Energy Efficiency Evaluation in Heterogeneous Computers

Borja Pérez, Esteban Stafford, José Luis Bosque, Ramón Beivide

Contents

- Motivation
- Goal
- Load Balancing Methods
- Benchmarks
- Experimental Results
- Conclusions

Motivation

- Energy efficiency is a real hurdle in the path to Exascale
  - Mont-Blanc
- GPUs offer a great FLOPs per Watt ratio but their truly efficient use is not trivial
  - Are based on Host-Device models
- How do work distribution and load balancing affect efficiency?

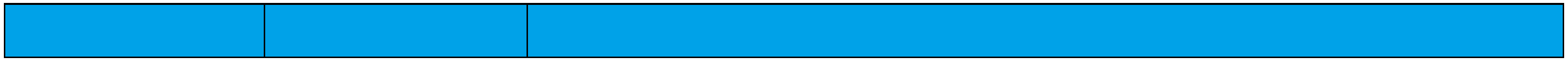
Goal

- Analyze the impact on both time and energy of distributing the load of a single, data-parallel kernel among several devices.
  - We use Maat [1], a library that provides device abstraction and load balancing for OpenCL kernels
    - Focus on data parallelism
    - Synchronisation has an overhead
    - Different applications may have different needs

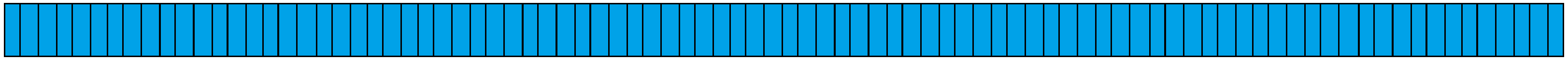
[1] Borja Pérez, José Luis Bosque, and Ramón Beivide. *Simplifying programming and load balancing of data parallel applications on heterogeneous systems*. GPGPU '16.

