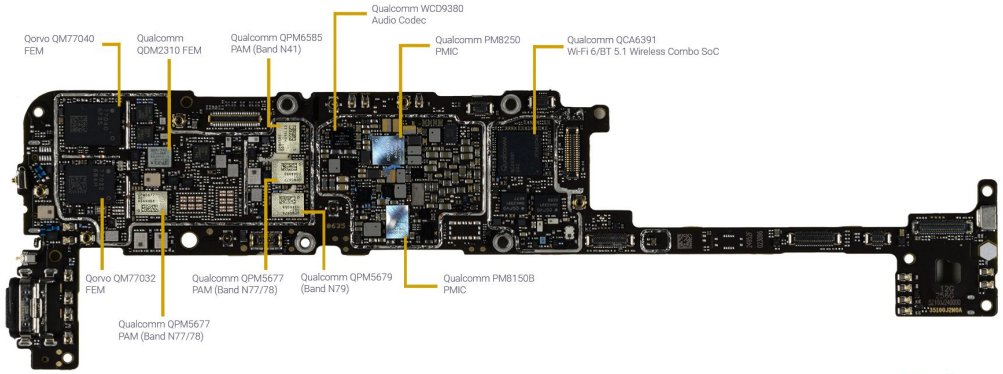
GPUs para ML

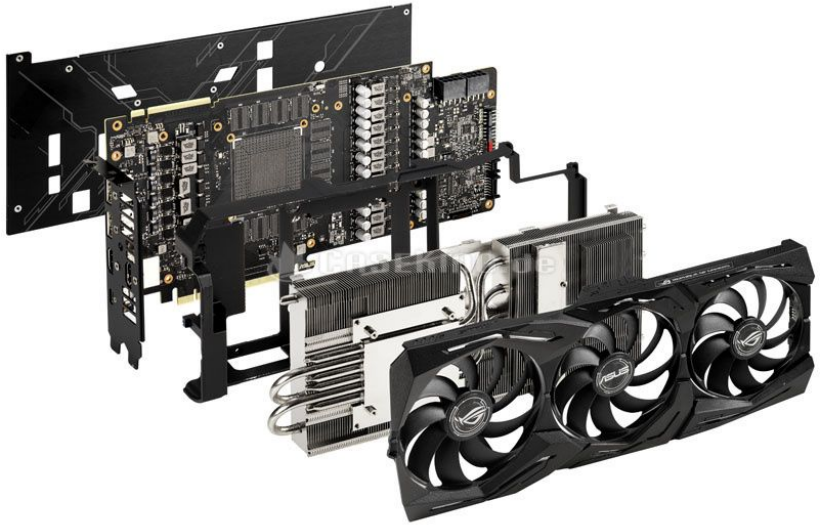
Un Enfoque Práctico

Centro de Computación de Alto Desempeño

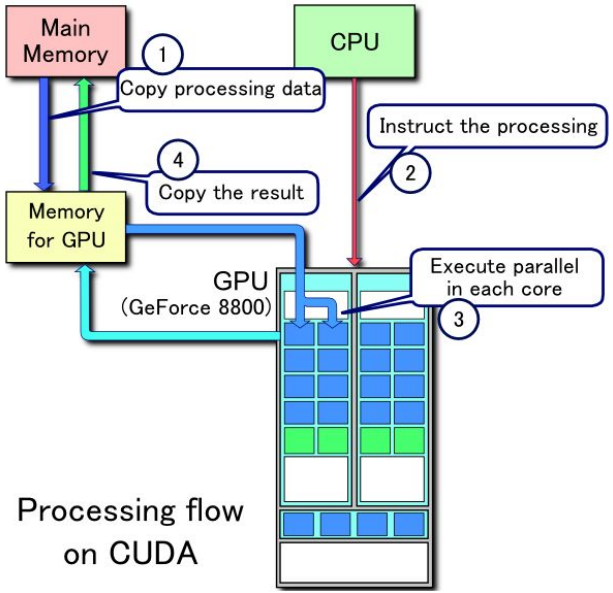
Nicolás Wolovick
GPGPU-FaMAF, CCAD-UNC



Tech Insights



Coprocesador esclavo de la CPU



Roofline model CPU vs. GPU

	TFLOPS float32 peak	GB/s peak
AMD EPYC 7532	2.44	204
TU102 (RTX 2080 Ti)	11.75	616
	5x	3x

CPU <-> GPU, 1 orden de magnitud abajo.

