

# AI Energy Elasticity and Data Center Energy Consumption Growth Scenarios to 2030

Aigumov T.G., <sup>a</sup>, Taha M.E.M., <sup>b</sup>, Dadaev Y.E. <sup>c</sup>

<sup>1</sup>PhD in Economics, Associate Professor, Associate Professor of the Department of Information Security and Software Engineering FSBEI HE "Dagestan State Technical University", Makhachkala, Russian Federation

<sup>2</sup>Postgraduate student of the Department of Electric Power Systems and Networks, Federal State Budgetary Educational Institution of Higher Education, Kazan State Power Engineering University; General Director of TMG LLC, Taha Mustafa Group Engineering and Project Management Limited Liability Company, Kazan, Russian Federation.

<sup>3</sup>Senior Lecturer, Department of Management, Kadyrov Chechen State University, Grozny, Russian Federation  
[915533@mail.ru](mailto:915533@mail.ru), [moustafa.taha@yandex.ru](mailto:moustafa.taha@yandex.ru), [dadaev.ya.82@mail.ru](mailto:dadaev.ya.82@mail.ru),


**Keywords:** AI energy elasticity, data centers, energy consumption, scenario modeling.

**Abstract:** The article introduces the concept of "AI energy elasticity" as the dimensionless sensitivity of data centers' total power consumption to changes in the volume of services based on artificial intelligence algorithms. A formal decomposition is proposed:  $E = E_{IT} \kappa_{inf}$ , where the IT component is determined by the training and inference profile, and the infrastructure multiplier is determined by the efficiency of cooling, distribution, and power conversion. Against the backdrop of the accelerated adoption of accelerators with high thermal packages and the transition to direct liquid cooling, three scenarios are assessed for 2030: moderate, baseline, and boosted. The methodological framework draws on Koomey's law for long-term reduction in energy per operation, systemic assessments of data center energy efficiency (Eric Mazanet, Arman Shehabi), work on the contrast between training and inference energy costs (Emma Strubell et al.), energy reporting protocols in machine learning (Peter Henderson), and recommendations on how to achieve savings through the choice of model, hardware, and deployment region (David Patterson et al.). It is shown that in the short term, energy elasticity remains elevated due to stagnation in the average market energy efficiency ratio and a structural shift towards AI workloads; in the long term, it decreases due to algorithmic and engineering improvements. Practical implications include standardization of measurements, mandatory reporting, spatiotemporal optimization of training, and proactive modernization of network infrastructure. Regional constraints are separately noted: shortages of connected capacity and water increase the importance of load flexibility and useful heat removal.

## 1 INTRODUCTION

The rapid spread of deep learning algorithms has transformed data centers (DCs) into structurally important consumers of electricity, and assessing "AI energy elasticity"—that is, the relative sensitivity of a DC's total electricity consumption to a relative change in the volume of AI-based services ( $\epsilon = (\partial E/E)/(\partial Y/Y)$ )—is becoming a key analytical tool. According to the International Energy Agency's baseline scenario, global data center electricity

consumption will more than double by 2030, reaching approximately 945 TWh. The largest contribution to this increase will come from the growth of AI-optimized workloads, for which a multiple increase in power demand is expected, with the growth rate of data center electricity consumption on average outpacing the global average by more than four times from 2024 to 2030. Historically, accelerated improvements in the energy efficiency of infrastructure and IT equipment have played a restraining role: analysis by Eric Mazanet, Arman

<sup>a</sup> <https://orcid.org/0000-0002-8737-0228>

<sup>b</sup>

<sup>c</sup>

<sup>h</sup>

<sup>h</sup>

<sup>t</sup>

<sup>p</sup>

<sup>ø</sup>

<sup>s</sup>

<sup>l</sup>

<sup>/</sup>

<sup>ó</sup>

Shehabi, and John Koomey showed that, despite the exponential growth of compute instances and storage in the 2010s, global data center power consumption increased by only a few percent thanks to advances in architecture, power electronics, and cooling.

