International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02229

Moore's Law & The AI Compute Bottleneck

Vishakha Agrawal, vishakha.research.id@gmail.com

***Abstract*—Moore's Law, the prediction that transistor density on integrated circuits would double approximately every two years, has been the driving force behind advancements in computing for decades. However, as semiconductor fabrication approaches the physical limits of silicon, the traditional scaling benefits of Moore's Law are diminishing. Simultaneously, the demand for AI compute has surged, leading to an unprecedented need for high-performance chips. This paper explores the decline of Moore's law, the resulting AI compute bottleneck, and potential technological breakthroughs that could redefine the future of AI hardware.**

Keywords - Moore's Law, Dennard Scaling , AI infrastructure , Neuromorphic Computing, GPU, TPU, Quantum Computing

I. Introduction

Moore's Law has long been the foundation of computing progress, enabling exponential growth in processing power at decreasing costs [13]. However, as transistors shrink to atomic scales, see figure 1, manufacturing constraints, rising costs, and energy inefficiencies threaten to slow down or even halt this trend [5]. Meanwhile, AI models such as GPT-4o and DeepSeek-R1 require exponentially more compute resources, exacerbating the limitations of current hardware[15]. This paper examines how the slowing of Moore's Law impacts AI development and explores potential solutions to overcome the AI compute bottleneck.

II. The Decline of Moore's Law

1) The Physical and Economic Limits of Scaling : The steady march of transistor miniaturization faces critical physical and economic barriers that are reshaping the semiconductor industry's future[4]. At sub-5nm scales, quantum effects such as electron tunneling introduce significant instability, as electrons can spontaneously bypass the physical barriers meant to control their flow. This quantum behavior leads to unpredictable current leakage, threshold voltage variations, and ultimately compromises the reliability of integrated circuits, making further miniaturization increasingly difficult without fundamental changes to transistor architecture or materials science.
   Meanwhile, the cost of semiconductor fabrication at advanced nodes has skyrocketed to unprecedented levels. TSMC's 3nm and 2nm process development requires multi-billion dollar investments in extreme ultraviolet lithography equipment, ultra-pure materials, and increasingly complex multi-patterning techniques. A single leading-edge fabrication facility now costs upwards of 20 billion dollar, requiring enormous production volumes to achieve economic viability and limiting access to only the largest semiconductor companies, thus reducing economic feasibility for widespread adoption across the industry.
   Additionally, increased transistor density inevitably leads to higher power consumption and thermal management challenges that threaten to undermine efficiency gains. As billions of transistors switch states in increasingly confined spaces, the resulting heat concentration creates hotspots that can compromise structural integrity, accelerate electron migration, and necessitate aggressive power throttling. These thermal issues have contributed to the end of Dennard scaling, where newer chips often cannot run all transistors at full power simultaneously, limiting chip efficiency improvements and requiring sophisticated power management techniques to maintain performance within thermal constraints.
2) The End of Dennard Scaling : The fundamental principle of Dennard scaling, which historically allowed power efficiency to improve as transistors shrank, has largely broken down in modern semiconductor development. This scaling law, formulated by IBM's Robert Dennard in 1974, predicted that as transistors became smaller, they would consume proportionally less power while operating faster—enabling the continuous performance improvements the industry enjoyed for decades [9]. However, as feature sizes approached atomic scales below 22nm, unavoidable physical limitations emerged involving gate oxide thickness and threshold voltages that prevented voltage scaling from keeping pace with size reduction. While transistor density continues to increase according to Moore's Law—doubling approximately every two years through advanced packaging techniques and 3D integration—power efficiency gains have stagnated dramatically.
   This power wall has created a situation where chip designers can place more transistors on silicon, but cannot power them all simultaneously at full performance, leading to diminishing performance improvements per watt and the rise of "dark silicon" where portions of a chip must remain inactive to maintain thermal limits. This breakdown has forced a paradigm shift in processor architecture toward specialized accelerators, heterogeneous computing, and domain-specific designs that prioritize efficiency over raw transistor counts.

