Feeding the Machine: Energy Demand of Data Centres and Artificial Intelligence

Onur Özgün | Erica McConnell | Sujeetha Selvakkumaran | Takuma Ono[1]

DNV - Veritasveien 1, 1363 Høvik, Norway

Abstract

The rapid expansion of AI and cloud computing is transforming data centres into one of the fastest-growing energy consumers globally. While past digital infrastructure growth has been moderated by efficiency gains, the computational intensity of AI may disrupt this pattern. This study examines how the interplay between technological progress, cost dynamics, and infrastructure constraints shapes the long-term trajectory of data centre expansion and energy consumption.

A system dynamics model is developed to capture the feedback loops governing data centre growth, semiconductor scaling, and grid constraints. The model tracks two types of data centres—AI and General Computing—and simulates how shifts in Process Node Size, Compute Power Efficiency, and Infrastructure Scaling influence energy demand, pricing, and market size over time.

Findings reveal a structural shift in data centre composition, with AI workloads surpassing general computing in the 2030s. While efficiency gains partially offset rising energy demand, they are insufficient to prevent an overall increase in power consumption. Grid constraints could become a bottleneck, limiting data centre expansion. Market-driven technological improvements drive the transition to Leading Edge Process Nodes, but cost reductions slow as fabrication reaches physical limits.

These results highlight the complex interaction between technology scaling, energy infrastructure, and AI adoption. The study suggests that AI-driven computing will not cause unchecked energy demand growth but will reshape how and where energy is consumed. Key uncertainties include semiconductor advancements, grid expansion rates, and AI efficiency improvements. The findings are relevant for policymakers, energy planners, and AI developers seeking to balance computational needs with sustainable energy growth.

Keywords: AI, energy, data centres

1 Introduction

When large language models and other generative artificial intelligence (AI) applications burst into the mainstream within mere months, discussions quickly turned to a growing concern: the potentially immense energy demand from the data centres that train and operate these models. Headlines warned of exponential electricity consumption and raised the spectre of AI-induced strain on national electricity grids (Goldman Sachs, 2024; SemiAnalysis, 2024; Deloitte, 2024). Yet, this is hardly the first time new digital technologies have provoked such alarming forecasts about their energy appetite, only for reality to fall short of dire predictions.

During the dot-com boom of the late 1990s, as personal computing and internet connectivity expanded rapidly, predictions of runaway electricity use were commonplace. One notorious example was a Forbes article by Huber & Mills (1999), suggesting that the Internet and personal computers would consume nearly half of the US electricity supply within a decade. Their forecast caused widespread concern, depicting a scenario in which rapid digital growth would drive a relentless need for coal-fired electricity (Lovins, 1999). In hindsight, these claims turned out to be grossly

[1] Özgün, Onur Onur.Ozgun@dnv.com | McConnell, Erica Erica.Mcconnell@dnv.com | Selvakkumaran, Sujeetha Sujeetha.Selvakkumaran@dnv.com | Ono, Takuma Tak.Ono@dnv.com

exaggerated. Later analyses demonstrated that the combined energy consumption of all computing and internet infrastructures stayed well below these catastrophic forecasts, only accounting for a few per cent of total electricity usage by the mid-2000s (Koomey, 2011).

A similar story unfolded as cloud services and global data centre infrastructure expanded through the early 2000s. Media reports and academic predictions issued warnings of unsustainable growth in electricity demand. In 2015, researchers predicted data centre electricity consumption would triple within a decade, potentially undermining global climate and energy goals (Bawden, 2016; Andrae & Edler, 2015). Yet again, actual data centre electricity use remained relatively modest—hovering around 1–2% of global electricity supply through the 2010s (Masanet et al., 2020). Historical trends show that although data centre electricity demand doubled globally between 2000 and 2005, its growth rate moderated significantly thereafter, increasing by about 56% between 2005 and 2010 and only marginally thereafter (Shehabi et al., 2016).

The exaggerated predictions of the past were largely due to linear thinking or so-called "bottom-up" estimation approaches, which extrapolated historical growth rates without adequately accounting for systemic innovation and efficiency improvements (Koomey, 2011; Masanet et al., 2020). Indeed, history consistently demonstrates that when digital technologies scale, efficiency innovations soon follow. The mid-2000s saw widespread adoption of server virtualisation, dramatically improving resource utilisation and reducing total server counts (Masanet et al., 2020). Data centre cooling and infrastructure efficiency improved significantly, reflected in improving metrics such as Power Usage Effectiveness (PUE), reducing overhead power requirements. Additionally, the industry transitioned from inefficient small-scale enterprise facilities to centralised hyperscale cloud providers with a strong economic incentive to minimise operating costs through greater efficiency (Masanet et al., 2020; Shehabi et al., 2016).

Despite these historical lessons, concerns resurfaced with the explosive growth of cryptocurrencies like Bitcoin in the 2010s. Extrapolations projected Bitcoin mining could soon consume as much electricity as entire countries or even the global electricity supply if its growth trajectory continued unchecked (Jezard, 2017). Yet Bitcoin mining energy use has remained comparatively small—below 0.5% of global electricity consumption in recent years (IEA, 2019)—due to market saturation, technological limits, and efficiency gains driven by high operating costs.

Is AI Different This Time?

With AI applications now quickly permeating everyday life, a familiar cycle of fear regarding digital energy demand repeats itself. Yet this raises a critical research question: is AI fundamentally different? Could AI workloads finally drive computing infrastructure to consume an increasingly large and potentially unsustainable share of global electricity?

Various forecasts highlight the uncertainty inherent in these predictions, with differing assumptions about technological advancement, infrastructure growth, and adoption rates. The International Energy Agency (2024) estimates global data centres currently consume approximately 1% of electricity, predicting moderate growth in the near term. However, in the US, the Federal Energy Regulatory Commission forecasts data centre electricity use nearly doubling from 19 GW to 35 GW by 2030, accounting for up to 9% of national electricity (FERC, 2024). Consultants also provide diverse perspectives: Boston Consulting Group (2024) and Goldman Sachs (2024) predict annual growth in US data centre energy demand of around 15–20%, potentially reaching approximately 8% of US electricity consumption by 2030. McKinsey & Company (2024), considering aggressive AI workloads, expects even faster growth, with AI-ready data centres potentially consuming 70% of total data centre capacity by decade's end.

Globally, the expansion of AI infrastructure appears even more pronounced. SemiAnalysis (2024) anticipates global data centre power usage nearly doubling from 49 GW in 2023 to 96 GW by 2026, with almost half of this growth attributable to AI workloads. Deloitte (2024) similarly projects global data centre electricity demand rising significantly, from 536 TWh in 2025 to 1,065 TWh by 2030, primarily driven by AI computing. IDC (2024) forecasts global data centre energy use more than doubling between 2023 and 2028, with AI as a significant contributor.

To clearly frame the discussion and establish a coherent "reference mode" for this system dynamics analysis—an essential baseline of past and predicted future behaviour over time—electricity use expressed as a percentage of total demand provides a consistent, scalable metric. Current estimates place global data centre energy use at around 2–3%, with forecasts ranging widely from 4.5% (SemiAnalysis, 2024) to over 12% (McKinsey & Company, 2024) by 2030. AI-specific forecasts exhibit even higher uncertainty, with Wells Fargo (2024) suggesting AI workloads alone could consume up to 5% of global electricity by 2030.

Insights from System Dynamics Literature

The system dynamics literature on data centre and AI electricity demand remains relatively limited, though several valuable studies exist. Dianati (2012), one of the earliest applications, used system dynamics for a Norwegian cloud computing firm's capacity planning. While insightful regarding energy consumption of individual data centre components, the study primarily emphasised operational details rather than overall growth dynamics or systemic efficiency improvements, leaving longer-term forecasts unexplored.

