Máquinas simuladoras de circuitos. Diseño y desarrollo de arquitecturas de simulación

J. Ranilla (Quantum and High-Performance Computing group, http://qhpc.uniovi.es)

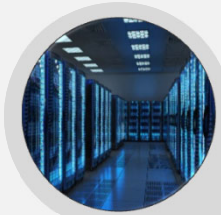
Universidad de Oviedo

QHPC GROUP

ACERCA DE NOSOTROS

Combarro E.F. Cortina R. Muñiz R. Muñoz A. Ranilla J. Revuelta P.

1998
Parallel Computing Group



2001
IA (Fuzzy, IR, TC, ML)

2004
Information Retrieval and Parallel Computing Group

2018
Quantum and High-Performance Computing Group

http://qhpc.uniovi.es
ranilla@uniovi.es

¿Por qué esta Presentación?

- Interés, demanda, ¿*esperanza*?, etc. en la CC

- Los ordenadores cuánticos tienen limitaciones
  - Disponibilidad y costes
  - QoS (errores y otros)

- La simulación tiene ventajas
  - Especialmente en la fase de desarrollo

- La simulación también tiene desventajas
  - Escalabilidad
  - Tiempo de respuesta

ESTA PRESENTACIÓN

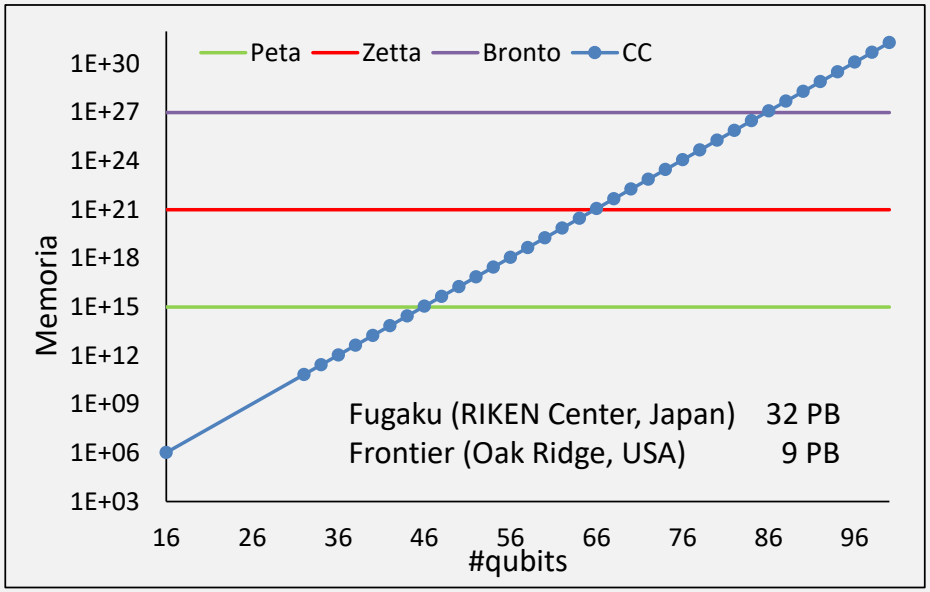
- No es una discusión sobre modelos: exactos, aproximados, generalistas, específicos, etc.
- No es una discusión sobre ¿es mejor nube o local?
- No es una disquisición sobre ¿el mejor software de simulación es…?
- La síntesis de un "trabajo": *General-purpose Quantum Circuit Simulator System Design (GQCS^2)*
  - Iniciado en 2018
  - NISQ (*noisy intermediate-scale quantum computing*)
  - *Full simulation,* Qiskit, etc.
- Y de su evolución para incorporar QoS y consumo energético

