eGPU for Monitoring Performance and Power Consumption on Multi-GPUs

XIII Workshop de Processamento Paralelo e Distribuído

John A. G. Henao (1), **Víctor M. Abaunza** (2)

Philippe O. A. Navaux (2), **Carlos J. B. Hernández**(1)

(1) **High Performance and Scientific Computing Center Industrial University of Santander**

(2) **Parallel and Distributed Processing Group, Informatics Institute, Federal University of Rio Grande do Sul**

August 21, 2015

Super Computación y
Cálculo Científico UIS

Introduction

The evaluation of performance
and power consumption
is a key step in
the design of applications
for large computing systems,
such as supercomputers, clusters
with nodes that have
manycores and multi-GPUs.

Background and Motivation

Develop a Monitor to analyze multiple tests
under different combinations of parameters to
observe the key factors that determine the energy efficiency
in terms of 'Energy per Computation'
on Cluster with Multi-GPUs.

Benchmark Used

- **The Standard Linpack widely used by the Green500 and the Top500.**
- **The linpack Benchmark HPL is representative for the applications that could be executed in large computing systems.**
- **The HPL allows test different combinations of parameters to find the performance numbers that reflect the largest problem can be run on a supercomputer.**

eGPU Monitor Structure

- **eGPU is formed by two levels:**

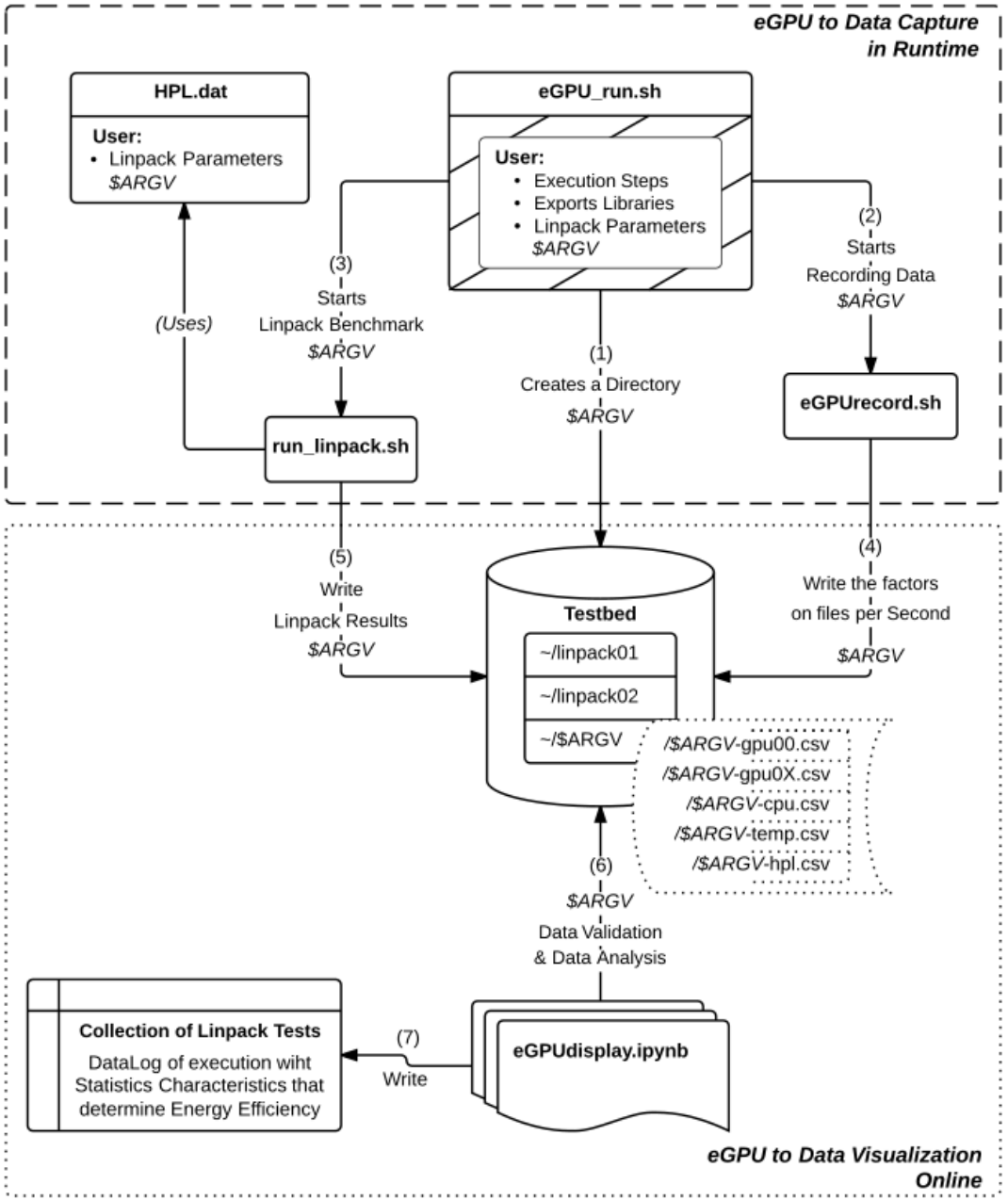
  I. **eGPU to Data Capture in runtime.**

