Sustainable Energy Abundance for AI

Applied Materials

Sundeep Bajikar, Aashiv Shah, Dean Ramos, Kostas Mallios, Chris Librie, Elena Kocherovsky, Nasreen Chopra, David Britz, Om Nalamasu, Tristan Holtam

Partners

Dallal Slimani (Schneider Electric), Rajiv Khemani (Auradine), Alexander Landia (The Mobility House), Junichiro Yaguchi (Kioxia), Kazuhiro Doh (Tokyo Electron), Claire Seo, Carrie MacGillivray (TechInsights), Pravin D'souza (Tata Electronics), Elizabeth Elroy (Micron), Marshall Chase (Micron), Brian Daigle (Micron), Neil Shah (Counterpoint), Herve Maury (STMicroelectronics), James Connaughton (JLC Strategies), Andy Karsner (Stanford University Precourt Institute), Peter Colohan (Lincoln Institute), Yannis Tsakiris (European Investment Bank), Handel Jones (IBS)

TABLE OF CONTENTS

INTRODUCTION

Strategic Imperative

Artificial Intelligence (AI) driven by advanced semiconductor technologies offers the single greatest lever to counteract slowing workforce growth and mounting sovereign debt in the world's largest economies. By boosting productivity at scale, AI can arrest declines in standards of living, shape equitable income distribution, and bolster political stability. Building the "computing superstructure" to support this ambition, however, will profoundly reshape global energy demand.

The incremental electricity demand burden will be so great that simply bolting on new streams of clean (or carbon-free) energy production or greatly expanding legacy sources will be insufficient and inefficient. The grid itself must be restructured with a holistic approach for sustainable abundant energy. A narrow focus on a subset of variables like emissions or "green" energy risks the creation of an unnecessarily fragile system. Scale-out of advanced computing infrastructure and the ability to cost-effectively and sustainably generate abundant supply are both vitally important to address macroeconomic, geopolitical, and climate risks.

This report outlines a new playbook to accelerate provisioning of sustainable abundant energy for AI through: (a) grid decarbonization, (b) energy-efficient computing, and (c) monetizing new opportunities for semiconductors. Grid decarbonization requires (i) more energy generation, (ii) grid modernization and scale-out, and (iii) mix shift to clean sources. Energy-efficient computing requires (i) a new playbook to optimize computing architecture, design, and manufacturing; (ii) fab solutions to raise efficiency and abatement; and (iii) fabs located where clean (or carbon-free) energy is abundant. Monetizing new opportunities for semiconductors requires a new collaboration model across the industry to enable high-velocity co-innovation.

Background and Motivation

Over the past decade, sustainability roadmaps have focused narrowly on emissions counts and renewable energy targets, often overlooking the looming surge in semiconductor- and AI-driven electricity demand. When Applied Materials launched its 2040 sustainability strategy in 2023, we recognized a critical gap – existing frameworks did not account for AI's transformative yet power-hungry trajectory. Our analysis presented in 2018 suggested that datacenter AI could account for 10-15% of global electricity consumption within a few years. After extensive discussion and debate, we established grid decarbonization as the foundational pillar of our approach. We explored semiconductor opportunities in power electronics to upgrade the grid.

This report presents a holistic strategy that realigns sustainability and decarbonization with energy abundance, ensuring AI's productivity gains are not sidelined by grid constraints. In practical terms, this entails embracing traditional (or carbon-intensive) energy sources like oil and gas while gradually transitioning toward cleaner sources such as nuclear, hydro, geothermal, solar and wind. We are forming partnerships to enhance our understanding of the upgrades needed for the electricity grid to enable rapid integration of existing and new energy sources.

When combined with progress in energy-efficient computing and abatement measures, we believe this holistic approach can lead to increased productivity while realistically accelerating the pursuit of sustainability. We believe such an approach would ensure continued growth of AI-related businesses, mitigating the risk of curtailing growth due to insufficient energy resources.

Acknowledgements

We would like to thank all our internal and external authors and partners for their guidance, insight and support for the development of this report. We would also like to thank many others who kindly supported our work with data and feedback but did not wish to be named.

Forward-Looking Statements

This document contains forward-looking statements, including those regarding trends in the semiconductor and technology markets, forecasted energy and electricity demand and consumption, compute needs and artificial intelligence workload, Applied’s playbook for reducing carbon emissions, and other statements that are not historical facts. These statements and their underlying assumptions are subject to risks and uncertainties and are not guarantees of future performance.

Factors that could cause actual results to differ materially from those expressed or implied by such statements include, without limitation: global economic, political and industry conditions; demand for semiconductor chips and electronic devices; the introduction of new and innovative technologies, and the timing of technology transitions; our ability to develop, deliver and support new products and technologies; market acceptance of existing and newly developed products; our ability to achieve sustainability strategies and targets; and other risks and uncertainties described in our SEC filings, including our recent Forms 10-K, 10-Q and 8-K filed with the U.S. Securities and Exchange Commission. All forward-looking statements are based on management’s current estimates, projections and assumptions, and we assume no obligation to update them.

THE MACRO PROBLEM:
Debt, Demographics, and Productivity

APPLIED MATERIALS®

THE MACRO PROBLEM: DEBT, DEMOGRAPHICS, AND PRODUCTIVITY

There are two major global macroeconomic challenges: debt and demographics. Higher productivity would be a potent cure for both. The chart below shows that global working-age population correlates strongly with GDP growth. Working-age population is expected to decelerate further through 2050. This poses a giant productivity challenge.

As growth in workers has slowed from +1.9% in 1965 to +0.8% in 2022, real GDP growth has slowed from +5.3% to +2.4%. Worker growth is projected to slow further with world population peaking in approximately the 2080 timeframe.

Chart 1: Demographics Challenge. Source: World Bank, Applied Materials Strategy & Market Intelligence (SMI).

Fertility rate measures the average number of live births per woman over her lifetime. A rate below 2.1 ("replacement rate") signals a shrinking population. Replacement rate is the fertility rate at which a population exactly replaces itself from one generation to the next.

The next chart shows that for all major economies, the fertility rate is trending below the replacement rate needed to keep population stable. Shrinking population = shrinking labor pools = lower GDP growth.

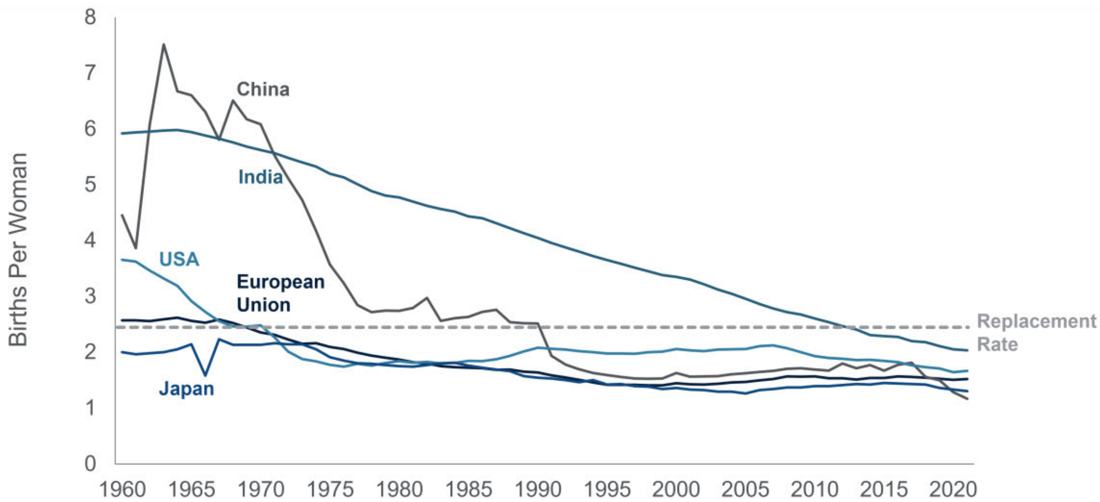


Chart 2: Fertility Rate < Replacement Rate. Source: World Bank.

The chart below shows year-over-year productivity correlating with GDP for G7 nations. Assuming AI can boost productivity, it has the potential to drive GDP growth.

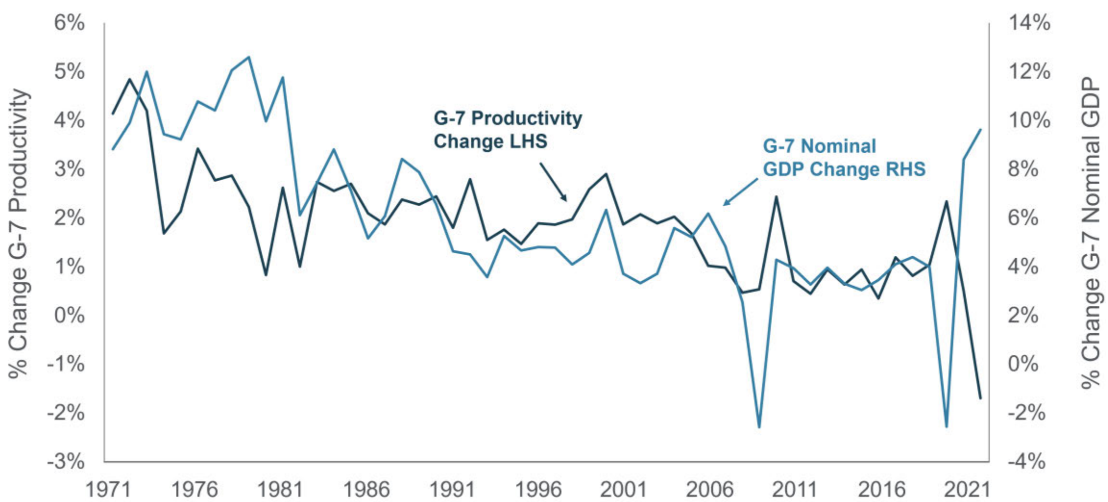


Chart 3: Productivity vs. GDP. Source: World Bank.

The next chart shows U.S. debt as a percentage of GDP. Gross U.S. debt as a percentage of GDP is now even higher than its peak during World War II, and likely to increase further driven by government spending.

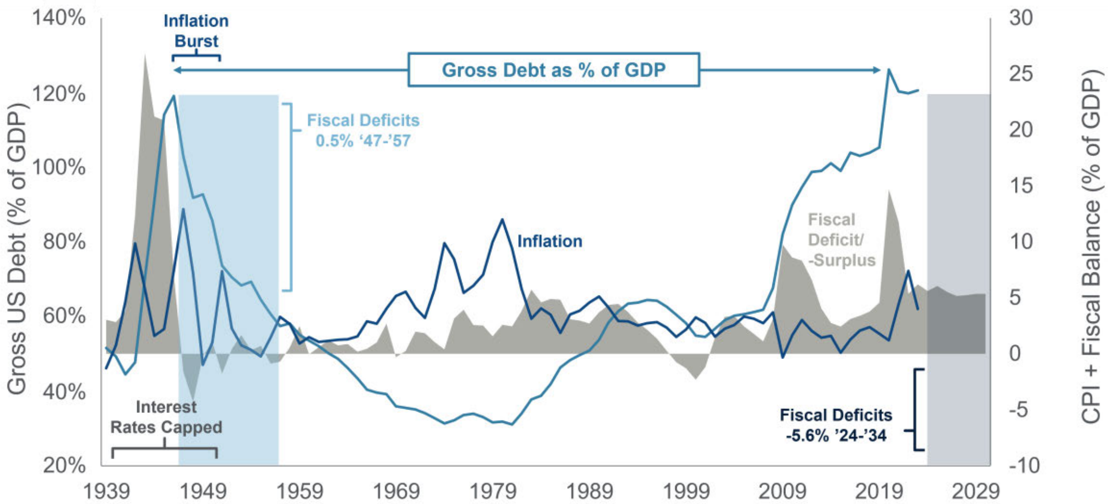


Chart 4: US Debt Challenge. Source: CBO, IMF WP 2024, Federal Reserve, Applied Materials SMI.

The chart below shows Marginal Revenue Product of Debt (MRPD). MRPD measures additional GDP generated per dollar of new government debt and is a measure of productivity (or diminishing returns) of fiscal spending. MRPD has been decreasing. In other words, each dollar of additional debt is generating diminishing returns due to lower productivity.

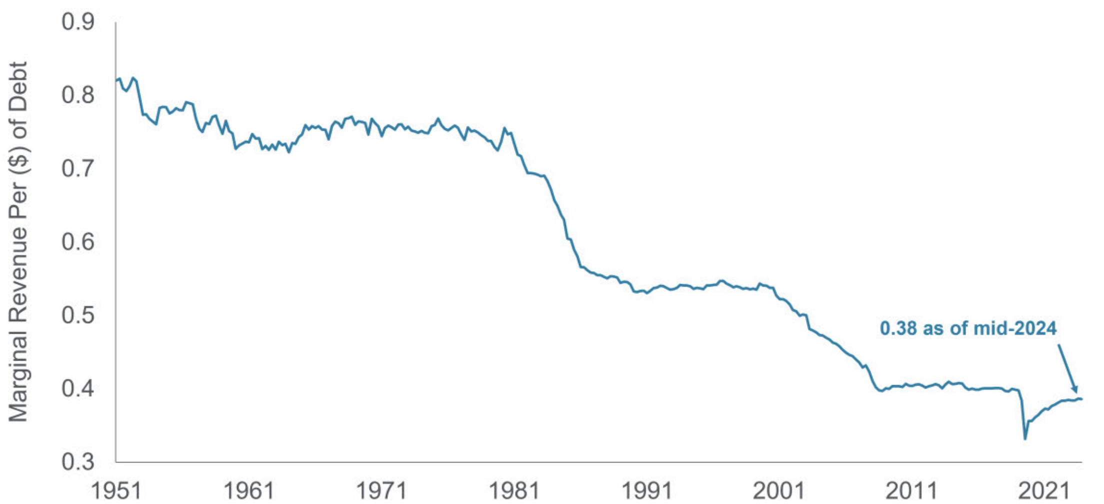


Chart 5: US MRPD. Source: Factset, Applied Materials SMI.

The next chart shows the growth of global debt. Debt growing faster than GDP puts pressure on GDP growth, and can fuel a debt spiral due to higher interest payments on debt taking on a higher share of government spending. When total debt expands faster than GDP, each dollar of borrowing yields progressively smaller increments of economic output. Higher debt levels mean larger interest payments, especially at higher interest rates. As debt service escalates, fiscal space narrows. Public and even private sector borrowing costs can rise -- if markets demand higher yields -- diverting capital away from productive investments.

Slower growth undermines tax revenues, widening deficits and prompting further borrowing. Higher interest rates or austerity measures can further suppress growth. This feedback loop, where debt issuance is driven by the need to sustain growth while simultaneously eroding growth prospects, creates a debt spiral that is both hard to arrest and detrimental to long-term economic prosperity.

