4E Efficient, Demand Flexible Networked Appliances EDNA

Total Energy Model 4.0 - Data Centres

MAY 2025

Technology Collaboration Programme
by iea

Energy Efficient End-use Equipment
International Energy Agency

The Technology Collaboration Programme on Energy Efficient End-Use Equipment (4E TCP), has been supporting governments to co-ordinate effective energy efficiency policies since 2008.

Fourteen countries and one region have joined together under the 4E TCP platform to exchange technical and policy information focused on increasing the production and trade in efficient end-use equipment. However, the 4E TCP is more than a forum for sharing information: it pools resources and expertise on a wide a range of projects designed to meet the policy needs of participating governments. Members of 4E find this an efficient use of scarce funds, which results in outcomes that are far more comprehensive and authoritative than can be achieved by individual jurisdictions.

The 4E TCP is established under the auspices of the International Energy Agency (IEA) as a functionally and legally autonomous body.

Current members of 4E TCP are: Australia, Austria, Canada, China, Denmark, the European Commission, France, Japan, Korea, Netherlands, New Zealand, Switzerland, Sweden, UK and USA.

Further information on the 4E TCP is available from: **www.iea-4e.org**

Electronic Devices & Networks Annex
EDNA

The Efficient, Demand Flexible Networked Appliances Platform of 4E (EDNA) provides analysis and policy guidance to members and other governments aimed at improving the energy efficiency and demand flexibility of connected devices and networks.

Further information on EDNA is available from: **www.iea-4e.org/edna**

This report was commissioned by the EDNA Platform of the 4E TCP and authored by Anson Wu and Fiona Brocklehurst. The views, conclusions and recommendations are solely those of the authors and do not state or reflect those of EDNA, the 4E TCP or its member countries.

Views, findings and publications of EDNA and the 4E TCP do not necessarily represent the views or policies of the IEA Secretariat or its individual member countries.

31 May 2025

Total Energy Model 4.0 - Data Centres

Final report

Anson Wu and Fiona Brocklehurst
HANSHENG LTD AND BALLARAT CONSULTING

Table of Contents

Table of Figures

Executive Summary

This report describes the Total Energy Model (TEM) 4.0, a novel bottom-up hybrid model designed to forecast data centre energy consumption and evaluate the effect of possible policy interventions on energy use. Moving away from traditional methods that rely on extrapolating equipment sales or data traffic, the TEM focuses on computational capacity and efficiency metrics derived primarily from CPU sales data and standardized Server Efficiency Rating Tool (SERT) performance measurements. It segments the market by data centre type (traditional, cloud, hyperscale, AI) and workload (general compute, AI) across five global regions.

Policy Modelling Capabilities

A key strength of the TEM is its structure, which is designed to support the analysis of various policy interventions. Policymakers can use the model, including a simplified dashboard, to assess the impacts of:

- Minimum efficiency standards or labels for ICT equipment.
- Server and infrastructure utilisation .
- Infrastructure efficiency improvements (e.g., PUE targets).
- Accelerated data centre consolidation (shifting workloads from traditional to cloud/hyperscale).
- Changes to server lifetimes to analyse lifecycle impacts.

This functionality allows evaluation of potential energy savings, feeding into cost-benefit analyses of different policy scenarios.

Potential for Improved Energy Consumption Modelling with Better Evidence

While the TEM offers a sophisticated framework, the report highlights significant uncertainties and areas where improved data could enhance modelling accuracy, particularly for energy consumption projections:

- **AI Sector Uncertainty**: Forecasting AI energy consumption is extremely challenging due to the sector's immaturity, rapid technological evolution, limited data availability (e.g., sales data, performance benchmarks beyond MLPerf), and the difficulty in predicting future efficiency gains. Current extrapolations produce highly uncertain, unrealistic results if practical constraints (chip supply, power availability, capital) aren't considered (which require economic modelling, beyond the scope of TEM).
- **Data Segmentation**: Currently the evidence base for allocating computational capacity accurately across different data centre segments (traditional, cloud,

hyperscale) is weak, affecting the accuracy of energy estimates due to varying utilisation rates for each segment.

- **Performance Benchmark Limitations**: For non-AI servers SERT provides standardized CPU performance data, but its representativeness of real-world configurations is not established – it needs validation. For AI, performance data from MLPerf data is limited, evolves rapidly, lacks comprehensive power data for all workloads (e.g., training), and its correlation with theoretical performance (like FLOPS) needs further analysis.
- **PUE Data**: Power Usage Effectiveness (PUE) data, though the best available metric for infrastructure efficiency, suffers from limitations and potential reporting bias towards more efficient centres.
- **Storage & Networking**: These components are currently modelled as a simple percentage of server energy; a more detailed, data-driven approach could improve accuracy.
- **Server Lifetimes**: Assumptions about server lifetimes, especially for rapidly evolving AI hardware, significantly impact installed base and energy estimates. Data on actual lifetimes would improve accuracy.

Conclusion on Model Utility

The TEM 4.0 provides a valuable framework for policy analysis, particularly for established non-AI server technologies. However, due to significant uncertainties, especially regarding AI, policymakers must use the model cautiously, understanding the limitations of the model and the current evidence base. The model is structured for relatively easy updates as better data becomes available (e.g., from reporting mandates or new research), allowing for improved projections and policy assessments over time. Integrating insights from financial/economic models that incorporate market constraints could further enhance its utility, though this would require careful verification.

Glossary

AI	Artificial Intelligence
CPU	Central Processing Unit
EDNA	IEA's Energy Efficient End-Use Equipment Technology Collaboration Programme: Efficient, Demand Flexible Networked Appliances
GPU	Graphics Processing Unit
LLM	Large Language Models
ML	Machine Learning
PUE	Power Usage Effectiveness
SEED	Server Energy Efficiency Database
SERT	Server Efficiency Rating Tool
TCO	Total Cost of Ownership
TGG	The Green Grid

1 Introduction

Data centre energy consumption continues to grow rapidly, driven by increased digitalization, cloud computing adoption, and the emergence of AI workloads. Current estimates suggest data centres account for approximately 1-2% of global electricity consumption, with projections indicating significant further growth through 2040. This growth presents challenges for electricity grid capacity, renewable energy adoption, and climate goals, making accurate forecasting and policy evaluation critical.

Previous modelling approaches have typically relied on extrapolating current trends in equipment sales or data traffic. However, these methods have led to widely divergent projections, with estimated 2030 data centre energy consumption varying by more than an order of magnitude across different studies. These disparities arise from several limitations in traditional approaches: dependency on short-term sales data that becomes increasingly unreliable for long-term forecasting, difficulty accounting for technological transitions like the rise of AI accelerators, and limited ability to model the impacts of policy interventions.

The rapid evolution of computing technology further complicates forecasting efforts. While the physical form factor of servers has remained relatively stable, internal components continue to advance rapidly, with significant changes in CPU architecture, memory technologies, and accelerator integration. Traditional models that rely on server unit sales struggle to account for these technological transitions, particularly when projecting more than 5-10 years into the future.

This report presents a novel bottom-up hybrid model that addresses these challenges by shifting focus from equipment sales to computational capacity and efficiency metrics. The model's foundation rests on three key principles:

First, it uses CPU sales data combined with standardized performance measurements from the Server Efficiency Rating Tool (SERT) as primary inputs. This provides accurate near-term data while enabling the model to track both performance improvements and energy efficiency trends. The standardized nature of SERT measurements ensures consistency across different hardware generations and manufacturers.

Second, the model converts these inputs into metrics of total computational capacity and system-level efficiency. This abstraction away from specific hardware implementations allows for more reliable long-term forecasting, as it focuses on fundamental computational demands rather than particular equipment form factors or architectures.

Third, the model incorporates detailed segmentation of the data centre market, tracking three distinct categories of general compute (traditional, cloud, and hyperscale) and separately modelling AI workloads. This granularity enables analysis of how different sectors evolve and allows policymakers to evaluate targeted interventions for specific market segments.

The model's structure specifically supports the analysis of policy interventions through several mechanisms:

- Evaluation of minimum efficiency standards for servers and infrastructure
- Modelling of utilisation improvements through workload consolidation
- Assessment of data centre consolidation scenarios
- Analysis of accelerated retirement programmes for legacy equipment
- Investigation of policies to improve infrastructure efficiency

Infrastructure energy consumption is modelled using Power Usage Effectiveness (PUE), with different trajectories for each data centre type based on technological capabilities and economic factors. While PUE has known limitations as a metric, it remains the most widely available measure of infrastructure efficiency and provides sufficient accuracy for long-term modelling purposes.

Our projections indicate that energy consumption will continue to grow. These results reflect several important market dynamics, including the rapid growth of AI workloads and continued migration to cloud environments. The model's results have been validated against historical data where available and compared with previous studies to provide a reality check.

However, modelling the growth in AI at this point in time is a particular challenge. There may be bottlenecks in chip and power supply, future efficiency gains are hard to model, and data availability is low. This means that there is high uncertainty in energy projections in this sector which becomes more extreme the further into the future the projections go.

The remainder of this report is structured as follows: Section 2 presents the detailed methodology, including data sources and key assumptions. Section 3 describes the policy modelling options. Section 4 describes the baseline scenario results and projections through 2040. Section 5 compares the results with recent projections by other authors. Section 6 presents sensitivity analyses and discusses key uncertainties. Finally, Section 7 offers conclusions and recommendations for policymakers.

2 Methodology

The model employs a bottom-up hybrid approach that combines CPU sales data with standardized performance metrics to estimate current energy consumption and project future trends. This section describes the model's structure, key calculations, and data sources.

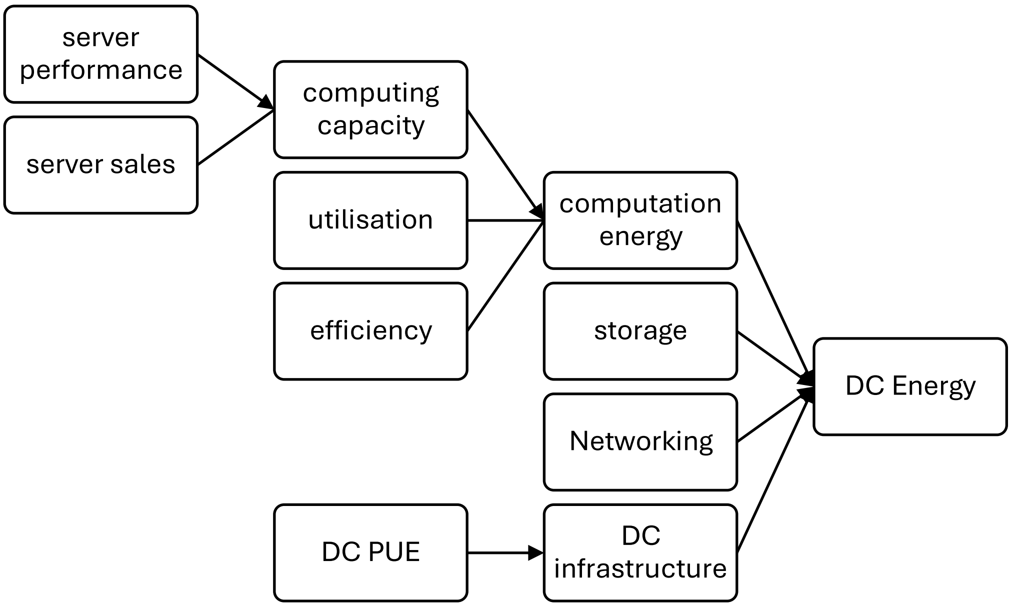
2.1 Model Overview

The fundamental calculation flow consists of four main stages:

1. Converting CPU sales and performance data into total market computational capacity
2. Determining computations and energy consumption based on utilisation and efficiency
3. Adding storage and networking energy based on a proportion of computational energy
4. Adding infrastructure energy based on PUE

A schematic of the model is shown in Figure 1.

Figure 1 Schematic of model structure



2.2 Market Segmentation and workload types

The model segments data centres into four categories:

1. Traditional (small sized, less efficient)
2. Cloud (large-medium scale, efficient)
3. Cloud (hyperscale very efficient)
4. AI (hyperscale, very efficient)

Each segment has distinct characteristics for:

- Computing capacity
- Computing efficiency
- Server utilisation

- PUE values

This segmentation is the weakest part of the model as there is no good evidence to determine how much of the computing capacity is found in each data centre segment. Because server utilisation varies by segment, this can affect the energy consumption estimates. The servers are therefore allocated by the total number of cores, with lower core counts in small DCs and highest in the hyperscale DCs. This results in different efficiency characteristics for each segment.

The original proposal was to also estimate the energy of edge data centres but the definition was too broad to determine the market size accurately. When comparing reports about edge data centres and computing, they ranged from 1+MW data centres to mobile phones and included servers located at mobile phone masts and driving-assisted cars but were rarely clear about what exactly was in scope or give a breakdown of devices. Given the difficulty in estimating computing capacity of the existing segments, it is excluded as a separate segment. However, the energy consumption should still be captured in the sales of servers and GPUs allocated to these segments.

2.2.1 Workload Types

Computational workloads are categorized as:

1. General compute
2. AI (artificial Intelligence)/ML(Machine Learning) workloads

AI covers workloads that require an accelerator card, usually a GPU, to perform the calculations at scale with sufficient performance and efficiency. General computing includes all other workloads, ie servers without a GPU.

We acknowledge that some non-AI computation (eg high performance scientific calculations) will use GPU and not all AI computation will be performed on GPUs, but we consider that this is a reasonable approximation.

2.2.2 Regional Considerations

While the model is global in scope, five regions are defined:

- North America
- South America
- Europe
- Asia (including Oceania)
- Middle East and Africa (MEA)

The model incorporates regional variations in:

- Computing capacity
- Computing efficiency

In total 40 subtypes are created for the base scenario as shown in Table 1.

Table 1 Subtypes used in model

Scenario	**Region**	**Computation**	**DC type**	**Product**
Base	N. America	Traditional	Small/medium	ICT
Base	S. America	Traditional	Small/medium	ICT
Base	Europe	Traditional	Small/medium	ICT
Base	Asia	Traditional	Small/medium	ICT
Base	MEA	Traditional	Small/medium	ICT
Base	N. America	Cloud	Large	ICT
Base	S. America	Cloud	Large	ICT
Base	Europe	Cloud	Large	ICT
Base	Asia	Cloud	Large	ICT
Base	MEA	Cloud	Large	ICT
Base	N. America	Cloud	Hyperscale	ICT
Base	S. America	Cloud	Hyperscale	ICT
Base	Europe	Cloud	Hyperscale	ICT
Base	Asia	Cloud	Hyperscale	ICT
Base	MEA	Cloud	Hyperscale	ICT
Base	N. America	AI	Hyperscale	ICT
Base	S. America	AI	Hyperscale	ICT
Base	Europe	AI	Hyperscale	ICT
Base	Asia	AI	Hyperscale	ICT
Base	MEA	AI	Hyperscale	ICT
Base	N. America	Traditional	Small/medium	DC Infrastructure
Base	S. America	Traditional	Small/medium	DC Infrastructure
Base	Europe	Traditional	Small/medium	DC Infrastructure
Base	Asia	Traditional	Small/medium	DC Infrastructure
Base	MEA	Traditional	Small/medium	DC Infrastructure
Base	N. America	Cloud	Large	DC Infrastructure
Base	S. America	Cloud	Large	DC Infrastructure
Base	Europe	Cloud	Large	DC Infrastructure
Base	Asia	Cloud	Large	DC Infrastructure
Base	MEA	Cloud	Large	DC Infrastructure
Base	N. America	Cloud	Hyperscale	DC Infrastructure
Base	S. America	Cloud	Hyperscale	DC Infrastructure
Base	Europe	Cloud	Hyperscale	DC Infrastructure
Base	Asia	Cloud	Hyperscale	DC Infrastructure
Base	MEA	Cloud	Hyperscale	DC Infrastructure
Base	N. America	AI	Hyperscale	DC Infrastructure
Base	S. America	AI	Hyperscale	DC Infrastructure

Scenario	Region	Computation	DC type	Product
Base	Europe	AI	Hyperscale	DC Infrastructure
Base	Asia	AI	Hyperscale	DC Infrastructure
Base	MEA	AI	Hyperscale	DC Infrastructure

2.2.3 Energy modes

The energy consumption for each subtype is broken down into three modes:

- High utilisation period
- Low utilisation period
- Other ICT (network and storage energy) which covers both high and low utilisation

This breakdown covers both the ICT energy use and the DC infrastructure. This allows the PUE to be changed between high and low utilisation, although for this report, it is kept the same.

