Dynamic Load Balancing in OpenFOAM

Roberto Ribeiro
University of Minho

Context

CFD + HPC

CFD needs computing power

the more the best

HPC systems can provide it

In particular, clusters (distributed memory systems) that are:

- **Easily extensible**
- **Cost-effective**

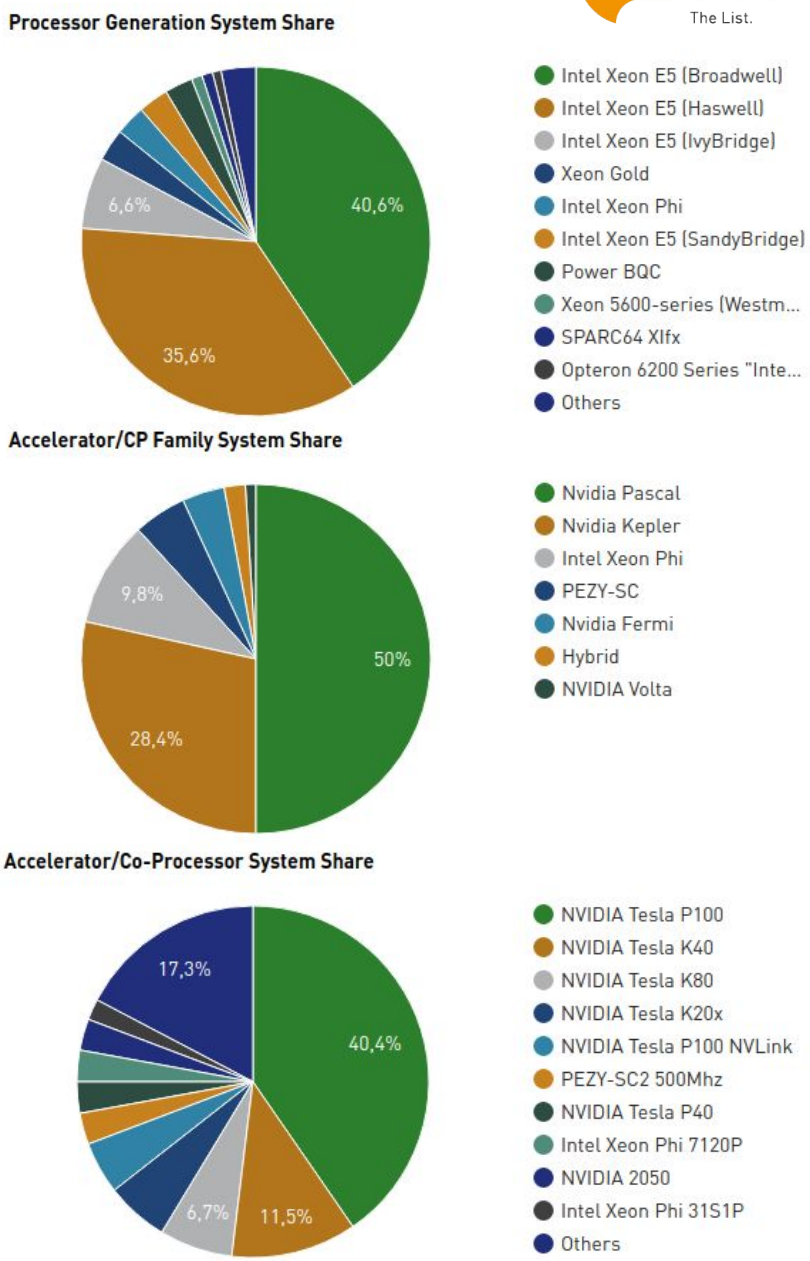
Context

Heterogeneous systems

Clusters are typically extended with new nodes from newer generations

There is also a plurality of computing devices

Systems are rendered highly heterogeneous



Top500 List Statistics November 2017

Heterogeneous Computing Era

Modern parallel computing systems are composed by a plurality of computing units from different generations and exhibiting different architectures and execution models.