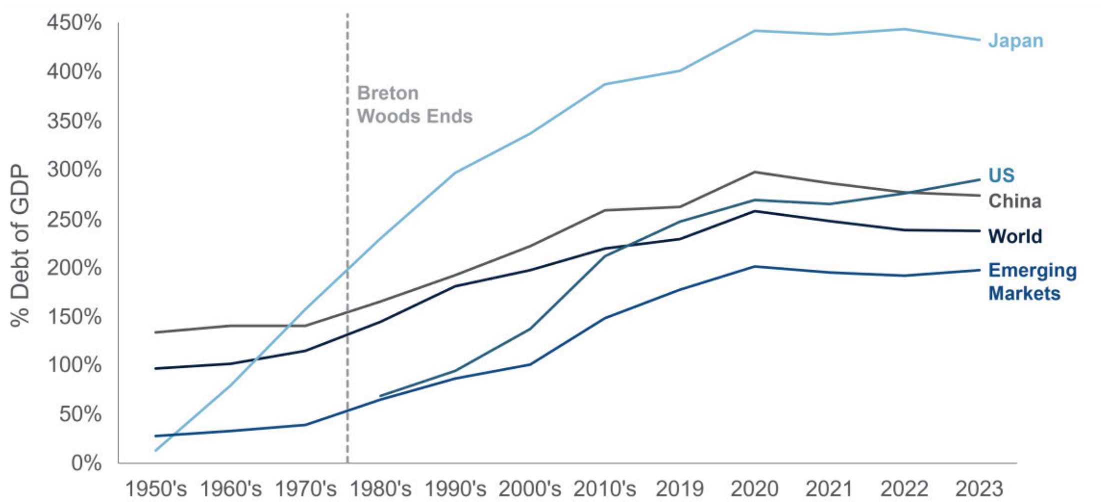


Chart 6: Global Debt Challenge. Source: IMF, Applied Materials SMI.

With demographics and debt conspiring to structurally lower productivity, it's imperative to think about a way to sustainably enable AI's vast and growing energy footprint. This requires us to rethink our approach towards decarbonization and sustainability to maximize investments in productivity-enhancing technologies like AI.

ACCELERATING SUSTAINABLE ENERGY ABUNDANCE FOR AI

APPLIED MATERIALS ®

ACCELERATING SUSTAINABLE ENERGY ABUNDANCE FOR AI

The conventional approach to achieving sustainability and decarbonization focused primarily on expanding renewable energy like wind and solar, while rapidly phasing out traditional (or carbon-intensive) energy sources. This strategy cannot sustainably support the enormous energy requirements demanded by the growth of AI. Likewise, improvements in energy-efficient computing alone will not sufficiently curb overall electricity use, as advances in efficiency inevitably lead to broader adoption across new applications, resulting in increased total energy consumption. To truly harness the transformative productivity and prosperity potential of AI, we need a comprehensive and pragmatic sustainability strategy that expands clean (or carbon-free) energy capacity, modernizes energy infrastructure, and maintains diverse energy generation sources to ensure stable, abundant power for continued technological innovation.

Conventional Approach to Sustainability

With the adoption of AI, the technology share of electricity consumption is set to skyrocket over the next 10-20 years. We forecast that electricity consumption by technology applications will increase from approximately 10% in 2021 to 25% in 2030 and 50% in 2050.

While AI is a significant driver of the increase in technology electricity consumption, the transition to electric vehicles, the move to digital, and the proliferation of IoT all play a role in the growth of technology electricity consumption.

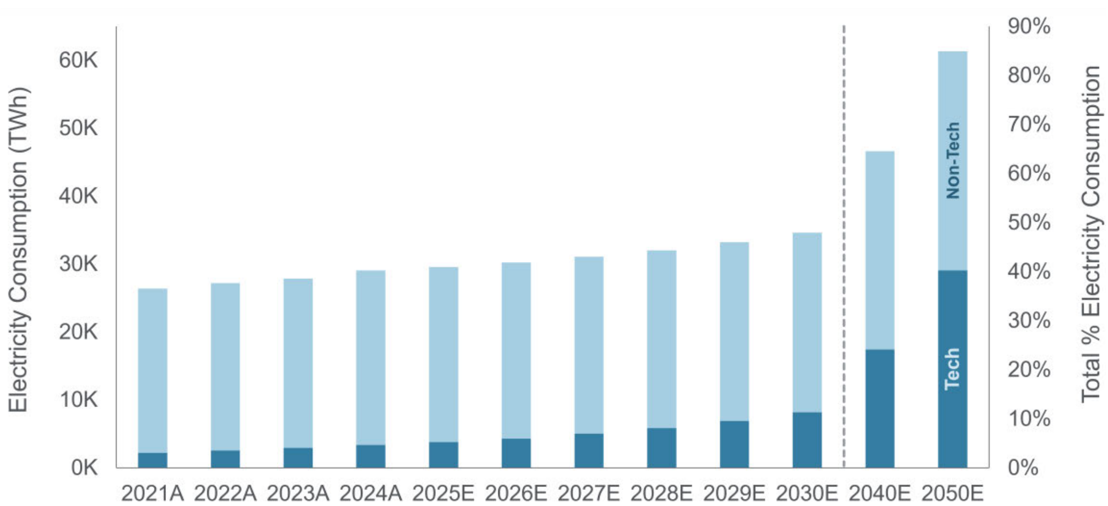


Chart 7: Power Consumption by Tech vs. Non-Tech. Source: Applied Materials SMI.

Using the U.S. Energy Information Administration's (EIA) reference scenario (Oct. 2023) for electricity supply, including the retirement of coal power plants, we estimate that electricity demand could be higher than supply as soon as 2030, if significant changes are not made to how energy supply is optimized.

Year	Total Demand (TWh, K)	Total Supply (TWh, K)
2030	34,628	31,739
2040	46,621	36,744
2050	61,350	42,298

Chart 8: Total Electricity Demand vs. Supply. Source: EIA (October 2023), Applied Materials SMI.

Datacenter AI Computing Energy Model

Increasing complexity and the growing number of parameters are driving exponential growth in computational requirements for training new Large Language Models (LLMs). Over the past four years, the compute requirement for training has increased over 10,000x, while the cost of training has increased even faster.

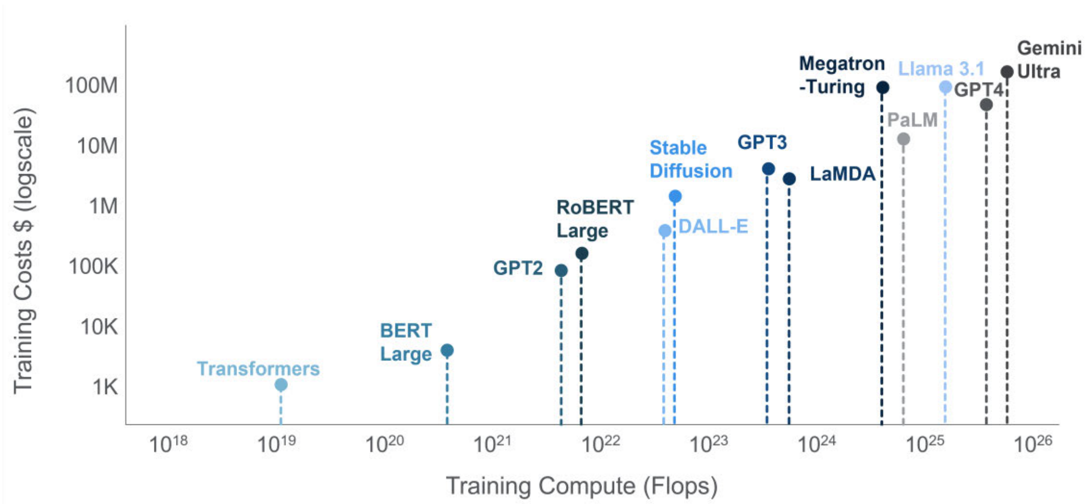


Chart 9: Computing requirements growing exponentially. Source: Artificial Intelligence Report 2024, Stanford University.

> **"The amount of computing needed to train Llama 4 will likely be almost 10x more than what we used to train Llama 3, and future models will continue to grow beyond that."**
> *— Mark Zuckerberg, CEO, Meta Platforms and Facebook (July 30, 2024)*

While multi-modal LLMs will need more compute, we believe new workloads such as humanoid robots and autonomous driving with high accuracy requirements will also drive increasing needs for training and compute beyond what we currently see. We are using CNNs as a proxy, but we note that we don't know yet what kind of models will be prevalent in the future.

No.	2016	2022	2030
MLP	61%	24%	15%
RNN/ LSTM	29%	2%	0%
CNN	5%	12%	20%
Transformer:	**0%**	**58%**	**65%**

Chart 10: AI Market Segmentation by Model Type. Source: Google, Applied Materials SMI.

As AI transitions to more multi-modal workloads, coupled with higher accuracy requirements, we anticipate that compute needs -- and thereby power consumption -- will increase exponentially. The chart below shows our forecast for datacenter electricity consumption to increase from 2% of global electricity consumption today to 12% by 2030.

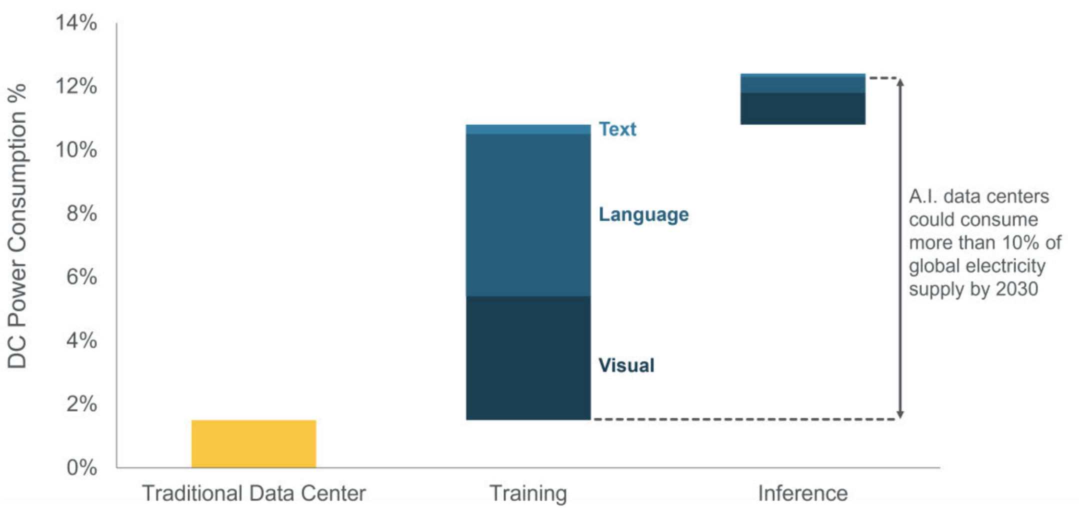


Chart 11: Datacenter power consumption as % of global electricity. Source: Applied Materials SMI.

As AI adoption increases, there will be a mix shift in electricity consumption and energy usage between traditional and AI datacenters. Over the next few years, AI datacenters will be the biggest driver of electricity usage for the datacenter category.

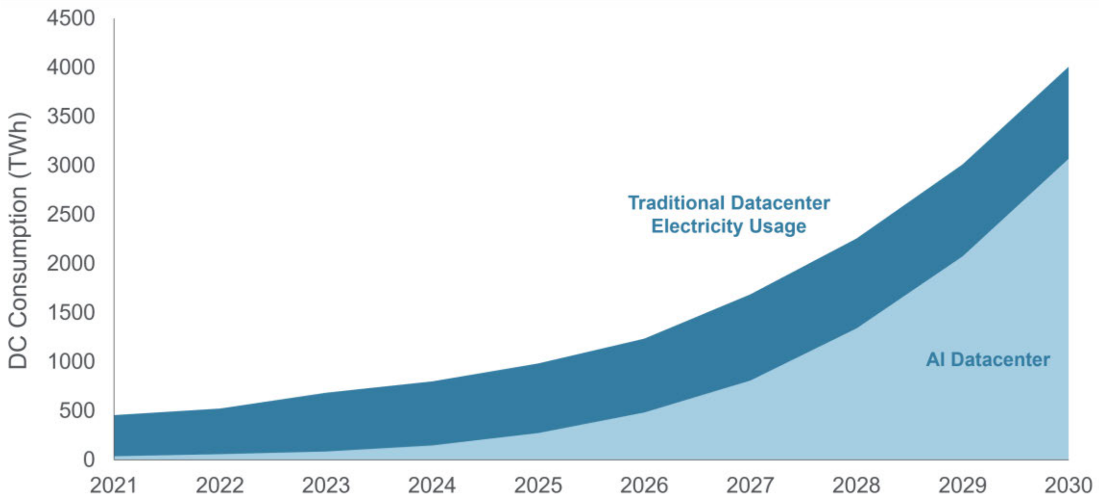


Chart 12: Datacenter Energy Use. Source: Applied Materials SMI.

Our estimate of AI datacenter electricity usage increasing to 10% of global electricity usage by 2030 is higher than some of the other estimates, likely due to our assumptions around the evolution of AI workloads. Note, IBS does have a higher estimate than Strategy & Market Intelligence (SMI), but underlying assumptions might vary.

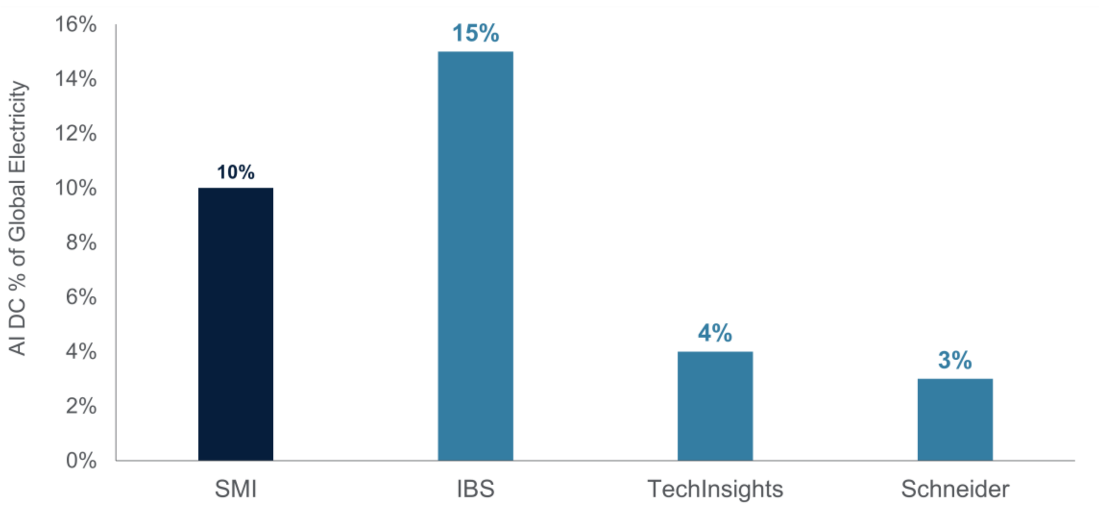


Chart 13: AI datacenter power consumption estimates. Source: Applied Materials SMI, TechInsights, Schneider Electric.

While AI datacenter consuming 10% of global electricity might seem like a high estimate, it is instructive to consider that according to a 2022 study, lighting alone consumes around 15% of global electricity. If AI is viewed as a significantly more productive technology, then it should be enabled to consume as much or more electricity to address global macro challenges.

Below, we show Schneider Electric's scenarios around AI and electricity consumption between 2025 and 2035. While all scenarios show an upward trend in AI energy usage, assumptions around AI-ready infrastructure cause scenario-specific shifts from 2027 and 2028. The energy crunch scenario is most interesting, as it shows how limited electricity supply can stunt AI development for multiple years. Note, the difference between AI electricity usage in Chart 12 verses Chart 14 is based on differing workload scenario assumptions used by the respective sources.