2.3 Computational Capacity

The computational capacity is a function of the performance of the product and the sales of the product. The product definition can vary as long as the performance is based on the same unit. For each DC segment i in year t:

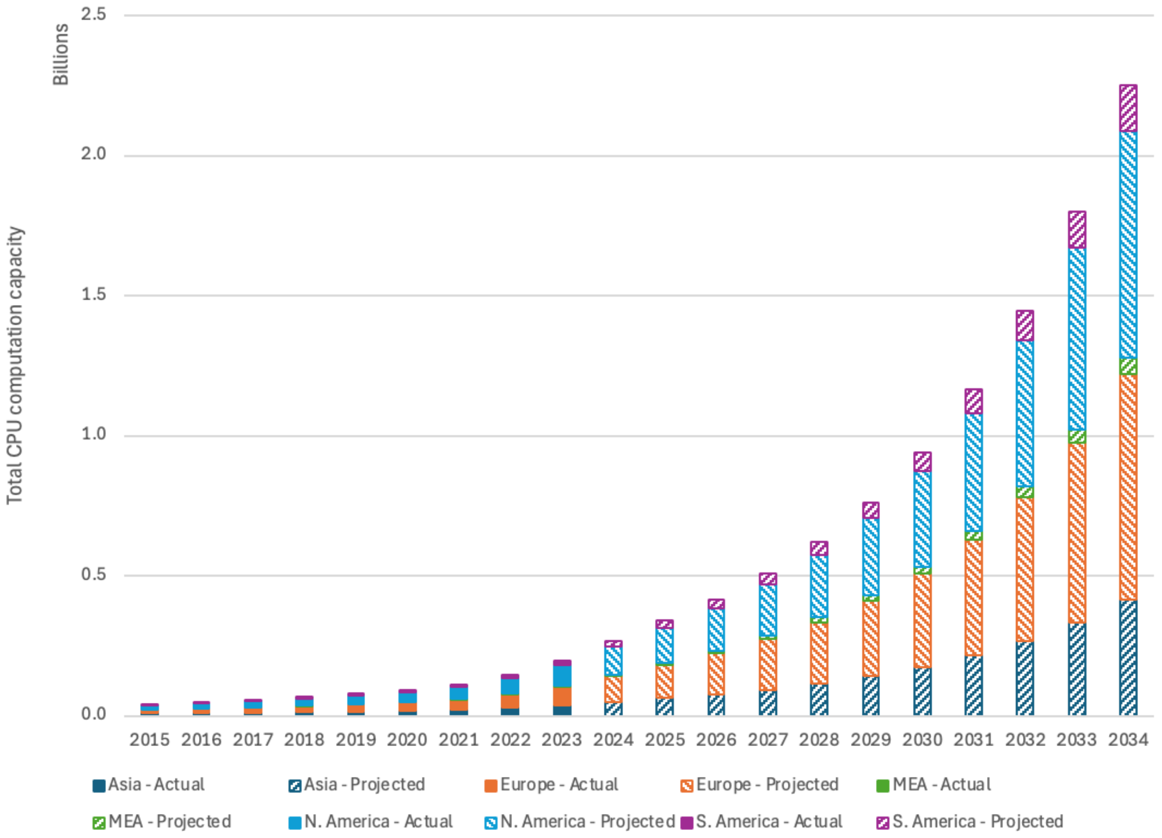
Computational_Capacity(i,t) = Sales(i,t) × Performance_Score(i,t)

2.4 Non-AI Computational Capacity Calculation

The model calculates total computational capacity using server sales data, broken down by number of cores per CPU, combined with SERT performance metrics. The number of cores is considered the primary factor determining the performance of the server. The performance per core has improved over time and the number of cores per server has also increased.

Where Performance_Score is derived from SERT worklet results for average number of cores per server. Figure 2 shows the projected global CPU capacity.

Figure 2 Modelled global CPU (that is, non AI) computational capacity



The server sales data are based on a report by FMI[1] report and adjusted using data from Statista[2,3,4]. This shows a sharp rise in sales over the short term, while previously it has been relatively flat. Short term forecasts show that sales are expected to remain at a higher level. We extrapolated future sales based on current trends.

CPU sales by cores

The number of CPU sales is from FMI (2024) and adjusted based on more detailed investor analysis data (Bruzzone, 2024)[5] which more closely matches our

[1] Future Market Insights (2024) *Global Data Center CPU market*

[2] IHS Markit; Statista estimates (2019) *Global server unit shipments 2016-2022, by technology* Statista Accessed January 2025 https://www.statista.com/statistics/934508/server-unit-shipments-by-technology-worldwide/

[3] MIC (2023) *Global server shipments 2018-2027* Statista Accessed Jan 2025 https://www.statista.com/statistics/1417940/global-server-shipments/

[4] IDC; TrendForce (2021) *Server unit shipments worldwide 2011-2021, by quarter* Statista Accessed Jan 2025 https://www.statista.com/statistics/287005/global-server-shipments/

[5] Bruzzone, M. (2024) *CPU Today*

understanding and broader discussions with industry. This is used to estimate the average number of cores per server based on segment and region.

The lifetime of non-AI servers is assumed to be 5.7 years, based on Shehabi et al 2024 [6]

2.4.1 SERT performance

SERT test data was obtained from The Green Grid (TGG) server energy efficiency database[7] (SEED). On a logarithmic scale performance scales linearly with the number of cores and over time as shown in Figure 3 .

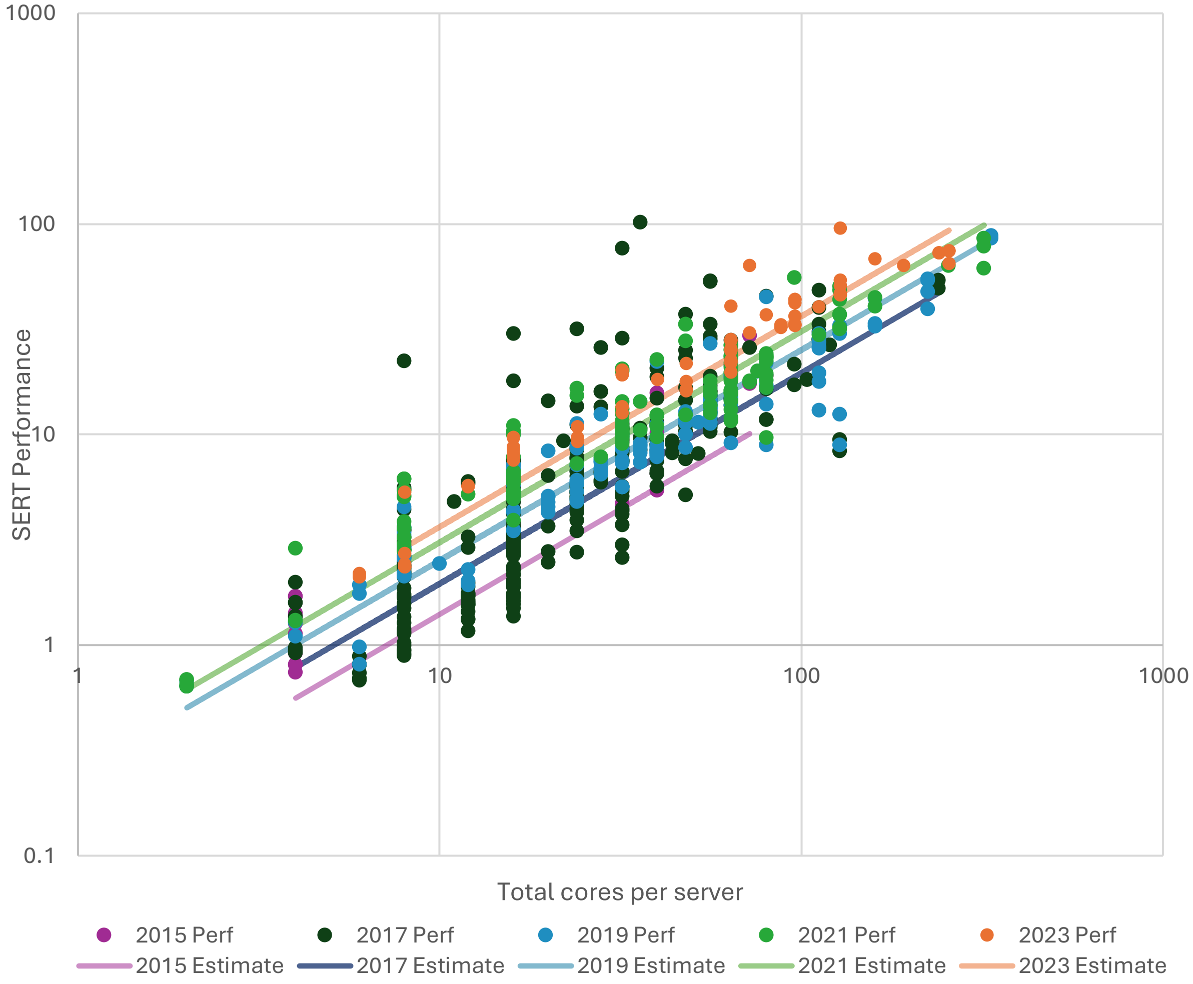


Figure 3 SERT performance trends by number of CPU cores and year of CPU launch

[6] Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D. 2024. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001637 -

[7] https://www.thegreengrid.org/en/resources/library-and-tools/581-TGG%27s-Server-Energy-Efficiency-Database. Accessed November 2024

2.5 AI computational capacity

AI computational capacity is calculated from the number of AI servers sold and performance using the MLPerf benchmark score explained in the next section.

AI sales data is very limited. Two sources were used: public shipment estimates from Techinsights[8] for GPUs and FMI estimates of sales of AI servers (not GPUs). The Techinsights data provides global shipments and importantly includes Google GPUs. These GPUs are made according to Google design and exclusively for Google use and shipped in high volume. The FMI data is used to estimate the regional breakdown for AI computational capacity.

GPU shipments were estimated to have increased from 0.55 million in 2019 to 6.3 million in 2023 with a very sharp increase from 2020. We extrapolated future sales based on current trends. This results in the total performance of GPUs (normalised against the Nvidia H100 SXM 80GB GPU) to increase rapidly (as shown in Figure 4). This shows N. America dominates the market, however, this might be underestimating the expected growth in Asia, particularly China which Chinese DIIRI (2023)[9] forecasts will have (31%) of global computational capacity in 2025, close to USA (36%). Conversely, Patel et al (2024)[10] , estimates that 70% of all AI computational capacity will be located in USA alone. Both reports predict much lower capacity in Europe in part due to high electricity costs and regulatory barriers. Geopolitics also restrict global access to current state of the art GPUs as reported by Reuters (2025)[11], but this could eventually be overcome by domestic production in China.

The lifetime of servers with GPUs is assumed to be four years (Patel & Nishball, 2023)[12], shorter than for 'conventional' servers. It is assumed that operators will plan to minimise the TCO[13] of the server and maximise revenue which increases the rate of server replacement because energy consumption is high, and efficiency is improving

[8] Balossier, E (2024) *Google is the third-largest designer of datacenter processors as of 2023—without selling a single chip*, Techinsights Accessed: August 2024 https://library.techinsights.com/search/wp-asset/1955818#code=DCC-2405-806&subscriptionId=null&channelId=null

[9] 中国通服数字基建产业研究院 (2023) *中国数据中心产业发展白皮书 (2023)*
China Communications Services Digital Infrastructure Industry Research Institute (2023) *White Paper on the Development of China's Data Center Industry (2023)* Accessed November 2024
https://aimg8.dlssyht.cn/u/551001/ueditor/file/276/551001/1684888884683143.pdf

[10] Patel, D. Nishball, D. Eliahou Ontiveros, J (2024) *AI Datacenter Energy Dilemma – Race for AI Datacenter Space* SemiAnalysis Accessed August 2024
https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/

[11] Freifeld, K (2025) *US tightens its grip on AI chip flows across the globe* Reuters. Accessed January 2025 https://www.reuters.com/technology/artificial-intelligence/us-tightens-its-grip-ai-chip-flows-across-globe-2025-01-13/

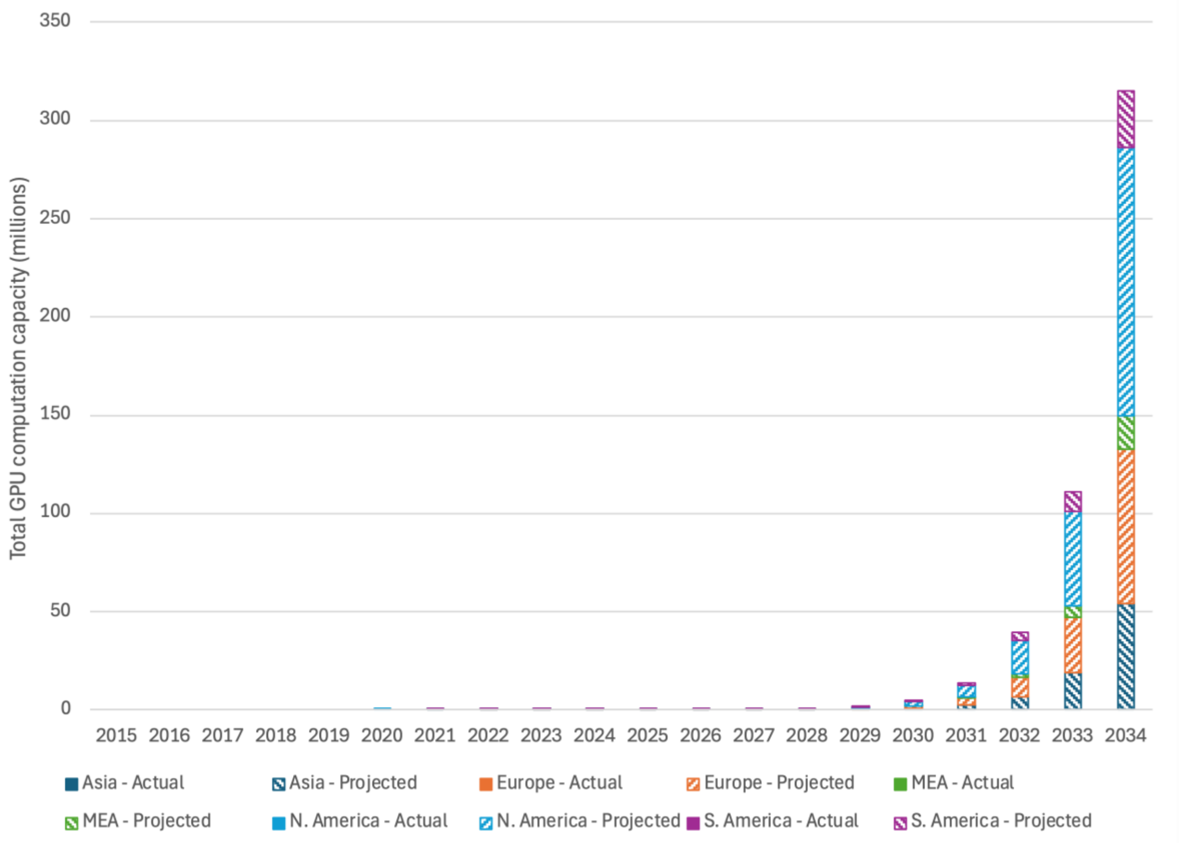
[12] Patel, D. Nishball D (2023) *GPU Cloud Economics Explained – The Hidden Truth* SemiAnalysis Accessed August 2024
https://semianalysis.com/2023/12/04/gpu-cloud-economics-explained-the/

[13] Total Cost of Ownership

rapidly. Furthermore, when power availability is limited and efficiency increases rapidly, faster replacement rates will increase available computing capacity at the same energy cost and allow them to sell the compute or services at lower cost than competitors, incentivising shorter lifetimes. Shehabi et al 2024 assumes a 5.7 year lifetime, and this substantially increases the installed base relative to our assumption of four years..

Figure 4 shows the estimated global GPU computational capacity by year and by region.

Figure 4 Modelled global GPU (AI) computational capacity



2.5.1 AI performance

There is no direct equivalent to SERT for AI performance and efficiency. MLPerf is used as this is the closest analogue that also publishes publicly available data[14]. MLPerf is a suite of standardized machine learning benchmarks developed by the MLCommons consortium. Unlike SERT, which focuses on general server efficiency, MLPerf specifically measures machine learning training and inference performance across different hardware platforms and frameworks.

MLPerf provides standardized performance metrics for AI workloads (eg llama, gpt, resnet), measuring both throughput and time-to-train a range of representative machine

[14] https://mlcommons.org/benchmarks/inference-datacenter/ Accessed December 2024

learning tasks including image classification, natural language processing, and recommendation systems.

However, MLPerf is relatively new compared to SERT, data are more limited and dominated by one manufacturer of accelerators, Nvidia GPUs. The benchmarks also evolve more frequently to keep pace with rapid developments in AI, including improvements in software efficiency, which can complicate long-term trend analysis. MLPerf also does not combine the different workloads in a similar way to SERT, and there can be wide divergence between workloads.

The analysis of MLPerf was based on the inference workloads with the most testing results, preferably with power testing. These were then normalised to the average performance of the NVidia H100 SXM 80GB GPU for which the most tests have been performed. Training workloads were not used because there was no power data available to analyse efficiency. Furthermore, training is performed on large clusters which makes the results more sensitive to other factors, such as network bandwidth and topology. The variation of GPU chips performance over time is shown in Figure 5 (own analysis).

The data shows an exponential increase in performance per GPU, particularly for the llama2 workload. However, the resnet workload shows less improvement comparatively. This is largely ignored because it is becoming less important than transformer models which are used for LLMs[15]. The relatively high performance 2020 GPU is also ignored because a highly performant and efficient GPU was launched but not widely sold. There are also very few datapoints for Google GPUs but we assume they must have competitive performance and efficiency.

