ADMIRE
malleable data solutions for HPC

Adaptive multi-tier intelligent data manager for Exascale

EuroHPC
Joint Undertaking

admire-eurohpc.eu

Energy-aware malleable scheduling techniques

Alberto Cascajo, David E. Singh, Alvaro Arbe Milara, and Jesus Carretero

PDP2023

Grant Agreement number: 956748 — ADMIRE — H2020-JTI-EuroHPC-2019-1

CABAHLA-CM: convergencia big data-HPC de los sensores a las aplicaciones (p2018/tcs4423)

Comunidad de Madrid

UNIÓN EUROPEA
Fondos Estructurales
Invertimos en su futuro

UNIÓN EUROPEA
Fondo Social Europeo
El Fondo Social Europeo invierte en tu futuro

CABAHLA

Outline

- ❑ Motivation
- ❑ Application energy profile
- ❑ Energy-aware malleable scheduler
- ❑ Results
- ❑ Conclusions

Introduction: contributions

❑ Work contributions:

- ➢Integration of the application use case with FlexMPI and an energy monitor
- ➢ Energy profile modeler
- ➢Energy-aware malleable scheduler
- ➢Practical evaluation on a real platform

Outline

- □ Motivation
- □ Application energy profile
- □ Energy-aware malleable scheduler
- □ Results
- □ Conclusions

Motivation

- ❑ Current schedulers lack of malleability support

- ❑ Energy-aware malleable techniques are an open research area

- ❑ Application monitoring combined with run-time algorithms can provide adaptability to application changing conditions

