Is AI the new Crypto in Terms of Energy Consumption?

Martin Flusberg
July 2024

Image Courtesy of eWeek

Almost 2 years ago I posted an article entitled *Is Ethereum's (Hopefully) Imminent "Merge" a Major Step Towards Lowering Blockchain's Energy Footprint?* The article discussed how much energy was being used to create cryptocurrencies – around the world crypto was then using about as much energy as Taiwan, the country with the 20th largest energy usage in the world. The article was focused on how a change about to be made by the second largest crypto company could have a significant impact on that.

Cryptocurrency is based on Blockchain, a way to maintain and update a database, typically across a network of computers. New data is added one 'block' at a time, and each block refers back to the previous block. To ensure each new block added to the chain is legitimate, a "consensus mechanism" is used that allows everyone on the network – known as *miners* - to check the legitimacy of every block. By far the most widely used consensus mechanism is called *Proof of Work (PoW).* The miner's goal is to be the first to get the answer to a complex problem, which gives them the right to record transactions in the blockchain and broadcast the new block to the rest of the network. Successfully adding a new block to the chain generates some new cryptocurrency that miners keep as a reward. In addition, miners are paid processing fees for the transactions in the block they added.

With the Merge, Ethereum was going to move from Proof of Work to an approach known as *Proof of Stake (PoS).* With PoS, anyone who owns cryptocurrency can pledge their tokens as collateral towards the development of the blockchain and become what is known as "validators". They are responsible for validating the legitimacy of transactions and deciding which ones will be processed first. In return, they are provided with a fixed percentage of the pledged assets when a new block is added to the chain.

There's no need for special equipment; the competition isn't about quickly solving a problem, but rather about how much each party is willing to put up as collateral. This process - called the 'staking' of crypto assets - requires much less computational power, and therefore much less energy is used by PoS than by PoW. The Ethereum Foundation estimated that this change would reduce the energy use of generating cryptocurrency by *more than 99%.* They made the change in September 2022 – and predicted that many other cryptocurrency companies would follow suit.

An article in Forbes published in October 2023 - *One Year After The Merge: Sustainability Of Ethereum's Proof-Of-Stake Is Uncertain* - made the argument that we really do not yet know how effective the Ethereum move was in reducing energy consumption.

The article noted that the *Crypto Carbon Rating Institution (CCRI)* had reported that Ethereum's energy usage did in fact drop by 99% immediately after the Merge. However, while there was very little reporting on Ethereum's electricity consumption in the 12 months following the Merge, CCRI reported that Ethereum's annualized electricity consumption was steadily increasing and *had increased by more than 300%* a year after the Merge.

The Forbes article went on to say that there are several likely explanations for this. For example, post-Merge it is relatively easy to become a validator on Ethereum; Ethereum currently has over 800,000 of them – more than twice the number before the Merge. While each validator may be consuming less electricity than pre-Merge, the sharp increase in the number of validators has contributed to increased overall electricity usage.

Moreover, the Merge changed the economic incentives and spawned an entire new "industry" of people pursuing *Maximum Extractable Value* (MEV). MEV is the maximum value that can be produced by including, excluding, or re-ordering the transactions in a block. Pre-Merge, under PoW, MEV was primarily extracted by miners since they control the block. Post-Merge, a large portion of MEV is extracted by "searchers" - a new group of actors in the Ethereum ecosystem "that run complex algorithms on blockchain data to detect profitable MEV opportunities and have bots to automatically submit those profitable transactions to the network," according to the Ethereum Foundation. While it is difficult to accurately calculate the electricity consumption of the computers being used to seek MEV opportunities, it is clear that simply focusing on how validators write transactions to the blockchain in PoS as compared to miners under PoW is insufficient.

So, the Merge has had a major effect on the energy use of crypto – although the overall reduction in energy use is not as large as expected. And Ethereum's prediction that other crypto companies would follow them has not happened to any meaningful extent. In particular, the largest crypto company, Bitcoin – is still using the PoW approach and has shown no indication of thinking of moving to PoS.

Clearly crypto has been one of the major technological movements in the past 10 years. But perhaps the hottest tech area to really emerge since the Merge is Artificial Intelligence – or AI.