GQCS²-QSV

***G**eneral-purpose **Q**uantum **C**ircuit **S**imulator **S**ystem based on **Q**uantum **S**tate **V**ector*

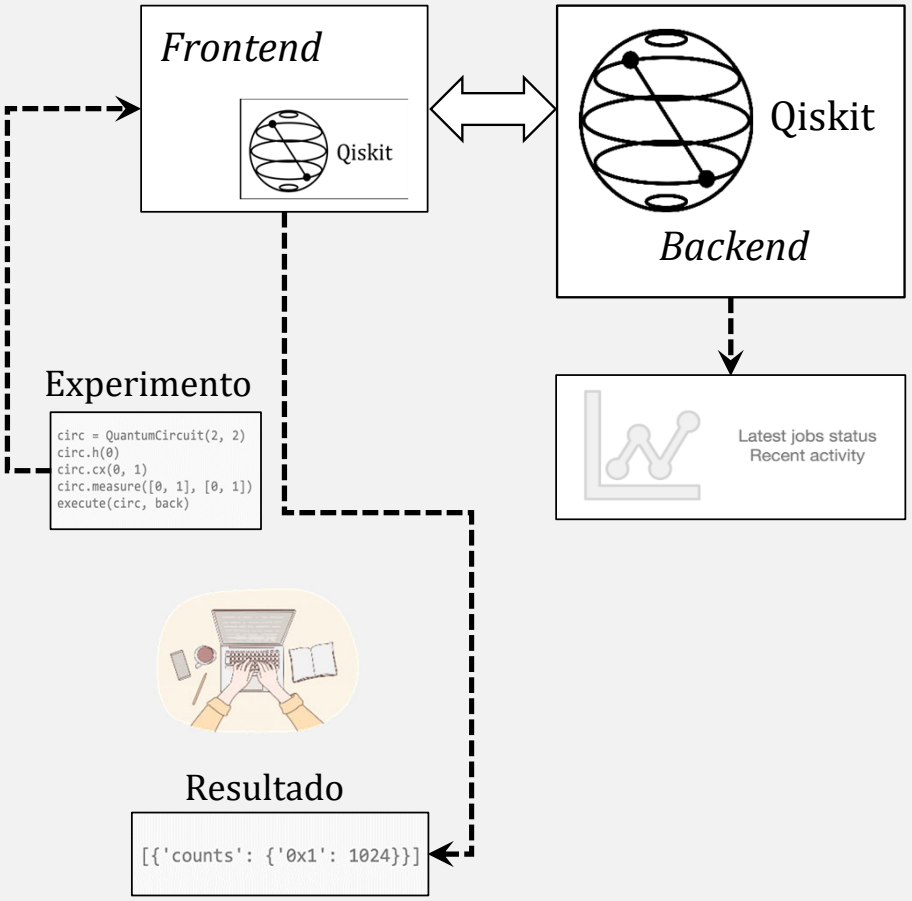
- *Memory Bound*

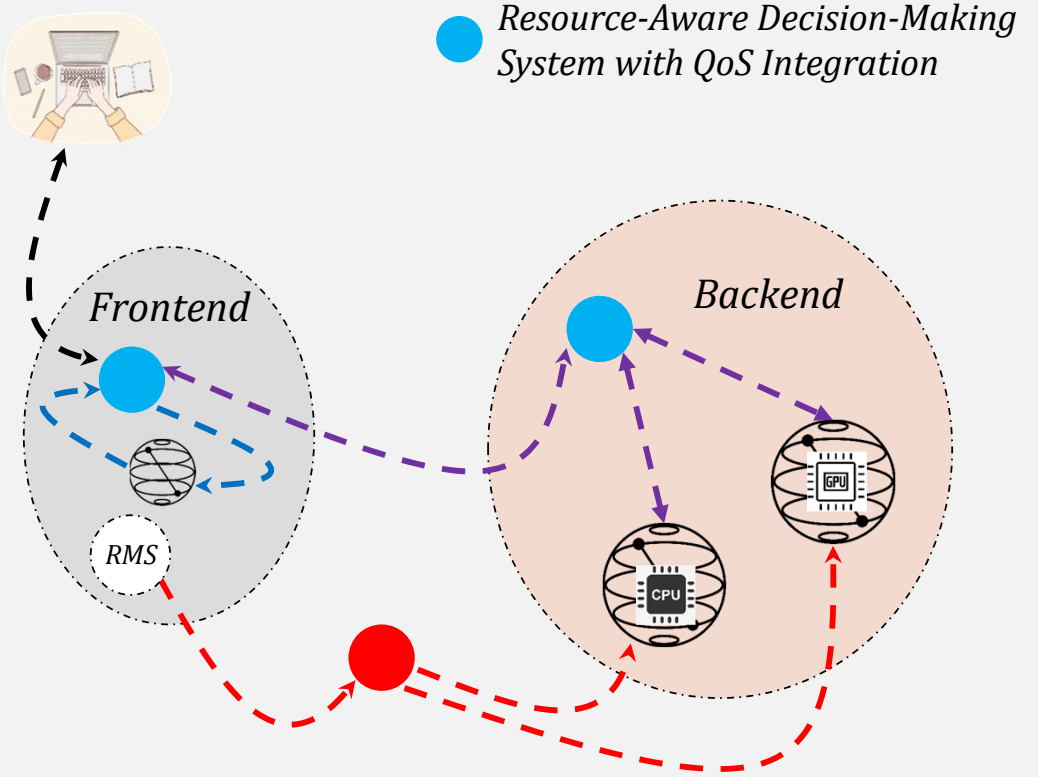
#qubits	Memoria RAM			
	MB	GB	TB	PB
16	1			
32		64		
38			4	
40			16	
46				1
50				16



- *¿Compute Bound?*
  - No necesariamente: núcleos matemáticos, *index bit swaps*, etc. Generalmente basados en librerías altamente eficientes y desarrollos específicos (p. ej. "*high-performance micro-kernels for matrix multiplication*")