[15] Large Language Models

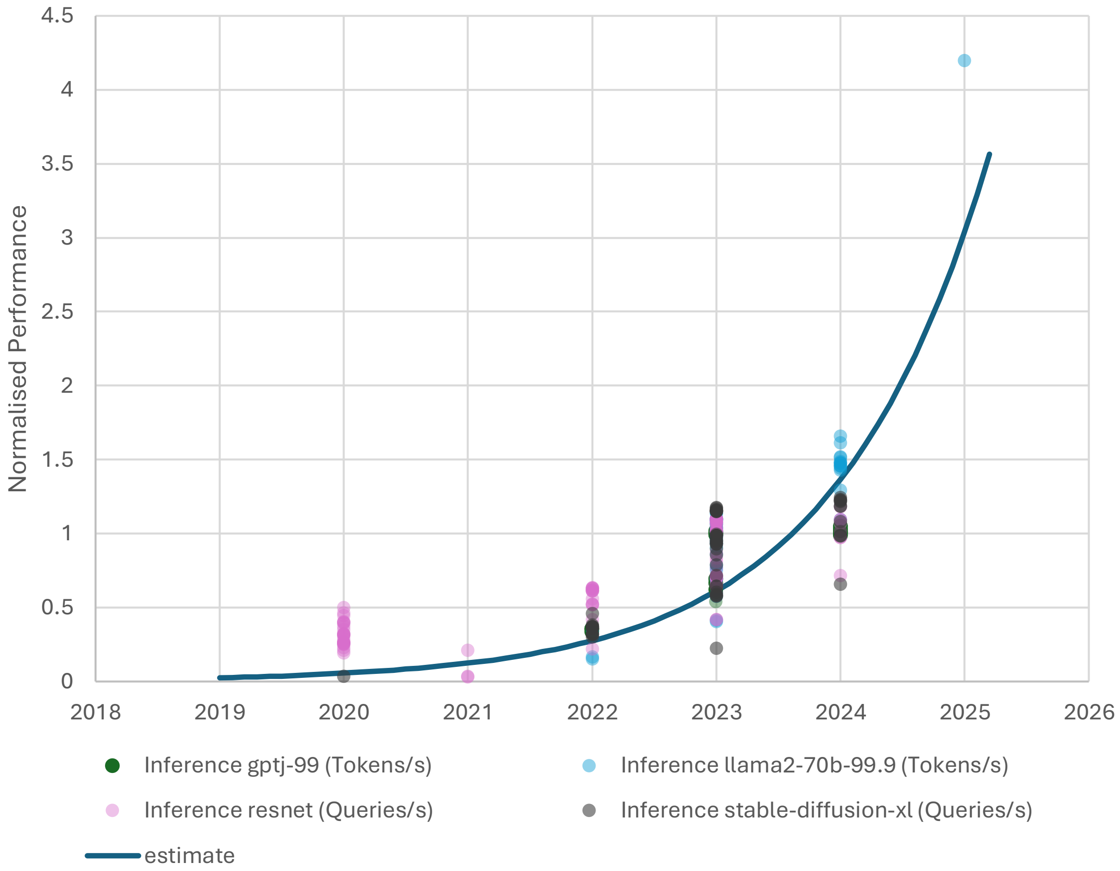


Figure 5 GPU normalised performance over time (own analysis)

2.6 Energy Consumption

Power demand at the expected utilisation is calculated using:

$$power_{load\%} = \frac{performance}{efficiency} * \frac{1 + idle\ ratio * util\_target}{1 + idle\ ratio * testing\ util}$$

This is derived from the basic formula efficiency = performance/power with an additional linear interpolation to account for the difference between the target utilisation and the utilisation at which efficiency was measured, where the performance, efficiency and testing utilisation must all be derived from the same test standard. For SERT[16] the test utilisation is 55%. It is assumed to be around 70% for MLPerf since no utilisation data is available and this is in line with the utilisation assumption in Shehabi et al 2024.

[16] While SERT does not have any units, this is simply a mathematical 'trick' because every worklet has its own units which do not add value to the overall metric. It is trivial to multiply by 1W or 1bit to remove the units because geometric means are used. The precise units of performance are not important as long as they are consistent.

Energy consumption is the product of the power and time, calculated for both the high and low utilisation levels.

The utilisation for each DC type is shown in Table 2.

Table 2 Assumed utilisation for each DC type

DC type	Utilisation mode	Utilisation %	Time (hr/yr)
Traditional, small DC	High	10%	5840
	Low	5%	2920
Cloud, large DC	High	30%	5840
	Low	15%	2920
Cloud, Hyperscale DC	High	40%	5840
	Low	20%	2920
AI, hyperscale DC	High	80%	5840
	Low	40%	2920

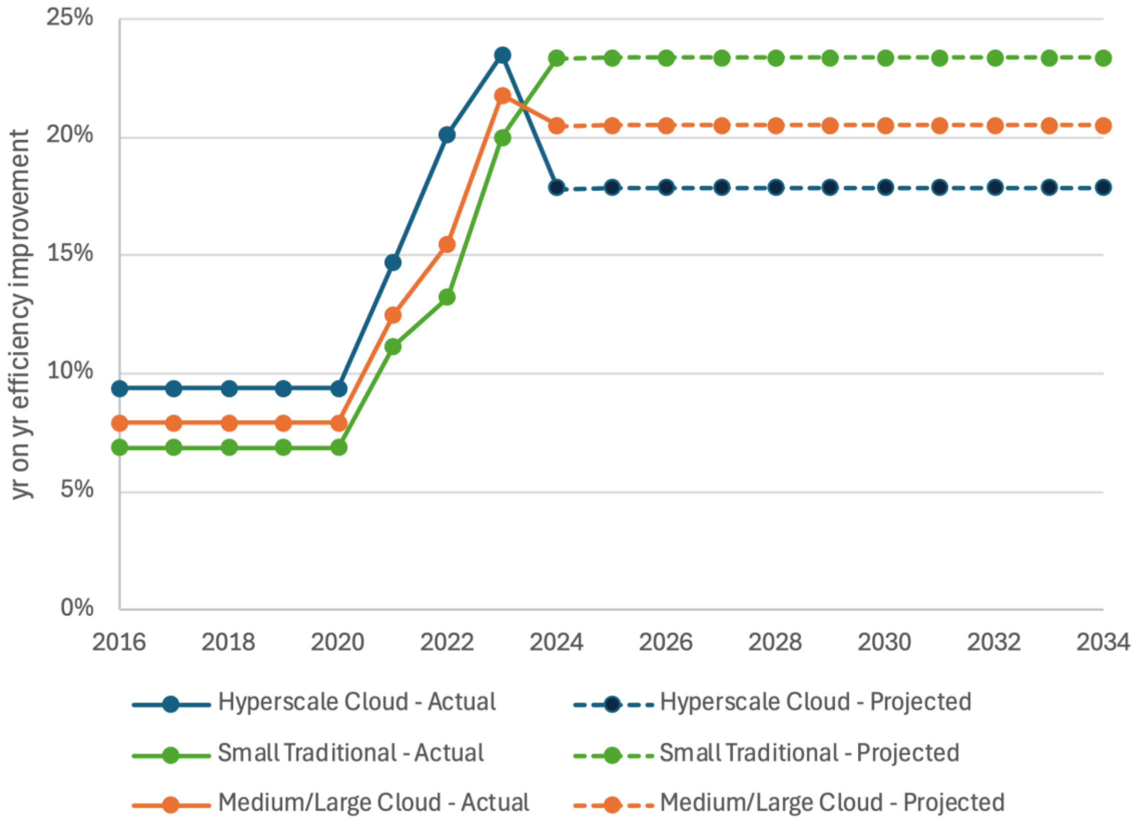
2.6.1 SERT efficiency and idle ratio

The efficiency is based on the server CPU launch year and SERT performance. It was not possible to derive a single formula that correlates with both year and performance so individual formulae were created for each year.

The efficiency, adjusted to take into account the sales weighting of server performance, ie taking into account the increasing proportion of sales of higher core servers, is shown in Figure 6. The annual efficiency improvement has been increasing in recent years from 6-8% in 2020 to around 18-23% in 2023, depending on the data centre type. Coroamă et al (2025)[17] estimate a similar efficiency value in 2023 but overall find a lower and consistent efficiency improvement trend of around 26% for the market dominant dual core servers, while identifying the very wide range in efficiency in a given year. Coroamă et al (2025)'s analysis shows that sales weighting has a considerable impact on the efficiency projections which could introduce further error if the sales data is poor.

[17] Coroamă, V C. Dumbravă, O. Hinterholzer, S. Progni, K. Hintemann, R (2025) *Energy efficiency of servers past and possible future trends*. IEA EDNA

Figure 6 *Modelled ICT efficiency improvements in TEM*



The idle ratio is the ratio of the max power (determined at 100% utilisation) and power at idle. Based on analysis of the SERT data, the idle ratio does not change over time.

2.6.2 AI efficiency and idle ratio

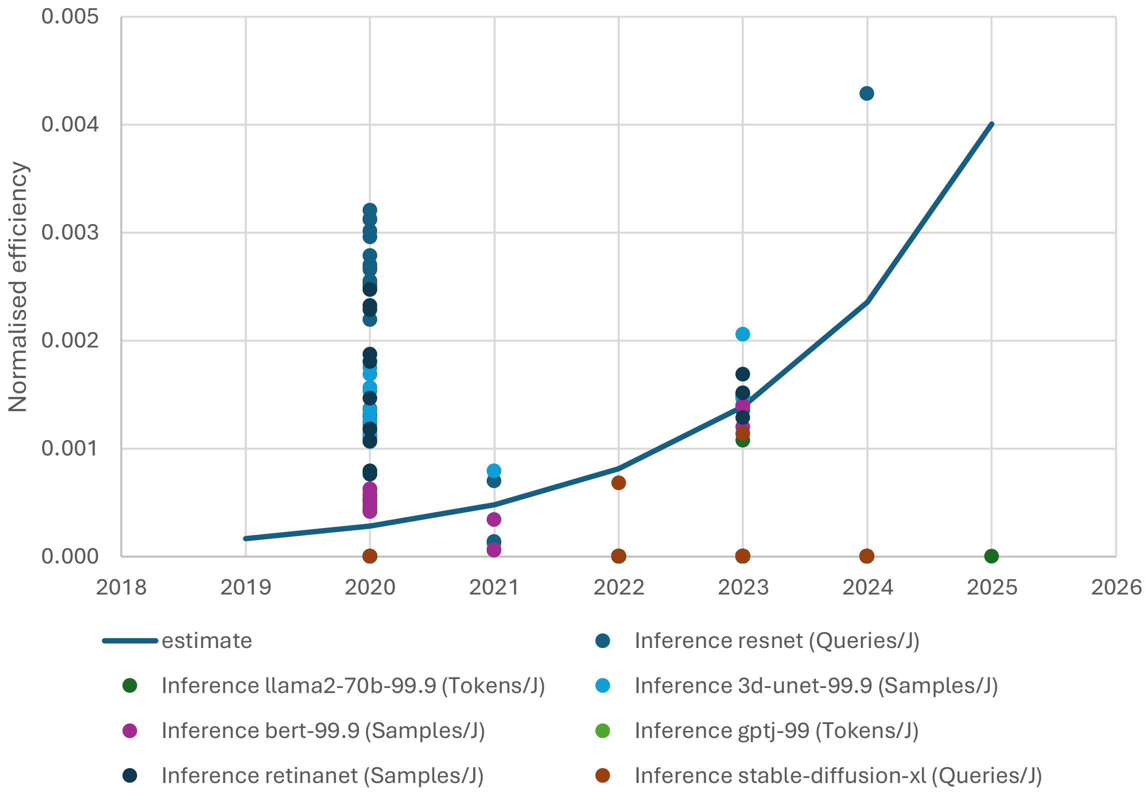
The trend in AI efficiency to date is analysed using the MLPerf inference workload data and normalised to the relative performance of an Nvidia H100 GPU (as shown in Figure 7). Relatively few test results included power measurements. The efficiency does not depend on the performance of the GPU because each year few GPUs are launched and they all aim to achieve maximum performance and efficiency. The large number of results in 2020 include that of the low sales GPU identified previously in the AI performance analysis (Section 2.5.1). As before this is not included in this analysis. These data may not include GPUs sold in China, which are subject to export restrictions.

Based on these data the efficiency improvement is exponential, as shown in Figure 7,with about a 10x improvement over the past few years. However, research from Tschand et al (2024 preprint)[18] analysis finds closer to 1000x efficiency improvement for

[18]Tschand, A. Tejusve Raghunath Rajan, A. Idgunji, S. Ghosh, A. Holleman, J. Kiraly, C. Ambalkar, P. Borkar, R. Chukka, R. Cockrell, T. Curtis, O. Fursin, G. Hodak, M. Kassa, H. Lokhmotov, A. Miskovic, D. Pan, Y. Prasad Manmathan, M. Raymond, L. St. John, T. Suresh, A. Taubitz, R. Zhan, S. Wasson, S. Kanter, D.

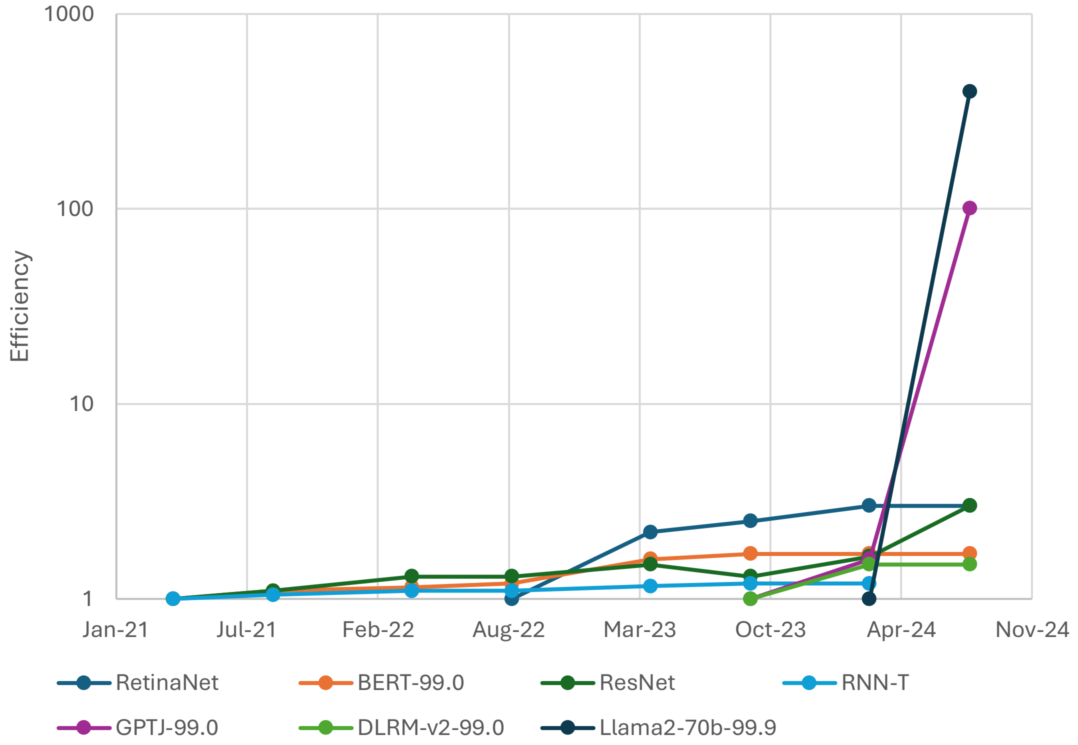
GPU chips using the same dataset (Figure 8, note log scale on y axis). This difference may be due to the way we have processed the data: as performance is normalised against the H100 chip results are excluded if the test is not performed on the H100 and the other chip in the launch year. In contrast, Tschand et al (2024 preprint) have normalised efficiency to 1 for the earliest test result and the x-axis is the date of the benchmark used. This results in a 10-100x magnitude difference in the efficiency improvement over the past 3-4 years alone as shown in Figure 8.

Figure 7 Normalised efficiency of GPUs over time



Janapa Reddi, V (2024 preprint) *MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from μWatts to MWatts for Sustainable AI* Arxiv Accessed December 2024 https://arxiv.org/html/2410.12032v1

Figure 8 Tschand et al (2024) DC AI efficiency analysis (log scale on y-axis)



Coroamă et al (2025) compared the efficiency trends using the manufacturer declared (theoretical) performance and thermal design power across a number of factors. It is inherently difficult to compare the MLPerf benchmarks with theoretical performance because many factors could affect real world performance/efficiency such as: software optimisation, what calculations are being made and memory bottlenecks. However, if we take what appears to be the most important factor, FP16/BF16 tensor efficiency, from Coroamă et al (2025) and compare it with the projected year on year annual energy for Hyperscale Cloud AI using the ML Perf analysis we see that they are close. While this could be coincidental, it provides additional confidence in the analysis.

Based on Shehabi et al 2024 we have assumed the Idle ratio for AI DCs is 5.

2.6.3 Storage and networking energy

Storage and networking energy are calculated as percentage on top of server energy and are assumed to be:

Storage energy + network energy = server energy * 25%

This is assumed to be unchanged over time. The percentage is also the same for AI servers which require very high speed (and high power) networking to link servers together for parallel processing during training but also inference workloads for very large models. Based on Shehabi et al (2024) storage and networking was approximately

35% of server energy in 2014, decreasing to 25% around 2021. The forecasts shows that the percentage drops to around 15% in 2028, due to the very high AI energy consumption. Modelling storage and networking can be improved using a similar approach to the server computational capacity but would require additional data and testing.

2.7 Infrastructure Energy

The DC infrastructure energy (cooling, power, lights security etc) is calculated from the PUE and ICT energy:

$$DC\ Infrastructure\ energy = ICT\ energy * (PUE - 1)$$

Where PUE values are segment-specific and time-dependent.