Load Balancing Methods

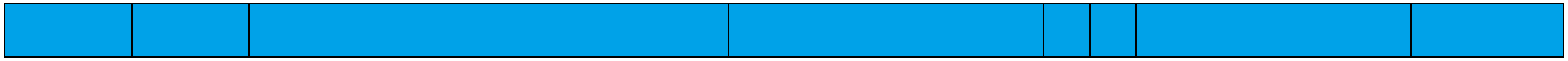
STATIC



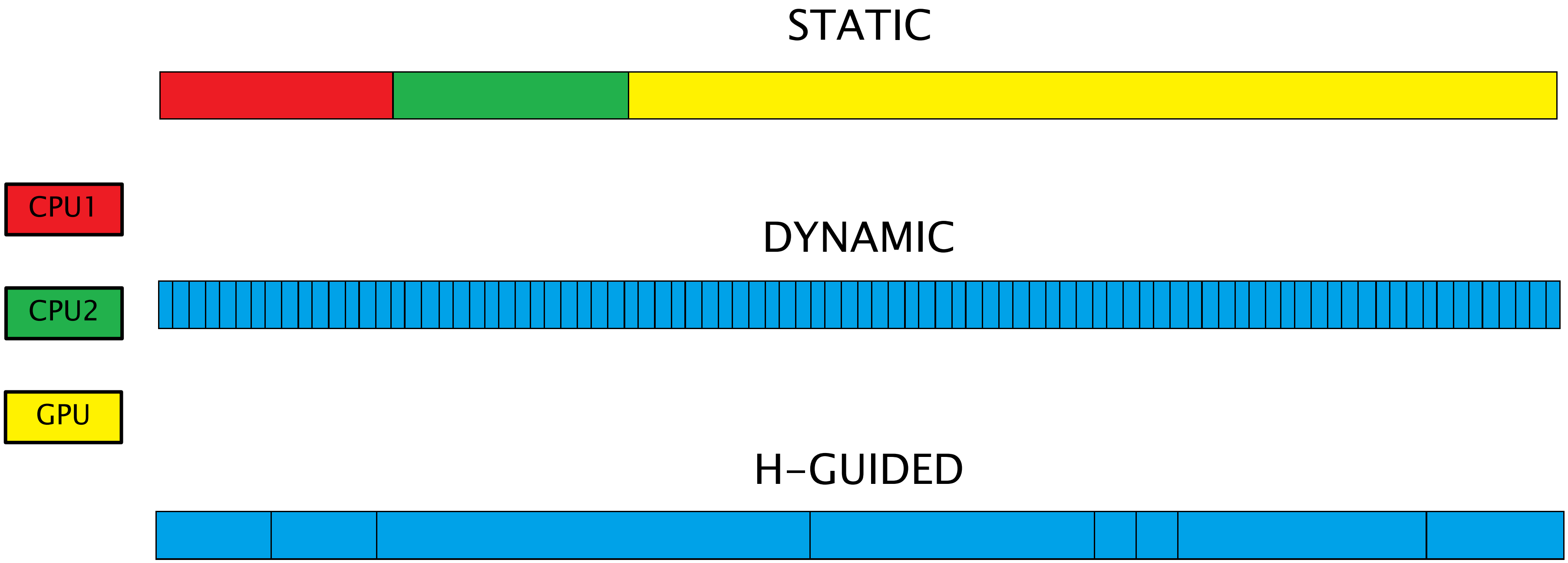
DYNAMIC



H-GUIDED



Load Balancing Methods


STATIC

CPU1

CPU2

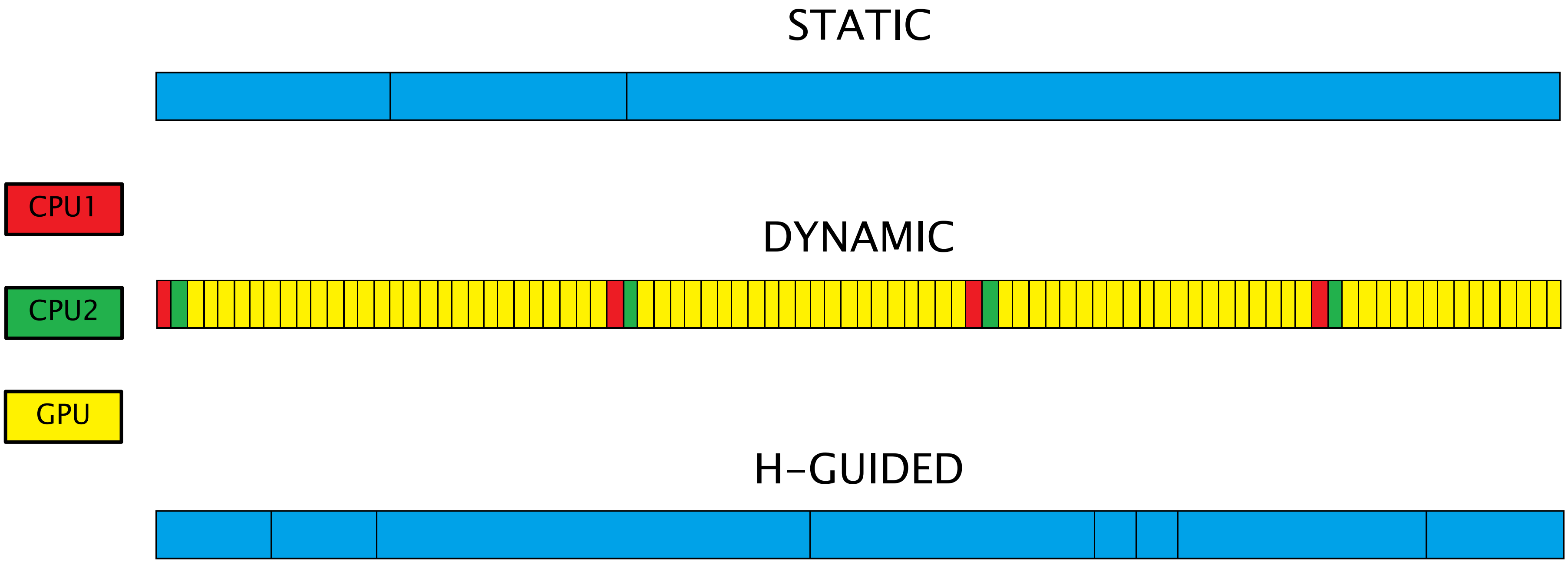
GPU

DYNAMIC

H-GUIDED

- Pros
  - Simple
  - Minimizes synchronisation points
- Cons
  - Determining computing powers
  - Irregular loads

Load Balancing Methods



STATIC

CPU1

CPU2

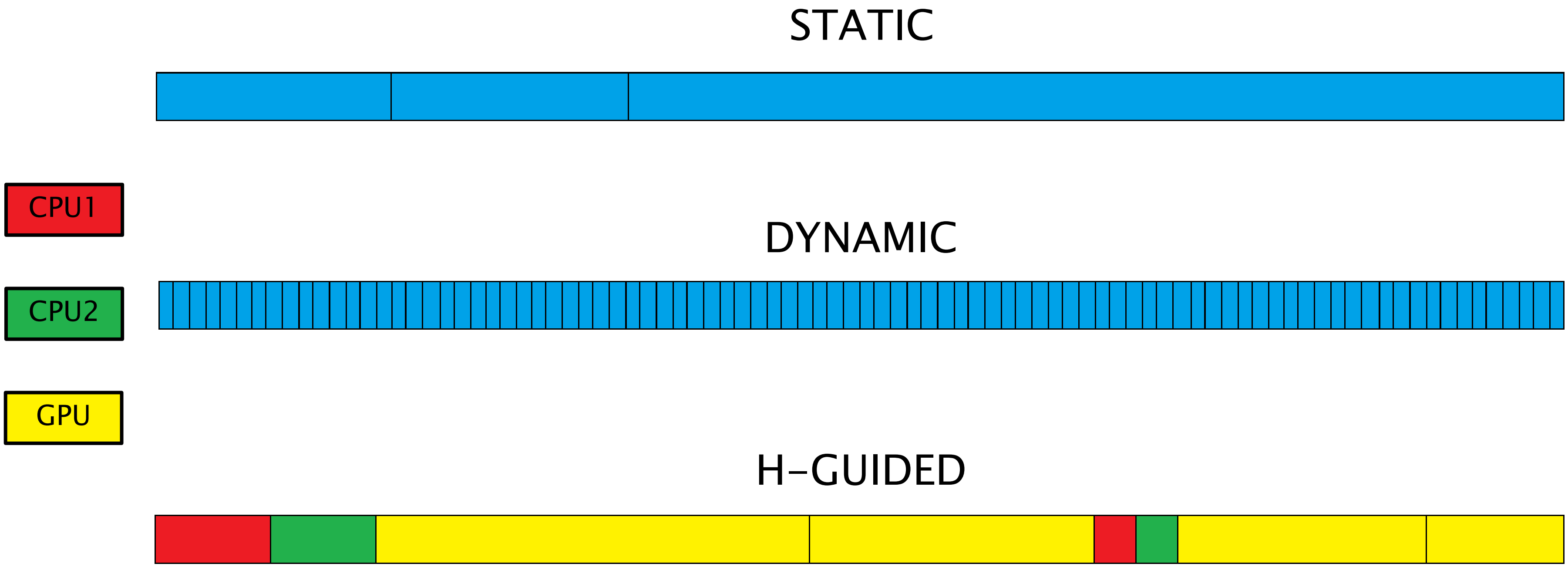
GPU

DYNAMIC

H-GUIDED

- Pros
  - Good for irregular loads
  - Does not use computing powers
- Cons
  - Too many synchronisation points

Load Balancing Methods



- Pros
  - Less synchronization points
  - Still dynamic
- Cons
  - It uses computing powers