And now there are huge concerns about the energy usage of AI.

Artificial Intelligence

While I suspect that anyone reading this article is well aware of what AI is, I'll just provide a relatively quick summary. AI enables computers to simulate human intelligence and perform tasks that, until recently, required human involvement. It works by taking in huge amounts of data and essentially looking for patterns in order to make inferences and predictions. AI has actually been around for quite a number of years, with early versions used for such things as spell checking, filtering out spam, and offering product recommendations.

A major turning point for AI came with the release of ChatGPT by OpenAI in November 2022. ChatGPT effectively engages in human-like conversations and generates meaningful content. Its users interact with it by typing in questions, or prompts, and based on those prompts ChatGPT can generate essays, write poetry, compose music, generate computer code, and more. Two months after its release ChatGPT had 100 million active users. Since ChatGPT was released, even more advanced AI platforms have emerged that can also be used to generate images and video. As I am sure you are aware there have been a lot of concerns raised about AI, but that is not the focus of this article.

The term that you will see being used most often today is *Generative Artificial Intelligence*. Generative AI is artificial intelligence capable of generating text, images, videos, "synthetic data" (information that's artificially created algorithmically rather than generated by real-world events) and more in response to prompts, using what are called generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics. The recent excitement around generative AI has been driven in large part by the simplicity of new user interfaces for creating text, graphics and videos extremely quickly.

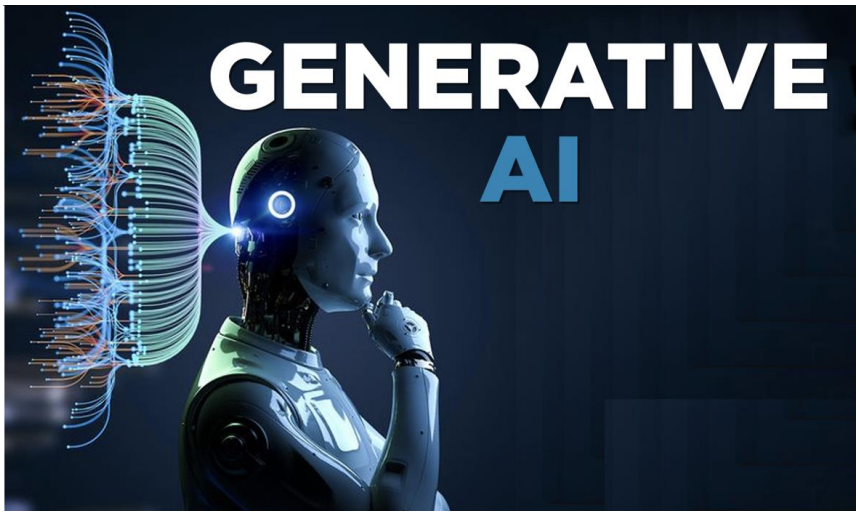


Image Courtesy of LinkedIn

But even Generative AI is not brand-new. It was actually introduced in the 1960s in what were effectively chatbots, but it was not until 2014, with the introduction of generative adversarial networks, or GANs -- a type of machine learning algorithm -- that generative AI could be used effectively.

Recent advances that have played a role in the wider adoption of generative AI include transformer models - and the language models they enabled. Transformers are in effect neural networks and a type of machine learning enabling researchers to train larger and larger models without having to label all of the data in advance. In addition, transformers enabled models to track the connections between words across pages, chapters and books rather than just in individual sentences. Advances in "large language models" (LLMs) - i.e. models with billions or even trillions of parameters -- have enabled generative AI models to write engaging text, paint photorealistic images and more. And innovations in multimodal AI – a new AI paradigm in which various data types (image, text, speech, numerical data, graphics and video) are combined with multiple intelligence processing algorithms to achieve higher performances - enable teams to generate content across multiple types of media. This is the basis for tools like Dall-E - by OpenAI - that automatically create images from text or generate text captions from images.