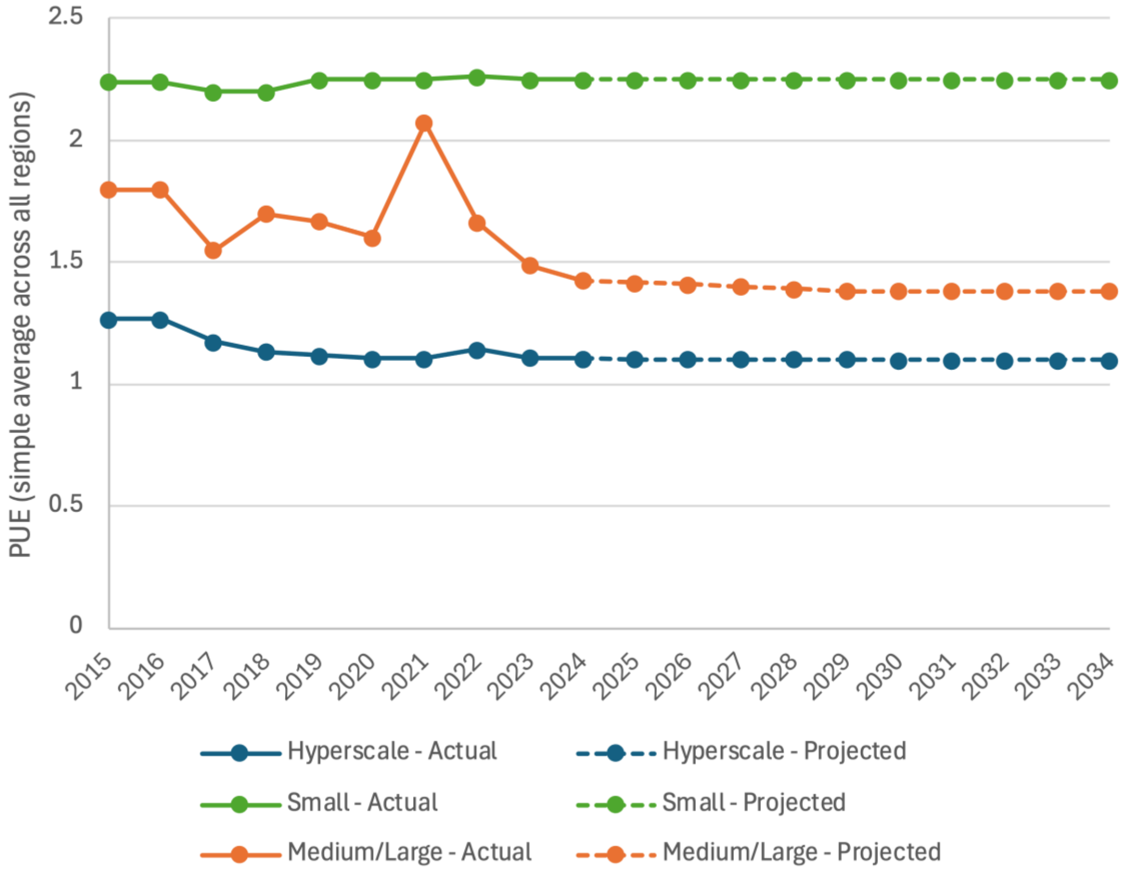
The PUE data is drawn from multiple sources for each region (see annex A). Where there is insufficient data for a region, the average global PUE is used.

PUE in USA and Europe has been relatively stable for the past few years and is not expected to improve, despite the expected introduction regulations setting minimum regulations, since the majority of DCs have already achieved the required performance. There is continued improvement in Asia and particularly China which had lagged behind in the past but recent data centres shows that they are already on par with best in class performance (Lanyang Technologies, 2023)[19]. Figure 9 shows the assumed average global PUE for three types of DC based on the data in Annex A. (AI is assumed to be in hyperscale DCs and have the same PUE. However, they are also being installed in Colo’s where dedicated DC’s are unavailable (eg Europe) and to distribute inferencing resources closer to the consumers and the ’edge’.)

Because the majority of the data are self-reported by hyperscale data centres in their sustainability reporting and large data centres in the EU CoC, the PUE data, although accurate, is likely to be better than the wider market, ie the less efficient data centres are less likely to report their efficiency.

[19] 兰洋科技 (2023) *数据中心能耗现状和能效水平分析*
Lanyang Technologies (2023) *Analysis of current energy consumption and energy efficiency levels in data centres* Accessed online November 2025
https://www.blueocean-china.net/faq3/234.html

Figure 9 Global data centre average PUE estimate



3 Policy modelling

The model supports analysis of policy interventions through adjustable parameters:

- Minimum ICT efficiency standards or labels
- Utilisation requirements
- Infrastructure efficiency
- Allocation of compute resources between DC types (accelerating shift away from small DCs)
- PUE improvements to traditional and cloud data centres.

In addition, it is possible to alter other factors which may be directly or indirectly affected by energy efficiency or other ecodesign policies, namely.

- Computing capacity (affected by increasing utilisation)
- Server lifetimes (to reduce lifecycle impacts)

While it is possible to change all the parameters in the base scenario and policy scenario (labelled as Alt scenario), for ease of use a simplified dashboard is available (See Annex B).

This simplified interface does not give full model functionality, for example minimum efficiency performance standards would result in a single step change in efficiency, while the dashboard can only model a continuous improvement. Depending on the policy maker requirements, it is possible to further refine and add functionality to the dashboard but this requires additional work and adds complexity.

4 Results

4.1 Note on the interpretation of the results

The model uses the data and assumptions described in the previous section to produce projections of energy use, split by DC type and region. The base scenario projections assume that trends in the demand for computing and the changes in efficiency continue into the future.

'Conventional' data centre use and technology is well established and understood – while they continue to evolve there are reasonable grounds to make projections into the future. The situation is different for AI which is still in its infancy; the future uptake and efficiency of AI is highly uncertain.

The model does not take into account factors that will limit the growth of data centres in general and AI data centres in particular, such as the constraints on the supply of chips/servers, the capital to build more data centres or the electricity to supply them with energy, all of which are likely to apply. These economic and practical factors are beyond the scope of this modelling but are expected to have a big impact on data centre development, particularly for AI data centres. Studies by other authors (Goldman Sach 2024 and SemiAnalysis 2024) do consider these constraints; and their projections are compared to those from the TEM and discussed in section 5.

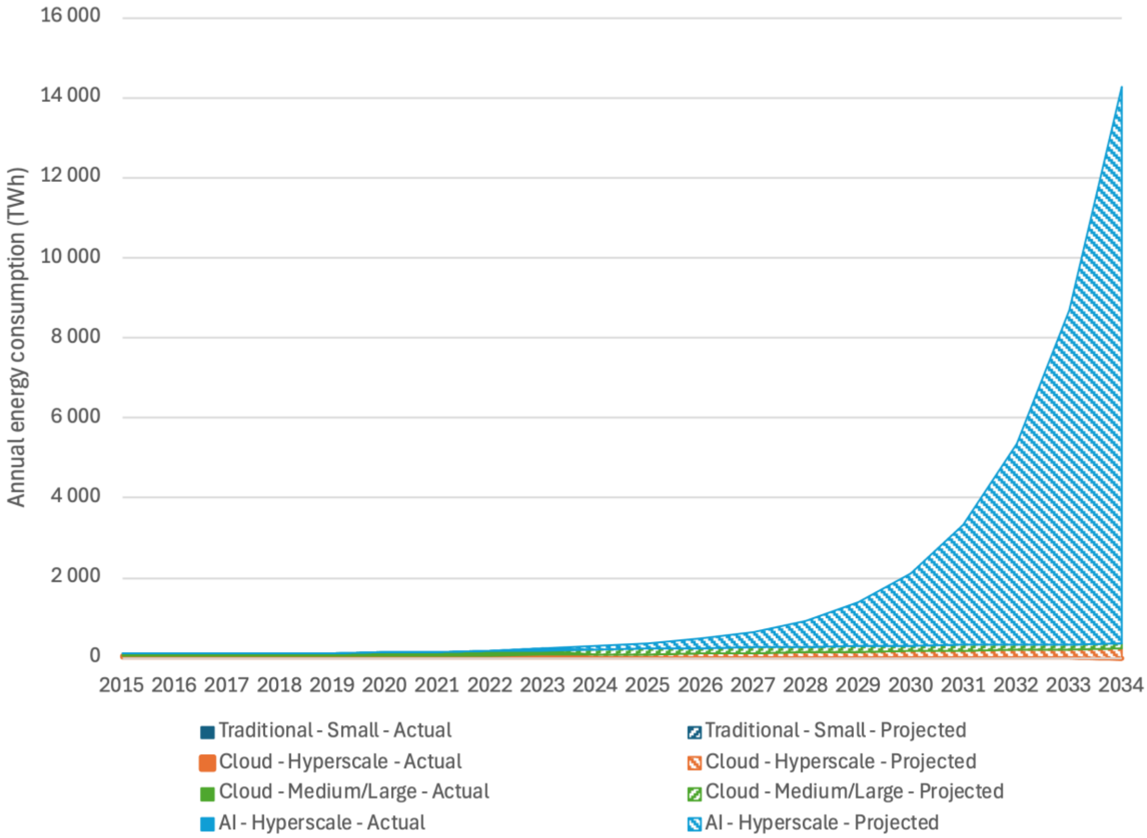
To summarise, these values are **not predictions**, they illustrate a hypothetical case in a fictional world where current trends in the growth of DC use for AI can continue without practical restraints.

The model has been intentionally structured so that additional, newer and better data can be integrated with relative simplicity. This will enable policymakers to create more accurate predictions for the market segments they are interested in and subsequently estimate the potential impacts of policies.

4.2 Overall projections

The results (Figure 10) show that from 2030 energy is entirely dominated by AI. This is due to the projected massive growth of AI compute capacity[20] which is forecast through a simple extrapolation of current trends. This outpaces the expected but highly uncertain improvements in efficiency. If this growth were to take place this consumption would have enormous impacts on the energy markets. The model does not account for power or budget constraints and other financial, supply or demand constraints which would set an upper limit on energy consumption. Because of this, Figure 11 and subsequent figures show a detailed view with the y-axis truncated at 1000 TWh to show the other trends.

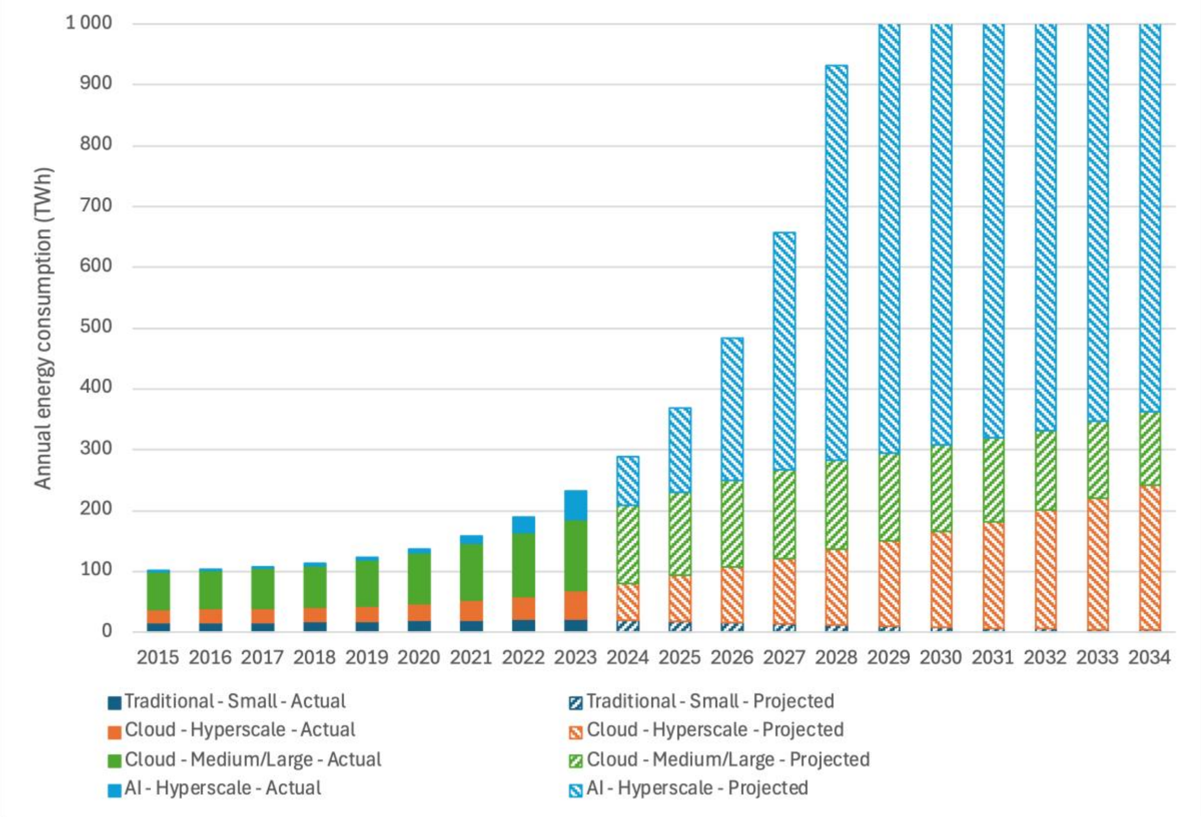
Figure 10 DC energy consumption by DC and compute type



Non-AI compute energy is also expected to triple by 2034 after relatively stable energy consumption from 2015 to 2020. Most of this increase is in hyperscale DCs, with use of small DCs declining over time.

[20] as described above, a combination of number of chips and their efficiency

Figure 11 DC energy consumption by DC and compute type (detail view)

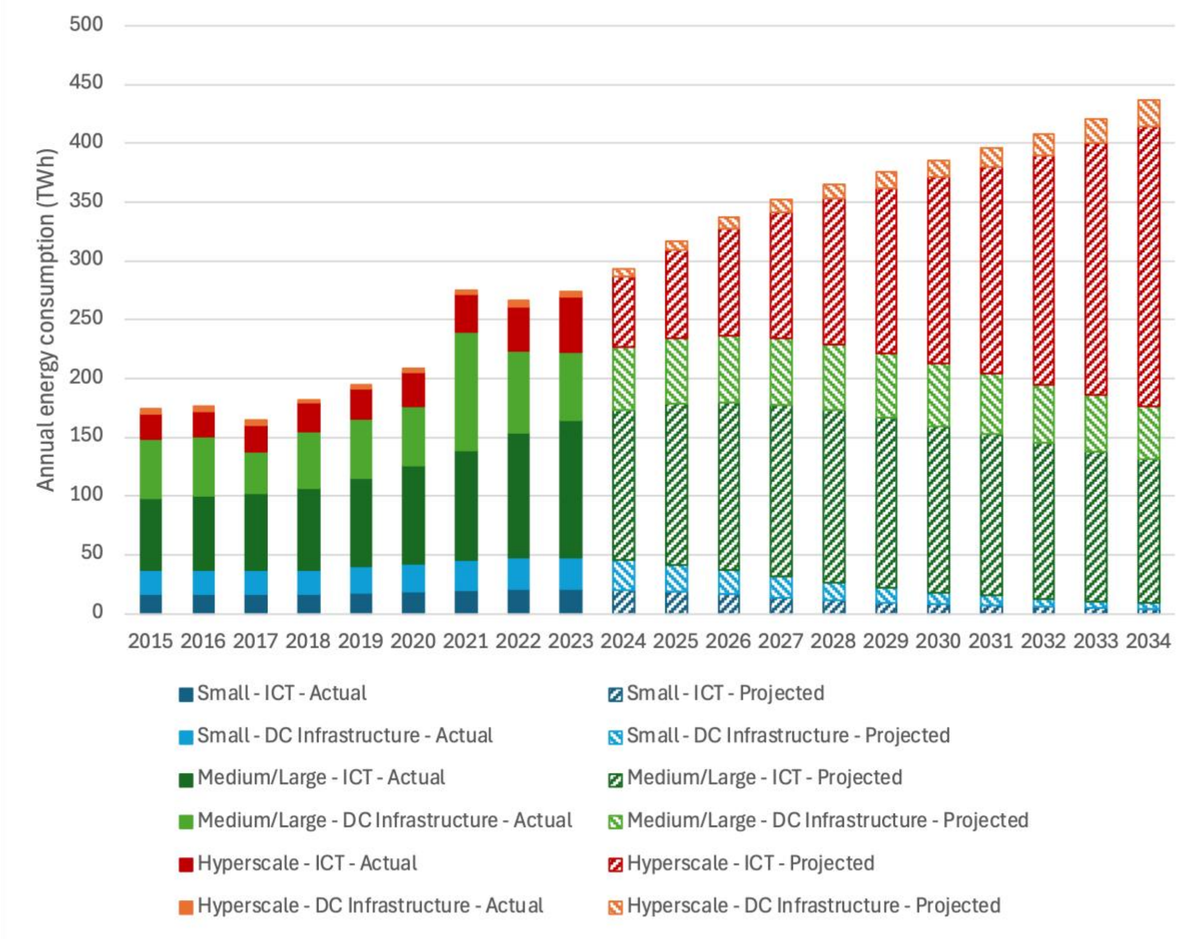


4.2.1 ICT vs DC infrastructure

Figure 12 shows that DC infrastructure[21] still has a significant contribution to the overall energy consumption when AI is removed. In particular, the medium/large DCs make a relatively large contribution to overall energy use and have scope to be reduced further by ambitious policy.

[21] ICT infrastructure is network and storage, DC infrastructure is non ICT energy use eg cooling equipment

Figure 12 non-AI ICT and DC infrastructure energy consumption



Infrastructure energy use is less significant for AI.

4.2.2 Energy consumption by region

The projection for non AI (Figure 13) and AI (Figure 14) is that North American DCs consume the most electricity, followed by Europe and then Asia . South America and the Middle East and Africa have extremely low energy consumption. This trend changes very little over time. However, the growth is extrapolated from existing trends and does not account for relative economic growth, industrial policy or other geopolitical factors that could affect this greatly. For example, there is substantial investment in AI in the Middle East (Benito, 2024)[22] and attempts to attract and grow AI capacity there (Newman et al, 2024)[23]. China is also building chip manufacturing and AI capability which could result in much more rapid growth there.