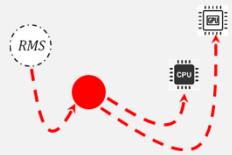
ARQUITECTURA



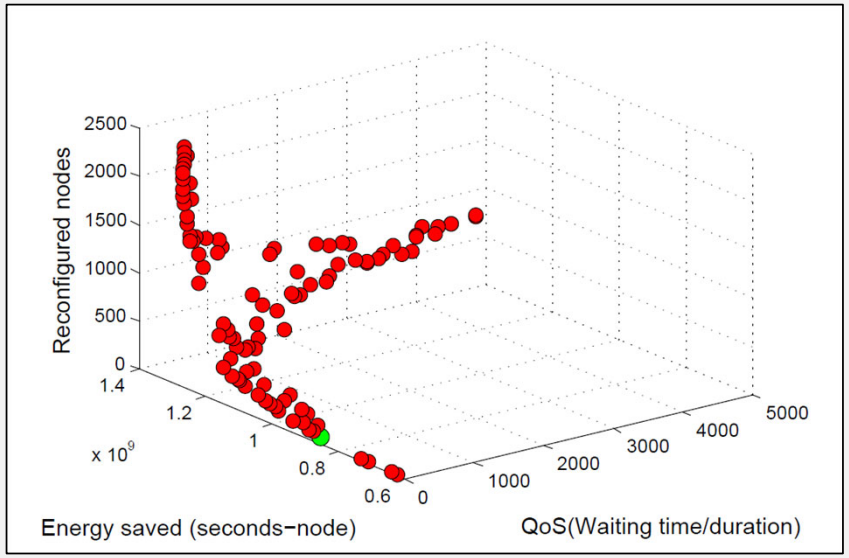


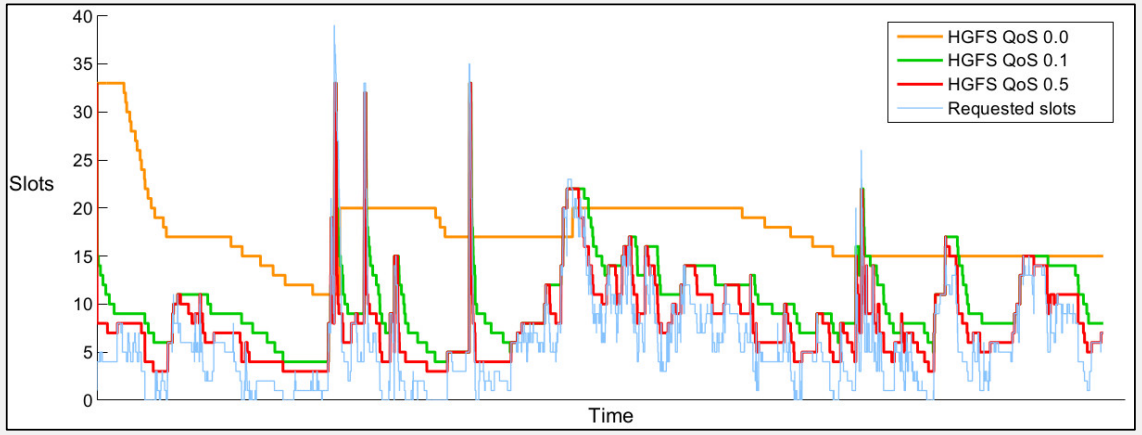
CODE-RPM

Energy-Conscious Decision System embedded in the RMS Power Module



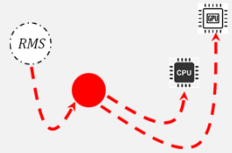
- **SOTA**: *EECluster*
  - Objetivo: optimizar el número de nodos activos y minimizar la fatiga del hardware
  - Basado en sistemas híbrido genético-difusos con pareto optimizado para QoS y carga
  - Efectivo pero difícil de entrenar y baja capacidad de adaptación a cambios



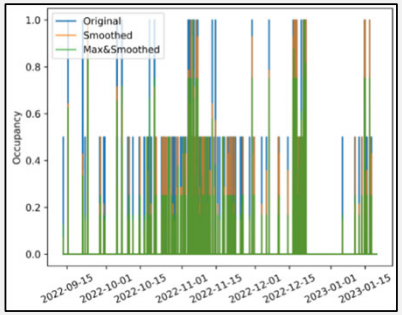


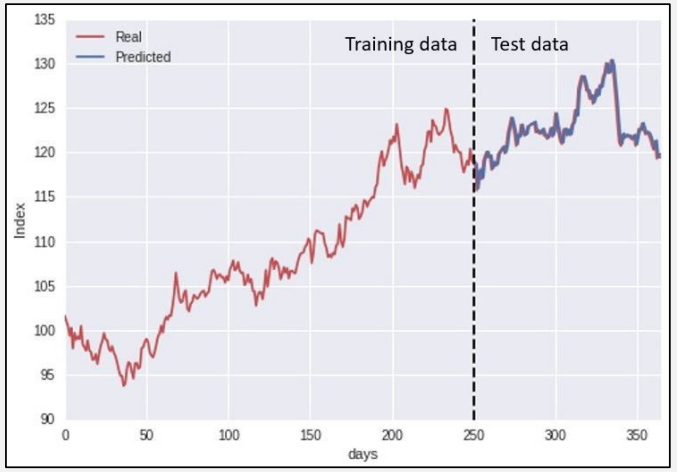
CODE-RPM

Energy-Conscious Decision System embedded in the RMS Power Module



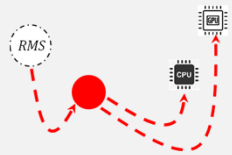
- **RECODE-RPM**
  - Plantea la predicción de la carga como problema de regresión
  - Trata la carga como valores numéricos y, basándose en la ventana de contexto, predice un horizonte





CODE-RPM

Energy-Conscious Decision System embedded in the RMS Power Module



- **NECODE-RPM**
  - Aborda la predicción como un problema de clasificación combinando NLP y LSTM
  - *Embedding layer* basada en el algoritmo *Word2Vec*