But all of this development is drawing concerns about how much energy AI is now consuming – and how that may rise considerably if AI is adopted at the levels expected in the future. And AI reportedly also requires the use of millions of gallons of water to cool the equipment at data centers. (Estimates are that Google's data centers used 20 percent more water in 2022 than in 2021, while Microsoft's water use rose by 34 percent. Shaolei Ren, a professor of electrical and computer engineering at UC Riverside, has estimated that engaging in a session of questions and answers with Chat GPT will consume about a half-liter of fresh water).

How Much Energy Does AI use?

A wide range of reports and articles over the past couple of years have talked about the amount of energy being used by AI. For example, an October 2023 article in Scientific American entitled *"The AI Boom Could Use a Shocking Amount of Electricity"* noted that the International Energy Agency has reported that, globally, data centers currently account for 1 to 1.5% of electricity usage and the boom in AI could drive that up significantly**.** The article also referenced a peer-reviewed analysis published in *Joule* that quantified the energy demand and stated that the continuation of the current trends in AI adoption would result in 1.5 million AI processor units a year being shipped by 2027 which could consume over 85 terawatt-hours of electricity annually—or more than what many small countries use.

The peer reviewed analysis that Scientific America referred to was performed by Alex de Vries, a data scientist and a PhD candidate at Vrije University, Amsterdam, where he studies the energy costs of emerging technologies. (He is also the founder of Digiconomist, a research company dedicated to exposing the unintended consequences of digital trends). De Vries had become known for sounding the alarm on the enormous energy costs of crypto and now is focusing on AI. A separate article that appeared on *Energy Technology Revolution* in April 2024 also references de Vries and states that the best estimate we have of global AI energy consumption comes from him

According to Energy Technology Revolution, de Vries took the energy specifications for its AI servers as well as sales projections provided by Nvidia – which provides perhaps 95% of all servers use for AI (more on Nvidia later) - and calculated that by the year 2027 AI could consume between 85-134 terawatt hours per year – which is about the amount of electricity that de Vries' home country, the Netherlands, consumes. (Note that it appears that Scientific America only referenced the lower end of this range). The Energy Technology Revolution article noted that this estimate doesn't account for the likelihood that

those servers don't run at 100% capacity all the time and servers running at partial capacity don't consume as much electricity as fully loaded ones. But de Vries' estimate also doesn't include several other avenues of AI energy consumption, including cooling, memory, and network equipment - and that additional energy consumption from these components probably more than offsets any error introduced by not accounting for capacity.

According to Energy Technology Revolution, exactly how much electric energy is consumed by AI is currently unknown but is clearly growing. The article noted that AI developers aren't eager to share such information, but that early estimates are concerning - according to one researcher, if AI were a country, by the year 2027 it would be about the world's 35th largest national electricity consumer. The article goes on to say that**,** according to the International Energy Agency (IEA), the combined electricity consumption of the major companies developing AI - notably Amazon, Microsoft, Google, and Meta - more than doubled between 2017 and 2021, but it's unknown how much of that energy went to AI.

An article in Vogue also reference de Vries and reported him as having said **"**You're talking about AI electricity consumption potentially being half a percent of global electricity consumption by 2027."

According to Boston Consulting Group, as noted in a LinkedIn article, the data-center share of U.S. electricity consumption is expected to triple from 126 terawatt hours in 2022 to 390 terawatt hours by 2030 - the equivalent usage of 40 million U.S. homes.

Image Courtesy of MIT Technology Review

A March 2024 article on Vox.com argued that "*AI already uses as much energy as a small country*. *It's only the beginning"*. The article went on to say that generative AI uses a lot of energy for training as well as for producing answers to prompts. It said that training ChatGPT uses about 1,300 megawatt-hours of electricity, which is equivalent to the annual consumption of about 130 homes in the US. It then noted that, according to the International Energy Agency (IEA), a single Google search takes 0.3 watt-hours of electricity, while a ChatGPT request takes 2.9 watt-hours. If ChatGPT were integrated into the 9 billion Google searches done daily, the IEA estimated that worldwide electricity demand would increase by 10 terawatt-hours a year — or about the same amount used by about 1.5 million households in the EU.

