ITK demand model

David Leitch

2024-10-14

# ITK’s central demand case and data centres

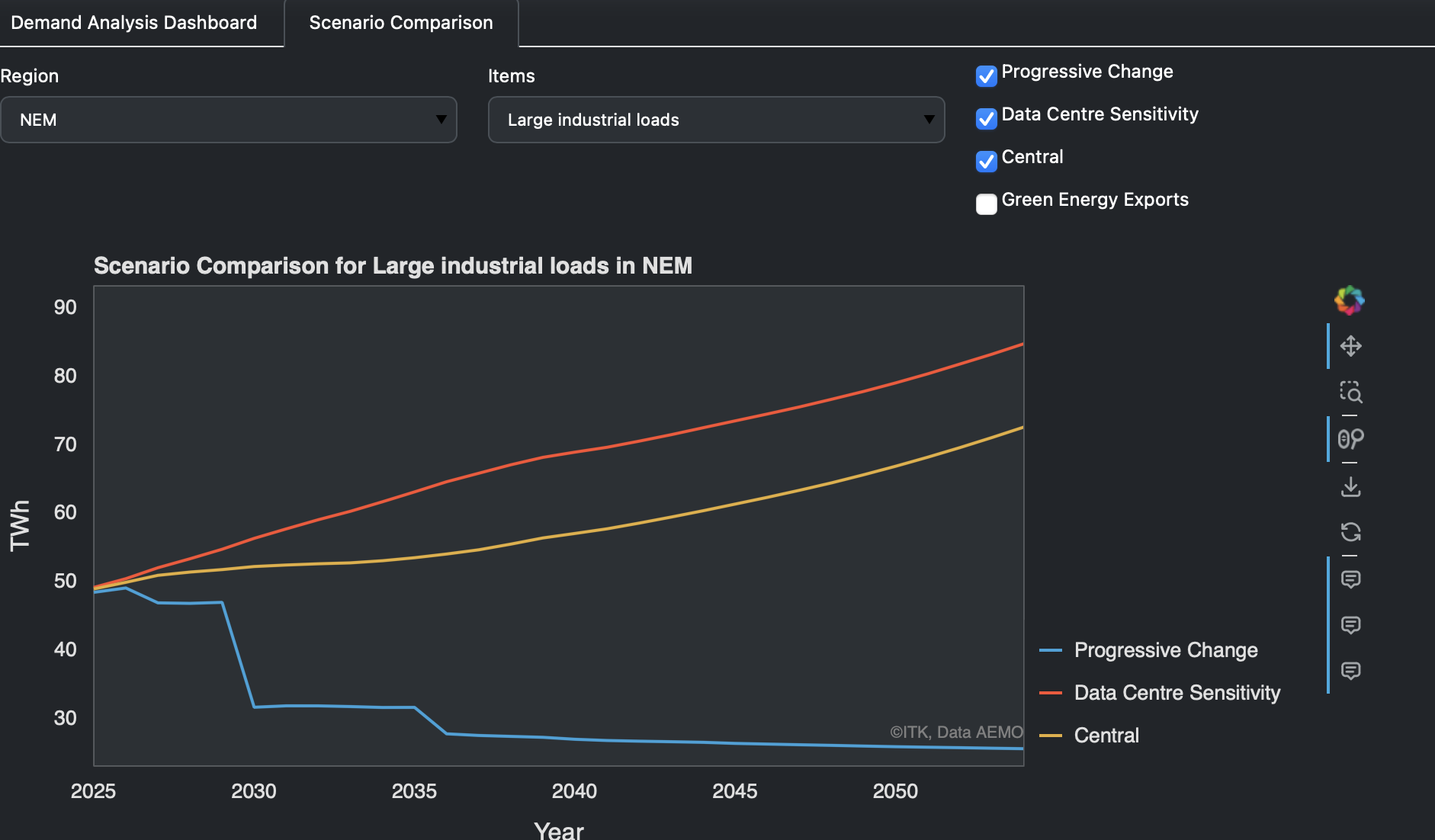
## AEMO’s central case is likely too aggressive

In summary ITK has adjusted AEMO’s ‘Central’ demand case down to allow for much less hydrogen, a bit less electrification and up to include 1/3 of the data centre sensitivity. This results in 20 less TWh of NEM wide demand by 2035 than the AEMO central case. Where the AEMO ‘Central’ case has 271 TWh of underlying demand in 2035 ITK’s forecast has 252 TWh. Out to 2030 the difference is trivial. Still our demand forecasts are not set in stone and we expect to revise our post 2030 numbers many times between now and then.

In thinking about the data centre sensitivity we took the opportunity to look at the commodity nature of the AI infrastructure models, the large investment >A$6 bn to stay on the leading edge and over the past 12 months the more than 50% drop in the price capacity providers can charge customers. A datacenter capable of training ChatGPT 4 in 4 days has a power consumption of only 150 MW. That’s a huge quantity of power going into one building but not material in the context of the NEM supply and demand in the NEM.