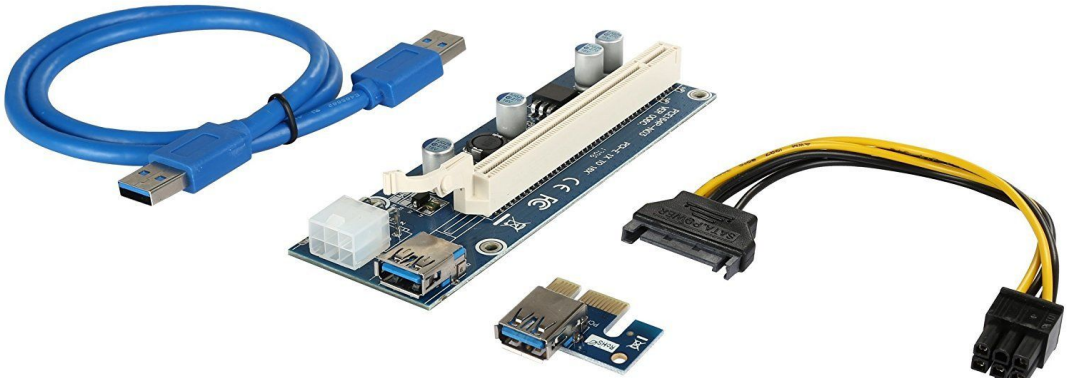
PCIe Bandwidth & Frequency

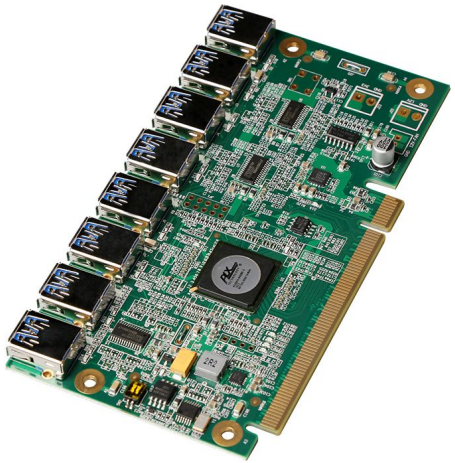
Year	Bandwidth	Frequency/Speed
1992	133MB/s (32 bit simplex)	33 Mhz (PCI)
1993	533MB/s (64 bit simplex)	66 Mhz (PCI 2.0)
1999	1.06GB/s (64 bit simplex)	133 Mhz (PCI-X)
2002	2.13GB/s (64 bit simplex)	266 Mhz (PCI-X 2.0)
2002	8GB/s (x16 duplex)	2.5 GHz (PCIe 1.x)
2006	16GB/s (x16 duplex)	5.0 GHz (PCIe 2.x)
2010	32GB/s (x16 duplex)	8.0 GHz (PCIe 3.x)
2017	64GB/s (x16 duplex)	16.0 GHz (PCIe 4.0)
2019	128GB/s (x16 duplex)	32.0 GHz (PCIe 5.0)

achtung, achtung!

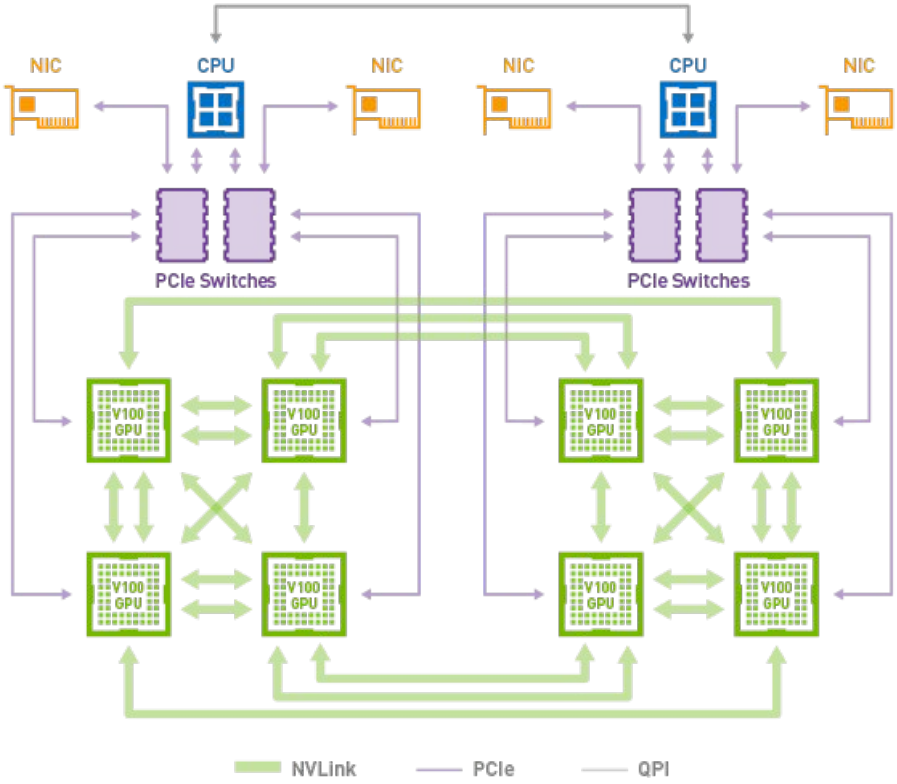
Achtung!

- PCIe lanes de CPU(s).
- PCIe slots físicos.
- Separación.
- Conexión a la CPU.