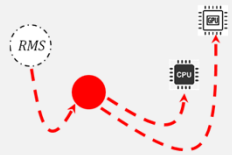
10 labels	4 labels	3 labels
D10 → [1.00,0.9)	VH (Very High) →[1.00,0.75)	H (High) -> [1.00,0.66)
D9 → [0.9,0.8)	H (High) → [0.75,0.5)	M (Medium) -> [0.66,0.33)
… … … … … … … … …	L (Low) → [0.5,0.25)	L (Low) -> [0.33,0]
D1 → [0.1,0]	VL (Very Low) → [0.25,0]	

  - Para comparar los resultados de la regresión se traducirán a símbolos

Resultado de regresión	Equivalente en Clasificación
0.25	L
0.60	M
0.80	H

CODE-RPM

Energy-Conscious Decision System embedded in the RMS Power Module



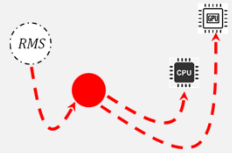
- **RECODE-RPM**

		ZeroR	RandomForest	SVM	XGBoost	LSTM	NLP
C3H3	Xeon	0.348	0.063	0.081	0.063	**0.038**	~
	I3	0.342	0.055	0.075	0.056	**0.031**	~
	GPU	0.194	0.058	0.07	0.058	**0.032**	~
C4H4	Xeon	0.35	0.076	0.092	0.077	**0.044**	~
	I3	0.34	0.065	0.082	0.066	**0.035**	~
	GPU	0.193	0.067	0.078	0.067	**0.036**	~
C6H3	Xeon	0.35	0.059	0.079	0.06	**0.034**	~
	I3	0.338	0.054	0.074	0.054	**0.028**	~
	GPU	0.192	0.056	0.071	0.055	**0.029**	~

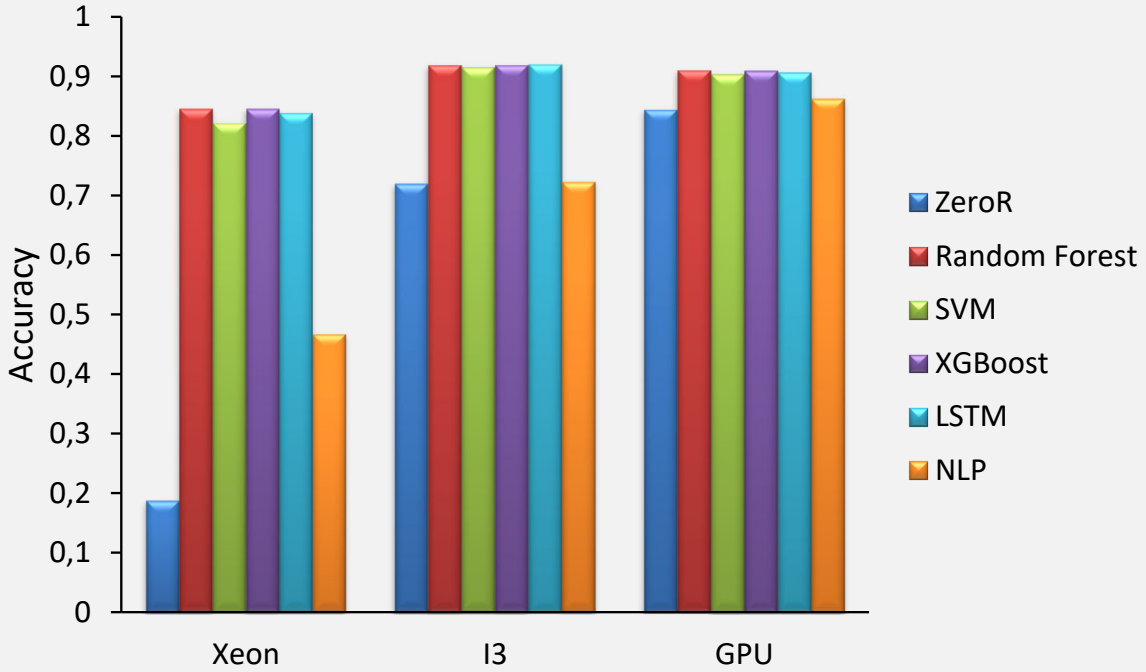
Table 6.1: Summary Table RMSE Cross Validation
(NLP being a classification model predicts discrete labels thus its RMSE is not comparable)

CODE-RPM

Energy-Conscious Decision System embedded in the RMS Power Module



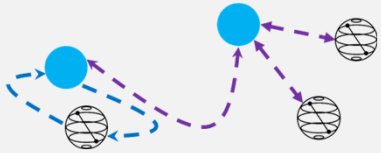
- **RECODE-RPM *vs.* NECODE-RPM**



- Futuro: ¿Mejorar los modelos de IA con aprendizaje por refuerzo?