However, this "flat" trajectory is ending: consensus forecasts for the US by 2030 already diverge by an order of magnitude, indicating high elasticity of power consumption relative to demand for AI services and the uncertainty of future workloads.

Demand for training large models and mass inference is fueled by exponential growth in the required computations: Dario Amodei and Danny Hernandez showed that from 2012 to 2018, the computational load for record-breaking models doubled approximately every 3.4 months, dramatically increasing the energy requirements per unit of model performance; subsequent work confirmed that this high rate of computation has persisted, albeit with signs of slowing. At the applied-discipline level, the problem was first articulated by Emma Strubell et al., who demonstrated the significant differences in energy costs between model tuning and pretraining, thus justifying the need to consider energy metrics alongside accuracy. Methodological frameworks and tools have been proposed to standardize measurements: Peter Henderson and colleagues introduced a protocol for systematically reporting the energy and carbon footprint of machine learning experiments, and Loïc Lannelongue developed a universal methodology for calculating "green algorithms" for comparable estimates based on time, processor type, memory, and site efficiency.

On the infrastructure side, improvements in data center energy efficiency have reached technological limits: although leading sites demonstrate values close to unity, the global average has stabilized in recent years, limiting further "engineering" reserves and increasing the sensitivity of overall consumption to AI traffic growth. Along with risks, there are also countertrends: David Patterson et al. demonstrated that careful selection of the model and equipment, localization of computation in low-carbon regions, and high utilization can reduce training energy consumption by up to 100 times, and specific emissions by up to 1000 times; With the widespread implementation of such practices, the energy load for training could stabilize by the end of the decade.

By "AI energy elasticity," we mean a dimensionless measure of the sensitivity of the total energy consumption of data centers  $E$  to changes in the volume of services provided based on artificial

intelligence algorithms  $Y$  under given quality and service standards:  $\varepsilon_{\{E,Y\}} \equiv d \ln E / d \ln Y$ . In practical assessments, the arc (average) elasticity over a finite demand interval is also used, and the qualitative interpretation of the value is traditional for economic and mathematical analysis:  $|\varepsilon| > 1$  is an elastic dependence,  $|\varepsilon| < 1$  is inelastic,  $|\varepsilon| = 1$  is unitary. This interpretation is directly based on the standard definitions of the elasticity coefficient and its logarithmic form.

To operationalize this indicator, it is convenient to decompose the data center's annual electricity consumption into the product of the "IT component" and the infrastructure multiplier:  $E = E_{IT} \cdot \kappa_{infr}$ . The latter corresponds to the data center's energy efficiency coefficient, normalized in the national standard: it is defined as the ratio of the total energy input at the facility boundary to the energy actually consumed by the IT load. The standard specifies metering boundaries, measurement categories, and a calculation scheme based on input metering readings, which is critical for the comparability of  $\kappa_{infr}$  between sites and over time. Logarithmic differentiation then yields the decomposition  $\varepsilon_{\{E,Y\}} = \varepsilon_{\{E_{IT},Y\}} + \varepsilon_{\{\kappa_{infr},Y\}}$ : the elasticity of total consumption is the sum of the elasticity of the IT component and the elasticity of infrastructure losses and subsystems (cooling, distribution, conversion). The IT component is naturally represented as  $E_{IT} = \sum_j e_j \cdot N_j$ , where  $e_j$  is the specific energy consumption for an elementary computational operation, transmission, or storage in the  $j$ -th subsystem (processors, memory, networks, storage devices), and  $N_j$  is the number of such operations determined by the load profile (model training, output, data preparation).

On this basis,  $\varepsilon_{\{E_{IT},Y\}}$  is decomposed into two partial components: by usage intensity (via  $\partial \ln N_j / \partial \ln Y$ ) and by technological parameters (via  $\partial \ln e_j / \partial \ln Y$ , reflecting the energy efficiency of hardware and software and algorithmic improvements). The long-term dynamics of  $e_j$  is limited by the empirical "Kumi's law": the electrical efficiency of computing (operations per unit of energy) has historically grown at a rate of approximately doubling every eighteen months; This reduces the energy elasticity of the IT component, all other things being equal. The structure of AI-based services creates an asymmetry between training and inference modes: the energy consumption of training large-scale models is orders of magnitude higher per unit of completed experiment; in addition to the actual parameter optimization, architecture search and a series of

additional training runs make a significant contribution.

