January 20254

ELECTRICITY FOR DATA CENTERS...

IS AI THE DRIVING FORCE?

Dr Lars Schernikau

Microsoft Meta Dropbox

CONTENT

Are we approching the AI revolution with our heads in "The Cloud"?

If so, what is actually necessary to handle the ever increasing demand for cloud storage, streaming, virtualization, artificial intelligence, and of course blockchain applications including digital currencies?

Data centers!

… which are massive server farms, consisting of hundreds, sometimes thousands, of computer racks stacked and clustered together… and all of this revolves around electricity?

These increasingly powerful computer chips (used to enable all of these functions and applications) require significant amounts of energy to execute their complex parallel operations. Now of course, the latest chip technology (inc. quantum computers) are becoming increasingly more energy efficiency, but the total power demand will still multiply.

Let's have a look and dig deeper into these industrial-scale "new" infrastructure projects that will run our future lives.

The Oxford dictionary states *"a large group of networked computer servers typically used by organizations for the remote storage, processing, or distribution of large amounts of data."*

Wikipedia goes further *and adds an important element: "Since IT operations are crucial for business continuity, it generally includes redundant or backup components and infrastructure for power supply, data communication connections, environmental controls (e.g., air conditioning, fire suppression), and various security devices.* ***A large data center is an industrial-scale operation using as much electricity as a medium town".***

1. Data centers, where it all started and where are we today?

Cloud computing took off like a rocket less than a decade ago. Platforms like Dropbox, enterprise solutions from Amazon and Microsoft and countless other new cloud systems, popped up during the Covid period. In fact, in 2019, cloud infrastructure services surpassed data centers' hard- and software for the first time (90 B USD, as per Synergy Research Group [1]).

Let's take a step back to the first data center (called a "mainframe") which was built in 1945 to house the ENIAC at the University of Pennsylvania. The Electronic Numerical Integrator and Computer (ENIAC) was the first electronic digital programmable general-purpose computer. The U.S. military designed the ENIAC to calculate artillery firing tables. However, it was not completed until late 1945. It's first program was apparently a feasibility study for nuclear weapons. [2]

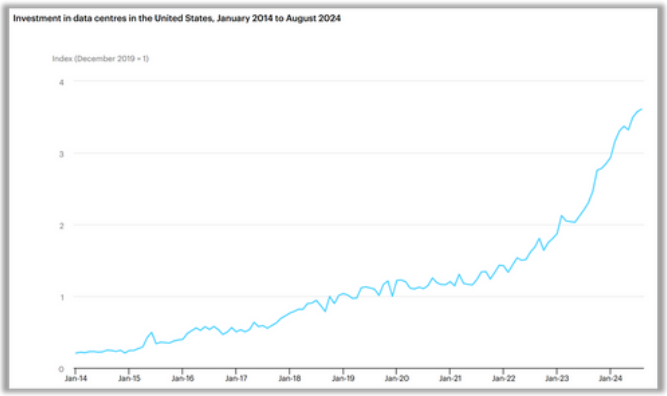
Mainframes from the 50s and 60s probably mark the legitimate birth of data centers, large buildings housing huge computer equipment with vacuum tubes and punch card systems.

In the 60s the first super-computer was introduced, the CDC6600. Transistors, magnetic tapes and networking advances changed the "data center" landscape in the 70s. Micro-computers of the 80s switched the focus to personal computing and it was not until the 90s that the internet "invention", and with it the importance of large "modern" connected data centers re-emerged.

Today's high speed data transmission allows for a different level of virtualization, streaming, cloud computing, AI, blockchain and more. ***Data centers have come back under the loop because their electricity demand is becoming a serious issue to be considered.***

As advanced computer hardware increase efficiency, the total energy demand will continue to surge and make up an ever-increasing share of total power consumption, like in Ireland where data centers already account for over 20% of electricity demand [3].

Figure 1: Data center investment in the US, IEA [6]



2. Charge me baby… data centers and AI need electricity

Apparently, there are already over 11.000 data centers worldwide, often spatially concentrated. The IEA [3] recently correctly stated that we have to put things into perspective. In 2024 ***data centers accounted for only around 1% or over 300 TWh of global electricity consumption*** (Germany consumes about 2%), and annual electricity consumption of data centers globally thus far, makes up about half of the electricity consumption from household IT appliances, like computers, phones and TVs. But there is no country growing faster than the "data center country" will in the years to come… so we will see some major changes taking place (Figure 1).