And one last reference; Sasha Luccioni - lead climate researcher at an AI company called *Hugging Face* – has noted that it is really hard to determine how much energy AI uses. With an appliance you know what energy grid it's using and roughly how much energy it's using. But AI's energy sources are widely distributed, making it very difficult to estimate. That being said, he goes on to state something to the effect that: "my own research found that switching from a nongenerative, older AI approach to a generative one can use 30 to 40 times more energy for the exact same task".

Bottom line; there are a lot of very different takes on how much energy AI uses – but widespread agreement that the usage is high and growing rapidly.

Why Does AI Use so Much Energy?

AI systems, especially those based on large-scale deep learning models, consume significant amounts of energy due to several factors:

1. Computational Intensity

- Training large neural networks requires immense computational power. Models like Chat GPT now have billions of parameters and training them involves processing vast amounts of data through multiple iterations (epochs), which demands extensive use of high-performance GPUs or TPUs.
- Even after training, running these models for *inference* (making predictions or generating text) can be computationally intensive, especially when serving millions of requests in real-time.

2. Data Requirements

- Preparing the vast datasets needed to train AI models involves collecting, cleaning, and preprocessing data, which also consumes significant computational resources and, consequently, energy.

3. Hardware Demands

- AI computations are typically carried out on specialized hardware like GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units), which consume more power compared to regular CPUs. These devices are optimized for the parallel processing required by AI tasks but are energy hungry.
- The high energy consumption of these devices generates a lot of heat, necessitating advanced *cooling systems* in data centers to maintain optimal operating temperatures, which further increases overall energy consumption.

4. Data Center Operations

- The data centers that house AI hardware require substantial energy not only for computing but also for lighting, security, and other infrastructure needs.
- Data centers often have redundant systems to ensure reliability and uptime, which can lead to additional energy usage.

5. Model Complexity

- The trend in AI research has been towards creating larger and more complex models to achieve better performance. Each new generation of models tends to be significantly larger and more computationally demanding than its predecessors, leading to higher energy consumption.
- Continual fine-tuning and updating of models to improve performance or adapt to new data also contribute to ongoing energy use.

6. Distributed Systems

- Distributed AI systems, especially those deployed across multiple data centers or edge devices, require robust networking infrastructure. Data transfer and synchronization across these networks also consume energy.
- Coordinating distributed computations, especially for large-scale training tasks, involves communication overhead and synchronization efforts, which add to the energy consumption.

Transparency Alert: the above items were generated entirely by ChatGPT in response to the question "why does AI use so much energy?" And ChatGPT did a pretty good job in answering the question.

Image Courtesy of codegeeks

Let's explore the questions a bit further.

How AI Uses Energy

The Scientific American article quoted earlier also included a separate interview with Alex de Vries. De Vries talked about the fact that AI has the two major phases (noted earlier): the *training phase* - where you're getting the model to teach itself how to behave – and the *inference phase* – where you place the model into a live operation and start feeding it prompts in order to produce original responses. He points out that both phases are very energy-intensive, and we don't really know what the energy ratio is. Historically, with Google, the balance was 60 percent inference, 40 percent training, but training with ChatGPT appears to involve less energy consumption compared to applying the model. That being said, he goes on to point out that the amount of energy usage is actually dependent on a lot of factors, including how much data is included in the models. For example, the large language models that ChatGPT is

powered by are notorious for using huge data sets having billions of parameters - and making these models larger is a factor in making them more robust and contributes to them needing more power.

A somewhat different position was taken in the Verge article noted earlier. The article talks about the difference between training a model for the first time and deploying it and claims that training, in particular, is extremely energy intensive, and consumes much more electricity than traditional data center activities. It says that training a large language model like Chat GPT-3 would use almost 1,300 megawatt hours (MWh) of electricity; or about as much power as used annually by 130 US homes. (GPT-3 was trained on 45 terabytes of text data according to McKinsey). It then compares that with streaming – noting that streaming an hour of Netflix requires .8kWh of electricity. This means you'd have to watch a staggering 1,625,000 hours on Netflix to consume the same amount of electricity as it takes to train GPT-3. The article then notes that GPT-3 energy use is lower than that of newer more state-of-the-art and complex AI models.