III. The AI Compute Bottleneck

1) Exponential Growth in AI Compute Needs: Modern AI models exhibit a voracious appetite for computational resources, requiring quadratically or even exponentially increasing compute power with each new generation

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02229

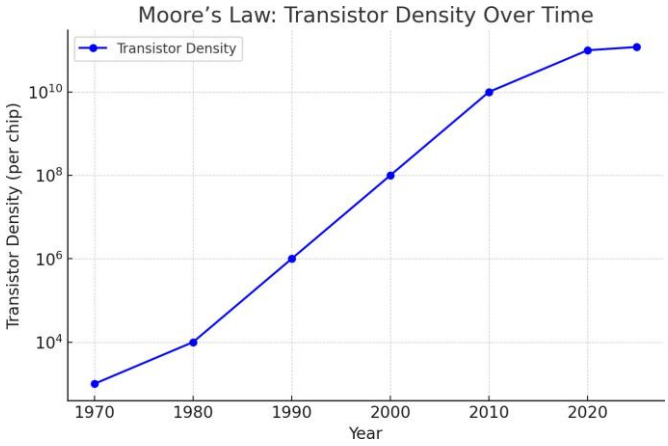

Fig. 1. Moore's Law : Transistor density over time

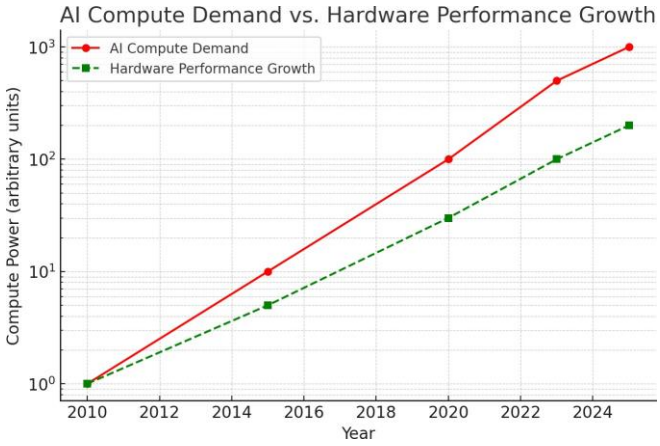

Fig. 2. AI compute demand Vs Hardware performance growth

as their parameter counts and architectural complexity grow. This scaling challenge stems from the fundamental mathematics of neural networks, where both training time and inference costs rise dramatically with model size—a phenomenon quantified in research showing that model capabilities often scale as power laws relative to compute.

The computational demands have reached such extreme levels that training large-scale AI models, such as OpenAI's GPT-4o, demands thousands of high-performance GPUs running for months, consuming vast amounts of energy. These massive training runs operate in specialized data centers with custom cooling infrastructure, often requiring hundreds of megawatts of power and costing tens or even hundreds of millions of dollars per model. The resulting carbon footprint has become a significant environmental concern [7], with single training runs potentially generating CO emissions equivalent to the lifetime emissions of multiple cars, spurring research into more efficient architectures and training methods.

Compounding these challenges, AI chip demand is outpacing production capabilities, leading to severe supply chain constraints across the semiconductor ecosystem. This shortage has created a situation where leading AI research organizations must secure chip allocations years in advance, while smaller startups struggle to access the hardware necessary for innovation, potentially concentrating AI development capabilities among a few well-resourced entities and creating national security concerns as countries compete for limited semiconductor manufacturing capacity.

2) The Energy Challenge : AI compute clusters are consuming gigawatts of electricity, raising concerns about sustainability as the environmental footprint of artificial intelligence continues to grow at an alarming rate. These massive computational facilities now rival small cities in their energy demands, with single AI research organizations sometimes requiring hundreds of megawatts of continuous power—enough to supply tens of thousands of homes[8]. The carbon impact varies dramatically depending on energy sources [11], with AI clusters in regions powered by coal or natural gas contributing significantly to greenhouse gas emissions, while those with access to hydroelectric, nuclear, or renewable energy sources fare better environmentally. This growing energy consumption has prompted increased scrutiny from environmental groups, regulators, and even AI researchers themselves[12], many of whom are working to develop more efficient algorithms and specialized hardware that can deliver comparable performance with reduced power requirements.

Data centers struggle with power limitations, and expanding AI infrastructure requires new energy solutions as existing electrical grids in many technology hubs approach capacity constraints. In areas like Silicon Valley, Dublin, and Singapore, data center operators face strict power allocation quotas, with some municipalities implementing moratoriums on new facilities due to grid capacity concerns. This infrastructure challenge has sparked innovation in several directions, including the strategic relocation of compute resources to regions with abundant renewable energy, investments in on-site power generation using solar, wind, or small modular nuclear reactors, and the development of advanced liquid cooling technologies that significantly reduce cooling-related energy overhead. Beyond technical solutions, the industry is also exploring new operational models, including workload scheduling that prioritizes non-time-sensitive AI training during periods of renewable energy abundance and innovative power purchase agreements that fund the development of new clean energy sources to offset consumption.