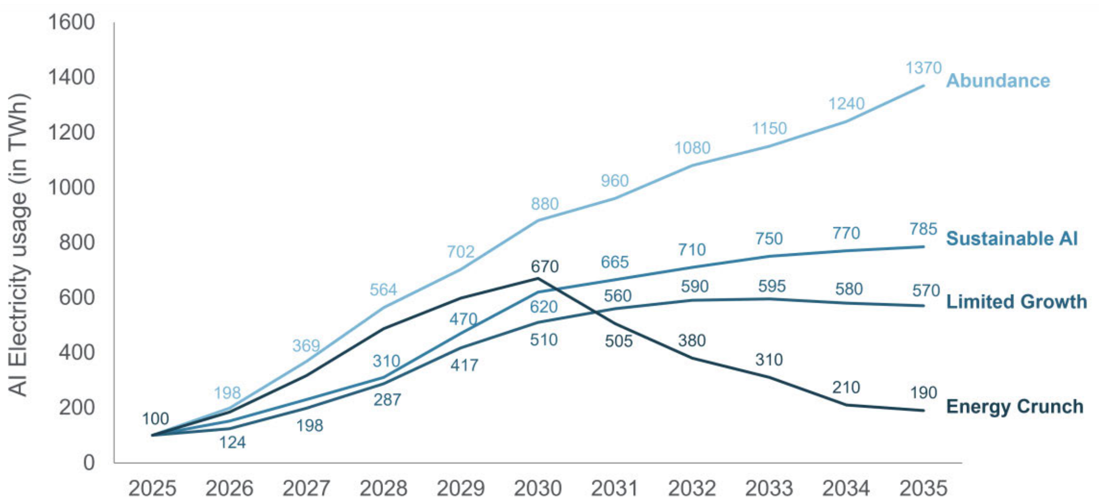


Chart 14: Global AI Electricity Use Forecasts. Source: Schneider Electric.

Datacenter Carbon Emissions

Translating electricity consumption to carbon emissions, we note that lifecycle carbon emissions will be driven by both the manufacturing and operation phase of AI accelerators and other components.

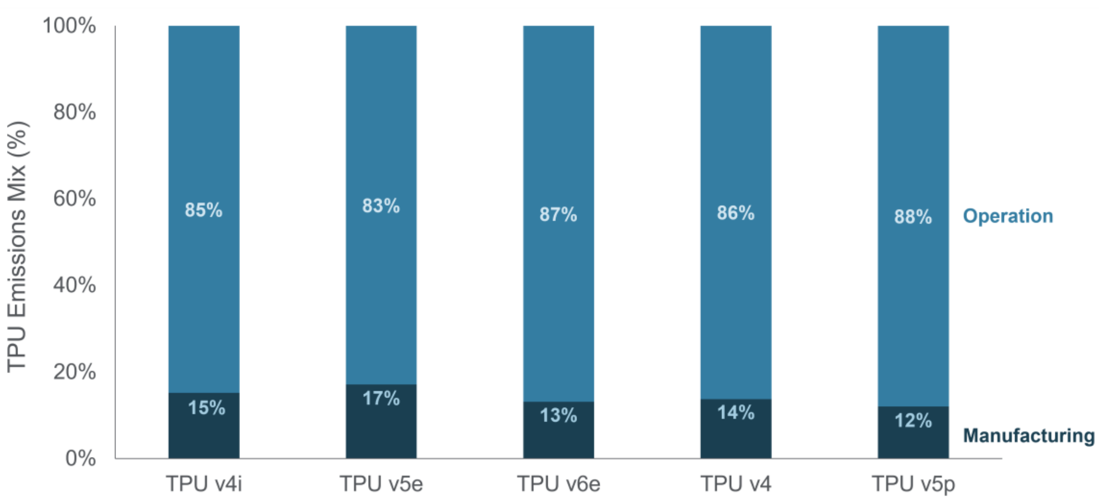


Chart 15: Tensor Processing Unit Carbon Emissions. Source: Google, Applied Materials SMI.

As GPUs become more powerful (500x increase in Tera Operations per Second (TOPS) over the last five generations), the rate of emissions per unit is increasing as well. As such, lowering emissions during both the operations and manufacturing phase will be important.

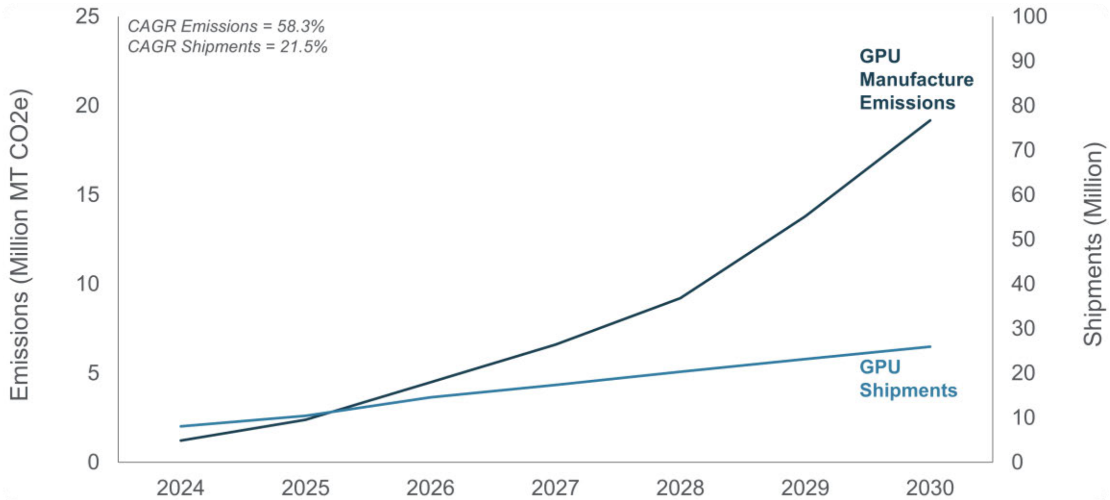


Chart 16: AI Chip Manufacturing Emissions. Source: TechInsights.

Solutions for Efficiency and Abatement in Manufacturing

Third-party studies have forecast that semiconductor manufacturing carbon emissions will reach 200M tons by 2030. This is the sum of both Scope 1 and Scope 2 emissions, with an underlying assumption of no changes in the current operating environment.

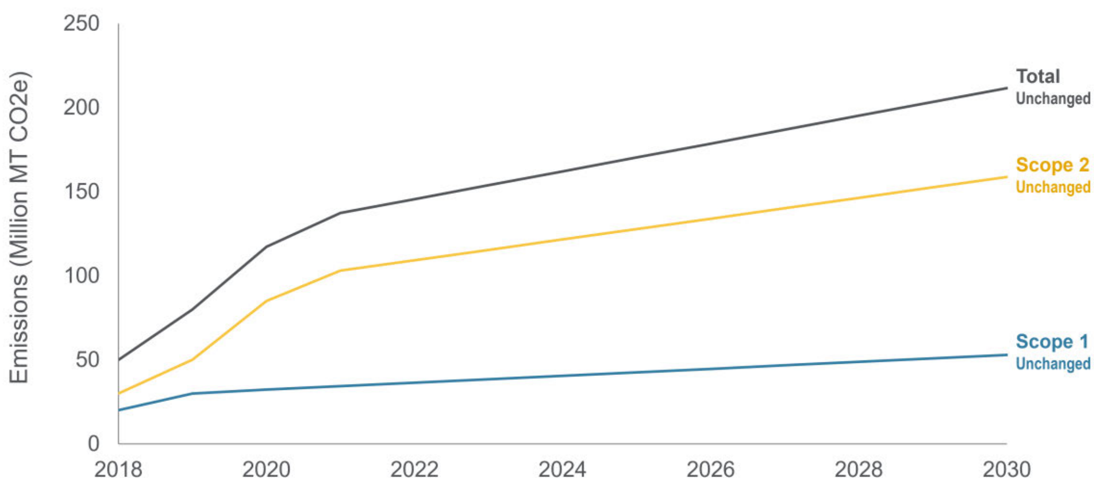


Chart 17: 2030 Current Trend. Source: McKinsey and Company, SEMI, BCG, TechInsights, Applied Materials, Imec.

To reduce carbon emissions from the manufacturing phase, grid decarbonization will be key. We believe total emissions will be lower by 25% to 150M tons in 2030, if the industry executes on current grid decarbonization plans. But the industry will need more than just grid decarbonization to minimize emissions from manufacturing.

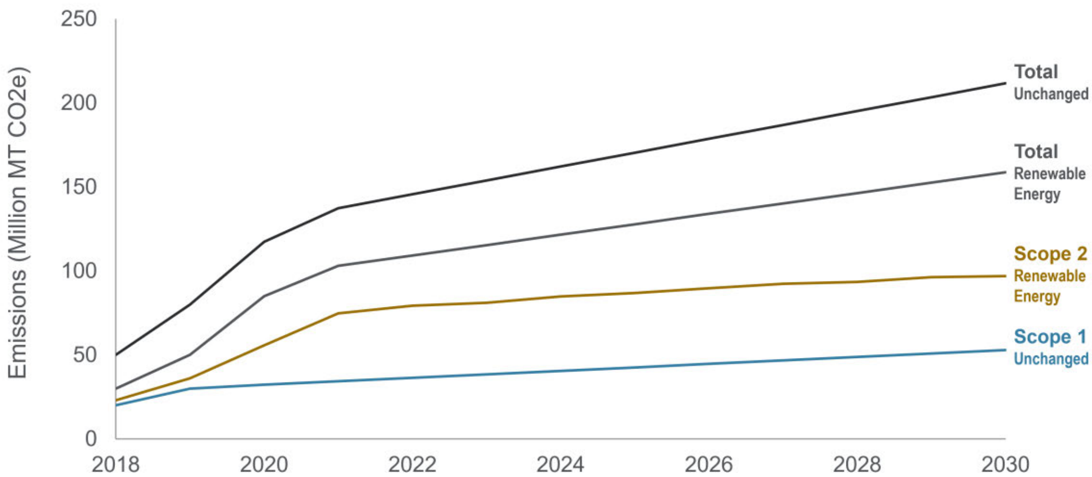


Chart 18: 2030 Challenge for Semi Fabs is 150M tons. Source: McKinsey and Company, SEMI, BCG, TechInsights, Applied Materials, Imec

In addition, there is an opportunity to reduce industry emissions to less than 100MTons with adoption of key technologies and using a minimally disruptive approach. The minimally disruptive approach focuses on four areas for near-term fab implementation:

1) Green implementation
2) Energy and utility management
3) Recipe optimization
4) High-efficiency abatement

Given the significant growth in AI demand and consequent rise in demand for semiconductors, adopting minimally disruptive solutions as soon as possible to minimize emissions while sustaining expanded manufacturing and operations will be critical.

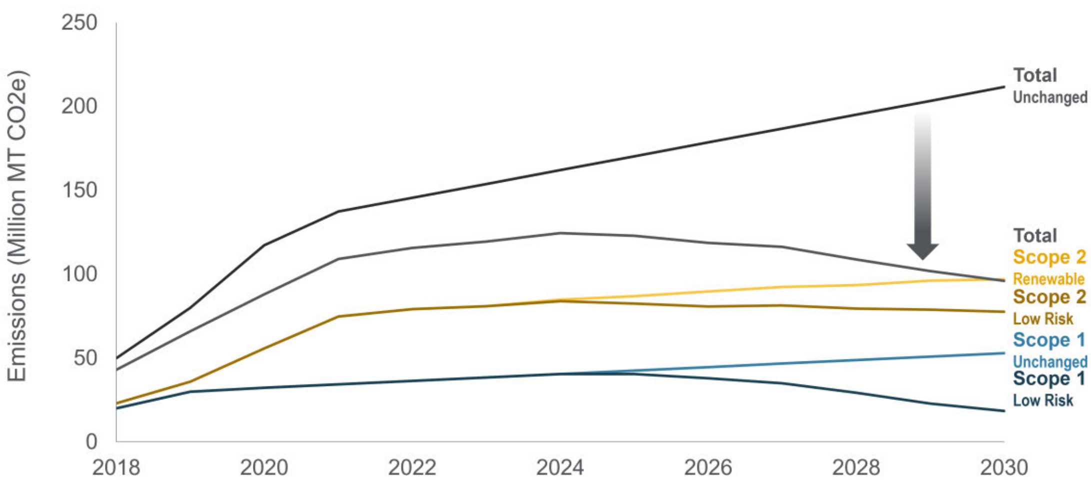


Chart 19: Path to <100M Tons per year. Source: McKinsey and Company, SEMI, BCG, TechInsights, Applied Materials, Imec.

A New Playbook for Sustainable Energy Abundance

Our proposed new playbook for accelerating provisioning of sustainable abundant energy for AI is composed of three pillars: (a) grid decarbonization, (b) energy-efficient computing, and (c) monetizing new opportunities for semiconductors.

Pillar A: Grid Decarbonization

Consuming electricity from the grid assumes the emissions profile of the grid itself. It should come as no surprise that emissions related to the grid represent the single biggest component of the emissions bill facing any given company. Still, in the conventional approach to sustainability, companies focused more on offsetting emissions derived from the grid rather than sourcing the grid with cleaner energy. Our new playbook prioritizes clean (or carbon-free) energy grid infrastructure for accelerating AI.

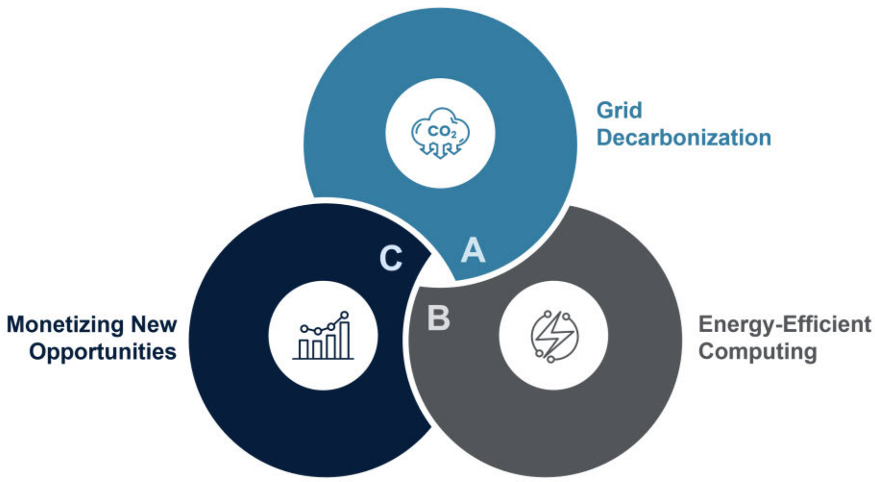


Chart 20: New Playbook – Sustainable Energy Abundance for AI. Source: Applied Materials SMI.

Pillar B: Energy-Efficient Computing

Energy-efficient computing maximizes computing intensity (measured in Tera Operations Per Second, or TOPS) per unit of power. This is done by optimizing computing architecture, as well as semiconductor design and manufacturing related to provisioning the chosen computing architecture.

Pillar C: Monetizing New Opportunities for Semiconductors

In the conventional approach, sustainability was largely regarded as a compliance checkbox item, with unclear ties to business results. In our new playbook, sustainability and decarbonization are expected to accelerate growth of AI and the future of computing while creating and shifting incentives toward generation, transmission, distribution, and storage of clean (or carbon-free) energy, along with modernization of the electricity grid – all of which are large and growing markets for semiconductors.