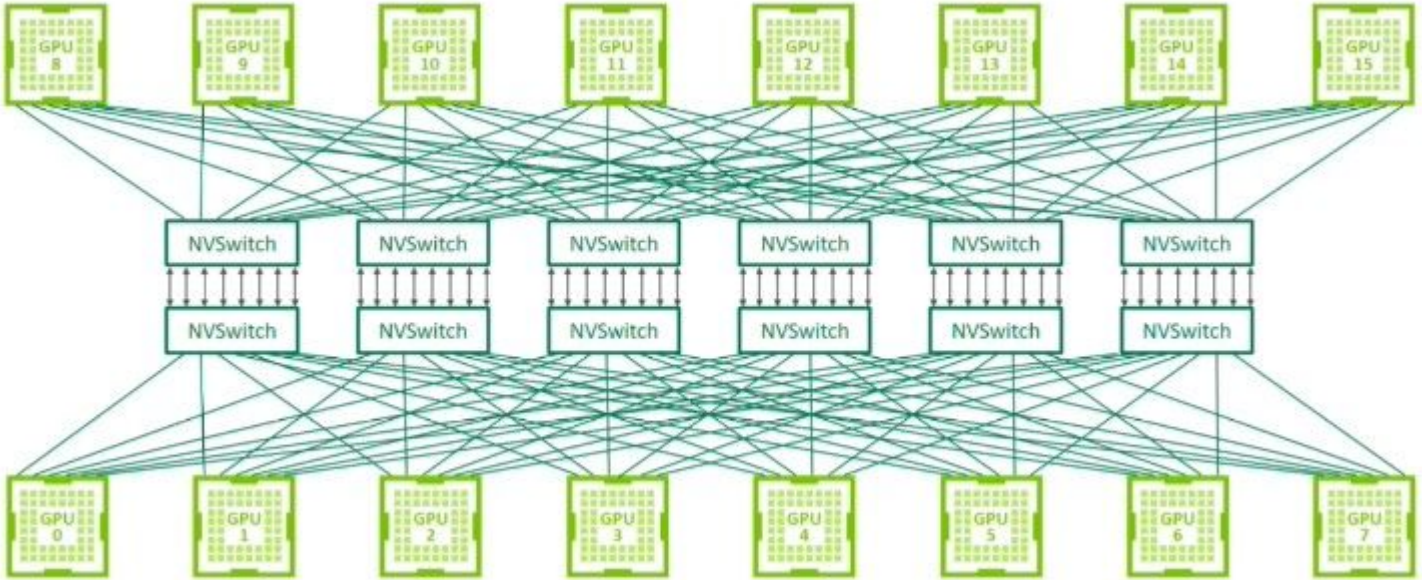




Y si tenés plata ...

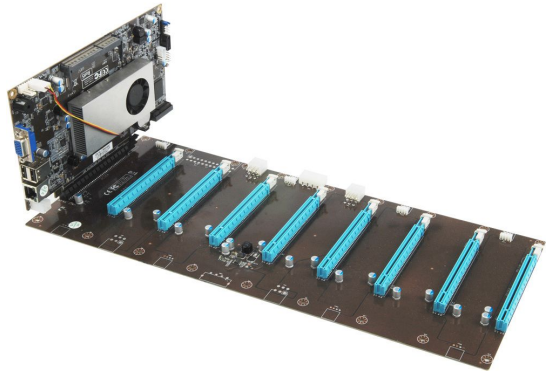


O aún más … 300 GB/s GPU-GPU, 2.4 TB/s total



Multi GPU

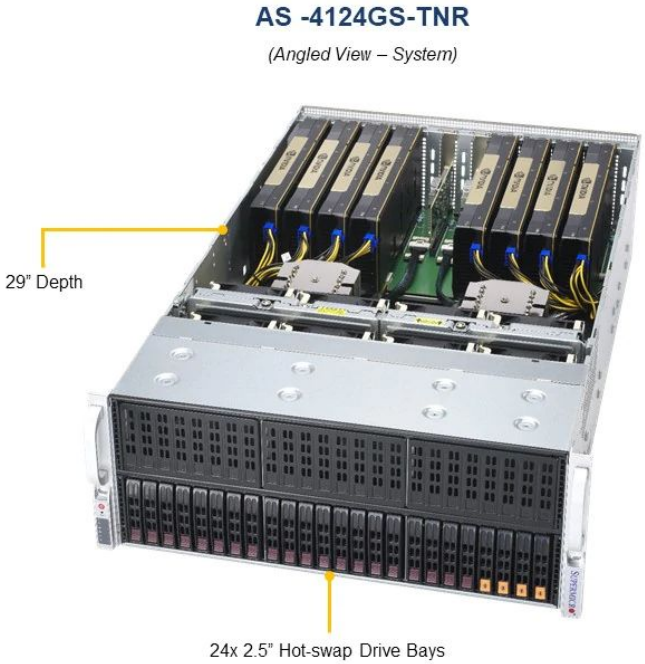
- 2 * (Xeon Gold 6248, 48 PCIe 3.0 lanes).
- 96 PCIe lanes
- Cada RTX 2080 Ti, 16x lanes.
- 96/16 = 6 < 8
- 🦆
- Nos reíamos de esto, pero "Está mal, pero no tan mal".



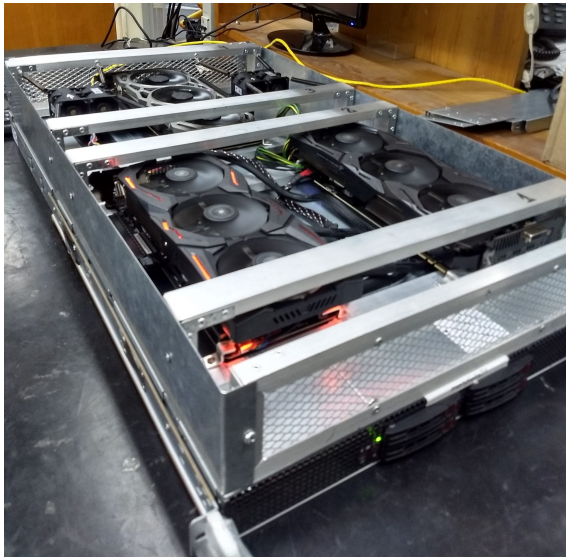
TensorFlow	Keras	PYTORCH
Caffe	Caffe2	theano