Outline

❑ Motivation

❑ Application energy profile

❑ Energy-aware malleable scheduler

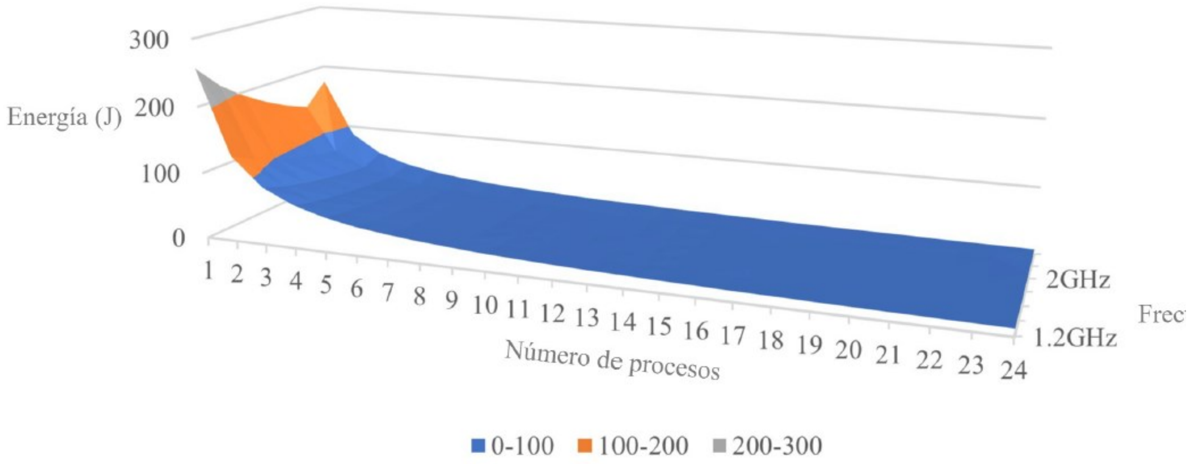
❑ Results

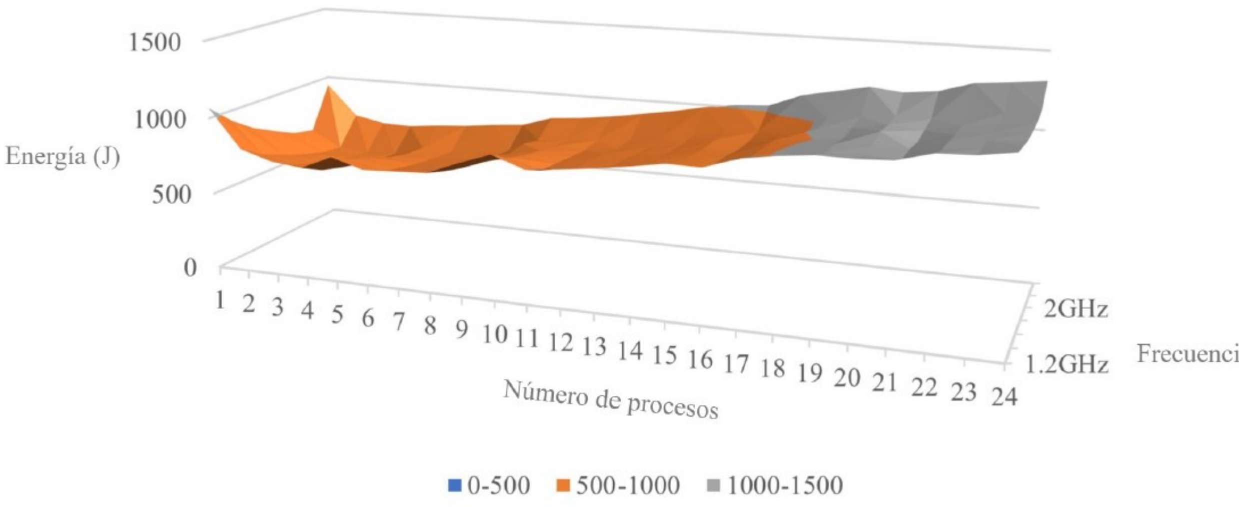
❑ Conclusions

Introduction: application energy profile

❑ Application energy profile is **application energy** for a certain:

1. DVFS value, compute-node dependent, RAPL interface.
2. Number of processes, application dependent, FlexMPI support.





Introduction: application energy profile

- ❑ Application energy profile: energy values for the range of two different parameters
  - ➢Compute node DVFS value
  - ➢Application number of processes

- ❑ Energy profile is different for each application

- ❑ Computing the application energy profile requires many evaluations

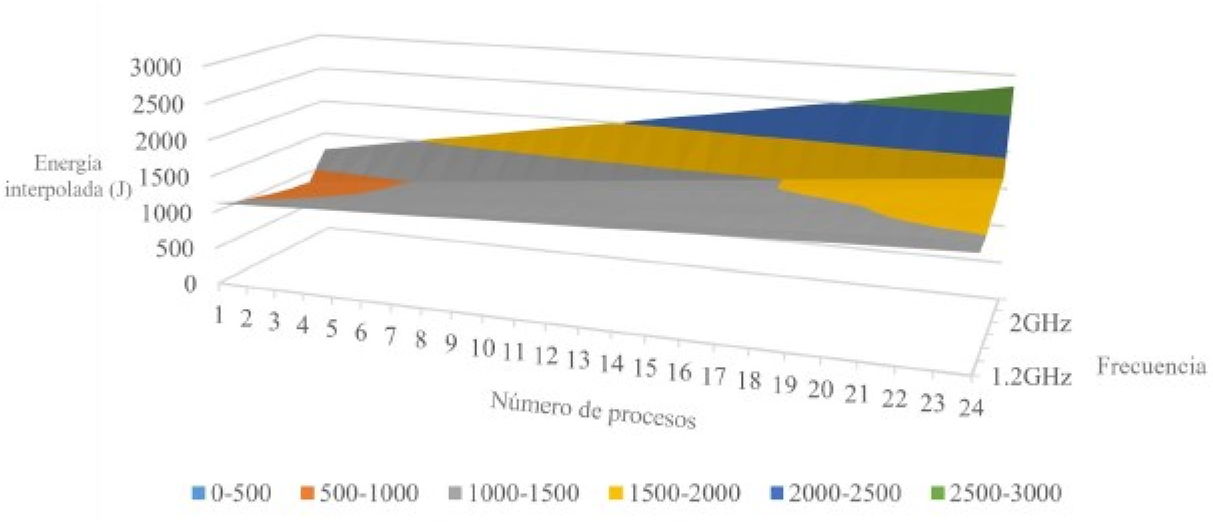
- ❑ Challenge: obtain the application energy profile in run-time with a reduced number of evaluation.