Therefore, as the share of training tasks increases, the partial elasticity  $\varepsilon_{\{E_{IT}, Y\}}$  increases, whereas with increasing mass inference, it is determined primarily by the energy intensity of the request and the target latency and quality constraints. The infrastructure component  $\kappa_{\text{infra}}$  depends on the power supply and cooling topology, operating modes, and measurement quality. The Russian standard prescribes a methodology for separating the IT load from associated losses and links  $\kappa_{\text{infra}}$  with a set of related data center efficiency metrics; The related documents in the same series also standardize the share of renewable energy, which is useful for parallel assessment of carbon – though not energy – elasticity.

Derived infrastructure metrics are also widely used in practical data center research; their correct application requires alignment with national measurement standards. Thus, AI energy elasticity is a parameter of the computing demand model, jointly determined by the technological trajectory of specific energy intensity, the AI workload profile, and the regulatory infrastructure efficiency. In the short term,  $\varepsilon_{\{E, Y\}}$  can be high due to infrastructure inertia and the abrupt introduction of capacity, while in the long term, it is reduced by algorithmic and hardware improvements. The difference between short- and long-term elasticity is a common feature of such dependencies. This concept provides a quantitative basis for scenarios up to 2030 and for assessing the effects of policy and engineering decisions.

The growth of data center energy consumption is determined by the combined effects of demand, hardware-architectural, operational, and infrastructure-energy factors, which are superimposed by climate constraints and regulatory frameworks. On the demand side, the accelerated adoption of AI systems plays a key role: both the specific computational complexity of training tasks and the frequency of inference requests are growing, which greatly increases the total energy consumption over the lifecycle. According to industry and research groups, in production use, the share of inference can exceed training and become the dominant energy consumption component, as a single trained set of models serves billions of user and service requests, while training only serves one-off campaigns (in some assessments on real-world platforms, inference accounted for approximately three-fifths of a company's machine learning energy consumption, with this share growing with the scale of deployments). This is explained by the "traffic multiplier": every slight increase in request length,

context growth, and model combination generates a superlinear load on accelerators and the data center network.

The hardware side of demand amplifies this dynamic. Modern accelerators for machine learning and highly parallel linear algebra are characterized by dramatically increased heat dissipation per module: typical maximum thermal power ratings for current modules reach hundreds of watts and approach kilowatts per device in server platform configurations. This dramatically increases the rack heat load density and the need for energy-intensive heat removal systems. For example, high-end accelerators were advertised with thermal power ratings of approximately 700 watts for server form factors, and higher for future generations. These figures are already beyond the comfort zone of traditional air cooling in computer rooms, directly leading to the widespread adoption of liquid heat removal and the restructuring of engineering infrastructure.

Architectural changes in data centers are manifested in two interconnected trends: the relentless growth of rack power density and the transition from air to direct liquid cooling of components. Field surveys and annual operator reviews show a steady shift toward heavier racks: the modal power level is increasing year over year, and the share of sites with rapidly increasing power density is particularly high among colocation and cloud service providers. At the same time, the industry's integrated energy efficiency metric (the ratio of a facility's total power consumption to its IT load) has been hovering around the lower limit achieved by the widespread "first wave" of measures (hot/cold aisle confinement, variable frequency drives, etc.) for several years now; no significant further reduction in this ratio is observed at the market average. This means that every additional watt of IT load translates almost linearly into an increase in the overall load on the facility's power system. Cooling systems engineering is undergoing a shift toward liquid paths. Professional associations and industry standards (e.g., energy efficiency standards for data center engineering systems and a special metric for the share of energy devoted to cooling—the "mechanical component of the load") recognize the need to account for increased accelerator heat fluxes and the associated revision of design targets for air conditioning and heat removal infrastructure. As rack densities and component thermal packages increase, the industry is rapidly introducing direct liquid cooling solutions for components, as well as immersion systems. These improve heat transfer but

require more complex electrical and hydraulic systems, increasing both the specific electrical load of engineering systems and reliability requirements.