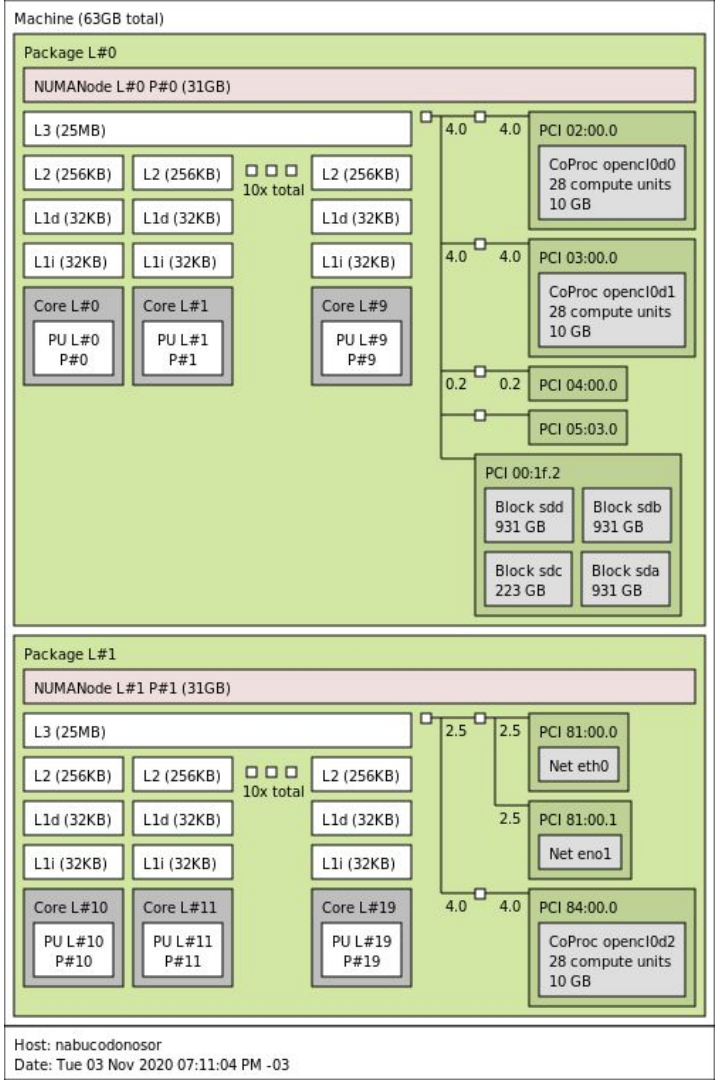
MultiGPU, otro

- 2 * (EPYC Rome, 128 **PCIe 4.0** lanes).
- 128 PCIe lanes libres.
- Cada RTX 2080 Ti, 16x lanes.
- 128/16 = 8 = 8 = 8
- 👋