Motivation

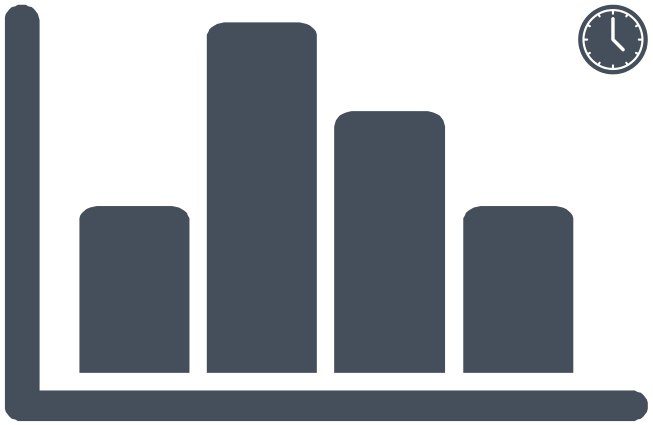
Challenges

Performance imbalances

Faster nodes wait for slower nodes

Resource idling

Results in overall resource underutilization



Motivation

Dynamic workloads

Dynamic workloads

e.g. Adaptive Mesh Refinement (dynamicRefineFvMesh)

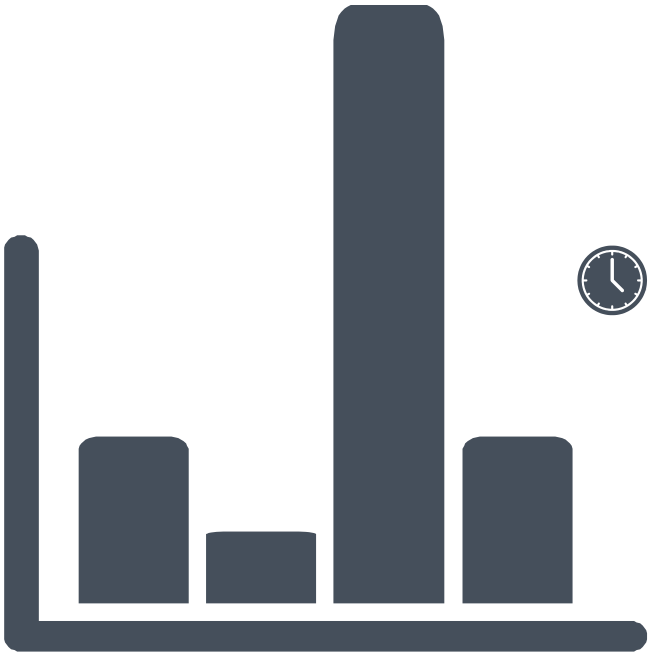
Cells are divided or merged in runtime

Depends on flow and other physical properties

Therefore, workload is dynamic and unpredictable

Motivation

More challenging with dynamic workloads



More imbalance

More resource idling

More resource underutilization

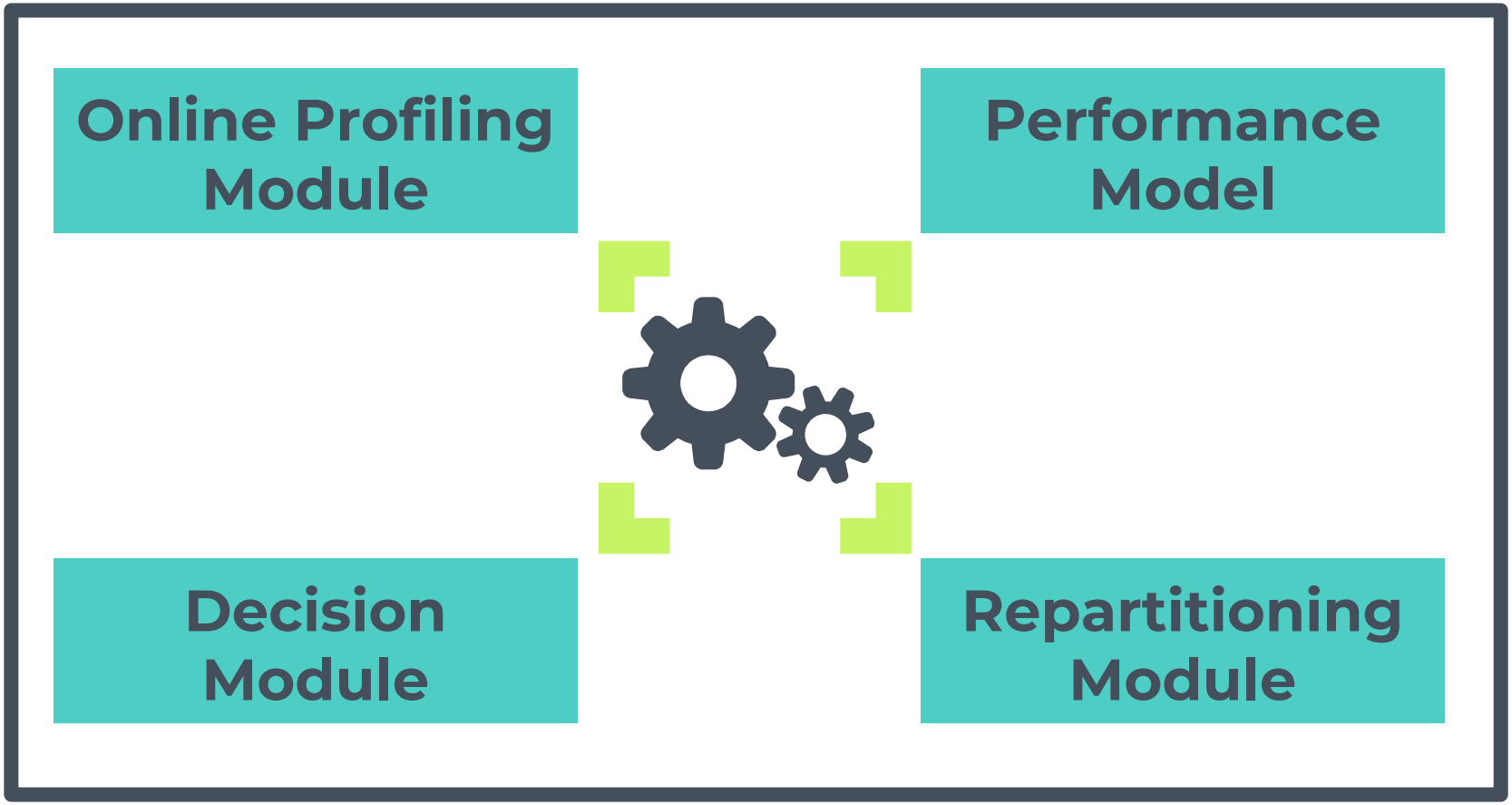
This time, unpredictable and steamed from a far more complex code/execution

Two-fold challenge

Heterogeneous Systems + **Dynamic workload**

How do we propose to address it

Heterogeneity-aware Dynamic Load Balancing



Online Profiling Module (OPM)

- OpenFOAM Instrumentation
- Low-level and relevant to execution time routines
- Separation/sieving of computation vs communication
- Information provided to PM

Performance Model (PM)

- Per CU performance characterization
- Defines a FV cell as work unit
- Basically, estimates the time required to process a cell for each CU
- Enables execution-time estimation of arbitrary workloads for each CU
- Information provided to DM

Decision Module (DM)

- Triggers re-balance based on the compute time (OPM) Relative Standard Deviation across CUs
- Linear equation system to determine a balanced distribution based on current load and PM info
- Requests RM re-distribution candidates
- Estimate re-distribution benefit based on migration cost (LR), iterations left and time gain (PM)
- Choose best redistribution and if beneficial, trigger migration

Repartitioning Module (RM)

- Partitioner interfaced as a 3rd party tool -- in this case ParMETIS
- Uses part of OpenFOAM ParMETIS routines plus newly introduced ones to support refined meshes
- Benefits from ParMETIS partitioning features:
  - Balanced **re-**distribution based on performance weights from DM
  - Boundary minimization
- Multiple decompositions requested to partitioner (learning process converging to one decomposition requested)