Spirova (2021) applied system dynamics to data centre electricity demand in the context of smart industry developments in the Netherlands. This study explicitly linked regional industrial growth and data centre energy requirements, predicting possible electricity shortages by the late 2020s—a strength in connecting systemic demand growth to broader infrastructure constraints, yet less applicable in global AI contexts.

More directly relevant, Wijnhoven and Koot (2021) modelled global data centre energy demands, clearly demonstrating that efficiency gains could partially but not fully offset rising electricity needs under certain scenarios. However, their analysis predated the explosive growth of generative AI, a crucial limitation in informing present dynamics.

Most recently, Paccou and Wijnhoven (2024) explicitly modelled AI's electricity impact through four distinct scenarios to 2035. Their model highlights critical factors influencing future energy use, such as public concerns, grid limitations, hardware efficiency improvements, and algorithmic advancements. The scenario-based approach provides a robust foundation for exploring diverse future outcomes, although specific assumptions on technological and social responses remain uncertain.

Research Question: Will AI Defy Historical Trends?

Historically, dire forecasts for digital energy demand repeatedly failed to materialise due to innovation-driven efficiency gains. Yet, the rapid and pervasive nature of AI might test these historical lessons. Is AI expected to push computing infrastructures into consuming unprecedented shares of global electricity?

This paper contributes to this ongoing debate by developing a comprehensive system dynamics model, explicitly capturing key feedback mechanisms driving AI and data centre energy dynamics. It builds upon and extends previous literature, explicitly addressing limitations in past modelling studies and integrating recent empirical insights to explore AI's energy demand trajectory.

2 Model

System Boundary

The model centres on three fundamental components driving the energy demand of data centres: (1) the demand for data centre services, (2) the efficiency of delivering these services, and (3) the constraints limiting the supply of data centres. Each of these components interacts dynamically to shape the overall energy consumption trajectory of data centres, but the model must also define clear boundaries to distinguish endogenous factors from external influences.

At the core of data centre energy demand is the **demand for data centre services**, which serves as the primary driver of electricity consumption. This demand is influenced by both endogenous and exogenous factors. Endogenously, data centre service prices, computing capabilities, and efficiency improvements feed back into demand, as more efficient and cost-effective data processing capabilities tend to spur greater usage. However, exogenous factors such as global Internet penetration, network infrastructure, and broadband speed improvements also play a significant role in shaping demand. These external factors enable more widespread use of digital services, but since they operate at a broader technological and policy level, they are outside the boundary of this system dynamics model. For example, the proliferation of high-speed 5G networks and edge computing can increase the need for data centre capacity, but their adoption is driven by telecommunications advancements that are not modelled explicitly here.

The second core component is the **energy efficiency of data centres**, which determines how much computational work can be performed per unit of electricity. Historically, efficiency improvements have been a counterbalancing force against rising demand. Advances in chip architecture, server utilisation, and cooling technologies have allowed data centres to process exponentially more data without proportional increases in power consumption (Masanet et al., 2020). A crucial historical driver of efficiency gains has been Moore's Law, which has enabled the doubling of transistor density roughly every two years, resulting in greater computational performance per watt. However, as the semiconductor industry faces physical limitations on transistor miniaturisation, the rate of efficiency gains from chip improvements is expected to slow. Instead, efficiency gains are now increasingly dependent on software-based optimisations, such as model compression and algorithmic efficiency, as well as specialised AI hardware accelerators, such as NVIDIA's AI-specific GPUs or Google's Tensor Processing Units (TPUs) (McKinsey & Company, 2024; SemiAnalysis, 2024). These efficiency dynamics are explicitly modelled endogenously, allowing the system to capture how improvements in computation per watt influence long-term energy demand trends.

The third major factor shaping data centre energy demand is the **constraints on data centre expansion**, which determine how quickly the industry can scale to meet growing demand. The model endogenously incorporates two key constraints: chip production capacity and power grid availability.

- **Chip production capacity** has historically expanded in tandem with efficiency improvements, ensuring a steady supply of more powerful processors. However, recent industry trends suggest that semiconductor manufacturing faces supply chain limitations, geopolitical risks, and long lead times for building new fabrication plants. These constraints introduce potential bottlenecks to computational scaling, which we model explicitly.
- **Power grid constraints** pose a more significant challenge. Unlike chip production, which is driven primarily by market forces, grid expansion is influenced by a mix of regulatory policies, renewable energy integration, and local infrastructure limitations (IEA, 2024). As data centres increasingly cluster in certain energy-constrained regions (such as Virginia in the U.S. or Dublin

in Ireland), electricity supply bottlenecks become a binding constraint on growth. While our model includes grid availability as a limiting factor, broader external influences—such as renewable deployment, transmission infrastructure buildout, and government regulations—are outside the model boundary.

Key demand-side variables have been regionalized such as market size and willingness to pay for services. Therefore, the model calculates the regional variability in demand for data centre services; however, the model currently does not regionalize the supply side dynamics, such as chip production. As a result, it does not capture localised efficiency disparities, such as the potential for regions with limited chip access to lag in energy efficiency improvements. Future iterations of the model could incorporate regional dynamics, particularly given the impact of semiconductor geopolitics (e.g., U.S.-China chip trade restrictions).

Model diagram and description

The model runs from 1980 to 2060, and most key variables are arrayed into two data centre types: AI and general purpose. **General purpose** focus on data services and cloud computing, such as data storage, retrieval, content delivery, enterprise applications, web hosting, and software as a service (SaaS). The **AI** array, on the other hand, involves training machine learning models through intensive iterative computations, requiring prolonged, high-performance compute power and large-scale data handling as well as their deployment, which is the inference stage. AI inference requires trained models to process new data in real-time or near real-time, requiring rapid response times and optimised hardware for efficiency. Unlike training, inference is often latency-sensitive and operates on distributed infrastructure to handle diverse end-user interactions.

We model **chip production capacity** across four generations of semiconductor technology, categorized by process node type: legacy, mainstream, advanced, and leading edge. These categories reflect the level of transistor miniaturization and fabrication complexity, with each newer generation offering higher transistor density, improved energy efficiency, and greater computational performance.

A newer process node generation represents a significant technological leap, enabling smaller, more power-efficient transistors that enhance chip performance while reducing power consumption per computation. The shift from one generation to the next is not just a matter of size reduction—it

involves fundamental changes in materials, lithography techniques (e.g., EUV for sub-7nm nodes), and transistor architectures (e.g., FinFET, GAAFET).

We define process node categories with the following reference sizes:

- **Legacy**: ≥28nm (older nodes used in mature applications such as microcontrollers and analog chips).
- **Mainstream**: 14-22nm (widely used in general-purpose computing and cloud CPUs).
- **Advanced**: 5-10nm (high-performance AI accelerators, GPUs, and data centre processors).
- **Leading Edge**: ≤3nm (cutting-edge chips for AI training, mobile processors, and HPC applications).

Within each process node category, we allow the average Process Node Size to continuously decrease over time, reflecting two distinct phases of technological progression:

1. R&D-Driven Reduction (Pre-Commercialization Stage): Before large-scale deployment, the process node shrinks as a function of time, driven by research advancements, breakthroughs in photolithography, and material innovations.
2. Experience-Driven Reduction (Learning-by-Doing Stage): Once a node enters mass production, its size further reduces with accumulated manufacturing experience, as higher wafer volumes lead to process refinements, yield improvements, and cost reductions.

This dynamic structure captures both planned technological roadmaps and emergent efficiency gains, ensuring a realistic representation of how semiconductor manufacturing evolves over time.

Figure 1 shows a causal loop diagram of the model. For the sake of simplicity, arrays, unit conversions (such as GWh to TWh, seconds per year), and certain exogenous variables have been omitted. All costs are in USD, and we measure data centre computation capacity in either tera or giga FLOP/s: floating point operations per second, representing the number of calculations a computer can execute per second. In the following model description, all **stocks, flows, and variables** are in **bold** to enable the reader to follow the causal loop diagram.