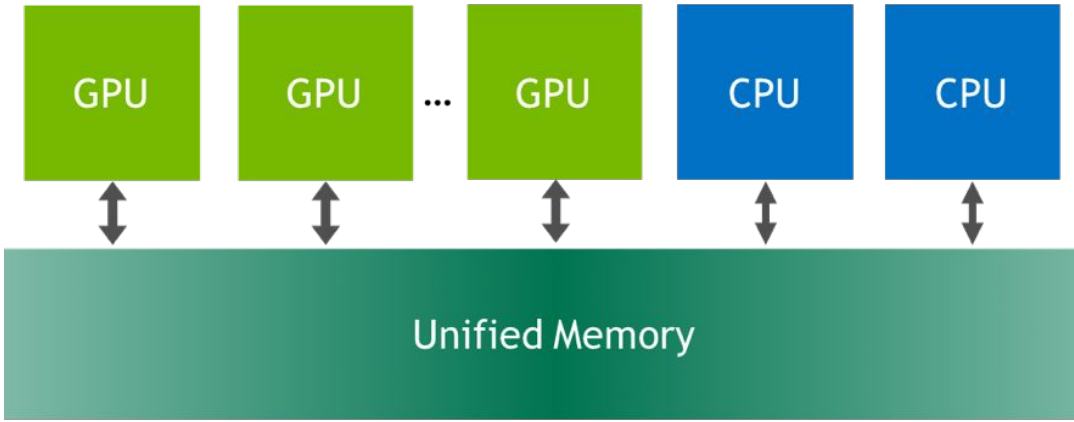
MultiGPU, Topología, Nabu



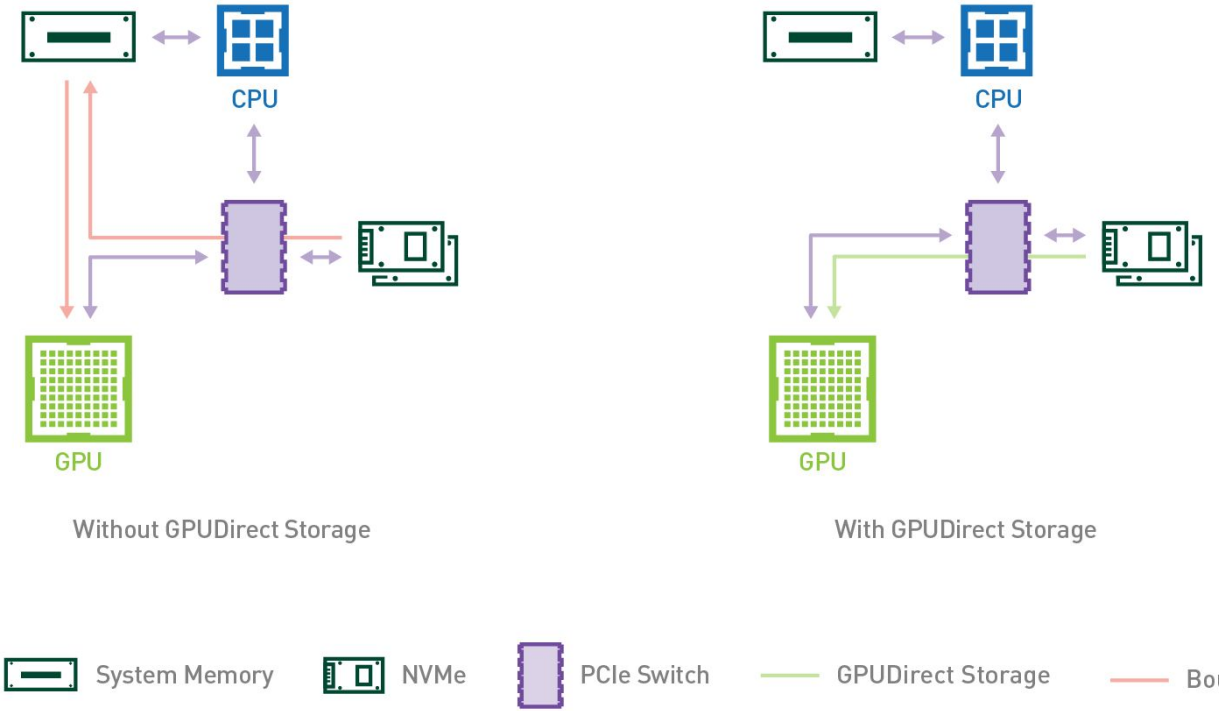


MultiGPU, HW page migration, unified memory

- Pedimos memoria y se mueve sola.
- Page out/in a CPU! ;) tenemos mucha memoria para GPU
  - Spoiler, da guano.
- Pascal, Turing, Volta, Ampere, cada vez más soporte del HW.



Comunicación directa por PCIe: SSD, IB. 1-copy



Filigranas eléctricas

- RTX 2080 Ti, **250W**
- RTX 3090, **350W**

16x PCIe cards, max 75W.

PCIe Power Connectors

- 75 W (6-pin)
- 150 W (8-pin)

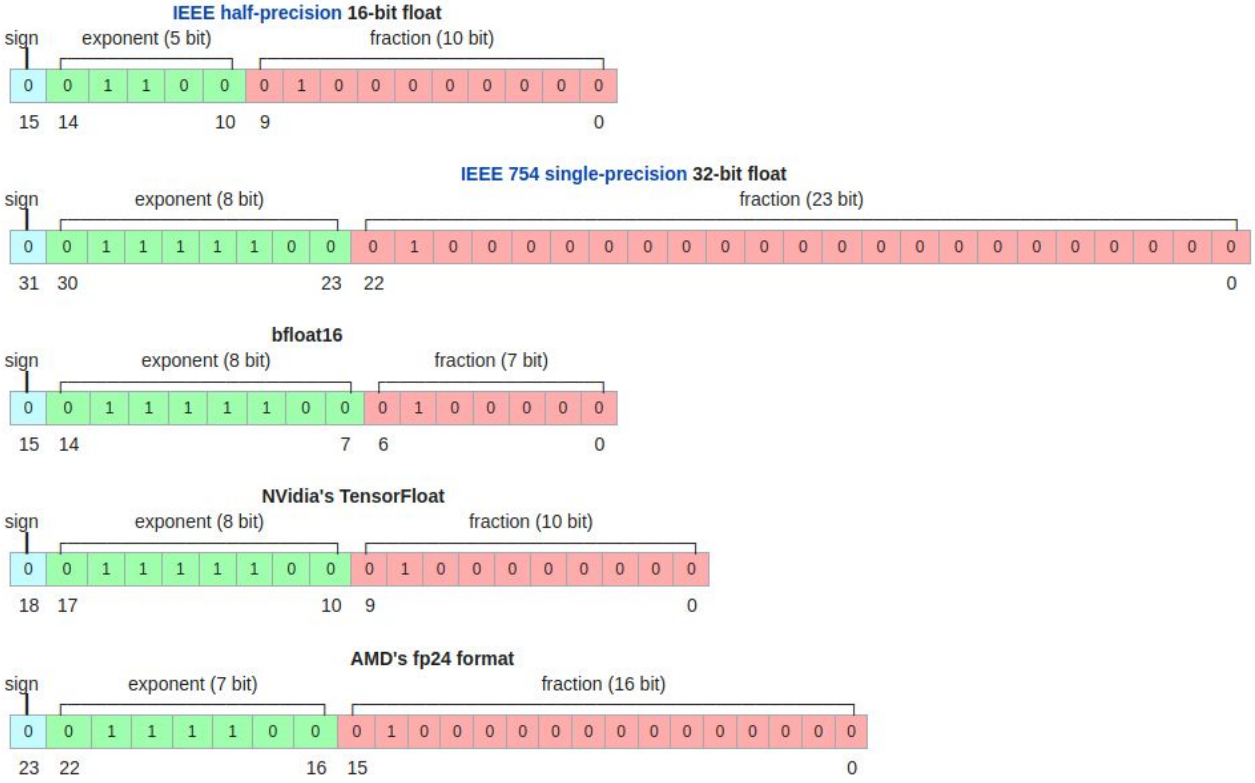
Máximo 75W+300W.

8-pin, 3 lineas 12V, 150W/12/3 = **4.16A**. Es un cable gruesito.

Representación de datos

- float64 no existe.
- float32, el ubícuo.
- float16, GPUs y CPUs modernas.
- bfloat16, GPUs y CPUs ultranuevas.
- bfloat19 (TF32 según NVIDIA), GPUs ultranuevas.

¡¡¡¡Duplican memoria y BW!!!!



Plétora de medidas según representación

	NVIDIA A100 para HGX
Peak FP64	9,7 TF
Peak FP64 Tensor Core	19,5 TF
Peak FP32	19,5 TF
Peak TF32 Tensor Core	156 TF | 312 TF*
Peak BFLOAT16 Tensor Core	312 TF | 624 TF*
Peak FP16 Tensor Core	312 TF | 624 TF*
Peak INT8 Tensor Core	624 TOPS | 1.248 TOPS*
Peak INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*