## Demand discussion

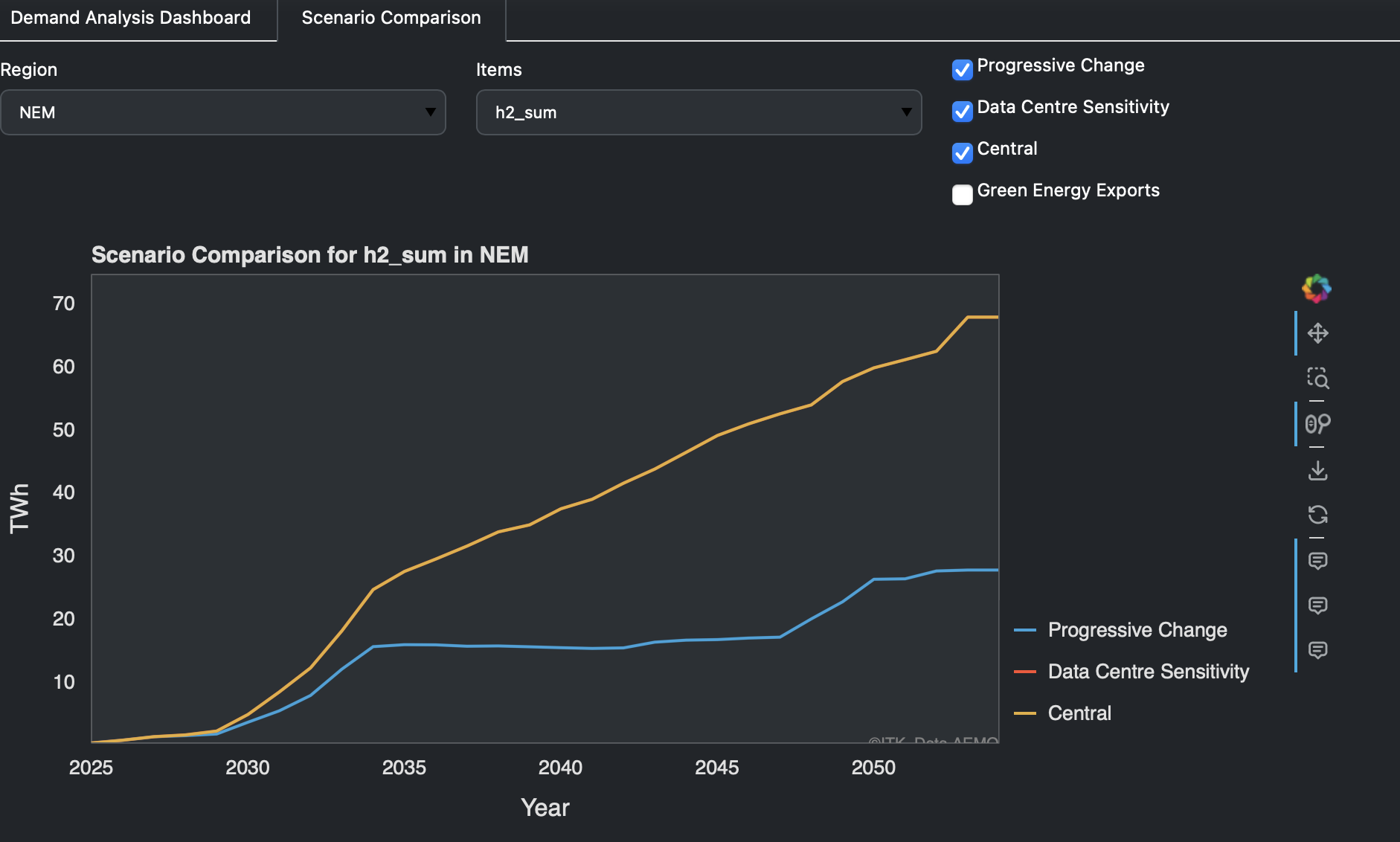
In order to appreciate AEMO’s demand forecasts more comprehensively we built a [demand dashboard](https://itkservices3.com/forecast.html) Sample output below. AEMO provides comprehensive demand but not always in a format that is easy to use. Specifically I have no interest in imaginary demand, eg demand that would occur except for energy efficiency. In any event sample output from the dashboard is shown below. After studying AEMO’s numbers ITK has produced its own, lower demand estimates. Of course this has implications for the new supply required and also the electricity price.



ITK demand dash. Data:AEMO

Looking at the numbers over an extended period of time and bearing in mind the costs of hydrogen and even the costs of process heat decarbonisation but also accepting that decarbonisation via electrification is achievable, affordable and above all essential, we have decided to push some of the electrification demand and hydrogen demand beyond our reasonable forecast horizon of 2035.