3) The Bottleneck in AI Hardware Supply :

The global chip shortage and increasing restrictions on semiconductor exports to China have exacerbated

ISSN: 2583-6129
DOI: 10.55041/ISJEM02229

hardware supply issues, creating a perfect storm for AI infrastructure expansion. This shortage, initially triggered by pandemic-related supply chain disruptions, has been prolonged by geopolitical tensions and export controls, particularly those imposed by the United States targeting China's access to advanced computing technologies. These restrictions have limited the flow of cutting-edge semiconductors and equipment to Chinese companies and research institutions, effectively bifurcating the global AI hardware ecosystem and forcing parallel development tracks. The ripple effects extend beyond China, as global manufacturers face increased scrutiny of their supply chains, longer lead times for critical components, and significantly higher prices for specialized AI accelerators. This situation has led to strategic stockpiling by major technology companies and nations treating semiconductor manufacturing capacity as a matter of national security, further straining an already tight market.

NVIDIA dominates AI compute hardware with its purpose-built GPU architectures and comprehensive software ecosystem, holding over 80% market share in AI accelerators and creating a near-monopoly that concerns both industry and government stakeholders. The company's specialized tensor cores, optimized memory hierarchies, and mature CUDA programming environment have created significant barriers to entry for potential competitors, allowing NVIDIA to command premium prices for its H100 and A100 chips that power most large-scale AI training. This dominance has prompted companies like AMD, Intel, and Google to invest billions in alternative solutions to reduce dependence on a single provider. AMD has accelerated development of its Instinct MI300 series[14], Google continues to expand its TPU (Tensor Processing Unit) architecture for internal use and cloud customers, and numerous startups are pursuing novel approaches including photonic computing, in-memory processing, and neuromorphic designs. These diversification efforts are critical for ensuring both competitive pricing and resilience against supply disruptions, though most industry analysts believe NVIDIA will maintain its leadership position for the foreseeable future given its substantial head start and continued aggressive investment in next-generation AI hardware

IV. Alternative Approaches to Overcoming the Compute Bottleneck

1) Specialized AI Hardware : Application-Specific Integrated Circuits (ASICs) are revolutionizing AI compute efficiency through dedicated hardware architectures optimized for the specific mathematical operations that underpin machine learning. Google's Tensor Processing Units (TPUs) exemplify this approach, with their matrix multiplication units specifically designed to accelerate the tensor operations central to neural network training and inference. These purpose-built chips achieve performance gains of 15-30x over general-purpose GPUs for certain workloads while consuming significantly less power. The TPU v4, for instance, delivers over 275 petaFLOPS of compute in a single pod configuration, enabling previously impossible model training scenarios. This specialization trend extends beyond the tech giants, with companies like Cerebras Systems developing wafer-scale chips containing trillions of transistors specifically arranged to minimize data movement—the primary energy bottleneck in AI computation.

Neuromorphic Computing represents a radical departure from conventional von Neumann architectures by mimicking the structure and function of biological neural systems[2], [3]. IBM's TrueNorth and Intel's Loihi chips implement spiking neural networks (SNNs) that process information through discrete events rather than continuous calculations, dramatically improving energy efficiency by up to 1000x for certain tasks. These neuromorphic systems excel at processing sensory data and pattern recognition while consuming mere watts of power, making them ideal for edge AI applications where energy constraints are paramount. The event-driven processing paradigm enables these chips to perform real-time learning and adaptation with minimal energy expenditure, opening new possibilities for autonomous systems that can operate without constant cloud connectivity. Research groups at universities and national laboratories are pushing this technology further by incorporating novel materials that more closely approximate biological neurons' adaptive properties.

Optical and Photonic Computing leverages light instead of electrons to perform calculations, promising transformative improvements in both speed and energy efficiency for AI workloads. Light-based processors can theoretically operate at frequencies in the hundreds of terahertz range—orders of magnitude faster than electronic systems—while generating minimal heat. Companies like Lightmatter and Luminous Computing are developing photonic chips that specialize in the matrix multiplication operations central to deep learning, with demonstrations showing 10-100x improvements in energy efficiency compared to electronic alternatives. These systems manipulate light through programmable interference patterns using components like Mach-Zehnder interferometers arranged in massive arrays. Beyond raw performance, photonic computing offers the potential for direct integration with fiber optic communication networks, eliminating energy-intensive conversions between optical and electronic domains that currently plague data center operations.