Ampere, la bestia

Feature	NVIDIA A100 40GB SXM4		NVIDIA A100 40GB PCI-Express	
GPU Chip	Ampere GA100			
TensorCore Performance*	19.5 TFLOPS	FP64	17.6 ~ 19.5 TFLOPS	FP64
	156 TFLOPS †	TF32	140 ~ 156 TFLOPS †	TF32
	312 TFLOPS †	FP16/BF16	281 ~ 312 TFLOPS †	FP16/BF16
	624 TOPS †	INT8	562 ~ 624 TOPS †	INT8
	1,248 TOPS †	INT4	1,123 ~ 1,248 TOPS †	INT4
Double Precision (FP64) Performance*	9.7 TFLOPS		8.7 ~ 9.7 TFLOPS	
Single Precision (FP32) Performance*	19.5 TFLOPS		17.6 ~ 19.5 TFLOPS	
Half Precision (FP16) Performance*	78 TFLOPS		70 ~ 78 TFLOPS	
Brain Floating Point (BF16) Performance*	39 TFLOPS		35 ~ 39 TFLOPS	
On-die HBM2 Memory	40GB			
Memory Bandwidth	1,555 GB/s			
L2 Cache	40MB			
Interconnect	NVLink 3.0 (12 bricks) + PCI-E 4.0		NVLink 3.0 (12 bricks) + PCI-E 4.0 *NVLink is limited to pairs of directly-linked cards*	
GPU-to-GPU transfer bandwidth (bidirectional)	600 GB/s			
Host-to-GPU transfer bandwidth (bidirectional)	64 GB/s			

¿Cuánto sale tener precisión doble?

- **RTX 2080 Ti**, 999 USD, 0.44 TFLOPS float64 1:32 ratio.
- **A100**, 9999USD, 9.7 TFLOPS float64, 1:2 ratio.

Double-precision FPUs in High-Performance Computing: an Embarrassment of Riches?

Jens Domke*,§, Kazuaki Matsumura†, Mohamed Wahib‡, Haoyu Zhang†, Keita Yashima†, Toshiki Tsuchikawa†, Yohei Tsuji†, Artur Podobas†,§, Satoshi Matsuoka§,†

*Global Scientific Information and Computing Center, Tokyo Institute of Technology
†Department of Mathematical and Computing Science, Tokyo Institute of Technology
‡AIST-TokyoTech Real World Big-Data Computation Open Innovation Laboratory, Tokyo, Japan
§RIKEN Center for Computational Science (R-CCS), RIKEN, Japan

***Abstract*—Among the (uncontended) common wisdom in High-Performance Computing (HPC) is the applications' need for large amount of double-precision support in hardware. Hardware manufacturers, the TOP500 list, and (rarely revisited) legacy software have without doubt followed and contributed to this view.**

In this paper, we challenge that wisdom, and we do so by exhaustively comparing a large number of HPC proxy applications on two processors: Intel's Knights Landing (KNL) and Knights Mill (KNM). Although similar, the KNL and KNM architecturally deviate at one important point: the silicon area devoted to double-precision arithmetics. This fortunate discrepancy allows us to empirically quantify the performance impact in reducing the amount of hardware double-precision arithmetic.

Our analysis shows that this common wisdom might not always be right. We find that the investigated HPC proxy applications do allow for a (significant) reduction in double-precision with little-to-no performance implications. With the advent of a failing of Moore's law, our results partially reinforce the view taken by modern industry (e.g., upcoming Fujitsu ARM64FX) to integrate hybrid-precision hardware units.

I. Introduction

It is becoming increasingly clear that the road forward in High-Performance Computing (HPC) is one full of obstacles. With the ending of Dennard's scaling [1] and the ending of Moore's law [2], there is today an ever-increasing need to oversee how we allocate the silicon to various functional units in modern many-core processors. Amongst those decisions is how we distributed the hardware support for various levels of compute-precision.

Historically, most of the compute silicon has been allocated to double-precision (DP; 64-bit) compute. Nowadays – in processors such as the forthcoming AA64FX [3] and NVIDIA Volta [4] – the trend, mostly driven by market/AI demands, is to replace some of the double-precision units with lower-precision units. Lower-precision units occupy less area (up to ≈3x going from double- to single-precision Fused-Multiply-Accumulate [5]), leading to more on-chip resources (more instruction-level parallelism), potentially lowered energy consumption, and a definitive decrease in external memory bandwidth pressure (i.e., more values per unit bandwidth). The gains – up to four times over their DP variants with little loss in accuracy [6] – are attractive and clear, but what is the impact on performance (if any) on existing HPC applications? What performance impact can HPC users expect when migrating their code to future processors with a different distribution in floating-point precision support? Finally, how can we empirically quantify this impact on performance using existing processors in an apples-to-apples comparison on real-life use cases without relying on tedious, slow, and potentially inaccurate simulators?

The Intel Xeon Phi was supposed to be the high-end for many-core processor technology for nearly a decade (Knights Ferry was announced in 2010), and has changed drastically since its first released. The latest (and also last) two revisions – the Knights Landing and Knights Mill – are of particular importance since they arguable reflect two different ways of thinking. Knights Landing has relatively large support for double-precision (64-bit) computations, and follows a more traditional school of thought. While Knights Mill follows a different direction, which is the replacement of double-precision compute units with lower-precision (single-precision, half-precision, and integer) compute capabilities.

In the present paper, we quantify and analyze the performance and compute bottlenecks of Intel's Knights Landing [7] and Knights Mill [8] architectures – two processors with identical micro-architecture where the main difference is in the relative allocation of double-precision units. We stress both processors with numerous realistic benchmarks from both the Exascale Computing Project (ECP) proxy applications [9] and RIKEN R-CCS Fiber Miniapp Suite [10] – benchmarks used in HPC system acquisition. Through an extensive (and robust) performance measurement process (which we also open-source), we empirically show the architecture's relative weaknesses. In short, the contributions of the present paper are:

1) An empirical performance evaluation of the Knights Landing and Mill family of processors – both proxies for previous and future architectural trends – with respect to benchmarks derived from realistic HPC workloads,
2) An in-depth analysis of results, including identification of bottlenecks for the different application/architecture combinations, and
3) An open-source compilation of our evaluation methodology, including our collected raw data.