[22] Benito, A (2024) *Saudi Arabia launches $100 Billion AI initiative to lead in global tech* CIO https://www.cio.com/article/3602900/saudi-arabia-launches-100-billion-ai-initiative-to-lead-in-global-tech.html

[23] Newman, M. Bergen, M. Solon, O. (2024) *Race for AI supremacy in Middle East is measured in data centres* https://www.bloomberg.com/news/articles/2024-04-11/race-for-ai-supremacy-in-middle-east-is-measured-in-data-centers

Figure 13 non-AI energy consumption by region

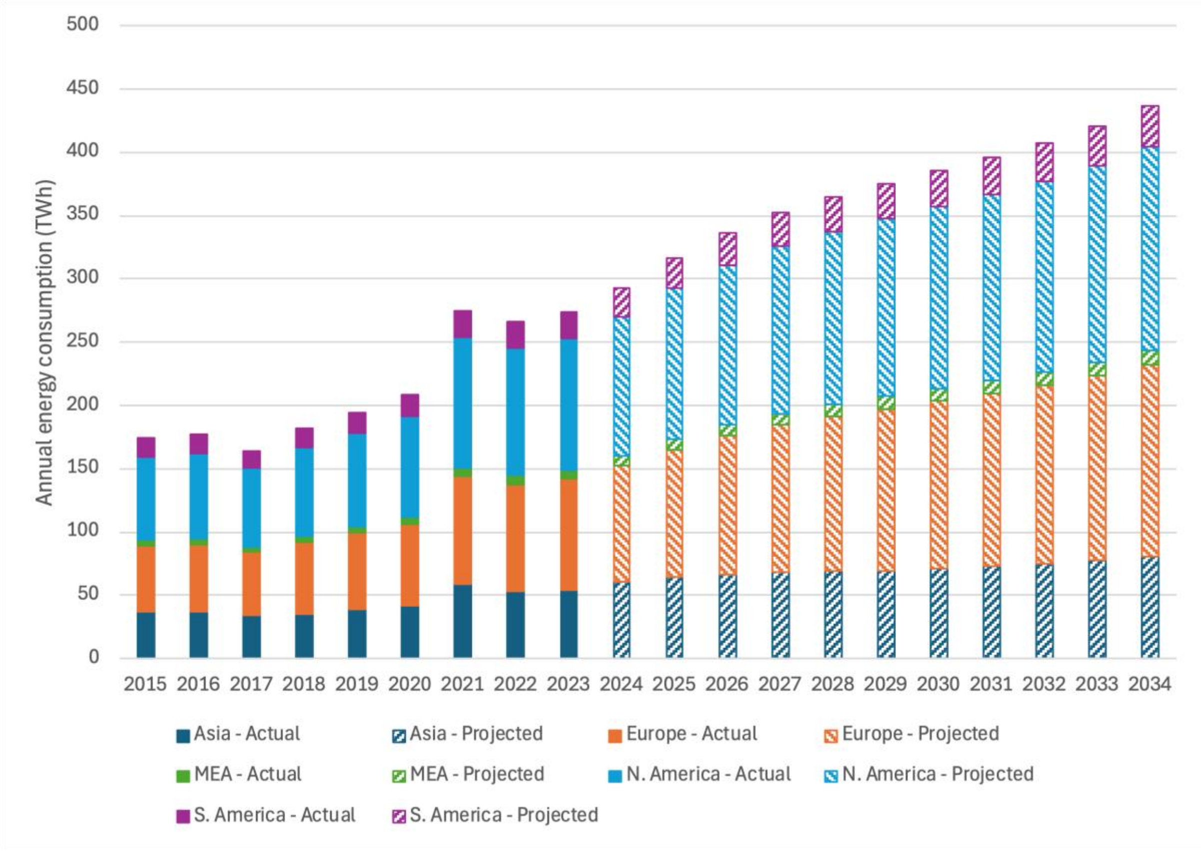
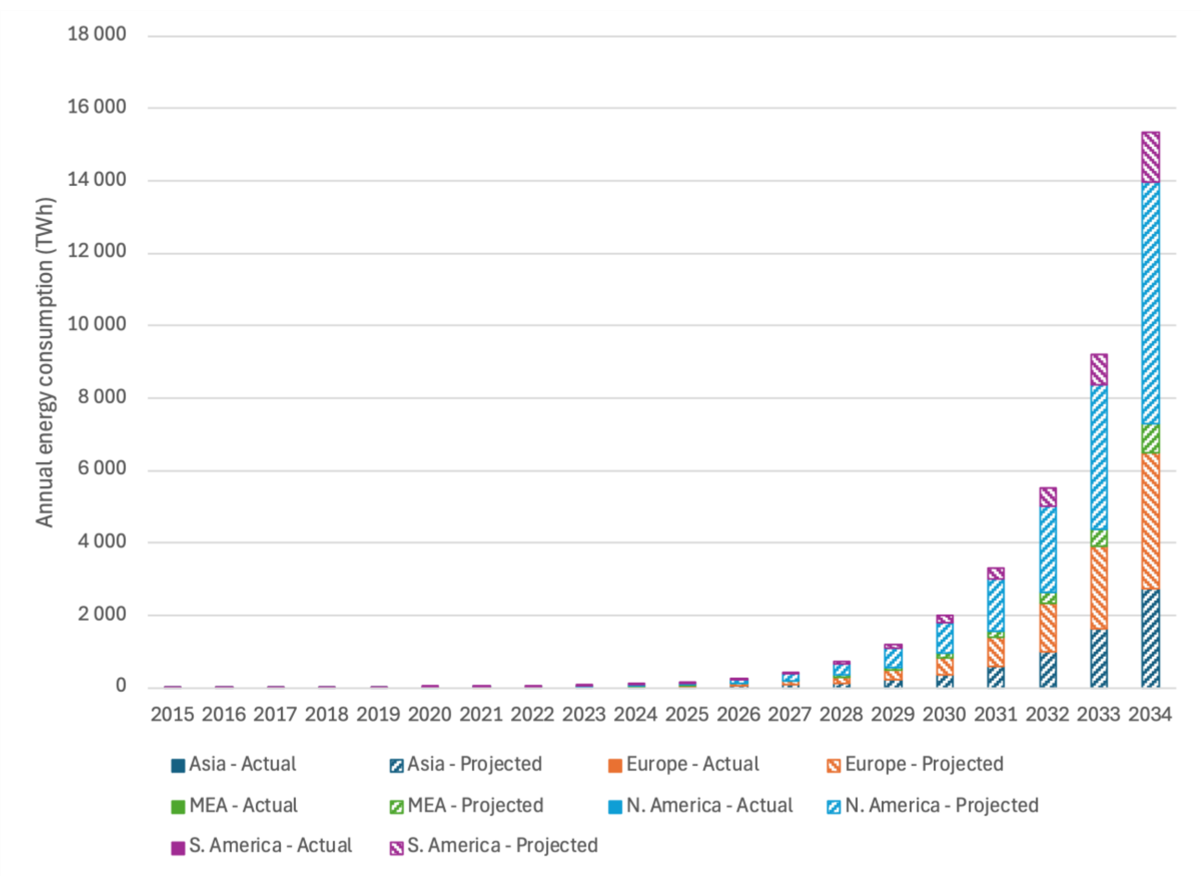


Figure 14 AI energy consumption by region



5 Comparisons with other projections

To validate the Total Energy Model (TEM) results and understand its limitations, we compared our projections with those from four recent studies that forecast data centre energy consumption. These studies were selected because they represent different methodological approaches and provide detailed projections through 2030, allowing for meaningful comparison with our model. They are:

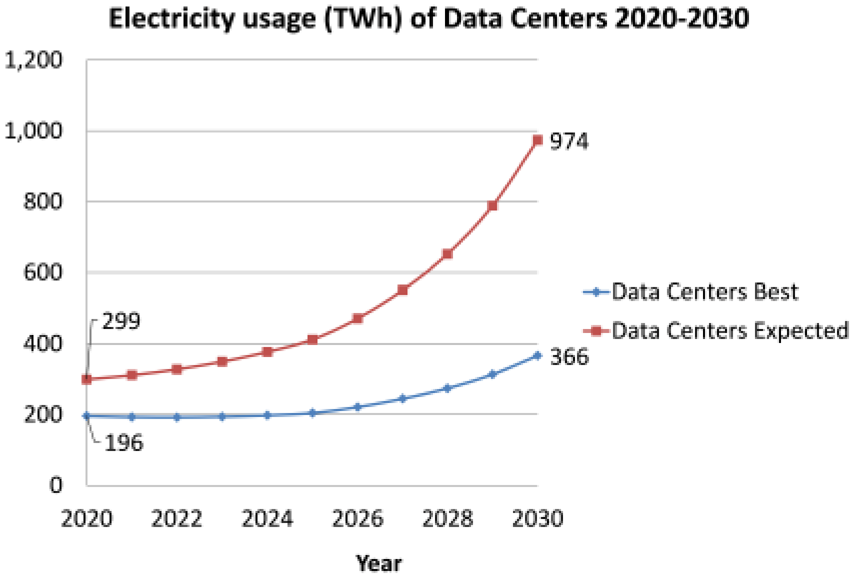
1. Andrae (2020): Uses data traffic-based methodology
2. Goldman Sachs (2024): Employs financial/market-based constraints
3. SemiAnalysis (2024): Utilises detailed technical analysis and power supply constraints
4. LBNL (2024): Uses a hybrid approach

Since the TEM base case is unconstrained, the energy consumption is expected to be significantly higher than the other projections.

5.1 Andrae, 2020

Andrae, A.S.G. (2020)[24] estimated future electricity consumption based on the estimated data demand, and the energy required per unit of data supplied. Two scenarios were presented, an expected case and best case as shown in Figure 15.

Figure 15 Electricity usage (TWh) of Data Centers 2020-2030 Source: Andrae, 2020



[24] Andrae, A.S.G. (2020). *New perspectives on internet electricity use in 2030.* Engineering and Applied Science Letters 3, 14.

This is a relatively old paper and published before the widespread availability and uptake of AI, particularly LLMs such as ChatGPT and these estimates do not explicitly account for AI. Table 3 compares these estimates with TEM projections.

Table 3 Comparison of Andrae and TEM 2023 and 2030 energy estimates

Model	**Energy use (TWh)**			
	2023		**2030**	
	expected case	**best case**	**expected case**	**best case**
Andrae (2020)	375	200	974	366
TEM 4	323	-	2510	-

Comparing the estimated energy consumption the TEM estimate in 2023 is between the expected and base case but in 2030 the TEM is much higher (by more than a factor of two) than the expected case.

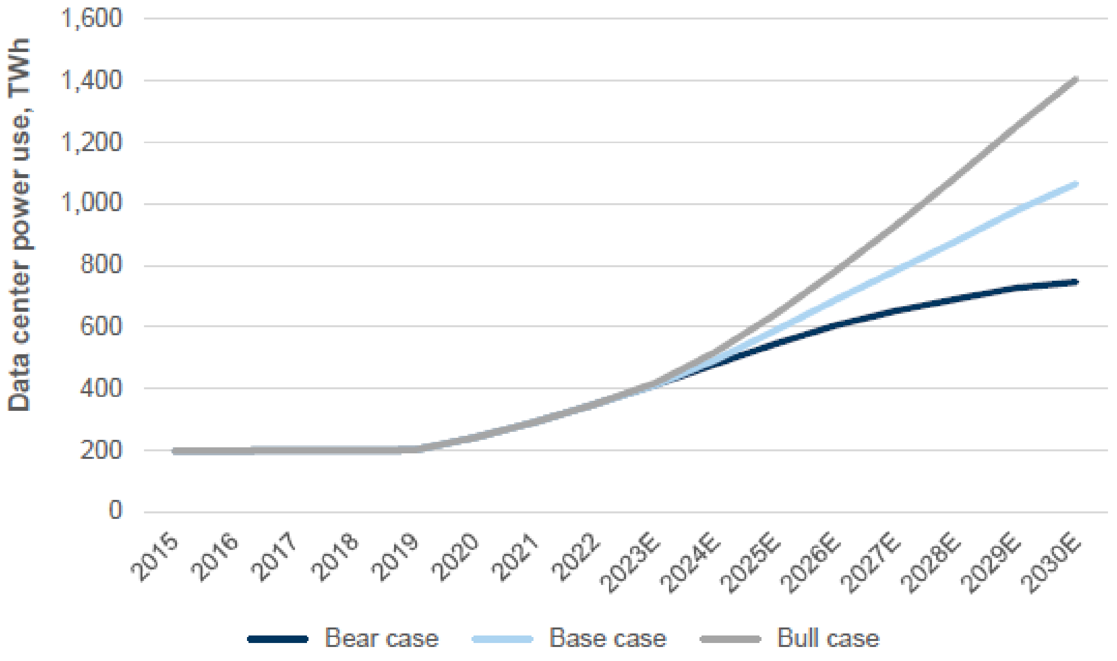
The data demand for AI servers undertaking computations is very high, especially for training but also when inferencing. However, the output data for the most common use, chatbots, is very low since it is mostly text. We are not aware of any research that has tried to quantify this data demand (or how it should be quantified) and therefore it is not possible to interpret the data demand used and so to adjust Andrae's estimates for AI.

5.2 Goldman Sachs, 2024

This report[25] was published in April 2024. The authors' projections for energy use are shown in Figure 16.

[25] Davenport, C. Singer, B. Mehta, N. Lee, B. Mackay, J. Modak, A. Corbett, B. Miller, J. Hari, T. Ritchie, J. Delaney, M. Revich, J. Jaiya, J. Venugopal, V. Cash, N and Halferty, O. (2024) *Generational growth, AI data centres and the coming US power demand surge*, Goldman Sachs Accessed January 2025

Exhibit 8: We see 2030 power use from data centers 1.8x-3.4x 2023 levels in our bear/bull case
Electricity demand from data centers in TWh, base case, bear case and bull case



Source: Masanet et. al (2020), Cisco, IEA, Goldman Sachs Global Investment Research

Figure 16 Goldman Sachs projections of data centre energy use

These are compared with TEM projections in Table 4 and Table 5. The energy consumption in the two models shows similar historical trends, approx. 200TWh in 2015 but has diverged in 2023, energy consumption is higher in the Goldman Sachs (GS) model. In 2030 the TEM projection is much higher than even the GS bull case with the difference being largely a much high projection in the TEM for AI energy use.

Table 4 Goldman Sachs and TEM estimates of data centre energy use in 2023 compared

	Base case		
	AI (TWh)	**Non-AI (TWh)**	**Total (TWh)**
Goldman Sachs	75	350	425
TEM 4	56	267	323

Table 5 Goldman Sachs and TEM estimates of data centre energy use in 2030 compared

	Base case			**Bear case**	**Bull case**
	AI (TWh)	**Non-AI (TWh)**	**Total (TWh)**	**Total (TWh)**	**Total (TWh)**
Goldman Sachs	500	650	1150	820	1400
TEM 4	2170	354	2510	-	-

The modelling approach is different to that of TEM, Andrae, Shehabi et al in that it uses financial modelling as well as technical data.

GS state that they consider two scenarios:

- server-supply driven, that is the supply of compute capacity either due to manufacturing or budget limits is the predominant constraint
- demand for computing power is the main driver.

GS applied both a server supply driven forecast and a compute demand driven forecast, with a heavier weight applied towards the supply-driven methodology.

Server capacity is unconstrained in the TEM, while Goldman Sachs (Davenport et al 2024) believe it to be the major factor limiting growth. The supply side growth (and constraint) is quantified in terms of revenue, not number or power of servers so comparing this to computing capacity would take additional analysis that is not in scope for this project.

In addition, there are some differences in inputs and assumptions between the GS and TEM models.GS expect data centre energy use excluding AI to rise due to an increase in workload only partially offset by modest power efficiency gains in cloud/hyperscale. Their predictions are more sensitive to this second factor. They note that efficiency gains slowed considerably in 2022 and 2023 but expect them to pick up slightly. Their predictions are shown in Figure 17.