### Hydrogen

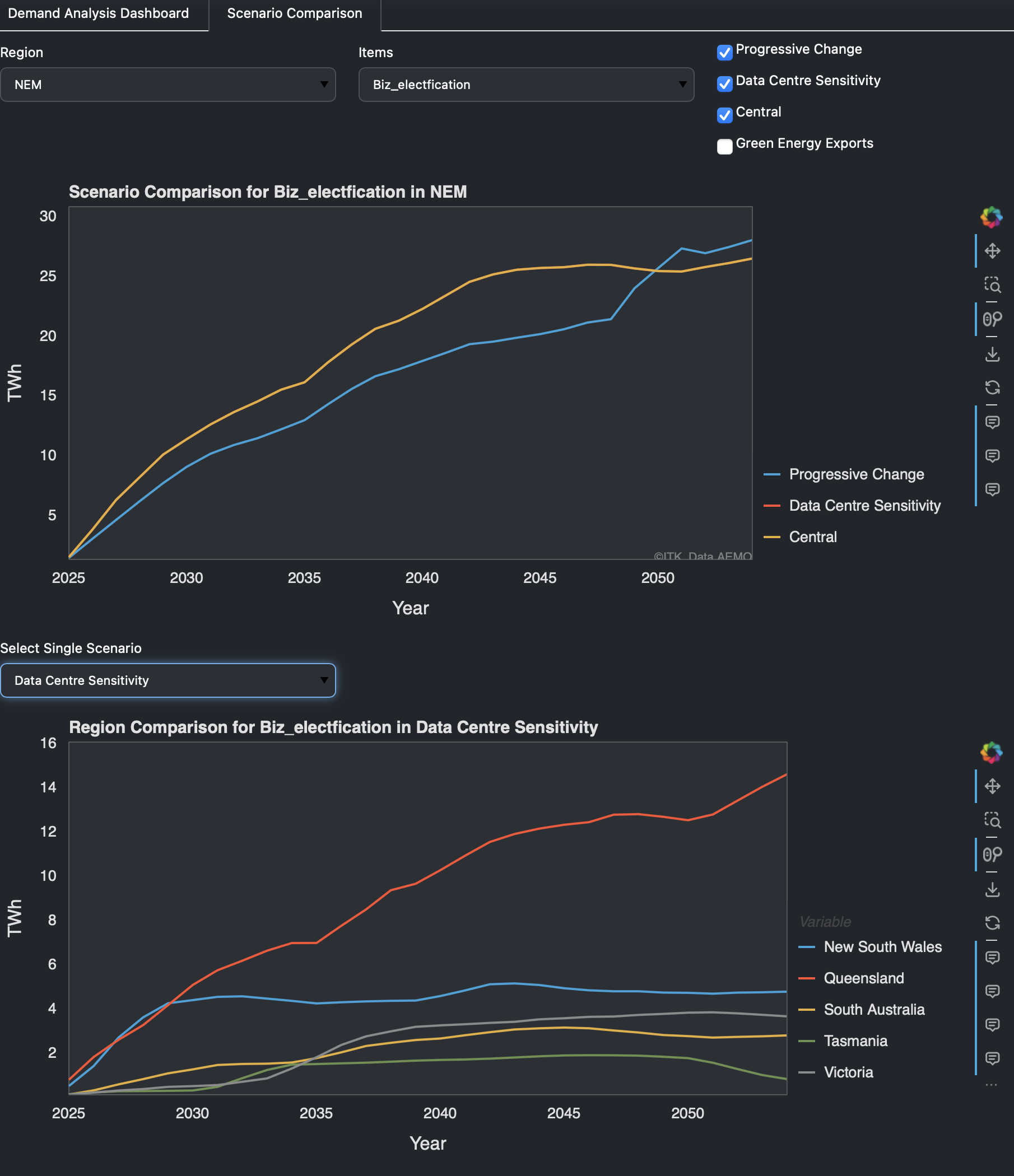
The electricity demand from Hydrogen production from the two main AEMO scenarios as follows: 

In our opinion even the hydrogen in the Progressive Change scenario cannot be forecast with confidence. Nevertheless in ITK’s base demand forecast we have adopted the progressive case for hydrogen demand. Further we have reduced even the Progressive Change case by 50%

### Business and residential electrification.

Although not the only way that business will decarbonise, the main thing is the replacement of gas used in process heat with electricity, as discussed in our prior note.

The way AEMO has forecast, the majority of this electrification happens in QLD.