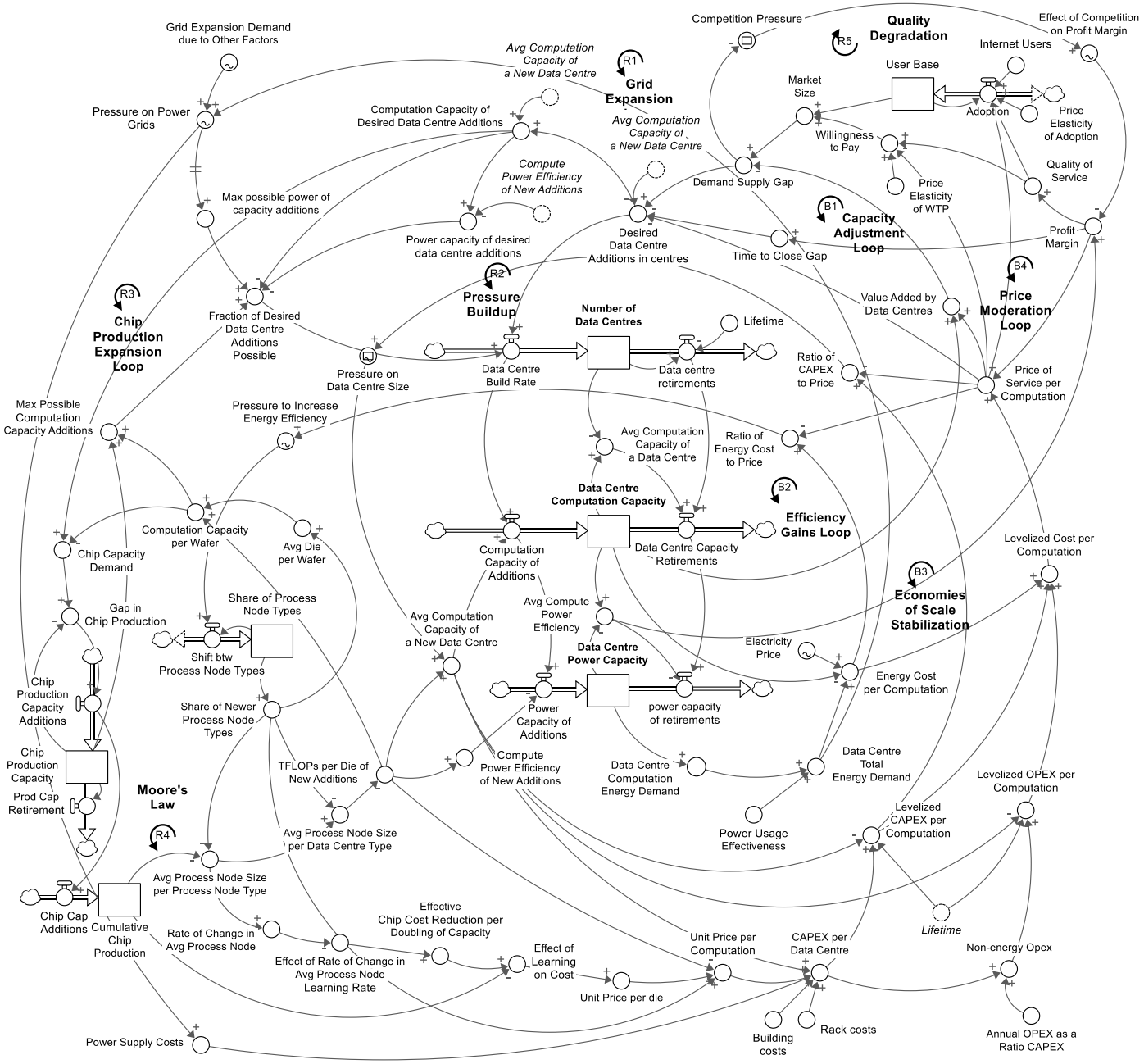


Figure 1. Causal loop diagram

At the heart of the model is a system that tracks the expansion of digital infrastructure by keeping a simultaneous record of **Number of Data Centres** (centres), **Data Centre Computation Capacity** (TFLOP/s), and **Data Centre Power Capacity** (GW). These three stocks evolve together, ensuring that when data centres are built, expanded, or retired, their computational and energy footprint are reflected accordingly.

Growth in data centres does not happen in isolation but is driven by the demand for computing services. The process begins with **Desired Data Centre Additions in Centres**, which represents the theoretical number of data centres that could be built based on user demand and industry trends. However, physical and economic constraints—most notably the availability of power infrastructure and semiconductor supply—limit how many of these desired additions can actually be realized. The **Fraction of Desired Data Centre Additions Possible** accounts for these constraints, restricting the flow into **Data Centre Build Rate** and thereby moderating expansion.

The characteristics of newly built data centres are not static. Over time, both their computing power and efficiency evolve. The model accounts for this by tracking **Avg Computation Capacity of a Data Centre** and **Compute Power Efficiency**, ensuring that newer facilities contribute more computation per unit of energy consumed. Two key pressures shape this process: **Pressure on Data Centre Size**, which pushes for higher **Avg Computation Capacity of a New Data Centre**, and **Pressure to Increase Energy Efficiency**, which drives up **Compute Power Efficiency of New Additions** (GFLOP/s/W). These improvements stem from advances in hardware, software optimization, and operational design.

The expansion of data centres brings with it a growing energy footprint. **Data Centre Total Energy Demand** is calculated by considering both direct power consumption and additional overhead requirements, captured through **Power Usage Effectiveness**. By multiplying **Data Centre Total Energy Demand** with electricity prices from the broader energy transition model, and applying the necessary unit conversions, we determine **Energy Cost per Computation** (USD/TFLOP). This is a crucial component of the cost structure, feeding into the **Levelized Cost per Computation**, which includes both capital expenditures (**CAPEX**) and operational expenditures (**OPEX**). **CAPEX** consists of **Chip Price**, **Building Costs**, **Rack Costs**, and **Power Supply Costs**, while **OPEX** is set as a ratio of **CAPEX**.

From **Levelized Cost per Computation**, we determine the **Price of Service per Computation**, which, after incorporating **profit margins**, defines the revenue potential of the industry. This revenue is captured as **Value Added by Data Centres** (USD/yr), which, in turn, feeds back into **Desired Data Centre Additions in Centres**, sustaining the cycle of expansion. As the industry scales, it becomes more sensitive to cost pressures. Two ratios—**Ratio of CAPEX to Price** and **Ratio of Energy Cost to Price**—determine whether the cost burden is becoming too high relative to revenue. If these ratios rise excessively, they reinforce **Pressure on Data Centre Size**, encouraging facilities to scale up, and **Pressure to Increase Energy Efficiency**, accelerating innovation in hardware and infrastructure.

Adoption of computing services is modelled through the interaction of **User Base** and **Adoption Rate**, which are driven by exogenous trends in **Internet Users**, **Price Elasticity of Adoption**, **Quality of Service**, and **Price of Service**. Together, these factors determine the **Market Size**, which then influences the **Demand Supply Gap**, signalling whether additional capacity is needed.

The semiconductor industry plays a pivotal role in determining the pace at which data centres can scale. The model tracks **Chip Production Capacity** and **Cumulative Chip Production**, both of which influence how many new facilities can be brought online. Beyond availability, **chip production dynamics also shape cost trends**, as the cost per processor die follows a learning curve. **Unit Price per Die** declines with cumulative output due to **Effect of Learning Rate**, which captures the reduction in cost per doubling of cumulative production capacity.

A key driver of the computational efficiency of data centres is **Process Node Size**, which determines how many transistors can be packed into a given area of silicon. As the semiconductor industry transitions from **Legacy (≥28nm)** to **Mainstream (14-22nm)** to **Advanced (5-10nm)** to **Leading Edge (≤3nm)** nodes, chips become more power-efficient and computationally powerful. Within each process node category, continuous refinement occurs, initially as a function of research and development and later through experience-driven manufacturing improvements.