Atención Primaria de la Salud de ML corriendo

- `poner la mano atrás de la máquina.`
- `htop`
- `perf top`
- `nvidia-smi`

1. Si la máquina está fría, tu código est
2. htop con barritas verdes al 100%; barrita roja malo.
3. Que no haya cosas raras en CPU, ej 200 hilos.
4. Que perf top muestre cosas razonables arriba.
5. Que se esté usando la GPU (nvidia-smi), con un duty-cycle alto y alta potencia.

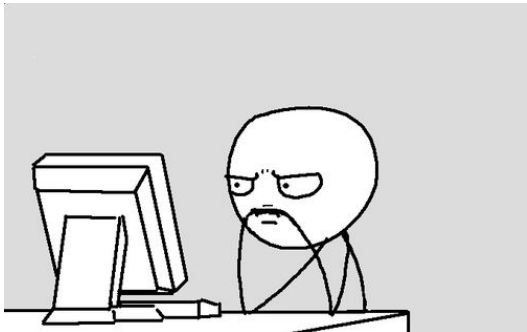
Medicina de precisión

- `nvprof`
- `nvvp`

Herramientas que muestran al detalle lo que pasa en todo el stack del software de la GPU.

- Migración excesiva de páginas CPU-GPU.
- Sub utilización del BW de memoria y de los TFLOPS de cálculo.
- No-superposición de computación y comunicación.

Muchas veces mirando el código y teniendo en mente el modelo de computación se mejora mmmucho el desempeño.



Multiprocesamiento

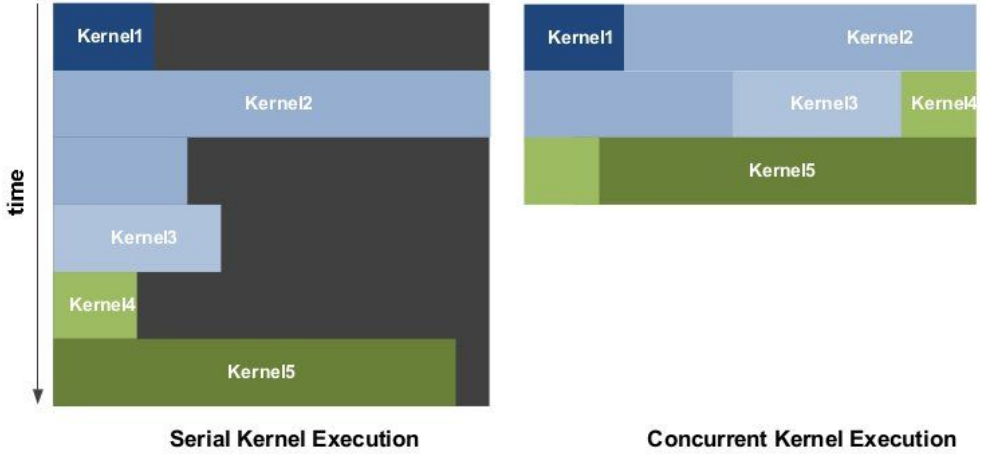
- Muy difícil vencer una TU102.
- Pero si le pegamos entre varios de vez en cuando cae.

¡Sobreventa de pasajes!

Si vemos

- No 100% GPU, no 100% W

Mandamos varios procesos al mismo chip.



Desempeño/USD, eficiencia económica

En GROMACs está superestudiado.

¿Algún trabajito asi para TF o esas cosas?

https://timdettmers.com/ es lo más parecido.

More Bang for Your Buck:

Improved use of GPU Nodes for GROMACS 2018

Carsten Kutzner,*,† Szilárd Páll,‡ Martin Fechner,† Ansgar Esztermann,† Bert L. de Groot,† and Helmut Grubmüller†

†*Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany*

‡*Center for High Performance Computing, KTH Royal Institute of Technology, 10044 Stockholm, Sweden*

E-mail: ckutzne@gwdg.de

Abstract

We identify hardware that is optimal to produce molecular dynamics trajectories on Linux compute clusters with the GROMACS 2018 simulation package. Therefore, we benchmark the GROMACS performance on a diverse set of compute nodes and relate it to the costs of the nodes, which may include their lifetime costs for energy and cooling. In agreement with our earlier investigation using GROMACS 4.6 on hardware of 2014, the performance to price ratio of consumer GPU nodes is considerably higher than that of CPU nodes. However, with GROMACS 2018, the optimal CPU to GPU processing power balance has shifted even more towards the GPU. Hence, nodes optimized for GROMACS 2018 and later versions enable a significantly higher performance to price ratio than nodes optimized for older GROMACS versions. Moreover, the shift towards GPU processing allows to cheaply upgrade old