Benchmarks

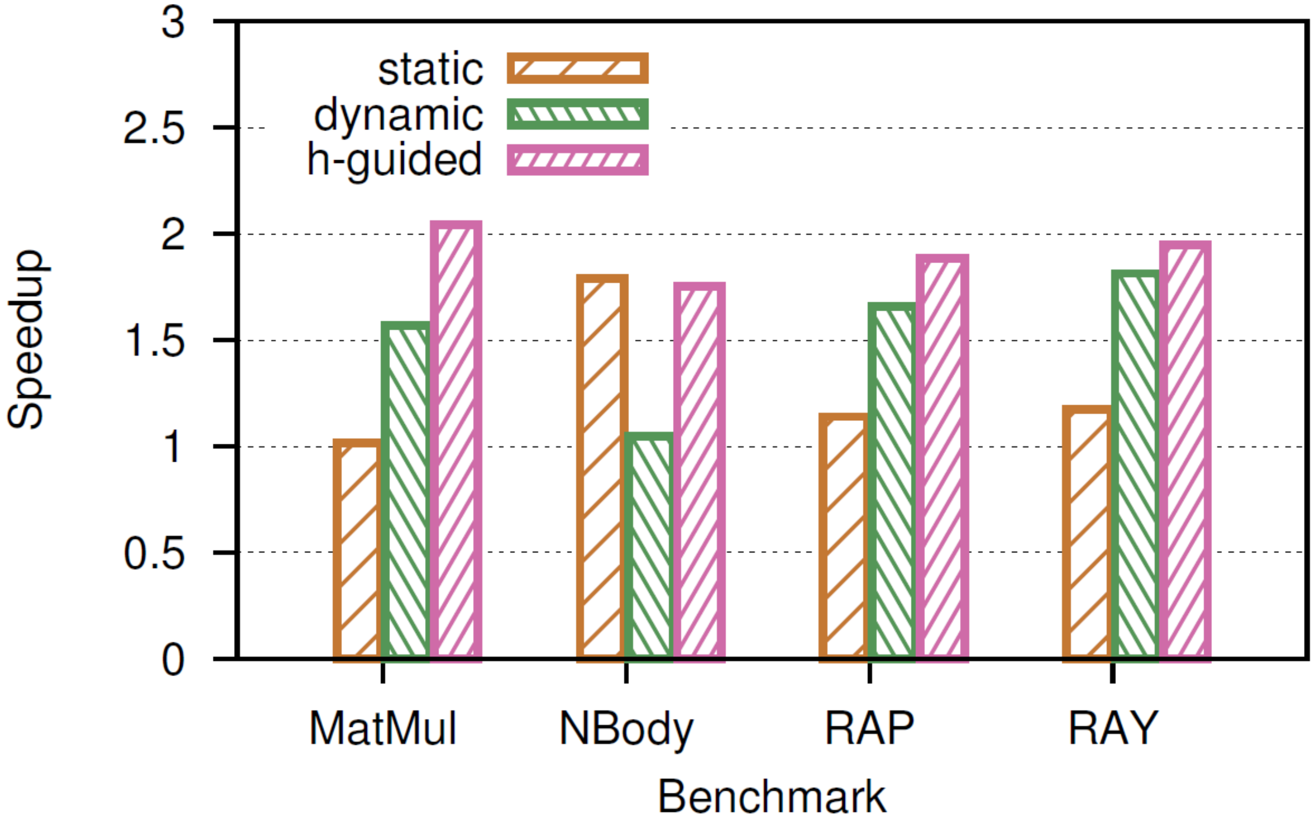
- Several available suites
  - Parboil, AMD APP SDK, Rodinia…
- Selected applications (have to be ported):
  - Regular
    - Nbody
    - MatMul
  - Irregular
    - RAP *(Resource Allocation Problem)*
    - Raytracing

Experimental Results

- The system:
  - 2xKepler K20m
  - 2xIntel Xeon E5-2670 (12 cores in total)
- Considered metrics
  - Speedup
  - Energy Consumption
  - EDP
- To measure energy, a monitor was developed that periodically samples the power consumption of each device
  - GPU power sensors through the NVIDIA Management Library (NVML)
  - Running Average Power Limit (RAPL) registers of the CPUs