Electrification. Data:AEMO

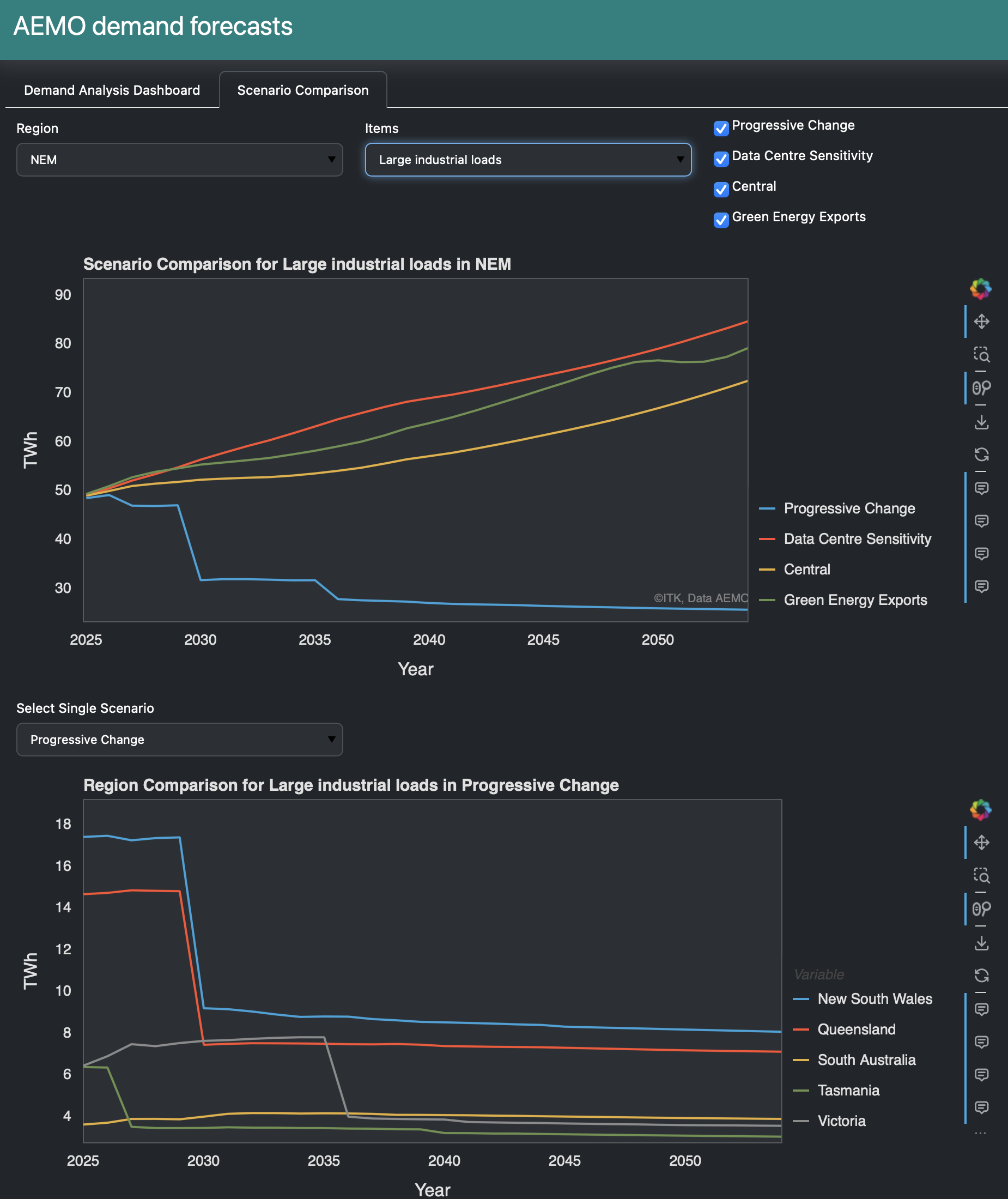
And most likely represents, simply put, the decarbonisation of Gladstone both at the alumina refinery, possibly the cement and probably the explosive business. Obviously what AEMO forecasts in detail is largely know only to them.

Broadly speaking by 2035 Business electrification is adding between 15 TWh (Central case) and 9 TWh(Progressive case). 15 TWh takes around 6 GW of renewable energy and 9 TWh is between 3 and 4 GW. As previously remarked the cost of business electrification is high as for process heat by and large the electricity cost needs to come in sub $50/MWh. However in ITK’s opinion this is plausible for interruptible processes, or at least those that don’t require firming.

As such the sensitivity of price to changing the business electrification case from ‘Central’ to ‘Progressive’ is likely to be low nevertheless we have adopted the Progressive Change case

### Large industrial loads

It’s worth noting that in the “Progressive Change” scenario large industrial loads decline precipitously from about 2030. Presumably this represents aluminium smelters closing down in QLD and NSW.



Data centre sensitivity. Data:AEMO

Here at ITK we have always worried about aluminium smelters closing as their subsidised electricity prices run out. Equally we have always believed that an unsubsidised firmed renewables powered smelter in Australia could outcompete an unsubsidised coal fuelled smelter in China.