2) Advanced Semiconductor Materials : Graphene and Carbon Nanotubes represent potential silicon replacements that could extend Moore's Law beyond its current physical limits. These carbon-based materials exhibit exceptional electrical properties, including electron mo-

ISSN: 2583-6129
DOI: 10.55041/ISJEM02229

bility up to 100 times greater than silicon and thermal conductivity superior to diamond. Carbon nanotubes can form transistors with channel lengths below 5nm without the quantum tunneling effects that plague silicon at these scales, potentially enabling another decade of transistor scaling. Research at universities like Stanford and MIT has demonstrated functional carbon nanotube processors, though significant manufacturing challenges remain in precisely positioning these molecular structures. Graphene, a single-atom-thick sheet of carbon, offers even more exotic possibilities, including ballistic electron transport that could enable terahertz-frequency operation. While commercialization timelines remain uncertain, these materials represent the most promising path to continue the exponential improvement in computing capabilities that has driven technological progress for generations.

3D Stacking and Chiplet Architectures represent a paradigm shift in semiconductor design philosophy—instead of shrinking transistors further, companies are exploring vertical integration and modular approaches to improve performance without requiring smaller manufacturing processes. Advanced packaging technologies like TSMC's System on Integrated Chips (SoIC) and Intel's Foveros enable the stacking of multiple silicon dies with thousands of connections between layers, dramatically increasing transistor density while improving data bandwidth and reducing energy consumption. This approach allows mixing and matching of components manufactured using different process nodes, optimizing each function for performance or cost. AMD's EPYC processors demonstrate the chiplet concept's effectiveness, using multiple smaller dies instead of one monolithic chip to improve manufacturing yields and performance scaling. For AI applications, chiplet architectures enable the integration of specialized accelerators, high-bandwidth memory, and even optical I/O components into unified packages that can be customized for specific workloads while maintaining manufacturability and thermal management advantages.

3) Quantum Computing for AI : Quantum computers have the potential to drastically accelerate AI workloads by exploiting quantum mechanical properties like superposition and entanglement to perform certain calculations exponentially faster than classical computers. Quantum machine learning algorithms [16], [1] could transform tasks like reinforcement learning, optimization problems, and generative modeling by sampling from probability distributions and exploring solution spaces in ways fundamentally impossible for classical systems. Recent theoretical work has demonstrated quantum advantage for kernel methods and dimensionality reduction techniques central to many AI approaches. However, practical implementation remains challenging, with current quantum processors limited by qubit coherence times, gate fidelities, and error rates that restrict their application to small-scale problems. Despite these challenges, quantum AI represents perhaps the most transformative long-term computing paradigm on the horizon.

Companies like IBM, Google, and startups such as PsiQuantum are working on hybrid quantum-classical systems that leverage each paradigm's strengths to tackle AI's compute needs. These hybrid approaches utilize quantum processors for specific subroutines where they excel, such as feature mapping or optimization steps, while relying on classical hardware for the remainder of the workflow. IBM's Quantum Conditional toolkit exemplifies this strategy, providing interfaces between quantum circuits and classical neural networks. Google's quantum supremacy experiments with its Sycamore processor demonstrated the potential for quantum advantage in sampling problems relevant to generative AI. Meanwhile, PsiQuantum is pursuing a photonic approach to quantum computing that promises improved scalability over superconducting or trapped-ion systems. These diverse technical approaches reflect the nascent state of quantum AI, with no clear winner yet emerging. Industry projections suggest that practical quantum acceleration for commercial AI applications may begin to appear within 5-10 years, initially focusing on specific bottlenecks in training and optimization before expanding to more general applications.