  II. **eGPU to Data Vizualization online.**

- **Composed by 7 events:**

  1) **Data Centralization.**
  2) **Starts *eGPUrecord.sh.***
  3) **Starts *runlinpack.sh.***
  4) **Write Computational Factors.**
  5) **Write the Performance.**
  6) ***eGPUdisplay.ipynb* used at post-processing.**
  7) **Write the Statistical Characteristics**

eGPU Monitor Structure



Experimental Procedures and Results

- **The computational resources used: One node of the ‘A’ settings.**

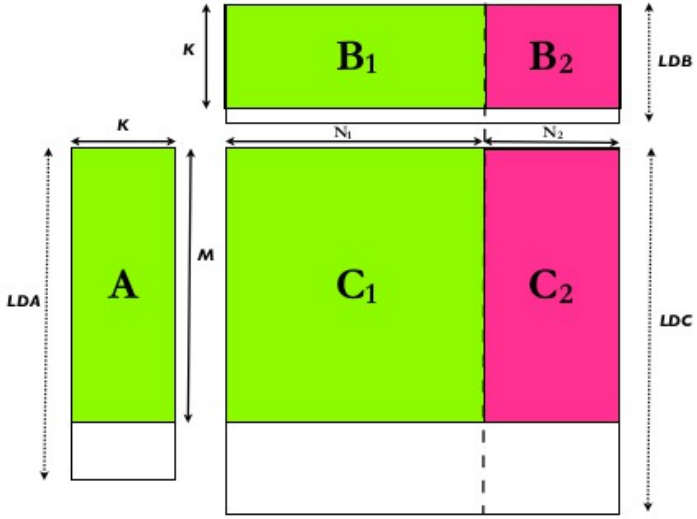
Setting GUANE	A	B	C
Node type	SL390s	SL390s	SL390s
Number of nodes	8	3	5
Processor Intel	Xeon	Xeon	Xeon
Processor Model	E5645	E5645	E5640
Processor by node (#)	2	2	2
Clock frequency (GHz)	2.40	2.40	2.670
Core/Processor (#)	6	6	4
Thread/Core (#)	2	2	2
GPUS Nvidia	Tesla	Tesla	Tesla
GPUS Model	M2075	M2050	M2050
GPUs by node (#)	8	8	8
Memory DDR (GB)	104	104	104
SAS disk (GB)	200	200	200
Gigabit Ethernet (Gbps)	10	10	10
InfiniBand (IB)	1	1	1

Experimental Procedures and Results

- **The Linpack used:**
  - **HPL.2.0 version configured for Tesla GPUs.**

Ref. Massimiliano Fatica. Accelerating linpack with CUDA on heterogenous clusters. ACM, 2009.

