Supercomputing, the heart of Deep Learning

Surely, at this point, some readers have already posed the question: why has a researcher in supercomputing such as me, started to investigate Deep Learning?

In fact, many years ago I started to be interested in how supercomputing could contribute to improving Machine Learning methods; Then, in 2006, I started co-directing PhD theses with a great friend, and professor at the Computer Science department of the UPC, Ricard Gavaldà[8] , an expert in Machine Learning and Data Mining.

But it was not until September 2013, when I already had a relatively solid base of knowledge about Machine Learning, that I started to focus my interest on Deep Learning. Thanks to the researcher from our Computer Architecture Department at UPC Jordi Nin, I discovered the article *Building High-level Features Using Large Scale Unsupervised Learning*[9], written by Google researchers. In this article presented at the previous International Conference in Machine Learning (ICML'12), the authors explained how they trained a Deep Learning model in a cluster of 1,000 machines with 16,000 cores. I was very happy to see how supercomputing made it possible to accelerate this type of applications, as I wrote in my blog[10] a few months later, justifying the reasons that led the group to add this research focus to our research roadmap.

[8] Ricard Gavaldà web page [online]. Available at: http://www.lsi.upc.edu/~gavalda/

[9] Quoc Le and Marc'Aurelio Ranzato and Rajat Monga and Matthieu Devin and Kai Chen and Greg Corrado and Jeff Dean and Andrew Ng, *Building High-level Features Using Large Scale Unsupervised Learning* .International Conference in Machine Learning, ICML 2012 [online]. Available at: https://arxiv.org/abs/1112.6209 [Accessed: 12/02/2018]

[10] Blog de Jordi Torres. "Barcelona Supercomputing Center starts to work on Deep Learning" June 2014). [online]. Available at: http://jorditorres.org/barcelona-supercomputing-center-starts-to-works-on-deep-learning/

Thanks to Moore's Law[11], in 2012, when these Google researchers wrote this article, we had supercomputers that allowed us to solve problems that would have been intractable a few years before due to the computing capacity. For example, the computer that I had access to in 1982, where I executed my first program with punch-cards, it was a Fujitsu that made it possible to execute a little more than one million operations per second. 30 years later, in 2012, the Marenostrum supercomputer that we had at the time at the Barcelona Supercomputing Center-National Supercomputing Center (BSC), was only 1,000,000,000 times faster than the computer on which I started.

With the upgrade of that year, the MareNostrum supercomputer offered a theoretical maximum performance peak of 1.1 Petaflops (1,100,000,000,000 floating point operations per second[12]). It achieved it with 3,056 servers with

[11] Moore's Lay. Wikipedia. [online]. Available at: https://en.wikipedia.org/wiki/Moore%27s_law [Accessed: 12/03/2018]
[12] FLOPS. Wikipedia. [online]. Available at: https://en.wikipedia.org/wiki/FLOPS [Accessed: 12/03/2018]

a total of 48,896 cores and 115,000 Gigabytes of total main memory housed in 36 racks. At that time the Marenostrum supercomputer was considered to be one of the fastest in the world. It was placed in the thirty-sixth position, in the TOP500 list[13], which is updated every half year and ranks the 500 most powerful supercomputers in the world. Attached you can find a photography where you can see the Marenostrum computer racks that were housed in the Torres Girona chapel of the UPC campus in Barcelona.[14].

The first GPU in the ImageNet competition

During that period was when I began to become aware of the applicability of supercomputing to this new area of research. When I started looking for research articles on the subject, I discovered the existence of the Imagenet competition and the results of the team of the University of Toronto in the competition in 2012[15]. The ImageNet competition (Large Scale Visual Recognition Challenge[16]) had been held since 2010, and by that time it had become a benchmark in the computer vision community for the recognition of objects on a large scale. In 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hilton used for the first time hardware accelerators GPU (graphical processing units) [17], which was already used at that time in supercomputing centers like ours in Barcelona to increase the speed of execution of applications that require the performance of many calculations.