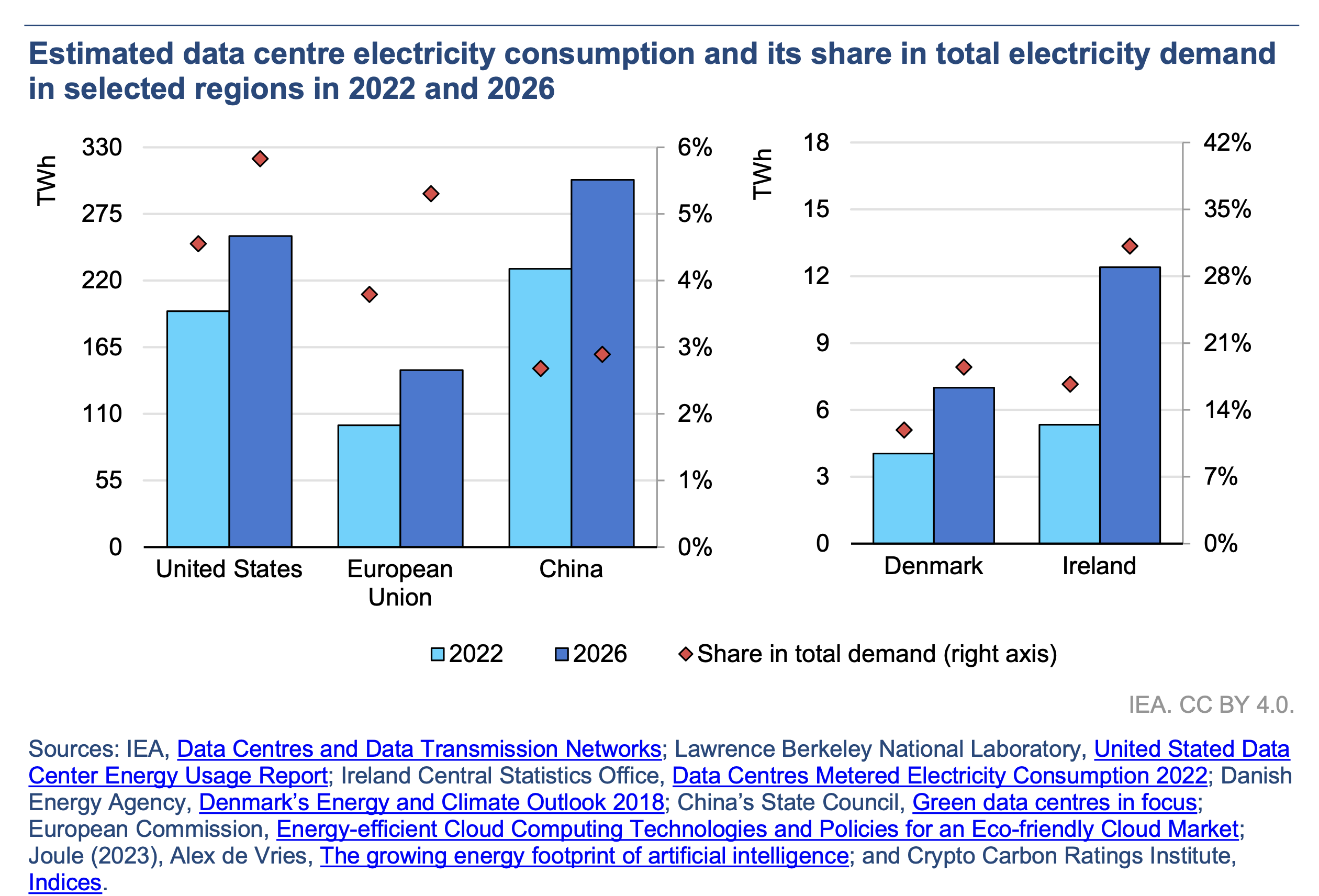
Its likely I’ll never know whether that assumption is true or false because for the most part aluminium smelters will continue to be subsidised and China, perhaps a few years later, will start to power its heavily energy intensive economy with renewables. So pre DataCentres I prefer the assumptions in the Central case for Large Industrial Loads.

### Data centre loads

On the more positive side in 2025 it looks like that Data Centre loads will indeed go beyond the 300 MW identified by AEMO based on metered data. AEMO models its data centre case to add about 8 TWh or roughly 3 GW of new wind and solar and perhaps very roughly 1 GW of firming to demand. Still that seems quite an aggressive sensitivity and so we have used 1/3 of the Data Centre sensitivity.

Note the following section draws heavily on work published by Dylan Patel, Daniel Nishiball and Jeremie Ontivereros at Semianalysis. In turn they reference an IEA study and some other academic pieces.

If we look at the IEA report, some small countries are projected to have up to 20% of electricity demand from data centres.