Smaller transistors result in lower power consumption per calculation, which we capture through **Power per Die** (W/die). This relationship follows a scaling law:

$$P_{die} = P_0 \times \left(\frac{S_0}{S}\right)^{\alpha}$$

where $\mathbf{P_o}$ is a reference power level, $\mathbf{S_o}$ is a reference process node size, **S** is the current process node size, and **α** is an empirically determined exponent. Because power consumption scales down with process node size, **Compute Power Efficiency** (GFLOP/s/W) follows an inverse relation:

$$\text{Compute Power Efficiency} = \frac{\text{TFLOP/s per Die}}{\text{P}_{\text{die}}}$$

Process miniaturization also affects cost. The **Unit Price per Die** follows a learning curve, declining as cumulative production increases:

$$C_{die} = C_0 \times \left(\frac{Q}{Q_0}\right)^{-\lambda}$$

where **Q** is cumulative production, $\mathbf{C_o}$ is the initial cost, and **λ** is the learning rate exponent. However, as process nodes shrink to the extreme sub-5nm range, fundamental physical constraints limit further reductions in cost and efficiency gains.

The same shrinking transistor size that improves efficiency also enhances computational performance. The relationship between process node size and **TFLOP/s per Die** follows:

$$\text{TFLOP/s per Die} = \beta \times \left(\frac{S_{ref}}{S}\right)^{n}$$

where **β** is a scaling factor and **n** is typically between 1.7 and 2.0. However, at the smallest scales, material physics and manufacturing precision impose practical limitations.

Two constraints govern the maximum feasible expansion of data centres. **Grid Power Availability** places a hard ceiling on growth, as data centres require guaranteed access to power before they can be built. In parallel, **Chip Production Capacity** determines how quickly computational capability can expand. The fraction of desired data centre additions that can actually be realized is given by:

$$\text{Fraction of Desired Data Centre Additions Possible} = \min\left(\frac{\text{Available Grid Power}}{\text{Power Required}}, \frac{\text{Available Chips}}{\text{Chip Demand}}\right)$$

Competition shapes both **Profit Margins** and the **speed at which the Demand Supply Gap is closed**. As competition intensifies, margins shrink, limiting reinvestment into new data centres. Lower margins also pressure pricing, influencing **Price of Service per Computation** and in turn affecting **User Base Growth**. Over time, as constraints on power and chip supply become more pronounced, data centre expansion is expected to transition from an early phase of rapid acceleration to a more moderated growth trajectory, much like the historical evolution of global electricity demand—where an initial period of exponential expansion slowed as structural constraints took hold.

This formulation ensures that the model captures both the technological acceleration of the semiconductor industry and the longer-term constraints that shape the evolution of data centre capacity.

Major Feedback Loops

There are **nine key feedback loops** in the model: five **reinforcing** and four **balancing**, each playing a role in shaping data centre expansion, efficiency improvements, competition, and cost dynamics.

Reinforcing Loops

The first reinforcing loop, **grid expansion**, captures how rising energy demand from data centres drives investment in grid capacity, enabling further expansion. As **Data Centre Total Energy Demand** increases, it places **Pressure on Power Grids**, prompting investments that, after some delay, raise

Max Possible Power of Capacity Additions. This allows a greater **Fraction of Desired Data Centre Additions Possible**, leading to new data centres coming online. These additions increase **Data Centre Computation Energy Demand**, further amplifying **Data Centre Total Energy Demand**, reinforcing the loop.

The **pressure buildup** loop describes how constraints on expansion exacerbate unmet demand. When **Desired Data Centre Additions in Centres** rise, they bring additional **Computation Capacity of Desired Data Centre Additions** and **Power Capacity of Desired Data Centre Additions**, which, in turn, reduce the **Fraction of Desired Data Centre Additions Possible** due to supply limitations. A lower fraction restricts **Data Centre Build Rate**, slowing **Computation Capacity Additions**, **Data Centre Computation Capacity**, and **Value Added by Data Centres**. As **Value Added by Data Centres** drops, the **Demand Supply Gap** widens, which further increases **Desired Data Centre Additions in Centres**, creating a self-reinforcing buildup of unmet demand.

The **chip production expansion** loop highlights how demand for computing power fuels semiconductor manufacturing. As **Computation Capacity of Desired Data Centre Additions** grows, it pushes up **Chip Capacity Demand**, leading to increased **Chip Production Capacity** and **Cumulative Chip Production** over time. Higher cumulative production reduces **Unit Price per Die**, lowering **CAPEX**, **Levelized Cost per Computation**, and ultimately **Price of Service per Computation**. A lower **Price of Service per Computation** makes computing more accessible, driving further **Desired Data Centre Additions in Centres**, which increases **Computation Capacity of Desired Data Centre Additions**, reinforcing chip demand.

The **Moore's Law** loop explains how improvements in semiconductor technology reduce energy demand while making computation more affordable, driving even more investment into chip production. As **Avg Process Node Size per Process Node Type** decreases, **TFLOP/s per Die of New Additions** increases, leading to higher **Avg Compute Power Efficiency** across data centres. This efficiency gain boosts **Profit Margin**, allowing better **Quality of Service** and increasing **Desired Data Centre Additions in Centres**. More data centre additions raise **Computation Capacity of Data Centre Additions**, leading to greater **Chip Capacity Demand**. As semiconductor manufacturers expand **Chip Production Capacity**, they drive further reductions in **Avg Process Node Size per Process Node Type**, sustaining the loop.

The **quality degradation** loop captures how intense competition erodes profit margins, leading to a decline in service quality. When there is **Competition Pressure** due to oversupply of computing services, **Profit Margin** shrinks, forcing providers to cut costs. This results in lower **Quality of Service**, which reduces **Willingness to Pay**, shrinking **Market Size** and widening the **Demand Supply Gap**. A larger gap intensifies **Competition Pressure**, reinforcing the downward cycle.

Balancing Loops

The **capacity adjustment** loop stabilizes expansion by limiting new data centre additions as demand is met. As **Desired Data Centre Additions in Centres** rise, they increase **Data Centre Build Rate**, expanding **Data Centre Computation Capacity** through the co-flow structure. This enhances **Value Added by Data Centres**, shrinking the **Demand Supply Gap**. A narrower gap reduces **Desired Data Centre Additions in Centres**, moderating growth and balancing the loop.

The **efficiency gains** loop drives improvements in energy efficiency as energy costs become a greater burden. When the **Ratio of Energy Cost to Price** rises, data centre operators face stronger **Pressure to Increase Energy Efficiency**, accelerating the adoption of smaller transistors in **Share of Process Node Type in New Data Centres**. This lowers **Avg Process Node Size per Data Centre Type**, improving **Compute Power Efficiency of New Additions**. More efficient additions lower **Power Capacity of Additions**, reducing **Data Centre Power Capacity**, **Data Centre Total Energy Demand**,

and ultimately **Energy Cost per Computation**. This decreases the **Ratio of Energy Cost to Price**, closing the loop.

The **economies of scale stabilization** loop limits excessive growth in data centre size. When the **Ratio of CAPEX to Price** climbs, data centre operators aim to cut costs by increasing **Avg Computation Capacity of a New Data Centre**, allowing them to achieve more computing power per facility. This lowers **CAPEX per TFLOP**, reducing the **Ratio of CAPEX to Price** and preventing unchecked size expansion.