Operational factors also play a significant role. Many operating facilities continue to experience underutilization of uninterruptible power supplies and distribution systems, which reduces the overall energy efficiency of both engineering systems and cooling (low utilization leads to converter stages operating far from their optimum efficiency). At the same time, operators are forced to implement redundancy schemes of N+1 and higher to ensure service fault tolerance, which increases the "ballast" share of consumption even with optimized system load management. Recent studies report that at a significant proportion of facilities, the actual load on available uninterruptible power supplies (UPS) was less than 40%, which directly impacts energy metrics at the site level.

The supply side of electricity is becoming an independent constraint. Inertial cycles of generating and grid capacity expansion are leading to a shortage of available connected capacity in key clusters, and waiting lists for grid connections are growing. Research by national laboratories and industry reviews document a multiple increase in the share of data centers in the electricity consumption balance of individual countries and regions in just a few years, as well as scenarios for doubling or tripling consumption by the end of the decade if the current pace of AI server deployment continues. This forces operators to seek alternative energy supply solutions – from long-term contracts for the supply of "green" energy to local generation and coordination of load flexibility, but in the short term it creates additional volatility in the tariff and technical environment.

Climate and water constraints place additional strain on energy consumption patterns. Locations in hot and arid zones impair the operating conditions of traditional evaporative cooling systems and increase both electricity and water consumption. Therefore, some operators are rapidly implementing closed-loop liquid cooling systems or completely abandoning water-intensive cooling modes, shifting the bulk of heat removal to recirculation systems with minimal specific water consumption. At the same time, some sites are integrating with district heating systems, returning excess heat to city networks and partially offsetting energy costs through useful heat removal. These technical solutions improve resilience, but generally do not reduce the electrical load on IT components and require additional pumping, heat exchange, and control capacity.

The structure of computing workloads and algorithmic practices are another source of change. After a "golden decade" of rapidly declining computing energy consumption (Kumey's Law, which describes the accelerated reduction in energy per operation due to systemic improvements), gains from hardware and software efficiency have plateaued: further improvements are becoming increasingly costly, and increased demand from scaling AI models and growing traffic outweighs energy savings per operation. At the same time, approaches to "energy benchmarking" and mandatory reporting on the carbon footprint of experiments and deployments are gaining ground in research and industry practice. Methodologies and tools are being proposed for systematically accounting for the energy costs and carbon intensity of computing, demonstrating orders of magnitude savings when switching to energy-optimal models, data centers, and regions with low-carbon generation. This creates the potential for "algorithmic decarbonization," but in practice requires disciplined development, reuse of learned weights, "distillation," and restrained hyperparameter search.

Infrastructure energy efficiency metrics are demonstrating the limits of previous approaches. The market average energy efficiency ratio has virtually plateaued in recent years, near the point where the effects of "low-hanging fruit"—organizational measures in computer rooms, airflow optimization, localized drive upgrades, etc.—have been exhausted. A further significant reduction in this ratio at the mass level requires a transition to a profound reconfiguration of cooling circuits (liquid paths at the rack and node level, immersion baths, and the elimination of evaporative cycles) and energy-optimal power distribution and conversion architectures, which significantly increases capital expenditures and implementation timelines. As a result, a "quasi-linear" transfer of IT load growth to the overall electrical load of facilities is most likely through 2030.

The market for cooling and power supply components is itself becoming a limiting factor. Rapidly growing demand for coolant distribution units, thermal modules, heat exchangers, and high-power power supplies is causing supply shortages and prolonging implementation times. This technological "friction" leads to the temporary operation of high-density racks in sub-ideal conditions, increasing specific energy consumption per unit of payload and reducing potential equipment utilization.