---

[13] Top 500 List – November 2012. [online] Available at: https://www.top500.org/list/2012/11/?_ga=2.211333845.1311987907.1527961584-375587796.1527961584 [Accessed: 12/03/2018]

[14] Marenostrum 3. Barcelona Supercomputing Center. [online]. Available at: https://www.bsc.es/marenostrum/marenostrum/mn3 [Accessed: 12/03/2018]

[15] Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada [online]. Available at:http://www.cs.toronto.edu/~kriz/imagenet_classification_with_deep_convolutional.pdf

[16] Russakovsky, O., Deng, J., Su, H. et al. Int J Comput Vis (2015) 115: 211. https://doi.org/10.1007/s11263-015-0816-y https://arxiv.org/abs/1409.0575

[17] Wikipedia. Graphics processing unit. [online] Available at: https://en.wikipedia.org/wiki/Graphics_processing_unit [Accessed: 12/02/2018]

For example, at that time BSC already had a supercomputer called MinoTauro, with 128 Bull505 nodes, equipped with 2 Intel processors and 2 Tesla M2090 GPUs from NVIDIA each one. With a peak performance of 186 Teraflops, launched in September 2011 (out of curiosity, at that time it was considered the most energy efficient supercomputer in Europe according to the Green500 list[18]).

Until 2012, the increase in computing capacity that we got each year from computers was as a result of the improvement of the CPU. However, since then the increase in computing capacity for Deep Learning has not only been credited to them, but also to the new massively parallel systems based on GPU accelerators, which are many times more efficient than traditional CPUs.

GPUs were originally developed to accelerate the 3D game that requires the repeated use of mathematical processes that include different matrix calculations. Initially, companies such as NVIDIA and AMD developed these fast and massively parallel chips for graphics cards dedicated to video games. However, it soon became clear that the use of GPUs for 3D games was also very suitable for accelerating calculations on numerical matrices; therefore, this hardware actually benefited the scientific community, and in 2007 NVIDIA launched the CUDA[19] programming language to program its GPUs. As a result, supercomputing research centers such as the BSC began using GPU clusters to accelerate numerical applications.

But as we will see later in this book, artificial neural networks basically perform matrix operations that are also highly parallelizable. And this is what Alex Krizhevsky's team did in 2012: he trained his Deep Learning algorithm "AlexNet" with GPU. Since then, some research groups have started using GPUs for this competition, and nowadays all the groups that do research in

[18] See https://www.top500.org/green500/
[19] Wikipedia. CUDA. [online]. Available at: https://en.wikipedia.org/wiki/CUDA

Deep Learning research field are using this hardware or equivalent alternatives that have appeared recently.

An exponential growth of computing capacity

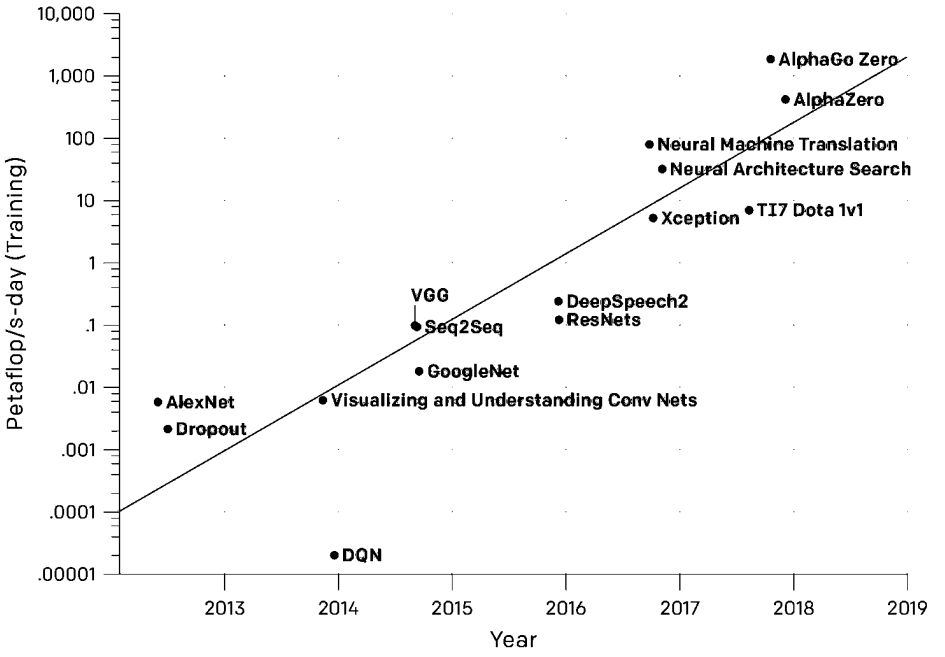
I have already said that Krizhevsky's team's milestone was an important turning point in the field of Deep Learning, and since then there have been spectacular results, one after another, with an exponential growth of increasingly surprising results.