PILLAR A:
Grid Decarbonization

PILLAR A: GRID DECARBONIZATION

We believe decarbonization of the electricity grid is the single most powerful lever for semiconductor, AI computing, and other technology companies to reduce their overall emissions footprint. It not only addresses Scope 1 and Scope 2 emissions -- often the largest portion of a technology company's direct operational emissions -- but also tackles a large portion of lifecycle emissions due to electricity consumption in downstream Scope 3 emissions.

In this section we take a deeper look at the components of Applied's Scope 3 emissions, highlighting grid decarbonization as the largest component. We note that the location of end use by semiconductor fabs or datacenters is the key underlying driver of grid emissions. We contemplate that strategically locating such key infrastructure in areas where the grid can be powered abundantly by clean (or carbon-free) energy would accelerate scale-out of AI.

Components of Decarbonization – 2024 Update

Applied Materials has laid out its baseline emissions for 2040, along with the factors that impact it the most. The biggest component of emissions is grid decarbonization, which is based on how clean or dirty the grid is in the locations where Applied's equipment is being used.

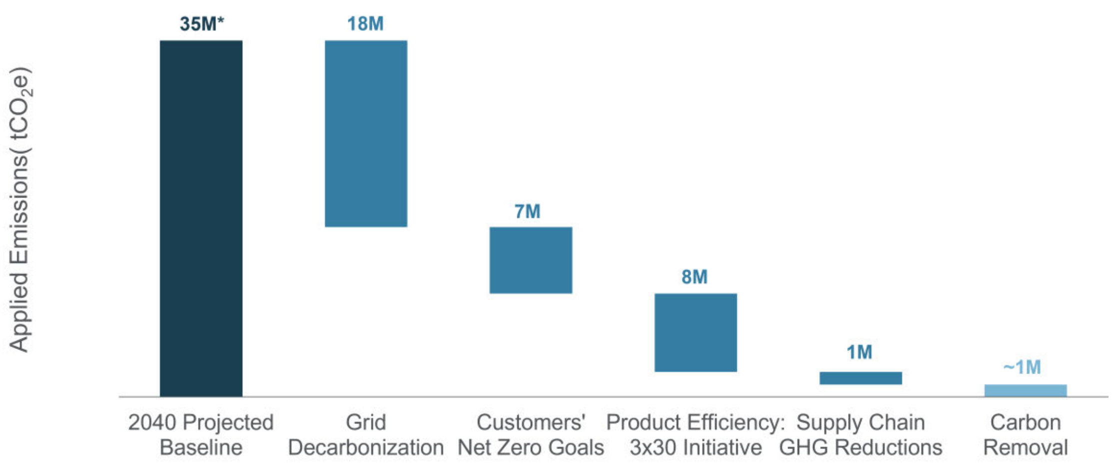


Chart 21: Components of Decarbonization – 2024 Update*. Source: Applied Materials.

**Covers 90% of Applied's Scope 3 per SBTi (90% product use emissions; 10% supply chain). GHG = Greenhouse Gas*

Locating Datacenters and Semiconductor Fabs Strategically

As companies look to minimize emissions, access to a clean grid will be one of the most important aspects in reducing carbon footprint. As such, strategically choosing datacenter and fab locations will be important.

The table below shows the relative mix of clean (or carbon-free) energy in the power grid as well as the carbon emission factor by country. The country emission factor shows the direct carbon emission per kWh of electricity across different countries, with a higher number showing higher emissions despite a similar renewal energy mix. For example, the United States has a country emission factor of 331 with a RE mix of 22%, while France has a country emission factor of 41 with a RE mix of 27%. This is because the majority of electricity in France is generated through nuclear power (with no direct carbon emissions), while the United States still gets a significant percentage of electricity supplied through coal and natural gas. In addition, we note there could be significant differences within a country such as the United States, with states like California having a higher mix of renewables compared to some other states.

Country	Number of datacenters	Share of RE in power generation 2023	Country Emission factor in gCO2/kwh electricity, 2023 (IEA database)
USA	3062	22%	331
Germany	406	52%	329
United Kingdom	377	48%	176
India	248	21%	748
Canada	248	67%	117
France	244	27%	41
Australia	243	34%	580
The Netherlands	178	47%	229
Brazil	161	89%	65
Indonesia	142	18%	777
Russia	90	18%	349

Chart 22: Datacenters vs. share of renewable energy generation. Source: Datacenter Map, Applied Materials SMI.

Cloud Service Providers Going Nuclear

The importance of grid decarbonization is demonstrated by the recent announcements from major Cloud Service Providers to power their datacenters through clean (or carbon-free) energy. We note the different CSP partnerships with nuclear energy providers (traditional and small modular reactors), with sustainability commitments being the key driver behind the new partnerships.

Cloud Service Providers	Nuclear Partners	Quotes
Microsoft	Constellation Energy	"This agreement is a major milestone in Microsoft's efforts to help decarbonize the grid in support of our commitment to become carbon negative." *— Bobby Hollis, VP of Energy, Microsoft (Sept 20, 2024)*
Google	Kairos Power	"This agreement is a key part of our effort to commercialize and scale the advanced energy technologies we need to reach our net zero and 24/7 carbon-free energy goals and ensure that more communities benefit from clean and affordable power in the future." *— Michael Terrell, Energy and Climate Sr. Director, Google (Oct 14, 2024)*
Amazon	X-energy	"Nuclear is a safe source of carbon-free energy that can help power our operations and meet the growing demands of our customers, while helping us progress toward our Climate Pledge commitment to be net-zero carbon across our operations by 2040." *— Matt Garman, CEO of AWS, Amazon Web Services (October 16, 2024)*
Oracle	NA	"So, we're in the middle of designing a datacenter that's north of a gigawatt, that has – the location and the power place we've located, they've already got building permits for three nuclear reactors. These are the small modular nuclear reactors to power the datacenter. This is how crazy it's getting. This is what's going on." *— Larry Ellison, Chairman, CTO, and Co-Founder, Oracle (Sept 09, 2024)*

Chart 23: Cloud Service Providers entered into agreements with Nuclear Power companies. Source: Company Reports.

In addition to the nuclear energy partnerships, we note that CSPs have started taking other steps to account for the increase in AI workloads. Google recently announced a partnership with Carrier Global to drive more efficient energy usage through enhanced grid flexibility and smarter energy management. Another Google partnership with Tapestry and PJM Interconnection is expected to bring energy capacity on the grid faster using AI, which could be particularly impactful given the large concentration of the United States' data centers in PJM.

CSPs are also looking into geothermal energy as another renewable source -- Google has announced a collaboration with Project Innerspace and SLB to accelerate geothermal energy adoption. Meta is partnering with Sage geosystems, while Microsoft announced an investment in G42 to build a datacenter powered by geothermal energy in Kenya.

Beyond locating facilities near clean (or carbon-free) energy sources, grid decarbonization can be advanced when companies analyze and consider good stewardship of the water footprint of the facility, which helps mitigate against regional water scarcity risks that periodically endanger local energy systems and to reduce local energy demand more generally, as water savings translate into energy savings.

Regional dependence on hydropower or forms of thermal energy generation that require substantial water cooling can be threatened by shocks from drought and low flow conditions, leading to costly blackouts and other disruptions. The OECD has found that it is more cost-effective to hedge energy systems against the risk from such water shocks than it is to recover from them.

Datacenter investors, therefore, should consider deploying water-saving cooling technologies like geothermal cooling or liquid cooling, versus more water-intensive practices like evaporative cooling, particularly in water-scarce regions. Even when liquid cooling can be more energy intensive than evaporative cooling at the site level, the tradeoff should be analyzed carefully at the regional level for opportunities to mitigate against costly water shocks to regional energy systems.

By lowering facility demand for water, datacenters also reduce the very high level of energy that water systems need for treating and delivering water. AI investors should also consider partnering directly with water providers near data center sites to create more energy-efficient water treatment and distribution, further advancing decarbonization.

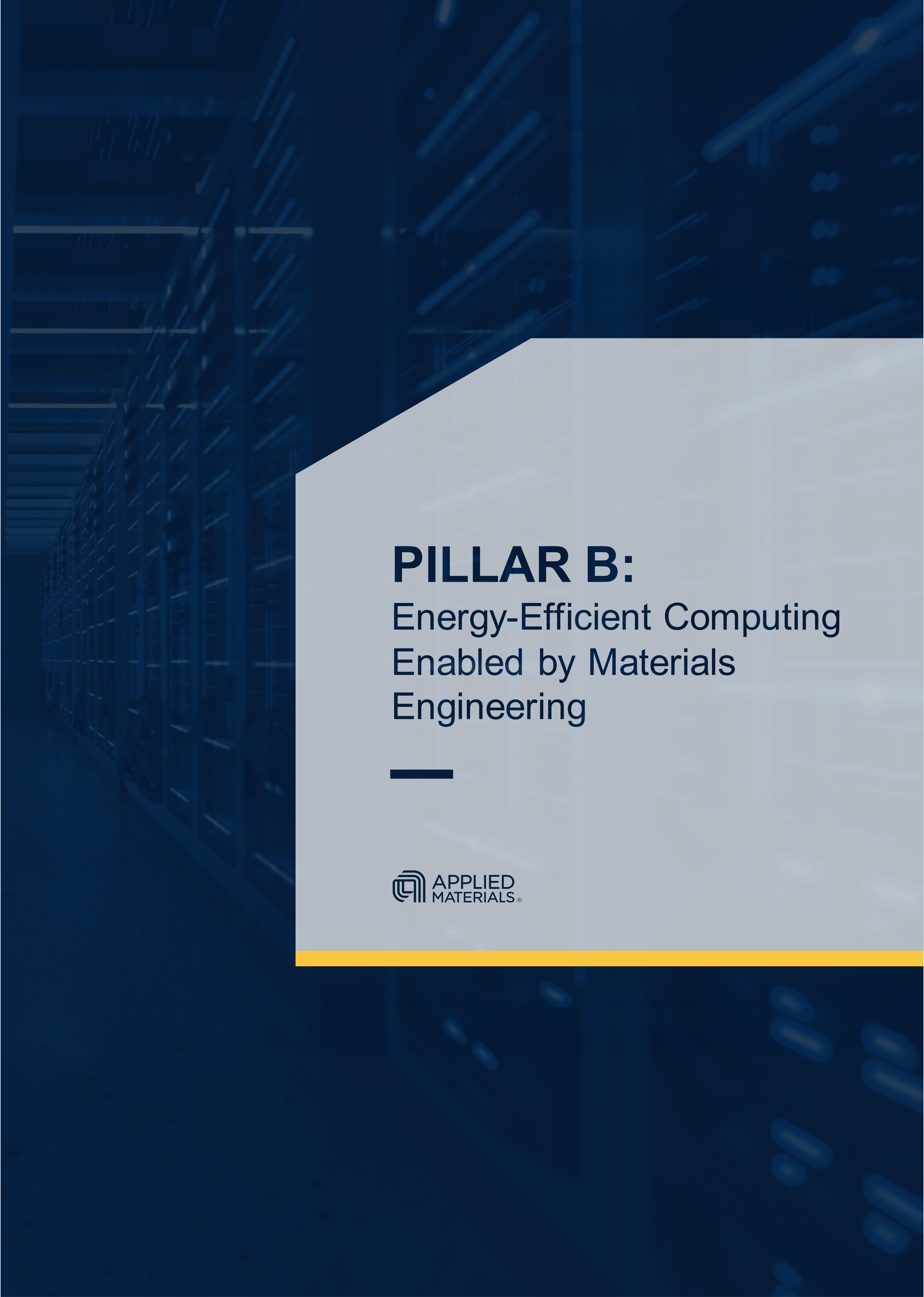
Grid Strategy

Many technology companies rely on Power Purchase Agreements (PPAs) and Virtual Power Purchase Agreements (VPPAs), and empirical studies suggest these agreements spur significant new wind and solar capacity additions. However, PPAs focus on energy generation and do not fund grid upgrades, transmission expansion, or smart grid deployments needed to deliver clean power reliably at scale.

A truly resilient grid strategy must blend three levers:

a) Generation scale-out: expand capacity for nuclear, wind, solar, and traditional forms of energy with a mix shift to cleaner sources over time
b) Transmission infrastructure modernization: upgrade lines, substations and control systems
c) Smart distribution: deploy digital sensors and AI-driven load management

By synchronizing these elements, we can deliver abundant power where and when AI needs it most in a way that is optimized to reduce emissions. The pace at which new wind, solar, nuclear, and other types of projects come online is too often bottlenecked by lengthy permitting reviews at federal, state, and local levels. To meet AI's accelerating energy demands, regulators should adopt fast-track approval pathways for projects that meet appropriate criteria.



PILLAR B:

Energy-Efficient Computing Enabled by Materials Engineering

APPLIED MATERIALS®

PILLAR B: ENERGY-EFFICIENT COMPUTING ENABLED BY MATERIALS ENGINEERING

New Semiconductor Playbook

In early 2017 we hypothesized that there would be an explosion of data generation from a variety of new categories of devices, most of which didn't even exist at the time. Data would become increasingly valuable as AI turned data into commercial insights. We hypothesized that to realize AI, the industry would need to enable new computing architectures. We further hypothesized that a new semiconductor playbook driven by materials engineering would be needed for delivering new computing architectures for AI and the future of computing.

In the PC and mobile eras, which spanned more than 20 years, the industry was driven by "classic 2D scaling," which refers to geometric transistor scaling as predicted by Moore's Law in 1965. As transistor features became smaller, performance increased and power decreased consistent with Dennard's Law. Higher transistor density also had the effect of lowering cost-per-transistor such that total chip cost decreased. During this period, we didn't see a lot of architecture innovation, as a lot of focus was put on geometric scaling of prevailing architecture, and this reinforced the rise of the x86 and Arm CPU architectures.

Classic 2D scaling is no longer working as well as before. It has become very complex and expensive, and the benefits it returns have been diminishing. As we advance in the era of AI and IoT, much emphasis is being placed on architecture changes to drive performance and power (energy efficiency) improvements. Architecture changes can drive a dramatically higher level of performance by fundamentally changing how data is processed.

We suggested that delivering energy efficiency improvements would involve a combination of approaches: new chip architectures, 3D design techniques, novel materials, new ways to continue shrinking transistors, and advanced packaging schemes to connect chips together in new ways. We call this the "new playbook" for semiconductor design and manufacturing. Enabling this new playbook requires major advances in materials engineering and greater collaboration across the industry ecosystem.

End of Moore's Law Classic 2D Scaling

In the sixth edition of their popular textbook published in 2017, John Hennessy and David Patterson (widely regarded as the fathers of modern computer architecture) have meticulously summarized 40 years of processor performance data to make the case that three of the popular "laws" related to CPU performance progression have all effectively petered out. According to the reference, Dennard scaling ended in 2003, Amdahl's Law reached its limits in around 2011, and Moore's Law progress ended in around 2015. Our interpretation of their analysis is that the free ride of performance improvement from classic 2D scaling has effectively ended, and we now must work hard to combine architecture innovation and materials engineering breakthroughs to enable the performance improvements required for AI.

The chart below shows computing performance improvements over time taking advantage of Moore's Law. Moore's Law "classic 2D scaling" has slowed dramatically, and performance improvements are increasingly more difficult to achieve.