Results

Evaluation systems

System	SeARCH			Stampede2
Nodes	Tag **641** - Ivy Bridge E5-2650v2 @ 2.60GHz, **16** cores p/node			Tag **KNL7250** - Intel Xeon Phi 7250 @ 1.4GHz ("Knights Landing"), **68** cores p/ node
	Tag **662** - Ivy Bridge E5-2695v2@ 2.40GHz, **24** cores p/node			
	Tag **421** - Nehalem E5520 @ 2.27GHz, **8** cores p/node			
	Tag **KNL7210** - Intel Xeon Phi 7210 @ 1.3GHz, **64** cores p/ node			
Multi-node configurations	**Homogeneous**	**Heterogeneous I**	**Heterogeneous II**	**Homogeneous**
	Multiple 641's	Pair(s) of 641+421	Pair 662+KNL7210	Multiple KNL7250's
Network	Myrinet (**myri**)	Myrinet (**myri**)	Ethernet(**eth**)	Intel Omni-Path (**OPA**)

damBreak

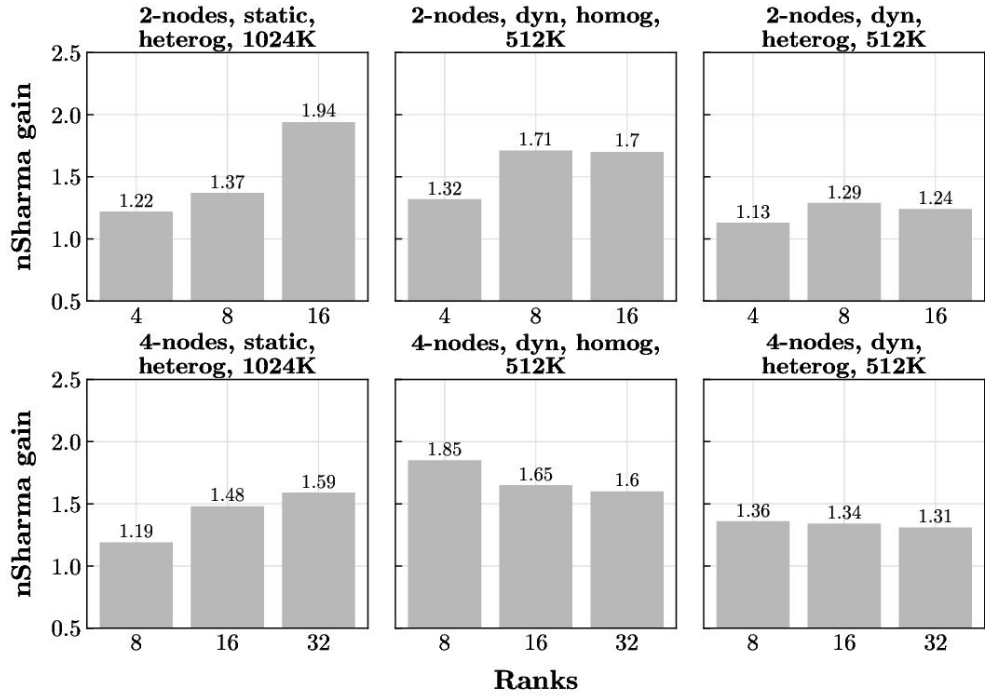
interDyMFoam

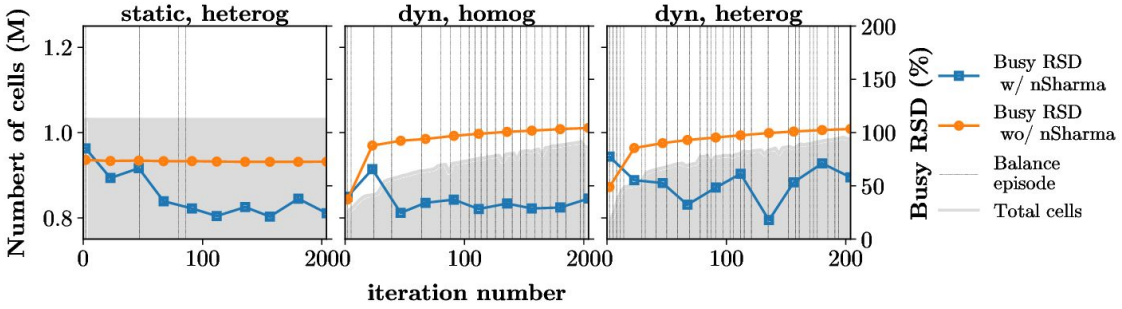
dynamicRefineFvMesh

Results

System	SeARCH			Stampede2
Nodes	Tag **641** - Ivy Bridge E5-2650v2 @ 2.60GHz, **16** cores p/node			Tag **KNL7250** - Intel Xeon Phi 7250 @ 1.4GHz ("Knights Landing"), **68** cores p/ node
	Tag **662** - Ivy Bridge E5-2695v2@ 2.40GHz, **24** cores p/node			
	Tag **421** - Nehalem E5520 @ 2.27GHz, **8** cores p/node			
	Tag **KNL7210** - Intel Xeon Phi 7210 @ 1.3GHz, **64** cores p/ node			
Multi-node configurations	**Homogeneous**	**Heterogeneous I**	**Heterogeneous II**	**Homogeneous**
	Multiple 641's	Pair(s) of 641+421	Pair 662+KNL7210	Multiple KNL7250's
Network	Myrinet (**myri**)	Myrinet (**myri**)	Ethernet(**eth**)	Intel Omni-Path (**OPA**)