RADMY-QoS

Resource-Aware Decision-Making System with QoS Integration

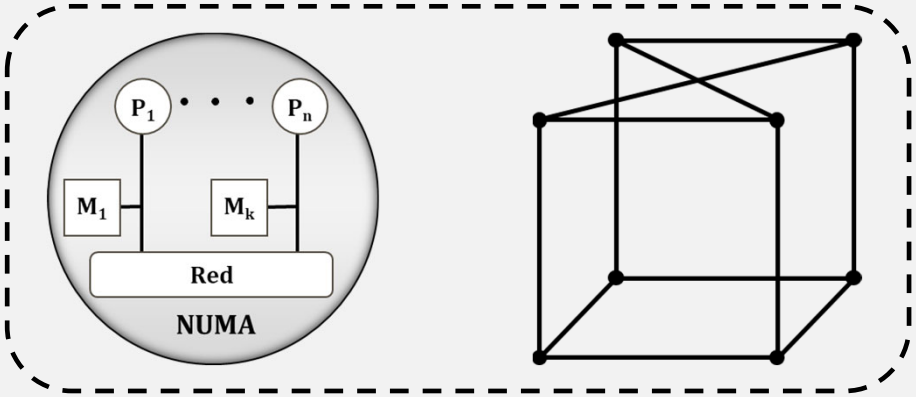


- Decide dónde y cómo se ejecutan las simulaciones
- Fase embrionaria
- Jerárquico
  - Primer nivel: Frontend vs. Backend
  - Segundo nivel: CPU *vs.* GPU
- Reglas
  - Pocas y sencillas. Diferentes según el nivel. *Condicionadas por…*
  - Basado en aspectos cuantitativos
- Futuro
  - Trabajo similar a **CODE-RPM**
  - ¿Por qué no QML? Hay pocas variables y datos

FINE-TUNE THE SETTINGS

El Hardware

8x Intel Xeon Platinum
8 TB RAM
2x TESLA P100



Cada CPU 1 TB RAM con 3x Intel UltraPath Interconnect a 10.4 GT/s máximo en cada dirección

- ¿Algo más a considerar?
  - *Runtime thread migration*
  - *Oversubscription threads – cores*
  - Etc.
- Es necesaria una “labor de campo”

FINE-TUNE THE SETTINGS

```
from qiskit.circuit.random import random_circuit
from qiskit.providers.aer import *
import numpy as np
seed = 2021
shots = 1024
n_layers =10
n_qubits = range(15, 39, 1)
reps = range(0,  10, 1)
n_proc   = [1, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 88, 96, 104, 112, 120, 128]
qctic_backend = Aer.get_backend("aer_simulator_statevector")
for n in n_qubits:
    for p in n_proc:
        qctic_backend.set_options(max_parallel_threads=p)
        circs = []
        for i in reps:
            circ = random_circuit(n, n_layers, measure = True, seed = 2021 + i)
            circs.append(circ)
        qctic_job = execute(circs, qctic_backend, memory=False, shots=shots)
        result = qctic_job.result()
        times = [result.results[i].time_taken for i in reps]
        print("nqubits=\t",n,"\tnproc=\t",p,"\ttime=\t",np.mean(times))
    print("")
```

```
#!/bin/sh
export OMP_DISPLAY_ENV=true
export OMP_PLACES=xxxx
export OMP_PROC_BIND=yyyy
python Test.py
```

< 36 qubits ≅ 1TB
< 37 qubits ≅ 2 TB
< 38 qubits ≅ 4 TB

Rojo → Board switch
Azul → System saturation

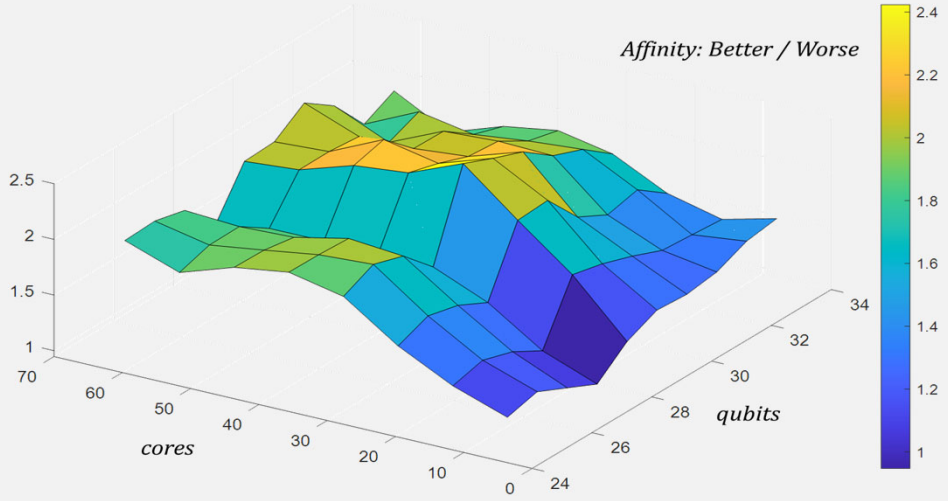
Affinity / Migration
Bind: Close, Spread, etc.
Places: cores, true, etc.
HT vs. No HT