Data centre demand. Source: IEA 2024

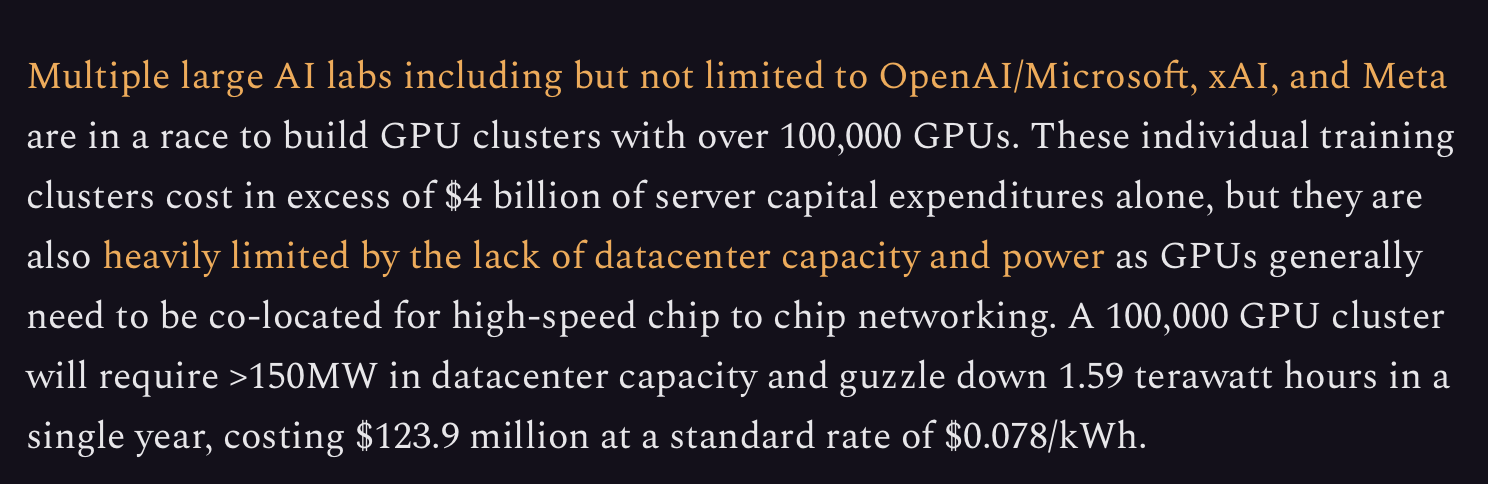
When comparing the average electricity demand of a typical Google search (0.3 Wh of electricity) to OpenAI’s ChatGPT (2.9 Wh per request), and considering 9 billion searches daily, this would require almost 10 TWh of additional electricity in a year.

AI electricity demand can be forecast more comprehensively based on the amount of AI servers that are estimated to be sold in the future and their rated power. The AI server market is currently dominated by tech firm NVIDIA, with an estimated 95% market share. In 2023, NVIDIA shipped 100 000 units that consume an average of 7.3 TWh of electricity annually. By 2026, the AI industry is expected to have grown exponentially to consume at least ten times its demand in 2023. (source: IEA 2024)

|  |
| --- |
| Figure 1: Global datacentre demand. Source:IEA 2024 |

As can be seen from [Figure 1](#fig-datacentre_demand) dedicated AI centres are the smallest piece of the data centre pie. The IEA numbers and projections are useful but at least for a dedicated AI centre “semianalysis” built the numbers from the ground up.

Its well worth reading this [article](https://www.semianalysis.com/p/100000-h100-clusters-power-network) from semi analysis which asserts that (1) there are four or five organisations trying to build next gen AI training clusters consisting of say 100,000 Nvidia H100 chips each of which costs say US$40,000 with an allup cost of say $US 4 billion. In my opinion these chips are actually remarkably power efficient and the total power to run this $US 4 billion site is just 150 MW. These calculations are based on each H100 chip consuming 400 W, the associated CPU 200 W and 600 W of cooling ie about 1.2 kw per chip and for 100,000 chips that’s 120 MW plus we add in some extras.



AI datacenter power. Source:semianalysis

The article estimates that such a system can train GPT4 style model in four days. There are lots of points to make about these AI clusters but two stand out. First the 150 MW is preferably available to a single building but realistically the AI centre has to cover a ‘campus’. Second the market is moving sharply into oversupply so that it’s likely going to be difficult to earn return on capital. Costs of renting H100 capacity for some job have halved over the past 12 months.

