Journal of Industrial Engineering and Applied Science
Vol. 4, No. 1, 2026 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online) | CODEN: JIEAAE

AI Chips and the Economics of Computer

LU, Wenqiang [1*]

[1] Nanchang Institute of Technology, China

**** LU, Wenqiang is the corresponding author, E-mail: 17379383151@163.com***

Abstract: Because chips exhibit large differences in speed, power consumption, and cost across tasks, performance, energy use, and time often involve intricate trade-offs under heterogeneous workloads. As frontier models continue to scale, compute is increasingly becoming a binding constraint, yet the economic meaning of "one additional unit of compute" remains insufficiently well defined. This paper links chip specialization to cost functions, market structure, and industrial policy, and explicitly incorporates the system-level division of labor between training and inference. It further emphasizes that leading-edge process capacity is scarce, so compute tends to concentrate in a small number of firms and countries, raising entry barriers in the compute–chip ecosystem and strengthening the extent to which industrial policy instruments such as subsidies, tax incentives, and export controls can generate salient shocks to the supply of compute.

Keywords: AI Accelerators, Semiconductor Value Chain, Compute Cost, Industrial Policy, Market Concentration.

Disciplines: Artificial Intelligence Technology. **Subjects:** Machine Learning.

DOI: https://doi.org/10.70393/6a69656173.333733 **ARK:** https://n2t.net/ark:/40704/JIEAS.v4n1a03

1 INTRODUCTION

Modern frontier AI has become increasingly compute intensive, not only because models have grown larger, but also because training cycles are embedded in competitive R and D routines where iteration speed carries strategic value and, in some cases, safety and reliability implications. Khan and Mann observe that training a leading AI algorithm can require roughly a month of computing time and may cost on the order of one hundred million dollars, an estimate that is striking precisely because it turns what used to be treated as a background engineering choice into a visible constraint on research trajectories. [1]

At the same time, the economic meaning of "one additional unit of compute" remains somewhat under specified. A unit of compute can be priced, but it is not a homogeneous commodity: throughput, latency, and energy consumption vary with workload characteristics, numerical precision, memory access patterns, and the surrounding software stack, which implies that identical nominal peak performance can translate into quite different effective production in real deployments. This heterogeneity does not merely complicate measurement, it also changes what it means to speak of "compute shocks" in policy debates, because a shock to high end accelerators may not be substitutable by an equal dollar shock to commodity processors. A further difficulty is that compute enters AI production jointly with complementary inputs.[2] If a lab attempts to compensate for slower chips by scaling out with many more devices, it may face algorithmic limits on parallelism as well as requirements for networking and systems engineering that are difficult to satisfy at scale, so the substitution margin is not frictionless even when budgets allow it. [3]

This leads us to further thinking about a pragmatic but under explored question: when compute is a bottleneck, is it the arithmetic capacity, the memory system, the interconnect, the energy envelope, or the organizational capability to operate large clusters that is binding, and how should an economist map these constraints into a coherent cost function and market structure narrative.

A recurring theme in the policy facing literature is that general purpose AI software, datasets, and algorithms are not always effective targets for controls, while the hardware required to implement modern AI systems appears, at least to some extent, more governable.

The argument is not that hardware governance is easy, but that the physicality of semiconductor supply chains, the concentration of certain high end capabilities, and the observability of large scale procurement can make hardware a comparatively tractable locus for intervention. Khan and Mann emphasize that leading edge specialized AI chips are essential for cost effective implementation at scale, and that the complex supply chains needed to produce them are concentrated in the United States and a small number of allied democracies, a configuration that creates an opportunity for export control policies. [4]

Still, this policy logic should not be treated as mechanically decisive. Supply chain concentration can create

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

leverage, yet it can also create incentives for circumvention, stockpiling, redesign, and the development of alternative architectures.[5] Moreover, the causal link from controls to outcomes such as innovation, diffusion, and welfare is rarely linear, partly because firms endogenize constraints through organizational adaptation and partly because measurement itself is unstable across time. In building the present framework, we found ourselves repeatedly revising how we classify chips and deployment contexts, not because the taxonomy is conceptually unclear.