Water aspects of operation are becoming strategically important. Estimates show that with

traditional evaporative cooling systems, each kilowatt-hour of electricity consumed by a data center can require a significant amount of direct water consumption for cooling. In regions with water shortages, this imposes severe restrictions on the choice of technologies and operating modes, and annual reports from major operators document a rapid increase in total water use amid the construction of "AI factories." The transition to closed liquid loops and the elimination of evaporative cycles is a response to these challenges, but, as noted, this increases the electrical loads of pumps and heat management systems.

Finally, the spatial and temporal structure of energy consumption plays a significant role. AI cluster load peaks—both during training and in mass inference—may occur during the same network peak hours, increasing system risks. In response, flexibility mechanisms are proposed: shifting energy-intensive training tasks to low-carbon nighttime windows, geographically distributing tasks across regions with low-carbon generation, coordinating consumption profiles with system operators, and using heat consumption from city grids for the efficient removal of excess heat. Regulatory and methodological initiatives and industry standards for energy efficiency in data center engineering provide a framework for such strategies, including clarifying the calculation and target values of the mechanical load component and reporting requirements [3, p. 112].

The initial modeling framework is based on a factor decomposition of annual energy consumption:  $E = EI \cdot k_{infr}$ , where  $EI$  is the energy directly consumed by the computing load (accelerators, central processors, memory, storage, intra-hall and backbone networks), and  $k_{infr}$  is an infrastructure multiplier reflecting additional costs for cooling, distribution, and power conversion (analogous to the facility's energy efficiency coefficient). Three key groups of parameters are key for the  $EI$  trajectory: (i) the introduction of new computing capacities and their utilization; (ii) the load profile (training and inference ratio in AI systems, query length, latency requirements); (iii) the specific energy consumption of elementary operations in hardware and software stacks. The  $k_{infr}$  values are related to the rack heat dissipation density, heat removal schemes (air, direct liquid cooling, immersion baths), redundancy modes, and input metering discipline. The set of parameters is calibrated against open international and national estimates: at the global level, the International Energy Agency expects the total electricity consumption of data centers to more than double by 2030, to

approximately 945 TWh, with AI-optimized sites demonstrating a significant increase. On the other hand, the contribution of data centers to the growth of global generation in the agency's baseline scenario is estimated at less than one-tenth of the total increase in electricity demand, which sets an upper limit on the feasibility of "explosive" forecasts. To improve the robustness of our findings, we use a three-pronged scenario approach: "baseline" (inertial trajectory), "moderate" (maximizing energy efficiency and organizational measures), and "accelerated" (high intensity of AI demand with limited infrastructure progress). The ranges are compared with a critical review of the international hardware energy efficiency (4E) program, which specifies corridors for 2030 from 600–800 TWh with moderate growth to 1,100–1,400 TWh with accelerated AI penetration. Regional estimates for the US, Europe, China, and individual power systems are also taken into account.

The baseline assumes: (a) accelerated but controlled expansion of AI clusters at the largest cloud operators; (b) a gradual shift in the load profile toward mass decommissioning with moderate growth in the share of long contexts; (c) stabilization of average market infrastructure metrics at the achieved levels due to a partial transition to direct liquid cooling in high-density zones, but without a total reconfiguration of the entire fleet; (d) maintaining the pace of algorithmic and hardware improvements that reduce the specific energy intensity of training and inference. At the global level, this is consistent with the IEA's estimate of a twofold increase and reaching approximately 0.9–1.0 PWh by 2030. Within this amount, a growing share comes from AI-optimized data centers, where, according to the agency's estimates, electricity demand by 2030 will more than quadruple relative to the baseline level of the middle of the decade. Thus, the global energy elasticity of AI remains above unity, but is smoothed out due to the efficiency of hardware and the partial replacement of air cooling with liquid cooling in the hottest zones.