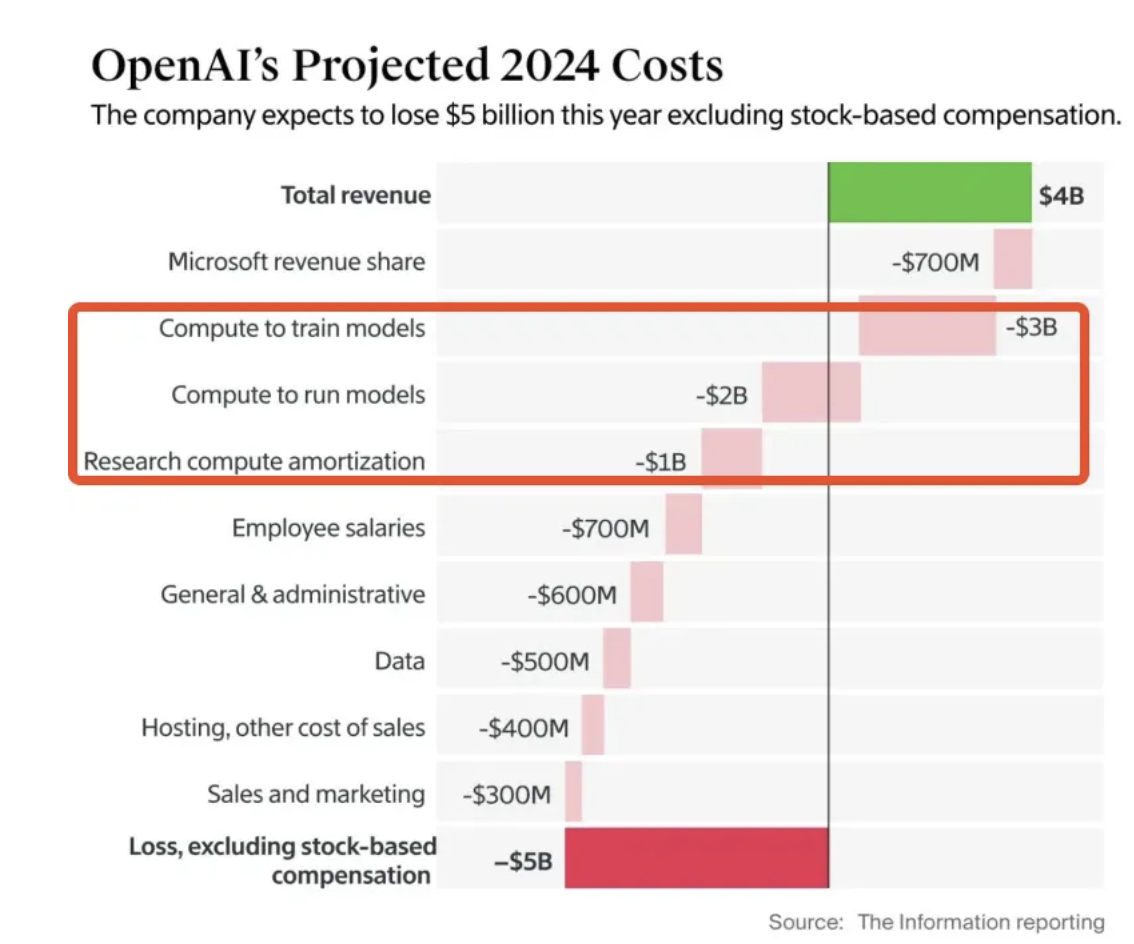
ChatGPT was launched in November 2022, built on the A100 series. The H100s arrived in March 2023. **The pitch to investors and founders was simple:** Compared to A100s, **the new H100s were 3x more powerful, but only 2x the sticker price**.

Market prices shot through the roof, the original rental rates of H100 started at approximately ***$4.70 an hour*** but were going for ***over $8***

In Aug 2024, if you’re willing to auction for a small slice of H100 time (days to weeks), you can start finding H100 GPUs for $1 to $2 an hour. (Source:Eugene Cheah)

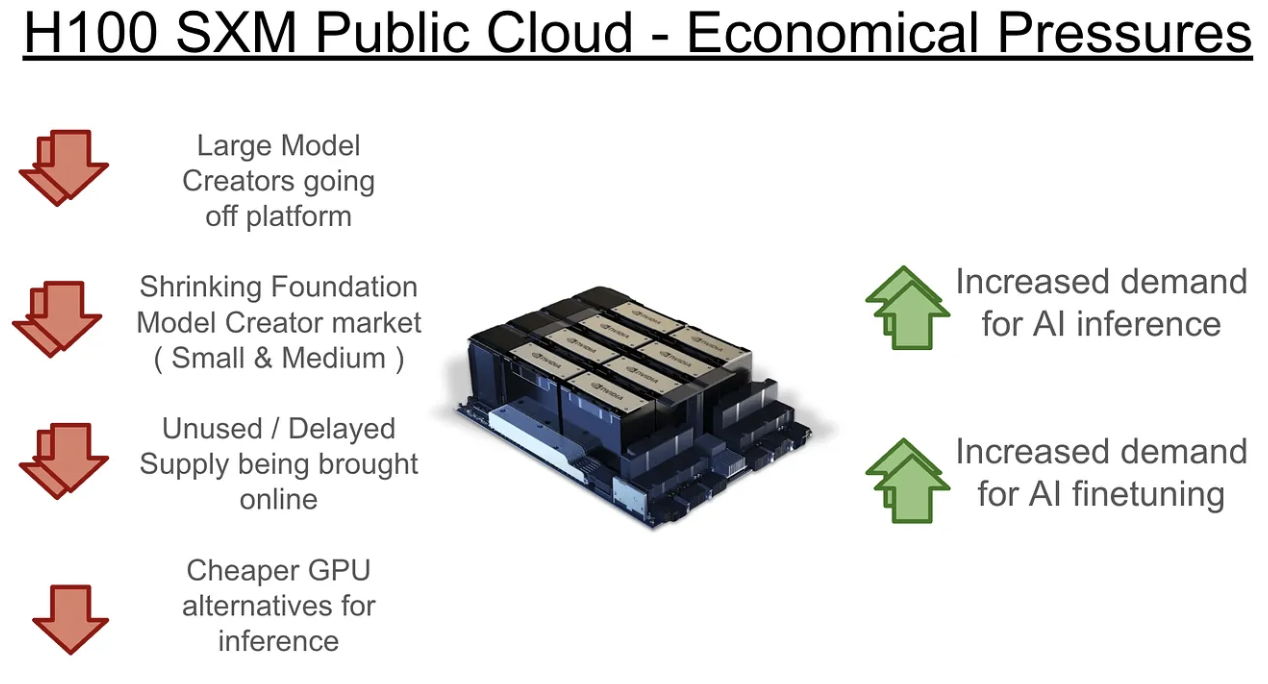
This is a very familiar story to any commodity follower, but it’s no less powerful for all that. And like most commodity markets suppliers will try and manage things by forming an oligopoly, eg iron ore.