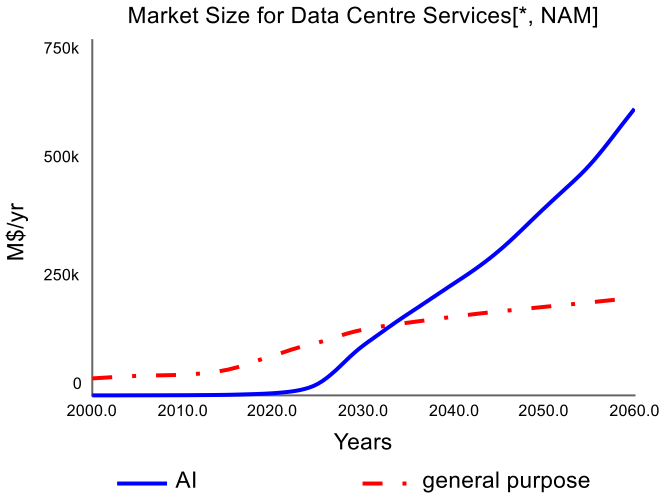
The **price moderation** loop regulates pricing dynamics in response to competition and demand shifts. As **Price of Service per Computation** falls, **Willingness to Pay** improves, leading to a larger **Market Size** and a shrinking **Demand Supply Gap**. As demand increases relative to supply, **Competition Pressure** eases, allowing for higher **Profit Margins**, which in turn supports a gradual rise in **Price of Service per Computation**, stabilizing the cycle.

These nine feedback loops together define the evolving interplay between technological progress, market forces, resource constraints, and economic pressures, shaping the trajectory of data centre expansion and efficiency gains over time.

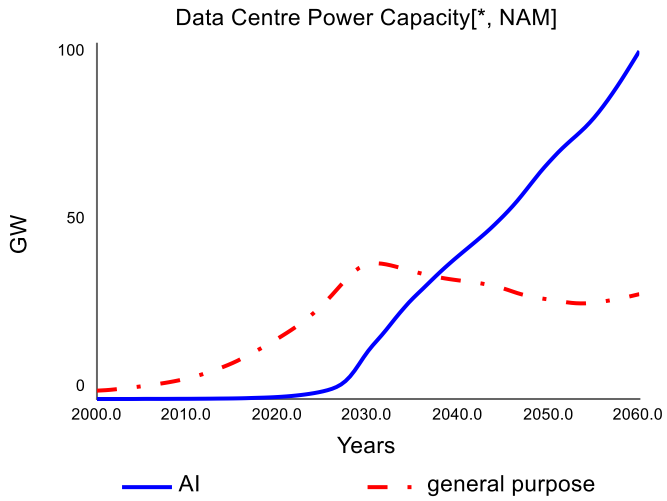
3 Results

Our base run results illustrate how rising demand for all types of data centres leads to continued expansion in **Data Centre Power Capacity**, improvements in **Compute Power Efficiency**, and increases in **Data Centre Total Energy Demand**, while at the same time, **Unit Price per Die** steadily declines. The NAM region (North America) will be presented in the base run results since it is the largest AI market today. The x axis of the figures will start from year 2000 to better focus on the period where computing demand and capacity substantially begin increasing.

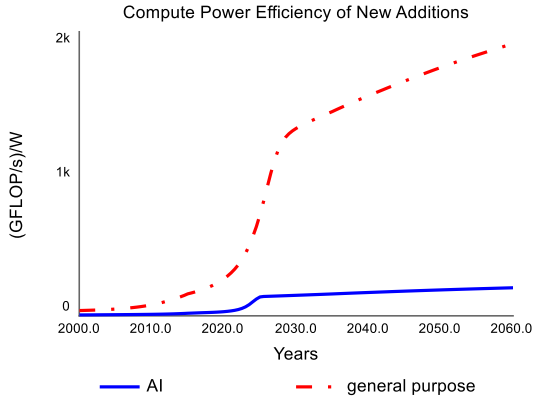
The evolution of **Market Size** for data centre services is influenced both exogenously—by population growth and the increasing share of the global population with internet access—and endogenously, through **Willingness to Pay for Data Centre Services**, which rises as these services become more essential to everyday life. The strongest growth occurs in the AI market, where computational demand increases more sharply than in general purpose markets, eventually surpassing general purpose in scale. However, this expansion is not just a function of exogenous demand drivers; endogenous system feedbacks, particularly those governing efficiency gains and hardware evolution, play an equally important role in shaping how computation capacity is distributed across data centre types and process nodes.



The persistent gap between **Market Size** and **Data Centre Computation Capacity** creates a sustained push for expansion, reflected in **Data Centre Power Capacity**, which grows significantly in both data centre types. AI workloads experiences increases severalfold, while general purpose grows at a slower pace, and eventually stabilizing. Although general purpose initially dominates in total **Data Centre Power Capacity**, AI segment eventually surpasses it as demand for high-performance computation accelerates in the 2040s.

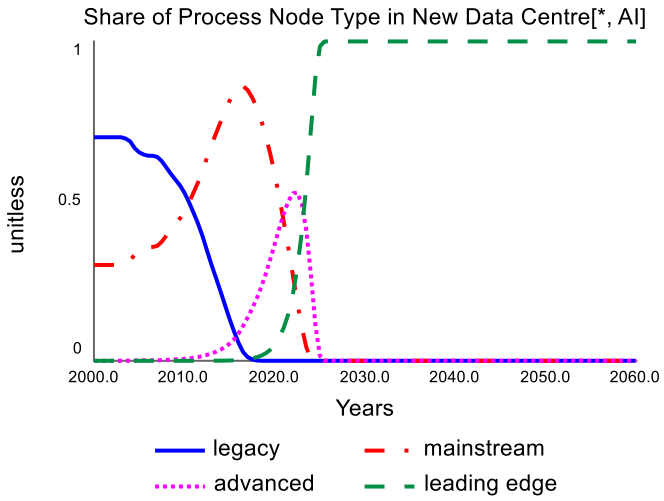


Despite this strong expansion, the rise in **Data Centre Power Capacity** is moderated by improvements in **Compute Power of New Additions**. Shrinking **Avg Process Node Size per Process Node Type** increases both **TFLOPs per Die** and **Power per Die**, making each unit of **Data Centre Power Capacity** more efficient over time. These effects are particularly pronounced for general purpose segment, where a steep rise in efficiency is seen in the late 2030s, followed by AI, which benefits from similar trends but stabilizes at a lower value.

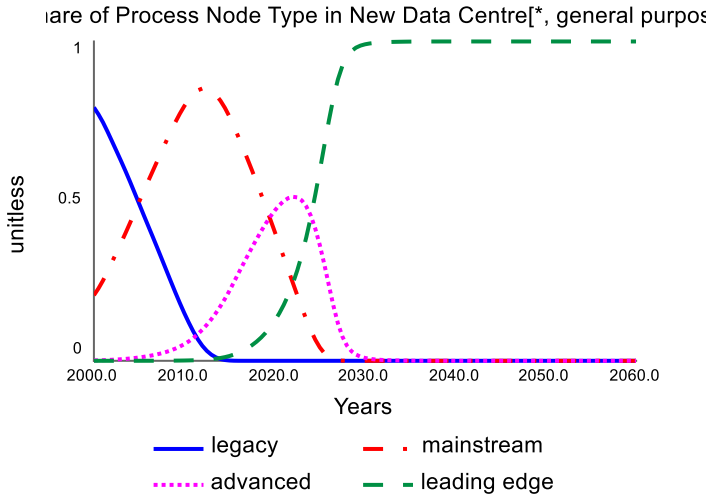


As semiconductor production scales up, there is a shift in **Share of Process Node Type in New Data Centres** towards smaller node sizes, particularly for AI applications. This transition occurs across all data centre types, though at different speeds, as each category responds differently to efficiency pressures and cost reductions.