DGEMM: LU Factorization



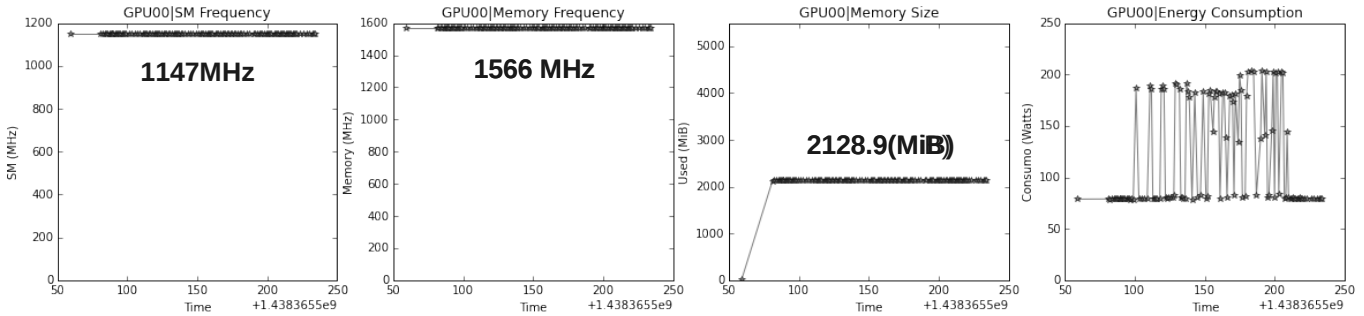
DGEMM('N','N',m,n1,k,alpha,A,lda,B1,ldb,beta,C1,ldc)

DGEMM('N','N',m,n2,k,alpha,A,lda,B2,ldb,beta,C2,ldc)