Efficiency translates into cost effectiveness once chip costs are interpreted as the sum of production costs and operating energy costs, operating costs can dominate production costs, while even for leading nodes, operating costs are similar in magnitude to production costs. This framing is essential for economic analysis because it suggests that electricity prices, cooling constraints, and data center power density can shift comparative advantage across chip types and across regions, even when nominal chip prices are held fixed. The cost breakdown further highlights that the foundry sale price embeds capital, materials, labor, foundry R and D, and profit margin, while the designer also bears design costs and then pays for assembly, test, and packaging through specialized suppliers.

Even if one does not accept every parameter in an idealized cost model, the decomposition itself suggests where entry barriers may arise: fixed costs in design, capacity constraints at leading nodes, and learning curves in operations.[6]

2 THE LAWS OF CHIP INNOVATION AND THE ECONOMIC CONSEQUENCES OF SLOWING SCALING

For several decades, the semiconductor industry offered an unusually generous background condition for digital innovation: transistor shrinkage delivered large, fairly predictable improvements in speed and energy efficiency, and the improvement arrived broadly enough that general purpose CPUs could remain the default compute substrate for many applications. Khan and Mann explicitly note that when transistor shrinkage progressed rapidly and produced large speed and efficiency gains through the late 2000s, specialized designs were of relatively low value and CPUs dominated.

This historical observation is not simply a technological anecdote. It implies an economic regime in which the marginal return to domain specific design was often dominated by the next process generation, so that specialization could look like an expensive detour rather than a durable advantage.

The authors go further by offering a counterfactual scale statement, namely that leading chips today would contain far more transistors if the earlier pace had continued. This leads us to further thinking about a foundational shift for economic analysis. When the baseline trend of general compute improvement slows, relative advantages become more persistent, and the market can reward specialization, system integration, and scale in a way that is arguably more structural than cyclical.

2.1 DENNARD SCALING EROSION, FREQUENCY SCALING FATIGUE, AND THE EMERGENCE OF BINDING CONSTRAINTS

A companion principle to Moore's Law, often associated with Dennard scaling, suggests that shrinking transistors should keep power per unit area roughly constant, yet the report emphasizes that transistor power reduction rates slowed around 2007.

The implications show up in CPU performance history. How early speed improvements were driven by frequency scaling, then by a mixture of frequency scaling and design improvements, then by multicore parallelism, and ultimately by a marked slowdown in speed improvements in the post 2015 period.

That regime appears to have weakened. Moore's Law, as the doubling of transistor count roughly every two years, held until the 2010s, after which transistor density improvements began slowing, and the authors note that leading chips today would contain far more transistors if the earlier pace had continued, a counterfactual that is directionally informative even if node naming and realized density are not always commensurate across firms and product mixes. In practice, what looks like a slowdown in transistor scaling has also shifted part of the productivity frontier toward manufacturing operations, reliability engineering, and data governance inside fabs, because effective capacity at advanced nodes is shaped not only by lithography limits but by uptime, yield learning, cycle time, and bottleneck management, which are increasingly mediated by data and automation rather than geometry alone. For instance, Yin proposes stacking-classifier based predictive maintenance with explainable AI under a synthetic data approach, a design that is useful for fault detection and severity classification while still raising questions about domain shift when models trained on synthetic regimes confront rare, evolving failure modes in production equipment. [7]In a related operational layer, automated ETL pipelines and class imbalance handling for data quality control can plausibly reduce measurement noise and improve traceability, yet the performance of such pipelines depends on governance choices and the stability of upstream sensing systems, which are often underreported in empirical studies. [8]At the plant-flow level, a data-driven method for real-time bottleneck detection and optimization using an active-period approach and visualization suggests how "capacity" can be reinterpreted as a dynamic object rather than a static constraint, though its generalizability across product mixes and dispatching rules remains an empirical question.