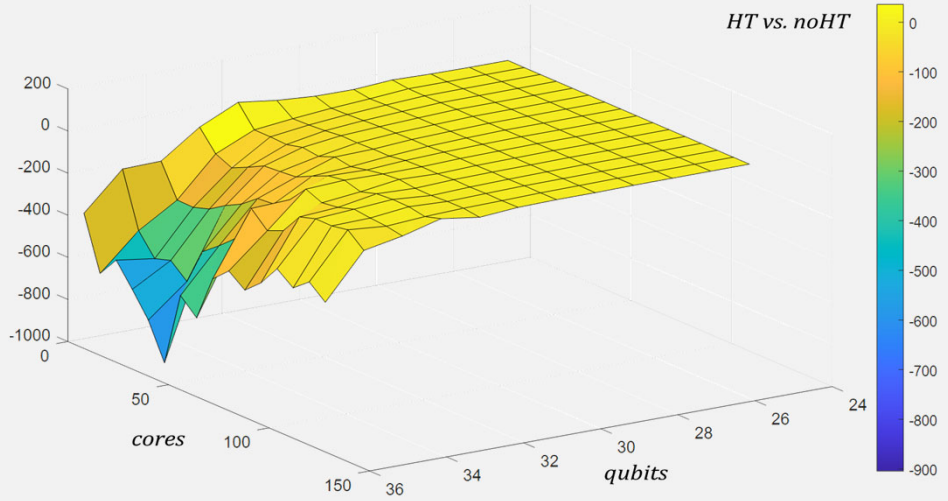
FINE-TUNE THE SETTINGS

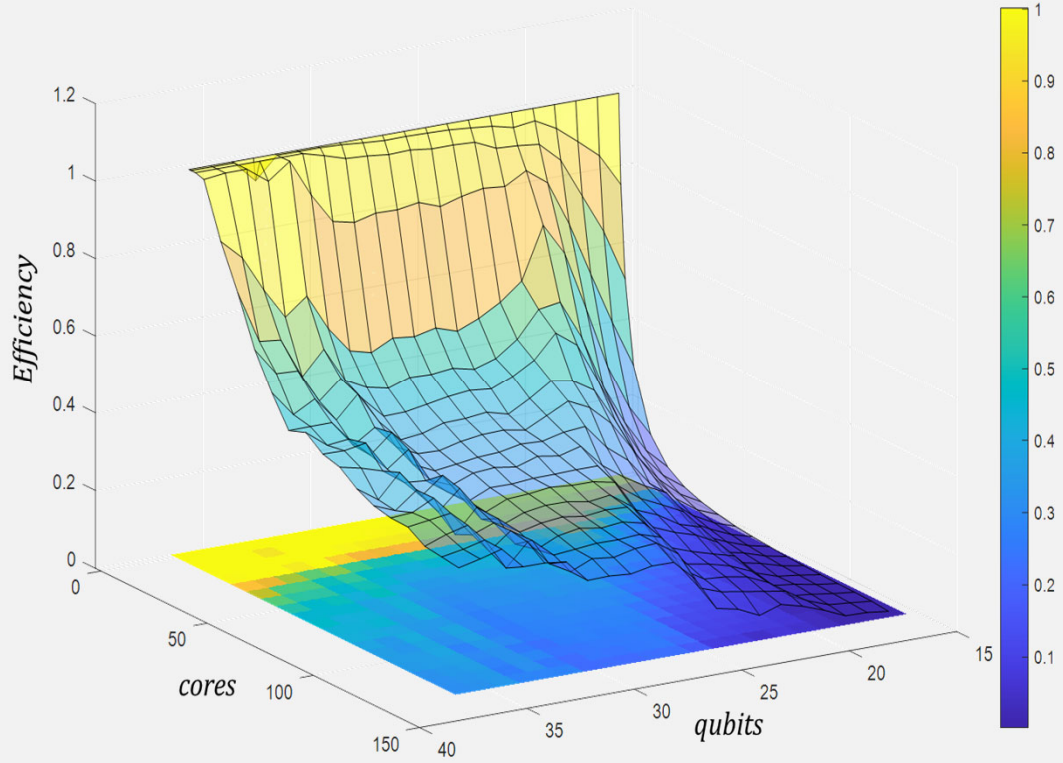
Qubits	Cores	Close	Spread	Loss function
25 (512 MB)	2	7,947	8,426	-5,68%
	4	4,122	4,920	-16,23%
	8	2,411	2,905	-17,02%
	16	1,910	1,668	12,64%
	24	1,498	1,274	14,96%
	32	1,208	0,970	19,70%
	40	1,083	0,860	20,60%
	48	0,954	0,797	16,47%
	56	0,836	0,738	11,69%
	64	0,768	0,666	13,27%
	72	0,714	0,757	-5,57%
	80	0,663	0,727	-8,80%
	88	0,630	0,704	-10,52%
	96	0,598	0,662	-9,74%
	104	0,595	0,729	-18,36%
	112	0,586	0,725	-19,07%
	120	0,599	0,760	-21,19%
	128	0,700	0,744	-5,90%
27 (2 GB)	2	35,391	37,478	-5,57%
	4	18,268	21,349	-14,43%
	8	10,593	12,700	-16,59%
	16	8,419	7,048	16,29%
	24	6,582	5,682	13,67%
	32	5,383	4,234	21,34%
	40	4,715	3,765	20,14%
	48	4,154	3,263	21,45%
	56	3,821	3,058	19,97%
	64	3,518	2,924	16,87%
	72	3,207	3,274	-2,06%
	80	2,904	3,121	-6,97%
	88	2,784	2,972	-6,31%
	96	2,705	2,854	-5,20%
	104	2,579	2,757	-6,45%
	112	2,476	2,715	-8,83%
	120	2,323	2,688	-13,56%
	128	2,400	2,604	-7,81%