According to the article in Energy Technology Revolution, while more research has been done on training energy usage, AI inquiry energy consumption is not well understood, and developers release little information about it. Inquiry energy consumption depends on a wide variety of factors, including type of task, length of response, and whether text or images are being processed. The article says that the only research they are aware of that measured the energy consumption of specific AI tasks was done by a team from Hugging Face, a company noted earlier. This research team defined a set of tasks, ran those tasks on a variety of AI models, and measured the energy consumed.

Among their findings: generating images consumes more energy than generating text. Generating text, on average, consumed the amount of electric energy needed to keep a 9W LED bulb lit for 19 seconds. Generating an image, on average, consumed enough energy to light that same bulb for 19 minutes, or 60 times as much as text. The expectation was that generating video would consume much more energy.

According to software developer N3XTCODER, Chat GPT-4 is assumed to have used about 25,000 GPUs in its training – which would require a supply of about 20 megawatts of electricity for up to 100 days. Once the model is fully trained, it probably uses nearly as much electricity generating responses from prompts – or inference. Meta attributed approximately 1/3 of their internal machine learning carbon footprint to model inference, with the remainder produced by data management, storage, and training. A 2022 study from Google attributed 60 percent of its machine learning energy use to inference and 40 percent to training.

They go on to note that generating a single AI image is estimated use as much energy as fully charging a smartphone, while text generation is much lower in its energy consumption, claiming that roughly 1/4000th of a smartphone charge is required for a text query.

They also discuss the fact that AI can either be trained on a specific task or on a multitude of different tasks. The range of tasks a model can solve has an impact on its energy consumption. While models with diverse ability appear to be more effective in terms of their overall training footprint, in the inference stage this ability requires more energy than task-specific models.

An article in *Semiconductor Engineering* also dealt with the energy usage of training vs. inference. It says that training consumes a huge amount of power because it iterates over the same dataset multiple times. And the amount of energy consumed by doing that is increasing rapidly. The article stated that: "If you

look at the amount of energy taken to train a model two years back, they were in the range of 27 kWh for some of the transformer models, but if you look at the transformers today, it is more than half a million kWh. The number of parameters went from maybe 50 million to 200 million. The number of parameters went up four times, but the amount of energy went up over 18,000 times".

And how does that compare to inference? According to this article: "Training involves a forward and backward pass, whereas inference is only the forward pass. As a result, the power for inference is always lower. Also, many times during training, batch sizes can be large, whereas in inference the batch size could be smaller." They also note that training is often done more than once in order to enhance the model.

That being said, inference may be replicated many times. Therefor they end by stating: "One prediction is that more than 70% to 80% of the energy will be consumed by inference rather than the training."

Once again there are wide discrepancies on estimates AI energy use.

Another way of looking at the reason for the extensive energy use of AI is to consider that it is essentially "manufactured" across multiple layers of equipment: chips, servers, and data centers. At the most basic level, AI's algorithms are processed by specialized computer chips that perform trillions of operations per second. As Chat GPT noted above, chips being used for AI are composed of discrete "graphical processing units" – or GPUs - rather than the central processing units (CPUs) found in standard computers. A GPU is a specialized processor, originally designed to accelerate graphics rendering, that is extremely effective at parallel processing - able to process many pieces of data simultaneously - making them useful for machine learning, video editing, and gaming applications and therefore making them the chip of choice for AI. And making them much more energy intensive. (That being said, Nvidia has reported to have claimed that, for AI workloads, two GPU servers can do the work of a thousand CPU servers at a fraction of the cost and energy).

And an overall power consumption increase will come on two fronts: an increase in the number of GPUs sold per year and a higher power draw from each GPU. Research firm 650 Group expects AI server shipments will rise from one million units last year to six million units in 2028. According to Gartner, most AI GPUs will draw 1,000 watts of electricity by 2026, up from roughly 650 watts on average today.

Rendering in Recent Forbes Magazine Article: *Futuristic AI Needs Futuristic Data Centers , and it's Just Starting*

Those chips are incorporated into "industrial strength" computers – i.e. servers. They handle data processing and storage and are networked with other servers and storage devices. The servers are housed in Data Centers, which contain backup power and the electric infrastructure required to essentially avoid outages, cooling equipment to ensure that the equipment doesn't overheat, and the cabling that physically connects the servers to each other and the internet. Most tech firms that run data centers don't reveal what percentage of their energy use processes AI. The exception is Google, which says that "machine learning" accounts for somewhat less than 15% of its data centers' energy use.