The IEA continues that in large economies like the United States, China and the European Union, data centers already account for around 2-4% of total electricity consumption and because they tend to be spatially concentrated, their local impact can be significant. This sector has already surpassed ***10% of electricity consumption in at least five US states*** in comparison to Ireland's 20% of all electricity consumption.

- The IEA also informs that large data centers can have a power demand equivalent to that of an electric arc furnace steel mill. However, steel plants are less likely to be clustered in the same geographic area like data centers are.

Bloomberg 2024 [4] commented on the recent boom: ***"AI Needs So Much Power That Old Coal Plants Are Sticking Around"***… and this is not an exaggeration. US electricity demand is expected to rise to 1.5% p.a. or more, making it the fastest growing rate since the 1990s. EVs, battery factories, computer chip plants and other "clean tech" manufacturing subsidized by governments, will only be a portion of this increasing demand. This creates a risk of energy grids collapsing. Obviously, old coal plants have to come to the rescue because there isn't a single data center that can run on an intermittent power source such as wind or solar or even varying river flows that is Hydro (see my blog on Hydro here).

Now accounting for the new AI technology, Brian Gitt once correctly said that today a single Nvidia AI GPU chip consumes about the same power as a typical US home without AI.

Even more interesting is the statement that peak power demand (in GW) for data centers may grow faster than the electricity consumption (TWh), which is what we have been witnessing globally already. McKinsey estimates that ***peak power demand for data centers will triple in the US alone to 80 GW*** [5], which equals all of Germany's current peak power demand. Such an uninterrupted peak power demand for serving supply, can of course never come from any combination of wind, solar and batteries, since batteries are only short duration energy storage.

The IEA's 2024 World Energy Outlook [6] discussed peak power demand for the first time in more detail. Independent of AI and data centers, primary energy demand will continue growing, with electricity demand expected to increase 6x faster, and peak power demand outpacing electricity consumption… very little has been mentioned about this crucial situation at national energy policy functions!

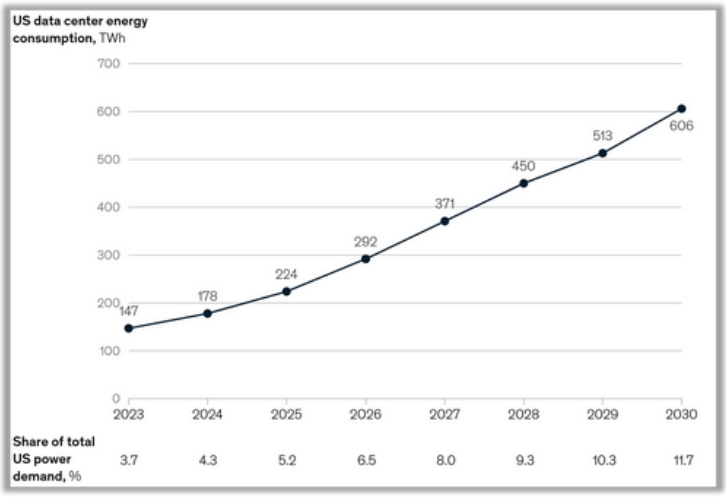
Energy is a three-tiered system (see my LinkedIn Post):

1. peak power,
2. electricity,
3. primary energy
    - 40% of primary energy provides the basis for electricity generation
    - Electricity provides the basis to serve peak power demand

Even though data center's development may be hampered by supply chain bottle necks and even the power grid itself, the IEA [3] estimates in its base case that data centers will make up 10% of electricity demand growth in the next decade, which I would question to be too low.

"Data centers have an exceptionally high spatial concentration, which has significant implications for local power grids, given their substantial power requirements". I would also argue that data center's development will increase power grid failures… see section 5 Data centers, AI, crypto, and the environment below.

Figure 2: US Data center power demand, McKinsey [7]



3. What do Google, Meta, Microsoft, Apple and Co. say?

Disclaimer upfront, I am an avid technology user and find products and services marketed by the stated companies not only helpful but highly important for the future of mankind despite questioning morality. Here I criticize only one aspect of all 4 big companies and that is ***energy policy***, which is not to be misinterpreted nor misunderstood.

Google, Meta, Microsoft, Apple, and many more claims to make use of 100% renewable power supply for their data centers, insinuating 100% "decarbonization" which is not only possible but also apparently already achieved [7].

Such claims are even made pertaining to their China based data centers. Clearly this is what we know as "green washing", as most of us are "buying" credits and applying them towards their power use.