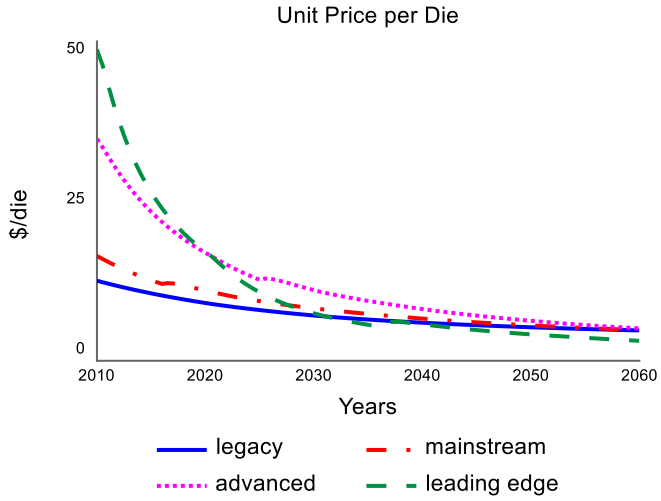
The shift toward smaller process nodes is evident in **Share of Process Node Type in New Data Centres**. At present, **Mainstream Process Nodes (≥28nm)** dominate, accounting for nearly 80% of new deployments. However, as efficiency concerns intensify—driven by a rising **Ratio of Energy Cost to Price**—there is a transition to smaller nodes. **Advanced Process Nodes (5-10nm)**, peaking in the late 2020s before giving way to **Leading Edge Process Nodes (≤3nm)**. By 2030s, almost all new data centres operate on **Leading Edge** technology, reflecting the AI segment's high computational intensity and strong incentives to maximize efficiency.



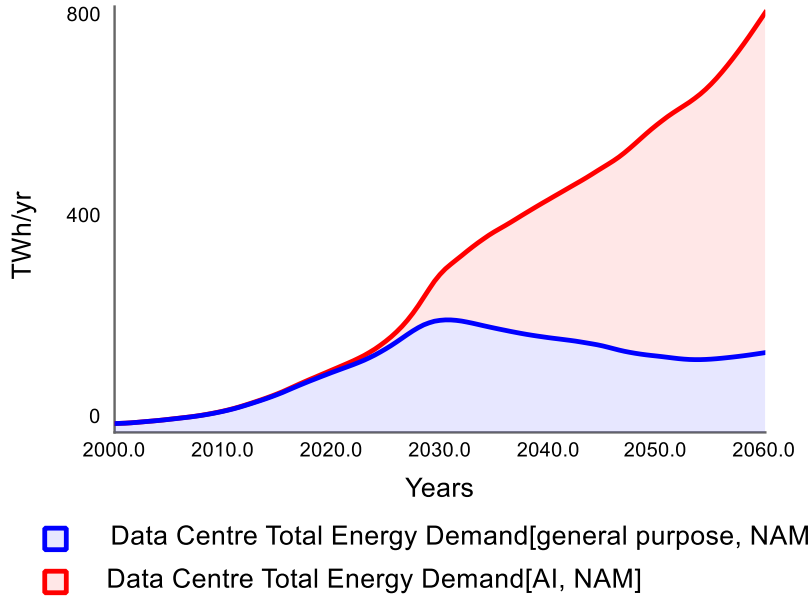
A similar shift occurs in general purpose , though at a slower pace due to its lower computational intensity. While **Mainstream Process Nodes** remain dominant throughout the simulation, they gradually cede ground to **Advanced** and **Leading Edge** nodes, which account for a growing share of new data centres. By 2030s, **Advanced Process Nodes** make up a majority of new cloud computing data centres but gives way to **Leading Edge Process Nodes**.



On the cost side, declining **Avg Process Node Size per Data Centre Type** and cumulative production effects drive reductions in **Unit Price per Die**. Learning effects amplify this trend, particularly for **smaller node sizes**, which see the sharpest cost declines. The cost advantage of transitioning to **Leading Edge Process Nodes** becomes increasingly apparent, though price reductions are tempered by the high complexity and capital intensity of manufacturing at these scales.



These trends collectively drive an increase in **Data Centre Total Energy Demand**, which more than quadruples from current year 2025 to end of simulation at 2060. While the total power required to sustain computation grows, the efficiency gains from **Process Node Size Reduction** prevent an even steeper rise, particularly in AI-focused workloads. Relative to global electricity demand projections, the share of data centres in total power consumption increases over time, though efficiency improvements mitigate this effect compared to a scenario without technological progress.



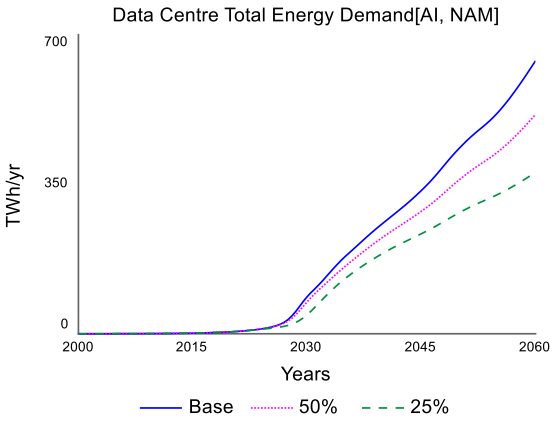
4 Scenario Analysis

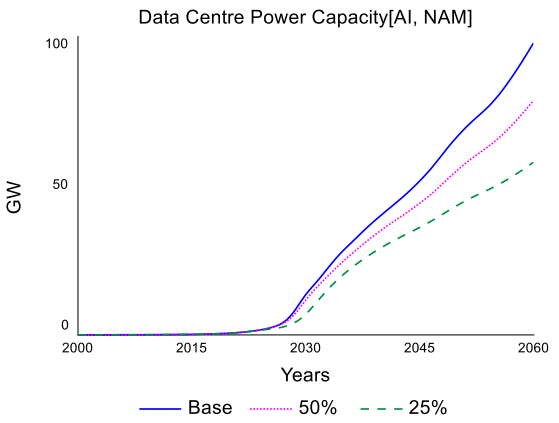
Of the various parameters and feedback loops that can be adjusted for scenario analysis, three were chosen due to policy implications that can be discussed. The adjusted parameters are 1) grid constraints, 2) efficiency gains, and 3) Moore's Law loop, or the performance gains. The presented graphs will show varying years, depending on the sensitivities highlighted.

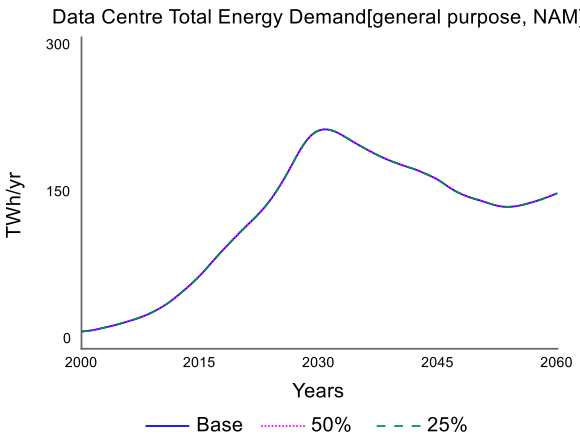
One of the most significant uncertainties in our results is the impact of **Grid Constraints**—specifically, whether electricity infrastructure can scale fast enough to accommodate the desired expansion of **Data Centre Capacity**. In the base run, grid expansion occurs at up to 15 GW per year, but alternative scenarios with lower rates (e.g., 5 GW per year) demonstrate how severe constraints could restrict growth. The base run is not constrained by the grid, thus in this scenario analysis availability of grid was adjusted to 50%, and down to 25%.

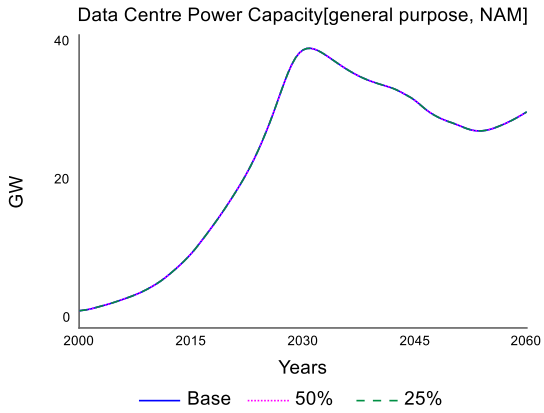
This limitation begins to emerge in the 2030s, disproportionately affecting AI workloads, which has the highest power demands. However, the constraints also general purpose, as competition for available grid capacity increases.