But things are starting to be done to address the issue of rising energy consumption.

What are AI, Computer, and Data Center Companies Doing to Reduce Energy Usage?

As AI began to expand rapidly and people began to recognize how much energy it was consuming, numerous companies have begun looking at what they may be able to do to lower energy usage.

Here are some of the actions being taken[1].

1. **Making the AI models more efficient**

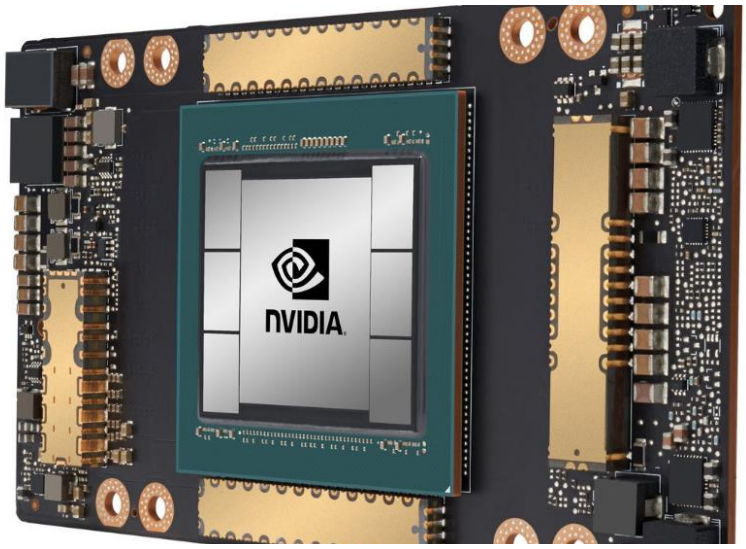
   While there is very little being written about this, several of the companies in the AI space have indicated that they are working on ways to lower the energy usage of AI models. Interestingly, the person from Hugging Face mentioned earlier has suggested that users be provided with information on the energy usage of different AI models so that they can choose one that is more energy efficient – and that he was actually working on something akin to an Energy Star rating for AI models to help make this happen.

2. **Developing more efficient hardware**

   Nvidia, the company that supplies the vast majority of GPU chips being used for AI, has expressed concerns about energy efficiency for a while. (Nvidia exploded to become the most valuable company in the world recently, but a sell-off put it back to the number 3 spot behind Microsoft and Apple in late June 2024). In March 2024 Nvidia announced that its next generation of GPU chips would be faster and more energy efficient than the current ones, estimating that the new chips would cut the energy consumption of training an AI model by nearly 75%.

   Other chip manufacturers are also developing more energy efficient AI products. For example, IBM has been trying to address one of the most energy intensive parts of the AI chip process, which relates to moving data between the computing and memory sections of the chips. They have reported that they found a way to more tightly integrate computing and memory so it will no longer be necessary to move information between them. IBM claims its new chips will be about five times as energy efficient as Nvidia's current GPUs (which would make them about 20% more efficient than Nvidia' new ones if Nvidia's claims are accurate).

[1] A significant amount of this section was taken from the Energy Technology Revolution article noted earlier.



Nvidia's H100 AI GPUs Projected to Surpass Energy Consumption of entire nations.
Source: Technovedas Semiconductor News

3. Implementing more efficient cooling systems

As noted earlier, a huge portion of the energy used to power AI chips and other electronic components ultimately turns into heat. Vertiv, a leading provider of power and cooling infrastructure equipment, has said that AI servers generate five times more heat than traditional CPU servers and require ten times more cooling per square foot. Data centers generally blow air over the hardware to reduce heat, while the HVAC systems cool the air down and try to keep outdoor heat out. Cooling systems account for about 40% of data center energy consumption. Because these systems generally leverage evaporation to help reject heat, they also consume a lot of water.