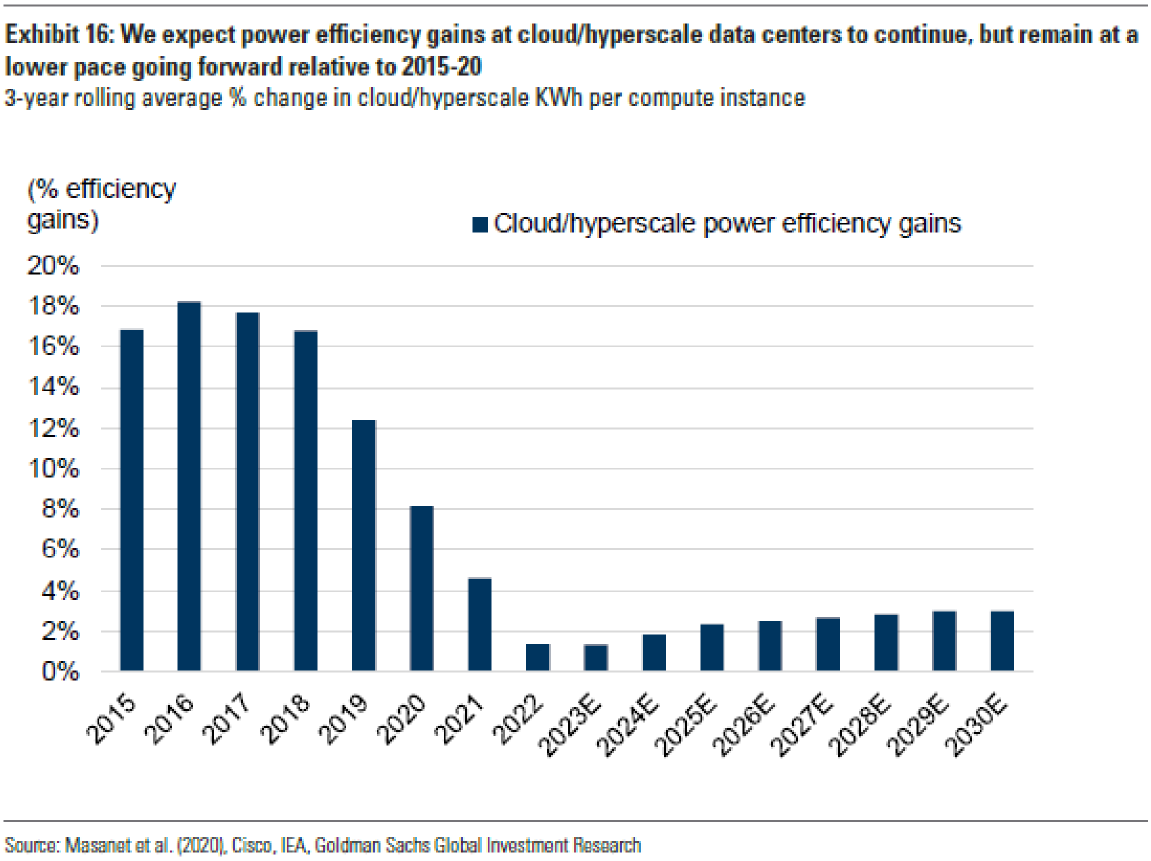


Exhibit 16: We expect power efficiency gains at cloud/hyperscale data centers to continue, but remain at a lower pace going forward relative to 2015-20

3-year rolling average % change in cloud/hyperscale KWh per compute instance

Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Global Investment Research

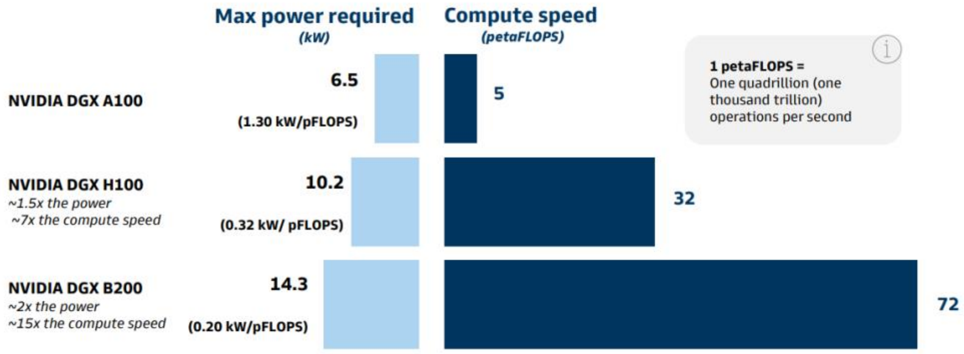
Figure 17 Goldman Sachs data and predictions of cloud/hyperscale data centre efficiency gains

The efficiency gains are much lower than identified in our analysis of SEED as shown previously in Figure 6. These take into account the efficiency gains from using CPUs with more cores and continued improvements in process node technology and chip packaging. Based on our estimates efficiency has been improving faster, at approximately 17% annually for hyperscale data centres vs 2% in the Goldman Sachs model. SEED is based on efficiency testing of only three configurations which may not be representative of real world configurations that could have lower efficiency but this difference in configuration cannot account for such a large difference. Overall, this use of SEED data in the TEM model may make it more realistic than other models. However the SERT data needs ground truthing with real measurements of energy use from operational data centres to validate it.

On the other hand, GS have present improvement in AI efficiency in terms of theoretical performance (FLOPS) as shown in Figure 18. We consider this to be less realistic than testing with MLPerf which includes limits from factors such as software architecture, network speed, memory bandwidth and cooling. Therefore, the FLOPS measurements will overestimate the efficiency gains made. However, if the GS model is solely supply constrained, the reduced efficiency would result in lower overall compute capacity rather than affecting energy consumed. It is not known if this would have changed the weightings of the scenarios used to increase total compute capacity and therefore energy consumed.

Exhibit 12: We have seen new AI innovations increase max power consumption per server but increase computing speed per server by an even greater level, representing meaningful reduction in power intensity

Recent evolution of NVIDIA server system specifications are indicative of increasing max power per server but with lower power intensity relative to computing speed (for training)



Source: NVIDIA, Goldman Sachs Global Investment Research

Figure 18 Data presented by Goldman Sachs on changes in power demand and computing speed for different generations of GPUs

5.3 SemiAnalysis, 2024

SemiAnalysis (2024) applied a detailed methodology which combined data on:

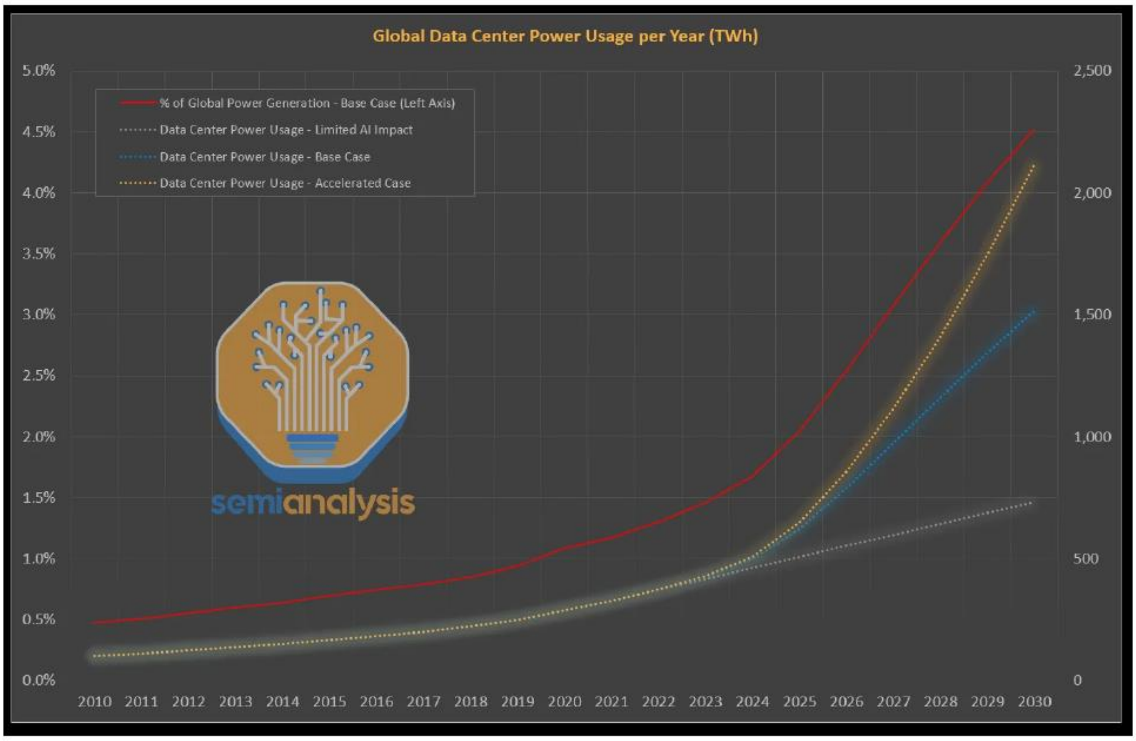
- Current conventional and AI data centre capacity (MW) including analysis of satellite imagery of buildings and new construction
- Projections of future conventional and AI data centre capacity based on demand for compute capacity constrained by:
    - Shipments of new chips
    - Constraints building and providing electricity to new data centres
- Chip and server energy use
- Server utilisation
- PUE

The data used were a combination of published and gathered by the organisation. A customised model was used to generate the energy projections. SemiAnalysis maintains a commercial product updating and selling their Data Center Industry Model[26].

Similar to Goldman Sachs, the authors have applied constraints to the supply of GPUs, but also considered how building and electricity supply could constrain energy consumption. The methodology is more detailed and seems to be more robust than

[26] https://semianalysis.com/datacenter-industry-model/

Goldman Sachs but there is insufficient detail to be certain. The resulting projections are shown in Figure 19.



SemiAnalysis Estimates

Figure 19 SemiAnalysis projections of data centre energy use

The results are compared with TEM projections in Table 6.

Table 6 Comparison of SemiAnalysis and TEM 2023 and 2030 energy projections

	2023	**2030**		
	Base case	**Base case**	**Limited AI impact**	**Accelerated case**
	Total (TWh)	**Total (TWh)**	**Total (TWh)**	**Total (TWh)**
SemiAnalysis	450	1 500	500	2 250
TEM 4	323	2 510	-	-

The SemiAnalysis estimate of energy consumption in 2023 is significantly higher than that from the TEM. In 2030 the projected base energy consumption from the SemiAnalysis model is significantly less than TEM due to constrained supply (as for GS). However, the Accelerated case and TEM are similar suggesting the unconstrained TEM represents a realistic upper limit to energy consumption.

SemiAnalysis present their estimates of compute capacity in FLOPS. As shown in Figure 20, they expect capacity growth to peak at 75% Quarter on Quarter (QoQ)(937% yr on yr if maintained over four quarters) and drop to approximately 25% in 4Q25 (244% yr on yr). The TEM model year on year capacity growth, based on the MLPerf benchmarked capacity is shown in Figure 21. 283% year on year is equivalent to 30% QoQ but, taking into account the different measurement approaches, the projected capacity growth from TEM is higher than that assumed by SemiAnlysis by 4Q25.

Additional analysis may be able to establish a correlation between FLOPS and MLPerf capacities across different hardware architectures and software that could be used as an adjustment factor so that values from the two measurements could be compared. This is outside the scope of this study.

Figure 20 AI compute capacity (in FLOPS) Growth %QoQ. Source SemiAnalysis (2024)

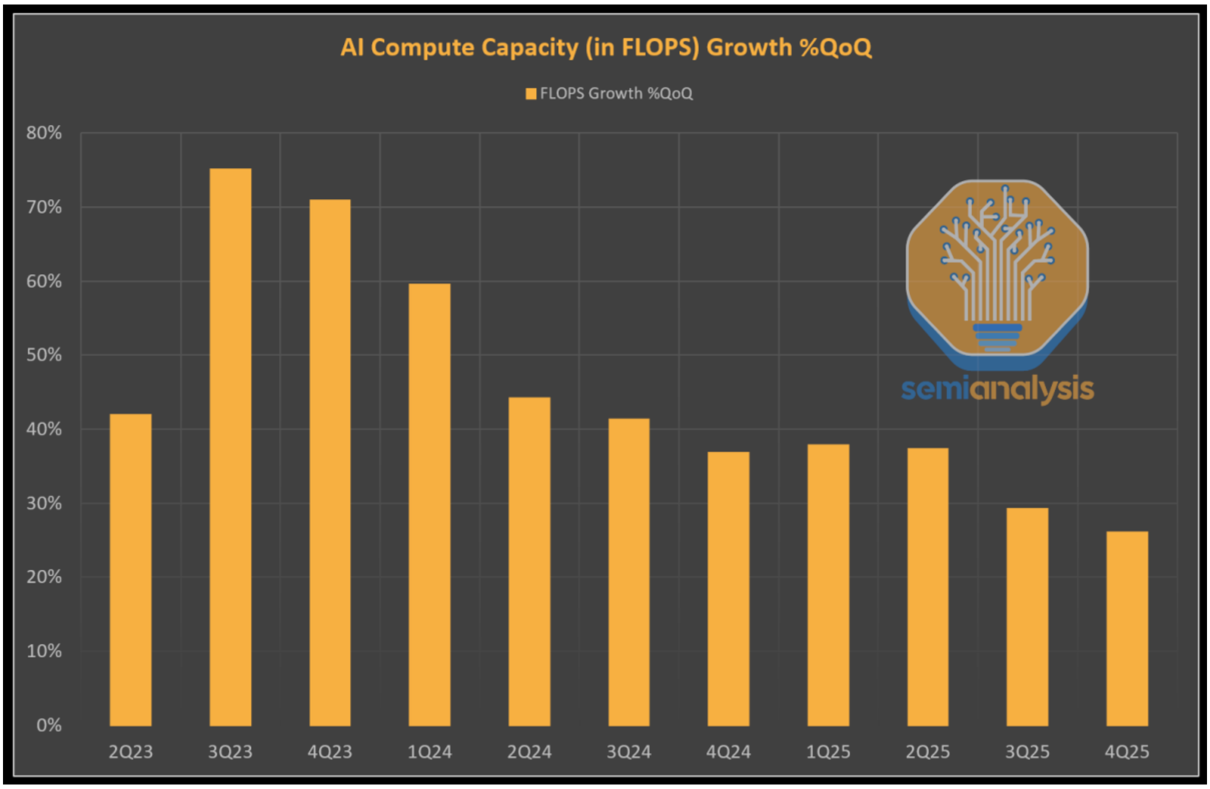
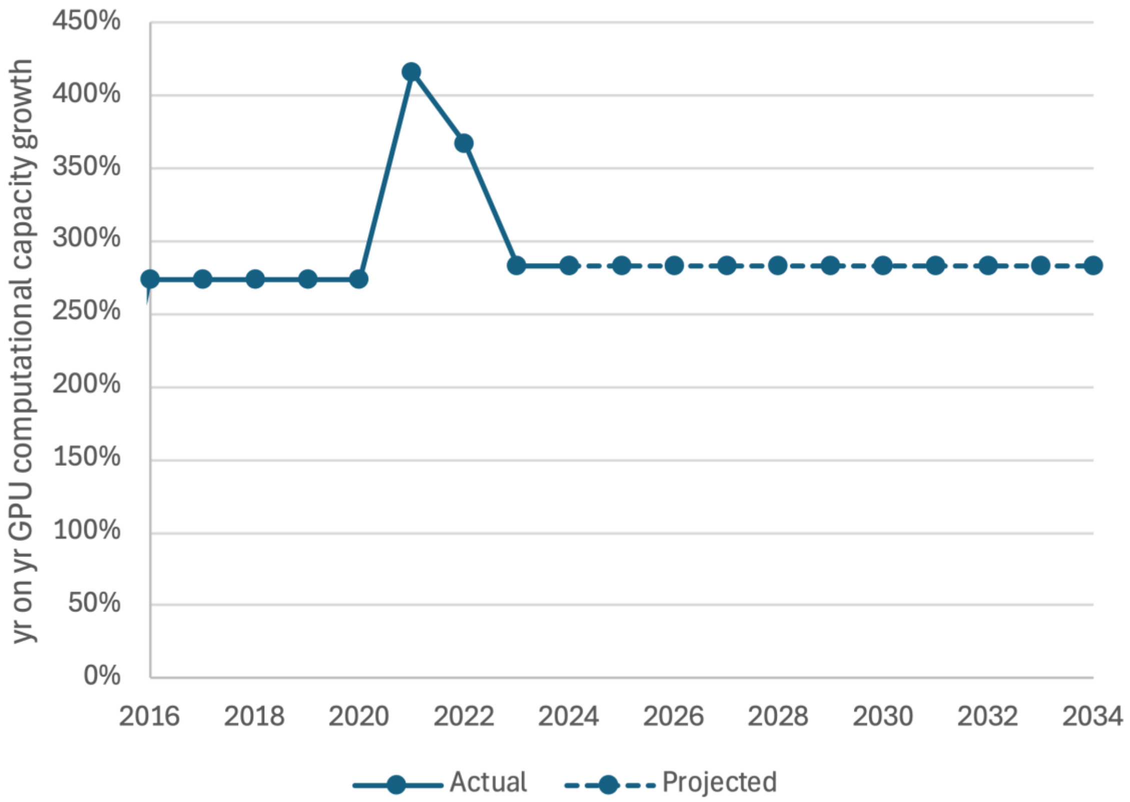


Figure 21 TEM estimates of yr on yr GPU compute capacity growth



5.4 LBNL 2024

The LBNL model[27] is a bottom-up model built on hardware equipment shipment (server, storage and networking). The server power consumption is then calculated from SEED values with ground truth testing. The power per server depends on the number of server sockets and is applied to all types of data centre. The networking power depends on port speed, with higher power for the growing market for very high speed connections. PUE is estimated based on the cooling architecture which is then associated with different types of data centre.

LBNL and TEM projections of data centre energy use in the USA and North America (USA, Canada and Mexico) are shown in Figure 22 and Figure 23. The server energy consumption shows similar trends. However, the growth trend in TEM lags behind the LBNL model, 50TWh is reached in 2019 and 100TWh in 2023, compared with 2021 and 2024 from the TEM. It would be expected that energy consumption from TEM to be higher given the wider geographic coverage (including Canada and Mexico as well as the USA).