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

[9]Considering these factors, the slowdown narrative should be read as a reallocation of technical effort across the stack, from device scaling toward monitoring, inference, and organizational coordination; in that sense, IoT-enabled ambient sensing combined with large language models offers an illustrative path for richer state observation, while still leaving open how such systems behave under adversarial noise, privacy constraints, and drift. [10]This shift also invites a broader economic interpretation in which supply disruptions and propagation risks become more central, and Bayesian network modeling of disruption probabilities provides one tractable way to formalize uncertainty, albeit with sensitivity to priors and network structure choices. [11]Finally, because these technical adjustments require coordination across teams and incentives across tasks, work on incentive mechanisms under unfairness-aversion preferences offers a complementary, if stylized, reminder that operational improvements are partly organizational technologies rather than purely physical ones.[12]

In such a setting, compute is less like a generic intermediate input and more like a differentiated input whose quality depends on multiple constraints at once, including power density, cooling, memory hierarchy, and interconnect. At very small scales, interconnect delays can become bottlenecks and resistance capacitance delays can increase, implying that simply shrinking transistors does not automatically translate into proportional system level speed improvements. These constraints matter because they provide microfoundations for why compute has become scarce in the specific sense relevant to frontier AI, namely scarce not only in quantity but in the quality dimensions that modern workloads actually demand.[13]

2.2 The Cost Side of Slowing Scaling: Capital Intensity, Capacity, and the Persistence of Old Nodes

If scaling were only a technological story, it would matter mainly to engineers. Its economic bite comes from cost and capacity. Fabs face great costs to build leading fabs or to upgrade existing ones, which makes an immediate transition of global capacity to leading nodes infeasible. Consequently, old nodes persist, often sold at lower prices to customers whose primary constraint is purchase cost rather than efficiency, and certain specialized low volume products remain on trailing nodes for cost effectiveness. [14]There is a subtle implication here that is easy to miss. The coexistence of many nodes implies a segmented market in which "state of the art" is not a uniform frontier, but a thin slice of capacity with distinct pricing and allocation mechanisms. That segmentation can have at least two interpretations. One interpretation is a supply side story: scarce leading node capacity makes the frontier expensive and rationed, giving large buyers and integrated platforms an advantage. Another interpretation is a demand side story: many applications rationally choose older nodes because the marginal value of energy efficiency and speed is low. [15]

In practice the two mechanisms can coexist, and empirical work that infers scarcity from observed concentration risks conflating them. Further research is needed to disentangle capacity constraints from heterogeneous demand when explaining market structure. Such examples are not definitive causal evidence, since firm exit can reflect strategy, customer composition, or capital market conditions, yet they sharpen the intuition that frontier participation is becoming less contestable, which is exactly the kind of condition under which industrial policy debates intensify.

2.3 Why This Chapter Matters for the Rest of the Framework

Considering the above factors, slowing scaling changes the economic meaning of specialization. When general purpose performance improvements decelerate, the relative payoff to AI optimized designs, specialized cores, and tightly integrated software stacks can rise, not only because they deliver higher throughput, but because they transform energy and time into binding constraints that can be relaxed. [16]

This leads naturally into the next chapters. The taxonomy of AI accelerators becomes more economically salient when specialization is valuable. Cost decomposition becomes more important when energy and capital costs rise. Market structure and policy analysis becomes more urgent when capacity constraints and capital intensity can plausibly reinforce concentration. The remaining open question is not whether Moore's Law slowed, which is widely believed, but how exactly the slowdown maps into effective compute supply, which variables best proxy that supply, and how policy shocks propagate when the underlying technology frontier is already constrained. Further research is needed to make these mappings credible.[17]

3 THE AI CHIP ZOO AND THE SYSTEM DIVISION OF LABOR BETWEEN TRAINING AND INFERENCE

A useful starting point is that "AI chips" are not merely faster versions of CPUs. A far higher degree of parallel computation than CPUs, systematic use of low precision arithmetic that remains adequate for many AI algorithms, and memory system choices that reduce the cost of moving data by storing an entire model on chip, complemented by programming languages and toolchains that translate AI code efficiently onto the device.