The baseline regional structure is characterized by double-digit capacity and energy growth rates in the United States and China, where key equipment supply chains and access to sites with large connected capacity are concentrated. According to analytical reviews, data center load in the United States could more than double by 2035, which is already evident in current plans of system operators and federal statistics forecasts: rapid revisions to load growth estimates and higher generation growth rates for 2025–2026 are directly related to new cluster demand. In China, industry estimates point to double-digit average annual growth rates and the possibility

of reaching levels comparable to the national electricity consumption of individual large countries by 2030. These facts reinforce global inertia, even with the implementation of energy efficiency measures.

In Europe, localized constraints on connected capacity are increasing in traditional hubs (the northwest of the continent, Ireland, and southeast England). National transmission operators' plans indicate accelerating demand growth and an increasing share of data centers in the balance of individual power systems, where, by the early 2030s, the share of electricity demand from data centers and associated digital clusters could reach significant proportions. This is confirmed by both the Irish grid operator's publicly available data on the balance to 2034 and the UK operator's analysis of digital infrastructure load trajectories. Taken together, this means that even a "slow-moving" world faces the need for accelerated network and generation development by 2030 to accommodate a new type of constant, "square-constant" data center load [2, p. 90].

A moderate option allows for the aggressive promotion of energy-efficient practices and technological solutions across the entire value chain: (1) standardized energy reporting for IT experiments and deployments, which disciplines research and product contours (limiting the depth of enumeration, reusing weights, "distillation"); (2) the selection of sites with low specific carbon intensity and high availability of renewable sources, which often coincides with lower specific cooling costs; (3) accelerated implementation of direct liquid cooling at the node and rack level (cold plates, inter-rack distribution units), ensuring a reduction in the mechanical component of consumption with increasing density; (4) alignment of the training profile in time and space (moving energy-intensive experiments to night windows of low carbon intensity and to areas with excess generation). With this discipline, the infrastructure multiplier  $kinf$  decreases in the hottest clusters, and the specific energy intensity of operations falls faster than under the inertial trajectory. The scale of the effect is confirmed by industry methodologies and reviews: average market values for infrastructure efficiency have stagnated in recent years, but new, large facilities are demonstrating better performance when transitioning to liquid solutions; this is a clear reserve for the "moderated" scenario. As a result, the global E trajectory by 2030 remains in the lower half of the 4E corridor: approximately 0.7–0.9 PWh.

Regionally, the most pronounced effect of the "moderated" scenario is expected in areas where

power system operators are already promoting load flexibility and rapidly expanding renewable generation: in several US states, in northern Europe, and in certain provinces of China, where "green" data cluster policies include preferential connection and target cooling system standards. Taking into account forecasts for China (accelerated trajectories until 2030 with the support of regional programs) and the United States (growth of connected capacity, but with active modernization of networks and construction of low-carbon generation), the "restrained" option does not cancel growth, but keeps energy elasticity closer to unity or below it due to intensive technological improvements and operational management of the load profile [5, p. 19].

The accelerated scenario assumes a sharp increase in the share of AI computational tasks with long contexts and extensive post-processing, widespread adoption of multimodal services, rapid growth in the frequency of AI service calls in mass-market products, as well as "compression" effects on the infrastructure side-limited availability of connected power in traditional nodes, shortages of cooling and power distribution components, and delays in reconfiguring racks for direct liquid cooling. Under these conditions, the load profile shifts toward high peaks and prolonged periods of near-maximum accelerator load, while the infrastructure multiplier (KINF) decreases more slowly due to the growing share of pumping and additional heat-exchange cascades. As a result, the elasticity of  $dlnE/dlnY$  for AI services rises above unity, and the global trajectory of E by 2030 approaches the upper limit of the 4E review - 1.1–1.4 PWh, which is consistent with the more "tough" market estimates of individual analytical groups for the mid-2030s.