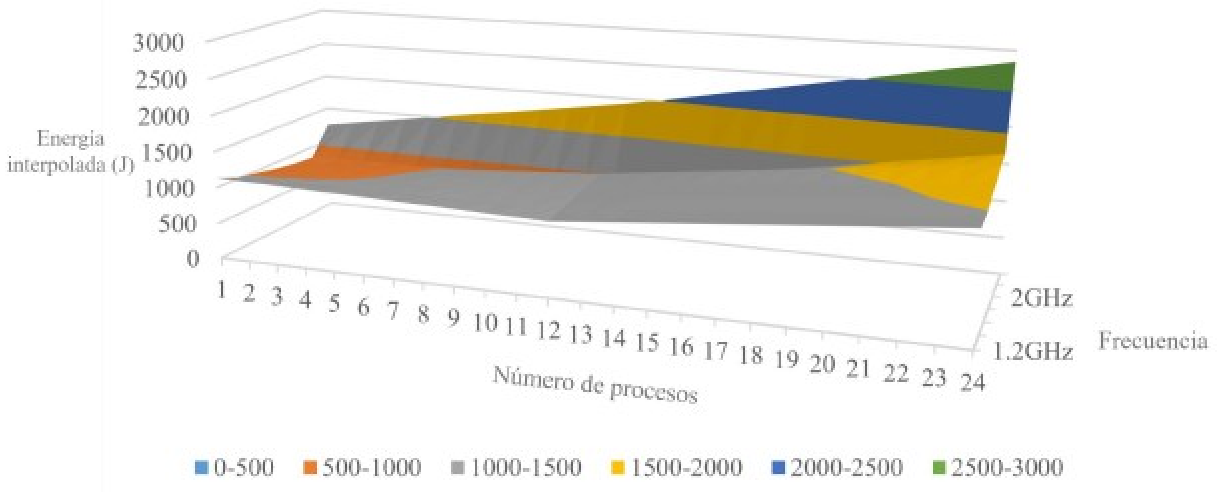
Energy profile modeler

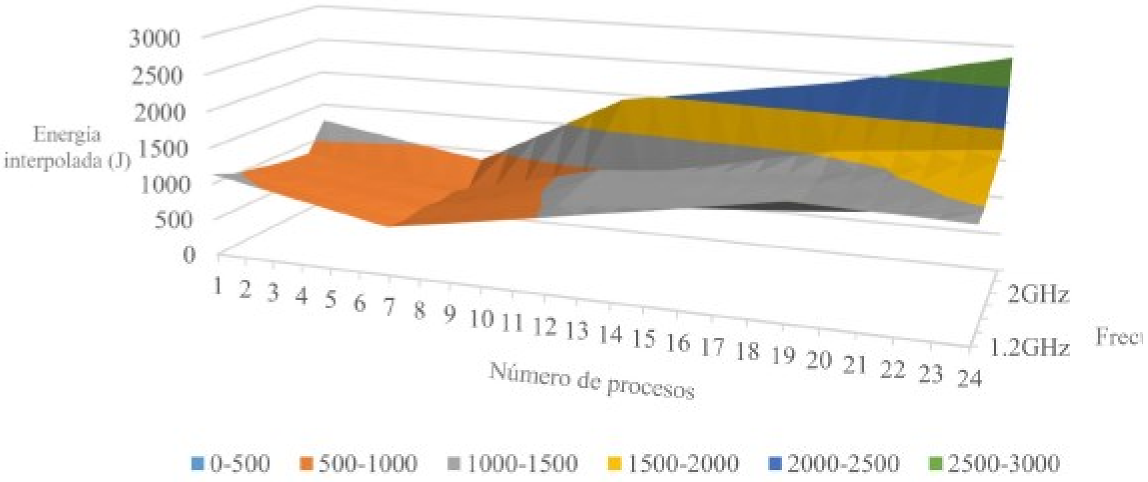
- □ Calculates the application energy profile in run-time combining:
  - ➢Application energy monitoring
  - ➢DVFS level
  - ➢Application's number of processes
  - ➢Interpolation algorithm
- □ Iterative algorithm
  - ➢Based on interpolation

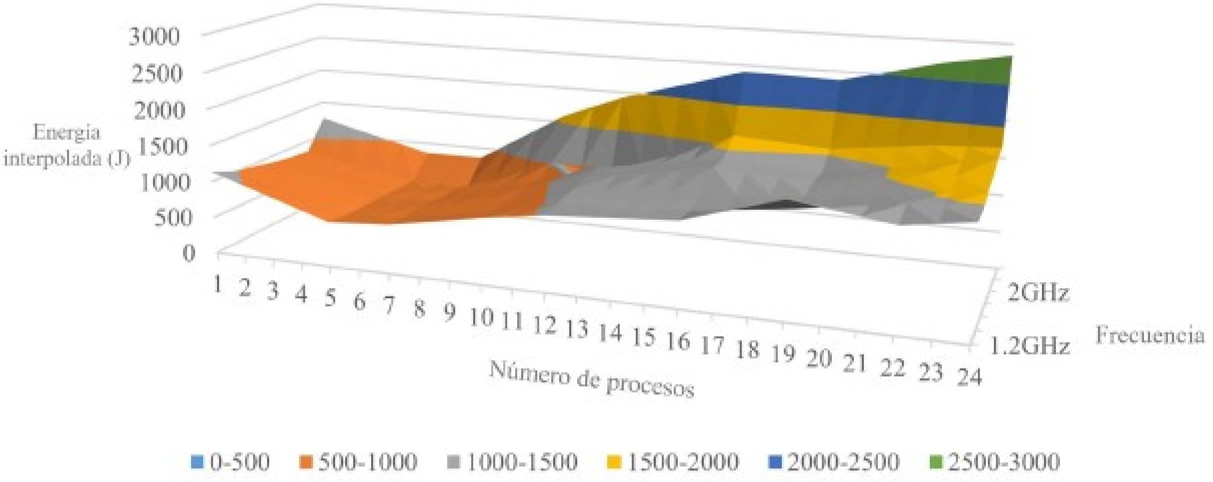
```
for i=1,2
    ($vf_i$,$np_i$) = SetSample(i)
     $E_i$ ← TakeSample($vf_i$,$p_i$)
end
AP = EProf($E_{i=1…N}$,$vf_{i=1…N}$,$p_{i=1…N}$)
err = ComputeError(AP)
while(error > threshold)
    i++
    ($vf_i$,$np_i$) = SetSample(i)
    $E_i$ ← TakeSample($vf_i$,$p_i$)
    AP = EProf($E_{i=1…N}$,$vf_{i=1…N}$,$p_{i=1…N}$)
    err = ComputeError(AP)
end
```

