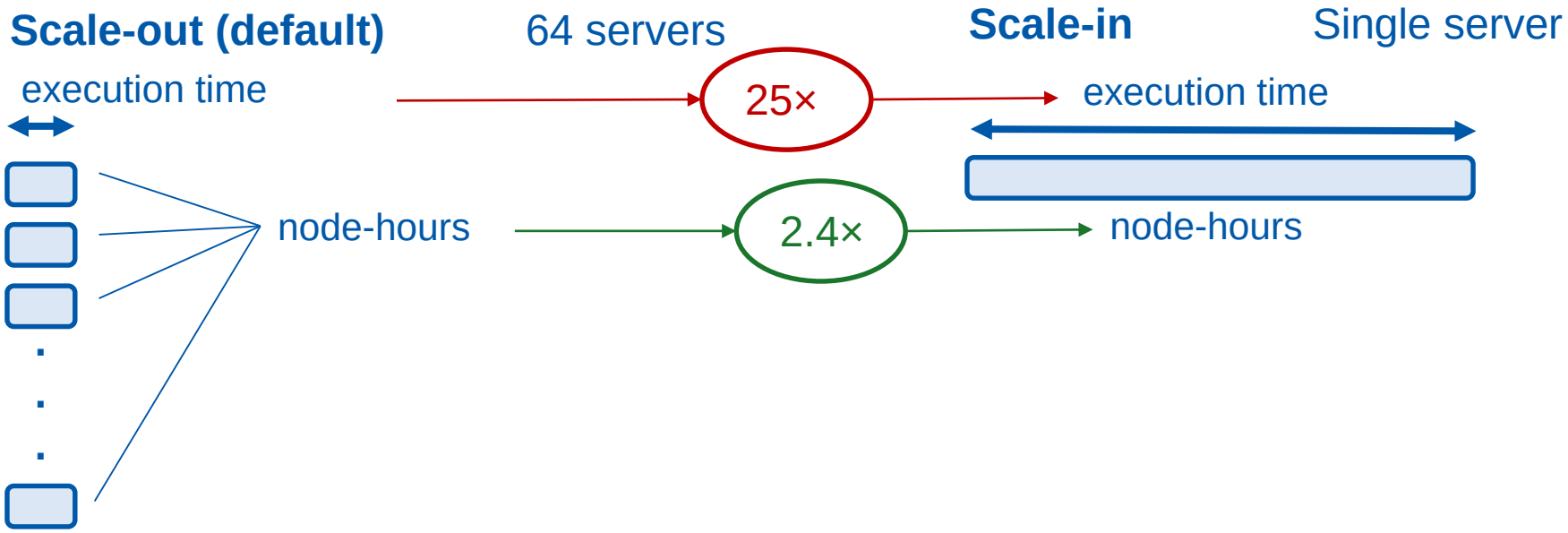
[2] Capacity computing refers to optimizing the efficiency of using a compute system to solve *as many mid-sized or smaller problems as possible at the same time* at *the lowest possible cost*



* ETP4HPC Strategic Research Agenda
Achieving HPC leadership in Europe
http://www.etp4hpc.eu/strategy/strategic-research-agenda/