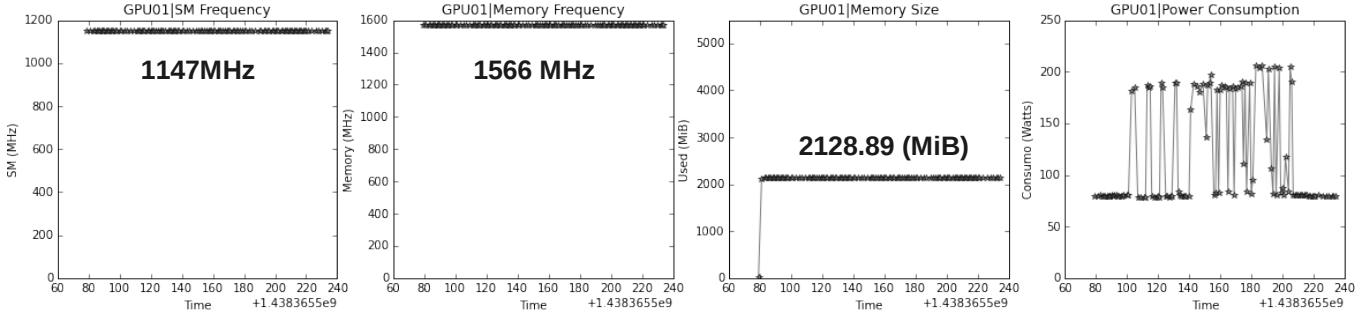
The Linpack parameters used

Matrix size	49152
Block size	1024
GPU Used	4
Cores per GPU	3
Process MPI	4

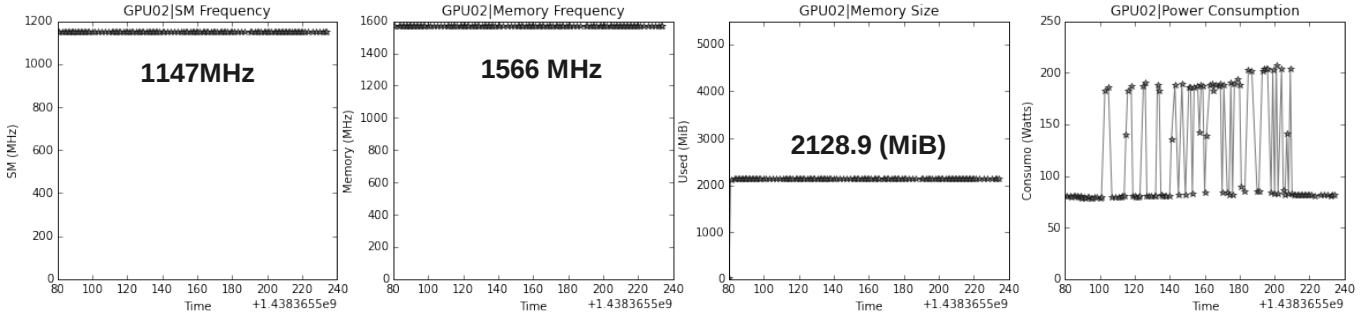
eGPU-Sequenceplot for 4 worker GPUs



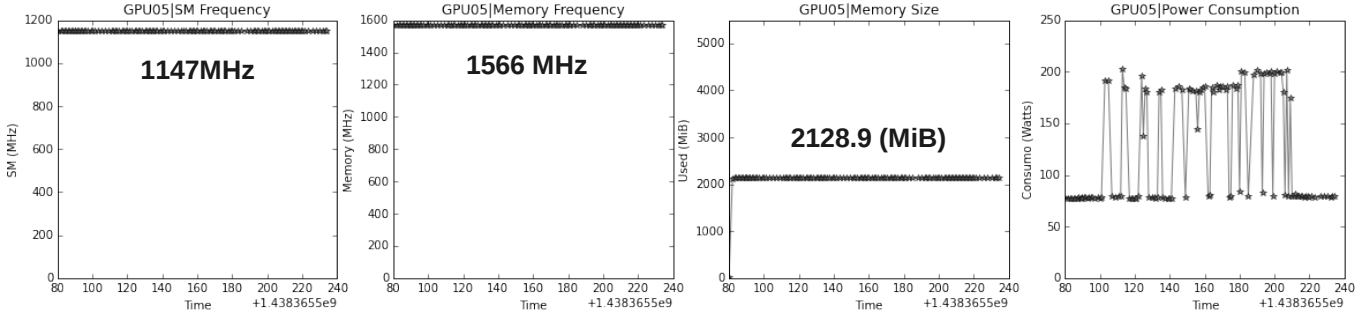
Mean: 120.55 (Watts)
Std: 51.29 (Watts)



Mean: 118.01 (Watts)
Std: 50.92 (Watts)