Experimental Results

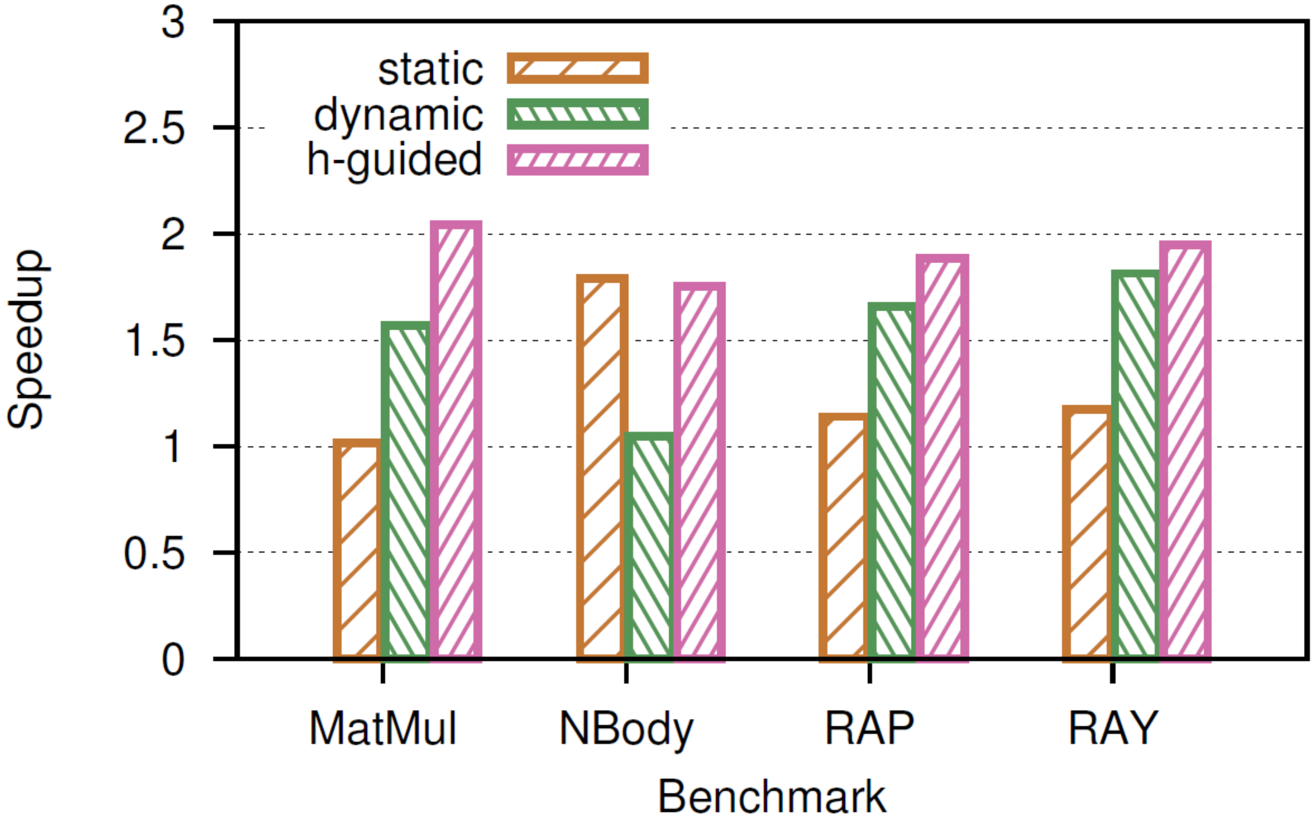
- Speedup



- Static succeeds at regular loads

Experimental Results

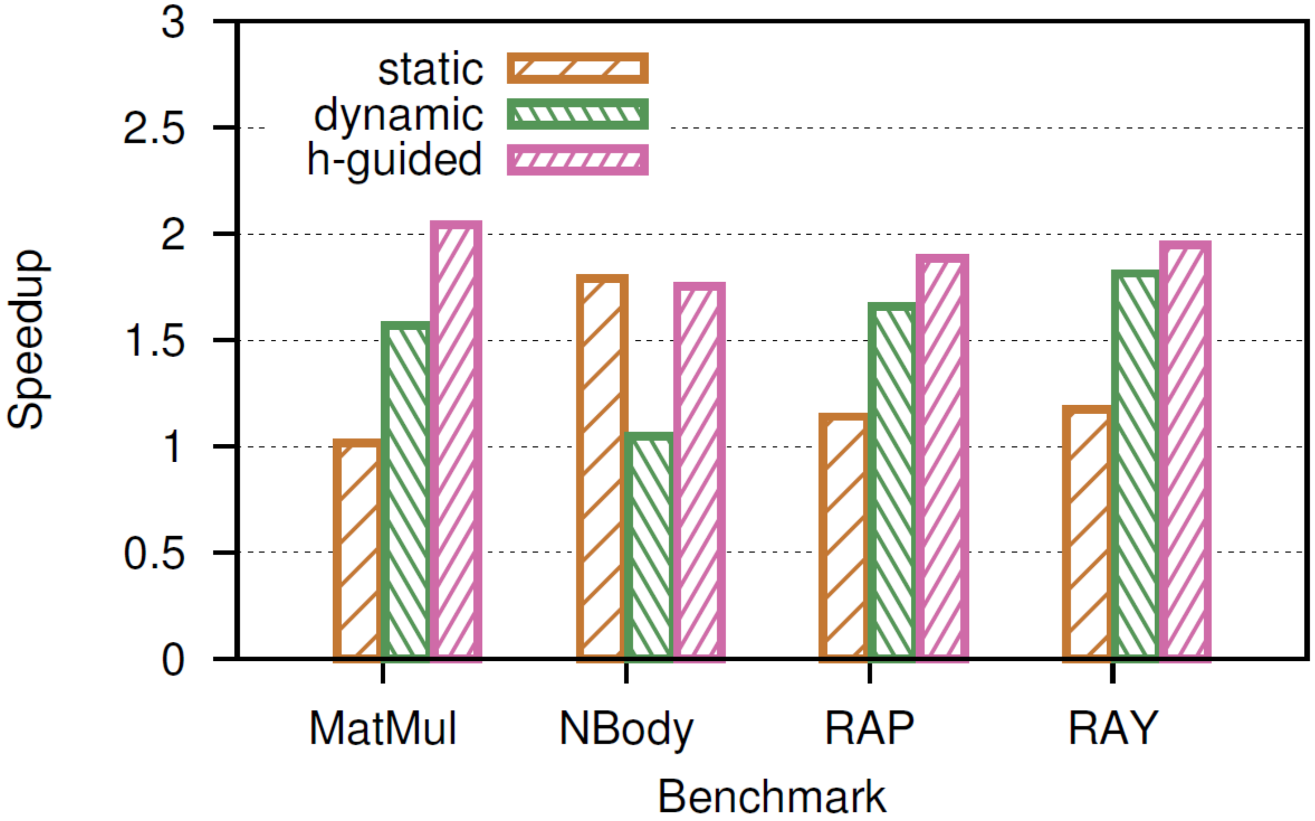
- Speedup



- Smallest package is too big for the CPU in MatMul

Experimental Results

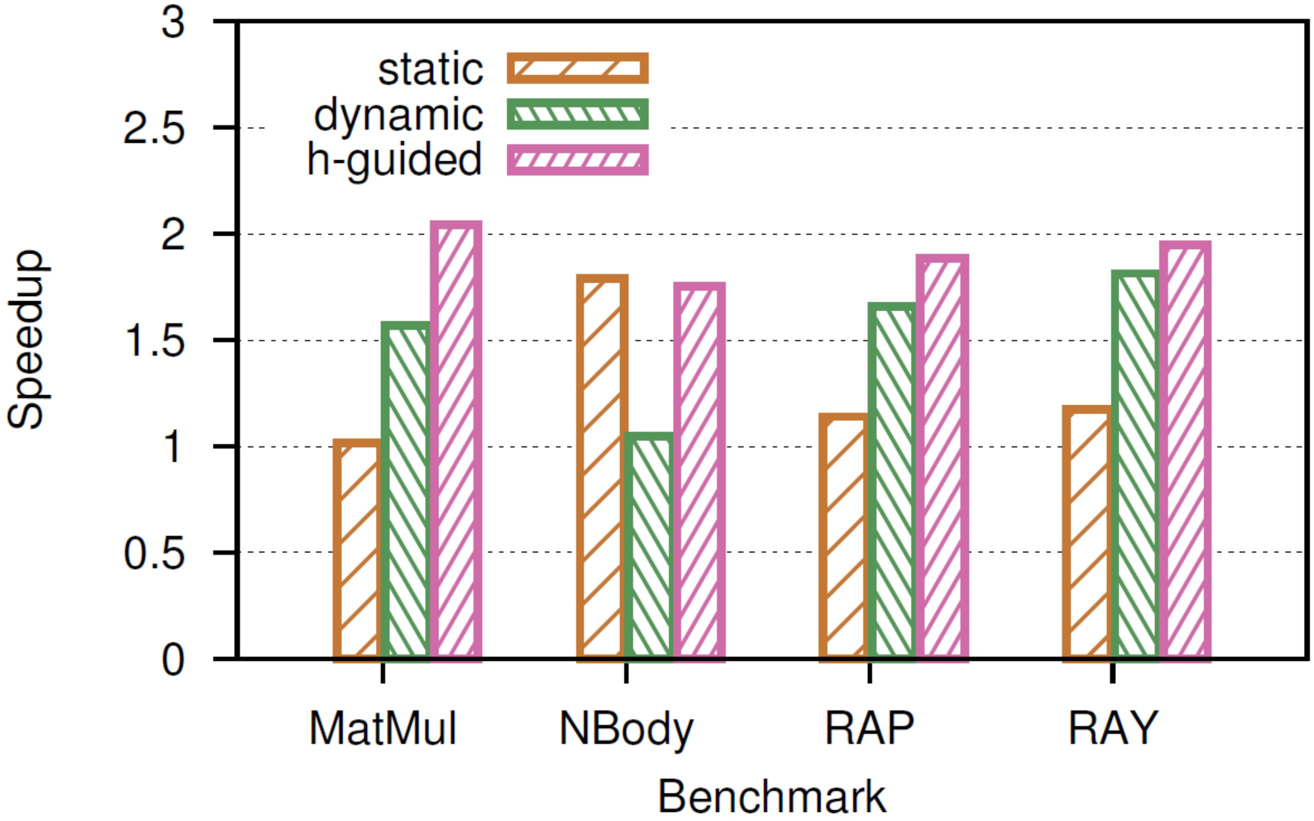
- Speedup



- Nbody is highly affected by overhead