[27] Shehabi et al (2024)

Figure 22 USA Server electricity consumption Source: LBNL, 2024

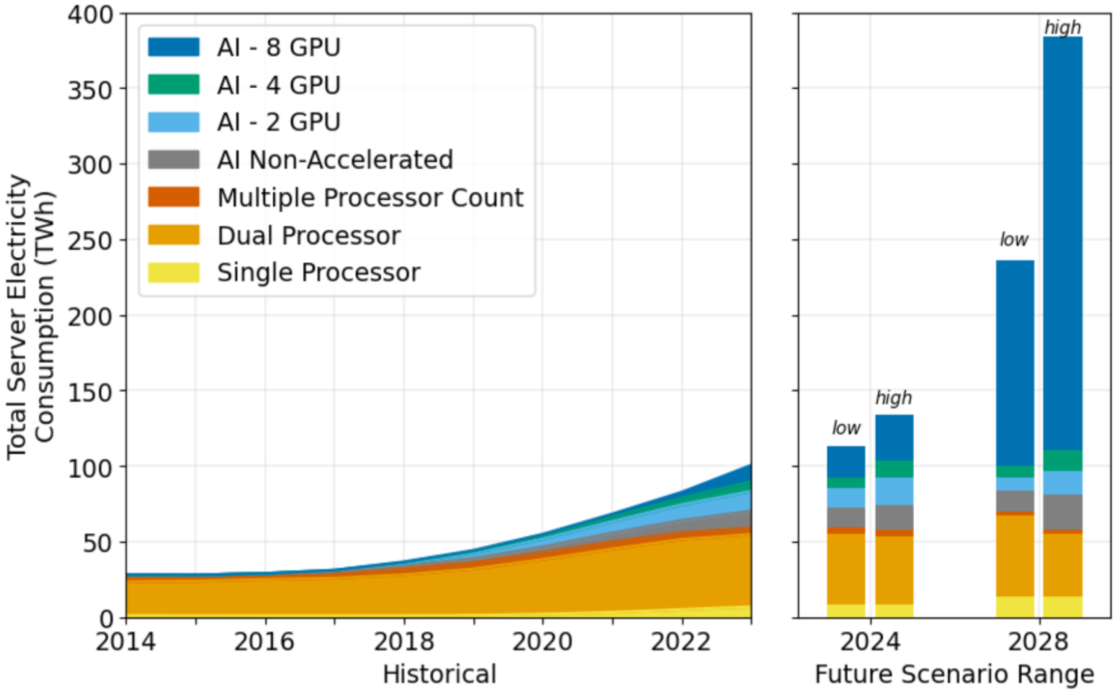
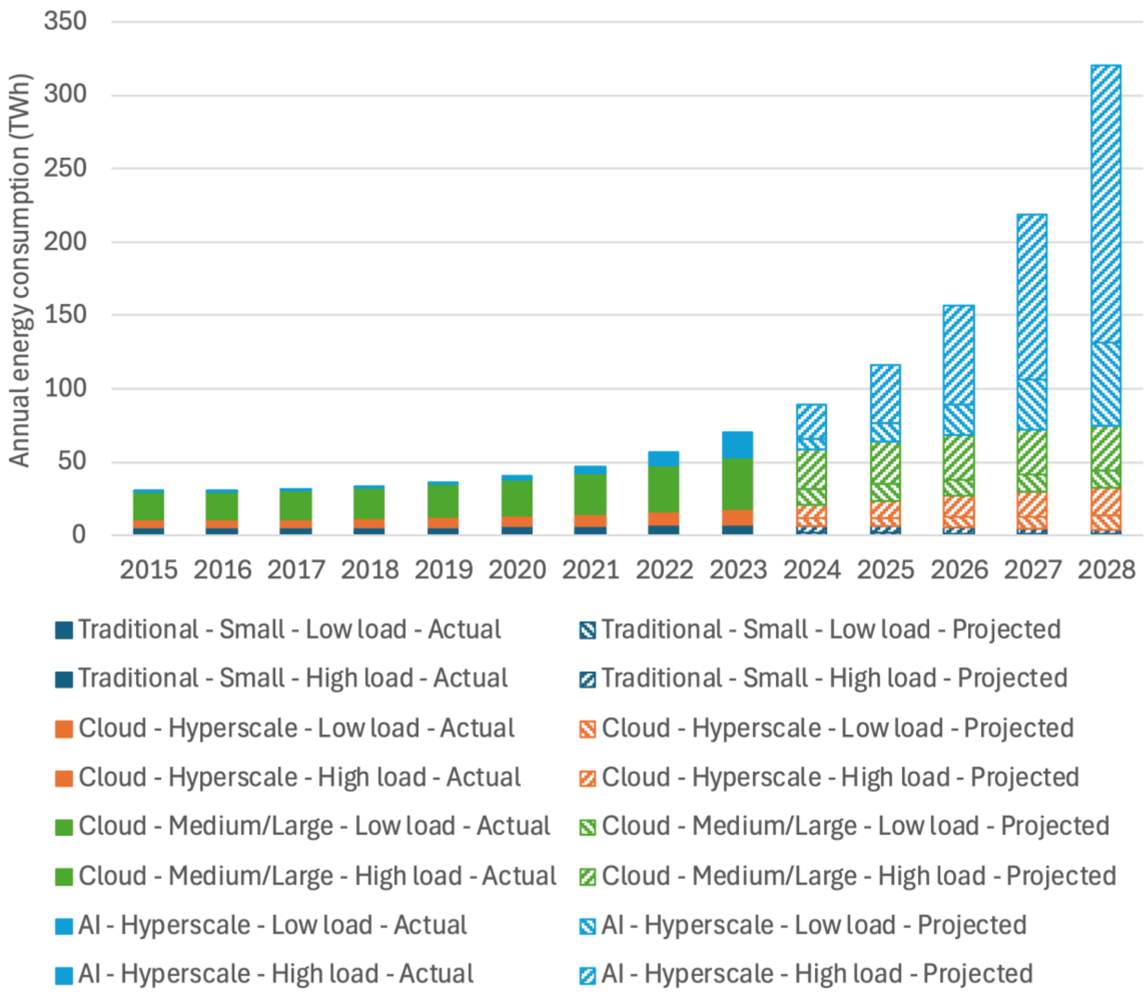


Figure 23 TEM North America server energy consumption



Some similarities between the results of LBNL and from the TEM is expected since the same datasets (SEED) are used and some assumptions are taken directly from LBNL model are used in the TEM. However, there are some notable differences in the inputs:

- Different server powers and efficiencies are assumed for different DC types in TEM.
- The lifetime of the AI GPUs is much lower in TEM, reducing the installed base.
- The estimated PUE in TEM for small DC is worse (higher) while that in cloud hyperscale PUE is better. However the differences are relatively small and so the impact will be too.
- The estimated PUE for AI hyperscale is slightly better in the LBNL model.

It is difficult to compare TEM compute capacity/efficiency with the LBNL shipment/power estimates as they are very different parameters.

5.5 Summary of comparisons of estimates

The energy projections from the different models are compared in Table 7 and Figure 24.

Table 7 Summary of model estimates of data centre energy use

	2023	2030		
	Base case	**Base case**	**Low**	**High**
Andrae, 2020	375	974	366	
Goldman Sachs, 2024	425	1 150	820	1 400
SemiAnalysis, 2024	450	1 500	500	2 250
TEM 4	323	2 510	-	-

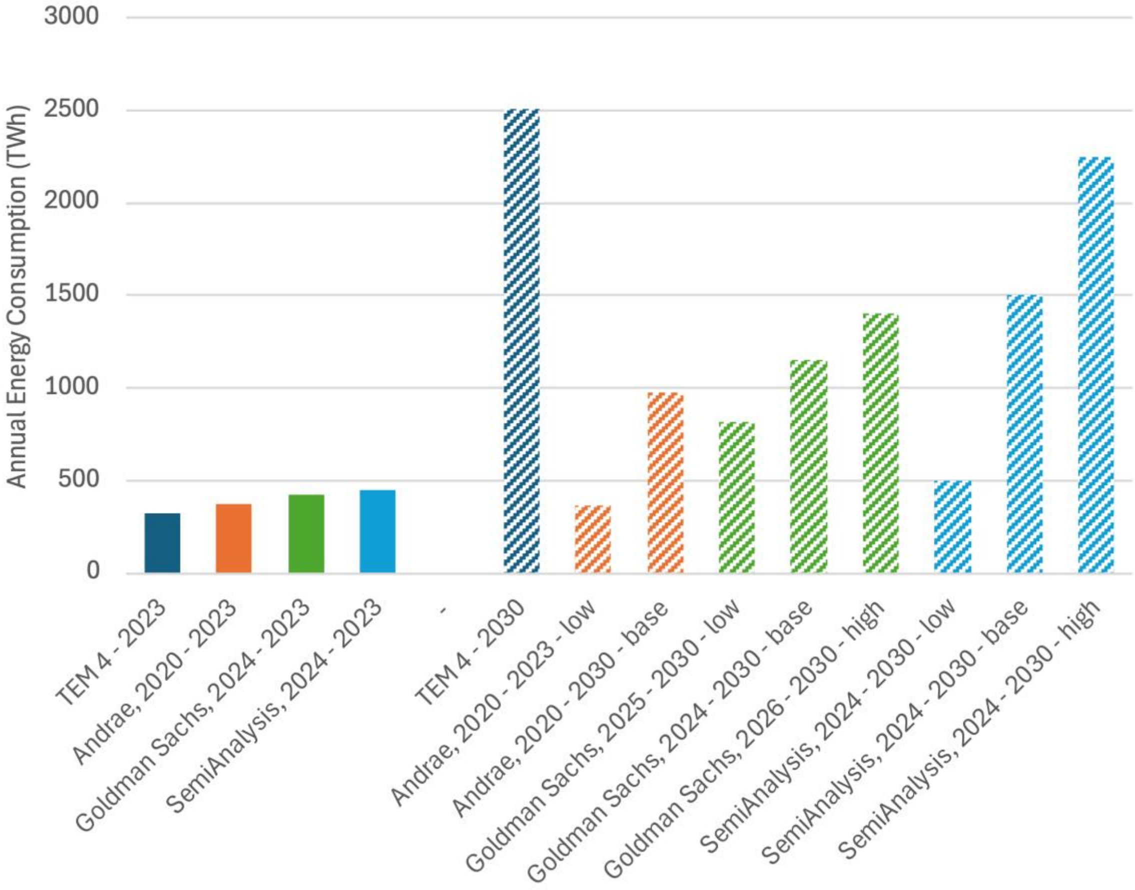


Figure 24 Comparison of model energy projections

6 Sensitivity analysis

From the calculations and comparisons, it is clear that the largest areas of sensitivity are the compute capacity (unconstrained and likely too high) and efficiency curves (potentially too pessimistic), particularly for AI. This is because they are modelled on exponential growth. Comparing the various models[28] (Table 7), the consensus would suggest that the projected energy consumption from the TEM is too high in 2030 (and beyond), due to the AI component. The uncertainty analysis will therefore focus on changes in AI compute capacity and efficiency and try to integrate the inputs from other models. These changes are applied from 2024 onwards.

In addition, the TEM projections of 2023 energy use seems low compared to other models. This may be due to assumptions on AI server lifetimes so this will also be investigated.

Table 8 shows the inputs varied in the sensitivity analysis.

Table 8 Inputs to sensitivity analysis

Parameter	**ICT yr on yr % efficiency improvement**	**AI yr on yr % efficiency improvement**	**AI yr on yr % capacity growth**	**AI lifetime yr**
Base value	18-23%	70%	283%	4.0
Variant used in sensitivity analysis	3%	200%	150%	5.7

6.1 Results

shows the projected energy consumption in 2023 and 2030 using the base assumptions and with the value of the parameter in the sensitivity analysis

Table 9 shows the projected energy consumption in 2023 and 2030 using the base assumptions and with the value of the parameter in the sensitivity analysis

Table 9 Projected energy consumption in base and sensitivity analyses

	2023 base Energy (TWh)	**2023 Energy with variant (TWh)**	**2030 base Energy (TWh)**	**2030 Energy with variant (TWh)**
Lower ICT efficiency improvement	323	323	2 510	2 900
Higher AI eff improvement	323	323	2 510	1 200
Lower AI capacity growth	323	323	2 510	400
Longer AI lifetime	323	350	2 510	3 600

[28] The projections from LBNL are not included in the table as they only extend to 2028.

	2023 base Energy (TWh)	2023 Energy with variant (TWh)	2030 base Energy (TWh)	2030 Energy with variant (TWh)
Combined scenario (all of above)	323	350	2 510	750

Figure 25 shows the alternative scenarios over time.

Figure 25 Energy consumption from sensitivity analysis

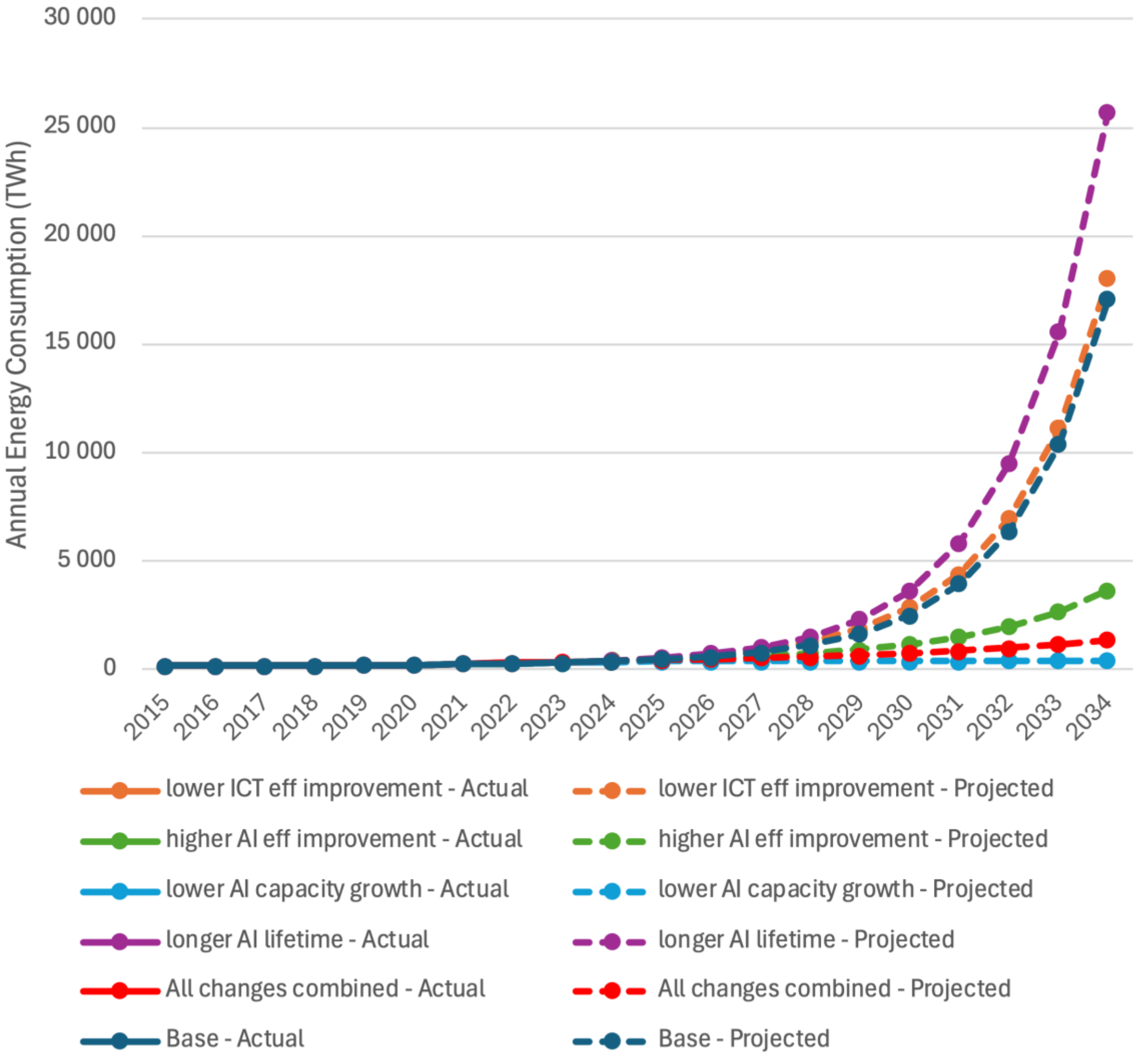
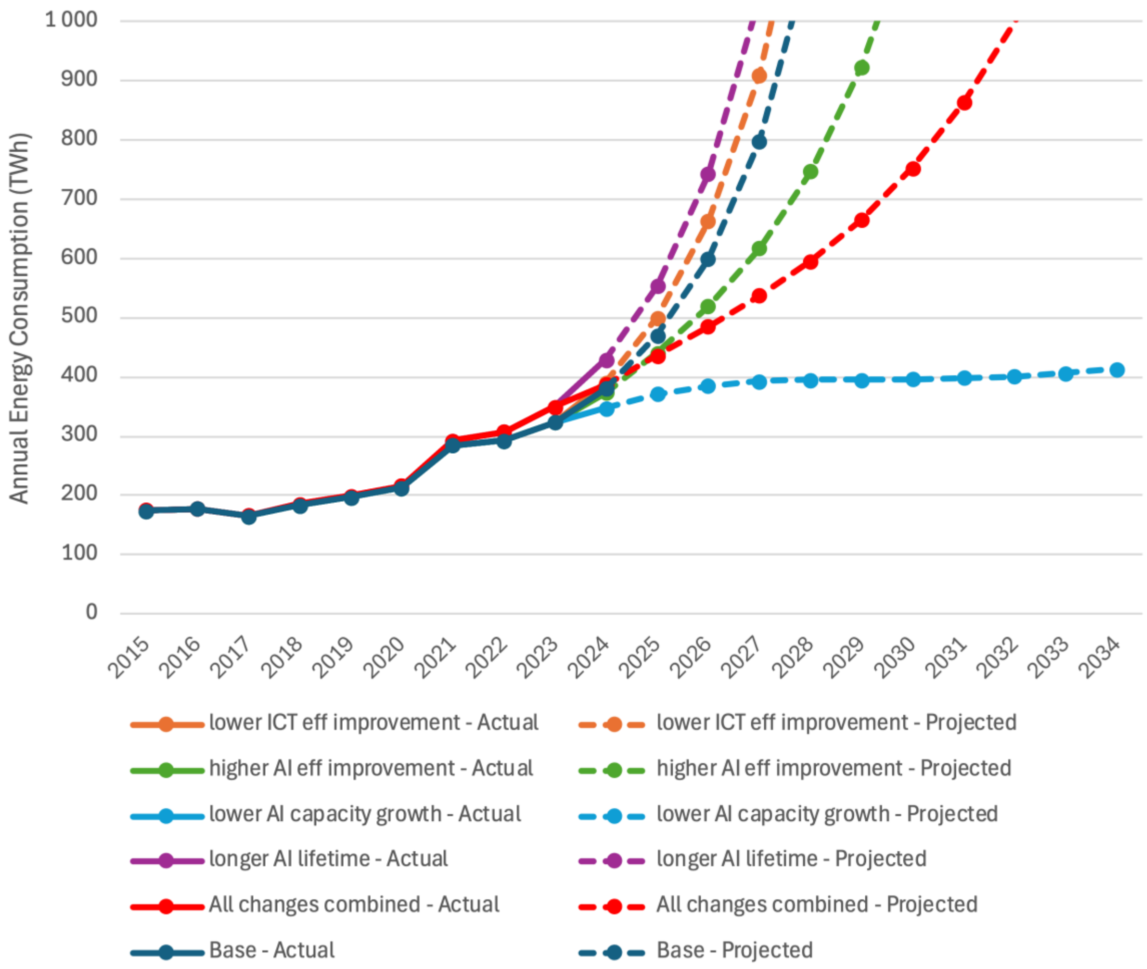


Figure 26 shows a detailed view of the same data.