The United States emerges as the "anchor" region in this scenario: rapid plans to build large campuses and contracts for low-carbon generation are failing to fully offset demand growth, leading to a significant increase in the share of data centers in national electricity demand by the mid-2030s, according to several researchers. Short-term forecasts already indicate an upward revision of load growth rates as new connections in individual synchronous zones are factored in. In China, the underlying growth of the digital economy is complemented by the challenges of sovereign AI development and data localization requirements, leading to sustained double-digit data center consumption dynamics. Industry reports indicate the possibility of achieving consumption levels comparable to those of large European countries by 2030. Under these conditions, regional water and climate restrictions are accelerating the

transition away from evaporative cooling systems, increasing the electrical component of cooling in hot clusters. For all three scenarios, three bottlenecks are critical: (1) the rate of introduction of accelerators with high thermal packages and the corresponding rack infrastructure; (2) the speed of transition to direct liquid cooling and the level of measurement standardization (which affects the actual reduction in  $k_{infr}$ ); (3) the availability of connected capacity and the local generation structure (which also determines the possible mode of learning equalization over time). According to industry surveys, average market values of infrastructure efficiency have shown virtually no improvement in recent years; this means that without rapid reconfiguration of "hot" nodes, the increase in IT load is almost linearly translated into an increase in the facility's electricity load. At the same time, national and regional operators are already factoring in increased demand growth rates into short-term balances, which confirms the feasibility of the "upper" corridor under an unfavorable combination of factors.

Small and open power systems are a special case. Ireland illustrates that a high concentration of data centers in a single network quickly changes the demand structure and reserve requirements: according to recent data from the transmission operator, the share of digital loads in total demand could reach about a third by 2034, and the short-term need for new generation and controlled capacity. This trend is already emerging in the middle of the decade. For the UK, specialized surveys point to a significant contribution of hyperscale data centers to demand growth by 2030, which requires accelerated development of high-voltage networks and the implementation of tariff signals that incentivize geographic redistribution of deployment. These observations are important for extrapolation to other isolated or "narrow" systems (island regions, individual interconnectors) [6, p. 31].

A comparison of the three trajectories allows us to outline a realistic corridor for global data center electricity consumption by 2030: from 0.7–0.9 PWh in the "moderate" scenario to 1.1–1.4 PWh in the "boosted" scenario, with a "baseline" target of approximately 0.9–1.0 PWh. A unifying feature of all scenarios remains the increased energy elasticity of AI in the coming years: AI-optimized clusters make a disproportionately large contribution to energy growth, which is also recognized in the IEA's thematic sections. The contribution of data centers to the overall growth of global electricity demand in 2024–2030 remains smaller compared to drivers such as industrial output, transport electrification, and the

growth of air conditioning, but the spatial concentration of load creates locally "hard" constraints and management challenges for system operators and regulators. The practical implications of the scenarios are twofold. First, for data center operators and their clients: the introduction of mandatory energy reporting, energy-oriented algorithm design, the deferral of training to periods of low carbon intensity, accelerated reconfiguration of "hot" rooms for liquid cooling, and localization in regions with abundant renewable energy sources and low outside temperatures. Second, for energy systems: accelerated expansion of network capacity, the development of manageable and low-carbon capacity (including flexible gas units, large battery stations, and new nuclear projects), and the implementation of tariff and market mechanisms that signal the cost of shortages based on location and time. Given the uncertainty of the AI demand scale, it seems reasonable to focus on a "readiness fork": planning infrastructure for the base path with the possibility of inexpensive "extension" to the upper corridor.

## 2 CONCLUSION

Thus, the concept of "AI energy elasticity" introduced in this paper—the marginal or arc sensitivity of the total energy consumption of data centers to changes in the volume of services based on artificial intelligence algorithms—allows us to link together three critical layers: the dynamics of demand for computing, the trajectories of technological efficiency, and the infrastructure constraints on energy supply. The decomposition  $E = EIT \cdot \kappa_{infr}$  showed that the contribution of the IT component is determined by both the intensity of use (training and inference profile, query length, latency constraints) and the specific energy intensity of operations, while the infrastructure multiplier captures the limits of the efficiency of cooling, distribution, and power conversion at the facility. In the short term, up to 2030, the overall elasticity remains elevated: any increase in AI traffic translates almost linearly into increased electrical load due to stagnation of average market infrastructure metrics and growth in accelerator thermal packages.