Despite what the IEA found if we look at OpenAI (ChatGPT) 2024 disclosed budget it shows that training cost is greater than the running and answering query cost, and this would go to electricity consumption.



Source: Eugene Cheah

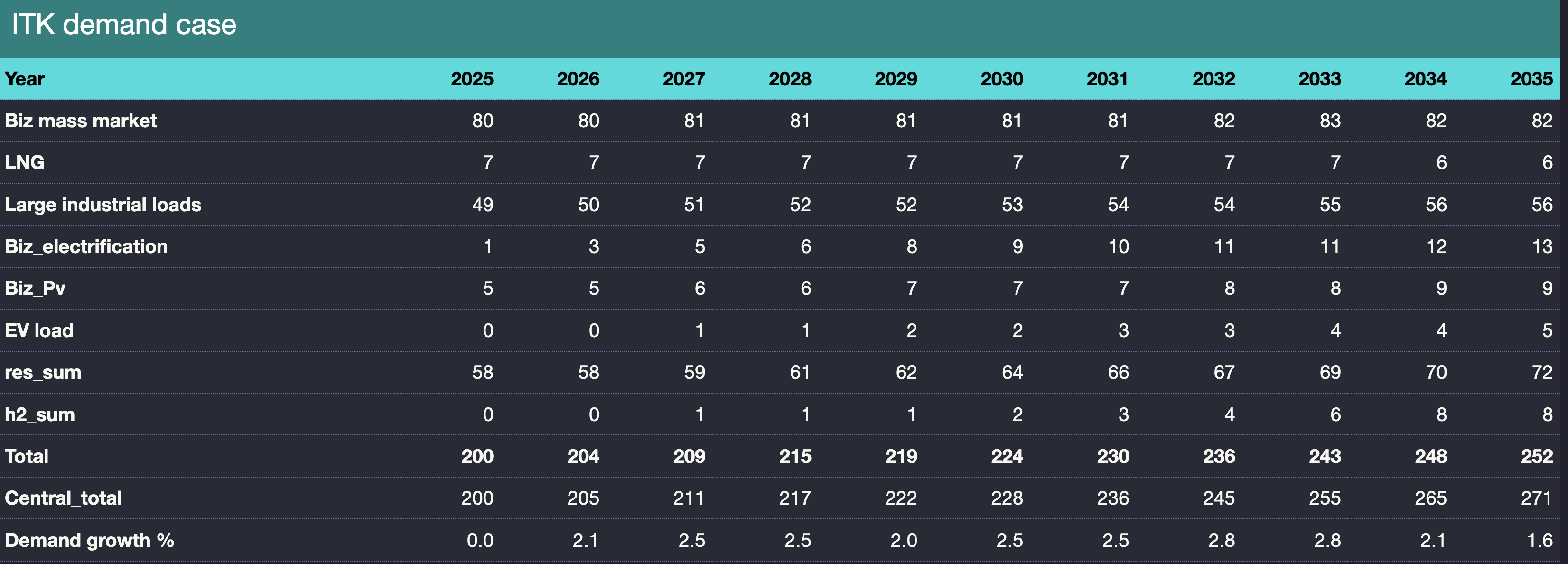
And this leads to the familiar commodity drivers that we see in many industries from iron ore to airline travel to lithium.



Source: Eugene Cheah

The point for ITK demand forecasting is that the cost of using data centres is falling towards cost as supply increases and that does increase demand, but that falling returns on capital will make investors cautious. Secondly as things stand AI training and usage is only a moderate piece of datacentre electricity consumption and for Australia I don’t see it as a major driver of new demand at least for a few years.

### ITK has about about 20 TWh less in 2035 than AEMO’s Central case

Following those adjustments the ITK central demand case at a NEM levels out to 2035 is: 

Of course it’s possible to extend the numbers out to the 2050s. However although we fully appreciate the investment model need to look at 20 and 30 year forecasts, ITK has focussed heavily on the next 10 years for two reasons.

1. The net present value of a project is of course driven more heavily by returns in the earlier years, even at a 10% discount rate the first 10 years are 68% of the value of a 25 year annuity.
2. Events that occur in the more distant future have uncertainly that make the forecasts essentially a guess and the optical illusion of growth that is forecast in the distant future tends to reduce focus on the nearer term.