SeARCH Homogeneous and Heterogeneous I configurations

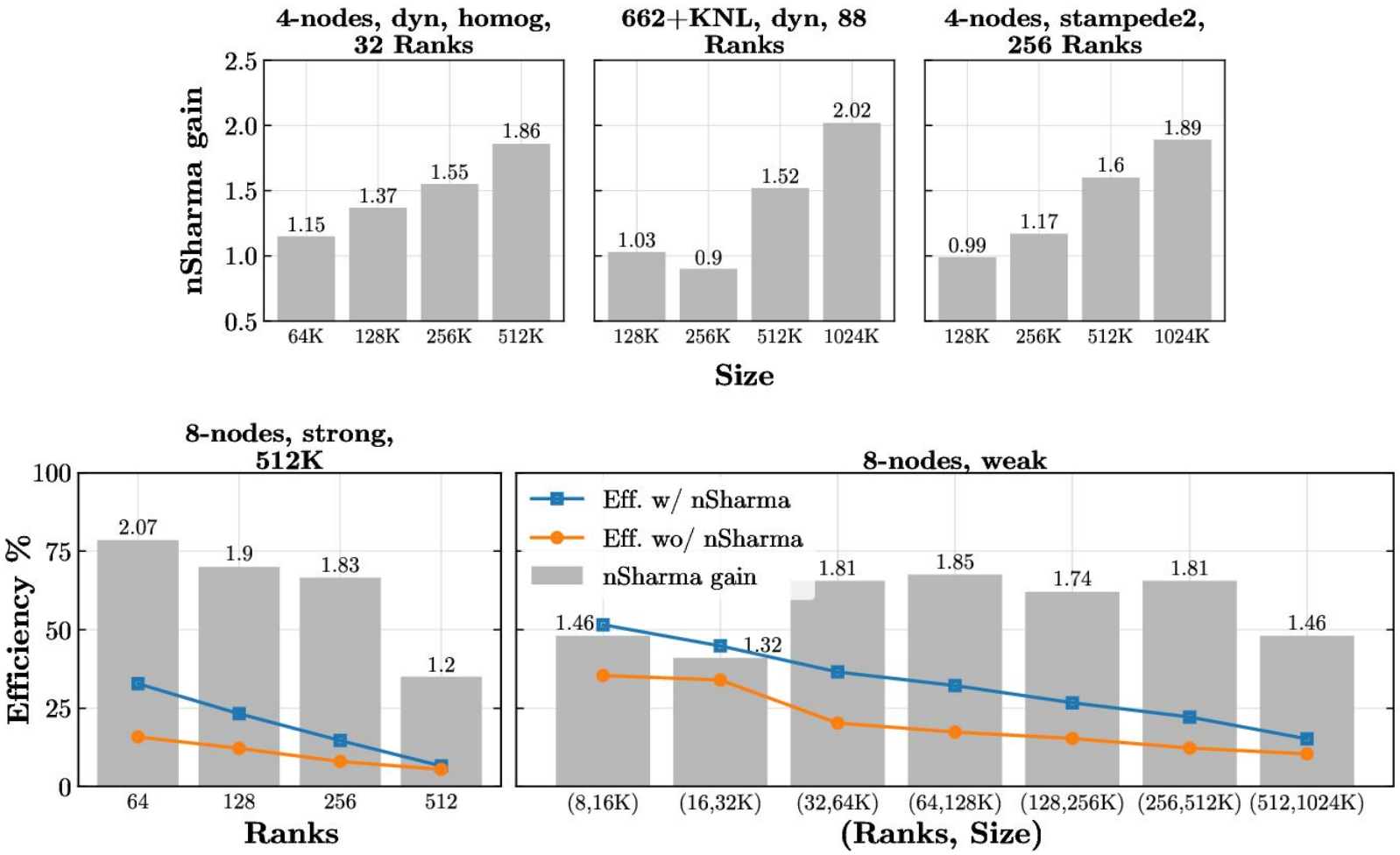




Results

Work and resource scalability

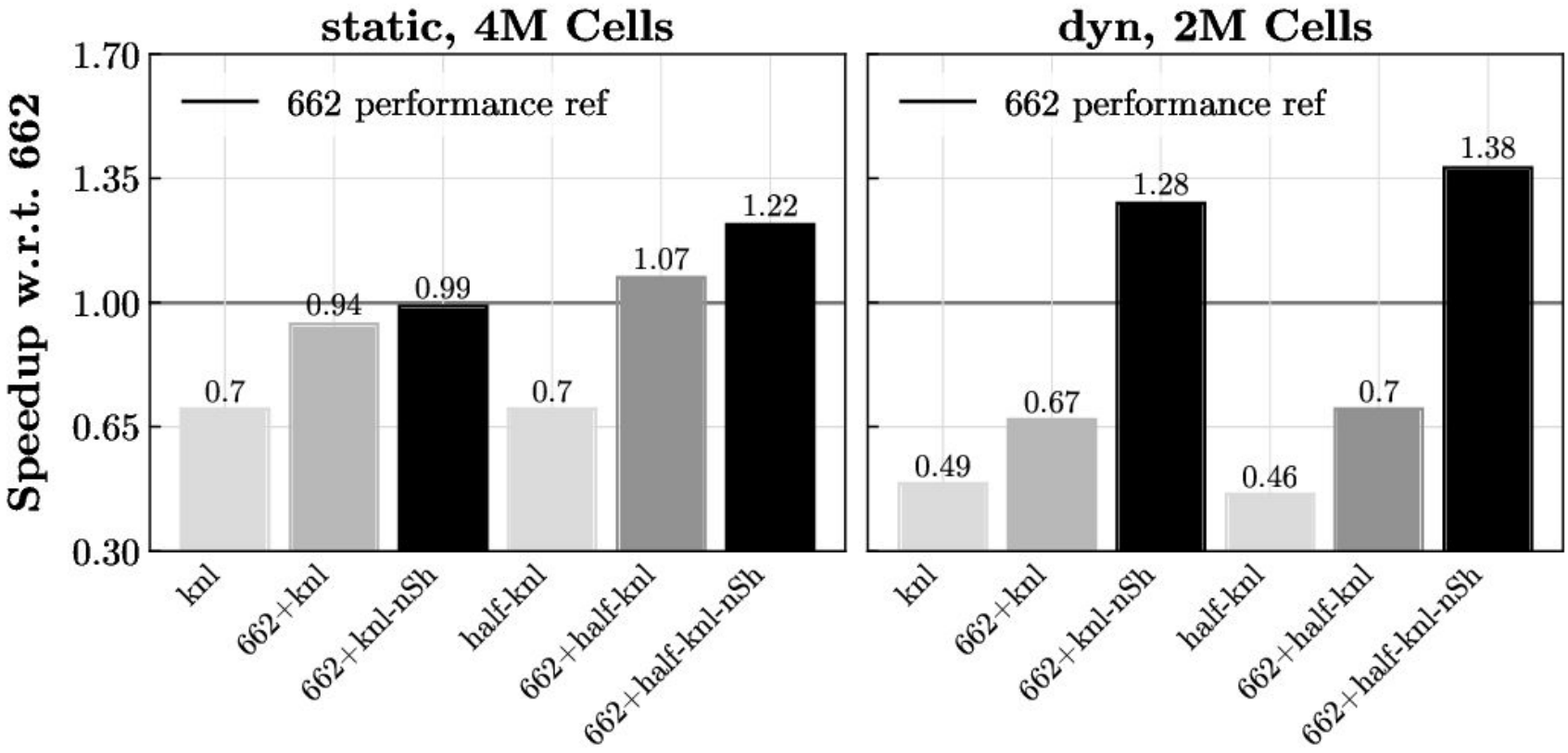
System	SeARCH			Stampede2
Nodes	Tag **641** - Ivy Bridge E5-2650v2 @ 2.60GHz, **16** cores p/node			Tag **KNL7250** - Intel Xeon Phi 7250 @ 1.4GHz ("Knights Landing"), **68** cores p/ node
	Tag **662** - Ivy Bridge E5-2695v2@ 2.40GHz, **24** cores p/node			
	Tag **421** - Nehalem E5520 @ 2.27GHz, **8** cores p/node			
	Tag **KNL7210** - Intel Xeon Phi 7210 @ 1.3GHz, **64** cores p/ node			
Multi-node configurations	**Homogeneous**	**Heterogeneous I**	**Heterogeneous II**	**Homogeneous**
	Multiple 641's	Pair(s) of 641+421	Pair 662+KNL7210	Multiple KNL7250's
Network	Myrinet (**myri**)	Myrinet (**myri**)	Ethernet(**eth**)	Intel Omni-Path (**OPA**)