Air-based cooling systems consume a lot of energy and also use a lot of noisy fans. To help keep chips cooler, reduce energy use, and generate less noise, data centers can move to liquid-based cooling. For example, in one type of liquid cooling system, servers are submerged in a non-electric-conducting liquid. The servers heat the liquid which is then pumped to equipment that cools the liquid and transfers heat outdoors. Because it's more efficient to move liquids than air, and cooler chips consume less energy than hotter chips, implementing liquid cooling can reduce overall data center energy usage by more than 10%. Liquid Cooling technology provider Super Micro has stated that they believe that switching to liquid cooling from traditional air-based cooling can reduce operating expenses by as much as 40%.

The UC Riverside professor studying water use of AI noted earlier has also pointed out that data centers' water use is a particular issue in drought-afflicted areas. To cool delicate electronics in the clean interiors of the data centers, water has to be free of bacteria and impurities that could gunk up the works. In other words, data centers often compete for the same water people drink, cook, and wash with.

4. Recovering data center waste heat

Instead of simply trying to transfer the heat removed by cooling systems into the atmosphere, some data centers actually circulate that heat to nearby buildings, including industrial facilities. For example, The Air, a data center located in Helsinki, uses heat absorbed from its electronic equipment to heat water to nearly 90°F, and then uses heat pumps to increase the temperature to nearly 200°F. That high-temperature heat is then sent into Helsinki's district heating and cooling system, which supplies heating and cooling to both homes and commercial buildings. The Air provides 1.3 MW of thermal power to the system, which is enough to meet the heating requirements of more than a thousand US households.

As another example, Stanford Energy System Innovations, a Stanford University program, applied the same approach for the Universitat Politècnica de València in Spain, (as well as some other universities including their own). They focused on a campus data center with an almost constant cooling demand across the year which used about 1.67 million kWh/year. To recover the waste heat generated by the data center, a method for assessing thermal performance of the system was implemented using a 300kW polyvalent heat pump (capable of easily switching from water-to-water to air-to-water working modes). It can simultaneously provide cooling and heating to a set of buildings on the campus: the data center itself and three buildings located nearby that currently use the university's central natural gas boilers for heating. A thermal storage system was added to balance cooling and heating needs. The results led to estimated thermal energy savings of more than 250,000 kWh/year.

5. Using AI to reduce data center energy consumption

Google has reported that researchers at its DeepMind lab had, ironically, trained a machine learning model to identify ways to reduce energy usage. They used data from thousands of Google data center sensors, including temperature, power meter readings, pump speeds, and setpoints, and trained the model to predict future operating conditions and attempt to reduce cooling system energy consumption. By adjusting setpoints to anticipate actual server processing loads, the Google researchers found that they could reduce cooling system energy consumption by 40%. Meta and Microsoft have also announced that they were using AI to improve data center energy efficiency, but they have provided little detail on what they have done or what effects it has had. According to the VP of Sustainability for Equinix, one of the nation's largest data-center companies: "AI can be used to improve efficiency, where you're modeling temperature, humidity, and cooling. It can also be used for predictive maintenance." Equinix has stated that using AI modeling at one of its data centers has already improved energy efficiency by 9%.

6. Transitioning data centers to renewable energy

Apple, Google, Meta, Microsoft, Amazon, and other data center companies have claimed to either already use 100% renewable energy or be in position to do so soon. The reality is that these companies typically buy carbon credits to offset their consumption. Moreover, because renewables are intermittent, there are numerous hours during which there isn't enough renewable energy available to power the data centers even if they are tied to renewables unless they can leverage data storage.

Google and Microsoft have pledged to achieve 24/7 carbon-free energy use by 2030. They have formed a partnership to help develop the technologies to achieve this and outlined their plans in a document they call the 24/7 Carbon Free Energy Compact.

Source - Spiceworks Article*" Combatting AI Energy Consumption through Renewable Sources"*

And What Does ChatGPT Have to Say About Reducing AI Energy Usage?