Most concerning may be that these large companies have been anti- coal/gas and pro wind and solar for a very long time. Only recently, as the concerns and total costs of wind and solar systems become more obvious, did they start to endorse nuclear. Let's be clear, ***there is not one single data center in the world that can run on intermittent wind and solar***, it is physically impossible and as wind and solar capture more market share, the total costs go through the roof. At this rate, not even nuclear, will independently be able to meet our future energy demands.

Another interesting point is data center operators claiming 100% "renewable" because of their utilization of hydro, then I ask you to consider the following: (1) hydro is a limited resource, if growing demand exhausts hydro power, someone else, like for example a hospital, falls short of energy supply forced to then turn elsewhere, usually to coal or gas. Also, an undisputed fact among people that deal with hydro power is (2) hydro is far from "zero CO2 or carbon" as it is often hailed. More on this in my blog post "the Unpopular Truth about Hydro Power"

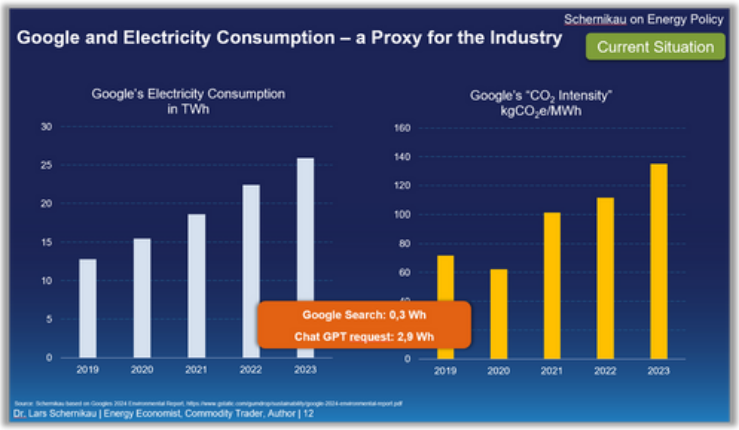
Google has been a particularly interesting case. Not only did Google's power demand double in 5 years, but its "CO2 intensity" also, simultaneously doubled per MWh. Thus, Google's "CO2 footprint" quadrupled (Figure 2). This appears to be completely contradictory to its "24/7 Carbon-Free" claims… all self-reported of course.

So, when I say that an ***AI request consumes 10x the energy that a standard Google search*** consumes… then you get a glimpse of what is heading our way…this is not even considering the "non-search" applications of AI which are far more important in this regard.

Operating on 24/7 Carbon-Free Energy by 2030.

Google Sustainability

Figure 3: Google's electricity consumption and "CO2 Intensity"



Critical Equipment: All data center operators and investors are facing struggles relating to the availability of critical equipment. Over the past 5 years, lead-times for generators, uninterruptable power supply, transformers, switchgears, and power distribution units have doubled and even quadrupled. Trained technicians are scares and in high demand due to limited services.

The raw materials and supply chains involved in providing these data centers, are rarely discussed.

Disposal … where all this goes after 3-7 years (which is how long the computers last)… is also not discussed. The environmental " friendliness" of "net-zero" data centers is laughable. They are industrial installations that have industrial footprints even if they don't have a smokestack… that smokestack can just be found somewhere else. Just like a "clean" EV driving in China charged with power that comes from a thermal coal power stations hundreds of kilometers away.

4. AI Bots: A DDoS attack on the entire Internet?

Artificial intelligence does not only require costly and energy hungry data centers, but also impacts "visited sites" and increases network traffic elsewhere impacting computer infrastructure and of course again…energy consumption. ***This is NOT accounted*** for in the numbers you see from the IEA or McKinsey.

As we are now becoming more technical, let me explain the scenario as follows… something called "AI crawlers" constantly crawl the internet, switching strategies, "learning", adjusting and reaping any information they can get their "hands on". A German network administrator called AI crawlers a ***"DDoS attack on the entire internet".***

The network administrator from Germany (Diaspora [8]) explained that 70% of data traffic on certain sites is caused by AI "crawler" bots. These requests use up server time and clog the system that they "crawl", resulting in the need for larger and more energy intensive infrastructure to deal with these additional requests.

Figure 4: DDoS attach on the Internet? [9]

Dennis Schubert
vor 20 Tagen

Excerpt from a message I just posted in a #diaspora team internal forum category. The context here is that I recently get pinged by slowness/load spikes on the diaspora* project web infrastructure (Discourse, Wiki, the project website, …), and looking at the traffic logs makes me impressively angry.

and the list goes on like this. Summing up the top UA groups, it looks like my server is doing 70% of all its work for these fucking LLM training bots that don't to anything except for crawling the fucking internet over and over again.