These features are often discussed as engineering details, yet they also provide a microfoundation for why compute is a differentiated intermediate input. A unit of compute, priced as a chip hour or a cloud instance hour, bundles arithmetic

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

throughput with memory access and software overhead, and these bundles differ materially across chip families, so the relevant economic object is not raw operations but effective work per unit cost. Even basic classification can become nontrivial once one asks whether a chip is optimized for training throughput, inference latency, energy efficiency under a power cap, or compatibility with a rapidly evolving software stack. That difficulty is not merely semantic; it suggests that economic measurement that relies on one or two headline performance metrics may be systematically incomplete.[18]

3.1 Three Core Chip Classes and Why Their Tradeoffs Are Structural

AI chips into three classes: graphics processing units, GPUs; field programmable gate arrays, FPGAs; and application specific integrated circuits, ASICs. The classification is not arbitrary. GPUs, originally designed for image processing with parallel computation, became increasingly used for AI training after 2012 and were dominant by 2017, while still retaining a relatively general purpose character compared with more specialized options. [19]

FPGAs, in turn, can be reconfigured after fabrication through programmable interconnections among logic blocks, while ASICs hardwire circuitry for specific algorithms, a design decision that typically yields higher efficiency at the expense of flexibility as algorithms evolve.

It is tempting to summarize these differences as a single continuum from flexible to efficient, at a more layered economic logic. Flexibility is valuable when model architectures and deployment requirements shift faster than capital can be redeployed, while efficiency is valuable when power budgets and operating costs are binding, and when the scale of deployment makes small energy savings aggregate into large cost differences. Under that interpretation, the AI chip zoo reflects not only technological diversity but also heterogeneity in users' objective functions, which include time to train, inference latency, total cost of ownership, and the option value of adapting to new models.[20]

3.2 Training Versus Inference as a System Level Division of Labor

Treat training and inference as distinct system tasks rather than two phases of the same computation. It notes that training involves additional computational steps beyond those shared with inference, and that the parallelism structure differs, with training often benefiting from data parallelism while inference frequently does not, especially when inference is performed on a single input at a time. This distinction matters because the binding constraint can shift. Training may be limited by throughput and distributed scaling, where interconnect and memory bandwidth become first order concerns, whereas inference may be limited by latency, energy per query, and reliability under tight power and thermal envelopes.[21]

Observed pattern in adoption. GPUs were historically dominant for training, but specialized FPGAs and ASICs have become more prominent for inference due to improved efficiency, and ASICs are increasingly used for training as well, although the risk of obsolescence remains.

If one reads this through an economic lens, a plausible interpretation is that the industry has been experimenting with where specialization is most valuable, sometimes in inference where workloads can be more stable, sometimes in training where scale advantages are decisive. Yet an alternative interpretation should be kept in view. Adoption patterns may reflect organizational complementarities, such as access to software talent, the ability to co design compilers, and the bargaining power to secure advanced node capacity, rather than purely technical superiority. Distinguishing these mechanisms is empirically difficult, and further research is needed to connect chip choice to credible measures of task specific constraints.[22]

3.3 Benchmarks, Multipliers, and the Measurement Fragility of "10 to 1,000 Times"

Because performance varies by workload, there is no common industry scheme for benchmarking CPUs versus AI chips, and that comparative speed and efficiency depend on the benchmark itself. AI chips typically provide a 10 to 1,000 times improvement in efficiency and speed relative to CPUs, with GPUs and FPGAs on the lower end and ASICs higher, and it presents Table 2 as a normalized comparison across training and inference, including generality and inference accuracy.