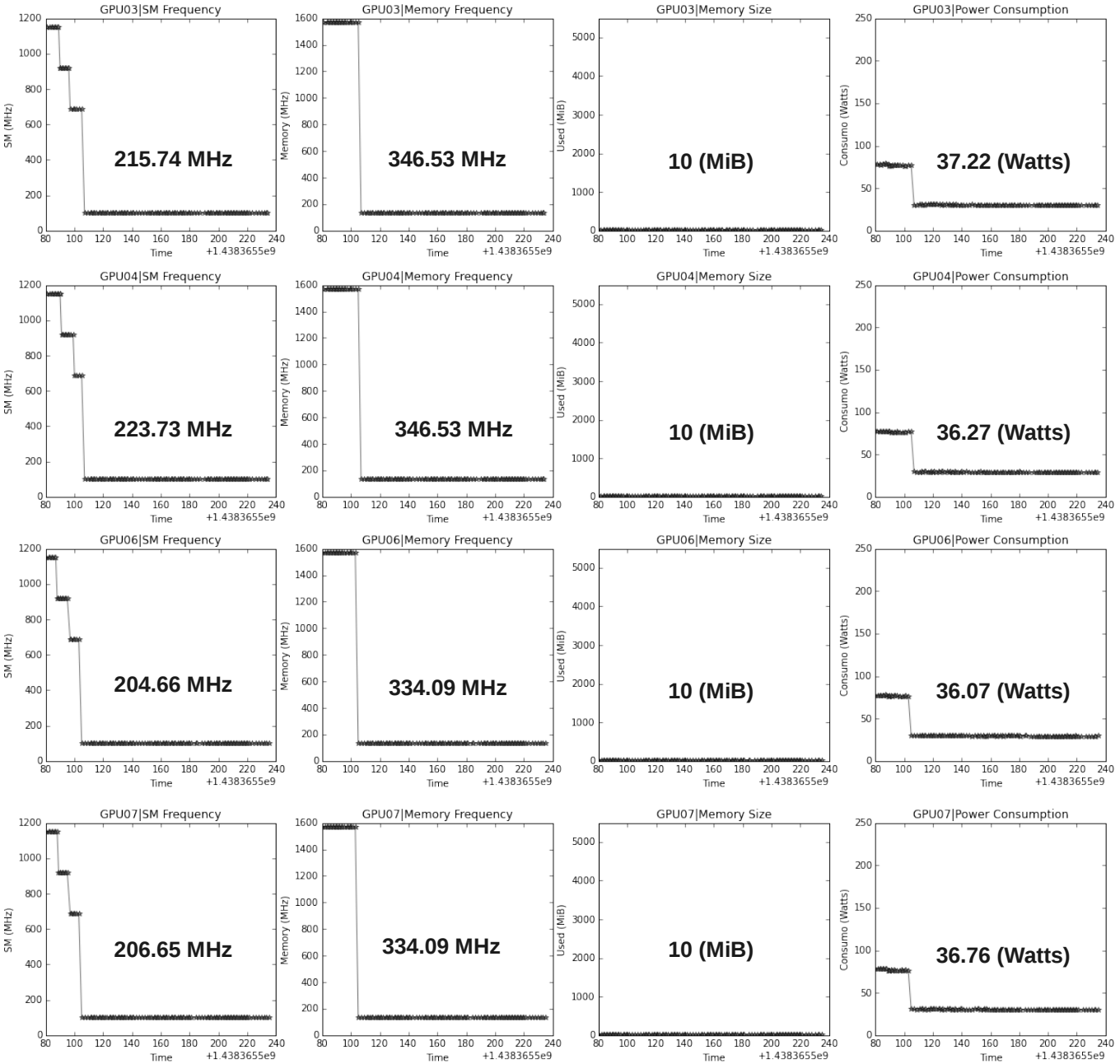


Mean: 119.12 (Watts)
Std: 51.36 (Watts)

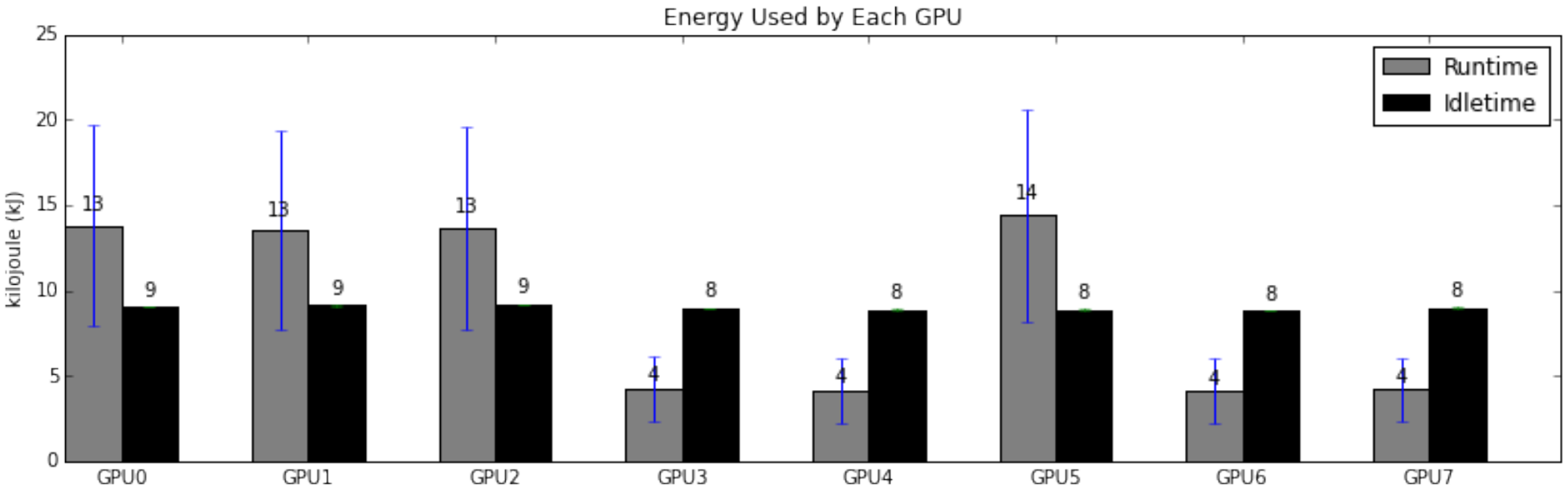


Mean: 125.65 (Watts)
Std: 54.32 (Watts)