But I believe that research in this field has been guided largely by experimental findings rather than by theory, in the sense that these spectacular advances in the area since 2012 have only been possible thanks to the fact that the computation that was required to be able to carry them out was available; In this way, researchers in this field have been able to test and extend old ideas, while they have advanced with new ones that required a lot of computing resources.

OpenAI[20] has recently published a study on its blog[21] that corroborates precisely this vision that I am defending. Specifically, they present an analysis in which it is confirmed that, since 2012, the amount of computation available to generate models of artificial intelligence has increased exponentially while claiming that improvements in computing capacity have been a key component of the progress of Artificial Intelligence.

In this same article they present an impressive graph[22] to synthesize the results of their analysis:

[20] See https://openai.com
[21] See https://blog.openai.com/ai-and-compute/
[22] See https://blog.openai.com/content/images/2018/05/compute_diagram-log@2x-3.png



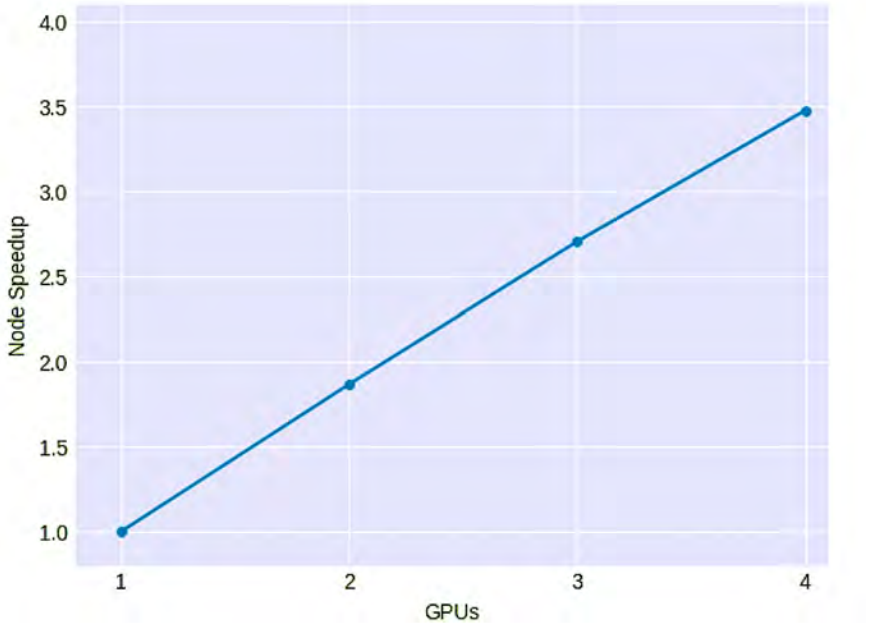
The graph shows the total amount of calculations, in Petaflop per day, that have been used to train neural networks that have their own name and are referents in the Deep Learning community. Remember that a "petaflop / s-day", the vertical axis of the graph that is in logarithmic scale, is equivalent to perform 1,000,000,000,000,000 neural network operations per second during a day (s-day), or a total of approximately 100,000,000,000,000,000 operations, regardless of numerical precision.

Accelerating Deep Learning with parallel systems

The tasks of training Deep Learning networks requires a large amount of computation and, often, they also need the same type of matrix operations as the numerical calculation intensive applications, which makes them similar to traditional supercomputing applications. Therefore, Deep Learning applications work very well in computer systems that use accelerators such as GPU or field-programmable gate arrays (FPGA), which have been used in

the Supercomputing field for more than a decade within the walls of the supercomputing research centers. These hardware devices focus on computational performance by specializing their architecture in using high data parallelism in supercomputing workloads. And precisely these techniques can also be used to accelerate the learning algorithms of Deep Learning.