These estimates are informed by benchmarking studies summarized elsewhere, and it notes a data gap for FPGA training because FPGAs are rarely used for training. These limitations do not invalidate benchmarking, but they do imply that economic claims about market power, entry barriers, or compute scarcity should ideally be grounded in deployment level accounting that includes utilization, overhead, and energy, rather than relying on peak multipliers alone.[23]

The software stack as part of the specialization story. It notes that domain specific languages can be specialized to program and execute efficiently on specialized chips, and it gives TensorFlow as a notable example, while also emphasizing that specialized code libraries, such as PyTorch libraries, can package chip specific knowledge into callable functions even when the user writes in a general purpose language. [24]

This observation matters because it suggests that hardware advantage is partly an ecosystem advantage. A chip design without a mature compiler, runtime, and developer tooling may fail to translate architectural potential into

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

realized performance, which in turn implies that market outcomes can be shaped by platform strategies and by the distribution of software capability across firms and countries.

4 FROM ENGINEERING METRICS TO ECONOMICS: COST FUNCTIONS, TOTAL COST OF OWNERSHIP, AND MARKET STRUCTURE

4.1 WHY EFFICIENCY BECOMES AN ECONOMIC VARIABLE ONCE OPERATING COST IS ADMITTED

A persistent simplification in discussions of AI chips is to treat benchmark performance as the relevant output and procurement price as the relevant input, as if "faster" and "cheaper" were stable attributes that travel unchanged across workloads, datacenter environments, and software stacks. That simplification becomes difficult to sustain once operating cost is treated as first-order rather than peripheral, because chip-related expenditure is more plausibly viewed as the sum of production cost and operating energy cost, and the latter can dominate quickly for older nodes while remaining comparable to production cost even at leading nodes.

Put differently, compute is not merely purchased, it is operated, and operation is a production process whose economics depends on utilization, cooling overhead, power delivery losses, and local electricity prices, all of which can vary across firms and countries in ways that are rarely harmonized in public data.

This leads to an uncomfortable but productive observation: two actors can buy ostensibly similar accelerators and still face very different marginal costs of compute if one runs high utilization fleets under tight power envelopes and another operates intermittent workloads with low utilization, or if their energy prices differ materially. In that sense, efficiency is not only a technical metric, it is also a transmission channel from engineering design into market structure, because lifecycle affordability can create entry barriers that procurement price alone does not reveal. Still, to some extent, the analyst must acknowledge that lifecycle accounting is itself a choice of conventions, and further research is needed to standardize which conventions best approximate real economic constraints.[28]

4.2 A STYLIZED COST MODEL AND THE LOGIC OF COST DECOMPOSITION

To move beyond slogans, prior work constructs a stylized model of chip costs across nodes from 90 nm to 5 nm by holding transistor count constant at a "5 nm-equivalent" level and anchoring the reference design to a server-grade accelerator with specifications similar to Nvidia's P100.

The model is presented from the standpoint of a fabless designer that purchases foundry services and then deploys the chips in its own servers, a structure that resembles the integrated design-and-deploy strategy of large platform firms, although the extent to which smaller firms can replicate this structure is exactly part of the empirical question.

The decomposition matters because it reveals where fixed costs hide. The foundry sale price embeds capital consumed, materials, labor, foundry research and development, and profit margin; the designer bears design cost; and outsourced assembly, test, and packaging services add a separate component, so total production cost per chip is the sum of these three line items.

Operating cost is then added using an explicit electricity price assumption.

While this structure is coherent, it also forces the reviewer to confront practical difficulties that are easy to underestimate. Design cost per chip depends on volume and on whether a firm reuses designs or intellectual property blocks, so a "design cost" line item is less a constant than a scale-dependent object.

Similarly, utilization rates and facility overhead are often proprietary, which means that any numerical estimate should be read as an illustrative mapping rather than a universal parameterization.

