12 June 2024

We'll Run out of Energy before we Run out of AI

Dear Investors,

In today's note we consider the implications of Artificial Intelligence (AI) on electricity demand and renewable energy requirements.

****

If we take a broader picture view of where AI might end up, we think it has dramatic consequences for US power demand. US utility companies, in the most recent quarter, have only just started to work this out.

AI technology is rapidly improving. The chances that we need 1 GPU per person is increasing quickly. For instance, a world where the productivity enhancements of AI mean that everyone uses search products embedded with a large language model (LLM) is fast-approaching. Chat-GPT released a new search function that uses AI and the Perplexity AI search tool has been found to be useful by many. Google is being forced to embed its Gemini LLM into its search function to compete. Companies like Google will need to facilitate peak search and LLM usage in a similar way to how utilities need to meet peak electricity demand in the morning and evenings.

In addition to search, the technology powering co-pilots is improving fast. At their recent technology conference, Google announced a plan to have a customised LLM for each individual person. The company's ultimate goal was to store and apply 'infinite context' about individuals or businesses whenever they asked the AI to do something. For instance, an AI writing an email for you would have data taken from every email you've ever written to help guide it. If co-pilots end up being used for everyday thinking, writing and spreadsheet tasks as the technology improves, then we may need more than 1 GPU per person in the developed world.

This approach on focusing on a personal assistant makes a lot of strategic sense to us. There are a few reasons. First, the threat to these Big Tech companies is becoming existential. If the LLM assistant becomes the interface between the user and what they want or need from technology applications, then the LLM co-pilot may become the operating system of operating systems. Microsoft's operating system may come under threat if Google can make it easier to complete work on its own operating system. Second, every technology that has been transformational has had an extremely wide reach and total addressable market. If AI is to have a similarly big impact, then it needs to become entrenched everywhere.

AI, though, isn't Siri. The reason is because the most productivity Siri could offer their users well and consistently was to reduce the need to use your hands. Siri couldn't write an email for you, or plan a holiday for you. Siri didn't have the potential to save the user a significant amount of time. AI, however, does. Whether the potential gets realised or not after the low hanging fruit is first picked will be the key determinative question.

But in addition to potential demand from search and co-pilots, AI is in demand from companies seeking to automate processes and reduce headcount. In its current form and point of development, AI can't be used for every task. AI won't replace lots of people until the infrastructure supporting AI use cases can lead to a machine that is more easily instruct-able. Humans, for the time being, are much easier to instruct because they understand context. Again, this is why we believe Google's customisation path makes sense. Customisation is designed to lead to a machine that is more easily instructed to your instructions and preferences because it has your *context*.

In addition to this, if we move towards a world of automated taxis like Google has via Waymo in San Francisco and Baidu has in Beijing, Chongqing, Wuhan, Shenzhen and Guangzhou, then the world will need a lot more GPUs and NPUs than today.

In addition, gaming companies are using AI to make their games more realistic for a huge global audience. As the relationship between humans and computers becomes increasingly important, the language that connects the two

will become more valuable. Education systems will likely be required to adapt and school children will also add to GPU demand.

Not many were able to see what would end up happening when the Apple iTunes and Google Play stores were first invented. There are apps that have since been created that some can't live without because of their usefulness, their convenience or their social networking aspects. Some examples of these apps include AirBnB, Uber and UberEats, Google Maps, internet banking, Instagram, Facebook, WhatsApp, Youtube, Tinder and Bumble to name a few. We believe that AI will enable an upgraded app store that will provide you with even better tools, and that we're at the start of this process. With the technology where it is today, we believe tasks that are highly contextualised or where the infrastructure can be built easily to allow for a significant number of inputs and consistent outputs will be the first productivity improvements to be made. As the technology gets better, we believe further improvements are likely to be added over time.

The next generation may find AI easier to use and integrate. A generation that may use AI to help them with university assignments or a personalised, multi-modal AI tutor for schooling may make the use of AI habitual.

Combined, search, co-pilots, summarisation, the automation of processes and service offerings of businesses, autonomous driving and gaming alone may increase the amount of GPUs we need over time to somewhere around 1 GPU (of current NVIDIA Hopper quality) per person. While the computational power of GPUs will increase and improve the ability to make computational networks, so will their price and power usage.