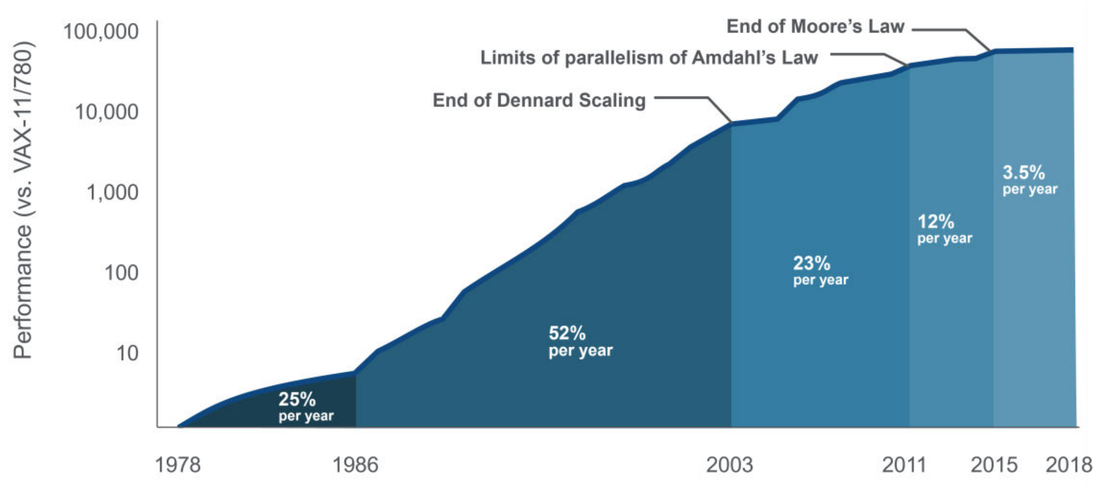


Chart 24: Chip performance improvements over time. Source: Computer Architecture: A Quantitative Approach, Sixth Edition, John Hennessy and David Patterson, December 2017.

Rise of New Computing Architectures

The figure below shows PCs and smartphones were only getting a 2x improvement in performance through Moore's Law, but architecture innovation through materials engineering can potentially drive performance improvement up to 2,500x despite limited benefits from the slowing of Moore's law.

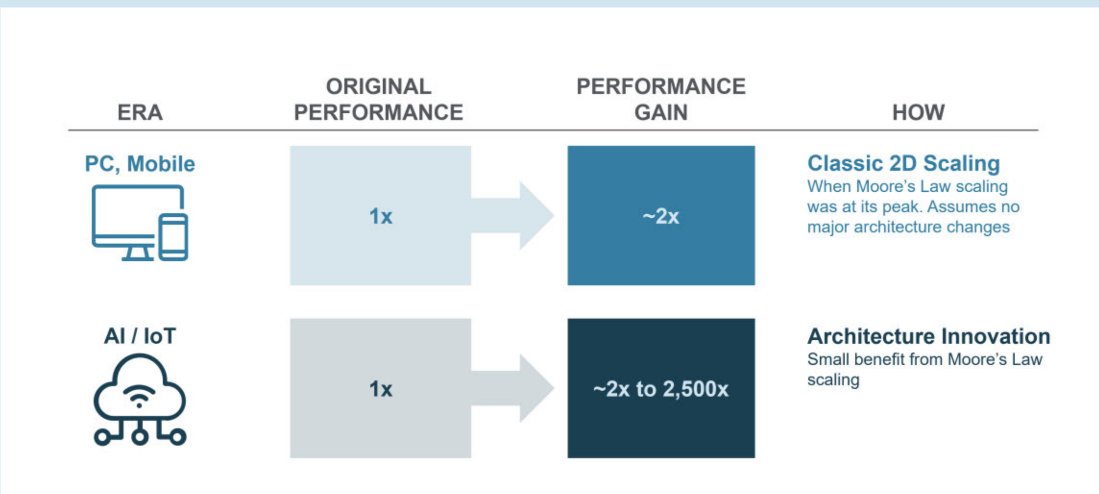


Chart 25: From Moore's Law classic 2D scaling to Architecture Innovation. Source: Applied Materials SMI.

GPUs are the primary example of a switch to using different types of architectures for AI workloads. As AI-driven workloads have increased, we can see the increased focus on GPU performance, which is demonstrated by the 500x increase in Floating-point operations per second (FLOPs) over the last five generations of NVIDIA GPU releases.

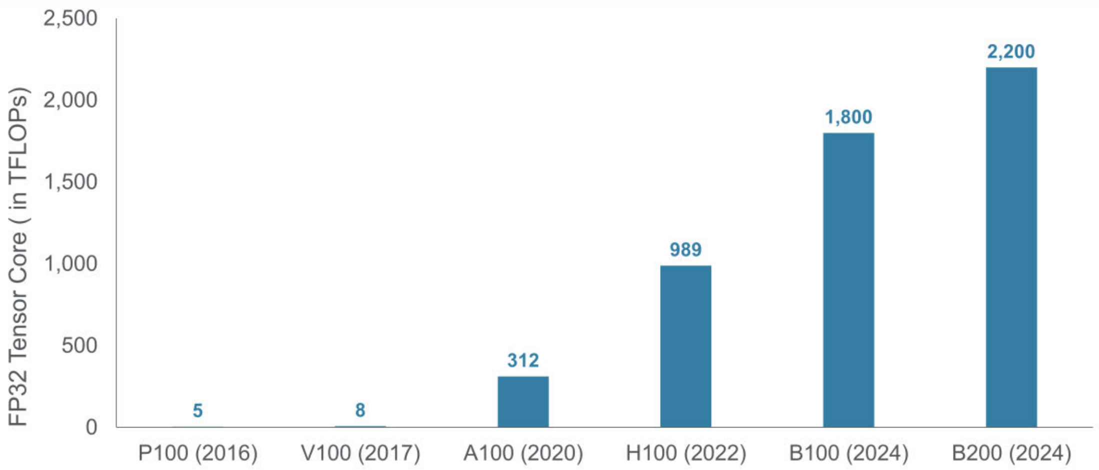


Chart 26: Performance comparison of FP32 Tensor Core (in TFLOPs) across different generations of NVIDIA GPUs. Source: NVIDIA, Applied Materials SMI.

The switch to GPUs from CPUs is demonstrated by the chart below, which shows average time to run an Apache Spark workload on CPUs versus GPUs. For this specific workload, there is a distinct benefit of 7x from an efficiency perspective when run on GPUs versus CPUs.

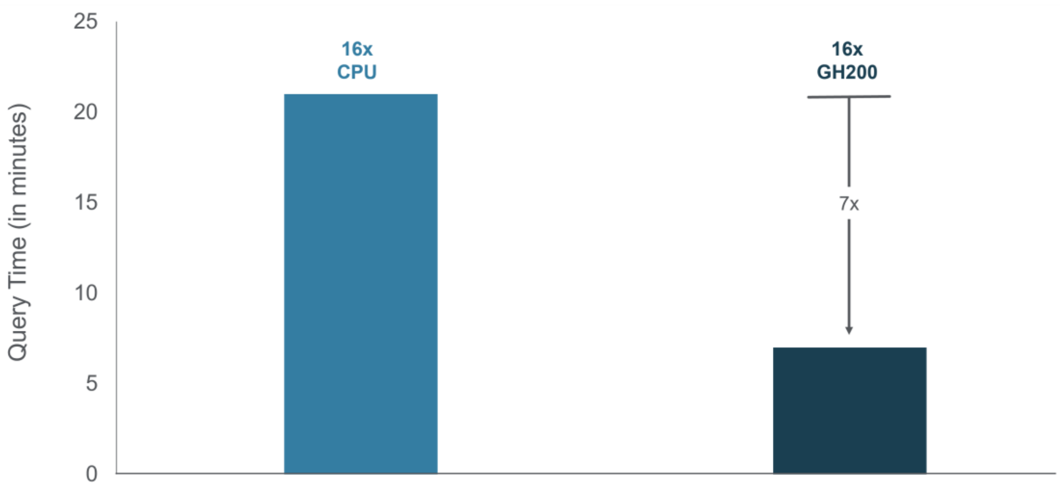


Chart 27: Query time comparison between 16x CPU and 16x GH200. Source: NVIDIA, Applied Materials SMI.

Additionally, we see that efficiency improves as the dataset gets bigger. Given the exponential growth of data and the larger model sizes in AI, scaling efficiently on new architectures will be key.

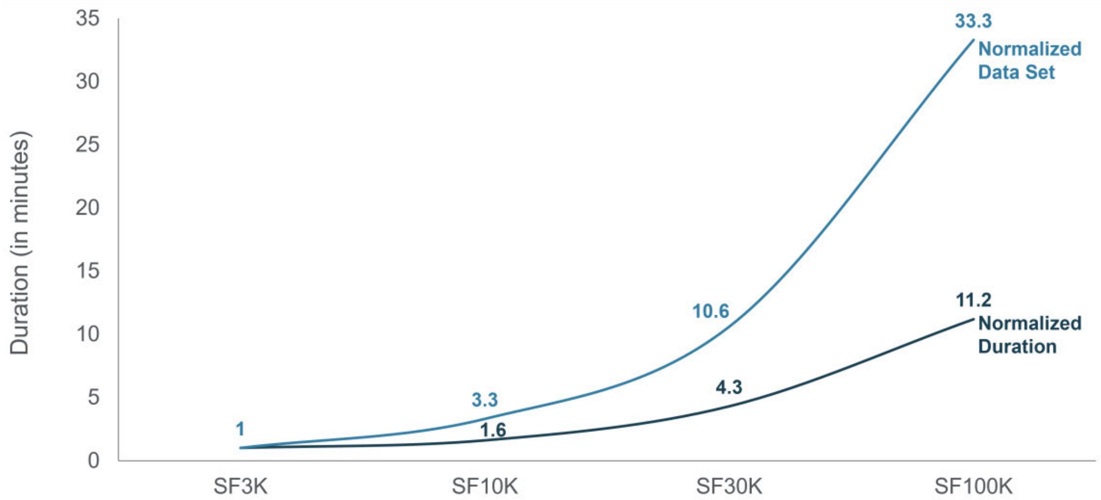


Chart 28: Scaling benefit of NVIDIA GH200. Source: NVIDIA, Applied Materials SMI.

In addition to GPUs, the table below shows new Application-Specific Integrated Circuits (ASIC) announced by Cloud Service Provider (CSP) and automotive companies over the past 10 years. Given the slowdown in Moore's Law and the increasing need for compute driven by AI, more and more companies are designing their own custom chips to deliver greater performance and energy efficiency.

Company	Chip Name	Announcement Time
Google	Tensor Processing Unit (TPU)	2015
Amazon	Graviton	Nov'18
Amazon	Inferentia	Nov'18
Amazon	Trainium	Dec'20
Tesla	Dojo	Aug'21
META	MTIA	May'23
MSFT	Maia	Nov'23
MSFT	Cobalt	Nov’23
Google	Axion	Apr'24
IBM	Spyre	Aug'24
OpenAI	[To Be Determined]	Oct'24 (News)

Chart 29: Custom ASICs announced by CSPs, Automotive and other companies. Source: Applied Materials SMI.

While new architectures have primarily focused on performance improvements, we note that the increasing power constraints are going to drive a hyper focus on power efficiency in conjunction with performance.

Below we see the technology stack for AI datacenters – starting from the application and platform layers (where the CSPs play) to the system/chip design including power delivery, where Applied's primary focus area lies. Energy-efficient computing can only be delivered through collaborative innovation across the entire tech stack.

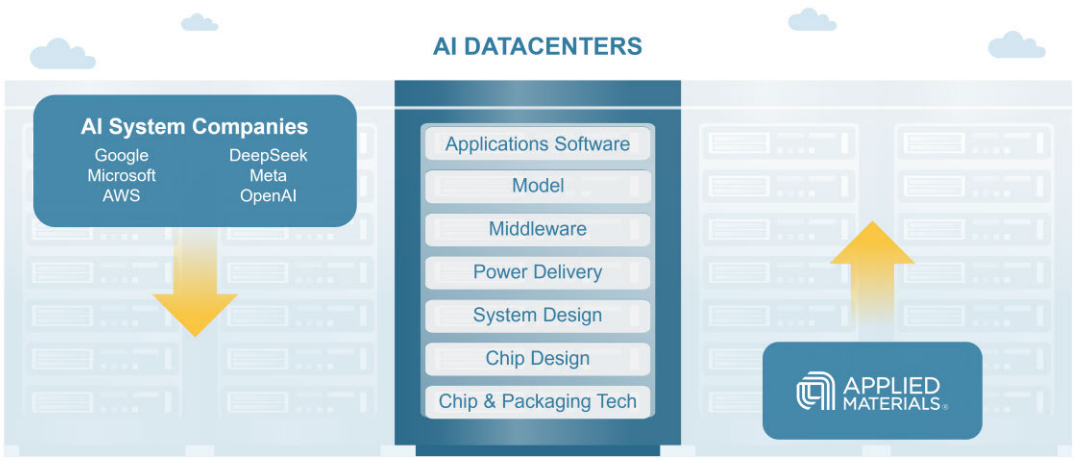


Chart 30: AI Datacenter Technology Stack. Source: Applied Materials SMI.

While a 10,000x performance improvement was delivered over the last 15 years, the exponential growth in compute requirements driven by AI means another 10,000x improvement in energy-efficient performance is required by 2040.

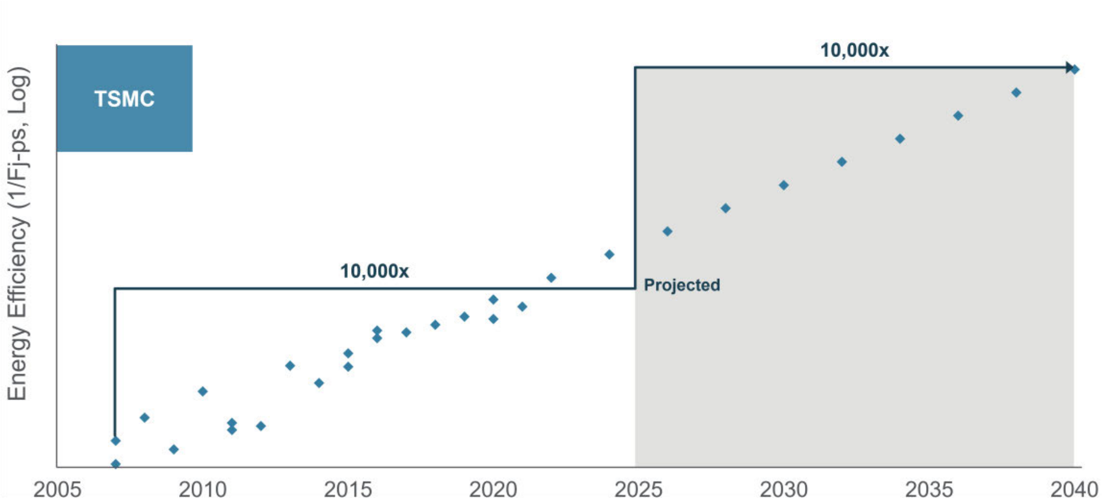
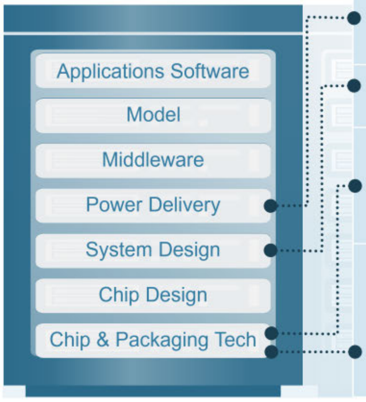


Chart 31: 10,000x improvements needed in energy efficient performance. Source: TSMC, Applied Materials SMI.