Experimental Results

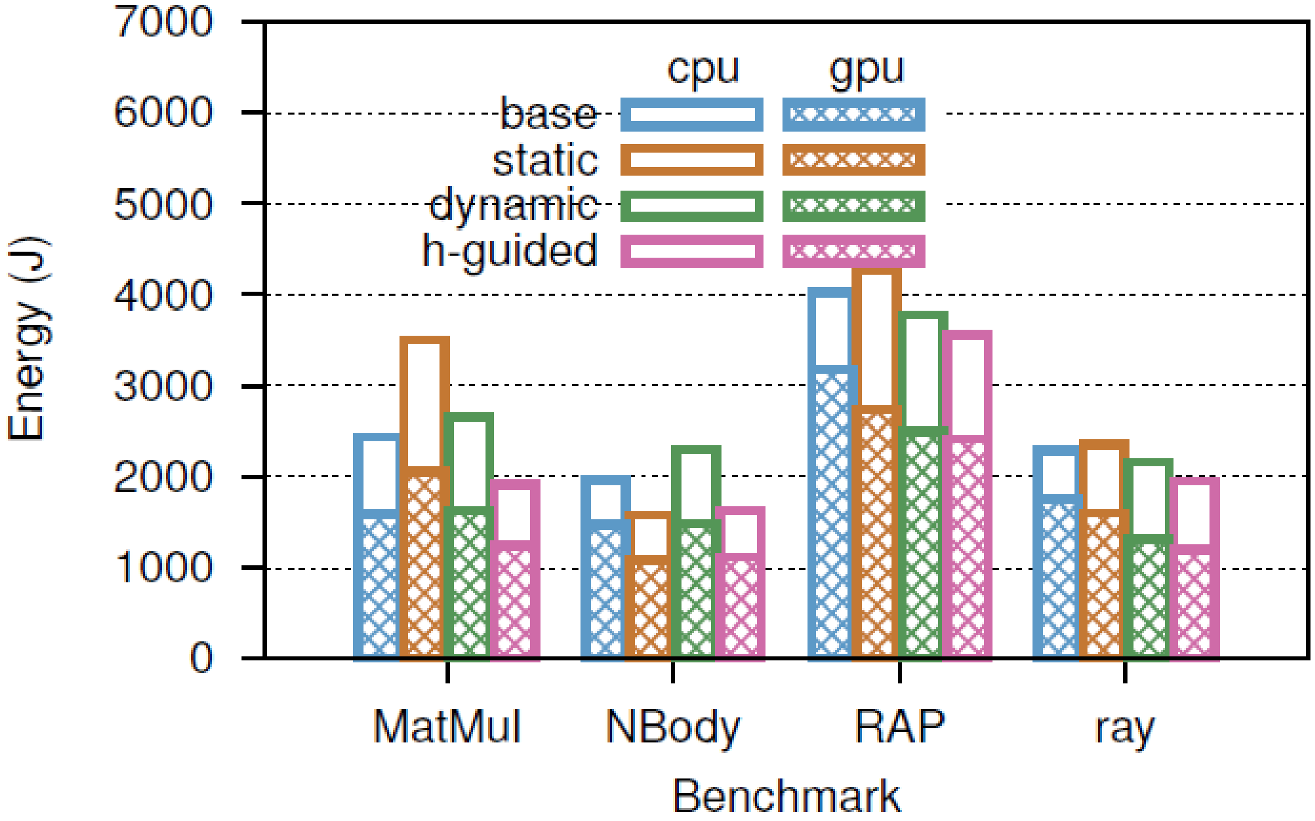
- Speedup



- Guided (and dynamic) succeed at irregular loads

Experimental Results

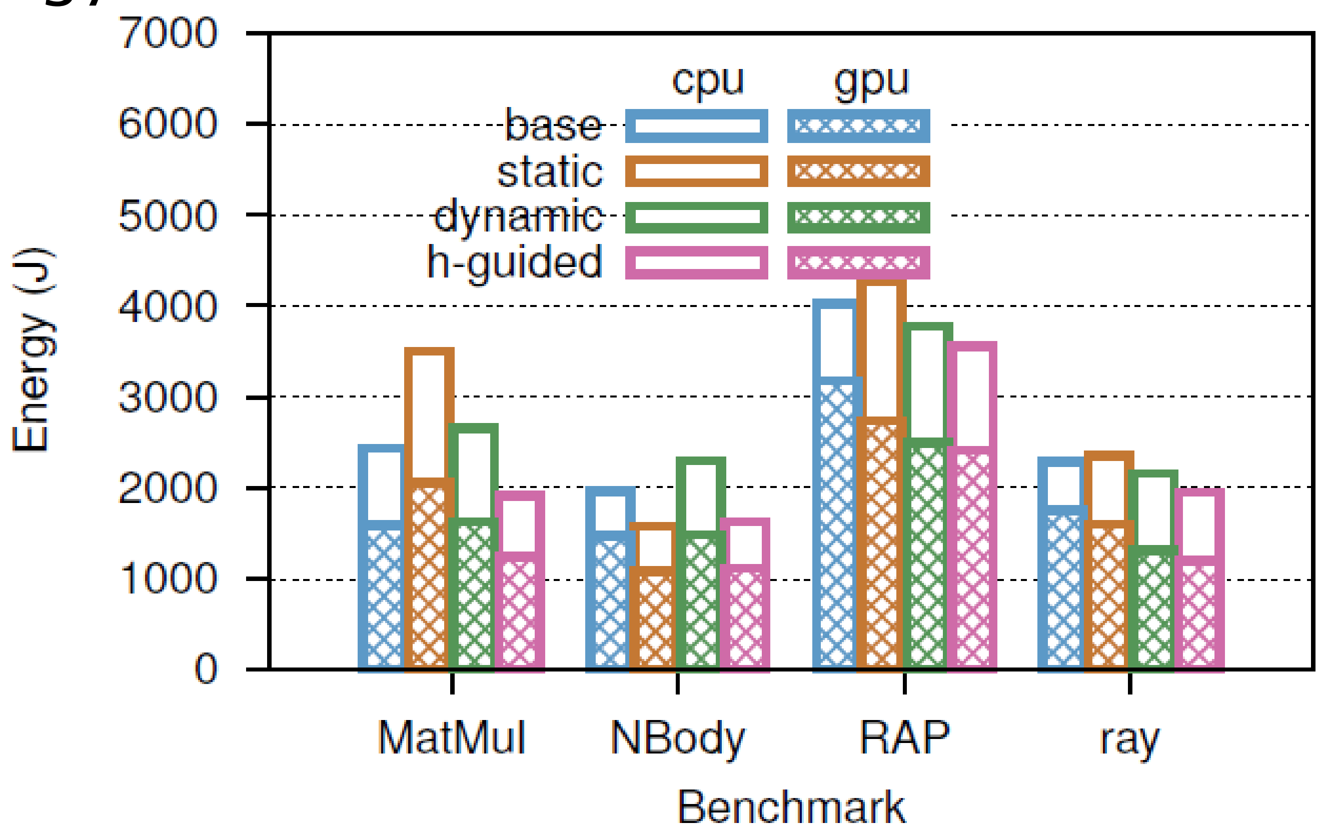
- Energy



- There is at least one option that improves energy for all benchmarks

Experimental Results

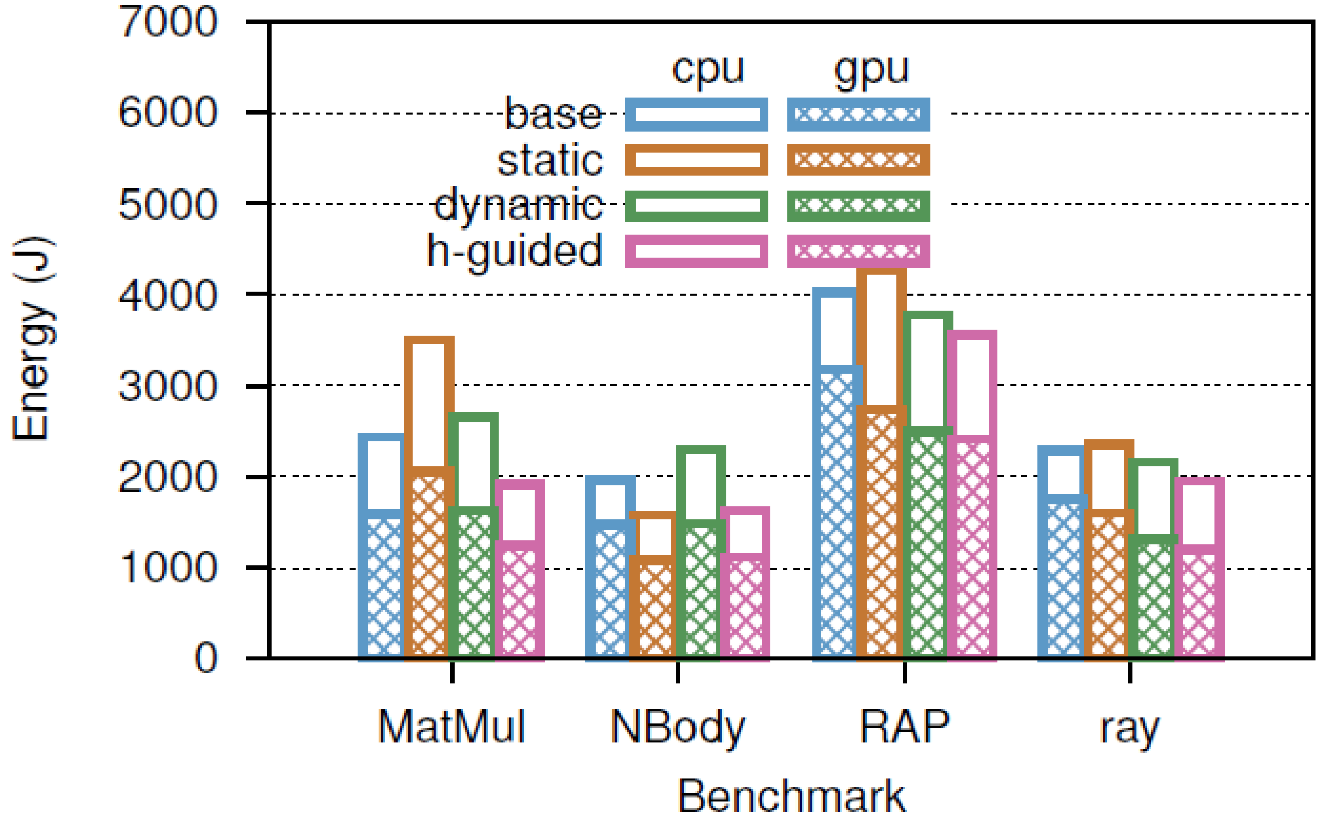
- Energy



- Best results with static for Nbody

