



## Unpacking the Complexity in AI Training, Energy Consumption, and Emissions

It is commonly known that AI training and testing requires a large amount of energy, and this has raised valid questions around AI's environmental impact. Training [large neural networks can result](#) in hundreds of metric tons of CO<sub>2</sub> emissions. The popular narrative is often that AI is at odds with environmental sustainability. This isn't the whole story.

The energy consumption and emissions associated with AI training and testing is a more complex calculation than it may at first appear. To bring more clarity to this complicated area, it is helpful to compare AI training energy consumption relative to the underlying microprocessors, namely, CPUs and GPUs. Importantly, not all servers are equal when it comes to AI training, and the time to compute variable has a significant impact on energy usage.

As a recap, a central processing unit (CPU) processes binary calculations, and while CPUs use numerous cores that can handle multiple tasks simultaneously, their function is one of serial processing, completing one task after the next. A graphics processing unit (GPU) contains a far larger number of cores suited for computations performed in parallel, which is an essential ingredient in machine learning and high performance computing.

How do these microprocessors compare when it comes to AI training energy requirements? We can calculate the energy consumption for AI training using a standard enterprise CPU server and a GPU server.

The calculations are made possible by assuming a server runs at thermal design power (TDP) during the training run. TDP is the maximum power a server can draw, and is very unlikely to be true for any sustained operation, but provides an upper bound for the amount of energy consumed.





## Natural Language Processing (NLP) Model Training

Type	Hardware	# Nodes	# Chips per node	NLP Training (Minutes)	TDP (Watts)	Energy Consumption (kWh)
GPU	DGX H100	1	8	6.38	10,200	1.1
CPU	Sapphire Rapids (Intel -8480)	16	2	47.27	1,200	15.1

The above comparison identifies the improvement in training time using similar BERT models and datasets. The GPU server resulted in approximately 14x lower energy consumption compared to a CPU server configuration.

**Table 1:** For the calculation, references and comparisons are from [MLPerf](#). The sample training set for comparison was a BERT model with data from Wikipedia

## Image Classification (ResNet)


Type	Hardware	# Nodes	# Chips per node	Resnet Training (Minutes)	TDP (Watts)	Energy Consumption (kWh)
GPU	DGX H100	1	8	14.75	10,200	2.5
CPU	Sapphire Rapids (Intel -8480)	16	2	89.02	1,200*	28.5

When we account for the disparity in training time, we find that the GPU server consumes significantly less energy than CPUs, despite the significant TDP (Thermal Design Power). Performing a similar comparison as performed by MLPerf, we identify the overall improvement in energy consumption using GPU is approximately 11x as compared to CPU server. GPUs provide a significantly lower energy consumption, lower training time and reduced carbon footprint.

**Table 2:** For the calculation, references and comparisons are from [MLPerf](#). The sample training set for comparison was a ResNet model with data from ImageNet

\* [https://www.supermicro.com/datasheet/datasheet\\_X13\\_Hyper.pdf](https://www.supermicro.com/datasheet/datasheet_X13_Hyper.pdf)

The point of this arithmetic is not to promote a GPU server or a CPU server for AI training. Rather, it adds nuance to the consideration of AI energy consumption and emissions, and it allows us to think more critically about how we use and optimize the hardware and software that makes all this AI possible.



The [Green500](#) (compiled by the TOP500 project) lists the most efficient supercomputers. [An analysis by NVIDIA](#) found that accelerated computing is used in all of the top 30 supercomputers. Accelerated computing in this case refers to specialized hardware (which includes GPUs) and optimized software that uses parallel processing to balance workloads. This has a significant impact on energy efficiency.

The top 500 systems consume more than 5 terawatt-hours annually. If the bottom 470 systems could be made as efficient as the top 30 by leveraging accelerated computing, the result would be an 80% reduction in energy consumption, some 4 terawatt hours fewer every year. What is more, energy savings have a commensurate impact on cost. The Green500 require a collective \$750 million worth of energy for their systems; with greater efficiency, that total cost could be reduced to \$150 million, according to NVIDIA's analysis.

Taking this more nuanced approach to understanding AI energy demands and lifecycle emissions, enterprises are positioned to think through how to prepare their technology infrastructure for high performance computing, as well as efficiency and environmental impact.

The two are not mutually exclusive. This has important implications for the business's Environmental, Social, and Governance (ESG) strategy. In Deloitte's case, our calculations are designed to optimize model training and efficiency, which reduces energy consumption in the on-premise hardware stack and at the cloud provider. In this, the organization can strive for the greatest possible use of its AI technologies while reducing as much as possible the overall energy demands and resulting emissions.

To be sure, there are other contributing factors to overall energy consumption in AI training: the efficiency of the model itself; the places where data is stored and accessed; and even the nature of energy production in a given geography (e.g., energy from a nuclear plant is "greener" than that from a coal plant). These factors will only become more important strategic considerations as the scope and scale of AI and the associated data continues to grow. Ultimately, enterprises can assess their AI training energy usage holistically, and look for opportunities to balance capabilities and corresponding emissions. Accelerated computing with GPUs is one area ripe with such opportunity.

## Get in touch

---

### Christine Ahn

Principal  
Deloitte Consulting LLP  
[chrisahn@deloitte.com](mailto:chrisahn@deloitte.com)

---

### Brandon Cox

Principal  
Deloitte Consulting LLP  
[brandoncox@deloitte.com](mailto:brandoncox@deloitte.com)

---

### Goutham Belliappa

Managing Director  
Deloitte Consulting LLP  
[gbelliappa@deloitte.com](mailto:gbelliappa@deloitte.com)

---

### Tanuj Agarwal

Senior Manager  
Deloitte Consulting LLP  
[tanuagarwal@deloitte.com](mailto:tanuagarwal@deloitte.com)

As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see [www.deloitte.com/us/about](http://www.deloitte.com/us/about) for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting. This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.