TABLE 4.1. CHIP COSTS AT DIFFERENT NODES WITH 5 NM-EQUIVALENT TRANSISTOR COUNT

Node (nm)	Year of mass production	Foundry sale price ($/chip)	Design cost ($/chip, 5M volume)	ATP cost ($/chip)	Total production cost ($/chip)	Annual energy cost ($/chip)
90	2004	2,433	630	815	3,877	9,667
65	2006	1,428	392	478	2,298	7,733
40	2009	713	200	239	1,152	3,867
28	2011	453	135	152	740	2,320
20	2014	399	119	134	652	1,554
16/12	2015	331	136	111	577	622
10	2017	274	121	92	487	404
7	2018	233	110	78	421	242
5	2020	238	108	80	426	194

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

4.3 The Mechanics of Energy Cost and Why Assumptions Are Not Harmless

Energy cost in the same framework is not a vague add-on; it is computed using explicit assumptions about throughput, thermal design power, utilization during training, and facility overhead. The reference accelerator runs at 9.526 TFLOPS for FP32 with a 250W thermal design power; a training utilization of 33 percent is adopted; power consumption is approximated as linear in utilization, yielding an estimate of 54 percent of thermal design power during training; energy costs are then increased to account for cooling and overhead using power usage effectiveness and power supply efficiency assumptions.

These assumptions are methodologically transparent, which is a strength, yet they also reveal a point that matters for policy-facing economics: if utilization, PUE, and electricity prices differ across regions and firm types, then the same hardware can imply quite different effective compute costs.

The challenge is that many public discussions treat electricity and utilization as background constants. In practice, they are strategic variables. Fleet operators can invest in scheduling, model optimization, and systems design to increase utilization and reduce wasted power, while firms without such capabilities may be forced into lower utilization regimes that inflate average cost. This creates a plausible channel through which operational capability becomes an entry barrier, although alternative explanations remain possible, including differences in demand uncertainty, workload intermittency, and risk tolerance.

Table 4.2. Parameters used to estimate annual energy cost

Parameter	Value or assumption
Accelerator reference	9.526 TFLOPS FP32, 250W thermal design power
Training utilization	33 percent
Utilization-power mapping	Linear, implying 54 percent of thermal design power during training
Electricity price	0.07625 dollars per kilowatt-hour
Cooling and overhead	Increase energy costs by 11 percent using PUE 1.11
Power supply efficiency	95 percent

4.4 When Does It Pay to Buy the Frontier and When Does It Pay to Stay Behind

The numerical pattern in Table 4.1 offers a compact economic narrative: annual operating energy costs can quickly exceed production costs, and they do so especially aggressively for trailing nodes. One stated finding is that in less than two years the cost to operate a leading-node AI chip such as 7 nm or 5 nm exceeds its production cost, while the cumulative electricity cost of operating a trailing-node chip such as 90 nm or 65 nm is three to four times the production cost.

The same calculations suggest that leading-node AI chips are roughly 33 times more cost-effective than trailing-node AI chips once both production and operating costs are counted.

These comparisons are informative, yet they should be interpreted with care. The “33 times” figure is conditional on a particular reference design, a particular accounting of transistor equivalence, and particular utilization and facility assumptions, and it is plausible that changes in precision regimes, memory bottlenecks, or workload types would alter effective cost-effectiveness.

There is also a selection issue that is easy to overlook. Observed deployment of advanced nodes can reflect procurement access and strategic positioning, but it can also reflect rational selection where only certain workloads justify frontier adoption, so adoption data alone cannot disentangle scarcity from heterogenous demand. To some extent, both forces can coexist, and empirical work that seeks to identify compute constraints should treat node choice as an outcome of multiple margins rather than a single “technology level” indicator.