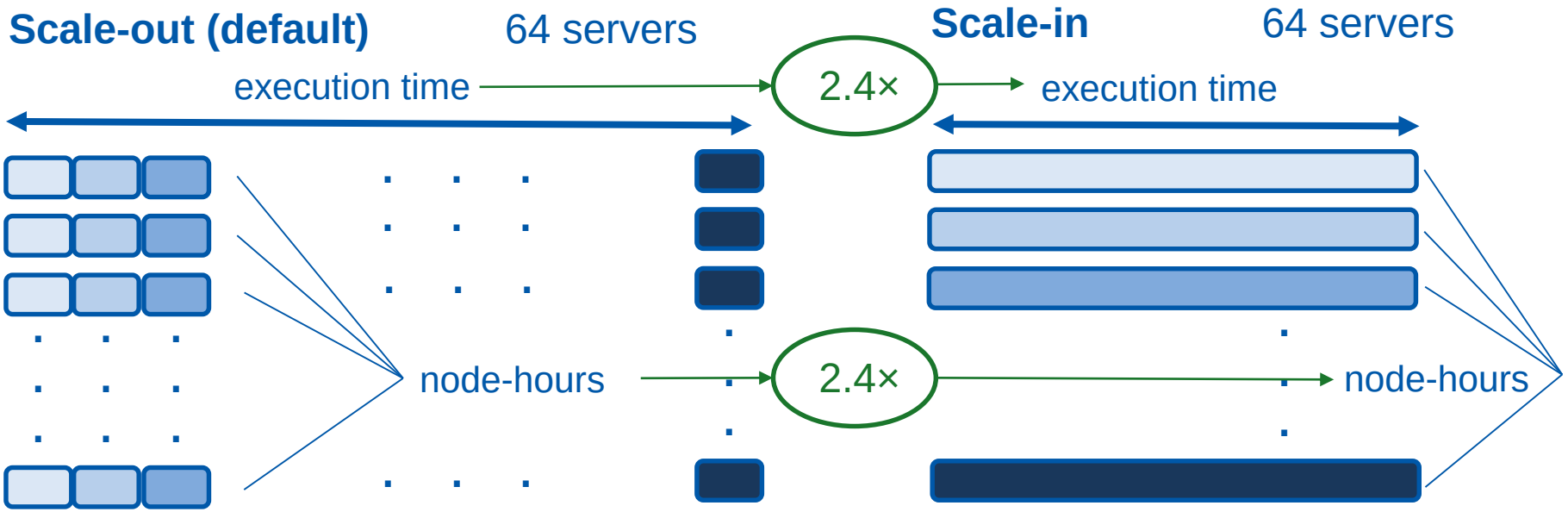
Scale-out *vs.* scale-in
Single job

- Single job – ALYA
- Scale-in improves node-hours & energy, but increases execution time



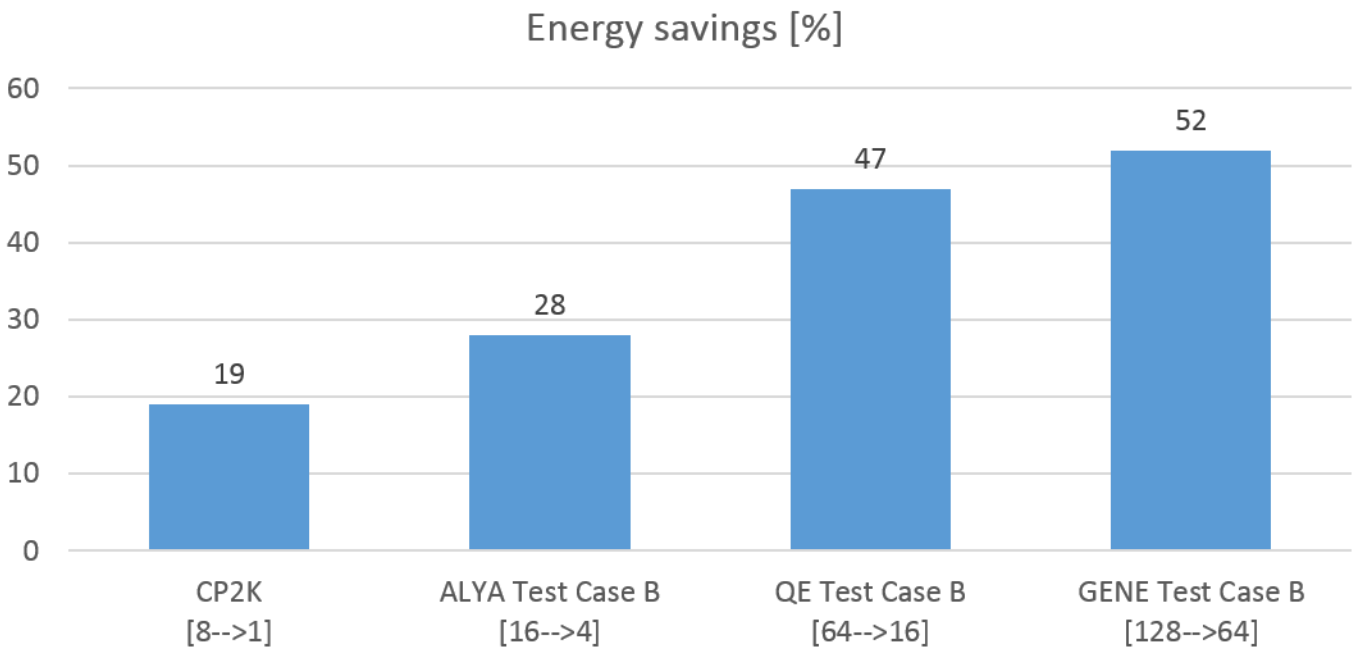
Capacity computing
One job → Batch of jobs

- Batch of jobs - ALYA, 64 jobs
- Scale-in improves node-hours & energy & **execution time**