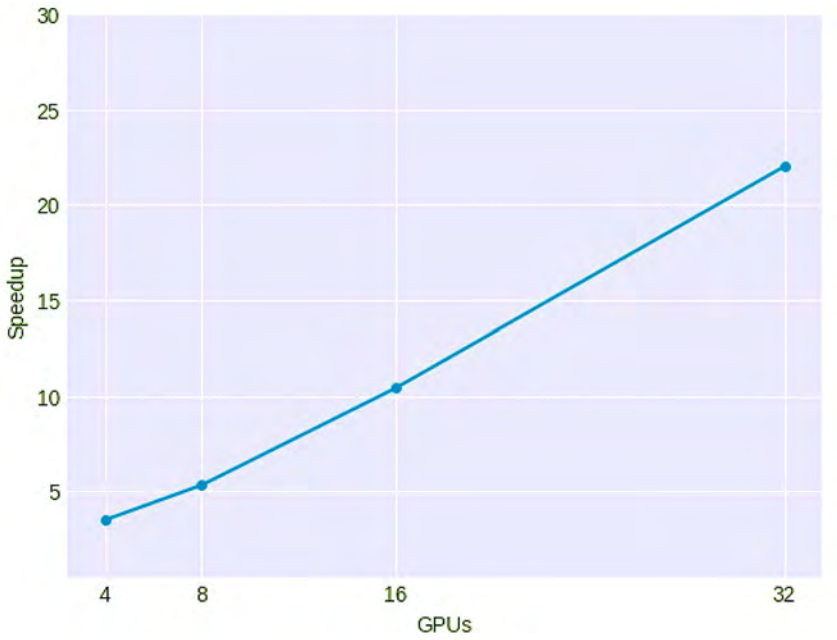
Therefore, from 2012 until 2014, Deep Learning researchers started using systems with GPUs. The advantage, in addition, is that these learning algorithms escalated perfectly when they could put more than one GPU in a node. The following graph, extracted from one of our research articles, shows how increasing the number of GPUs can accelerate the learning process[23]:



[23] Campos, V., F. Sastre, M. Yagues, J. Torres, and X. Giro-I-Nieto. Scaling a Convolutional Neural Network for Classification of Adjective Noun Pairs with TensorFlow on GPU Clusters. 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)

Accelerating Deep Learning with distributed systems

The large computational capacity available at that time allowed the Deep Learning community to move forward and be able to design increasingly complex neural networks, requiring more computing capacity than a server with multiple GPUs could offer. Therefore starting in 2014 to accelerate even more the calculation required, computing began to be distributed among multiple machines with several GPUs connected by a network. This solution had been adopted again previously, and very well known, in the community of researchers in supercomputing, specifically in the interconnection of machines through optical networks with low latency, which made it possible to do the distribution in a very efficient manner. The following graph shows how the same algorithm can be accelerated with several machines that each have 4 GPUs[24]:



[24] Campos, V., F. Sastre, M. Yagues, M. Bellver, X. Giro-I-Nieto, and J. Torres. Distributed training strategies for a computer vision deep learning algorithm on a distributed GPU cluster. Procedia Computer Science. Elsevier. Volume 108, Pag. 315-324. https://doi.org/10.1016/j.procs.2017.05.074