Looking across the end-to-end computing capability stack, Applied is focused on innovations in chip technology, packaging, circuit board and rack interconnect as well as power delivery. A combination of technology inflections will together deliver advances in energy-efficient computing performance.



		Inflection	Applied Focus
Power	Compound Semiconductors	IPD and IVR GaN SiC	✓
Board & Rack	Optical Interconnect	Si Photonics µLED	✓
Package	3D Integration	HBM W2D Bonding D2D Bonding Advanced Substrates	✓
Chip	Compute Memory	$4F^2$ DRAM 3D DRAM	✓
Chip	Leading Edge Logic	GAA BSPD CFET	✓

Chart 32: Applied's areas of focus. Source: Applied Materials SMI.

Materials Engineering Innovations Enabling AI

Materials engineering device inflections are playing a major role in powering AI computing, with Applied being a critical enabler. Gate-All-Around (GAA) transistors are the next major advancement in leading-edge logic technology after FinFET. Semiconductor fabs are currently in the early stages of ramping production for GAA to power next-generation AI computing chips. With GAA and Backside Power Delivery, transistor energy efficiency is determined by precise combinations of nanoscale exotic materials that can only be done under high vacuum with Applied's most advanced Integrated Materials Systems (IMS).

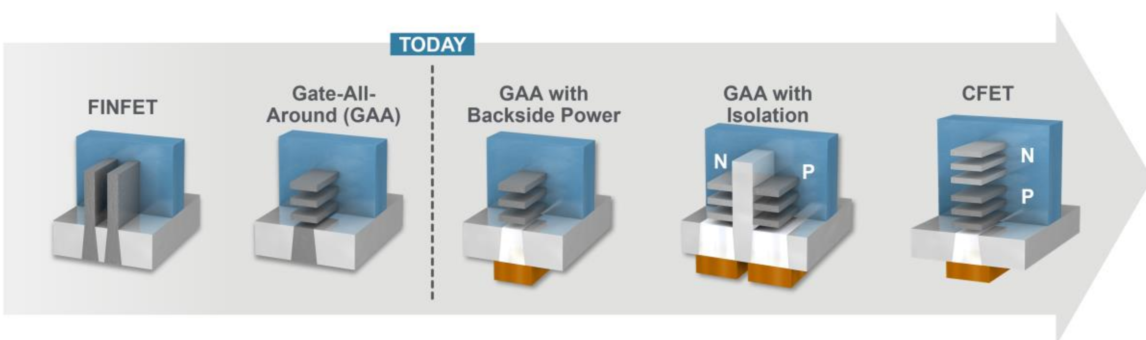


Chart 33: Major device inflections underway. Source: Applied Materials SMI.

The chart below shows that GAA is estimated to provide 25-30% improvement in energy efficiency. GAA is enabled by materials engineering, with five key materials in the channel, and layers that are 1-2nm thin.

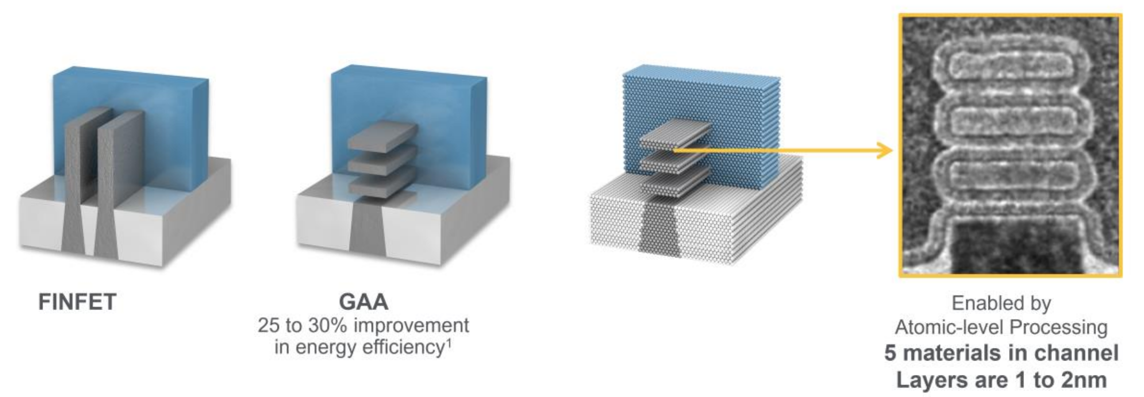


Chart 34: Critical device inflections enabled by Materials Engineering. Source: TSMC, Applied Materials.

New inflections in compute memory (i.e. DRAM) such as 4F2 and 3D DRAM will play a major role in improving the energy efficiency of AI computing. Materials engineering will be the critical enabling semiconductor technology for these advancements, as chipmakers bring elements of advanced transistor technology (e.g. High-K Metal Gate) to DRAM. Below, we see the anticipated technology roadmap for compute memory leading up to 3D DRAM near the end of the decade.

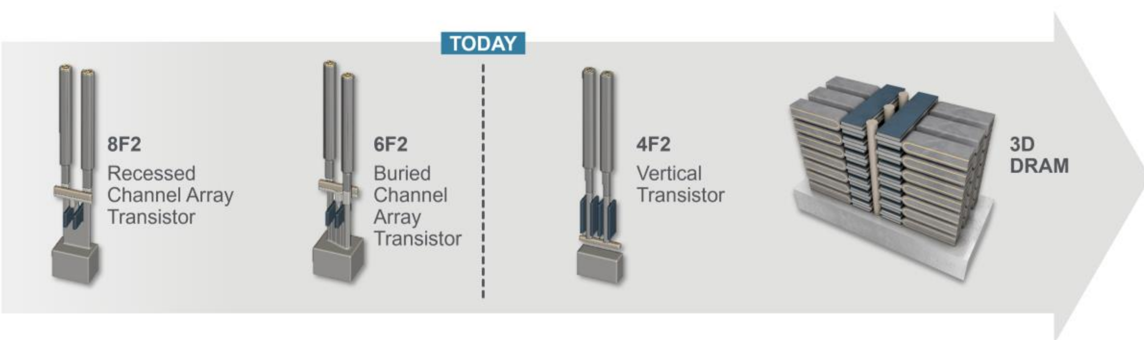


Chart 35: Compute Memory inflections enabled by Materials Engineering. Source: Applied Materials.

3D DRAM is a major inflection that will be enabled by materials engineering, with Applied playing a key role. The chart below illustrates that 3D DRAM involves more than 100 layers of materials, high-aspect-ratio device construction, with extreme precision needed at the nanoscale to deliver electrical properties for greater energy efficiency.

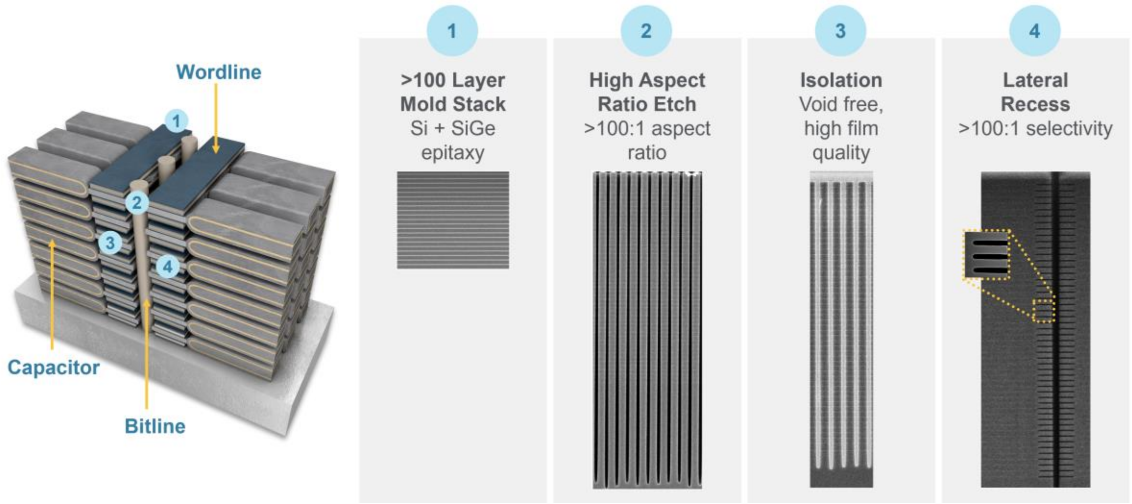


Chart 36: Materials Engineering to enable 3D DRAM. Source: Applied Materials.

These are just a few examples of how materials engineering is powering the new semiconductor playbook to enable continued improvements in energy efficiency. Applied has the industry's broadest portfolio of materials engineering solutions across the steps required to create, shape, modify, analyze, and connect materials at the atomic scale. We are the leader in materials engineering solutions that make up almost every chip in the world.

Looking to the future, chip nodes and features will be described in angstroms, or tenths of a nanometer. At these dimensions, materials behave in entirely different ways as surface properties begin to dominate bulk properties. Applied has a big role to play in delivering innovative materials engineering solutions that enable customers to continue building high-performance, energy-efficient chips for AI and the next-generation of electronics applications.

PILLAR C:

Monetizing New Opportunities for Semiconductors

APPLIED MATERIALS®

PILLAR C: MONETIZING NEW OPPORTUNITIES FOR SEMICONDUCTORS

To successfully embed sustainable abundant energy into corporate strategy, organizations must establish clear linkages between sustainability initiatives and measurable business outcomes. When sustainability is seen merely as a compliance obligation, management teams and boards of directors would default to fulfilling minimum reporting requirements, limiting broader investments. However, by demonstrating a direct connection between decarbonization strategies and tangible business results -- such as opportunities for revenue growth, competitive advantage, cost savings, increased industry connectivity and enhanced brand reputation -- companies can motivate deeper commitments, unlock innovation, and foster strategic integration of sustainability across their entire operations and value chains. Below we consider market opportunities for power semiconductors for the datacenter, renewable energy generation, and smart infrastructure.

Datacenter Power Semiconductors

GPUs and ASICs will drive significant growth in datacenter power semiconductors, along with increasing semiconductor content in other components, such as the motherboard and the power supply.

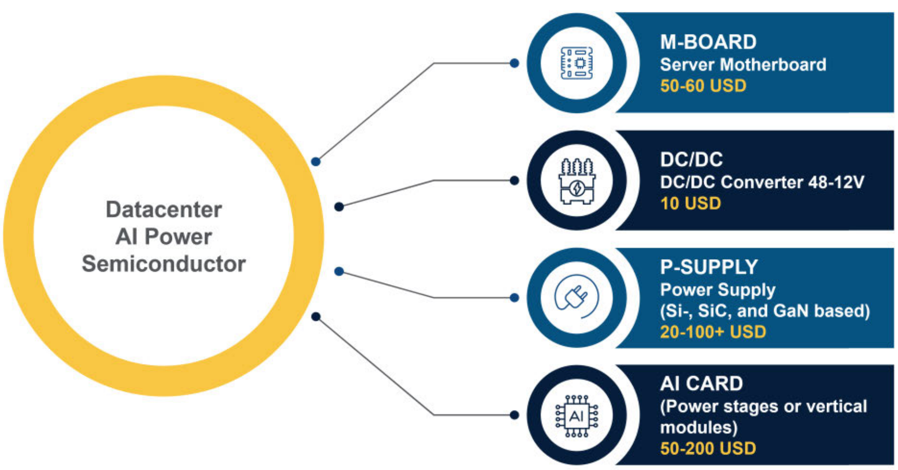


Chart 37: AI Datacenter Power Semi Components. Source: Infineon, Applied Materials SMI.

We forecast that the datacenter power semiconductor market could reach $9B by 2030, driven by an increase in servers as well as the number of accelerators per server. Approximately $7B of the opportunity is driven by the GPU/ASIC cards, with the rest being driven by the power supply and other areas. As the technology matures, we believe gallium nitride (GaN) will play an important role in this market, given its higher efficiency.

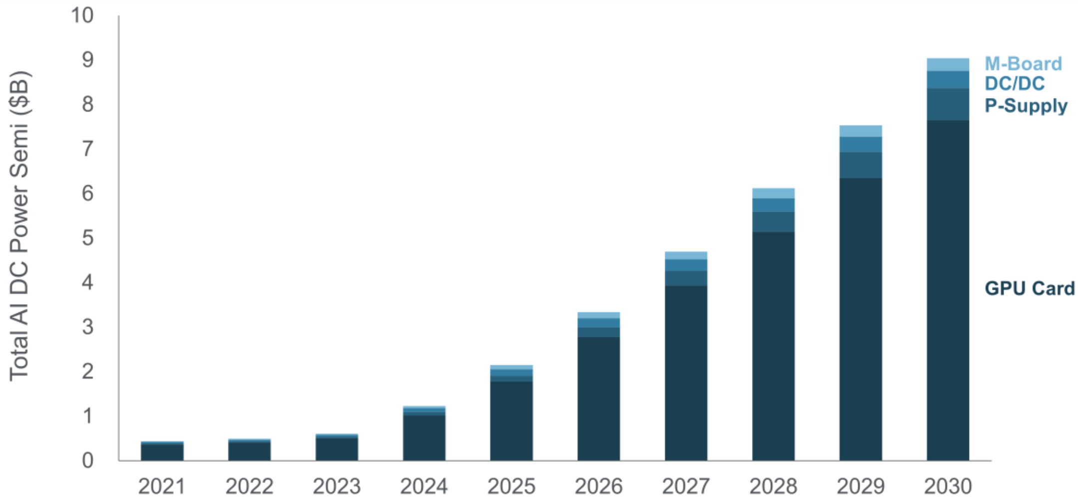


Chart 38: AI Datacenter Power Semis Opportunity. Source: Infineon, Applied Materials SMI.

Semiconductors for Renewable Energy

Renewable generation is another area of opportunity for semiconductor companies. As the grid incorporates increasing levels of renewables, more semiconductors will be needed for energy generation.

Silicon carbide (SiC) will play an important role in solar generation and potentially wind as well. Renewable energy generation is projected to have a TAM of approximately $23B by 2030 for semiconductor revenues.

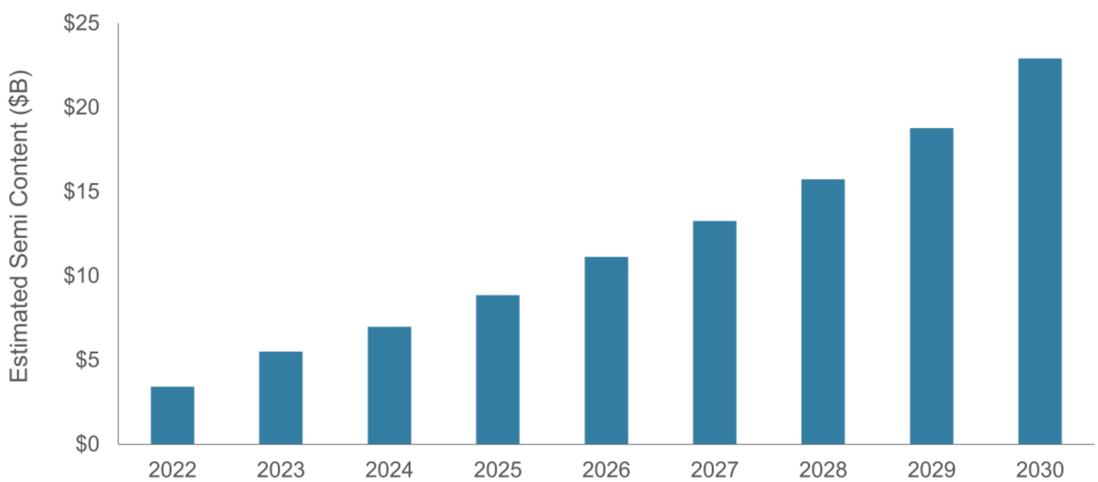


Chart 39: Semiconductor revenues driven by renewable sources. Source: Applied Materials SMI.