Energy profile modeler

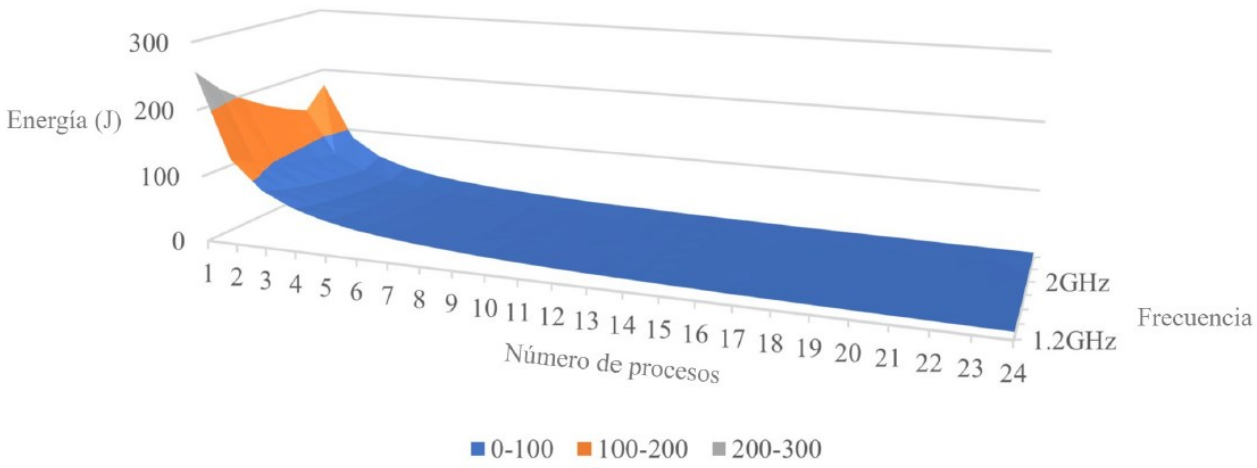


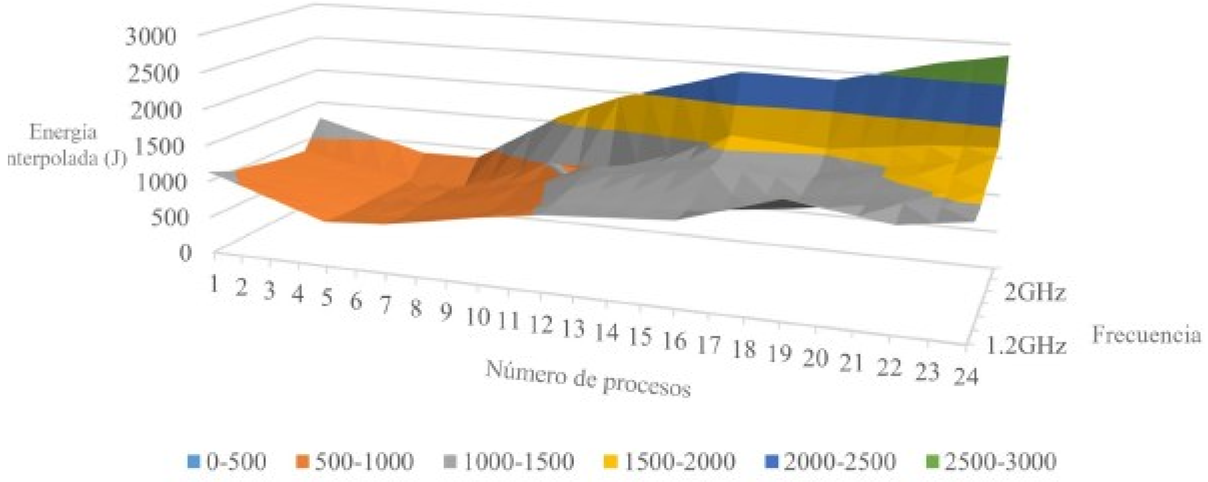


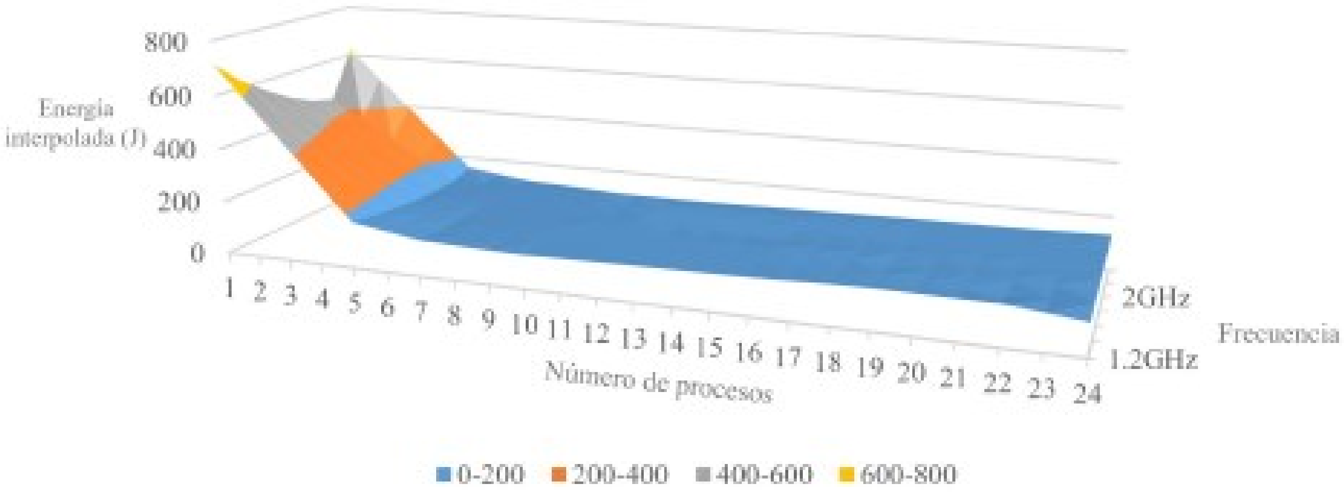


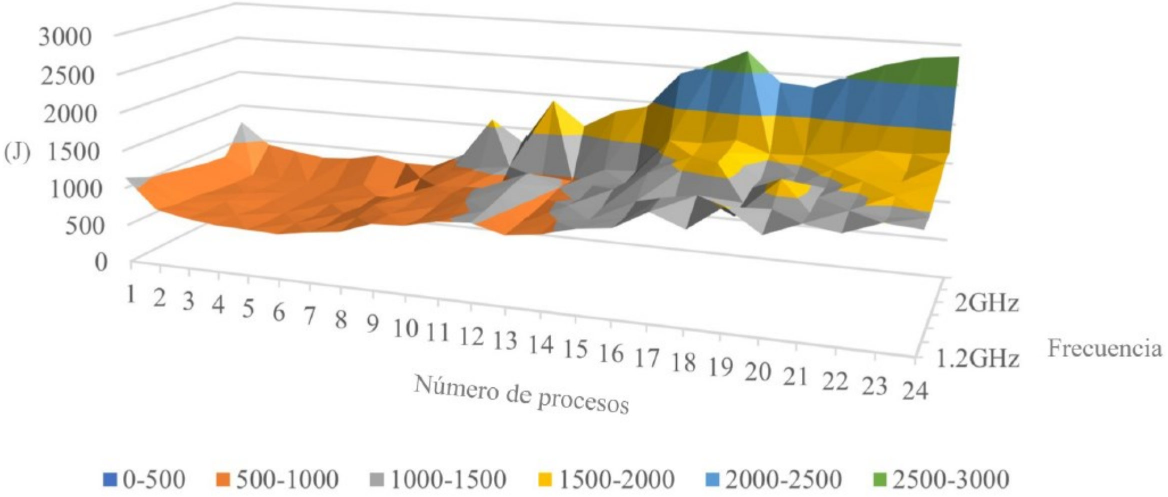


Energy profile modeler









Outline

□ Motivation

□ Application energy profile

□ Energy-aware malleable scheduler

□ Results

□ Conclusions

Malleable scheduler

- ❑ Considers both the application energy profile (*E*) and execution time (T)
- ❑ E_{max}, T_{max} are the application maximum values
- ❑ W_1 and W_2 are weights
- ❑ Optimization algorithm searches the minimum F value
- ❑ Balances two goals: energy and execution time minimization

$$F(NP, freq) = W_1 \frac{E(NP, freq)}{E_{\max}} + W_2 \frac{T(NP, freq)}{T_{\max}}$$