****

As a thought experiment, let's assume that by 2032, the US will need 1 GPU (Hopper-equivalent) for every 3 people. We know that, as a rough rule of thumb, *4,000,000 NVIDIA Hopper GPUs increases total US power demand by around 1.0%*. We also know that estimates from SemiAnalysis (who break down the power needed from AI at the server level) suggest that we can currently make around 7,000,000 GPUs per year between Google, NVIDIA and others. The following table shows how much extra power will be needed by 2032:

Year Ending	GPUs Made	Cumulative GPU Stock	Power Demand (PD)	Efficiency Factor	Cumulative PD
2024	7,000,000	3,000,000			5.00%
2025	9,000,000	12,000,000	2.25%	1.00	7.25%
2026	10,000,000	22,000,000	2.50%	1.40	9.04%
2027	11,000,000	33,000,000	2.75%	1.00	11.79%
2028	12,000,000	45,000,000	3.00%	1.35	14.01%
2029	13,000,000	58,000,000	3.25%	1.00	17.26%
2030	14,000,000	72,000,000	3.50%	1.30	19.95%
2031	15,000,000	87,000,000	3.75%	1.00	23.70%
2032	16,000,000	103,000,000	4.00%	1.25	26.90%

Source: SemiAnalysis and Own Estimates

In 8 or so years, the amount of electricity needed by data centres will increase to 27% to total US electricity consumption from around 5% today.

Moore's law, which described how humanity could increase the number of transistors that could fit onto a single chip (so that compute power and battery life could improve as electrons had less distance to travel) is now being applied at the network level. NVIDIA recently discovered something Google had known for 5 years – the most efficient way to compute and infer is to join as many chips together as possible in as small a space as possible. It's Moore law applied to a group of chips rather than a single chip. Using passive copper to interconnect as many GPU and CPU chips together reduces electricity consumption. All of this is now possible because of the ability to liquid cool chips within the datacentre, allowing for many more chips to fit within one server rack. We estimate that this will improve

electricity usage, at a macro level, by around 30% to start but will gradually get harder to achieve further gains. We've estimated for these improvements via an efficiency factor in the table above.

The implications of the forecast are significant. To have 1 GPU for every 3 people, we need to increase the grid by around another 1/3rd. To put this into context, US electricity demand has been growing by around 0.5% per annum for many decades, but that might ramp 6x to around 3% per annum in just a couple of years. The increase in the number of gas-fired power plants and renewable energy farms needed to enable this technology will be significant relative to how many we have in place today.

None of this demand for electricity accounts for the increasing demand from our need to reduce our fossil fuel consumption. The electrification of cars, engines more generally, other transport (ships, planes and trains) and other equipment that currently uses oil will add to this demand for electricity. Neither does this forecast account for the onshoring that's happening within countries because of an increasingly bifurcated geopolitical world.

****

The implications for our transition to a planet powered by renewable energy is hard to understate. Before the advent of AI, onshoring and autonomous vehicles, Bill Gates estimated that we'd need to add 75 gigawatts (GW) of renewable capacity ***every year*** for the next 30 years to decarbonise America's power grid by 2050. The demand for power driven by these new trends will likely increase this by at one-third more. This means that we would need to add 100 GW of renewable energy each year. Is that a lot? Last year, the US added around 37 GW of renewable energy. That's only around one-third of what we need to add every single year for the next 30 years. This deficit will accumulate.

Added to that, the Biden Administration has ruled that coal power plants will need to decarbonise by around 2032. Currently, coal still supports around one-sixth to one-fifth of total US electricity power generation. In other words, around 20% of US total electricity supply will need to be replaced with clean sources on top of the heightened speed of demand for electricity.

****

In the most recent quarterly earnings season, we're starting to see US utility companies recognise the scale of the problem. They have triangulated the picture. In the following section, we provide you with commentary from some of the largest companies in the space:

- Dominion Energy is a utility that serves many of the data centres in Northern Virginia, which is the largest data centre market in the world (because it's closest to where the internet was invented and also where undersea cables were first linked) and where around 35% of hyperscaler data centres are located. In their latest quarterly earnings report, Dominion said, "we're ramping into the very substantial and growing multi-decade utility investment required to address resiliency and decarbonisation public policy goals, plus the very robust demand growth we're observing in real time across our system. Our year-over-year sales growth rate through March was 4.8%...driven by economic growth, electrification and accelerating data centre expansion. The data centre industry has grown substantially in Northern Virginia in recent years. In aggregate, we've connected 94 data centres with over 4GW of capacity over the last approximately 5 years. We expect to connect an additional 15 data centres in 2024…In recent years, this growth has accelerated in orders of magnitude, driven by 1) number of data centres requesting to be connected to our system; 2) the size of each facility; and 3) the acceleration of each facility's ramp schedule to reach full capacity. For some context, historically, a single data centre typically had a demand for 30 megawatts (MW) or greater. However, we're now receiving individual requests for demand of 60 to 90 MW or greater and it hasn't stopped there. We get regular requests to support larger data centre campuses that include multiple buildings and require total capacity ranging from 300 megawatts to as many as several gigawatts."
- Kidder Morgan, the largest mover of gas in the US, said that if 40% of AI demand is served by natural gas, that would result in incremental demand of 7 to 10 billion cubic feet of gas a day. "Utilities throughout America are sounding alarm, one south-east utility announced its expectation that its winter demand would increase by 37% by 2031.