Experimental Results

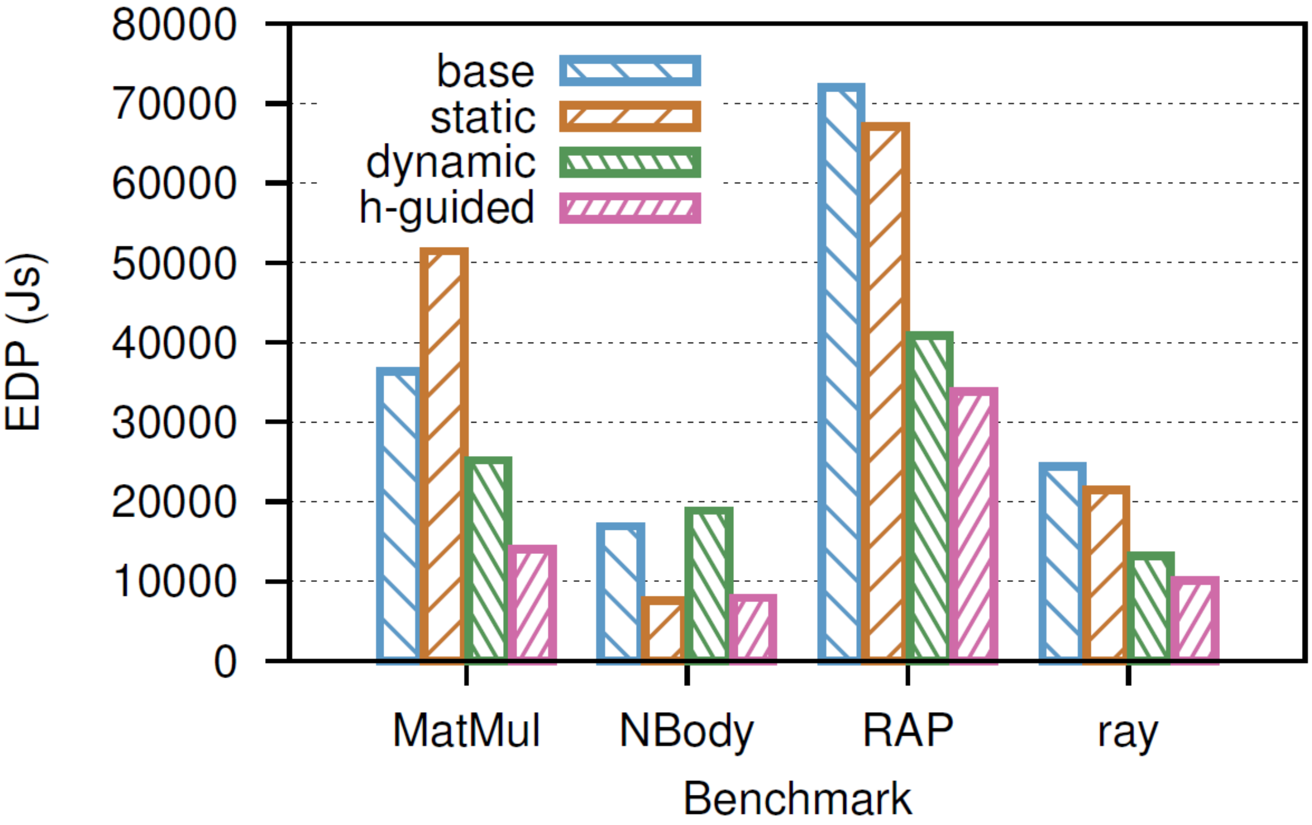
- Energy



- Almost same energy with guided and dynamic

Experimental Results

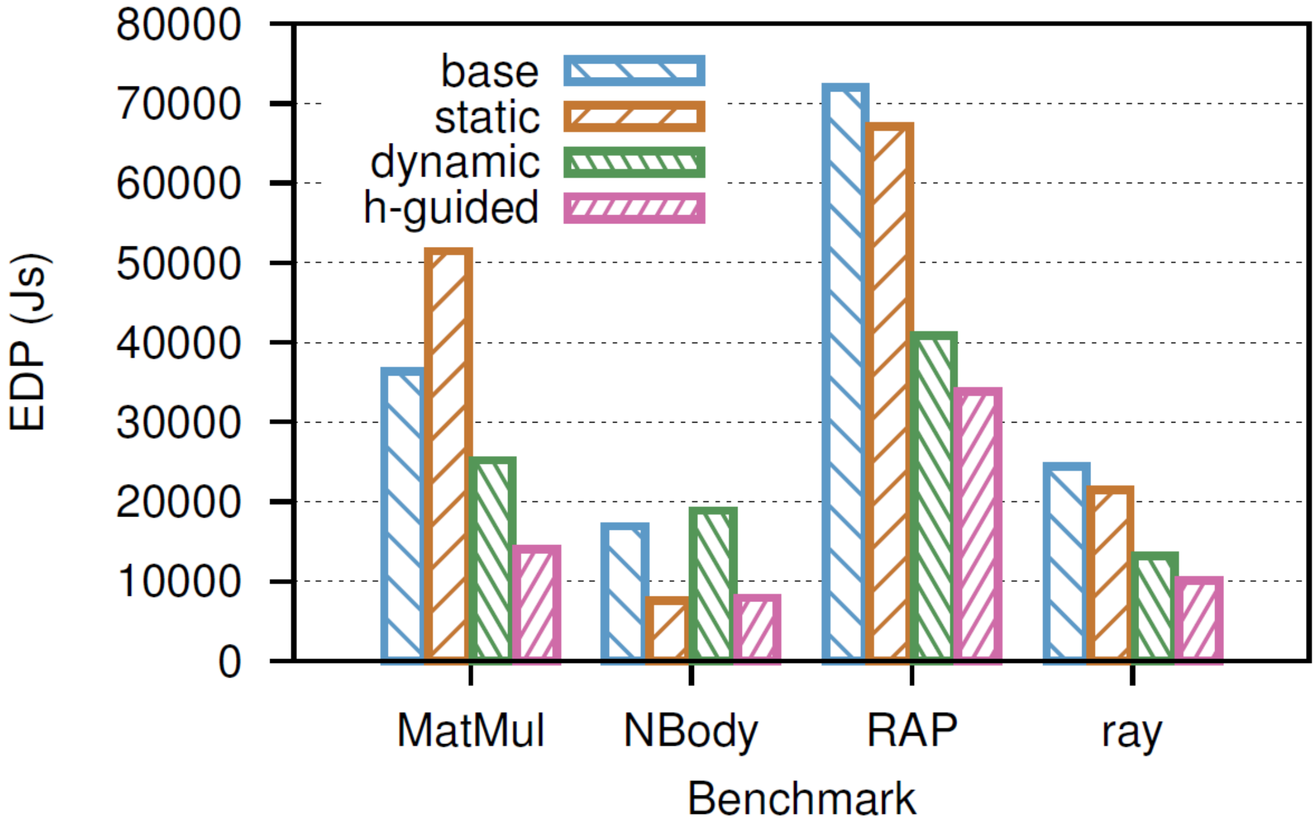
- EDP



- Confirms the previous results

Experimental Results

- EDP



- Improves even if the balancing is not the best

Conclusions