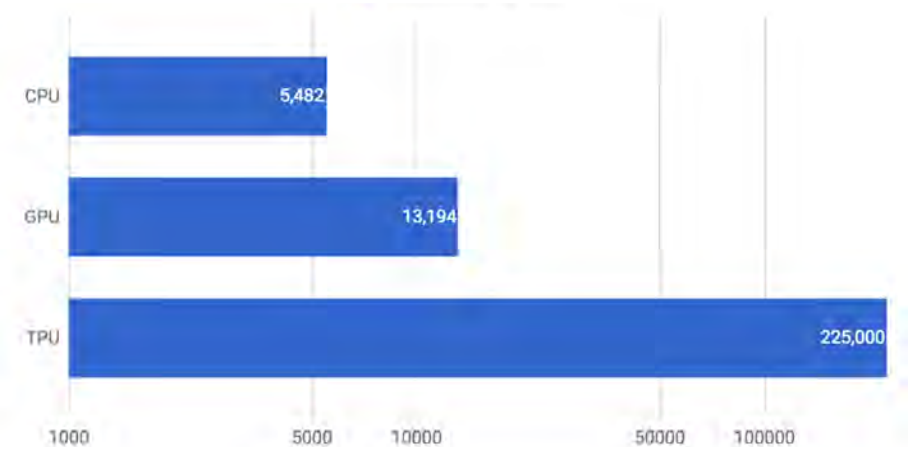
Also libraries of the standard of communication like Message Passing Interface (MPI) [25], used in the scientific community of supercomputing for decades, are spreading in distributed Deep Learning, requiring the knowledge of experts in supercomputing to accelerate these algorithms.

Will specialized hardware for deep learning be a game changer?

As of 2016, in addition to all the previous innovations in supercomputing, processing chips began to appear that were specially designed for Deep Learning algorithms. For example, in 2016 Google announced that it had built a dedicated processor called the Tensor Processing Unit (TPU) [26]. Since then Google has already developed 3 versions of TPU, the last one presented in its IO[27] conference, where they claimed that it is 8 times more powerful than the previous version. In addition, now not only the architecture is specific to train neural networks, but also for the inference stage (use of the previously trained model).

In the following graph obtained from the Google Cloud blog[28], we can see a comparison of predictions per second obtained (on a logarithmic scale) for the three different types of architecture mentioned above.

[25] MPI Wikipedia. https://en.wikipedia.org/wiki/Message_Passing_Interface
[26] Wikipedia. Tensor Processing Unit. [online]. Available at: https://en.wikipedia.org/wiki/Tensor_processing_unit [Accessed: 20/04/2018]
[27] Google IO 2018 conference (8-10 May 2018). Videos. [online]. Available at: https://www.youtube.com/playlist?list=PLOU2XLYxmsIInFRc3M44HUTQc3b_YJ4-Y [Accessed: 16/05/2018]
[28] See https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu



The acceleration of Deep Learning with specialized hardware has only just begun, both for the training stage and the inference stage, if we take into account that numerous companies are appearing that are designing and starting to produce specific chips for Artificial Intelligence[29].

Specialized hardware is sure going to be a big factor in Deep Learning race. We will see great progress soon, I am sure. However, even more interesting will be to see what role Deep Learning plays in changing hardware in the near future.

Tapping the Next Generation of Supercomputers

And now we are at the point of convergence of Artificial Intelligence technologies and supercomputing technologies. The result will soon be part of the portfolio that companies providing computer systems (and Cloud services) will offer to the industrial and business world.

[29] Big Bets on A.I. Open a New Frontier for Chip Start-Ups, Too. The New York Times. January 14, 2018 [online]. Available at: https://www.nytimes.com/2018/01/14/technology/artificial-intelligence-chip-start-ups.html [Accessed: 20/04/2018]

An example of what will be on the market in a short while, is a part of the current Supercomputer Marenostrum of the Barcelona Supercomputing Center (BSC). MareNostrum is the generic name used by the BSC to refer to the different updates of its most emblematic and the most powerful supercomputer in Spain, and until today four versions have been installed since 2004[30]. At present, the Marenustrum is the most heterogeneous supercomputer in the world, with all kinds of experimental hardware available on the market, since its purpose is to serve as an experimental platform to design future supercomputers.

This is structured in that the calculation capacity of the current MareNostrum 4 is divided into two parts of totally differentiated hardware: a block of general purpose and a block of emerging technologies. The emerging technologies block is made up of clusters of three different technologies that will be incorporated and updated as they become available. These are technologies that are currently being developed in the United States and Japan to accelerate the arrival of the new generation of pre-exascale supercomputers.