According to ChatGPT, to address the high energy consumption of AI, several strategies can be employed:

- **Improved Algorithms:** Developing more efficient algorithms that require less computational power.
- **Energy-Efficient Hardware:** Investing in hardware specifically designed to be more energy-efficient for AI workloads.
- **Optimized Data Centers:** Implementing more energy-efficient cooling and power management solutions in data centers.
- **Renewable Energy:** Shifting to renewable energy sources to power data centers can reduce the carbon footprint of AI operations.
- **Model Compression:** Techniques like model pruning, quantization, and distillation can reduce the size and complexity of AI models, making them more efficient.

By understanding and addressing these factors, the AI community can work towards more sustainable and energy-efficient AI systems

Thanks Chat!

Are Governments Doing Anything to Address the Increasing Energy Usage of AI?

A February 2024 article in Yale Environment 360, entitled *As Use of A.I. Soars, So Does the Energy and Water It Requires* had the following statement as its subtitle: Generative artificial intelligence uses massive amounts of energy for computation and data storage and millions of gallons of water to cool the

equipment at data centers. Now, legislators and regulators — in the U.S. and the EU — are starting to demand accountability.

And, indeed, there do seem to be things beginning to happen.

"The development of the next generation of AI tools cannot come at the expense of the health of our planet," was a statement by Massachusetts Senator Edward Markey early in 2024 after he and other senators - and representatives - introduced a bill that would require the federal government to assess AI's current environmental footprint and develop a standardized system for reporting future impacts. The **AI Environmental Impacts Act of 2024** would require the US EPA and other federal agencies to study the environmental consequences of AI, including energy and water consumption, and make recommendations for legislation to help address the issue. The Act also requires the National Institute of Standards and Technology to set standards for quantifying AI's environmental impacts and maintain a voluntary reporting system for AI-related companies. The data collected by the reporting system would form the basis of a report to Congress.

The European Union's "A.I. Act,", approved by member states earlier in the year, will require "high-risk AI systems" (which includes ChatGPT) to report their energy consumption, resource use, and other impacts throughout their systems' lifecycle. The EU law takes effect in 2025.

So governments are beginning to act.

Where Is AI Energy Usage Headed?

There are a lot of different possible responses to this question.

On the pessimistic side, if you project that AI use continues to grow at current rates and assume that the equipment that produces it continues to operate at current efficiencies, the calculations suggest that energy use will increase astronomically – likely by at least a factor of 10. Some have argued that as AI becomes even more advanced it will require even more computational power and intricate hardware, consuming even more energy than it already does. And data centers, excluding crypto mining, are already responsible for about 1 percent of global electricity consumption.

On the optimistic side, the Energy Technology Revolution article cited earlier took the position that: "in the end, AI energy consumption will probably follow the same trajectory as overall internet-related energy did. Around the turn of the century, analysts predicted that internet energy consumption would grow to levels that would overwhelm the power grid. Instead, data center operators and hardware manufacturers improved efficiency, so that even though internet data usage grew immensely, data centers consume only about one percent of overall global electricity consumption".

The article noted that, in 1999, two analysts working for the Greening Earth Society, a now defunct public relations organization funded by coal companies, claimed that the Internet accounted for 8% of US electricity consumption. Furthermore, they predicted that by the end of the following decade, energy consumed to manufacture, use, and network computers would account for about half the output of the electric grid. These claims led a group of scientists at Lawrence Berkeley National Laboratory (LNBL) to conduct a study that concluded that all Internet and other computing equipment was only consuming

about 3% of US electricity. LBNL kept updating this and in 2016, concluded that in 2014 data centers accounted for 1.8% of total U.S. electricity consumption. A report by IEA stated that from 2015 to 2022 worldwide data center workloads went up 340%, but that data center energy use only went up 20% to 70% -and that global data center electricity consumption stood at 1-1.3% of total energy use.

The reason that the contribution of data centers to overall energy use went down rather than up was that the energy efficiency of the equipment improved, as well as the effectiveness of the cooling systems, and the industry shifted from many small data centers, to fewer huge, but more efficient facilities**.**

Assuming the same thing happens with respect to AI, we can expect its energy usage to drop significantly. And, if we see a huge shift to renewable energy to power the data centers, which is something that is already happening, the environmental impact of AI can actually go down rather than up.

So, I'll end the article here, on this optimistic note.