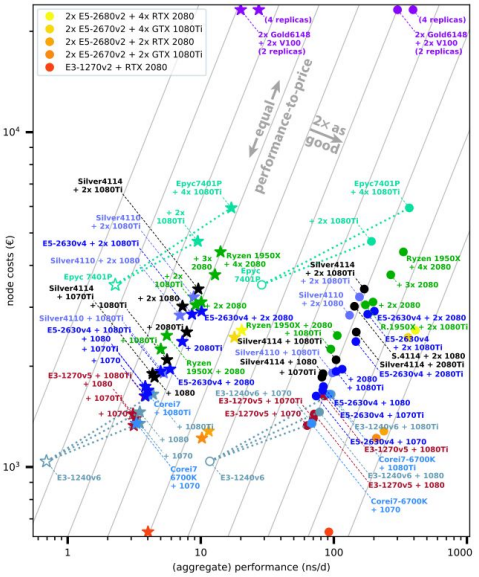
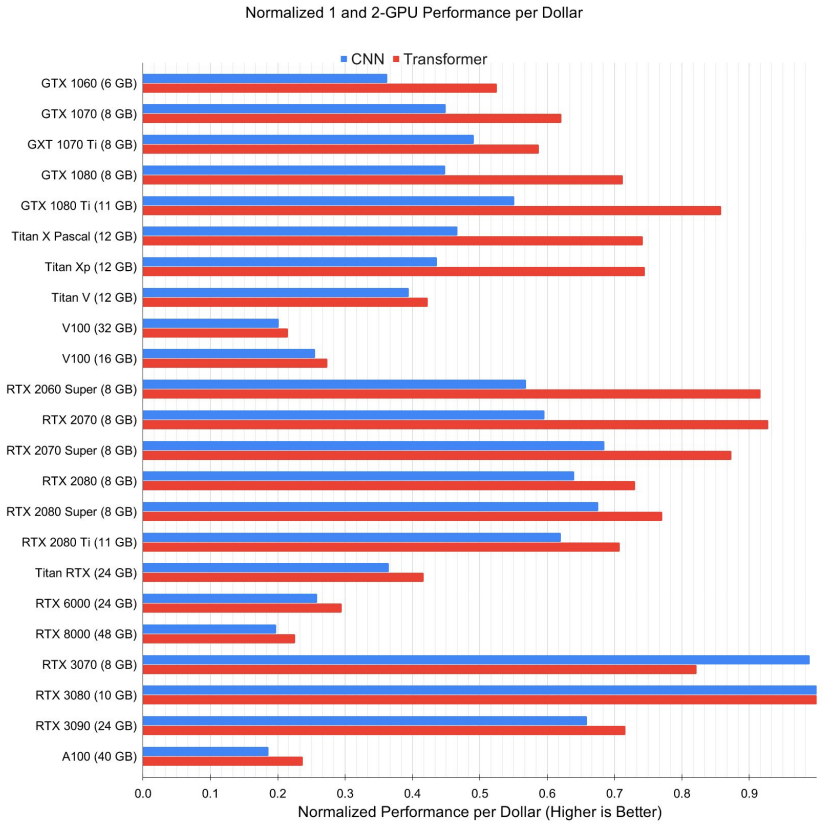
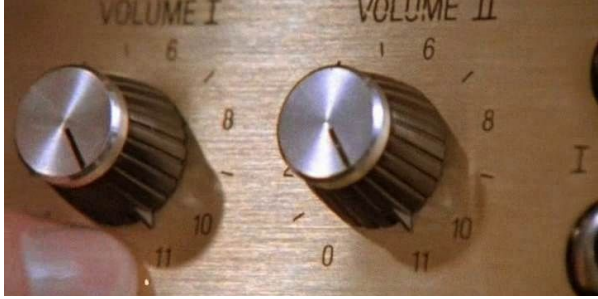


Fig. 9: (Aggregate) simulation performance in relation to net node costs. MEM (circles) and RIB (stars) symbols are colored depending on CPU type. Symbols with white fill denote nodes without GPU acceleration; dotted lines connect GPU nodes with their CPU counterparts. Grey: isolines of equal P/P ratio like in Fig. 1 with superior configurations to the lower right. Old nodes with upgraded GPUs from Table 4 are shown in yellow-orange colors (legend).

Tienda La Rosita, donde cada $ amerita

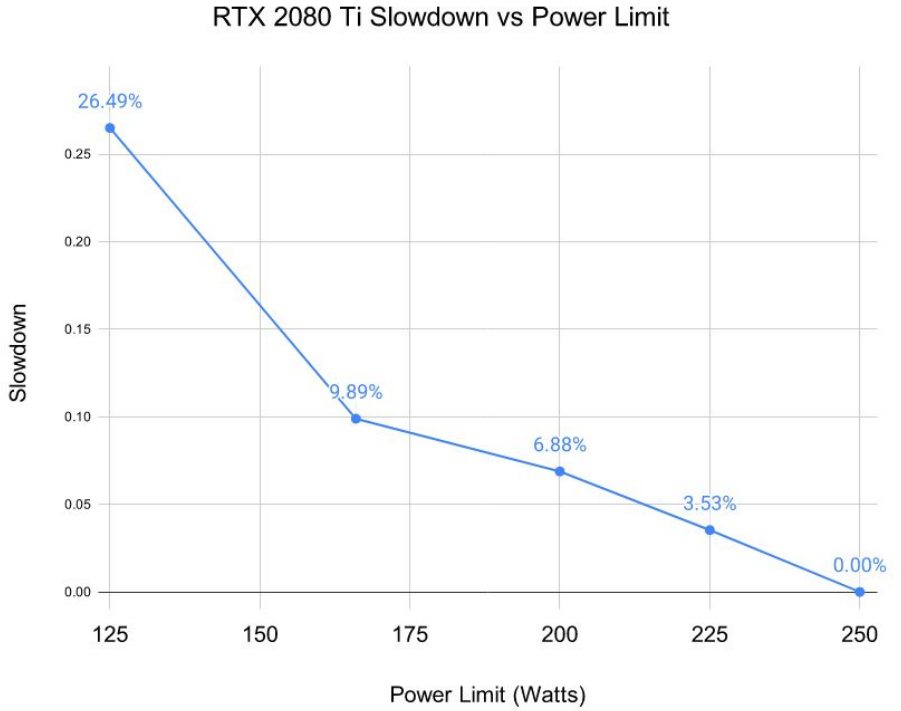


Powercapping



- Limitar la potencia consumida.
- Usualmente no lineal con la performance => mejora performance/W, eficiencia energética.

- Si dudás de tu fuente.
- Si precisás bajar la temperatura.
- Si te hartaste del ruido.



Workloads similares