Semiconductors for Smart Infrastructure

The chart below shows the opportunity driven by smart infrastructure (smart meters, sensors, heat pumps, etc.). We estimate that smart infrastructure will drive approximately $19B of semiconductor revenues by 2030.

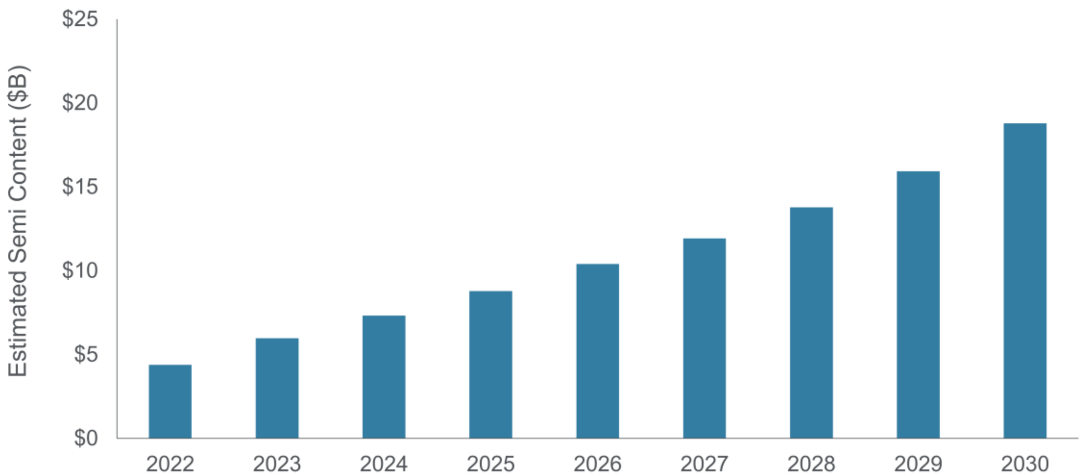


Chart 40: Semiconductor revenues driven by Smart Infrastructure. Source: Applied Materials SMI.

The chart below shows the estimated semiconductor opportunity from the buildout of electric vehicle (EV) charging infrastructure. Semiconductors, and specifically SiC, will play a crucial role in this ecosystem by enabling advanced power electronics, which are essential for fast and efficient charging stations. We assume a cumulative installed base of 25M chargers by 2030, which results in semiconductor revenue opportunity of $4B.

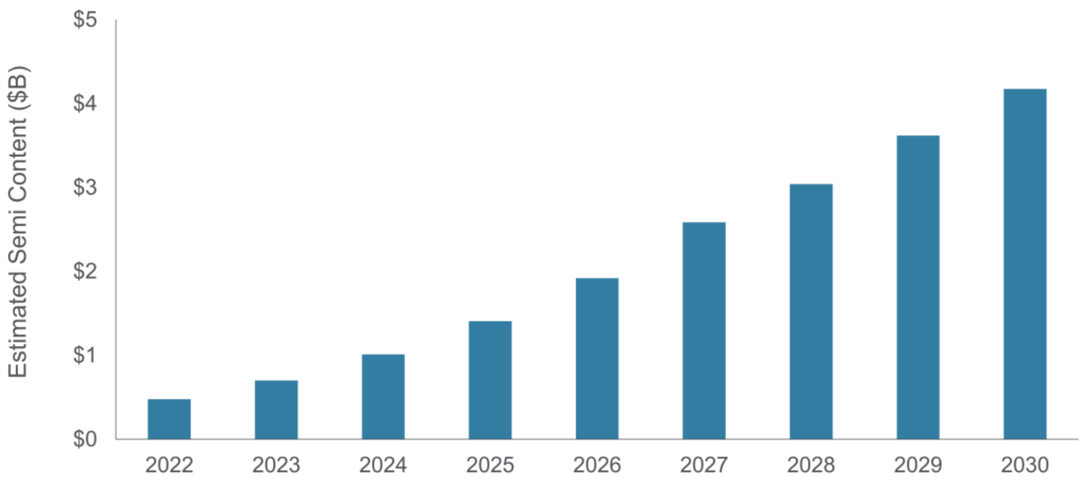


Chart 41: Semiconductor revenues driven by EV charging infrastructure. Source: Applied Materials SMI.

The chart below shows the estimated semiconductor opportunity from grid transformation split across generation, distribution, and transmission. We estimate a grid transformation TAM of $50B by 2030, including $23B from renewable generation, $24B from distribution (charging infrastructure, smart meters, smart sensors), and the rest from transmission (bidirectional grids, microgrids).

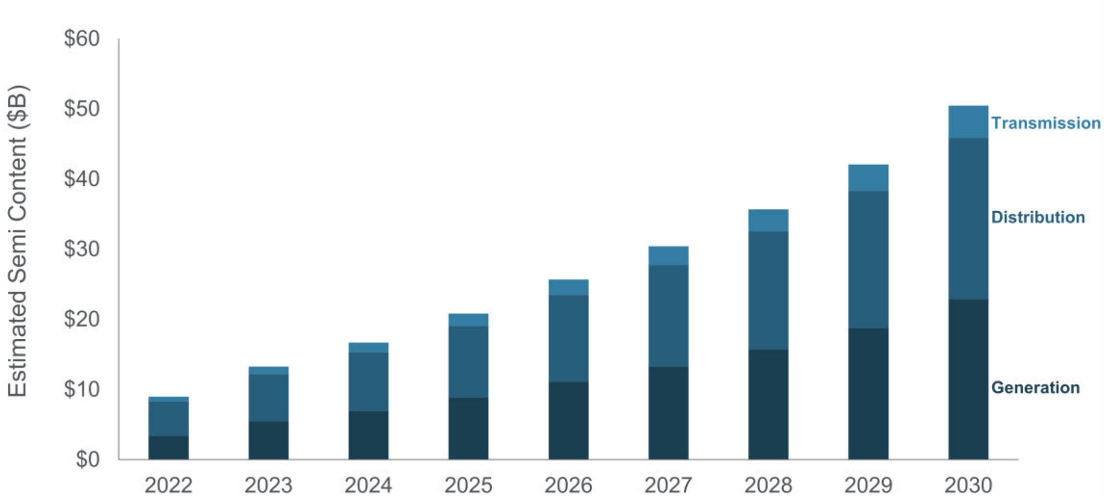


Chart 42: Semiconductor revenues driven by Grid Transformation. Source: Applied Materials SMI.

Companies are already identifying these opportunities and filling market gaps that have arisen. For example, The Mobility House (TMH) is working to improve the economics of Net Zero by connecting Automotive EVs to the grid through their Vehicle-to-Grid (V2G) solution. V2G, part of a broader Vehicle-Grid Integration (VGI) portfolio, combines bidirectional charging hardware and intelligent energy-management software to turn electric vehicles into flexible grid assets.

Through offerings like the Mobilize Power package of Renault in European countries, which integrates a bidirectional charger, V2G energy contract, and mobile app for vehicles and TMH's ChargePilot® platform, EVs can both draw energy and dispatch power back to the grid for services like peak shaving, solar self-consumption optimization, and frequency regulation. By interconnecting with utility systems and leveraging dynamic rate and proprietary trading programs, this V2G capability not only unlocks new revenue streams for EV owners but also enhances grid stability and resilience. The total market size is estimated at $54B in 2030 (Chart 42) and $540B in 2040, and will need significant increase in computational power needs to fully unfold.

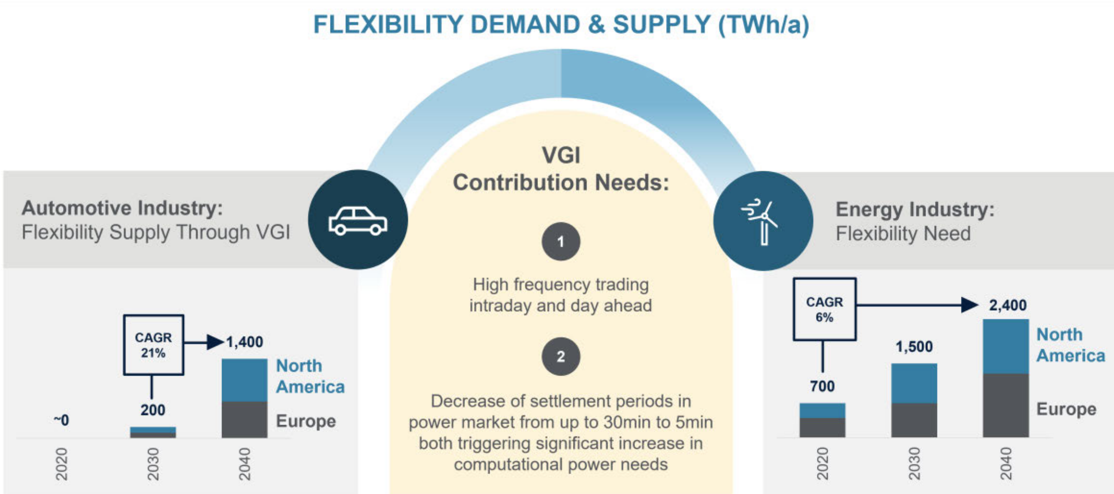


Chart 43: VGI offers flexibility and economics to the energy industry. Source: The Mobility House.

Vehicle Grid Integration (VGI) delivers measurable emissions reductions by intelligently coordinating electric vehicle charging and discharging with clean (or carbon-free) energy availability and grid needs and displaces carbon intensive peaker plants through bidirectional power flows. Through that, TMH's platform can reduce CO_2 emissions by 50% compared to unmanaged charging by shifting load to periods of high renewable output and feeding back power into the energy system when renewables output is low (Chart 43). Across numerous deployments, VGI projects have not only integrated greater shares of clean generation, but also enhanced grid stability and delivered lasting greenhouse gas reductions, underscoring VGI's pivotal role in the energy transition.

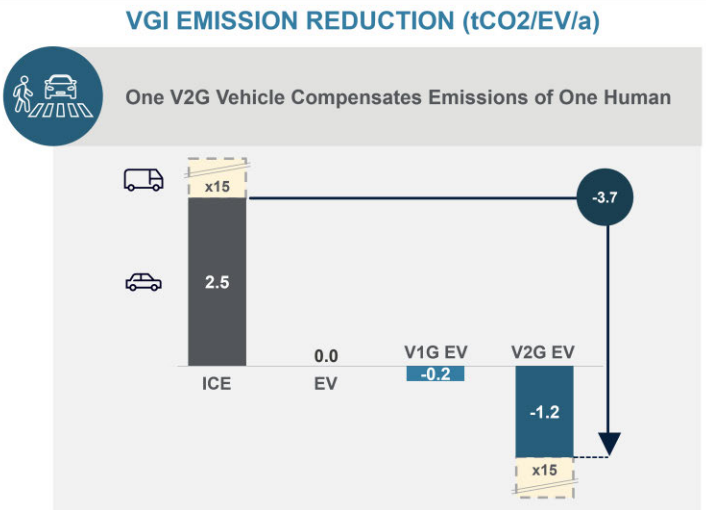


Chart 44: VGI reduces emissions beyond the drive train. Source: The Mobility House.

THE FUTURE OF COMPUTING

THE FUTURE OF COMPUTING

Looking beyond AI datacenters, Bitcoin mining and humanoid robots represent two examples that offer a window into the future of computing. Both demand unprecedented levels of energy that is available in a distributed topology. Bitcoin's proof-of-work consensus drives high-intensity power use, while training and operation of humanoid robots, especially for highly precise applications, could multiply compute and energy requirements many-fold. Yet these innovations also open new pathways for synergy: flexible Bitcoin loads can balance energy usage at the grid, while humanoid robots can build upon advancements made by the EV industry.

AI and Bitcoin: Diverging Energy Needs

AI-driven workloads, such as training large language models and running inference engines, require uninterrupted and high-priority energy access. These workloads typically peak during business hours or other seasonal high-consumption periods overlapping with conventional energy consumption patterns. This inflexibility contributes to higher grid stress, particularly during demand surges.

Unlike AI data centers, Bitcoin mining can dynamically adjust its energy consumption based on grid conditions. Mining operations, co-located with AI data centers, can be scaled up during periods of excess electricity (such as midday solar overproduction) and curtailed during peak demand, acting as a flexible load-balancing mechanism.

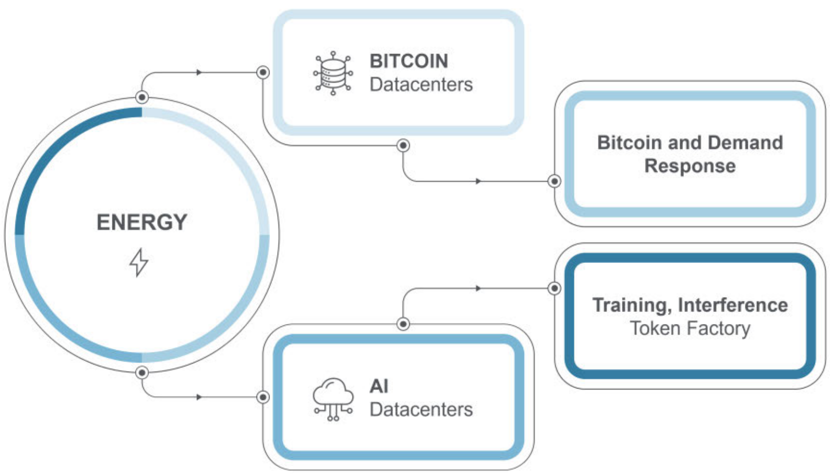


Chart 45: Diversified Use of Energy for Economic Value. Source: Auradine.

Grid Stability and Challenges

The grid faces instability due to daily energy demand fluctuations—especially when demand surges. This rapid fluctuation causes grid instability, leading to energy curtailment, reliance on peaker plants, and inefficiencies in renewable energy utilization.

Bitcoin mining can absorb excess solar or other energy during midday, preventing curtailment, and then scale down when AI and residential loads peak. Technologies such as EnergyTune™ from companies like Auradine enable Bitcoin miners to optimize their energy consumption by aligning it with real-time grid conditions. This ability to act as an "energy sponge" helps smooth out fluctuations in grid demand, effectively reducing strain on the grid and enhancing stability.