Results

Increased extracted performance

System	SeARCH			Stampede2
Nodes	Tag **641** - Ivy Bridge E5-2650v2 @ 2.60GHz, **16** cores p/node			Tag **KNL7250** - Intel Xeon Phi 7250 @ 1.4GHz ("Knights Landing"), **68** cores p/ node
	Tag **662** - Ivy Bridge E5-2695v2@ 2.40GHz, **24** cores p/node			
	Tag **421** - Nehalem E5520 @ 2.27GHz, **8** cores p/node			
	Tag **KNL7210** - Intel Xeon Phi 7210 @ 1.3GHz, **64** cores p/ node			
Multi-node configurations	**Homogeneous**	**Heterogeneous I**	**Heterogeneous II**	**Homogeneous**
	Multiple 641's	Pair(s) of 641+421	Pair 662+KNL7210	Multiple KNL7250's
Network	Myrinet (**myri**)	Myrinet (**myri**)	Ethernet(**eth**)	Intel Omni-Path (**OPA**)



Future work

Evaluate with larger node counts

Validate with more/different problems

Devise support and evaluate different dynamic workloads (e.g. particles, moving meshes)

Deploy

Acknowledgements

- This work is funded by FEDER funds through the COMPETE 2020 Programme and National Funds through FCT - Portuguese Foundation for Science and Technology under the project UID/CTM/50025/2013.

- Minho University cluster under the project Search-ON2
  - Revitalization of HPC infrastructure of UMinho, (NORTE-07-0162-FEDER-000086), co-funded by the North Portugal Regional Operational Programme (ON.2-0 Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF).

- PT-FLAD Chair on Smart Cities & Smart Governance

- School of Engineering, University of Minho within project *Performance Portability on Scalable Heterogeneous Computing Systems*

- Texas Advanced Computing Center (TACC) at The University of Texas at Austin

nSharma: Numerical Simulation Heterogeneity Aware Runtime Manager for OpenFOAM,
R. Ribeiro, L. P. Santos, and J. M. Nóbrega,
accepted in International Conference on Computational Science (ICCS),
2018

rribeiro@di.uminho.pt