V. The Future of AI Compute Beyond Moore's Law

1) AI-Driven Hardware Innovations : AI itself is now being used to design better chips, creating a fascinating recursive relationship where artificial intelligence improves the very hardware it runs on. Advanced machine learning systems from companies like Google, NVIDIA, and Synopsys are revolutionizing the chip design process by optimizing floor plans, power delivery networks, and routing configurations that would take human engineers months to perfect. These AI-driven design tools can explore solution spaces containing billions of potential configurations, identifying non-intuitive architectures that maximize performance within current physical constraints. Google's work on TPU [10] design using reinforcement learning has demonstrated up to 30% improvement in power efficiency compared to human-designed layouts. This approach is particularly valuable as traditional scaling approaches falter, allowing designers to extract maximum performance from existing fabrication technologies. The recursive improvement cycle—where better AI leads to better chips which enable better AI—creates an accelerating feedback loop that could dramatically speed hardware evolution beyond what traditional engineering approaches could achieve.

AI-automated semiconductor manufacturing is emerging as a key area of research as the complexity of chip fabrication reaches levels that challenge human oversight capabilities. Modern semiconductor fabs

ISSN: 2583-6129
DOI: 10.55041/ISJEM02229

contain thousands of sophisticated machines performing hundreds of processing steps, each with dozens of parameters that must be precisely controlled. Machine learning systems are now being deployed to monitor and optimize these processes in real-time, detecting subtle patterns in manufacturing data that indicate potential yield issues before they become critical problems. Applied Materials, ASML, and other equipment manufacturers are integrating ML capabilities directly into their tools, allowing for adaptive process control that continuously optimizes parameters based on incoming sensor data. These systems can predict equipment failures before they occur, automatically adjust for process drift, and even suggest novel process recipes that improve yield or performance. As feature sizes shrink to atomic scales, AI assistance becomes not just advantageous but necessary, as the physics of semiconductor manufacturing becomes too complex for traditional analytical approaches.

2) The Role of Cloud AI Distributed Compute: AI companies are shifting toward distributed training across multiple data centers to reduce reliance on single hardware clusters, improving both scalability and resilience in the face of supply chain constraints. This architectural shift requires sophisticated orchestration systems that can manage model training across geographically dispersed computing resources while minimizing the latency penalties inherent in long-distance data transfer. Companies like Meta have pioneered techniques such as their ”Sharded Data Parallel” approach, which allows models with trillions of parameters to be trained across thousands of GPUs in different physical locations. This distributed paradigm also enables more efficient resource utilization, as training workloads can be dynamically allocated to regions with excess capacity or lower energy costs. Beyond technical advantages, this approach provides strategic benefits by allowing AI development to continue even if specific hardware components become unavailable due to supply chain disruptions or export restrictions. The most advanced implementations even incorporate heterogeneous computing resources, combining different accelerator types to optimize for specific portions of the training process.
Innovations like federated learning and edge AI reduce the need for centralized compute power by fundamentally rethinking where and how AI processing occurs. Federated learning enables model training across thousands or millions of devices without centralizing sensitive data, allowing each device to contribute to model improvement while keeping raw data local. This approach, pioneered by Google and now adopted by Apple, Samsung, and others, dramatically reduces the bandwidth and centralized compute requirements for model training while simultaneously addressing privacy concerns. Concurrently, advances in model compression, quantization, and neural architecture search have enabled sophisticated AI capabilities to run directly on edge devices with limited computational resources. Qualcomm’s Hexagon processors and Google’s Edge TPUs exemplify this trend, bringing multi-TOPS (trillion operations per second) performance to devices consuming mere watts of power. These distributed approaches not only alleviate the pressure on centralized data centers but also enable entirely new applications where latency, connectivity, or privacy constraints would otherwise make AI deployment impossible.

3) Policy Industry Collaboration : Governments and private sector players must collaborate to ensure semiconductor supply chain stability as the strategic importance of advanced computing becomes increasingly apparent. The highly specialized nature of semiconductor manufacturing—where a single advanced chip might require components and expertise from dozens of countries—makes this industry particularly vulnerable to geopolitical tensions and trade disruptions. Effective collaboration requires creating resilient, diversified supply networks while addressing national security concerns through carefully calibrated export controls and technology transfer policies. Organizations like the newly formed Chip 4 Alliance (U.S., Japan, South Korea, and Taiwan) represent emerging frameworks for coordinating semiconductor policy across democratic nations. These collaborative efforts must balance competing priorities: maintaining technological leadership, ensuring economic competitiveness, protecting intellectual property, and promoting sufficient global access to prevent technological fragmentation. Industry consortia like the Semiconductor Industry Association play a crucial role in articulating technical requirements and implementation challenges to policymakers, helping to shape regulations that protect strategic interests without unnecessarily constraining innovation.
Investments in domestic chip manufacturing, exemplified by the CHIPS Act in the U.S. and similar initiatives in Europe, Japan, and South Korea, are crucial for maintaining long-term AI compute capabilities in an increasingly fractured geopolitical landscape. These programs, which commit hundreds of billions of dollars to semiconductor research and fabrication capacity, represent a significant shift from the globalized supply chain model that dominated the industry for decades. The U.S. CHIPS Act allocates $52.7 billion for semiconductor research, development, and manufacturing, including $39 billion in manufacturing incentives and $13.2 billion for R&D and workforce development. These investments aim to increase domestic production capacity for the most advanced logic and memory chips[6], reducing vulnerability to supply disruptions and foreign depen-