One of these technologies is based on an IBM system designed especially for Deep Learning and Artificial Intelligence applications[31]; IBM has created all the software stack necessary for it. At the time of writing this book, the hardware is available and the PowerAI[32] software package will soon be installed, which will convert this supercomputing hardware into a machine specially designed for Artificial Intelligence. Through this software, the main frameworks of Deep Learning such as TensorFlow (and Keras, included in

[30] Marenostrum. Barcelona Supercomputing Center [online] Available at : https://www.bsc.es/marenostrum/marenostrum [Accessed: 20/05/2018]
[31] IBM Power System AC922 Introduction and Technocal Overview- IBM RedBooks by Andexandre Bicas Caldeira. March 2018. [online]. Available at http://www.redbooks.ibm.com/redpapers/pdfs/redp5472.pdf [Accessed: 20/05/2018]
[32] IBM PowerAI: Deep Learning Unleashed on IBM Power Systems Servers. IBM RedBooks. March 2018. [online]. Available at https://www.dropbox.com/s/fd3oiuttqdeilut/IBMpowerAI.pdf?dl=0 [Accessed: 20/05/2018]

the Tensorflow package), Caffe, Chainer, Torch and Theano will be available to researchers of Artificial Intelligence.

In terms of hardware, this part of the Marenostrum consists of a 54 node cluster based on IBM POWER9 and NVIDIA V100 with Linux operating system and interconnected by an infiniband network at 100 Gigabits per second. Each node is equipped with 2 IBM POWER9 processors that have 20 physical cores each and 512GB of memory. Each of these POWER9 processors are connected to two NVIDIA V100 (Volta) GPUs with 16GB of memory, a total of 4 GPUs per node. See the following picture:

The new NVIDIA V100 GPUs are the most advanced GPUs yet[33] to accelerate Artificial Intelligence applications equivalent to 100 CPUs according to NVIDIA[34]. This is achieved by matching their CUDA cores with 640 core tensor, which did not have the previous family of GPU Pascal from NVIDIA. The core tensor is specifically designed to multiply two matrices of 4×4 elements in floating point format and also allows the accumulation of a third matrix, thus being able to execute very quickly the basic operations of neural networks both in the inference phase and training phase.

In addition, this new version of GPU updates the bus with NVLINK 2.0[35] system that allows a high bandwidth with six links that can transfer 50 Gigabytes per second. Although traditionally the NVLINK bus was originally designed to connect the GPUs, this version also makes it possible to connect GPU and CPU. Another important element is access to memory, which has improved compared to previous versions and allows bandwidths of up to 900 GigaBytes per second. Something awesome!.

I describe all this detail so that it does not surprise you that with only 3 racks of the current Marenostrum (the ones seen in the following picture) there is 1.5 Petaflops of theoretical maximum performance, much more than the 1.1 Petaflops that in 2012, the Marenostrum 3 had, in 36 racks (shown in the previous photo in page 30).

[33] Just at the time of finalizing the contents of this book NVIDIA has presented a new version of the V100 with 32 GB of memory. [online]. Available at https://www.top500.org/news/nvidia-refreshes-v100-gpus-upgrades-dgx-lineup/

[34] Tesla V100 NVIDIA. [online]. Available at http://www.nvidia.com/v100 [Accessed: 20/03/2018]

[35] CTE-POWER User's Guide. Barcelona Supercomputing Center 2018 [online]. Available at https://www.bsc.es/user-support/power.php [Accessed: 20/05/2018]

In summary, it is not my intention to give you a computer architecture class, but I would like to explain and make you aware, with real examples close to us, that the computing capacity is exponentially evolved. And it has allowed us , as I said before, to try new ideas or to extend the old ones, since many of the advances in the Deep Learning area since 2012 have been guided by the experimental findings with hardware equivalent to the one that was being used in the world of supercomputing.

However, without any doubt, other factors have contributed to trigger the resurgence of Artificial Intelligence, and I must be fair and say that it is not only due to Supercomputing. If I asked you one of the other key factors, the Big Data phenomenon comes to mind as another of the facilitators of the resurgence of this new stage of Artificial Intelligence. But there are others that perhaps the reader is not so familiar with . In the next chapter we will present them.