A comparison of the "constrained," "baseline," and "boosted" scenarios revealed a realistic range for global data center electricity consumption by 2030 and highlighted key uncertainties. In the "constrained" scenario, development and operational discipline—standardized energy and carbon footprint

reporting, aligned training across time and space, prioritizing low-carbon sites, and switching to direct liquid cooling—reduce  $\kappa_{\text{infr}}$  and accelerate the decline in specific energy intensity, bringing the elasticity closer to unity. The "base" path reflects an inertial combination of modernization and scaling: energy consumption grows faster than the digital sector's GDP, but slower than with simple extrapolation of AI traffic. The "boosted" scenario demonstrates the upper bound: accelerated integration of AI into mass services, shortages of connected power and cooling components, and delays in rack reconfiguration maintain elasticity significantly above unity.

The theoretical framework is consistent with the empirical body of work: "Kumey's Law" of multi-year energy-per-operation decline (John Koomey), systemic assessments of infrastructure efficiency and saturation of first-wave effects (Eric Mazanet, Arman Shehabi), evidence of the high energy intensity of training large-scale models and the gap between training and inference (Emma Strubell et al.), energy reporting protocols in machine learning (Peter Henderson et al.), and engineering recommendations on how to achieve savings through model, hardware, and deployment region selection (David Patterson et al.). Significant risks remain on the demand side for computing: the exponential growth in computing requirements for record-breaking models documented by Dario Amodei and Danny Hernandez, albeit slowing, continues to push up energy elasticity.

The practical implications for policy and management are clear. First, mandatory energy reporting for research teams and developers must be established, including the publication of specific energy intensity metrics, load profiles, and experimental conditions; this creates a "closed loop" between scientific quality metrics and operating costs. Second, accelerated standardization of measurements at facilities is essential: unification of IT and engineering load accounting boundaries, regular verification of metrics, and target energy efficiency coefficients for site classes. Third, spatiotemporal optimization should be implemented: shifting energy-intensive training tasks to periods of low carbon intensity, geographic dispersion of "hot" clusters, flexibility agreements with system operators, and the transfer of useful heat to urban networks. Finally, on the energy system side, this requires accelerated development of network infrastructure, manageable and low-carbon capacity, as well as tariff and market mechanisms that reflect the cost of shortages by location and time.

The final conclusion is as follows: through 2030, data centers form a new class of large, high-density, and weakly compressible loads. The fate of the global energy consumption trajectory is no longer determined so much by individual innovations as by the overall organization of the industry: the maturity of energy reporting, the speed of transition to liquid cooling, the discipline of algorithm design, and the readiness of energy systems to accommodate clusters with high baseloads. If these elements are coordinated, the "base" corridor appears achievable; ignoring them leads to a "forced" path and increased system costs.

## REFERENCES

- Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J., 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21 (248), 1–43.
- Koomey, J.G., Berard, S., Sanchez, M., Wong, H., 2011. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, 33 (3), 46–54.
- Mazanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J., 2020. Recalibrating global data center energy-use estimates. *Science*, 367 (6481), 984–986.
- Mytton, D., Ashtine, M., 2022. Sources of data center energy estimates: A comprehensive review. *Joule*, 6 (9), 2032–2056.
- Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., Dean, J., 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55 (7), 18–28.
- Sevilla, J., Villalobos, P., Cerón, J.F., Burtell, M., Heim, L., Nanjajjar, A.B., Ho, A., Besiroglu, T., Hobbhahn, M., 2022. Compute trends across three eras of machine learning. *arXiv preprint*, arXiv:2202.05924, 1–25.
- Shehabi, A., et al., 2024. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory (LBNL), Report, 1–79.
- Strubell, E., Ganesh, A., McCallum, A., 2019. Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
- Van Geet, O., Sickinger, D., 2024. Best Practices Guide for Energy-Efficient Data Center Design. U.S. Department of Energy / NREL, 48 pp.