ISSN: 2583-6129
DOI: 10.55041/ISJEM02229

dency. However, building a robust domestic semiconductor ecosystem requires more than capital investment—it demands coordinated workforce development, university research partnerships, and supporting infrastructure. The long-term success of these initiatives will depend on sustaining investment beyond initial funding cycles and fostering the specialized talent pipeline necessary to operate increasingly sophisticated fabrication facilities.

VI. Conclusion

Moore's Law is approaching its limits, posing a significant challenge to the future of AI compute[17]. The increasing computational demands of AI models have outpaced traditional semiconductor advancements, leading to the AI compute bottleneck. While new materials, specialized AI hardware, and quantum computing offer potential solutions, the industry must also embrace alternative strategies like distributed compute, energy-efficient architectures, and AI-optimized hardware. The future of AI will depend not only on overcoming technological constraints but also on strategic investments in next-generation computing infrastructure.

References

[1] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[2] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018.

[3] Mike Davies, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A Fonseca Guerra, Prasad Joshi, Philipp Plank, and Sumedh R Risbud. Advancing neuromorphic computing with loihi: A survey of results and outlook. *Proceedings of the IEEE*, 109(5):911–934, 2021.

[4] Nicola De Liso, Serena Arima, Antonio Troisi, and Giovanni Filatrella. Semiconductors' miniaturization through time: from moore's law to eroom's law? In *2023 IEEE Nanotechnology Materials and Devices Conference (NMDC)*, pages 621–625. IEEE, 2023.

[5] Kenneth Flamm. Measuring moore's law: evidence from price, cost, and quality indexes. In *Measuring and Accounting for Innovation in the 21st Century*. University of Chicago Press, 2019.

[6] Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W Mahoney, and Kurt Keutzer. Ai and memory wall. *IEEE Micro*, 2024.

[7] Peter Henderson, Jie Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *ArXiv*, abs/2002.05651, 2020.

[8] Yu hsin Chen, Tien-Ju Yang, and Joel S. Emer. Understanding the limitations of existing energy-efficient design approaches for deep neural networks. 2018.

[9] Lennart Johnsson and Gilbert Netzer. The impact of moore's law and loss of dennard scaling: Are dsp socs an energy efficient alternative to x86 socs? In *Journal of Physics: Conference Series*, volume 762, page 012022. IOP Publishing, 2016.

[10] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*, pages 1–12, 2017.

[11] Alexandre Lacoste, Alexandra Sasha Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *ArXiv*, abs/1910.09700, 2019.

[12] David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Llu´ıs-Miquel Mungu´ıa, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *ArXiv*, abs/2104.10350, 2021.

[13] John Shalf. The future of computing beyond moore's law. *Philosophical Transactions of the Royal Society A*, 378(2166):20190061, 2020.

[14] Alan Smith, Eric Chapman, Chintan Patel, Raja Swaminathan, John Wuu, Tyrone Huang, Wonjun Jung, Alexander Kaganov, Hugh McIntyre, and Ramon Mangaser. 11.1 amd instincttm mi300 series modular chiplet package–hpc and ai accelerator for exa-class systems. In *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 67, pages 490–492. IEEE, 2024.

[15] Neil Thompson and Svenja Spanuth. The decline of computers as a general purpose technology: why deep learning and the end of moore's law are fragmenting computing. *Available at SSRN 3287769*, 2018.

[16] Fevrier Valdez and Patricia Melin. A review on quantum computing and deep learning algorithms and their applications. *Soft Computing*, 27(18):13217–13236, 2023.

[17] Nick Zhang. Moore's law is dead, long live moore's law! *arXiv preprint arXiv:2205.15011*, 2022.