Qubits	Cores	Close	Spread	Loss function
29 (8 GB)	2	143,764	153,892	-6,58%
	4	74,124	86,688	-14,49%
	8	44,094	54,974	-19,79%
	16	35,991	28,858	19,82%
	24	28,038	21,449	23,50%
	32	23,185	17,145	26,05%
	40	20,353	15,519	23,75%
	48	18,129	14,966	17,45%
	56	16,193	13,365	17,47%
	64	14,238	12,135	14,77%
	72	13,337	13,727	-2,84%
	80	12,315	12,976	-5,09%
	88	11,734	12,627	-7,08%
	96	10,984	11,677	-5,93%
	104	10,568	11,019	-4,10%
	112	10,329	10,511	-1,73%
	120	9,754	10,835	-9,97%
	128	10,101	10,434	-3,19%
31 (32 GB)	2	678,126	745,928	-9,09%
	4	338,472	442,073	-23,44%
	8	190,517	336,992	-43,47%
	16	148,329	140,362	5,37%
	24	122,851	104,535	14,91%
	32	97,670	83,011	15,01%
	40	85,672	77,980	8,98%
	48	79,950	67,523	15,54%
	56	73,382	63,554	13,39%
	64	67,998	57,650	15,22%
	72	62,086	61,554	0,86%
	80	56,149	59,825	-6,15%
	88	55,799	54,075	3,09%
	96	47,315	53,779	-12,02%
	104	51,058	48,700	4,62%
	112	42,353	48,287	-12,29%
	120	45,542	44,127	3,11%
	128	38,694	50,316	-23,10%

FINE-TUNE THE SETTINGS

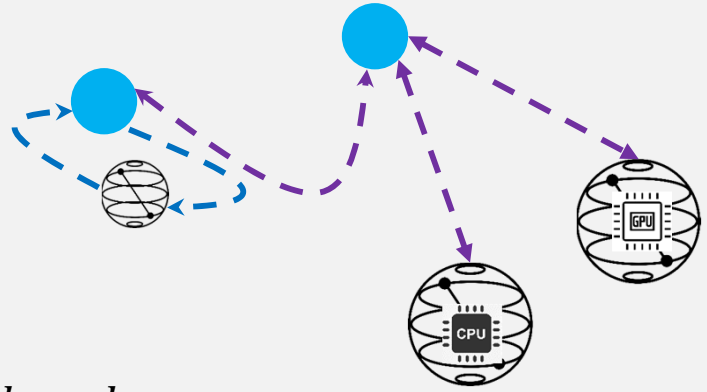






FINE-TUNE THE SETTINGS

RADMY-QoS



- Reglas… *Condicionadas por…*
  - Definir colas de alta y baja prioridad
  - Si el $\#qubits<x\rightarrow$ ejecutar en frontend (o nodo usuario)
  - Si el $\#qubits<y\rightarrow$ ejecutar en GPU
  - En la CPU
    - Los circuitos con consumo $< 1TB \rightarrow$ QSV en RAM de una única *board*
    - Mayor productividad (incluso QoS) limitando el #cores (cuando sea posible)
    - Ajustar dinámicamente (por circuito) el tipo de afinidad
    - Siempre libre un #cores o el sistema colapsará
    - Etc.

- Cambios “cualitativos” en el hardware → Nuevo ajuste → nuevas reglas

FINE-TUNE THE SETTINGS

¿Qué faltaría?

- Software
  - Software de terceros, poco se puede (debe) hacer (p. ej. ¿modificar Qiskit?)
  - Revisar la calidad de las librerías soporte

- GPUs
  - No todo es susceptible de ser ejecutado eficientemente en GPU

$$T(n,p,k) = T_{SendGPU}(n) + T_{RecGPU}(n) + T_{Exchange}(n) + \max\left(T_{ar\_GPU}(n,p,k), T_{ar\_CPU}(n,p,k)\right)$$

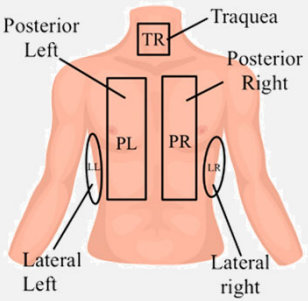
- Más seguro, pero hasta aquí se ha llegado

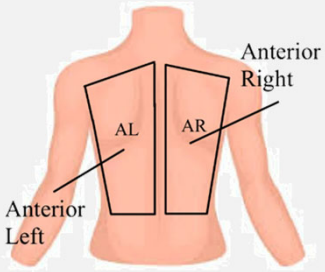
¿DÓNDE LO HEMOS USADO?