Capacity computing
Large-memory nodes → Batch of jobs

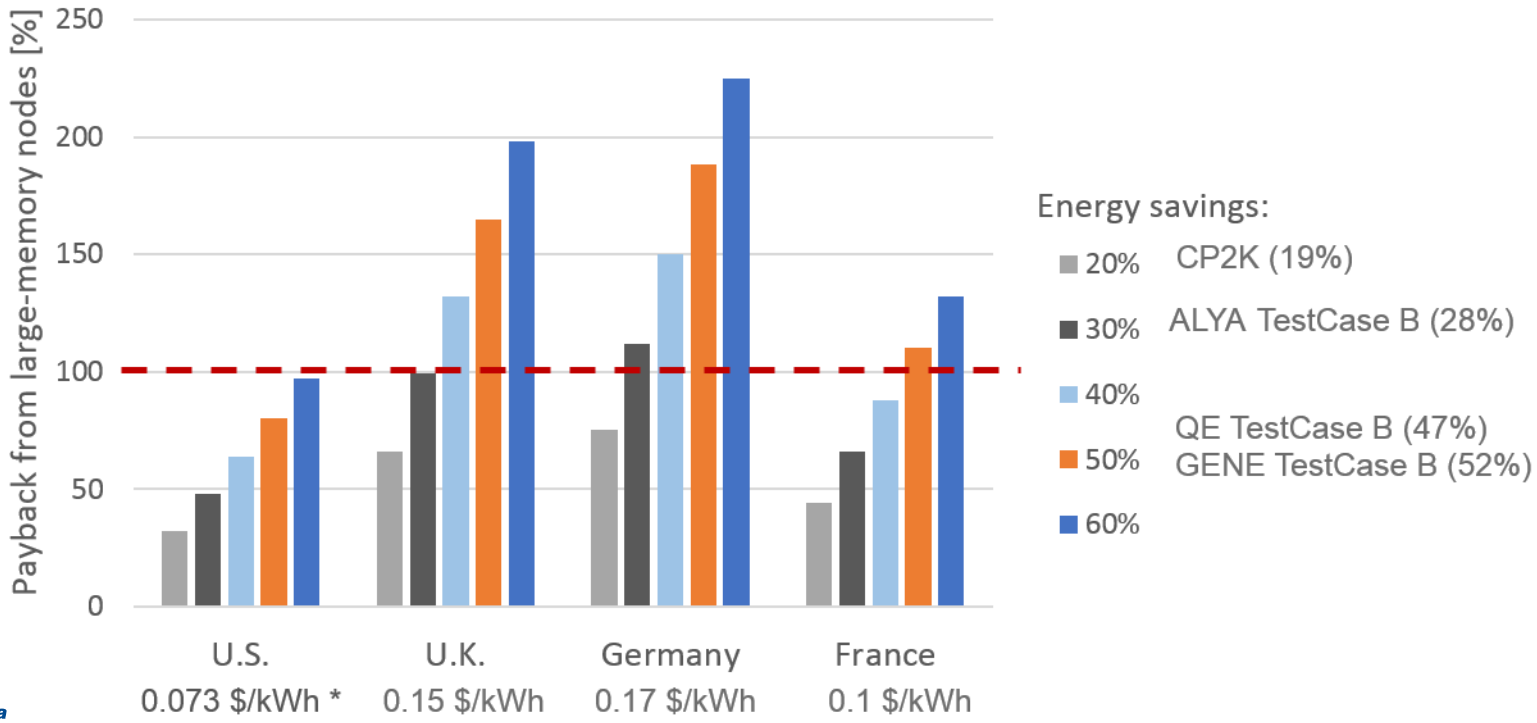
- Average benefits for batches of jobs:
  - Energy: 36%
  - Node-hours: 40%
  - Execution time for job batches: 40%



Are we cheating, again?

- Large memory nodes are more expensive
  - *Example: 4GB DIMM → 16 GB DIMM ⇒ ~ 2× more expensive*
  - *10% server cost is main memory*

- Energy savings ($) *vs*. Memory upgrade cost ($)



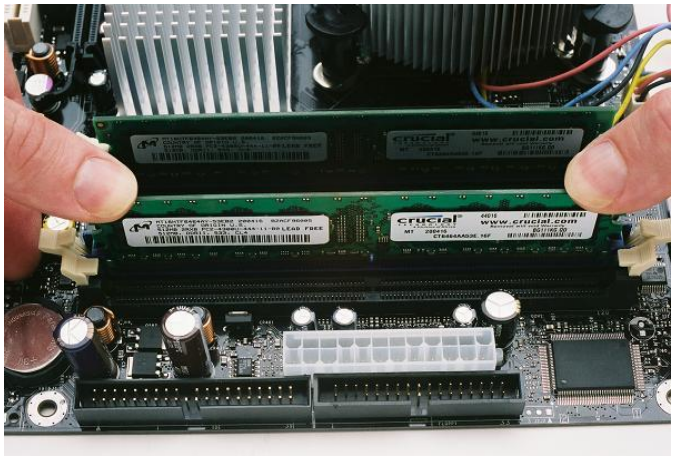
* Average industry price, publically available