- Using all the available devices is worth it both performance-wise and efficiency-wise
  - Contrary to other authors
- There is always a load balancing approach that improves efficiency
  - Usually several
- Currently working on analyzing different frequencies for the GPU

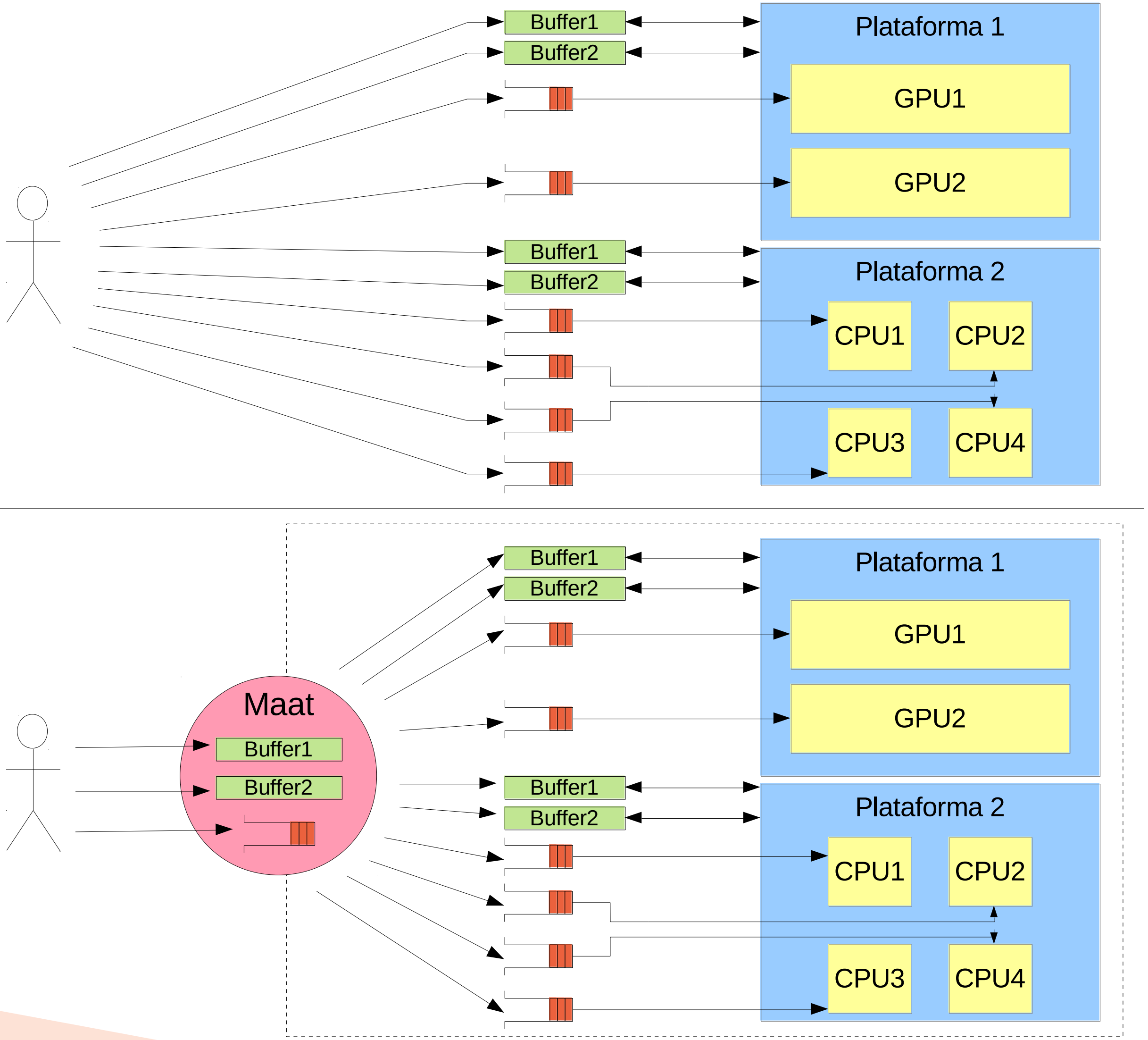
Energy Efficiency Evaluation in Heterogeneous Computers

Borja Pérez, Esteban Stafford, José Luis Bosque, Ramón Beivide

UNIVERSIDAD DE CANTABRIA

Abstraction of the system

- The programmer communicates with the whole system
- Transparent system management


Buffer1
Buffer2
Plataforma 1
GPU1
GPU2
Buffer1
Buffer2
Plataforma 2
CPU1
CPU2
CPU3
CPU4
Maat
Buffer1
Buffer2
Buffer1
Buffer2
Plataforma 1
GPU1
GPU2
Buffer1
Buffer2
Plataforma 2
CPU1
CPU2
CPU3
CPU4