Figure 26 Energy consumption from sensitivity analysis (detailed view)



6.2 Discussion of results

Figure 27shows the relative sensitivity of the projected energy to each changed parameter. It is clear that the ICT efficiency has a low sensitivity because non-AI energy consumption is a small proportion of the total, while the AI efficiency and capacity have the largest impacts.

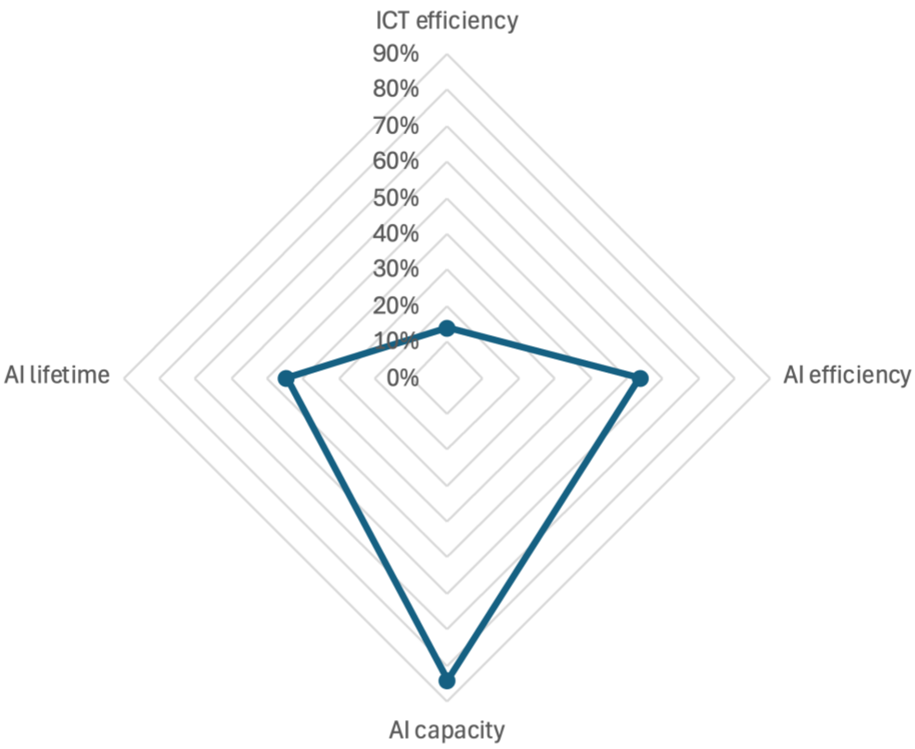


Figure 27 Sensitivity of projected energy to each parameter

Considering the effect of each variant in turn:

Reducing the annual increase in non-AI ICT efficiency improvement from 18 to 23% causes the energy consumption to increase by 390TWh in 2030. This is double the non-AI energy in the base case in this year but a relatively small proportion of total energy since that is mostly AI related.

Changes to the AI lifetime has a linear impact on the energy consumption: an increase of approximately 50% from 4yr to 5.7 yr increases the energy 45% for every year modelled, rather than compounding over time.

The AI efficiency yr on yr improvement could be as high as 1000% per year based on Tschand et al (2024 preprint). A more moderate 200% is used for the sensitivity analysis. This has the second largest impact on the energy consumption by 2030, reducing the energy consumption by 54% to a more moderate figure, although there is still substantial growth in energy use from 2023.

Reducing the AI capacity yr on yr growth to 150% has the single largest impact on the energy consumption, reducing energy by 84% relative to the base case in 2030, similar in scale to the AI efficiency improvement.

The combination scenario combines the effect of all the sensitivity changes results in projected energy consumption of 808 TWh in 2030. The combination of reduced capacity growth and improved efficiency of AI results in a huge reduction of energy. This is slightly offset by the reduced efficiency of the non-AI ICT efficiency.

Comparing the base and alternative scenario suggests that the energy consumption of data centres lies within a very wide range of between 800TWh to 2 400 TWh in 2030.

7 Conclusions

The purpose of the project was to create a new model to model data centre energy efficiency policies, and project the long term energy consumption.

7.1 Policy modelling

The TEM provides a framework on which data centre policies can be modelled. It provides a relatively simple dashboard through which basic policy impacts can be modelled, including changes to server efficiency, market growth rate, lifetimes, utilisation rates and shifting traditional server loads to cloud. Having this ability is very important for assessing policy impacts and cost-benefits analysis.

However, these must be compared against a base case. Long term projections of data centre energy consumption, especially AI, is very hard to do with confidence and accuracy due to the highly uncertain market. In the absence of reliable projections, policy makers must use the model wisely, this may include assessing the credibility and updating the inputs, more extensive sensitivity analysis and quantifying ranges of energy savings and consumptions.

For AI, it may be the case that the sector is too immature and projections are too uncertain for some types of policies. However, non-AI servers can be more accurately modelled and policies already exist. In any case the model is only a tool and it is the responsibility of the policy maker to use it with full knowledge of the limitations, in particular the evidence base available.

7.2 Energy consumption projections.

Projections from the TEM seem to give results in the short term which are comparable with those from other models which helps to validate the modelling approach. The figures in the medium term are unrealistically high due to unconstrained in AI energy use. In reality economic and supply limits will restrict this growth.

It is extremely difficult to predict the growth of AI capacity, efficiency and energy use. Without long term market forecasts, simple extrapolation of short-term forecasts similar to those used for other products produce wildly unrealistic results. Assuming no change in capacity or efficiency is equally unrealistic. This is mostly due to the rapid growth and developments in AI technology at this stage of immaturity in the AI industry coupled with geopolitical conditions.

However, as the AI market matures the quality of projections of parameters data will improve, and the model could become more accurate. The model structure should make it relatively easy to update as new data becomes available, for example through the requirement for data centre reporting under the EU Energy Efficiency Directive[29]. Another example is the new model from EpochAI, GATE[30], published in March 2025 that projects AI computational capacity and efficiency to 2045 (but not energy) using economic modelling, showing how frequently and rapidly new data is becoming available.

Financial/economic models (such as those used by Goldman Sachs and SemiAnalysis) seem to give the most reasonable results and are likely the best way to project future energy consumption in the short term because they are able to take into account market constraints, such as the supply of chips, or access to capital or electricity. However, financial models need extensive research and knowledge and existing models are not suitable for policy modelling or easily accessible to policy makers.

In the short term, we need to find a way to integrate the results of financial modelling in a non-arbitrary way into the model, that is verifying their assumptions and not just adjusting our inputs to get the outputs to match. The sensitivity analysis shows the range of values that could be expected from using the inputs from other models and provides upper and lower bounds to future energy consumption which are in line with other projections.

[29] Article 12 of the recast Energy Efficiency Directive, EU/2023/1791 and COMMISSION DELEGATED REGULATION (EU) 2024/1364 of

[30] https://epoch.ai/gate

Annex A PUE data sources

Cloudscene (2024) *Market Profile Australia*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Australia/all

Cloudscene (2024) *Market Profile Canada*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Canada/all

Cloudscene (2024) *Market Profile France*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-France/all

Cloudscene (2024) *Market Profile Germany*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Germany/all

Cloudscene (2024) *Market Profile India*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-India/all

Cloudscene (2024) *Market Profile Italy*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Italy/all

Cloudscene (2024) *Market Profile Japan*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Japan/all

Cloudscene (2024) *Market Profile Netherlands*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Netherlands/all

Cloudscene (2024) *Market Profile Singapore*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Singapore/all

Cloudscene (2024) *Market Profile South-Africa*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-South-Africa/all

Cloudscene (2024) *Market Profile Spain*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Spain/all

Cloudscene (2024) *Market Profile Sweden*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-Sweden/all

Cloudscene (2024) *Market Profile United-kingdom*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-United-kingdom/all

Cloudscene (2024) *Market Profile USA*. Accessed Nov 2024 https://cloudscene.com/market/data-centers-in-USA/all

中国通服数字基建产业研究院 (2023) *中国数据中心产业发展白皮书 (2023)*

China Communications Services Digital Infrastructure Industry Research Institute (2023) *White Paper on the Development of China's Data Center Industry (2023).* Accessed Nov 2024 https://aimg8.dlssyht.cn/u/551001/ueditor/file/276/551001/1684888884683143.pdf

兰洋科技 (2023) *数据中心能耗现状和能效水平分析*

Lanyang Technologies (2023) *Analysis of current energy consumption and energy efficiency levels in data centres*. Accessed ov 2024 https://www.blueocean-china.net/faq3/234.html

Economic Times of India (2024) *Data centres are wary of sustainability push in govt's AI mission GPU tender*. Accessed Nov 2024 https://economictimes.indiatimes.com/tech/technology/sustainability-push-in-gpu-tender-worries-data-centres/articleshow/114053718.cms?from=mdr

NTT (2024) *Environment Data*. Accessed Nov 2024 https://www.nttdata.com/global/en/about-us/sustainability/esg-data/environment-data

Facebook (2020) *2020 Sustainability Report*. Accessed Nov 2024 sustainability.fb.com/sustainabilityreport2020

Meta (2024) *Sustainability Report*. Accessed Nov 2024 https://sustainability.atmeta.com/2024-sustainability-report/

Google (2024) *Google data center PUE performance*. Accessed Nov 2024 https://datacenters.google/efficiency/

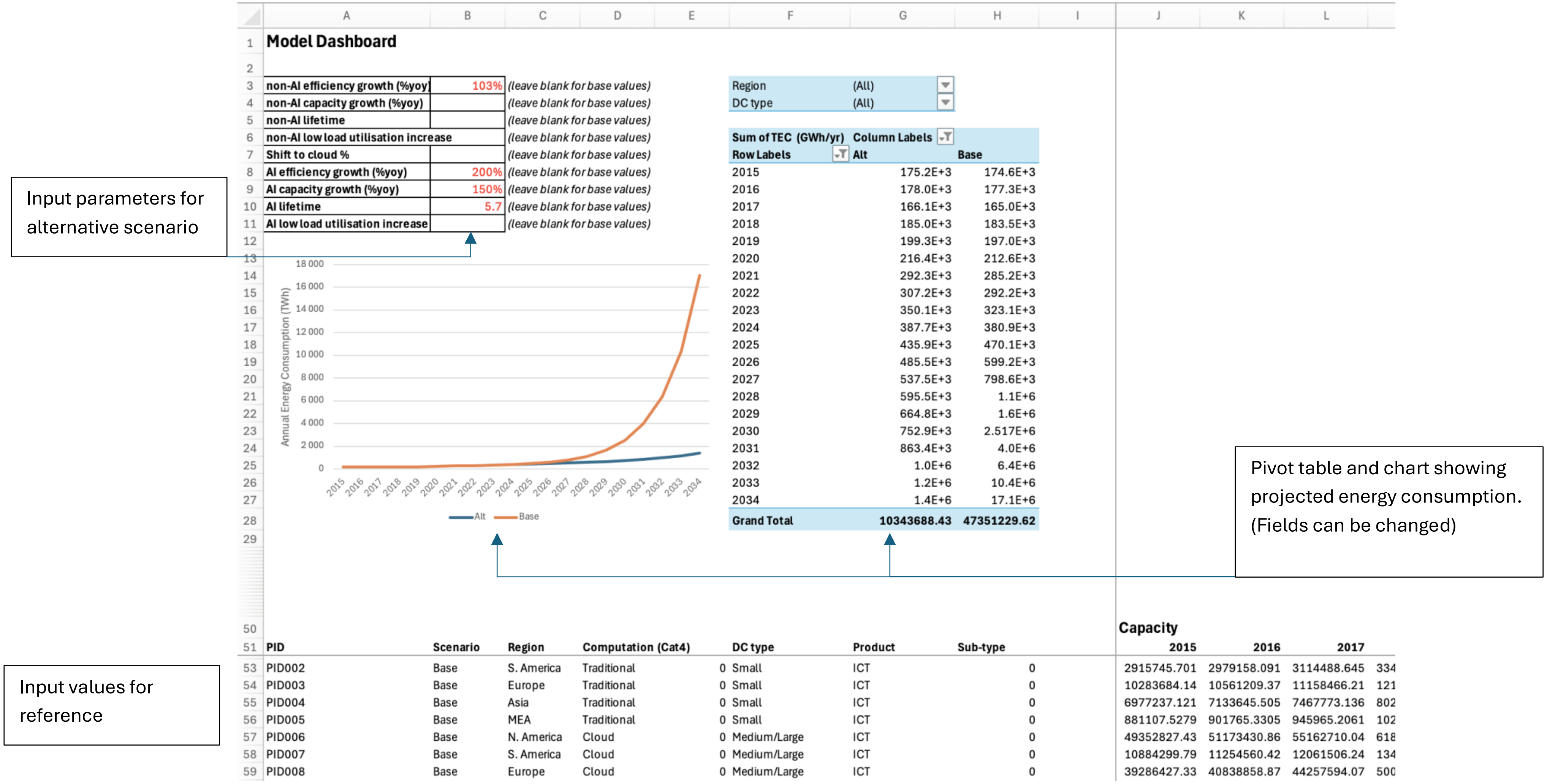
Walsh, N (2022) *How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability*. Microsoft. Accessed Nov 2024 https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/ 1/4

Uptime Institute (2024) *Uptime Institute Global Data Center Survey 2024 Executive Summary*. Accessed Nov 2024 https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2024

Supermicro (2022) *Data centers & the environment 2021 report on the state of the green data center*. Accessed Nov 2024 https://www.supermicro.com/white_paper/DataCenters_and_theEnvironmentFeb2021.pdf

EU Code of Conduct for data centres (unpublished). Accessed Nov 2024

Annex B Model Dashboard



Model Dashboard

3	non-AI efficiency growth (%yoy)	103%	*(leave blank for base values)*
4	non-AI capacity growth (%yoy)		*(leave blank for base values)*
5	non-AI lifetime		*(leave blank for base values)*
6	non-AI low load utilisation increase		*(leave blank for base values)*
7	Shift to cloud %		*(leave blank for base values)*
8	AI efficiency growth (%yoy)	200%	*(leave blank for base values)*
9	AI capacity growth (%yoy)	150%	*(leave blank for base values)*
10	AI lifetime	5.7	*(leave blank for base values)*
11	AI low load utilisation increase		*(leave blank for base values)*

Region	(All)
DC type	(All)

Sum of TEC (GWh/yr)	Column Labels	
Row Labels	Alt	Base
2015	175.2E+3	174.6E+3
2016	178.0E+3	177.3E+3
2017	166.1E+3	165.0E+3
2018	185.0E+3	183.5E+3
2019	199.3E+3	197.0E+3
2020	216.4E+3	212.6E+3
2021	292.3E+3	285.2E+3
2022	307.2E+3	292.2E+3
2023	350.1E+3	323.1E+3
2024	387.7E+3	380.9E+3
2025	435.9E+3	470.1E+3
2026	485.5E+3	599.2E+3
2027	537.5E+3	798.6E+3
2028	595.5E+3	1.1E+6
2029	664.8E+3	1.6E+6
2030	752.9E+3	2.517E+6
2031	863.4E+3	4.0E+6
2032	1.0E+6	6.4E+6
2033	1.2E+6	10.4E+6
2034	1.4E+6	17.1E+6
Grand Total	**10343688.43**	**47351229.62**

Capacity

	PID	Scenario	Region	Computation (Cat4)		DC type	Product	Sub-type	2015	2016	2017	
53	PID002	Base	S. America	Traditional	0	Small	ICT	0	2915745.701	2979158.091	3114488.645	334
54	PID003	Base	Europe	Traditional	0	Small	ICT	0	10283684.14	10561209.37	11158466.21	121
55	PID004	Base	Asia	Traditional	0	Small	ICT	0	6977237.121	7133645.505	7467773.136	802
56	PID005	Base	MEA	Traditional	0	Small	ICT	0	881107.5279	901765.3305	945965.2061	102
57	PID006	Base	N. America	Cloud	0	Medium/Large	ICT	0	49352827.43	51173430.86	55162710.04	618
58	PID007	Base	S. America	Cloud	0	Medium/Large	ICT	0	10884299.79	11254560.42	12061506.24	134
59	PID008	Base	Europe	Cloud	0	Medium/Large	ICT	0	39286427.33	40838858.87	44257594.07	500

Input parameters for alternative scenario

Pivot table and chart showing projected energy consumption. (Fields can be changed)

Input values for reference