Notable observation is that AI is affected capacity constraints. The data centre energy demand and power capacity are reduced proportionally. However, General purpose data centres are unaffected by this scenario. This indicates that expansion of the AI sector is more sensitive to the availability of power to support it.

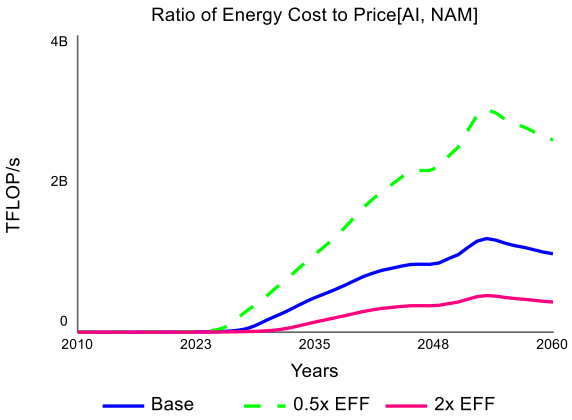


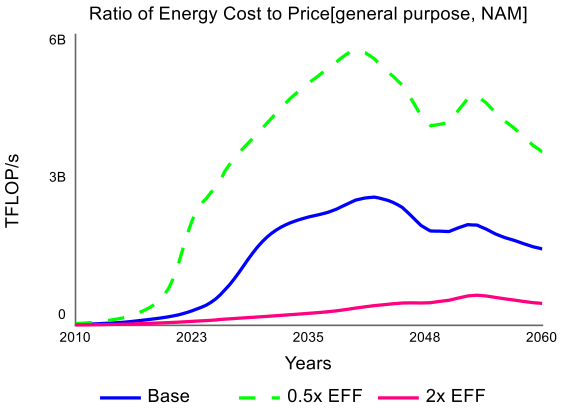






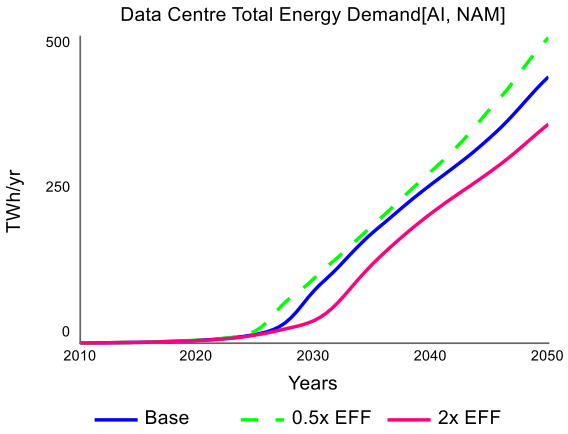
Efficiency gain sensitivity can be tested by scaling the Power per Die variable. Similar to grid constraints, a low efficiency and high efficiency are compared to the base case, at 0.5 and 2.0 efficiency multipliers, respectively.

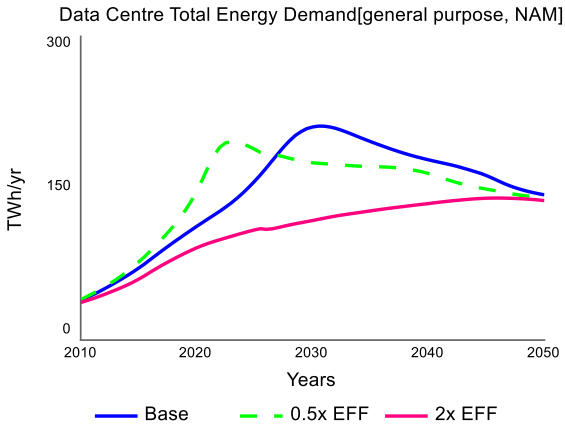




Efficiency sensitivities have a strong influence on costs for both AI and general purpose data centre markets. This is due the crucial role that crucial role that processors and their energy performances have on the core operations of data centres. In the 2x efficiency case, this outdated hardware are

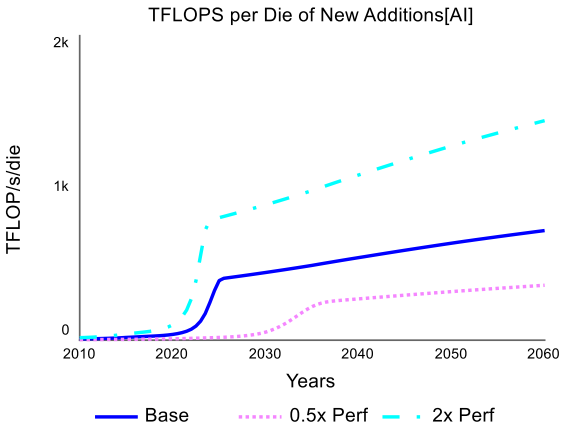
swapped out quickly, and this decision is quickly rewarded by lower costs. The 0.5x case, on the other hand, leads to data centres retaining older hardware, which in turn becomes more costly and less efficient.

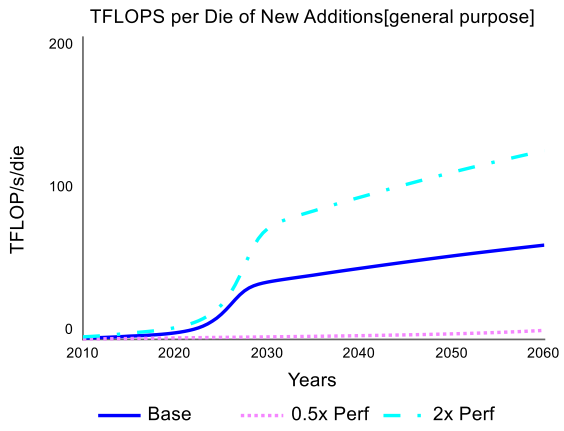




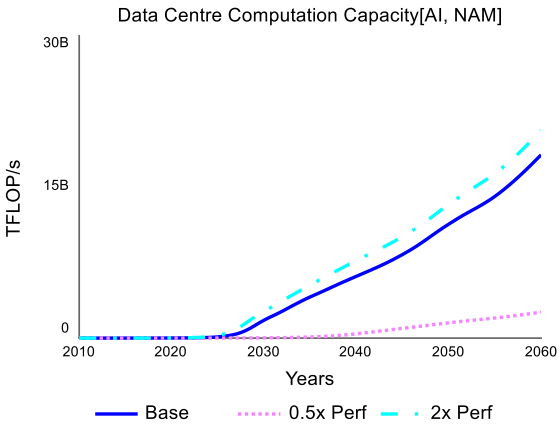
The loss in efficiency is also reflected in the total energy demand for both AI and general computing. Direct adjustments in efficiencies affect energy demand of general purpose data centres more dynamically than with AI data centres, which respond proportionally to the efficiency changes. In general computing market, the overall energy demand is consistently lower for 2x efficiency, but base case and 0.5x efficiency eventually catch up and converge around 2050s.

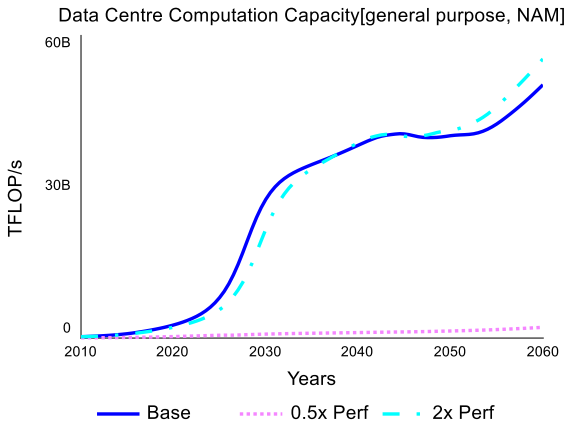
To test the effect of computing performance gain or hindrance, the variable TFLOPS per Die of New Additions was scaled by 2 and 0.5. Both AI and general computing performances benefit greatly from the 2x performance boost. For the 0.5x low performance scenario, the AI market recovers moderately but at a consistently lower performance. General computing suffers greatly from loss in performance for the remainder of the simulation.