Oh, and of course, they don't just crawl a page once and then move on. Oh, no, they come back every 6 hours because lol why not. They also don't give a single flying fuck about `robots.txt`, because why should they. And the best thing of all: they crawl the stupidest pages possible. Recently, both ChatGPT and Amazon were - at the same time - crawling the entire edit history of the wiki. And I mean that - they indexed every single diff on every page for every change ever made. Frequently with spikes of more than 10req/s. Of course, this made MediaWiki and my database server very unhappy, causing load spikes, and effective downtime/slowness for the human users.

If you try to rate-limit them, they'll just switch to other IPs all the time. If you try to block them by User Agent string, they'll just switch to a non-bot UA string (no, really). This is literally a DDoS on the entire internet.

Just for context, here's how sane bots behave - or, in this case, classic search engine bots:

- 16.6k requests - 0.14% - `Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)`
- 15,9k req - 0.14% - `Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm) Chrome/116.0.1938.76 Safari/537.36`

Because those bots realize that there's no point in crawling the same stupid shit over and over again.

I am so tired.

This example shows that only 0,5% of data traffic comes from classic Google and BING bots and 70% from GPTBot, ClaudeBot and other "AI crawlers". Here a quote from the web administrator on the effected site:

"Oh, and of course, they don't just crawl a page once and then move on. Oh, no, they come back every 6 hours because lol why not. They also don't give a single flying fuck about robots.txt, because why should they. And the best thing of all: they crawl the stupidest pages possible. Recently, both ***ChatGPT and Amazon were – at the same time – crawling the entire edit history of the wiki.***

If you try to rate-limit them, they'll just switch to other IPs all the time. If you try to block them by User Agent string, they'll just switch to a non-bot UA string (no, really). ***This is literally a DDoS on the entire internet." [8]***

5. Summary and implications

It doesn't take much more than common sense to understand that technological progress, although crucial to our further development, has tremendous energy demand implications, whether it be artificial intelligence, the "cloud", space travel, blockchain, or robotics.

Mark Mills [9] noted that capital spending on global cloud infrastructure now exceeds global capital spending by all electric utilities combined, even though we're electrifying everything.

The cloud is not the internet. The cloud uses the internet in combination with data centers and data transmission lines as factories and highways.

The impact of technological progress on the environment is manyfold, whether it be "just" the energy, raw materials and water required, or the heat emitted by large server farms. *"With drought spreading around the globe, battles over water are erupting between AI companies seeking more computing power and communities where their facilities are located."* (Bloomberg [2]).

Global primary energy consumption more than doubled in the last 40 years, while the energy intensity of GDP improved/reduced by a little less than 1%. This confirms the Jevon's Paradox that ***energy efficiency improvements are in principle always offset by higher energy demand*** in this case from technological progress such as data centers for the cloud and AI (Polimeni et al. 2015).

I predict that data centers, or more generally "cloud computing", encompassing the AI revolution, has the potential to consume 10% of global electricity within the next 10-15 years, making a much bigger impact on peak power. That peak power demand needs to be ***met and this can never be done by wind and solar without expensive and extensive short and long duration storage, thermal power backup, and vast network and transmission infrastructure.***

Even nuclear will not be the solution. Consider that when a new datacenter uses nuclear, a scarce resource, then the hospital next door has to go its power elsewhere (see my article on nuclear here, published at Oxford Institute)

Either (1) the lights go off, (2) energy costs skyrocket, or (3) we pull up our sleeves and build technologically advanced and clean mix of coal, gas, hydro, and nuclear infrastructure to give us what we need for human progress!

Links and Resources

[1] The Cost of Cloud, a Trillion Dollar Paradox, Andreessen Horowitz, May 2022. link
[2] Bloomberg; Thirsty Data Centers Are Making Hot Summers Even Scarier. July 2023. link
[3] Bloomberg: AI Needs So Much Power That Old Coal Plants Are Sticking Around (1)." January 2024. link
[4] More details with links at Alex Epstein post, May 2024, link
[5] Econnex, link
[6] IEA: What the Data Centre and AI Boom Could Mean for the Energy Sector – Analysis," October 2024. link
[7] McKinsey: Data Centers and AI: How the Energy Sector Can Meet Power, September 2024. link.
[8] IEA World Energy Outlook 2024, October 2024, page 45, link
[9] Dispora AI bots crawl the internet, German language summary, English language primary source
[10] Mark Mills on the Cloud, the Robot Revolution, and Machines That Think, 2023, link