eGPU-Sequenceplot for 4 idle GPUs



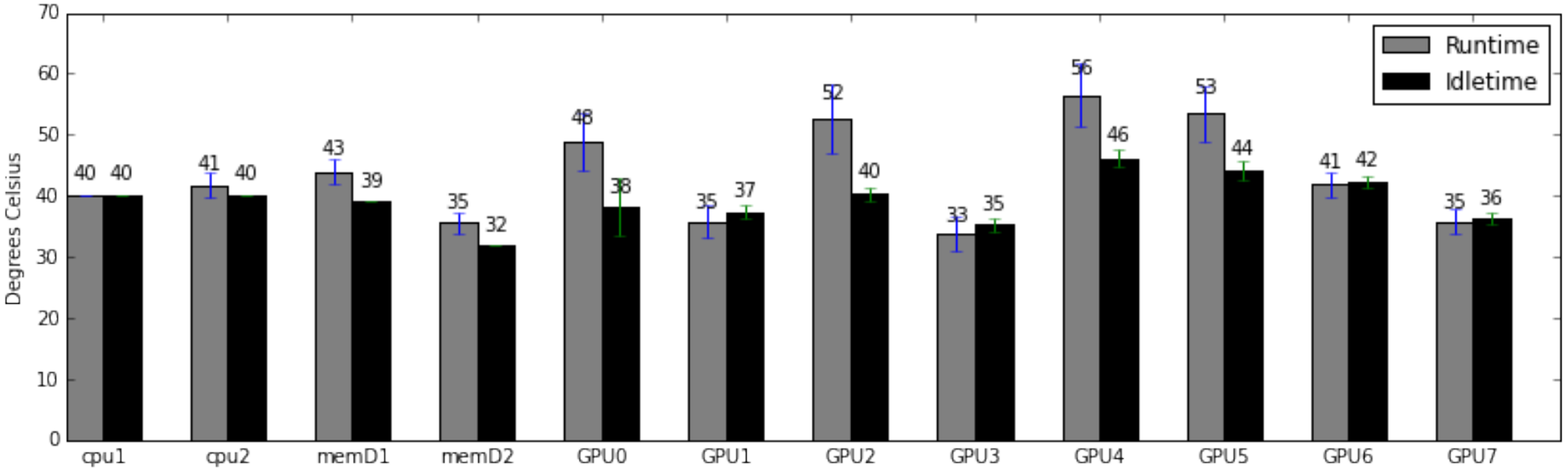
eGPU-Bar graph to Analysis of Energy



Energy Consumption Between Idle time and Runtime by each GPU.

- **Average Energy Consumption Idle time: 61Kj**
- **Average Energy Consumption Runtime: 69Kj**

EGPU-Bar graph to Analysis of Temperature



Temperature used in the node Between Idle time and runtime.

- **Average Temperature Idle time: 469 DC**
- **Average Temperature Runtime: 512 DC**

eGPU-Results

eGPUrecord Time:	121 sec
eGPUrecord Latency:	6 sec
HPL2.0 Time:	114.54 sec
HPL2.0 Performance:	691.20 GFLOPS
Power Consumption:	629.65 Watts
Energy Consumption:	72.12 kJ
Energy Efficiency:	1097.68 MFLOPS/W

eGPU writes a datalog by each test with Statistical Characteristics That determine of Energy Efficiency.

Conclusions

- **eGPU facilitates the collection and visualization of data to analyze many tests under different combinations of parameters and observe the granularity of the factors that determine energy efficiency in clusters with multi-GPUs.**

- **The method we use is focused ony analyzing previously compiled applications, where researchers do not need to orchestrate the code to execute eGPU, ensuring the integrity of the results.**

- **Based on the experiment procedures and results presented, eGPU is a good alternative to analyze power consumption in clusters with multi-GPUs from a software level, and can be complemented with other energy monitors that are designed to be plugged-in directly into the power supply to make holistic measures in clusters with multi-GPUs.**

Questions?

Obrigado pela sua atenção!

eGPU for Monitoring Performance and Power Consumption on Multi-GPUs

XIII Workshop de Processamento Paralelo e Distribuído