Outline

❑ Motivation

❑ Application energy profile

❑ Energy-aware malleable scheduler

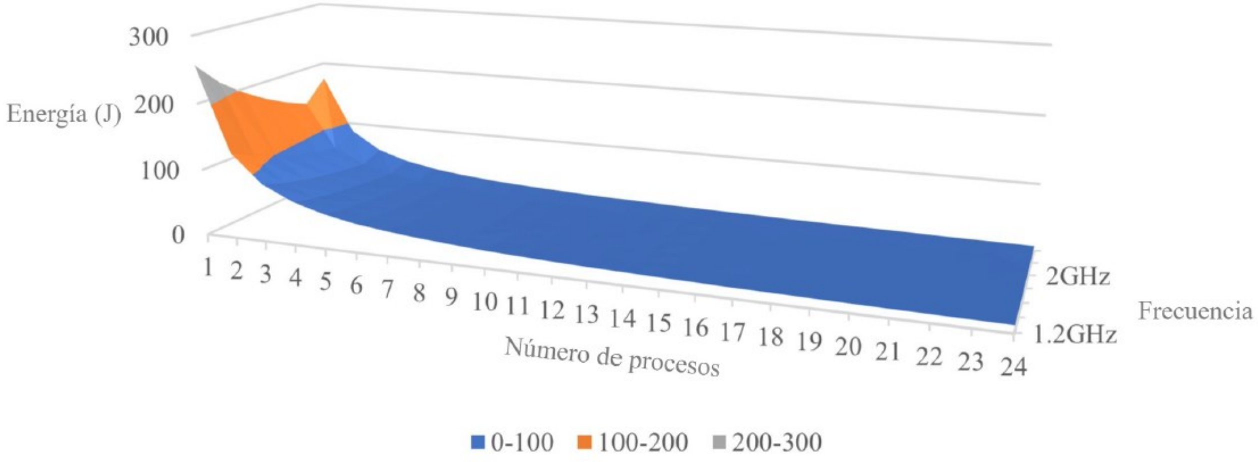
❑ Results

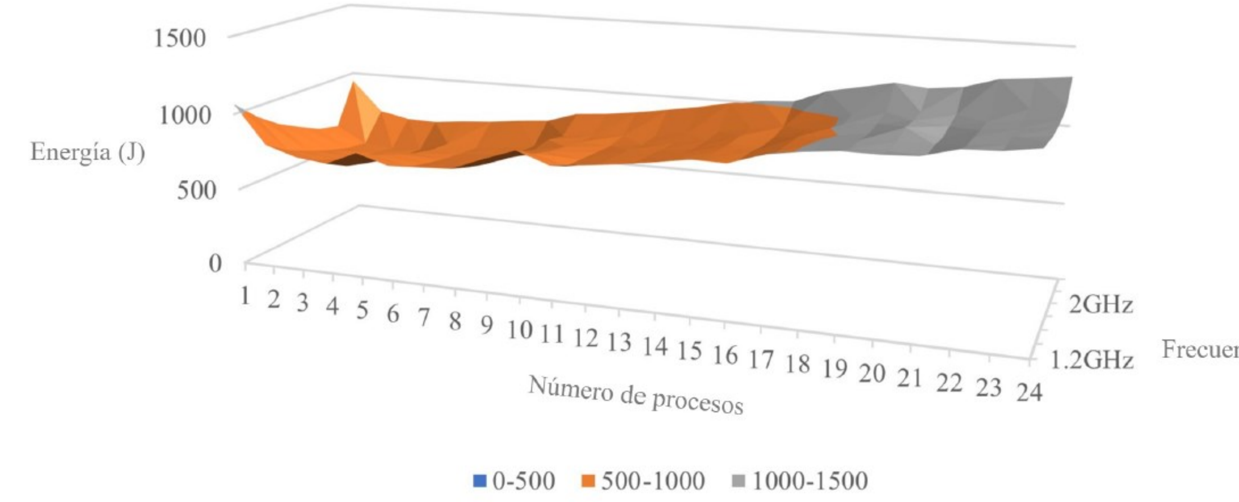
❑ Conclusions

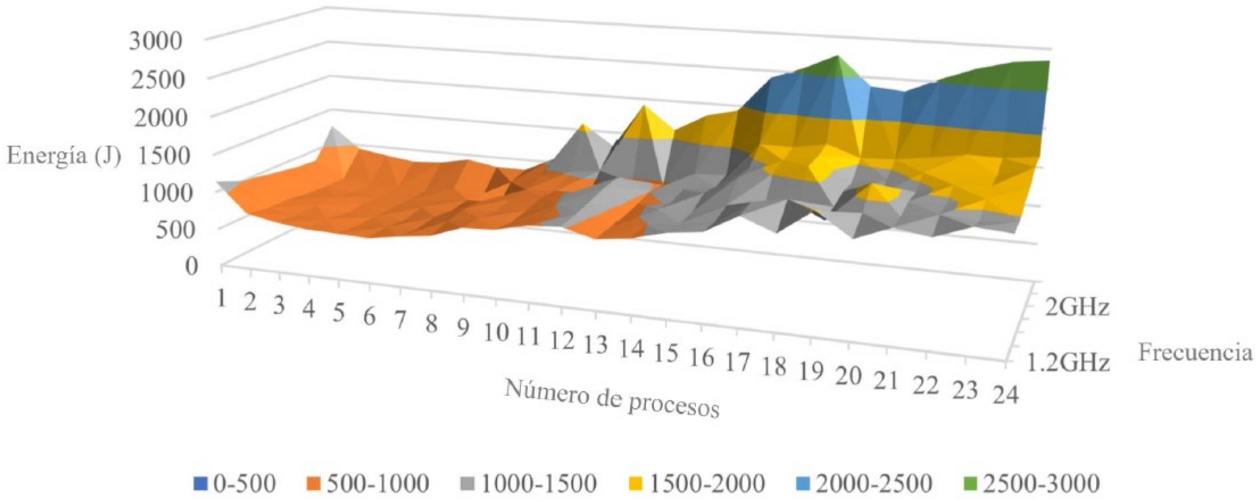
Evaluation: platform

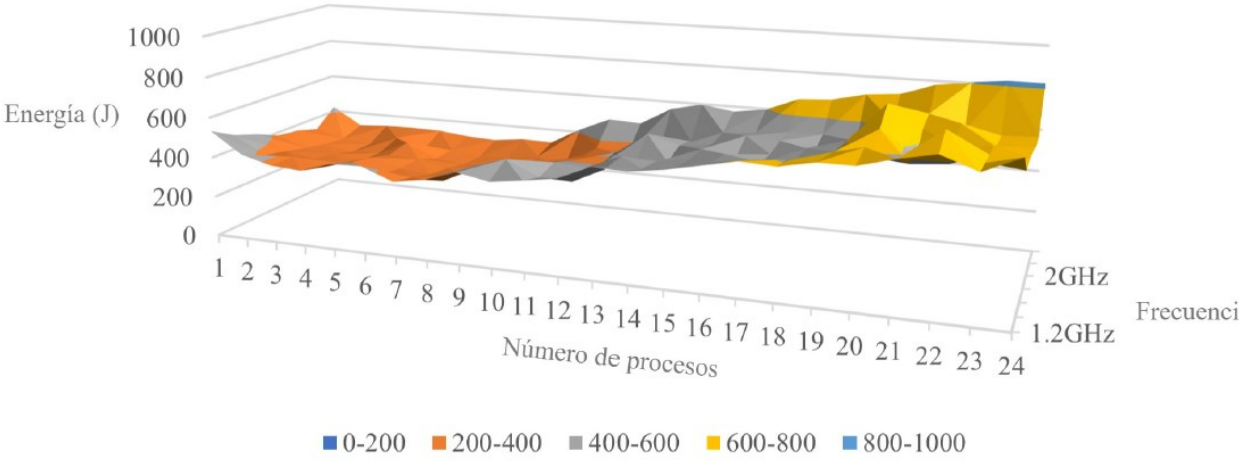
- ❑ Intel Xeon Gold 6212U, 24 cores, 314 GB RAM.
- ❑ Use cases:
  - ➢Use case A: CPU-intensive with high locality data accesses
  - ➢Use case B: CPU-intensive with low locality data accesses
  - ➢Use case C: communication-intensive with low locality data accesses
  - ➢Use case D: I/O-intensive with low locality data accesses
  - ➢Use case E: mixed CPU, communication and I/O phases with low locality data accesses