4.5 Replacement Horizons, Scale Economies, and the Market-Structure Implications of Fixed Costs

Node choice is, in this framework, an intertemporal problem. One estimate implies it takes 8.8 years for the cost of producing and operating a 5 nm chip to equal the cost of operating a 7 nm chip; below this horizon the 7 nm chip is cheaper, while above it the 5 nm chip becomes cheaper, suggesting an incentive to replace 7 nm chips only when expecting to use 5 nm chips for roughly that duration.

Yet typical replacement cycles for server-grade chips are described as about three years, a practice consistent with the cadence of new node introductions, which implies that some firms purchase newly introduced node chips as soon as they become available.

The tension is analytically useful: it suggests that private decisions may be driven not only by lifecycle cost minimization but also by competitive pressures, option value, and supply allocation dynamics, especially if access to leading-node capacity is itself rationed.

Cost decomposition also clarifies why compute markets may exhibit scale economies and concentration. Design costs are fixed and scale-dependent, and the foundry price embeds large capital and research components, while assembly, test, and packaging introduces an additional layer of specialized supply. At the foundry level, an additional empirical anchor is that revenue can be decomposed into capital consumed,

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

other costs, and operating profit, with unweighted averages indicating roughly 24.93 percent capital consumed, 39.16 percent other costs, and 35.91 percent operating profit, alongside a capital depreciation rate of 25.29 percent.

These shares are not a complete industrial organization model, yet they hint at a structural feature: when capital consumption and profit margins are material components of revenue, frontier production tends to reward scale, financing capacity, and risk-bearing, which may reduce contestability.

Considering these factors, policy instruments that target any component of the total cost stack, subsidies for fabrication and equipment, tax incentives for datacenter buildout, or constraints on advanced components, can plausibly transmit through lifecycle costs rather than through purchase prices alone, reshaping the boundary between firms that internalize fleet operations and those that must buy compute as a service. The direction and magnitude of these effects, however, remain uncertain because firms can adapt through software optimization, alternative architectures, and changes in deployment strategies, and these adaptations are not well measured in standard datasets. Further research is needed to connect policy shocks to observed changes in utilization, energy intensity, and effective throughput, ideally using research designs that can separate true supply constraints from correlated demand surges and selective disclosure in performance reporting.

5 CONCLUSION

This paper examines AI chips as an object with both technical and economic characteristics. It points out that as the scale of cutting-edge models increases, computing power is gradually becoming a key constraint. However, the economic significance of "adding one unit of computing power" is difficult to explain because chip speed, energy consumption, and cost vary depending on the workload, and training and inference also impose different system constraints. Furthermore, this paper categorizes AI chips into three main types: GPUs, FPGAs, and ASICs. It explores why benchmark-based performance multipliers are valuable but only partially comparable, and emphasizes that software toolchains and deployment environments affect actual performance. In addition, chip operating energy consumption is comparable to or even higher than production costs, and the scarcity of leading nodes may lead to high barriers to entry and market concentration. Subsidies, tax incentives, and export controls may impact the effective supply of computing power. The paper calls for more sophisticated measurement methods and causal designs to track spillover effects and welfare trade-offs under uncertainty.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

LU, Wenqiang

School of Compuer Science, Nanchang Institute of Technology, China.

REFERENCES

[1] Toews, R. (2023). The geopolitics of AI chips will define the future of AI. Horizons: Journal of International Relations and Sustainable Development, (24), 126-138.

[2] Russell, A. L., & McGravey, K. (2025). Digital oil: chips, artificial intelligence and US national security. International Affairs, 101(3), 1087-1101.

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.

[3] Leventhal, I., & Lathrop, R. (2025). AI Chips Pose Demanding Test Challenges: An Exploration of New Methodologies. IEEE Electron Devices Magazine, 3(1), 18-23.

[4] Yin, M. (2025). Predictive Maintenance of Semiconductor Equipment Using Stacking Classifiers and Explainable AI: A Synthetic Data Approach for Fault Detection and Severity Classification. Journal of Industrial Engineering and Applied Science, 3(6), 36-46.