Large-memory nodes for energy-efficient HPC
Summary

- Scale-out *vs.* scale-in in HPC
  - Execution time *vs.* Cost *vs.* Energy

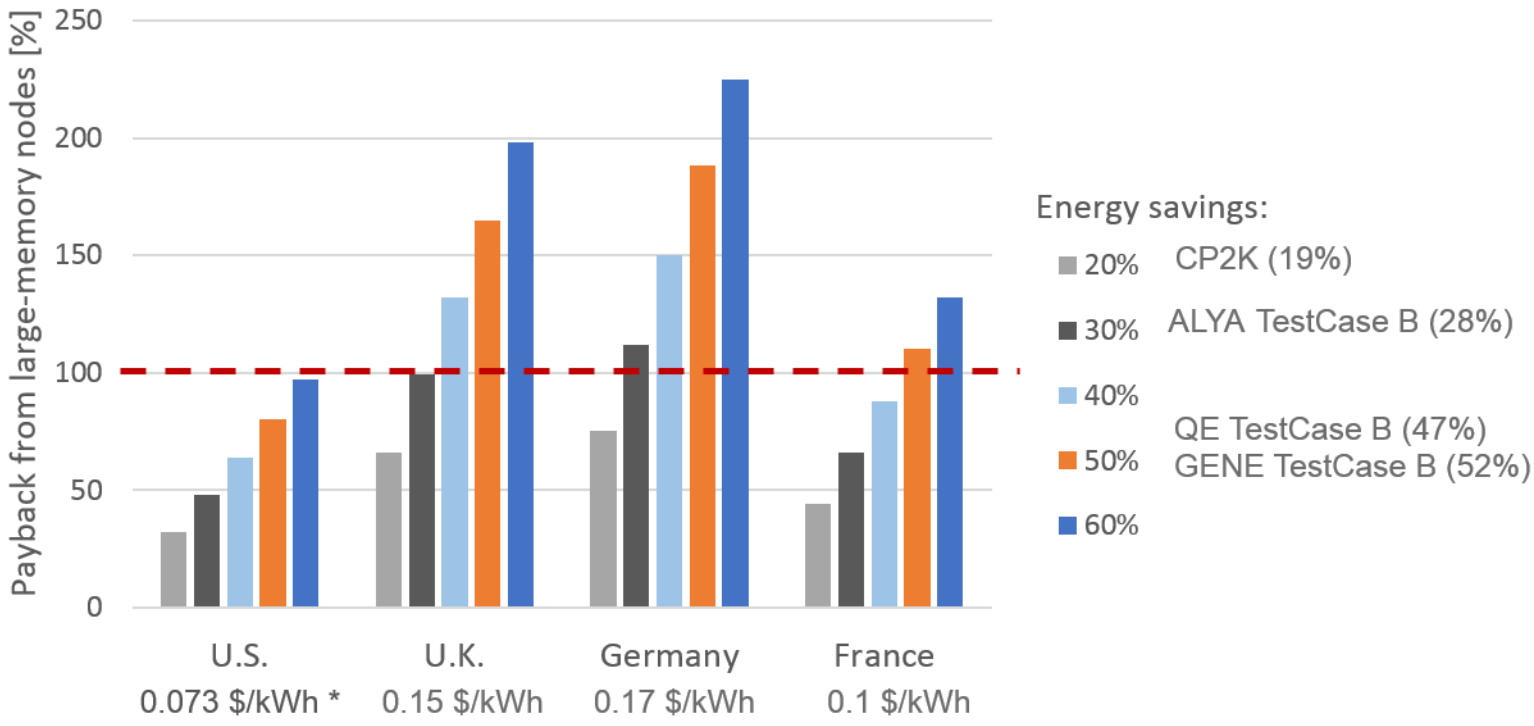
- Large-memory nodes → Further scale-in → Savings
  - Energy consumption – 36%
  - Experimentation cost – 40 %
  - Execution time for job batches – 40 %
  - System throughput – 40 %

- Just do it!
  - Plug & Play
  - Legacy codes proof



Questions?

- Large-memory nodes for energy-efficient HPC



* Average industry price, publically available

Large-Memory Nodes for Energy Efficient High-Performance Computing

THANK YOU!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Large-Memory Nodes for Energy Efficient High-Performance Computing

Darko Zivanovic, Milan Radulovic, Germán Llort, David Zaragoza, Janko Strassburg, Paul M. Carpenter, Petar Radojković, Eduard Ayguadé

The International Symposium on Memory Systems (MEMSYS16),
3-6 October 2016