For the energy consumption of AI and general computing markets, similar observations are made. Even with a doubling of expected performance gains, there is not a substantial increase in energy demand over the simulation years. However, loss in Moore's Law performance gains lead to a drastically lower overall energy demand for both markets. This scenario test can be interpreted that increasing performance from the base case do not incur significant energy penalties, but loss in performance can lead to an almost "bottoming out" of total energy demand.





5 Conclusion

The rapid expansion of artificial intelligence has placed data centres at the centre of global energy discussions. While historical trends suggest that efficiency improvements have consistently moderated digital infrastructure's power demand, the scale and computational intensity of AI introduce new dynamics that may test these historical patterns. This study sought to answer whether AI-driven data centres will follow past trends of efficiency-led moderation in energy demand or whether their rapid expansion will lead to a fundamental shift in electricity consumption.

Our results suggest that while efficiency gains will significantly offset growth in computation demand, they will not be sufficient to prevent an overall increase in Data Centre Total Energy Demand. The rapid rise of AI workloads leads to a shift in Data Centre Computation Capacity, eventually surpassing general computing in scale. However, this shift is not just a function of exogenous demand growth; endogenous system dynamics, particularly those governing Process Node Size, Compute Power Efficiency, and Infrastructure Scaling, will determine the trajectory of this expansion. The model highlights a complex interplay between efficiency improvements, cost reductions, and infrastructure constraints, which together shape the future of AI-driven computation.

For the energy industry, these findings underscore the growing importance of Grid Constraints in determining the future capacity of data centres. Unlike previous waves of digital infrastructure growth, which largely scaled within existing power availability, AI workloads require high-density, continuous power consumption that may exceed local grid capacity in key regions. If grid expansion does not keep pace with demand, it could create bottlenecks that slow AI adoption and shift data centre expansion to regions with excess power supply. This raises critical questions for energy planners, particularly regarding the balance between expanding renewable energy generation and ensuring stable, high-capacity power delivery for data centres.

For the AI industry, the results highlight both opportunities and challenges. The increasing computational intensity of AI workloads places strong economic pressure on efficiency improvements, leading to the accelerated adoption of Leading Edge Process Nodes and specialised hardware. This means AI providers will need to navigate a rapidly evolving semiconductor landscape, balancing performance gains with rising chip production costs and potential supply chain constraints. Additionally, while Compute Power Efficiency improvements will help contain costs, the growing Ratio of Energy Cost to Price may force AI companies to optimise model architectures, reduce redundancy, and prioritise energy-efficient inference over brute-force training.

Despite the robustness of the model in capturing key feedback loops, there are several uncertainties and limitations. The pace of technological change, particularly in semiconductor fabrication and AI algorithmic efficiency, remains difficult to predict. While Process Node Size Reduction follows

established scaling laws, its future trajectory could be affected by fundamental physical limits and shifts toward alternative computing paradigms, such as neuromorphic computing or quantum acceleration. Additionally, the model assumes a relatively smooth progression of grid expansion, but real-world deployment of energy infrastructure is often subject to regulatory delays, investment cycles, and geopolitical risks.

Ultimately, the question of whether AI will fundamentally alter the trajectory of global energy demand remains open. While efficiency gains have historically tempered digital expansion, the unique scale and power demands of AI workloads may push infrastructure growth into uncharted territory. The results of this study suggest that AI-driven energy demand will not grow unchecked, but neither will it plateau quickly—it will be shaped by the continuous interplay of technological progress, market forces, and resource constraints. How these factors evolve over the coming decades will determine whether AI follows historical efficiency trends or drives a step-change in computing's energy footprint.

References

Andrae, A. S. G., & Edler, T. (2015). On global electricity usage of communication technology: trends to 2030. *Challenges*, 6(1), 117-157. https://doi.org/10.3390/challe6010117

Bawden, T. (2016). Global warming: Data centres to consume three times as much energy in next decade, experts warn. *The Independent*. https://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html

Boston Consulting Group. (2024). *Power Moves: How CEOs Can Achieve Both AI and Climate Goals.* https://www.bcg.com/publications/2024/how-ceos-can-achieve-both-ai-and-climate-goals

Deloitte. (2024). *As Generative AI Asks for More Power, Data Centers Seek Sustainability.* https://www2.deloitte.com/us/en/insights/industry/technology/as-generative-ai-asks-for-more-power-data-centers-seek-sustainability.html

Dianati, K. (2012). *A System Dynamics Approach to Data Center Capacity Planning – A Case Study.* University of Bergen. https://hdl.handle.net/1956/6124

Federal Energy Regulatory Commission. (2024). *Assessment of Demand Response and Advanced Metering.* https://www.ferc.gov/sites/default/files/2024-02/2024_demand_response_assessment.pdf

Goldman Sachs. (2024). *AI, Data Centers and the Coming US Power Demand Surge.* https://www.goldmansachs.com/intelligence/pages/ai-data-centers-and-us-power-demand.html

Huber, P., & Mills, M. (1999). Dig more coal – the PCs are coming. *Forbes Magazine*, May 31, 1999. https://rmi.org/wp-content/uploads/2017/05/RMI_Document_Repository_Public-Reprts_E99-18_MMABLInternet.pdf

International Data Corporation. (2024). *IDC Report Reveals AI-Driven Growth in Datacenter Energy Consumption.* https://www.idc.com/getdoc.jsp?containerId=US50755623

International Energy Agency. (2024). *What the Data Centre and AI Boom Could Mean for the Energy Sector.* https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector

Jezard, A. (2017). In 2020 Bitcoin will consume more power than the world does today. *World Economic Forum*. https://www.weforum.org/agenda/2017/12/bitcoin-consume-more-electricity-than-world-2020/

Koomey, J. G. (2011). *Growth in data center electricity use 2005 to 2010.* Analytics Press. https://www.koomey.com/post/8323374335

Lovins, A. (1999). *Exchanges between Mark Mills and Amory Lovins about the electricity used by the Internet.* Rocky Mountain Institute. https://rmi.org/insight/the-internet-and-electricity-use/

Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984-986. https://www.science.org/doi/10.1126/science.aba3758

McKinsey & Company. (2024). *AI Power: Expanding Data Center Capacity to Meet Growing Demand.* https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity

Paccou, R., & Wijnhoven, F. (2024). *Artificial Intelligence and Electricity: A System Dynamics Approach.* Schneider Electric Sustainability Research Institute. https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/artificial-intelligence-electricity-system-dynamics-approach/

SemiAnalysis. (2024). *AI Datacenter Energy Dilemma – Race for AI Datacenter Space.* https://semianalysis.com/p/ai-datacenter-energy-dilemma

Shehabi, A., Smith, S. J., Sartor, D. A., Brown, R. E., Herrlin, M., Koomey, J. G., ... & Lintner, W. (2016). *United States Data Center Energy Usage Report.* Lawrence Berkeley National Laboratory. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

Spirova, T. (2021). *The Role of Data Centres and Energy Available to Support the Growing Smart Industry in Twente.* University of Twente. https://purl.utwente.nl/essays/86656

Wijnhoven, F., & Koot, M. (2021). Usage impact on data center electricity needs: A system dynamic forecasting model. *Applied Energy*, 288, 116798. https://doi.org/10.1016/j.apenergy.2021.116798

Wijnhoven, F., & Paccou, R. (2024). *Artificial Intelligence and Electricity: A System Dynamics Approach.* Schneider Electric Sustainability Research Institute. https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/artificial-intelligence-electricity-system-dynamics-approach/