[5] Yin, M. (2025). Data Quality Control in Semiconductor Manufacturing through Automated ETL Processes and Class Imbalance Handling Techniques. Journal of Industrial Engineering and Applied Science, 3(6), 13-22.

[6] Agrawal, A., Gans, J., & Goldfarb, A. (Eds.). (2019). The economics of artificial intelligence: An agenda. University of Chicago Press.

[7] Pang, F. (2025). Animal Spirit, Financial Shock and Business Cycle. European Journal of Business, Economics & Management, 1(2), 15-24.

[8] Mills, S. (2024). Algorithms, bytes, and chips: The emerging political economy of foundation models. Available at SSRN 4834417.

[9] Aarne, O., Fist, T., & Withers, C. (2024). Secure, Governable Chips. Center for a New American Security, Jan.

[10] Wang, J., Cao, S., Tim, K. T., Li, S., Fung, J. C., & Li, Y. (2025). A novel life-cycle analysis framework to assess the performances of tall buildings considering the climate change. Engineering Structures, 323, 119258.

[11] Yin, M. (2025). A Data-Driven Approach for Real-Time Bottleneck Detection and Optimization in Semiconductor Manufacturing Using Active Period Method and Visualization. Academic Journal of Natural Science, 2(4), 19-26.

[12] Sun, Y., & Ortiz, J. (2024). An ai-based system utilizing iot-enabled ambient sensors and llms for complex activity tracking. arXiv preprint arXiv:2407.02606.

[13] Wang, J., Tse, T. K., Li, S., & Fung, J. C. (2023). A model of the sea–land transition of the mean wind profile in the tropical cyclone boundary layer considering climate changes. International Journal of Disaster Risk Science, 14(3), 413-427.

[14] Pang, F. (2020, November). Research on Incentive Mechanism of Teamwork Based on Unfairness Aversion Preference Model. In 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (pp. 944-948). IEEE.

[15] Chen, Y. (2025). Interpretable Automated Machine Learning for Asset Pricing in US Capital Markets. Journal of Economic Theory and Business Management, 2(5), 15-21.

[16] Chen, Y. (2025). Leveraging LSTM Networks for Vehicle Stability Prediction: A Comparative Analysis with Traditional Models under Dynamic Load Conditions. Computing and Interdisciplinary Science, 1(2), 15-22.

[17] Rocha, E. M., & Lopes, M. J. (2022). Bottleneck pr

[18] Chen, Yinlei. "Daily Asset Pricing Based on Deep Learning: Integrating No-Arbitrage Constraints and Market Dynamics." Journal of Computer Technology and Applied Mathematics 2.6 (2026): 1-10.

[19] Davies, M., & Sankaralingam, K. (2025). Defying moore: Envisioning the economics of a semiconductor revolution through 12nm specialization. Communications of the ACM, 68(7), 108-119.

[20] Chen, Y. (2025). Artificial Intelligence in Economic Applications: Stock Trading, Market Analysis, and Risk Management. Journal of Economic Theory and Business Management, 2(5), 7-14.

[21] Brochado, A. F., Rocha, E. M., Almeida, D., de Sousa, A., & Moura, A. (2023). A data-driven model with minimal information for bottleneck detection-application at Bosch thermotechnology. International Journal of Management Science and Engineering Management, 18(4), 318-331.

[22] Comunale, M., & Manera, A. (2024). The economic impacts and the regulation of AI: A review of the academic literature and policy actions.

[23] Salim, M. (2025). Quantum Revolution: Redefining Industry and the Global Chip Race. Quantum.

[24] Erdil, E., Potlogea, A., Besiroglu, T., Roldan, E., Ho, A., Sevilla, J., ... & Sandler, R. (2025). GATE: An Integrated Assessment Model for AI Automation. arXiv preprint arXiv:2503.04941.

Copyright © 2026 The author retains copyright and grants the journal the right of first publication.
This work is licensed under a Creative Commons Attribution 4.0 International License.