- *Classification of Heart Sounds using Quantum Machine Learning Models*. 5th International Conference on Advances in Signal Processing and Artificial Intelligence, 7-9 June, Tenerife 2023
- *Quantum Neural Networks in Lung Sounds Analysis: Evaluating Performance with Signal Representation Techniques*. 54º Congreso Español de Acústica, 18-20 octubre, Cuenca 2023

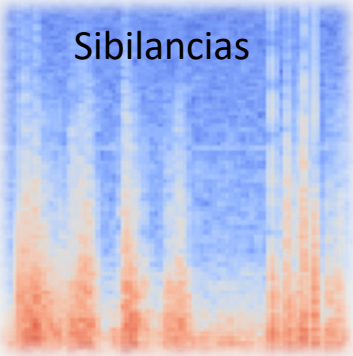
La BBDD: ICBHI

- Universidades de Aveiro, Thessaloniki y Coimbra
- 920 audios, 6898 ciclos respiratorios. Ruidosa, mal balanceada, compleja
- 7 zonas de grabación, 4 dispositivos de grabación





	N.º Ciclos
Sanos	860
Crepitaciones	704
Sibilancias	212
Ambos	183



¿Dónde lo hemos Usado?

Preprocesado

- Preénfasis
- Enventanado
- Hamming Window

Hiperparámetros

- N.º de puntos (STFT) $\in [64, \dots, 4096]$
- % solapamiento (STFT, MFCC, CCGR) $\in [10, \dots, 90]$
- Banco de Filtros (MFCC) $\in [10, \dots, 100]$
- N.º coeficientes (MFCC, CCGR) $\in [10, \dots, 30]$

$\cong$ 50.000 características

Representaciones de la señal

- Short-Time Fourier Transform (STFT)
- Mel-Frequency Cepstral Coefficients (MFCC)
- Cochleograms (CCGR)

Método de Selección

- Validación cruzada estratificada 10 *folders*
- SVM con configuración por defecto (estática)

Resultado

- MFCC
- N.º de coeficientes 18, N.º filtros 20
- Solapamiento 10%

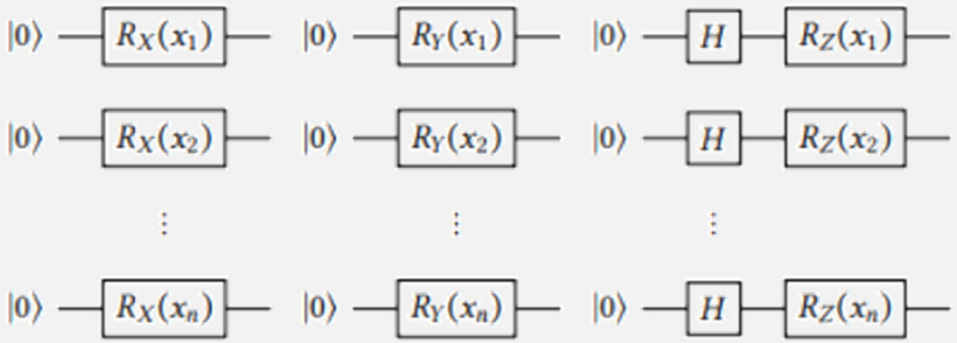
¿DÓNDE LO HEMOS USADO?

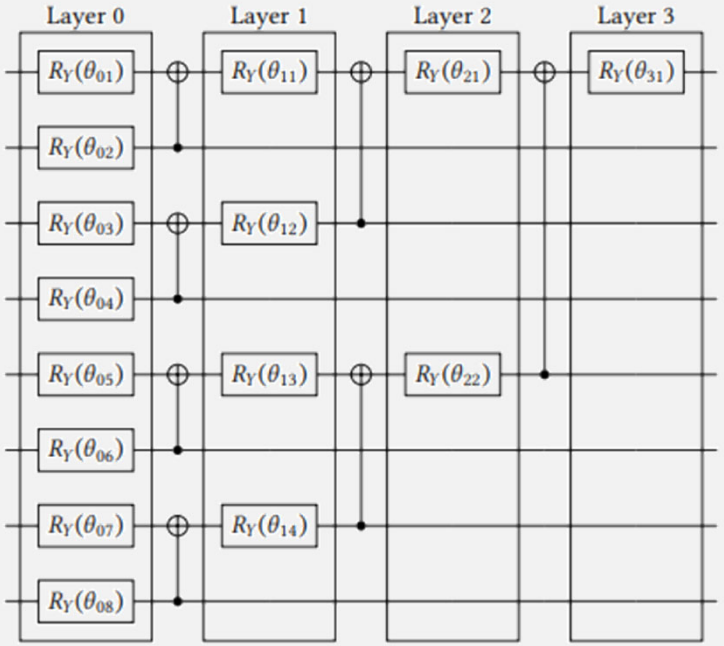
NN

- Dense(num_units, relu) + Dense(num_classes, softmax) / Dense(1, sigmoid)
- Dense(16, relu) + Dense(2, relu) + Dense(num_classes, softmax) / Dense(1, sigmoid)
- Earlystoping; funciones de perdida: Categorical y Binary CrossEntropy; optimización: Adam
- Hiperparámetros: batch_size, num_epochs, patiencie, learning rate, num_units

QNN

Feature map: Angle Encoding





Forma variacional Tree Tensor

¿Dónde lo hemos Usado?

Resultados

Modelo	Sistema	Aciertos Test
MFCC	NN sin límite de parámetros	65,6%
MFCC	NN con PCA(n_components = nqubits_ang = 16)	60,7%
MFCC	QNN con 16 qubits y batch size el usado en NN	**66,1%**

El otro tipo de resultados

- Aceleración ≅ 5x
- La desviación típica reducida en un orden de magnitud