Evaluation: energy profile





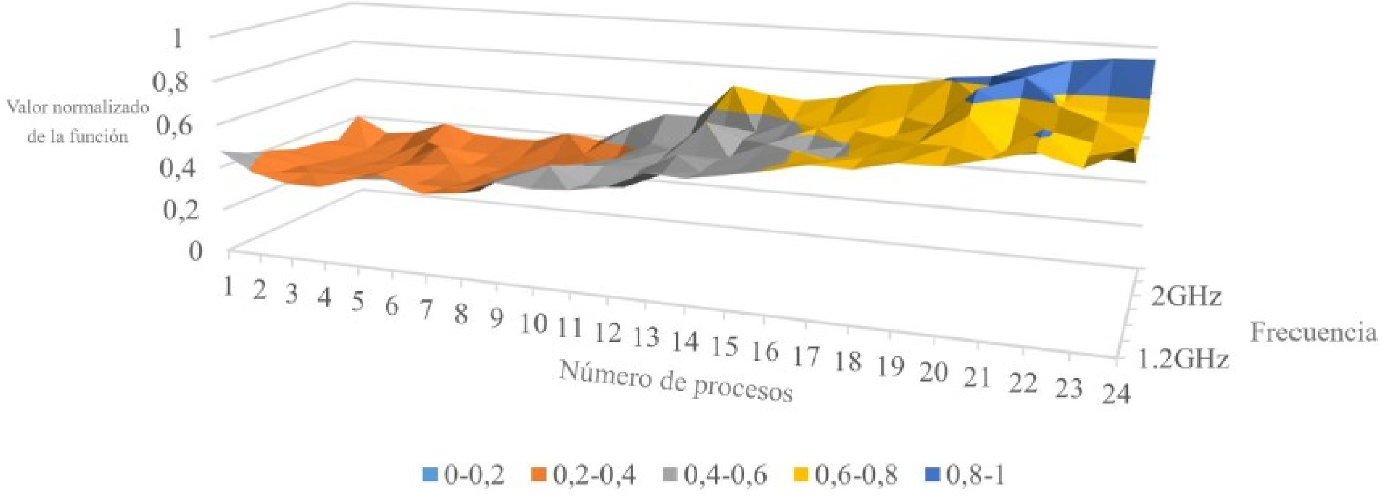


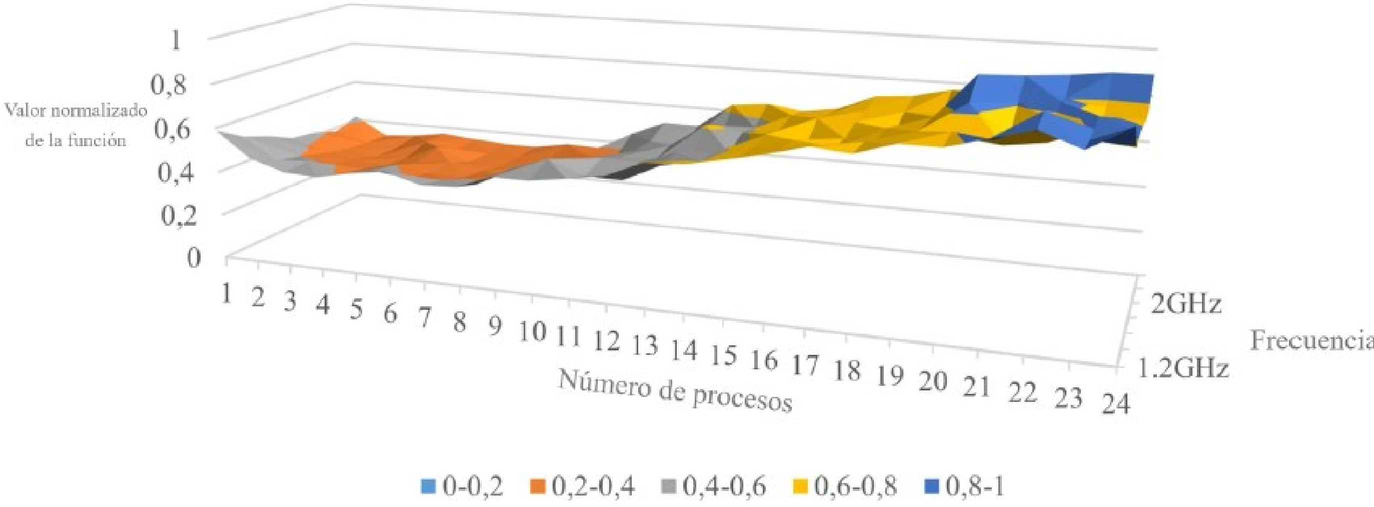


- Optimization *F* function for use case E
  - Only energy optimization

    $W_1=1, W_2=0$
  - Only execution time optimization

    $W_1=0, W_2=1$





Results

- □ Scheduler solutions for the energy minimization ($W_1=1, W_2=0$)

- □ Full search vs interpolation with 5 values.

Use case	Configuration (full detail)	Saving (full detail)	Configuration (interpolation)	Saving (interpolation)
A	24 procs, 2.2 GHz	93%	24 procs, 2.0 GHz	92%
B	24 procs, 2.2 GHz	92%	10 procs, 2.4 GHz	85%
C	3 procs, 2.2 GHz	59%	5 procs, 2.2 GHz	54%
D	8 procs, 2.0 GHz	81%	7 procs, 2.2 GHz	81%
E	3 procs, 2.0 GHz	77%	1 proc, 2.,2 GHz	75%

ADMIRE
malleable data solutions for HPC

Outline

- ❑ Motivation
- ❑ Application energy profile
- ❑ Energy-aware malleable scheduler
- ❑ Results
- ❑ Conclusions

EuroHPC
Joint Undertaking

Conclusion

- ❑ We have developed a dynamic energy-profile model
  - ➢Accurate for the considered use cases
  - ➢Only a few iterations produce a good model (in terms of detail level)

- ❑ We have implemented a malleable scheduler
  - ➢That uses the previous model to determine the best application configuration

- ❑ We have completed an evaluation on a real platform
  - ➢By means of this approach it is possible to minimize either the energy consumption or the execution time.
  - ➢Intermediate optimization levels that balance both terms are also possible