- NRG Energy, one of the largest utilities in the US, said that "electrification trends compounded by GenAI data centre growth forecast a signal of transformative rises in power demand…We're seeing clear signs of a step change in the long-term fundamentals of power demand for multiple catalysts. This marks the departure from an extended period of stagnant power demand during which energy efficiency outpaced usage growth. For the first time in decades, and perhaps in my 40 years in this business, we are experiencing fundamental improvements driven by demand rather than commodity prices…This increase in demand is attributed to several factors, including electrification, manufacturing, onshoring, LNG, crypto, greater industrial loads and data centre growth. Recent advancements in GenAI are compounding and accelerating these factors, leading to the formation of the next power demand super cycle…To be clear, it takes only a fraction of what is being forecasted to be in this super cycle."
- Willdan Group, a consultancy to utilities and governments, said, "[t]oday, you'll hear us talk about new data centre load driven by AI processing that is adding demand for electricity far faster than most predicted…*In our upfront work, we see that customers are beginning to rapidly prepare for new electric load on the power grid*. According to the Federal Energy Regulatory Commission (FERC) data, over the past year grid planners nearly doubled the five-year load growth forecast. The main drivers are investments in new industrial manufacturing and data centre facilities…Since these forecasts were filed with FERC, Willdan customers like Puget Sound Energy, Duke Energy, Dominion and TVA have stated that their load expectations have grown even higher due to data centres. This indicates that the current FERC load forecast is likely to be an underestimate…[Deal flow is] is already starting to occur. One of our technical people was at a conference just two days ago where utilities were confessing the load growth they are already seeing on their networks. So this has started. I met with two utilities over the last 2 weeks and also had that kind of discussion about increased load growth and how they need to reconfigure their programs, their forecasts, some of their software, all to prepare for this pretty rapidly changing environment. So, it's starting to occur, *and I don't think it was expected even several months ago*."

Along with the rollout of AI technology, we think the need for significant increases in power generation will be one of the major trends of our generation. There are many ways to benefit, from investing in companies that build electrical products to support the grid, to investing in companies that build the grid, to investing in companies that build power generation, to investing in companies that deliver power to the end user. In addition, we believe the amount of copper needed to support this activity will likely be a step-change higher. While the quantum of energy that will ultimately be needed will depend on how AI technology progresses, we continue to monitor this closely.

Before these technologies (AI, semiconductors, autonomous vehicles, search, co-pilots, automation of processes, summarisation, gaming to name some) can reach their full potential, we believe there's a high chance that we'll first run out of cheap power.

Kind Regards,
Fawkes Capital Management

Please click *here* to subscribe to receive future updates.

Fawkes Capital Management Disclaimer

The information contained in this report has been prepared by Fawkes Capital Management Pty Ltd ("Fawkes"). Fawkes is a Corporate Authorised Representative of One Wholesale Fund Services Ltd ("OWFS"), ACN 159 624 585, AFSL 426503. Fawkes offers financial services in Australia only to 'wholesale clients' as defined by the Corporations Act 2001. Fawkes is the investment manager for the Fawkes Capital Fund (the "Fund"). The issuer and trustee of the Fund is One Funds Services Limited ("OFSL"), ACN 615 523 003, AFSL 493421, which is only available to wholesale clients. The information in this article is current as at the date of publication and is subject to change. Fawkes and/or the Fund may hold or intend to hold positions in any of the securities mentioned in this report. Fawkes has no obligation to inform anyone of any changes to its view of, or holdings in any securities mentioned in this report. This information is general in nature. It doesn't take into account a person's objectives, financial situation or needs. Because of that, any persons relying on this information should consider obtaining independent advice before making any investment decisions based on this information. The reader agrees not to invest based on this article, and to perform his or her own due diligence and research before taking a position in any securities mentioned. Information in this article may constitute Fawkes' judgement at the time of publishing and is subject to change. Whilst Fawkes believes this information is correct, no warranty is made as to its' accuracy or reliability. Fawkes doesn't accept responsibility for any loss or liability incurred by you in respect of any error, omission, reliance, or misrepresentation in the information contained in this article. Past performance is not a reliable indicator of future performance. The value of an investment may rise or fall with the changes in the market. Any projection or forward-looking statement in this article is provided for information purposes only. Whilst reasonably formed, no representation is made as to the accuracy of any such projection or that it will be met. Actual events may vary materially. Investors should consider the Fund's Information Memorandum ("IM") dated 24 May 2024 issued by OFSL before making any decision regarding the Fund. The IM contains important information about investing in the Fund and it is important investors obtain and read a copy of the IM before deciding about whether to acquire, continue to hold or dispose of units in the Fund.