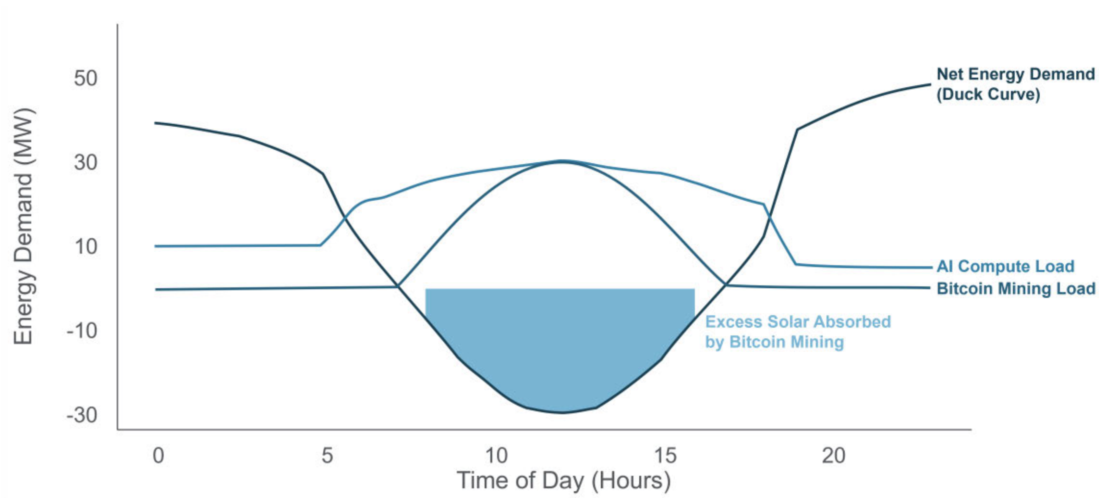


Chart 46: Duck Curve + AI and Bitcoin Energy Demand. Source: Auradine.

AI and Bitcoin data centers, though energy-intensive, offer a unique synergy in addressing grid stability challenges. AI's rigid energy demand contrasts with Bitcoin mining's flexible load capabilities, allowing for strategic demand-side management. By integrating demand-response solutions, we can create a scalable energy future that harmonizes AI growth with sustainable energy practices.

Humanoid Robots

The next revolution in computing is not confined to server farms – it is walking, talking, and interacting in the real world. Humanoid robots, championed by industry leaders from NVIDIA's GTC demos to Tesla's Optimus vision, promise to transform manufacturing floors, healthcare settings, classrooms, battlefields, and even our homes. These machines will combine edge-level intelligence with large-scale training in the datacenter, multiplying compute and energy requirements by orders of magnitude. Humanoid robots are often referred to as "physical AI" because they represent a major practical delivery mechanism for the productivity boost envisioned from AI. As some of us like to say: *Robots x AI = (virtually) Infinite Productivity*.

As shown in the next chart, we project an installed base of 30 million humanoid robot units by 2030, growing to over 360 million by 2040, driving an incremental $100 billion in semiconductor revenue by 2030 and $700 billion by 2040. Our numbers could prove to be conservative especially if we compare our estimates to IBS, which forecasts 1B robots in China by 2040.

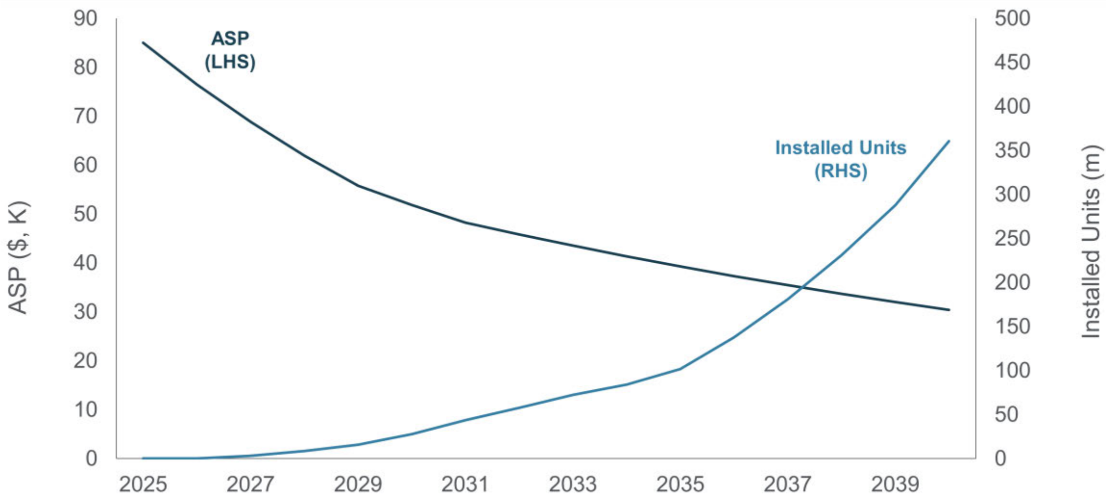


Chart 47: Humanoid Robot Market - Units and ASPs. Source: Applied Materials SMI.

Our assumptions about humanoid robot adoption across different industries is shown in the chart below. Adoption starts in labor-intensive industries such as manufacturing (e.g. already happening at BMW, Tesla, Amazon) and in developed economies where high labor cost justifies robot replacements. We are not expecting significant adoption of consumer robots until after 2035, as prices would need to move lower to justify adoption, in our view.

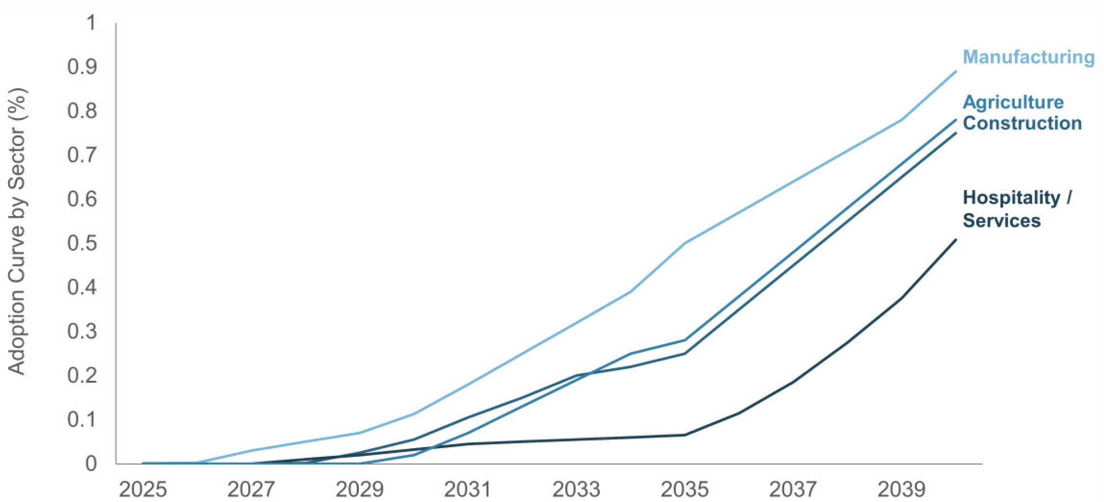


Chart 48: Macro assumptions for Humanoid Robots. Source: Applied Materials SMI.

We project that by 2040 electricity demand for humanoid robots (Chart 49) will surge from near zero in the mid-2020s to over 1,300TWh, driven by the transition from pilot deployments to mass-market adoption after 2030. This trajectory implies that charging for humanoid robots alone could account for roughly 3% of global electricity – on par with today's entire datacenter sector.

Elon Musk has said training a humanoid robot will require 10x the compute needed to train autonomous cars, given the wider range of functions that a robot will perform. We view humanoid robots as an extension of edge AI, and another advancement whose growth might be stunted without sustainable energy abundance.

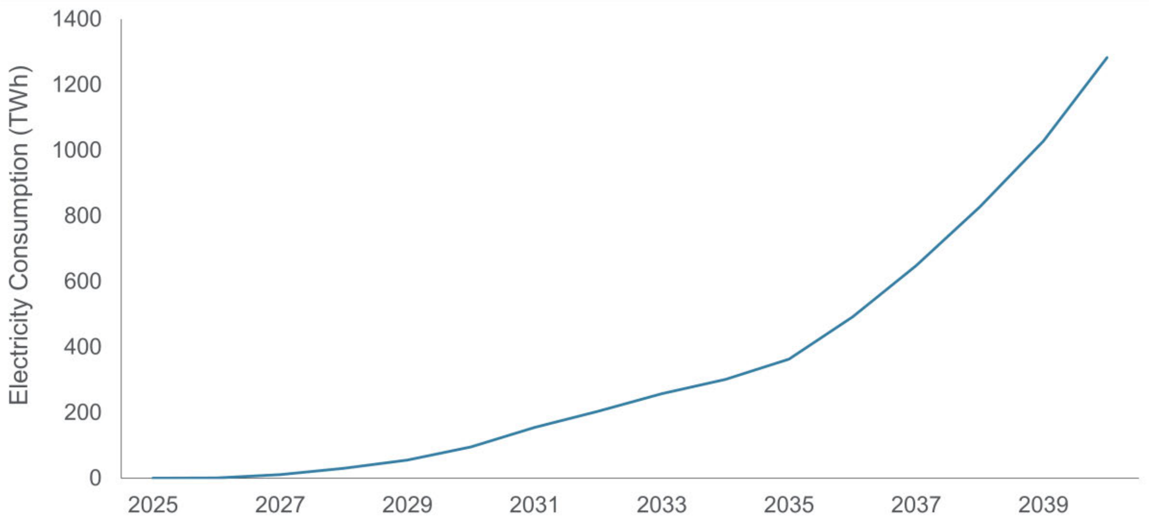


Chart 49: Humanoid Robot Electricity Usage. Source: Applied Materials SMI.

Together, these dynamics make clear that scaling humanoid robotics sustainably will require not only expanded clean (or carbon-free) energy capacity and grid modernization, but also flexible load strategies, such as demand-response and dynamic scheduling, to integrate this productivity-driving growth segment of computing without compromising reliability or decarbonization.

CALL TO ACTION

BUILDING THE FOUNDATION FOR SUSTAINABLE ENERGY ABUNDANCE

This paper has presented a vision where **AI and semiconductors unlock unprecedented productivity**, enabling societies to address the mounting macroeconomic pressures of **debt and demographics**. However, this vision can only be realized if we confront one of the most urgent constraints of our time: **energy availability**.

We advocate for a new paradigm — **sustainable energy abundance** — achieved not by sacrificing growth, but by aligning it with a **holistic decarbonization and sustainability strategy**. This strategy recognizes that:

AI and future computing technologies are both dependent on and uniquely capable of accelerating sustainability.

The energy required to power the next generation of computing must come from diverse, increasingly clean, and intelligently deployed sources.

A modernized, decarbonized grid is essential infrastructure for the digital age. But this is not a challenge any one company or country can solve alone.

We call on:

Industry leaders to move beyond isolated sustainability initiatives and collaborate across the value chain, from materials engineering to chip design to data centers, to co-innovate at high velocity.

Policy makers to incentivize long-term investment in all forms of energy, while steering the global energy mix toward cleaner, more resilient sources.

Utilities and infrastructure providers to partner with technology innovators in modernizing the grid and integrating intelligent computing to optimize energy use and distribution.

Investors and shareholders to align capital with strategic visions for **monetizing new opportunities for semiconductors**, competitive advantage, and long-term value creation.

Having a published plan to achieve sustainability is not enough. We must accelerate it **with AI, for AI**.

Let us seize this moment to rethink what’s possible, work together to scale sustainable solutions, and **build the energy foundation for a more productive, prosperous, and resilient future**.

REFERENCES

1. *Materials Engineering is Key to Unlocking Commercial Value from AI*, December 2017, Sundeep Bajikar.
2. *"Computer Architecture – A Quantitative Approach Sixth Edition"*, December 2017, John Hennessy, David Patterson.
3. *"Enabling the A.I. Era – From Materials to Systems"*, New Street Research Conference, May 2018, Sundeep Bajikar.
4. *Why A.I. Workloads Require New Computing Architectures*, June 2018, Sundeep Bajikar
5. *Why Materials Engineering Breakthroughs are Required for New AI Computing Architectures*, June 2018, Sundeep Bajikar.
6. *"Navigating the Perfect Storm - Enabling the A.I. Era"*, A.I. Design Forum, SemiCon West, July 2018, Gary Dickerson.
7. *"Enabling the A.I. Era – From Materials to Systems"*, A.I. Summit at SemiCon Korea, January 2019, Sundeep Bajikar.
8. *"International Energy Outlook 2023"*, October 2023
9. *"The A.I. Paradox – Biggest Opportunities and Challenges of our Lifetimes"*, A.I. Design Forum, SemiCon West, July 2019, Gary Dickerson.
10. "*The 2024 AI Index Report*"- Stanford University
11. "*TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings",* June 2023
12. *"Make Possible a Better Future"*, Sustainability Summit, SemiCon West, July 2022, Gary Dickerson.
13. *"Make Possible a Better Future - Panel Discussion"*, Sustainability Summit, SemiCon West, July 2022, Sundeep Bajikar, Nasreen Chopra, Betty Jiang.
14. *Artificial Intelligence and Electricity- A System Dynamics Approach,* December 2024, Rémi Paccou and Fons Wijnhoven.
15. "*Global AI GPU Manufacture Carbon Emissions Forecast, 2025-2030*", TechInsights, February 2025, Stephen Russell.
16. *"Wafer Fab Equipment Positioned for a New Wave of Growth"*, June 2024, Sundeep Bajikar.
17. "*Race for AI Leadership Fueled by Materials Engineering*", SemiCon West, July 2024, Prabu Raja.
18. "*Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends*", February 2024, Ian Schneider, Hui Xu, Stephan Benecke, David Patterson, Keguo Huang, Parthasarathy Ranganathan, and Cooper Elsworth∗ (Google).
19. *Data center map*
20. *"Materials Engineering: The True Hero of Energy-Efficient Chip Performance"*, August 2024, Balaji Chandrasekaran, Po-Wen Chan.
21. *NVIDIA GH200 Superchip Delivers Breakthrough Energy Efficiency and Node Consolidation for Apache Spark,* August 2024, Amr Elmeleegy, Ivan Goldwasser and Karthikeyan Rajendran
22. *Energy and Policy Considerations for Deep Learning in NLP*, Proceedings of ACL 2019, Emma Strubell, Patrick Almeida, Andrew McCallum
23. *Recalibrating Global Data Center Energy-Use Estimates*, Science 2020, Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, Jonathan Koomey

24. *Net Zero by 2050: A Roadmap for the Global Energy Sector*, IEA May 2021
25. *Green AI*, Communications of the ACM April 2020, Roy Schwartz, Jesse Dodge, Noah A. Smith, Oren Etzioni
26. *The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010-2015*, Sustainability 10(9):3027 2018, Jens Malmodin, Dag Lunden
27. *100 Days: Bitcoin America Policy Recommendations*, Auradine January 2025
28. *TED 2025 Humanity Reimagined conference sessions, brain dates and key discussions*, April 2025.
29. *Power Roadshow*, November 2023, Infineon
30. *Clean Energy will be critical to winning the AI race with China*, April 2025, Hank Paulson
31. The Mobility House (http://www.themobilityhouse.com).
32. *Situational Awareness – The Decade Ahead*, June 2024, Leopold Aschenbrenner
33. *A Look Ahead to the Sustainability of 2nm Processes*, TechInsights January 2025, Stephen Russell.
34. *Hybrid Bonding Increases Complexity and Carbon Intensity*, TechInsights March 2025, Stephen Russell.
35. *Artificial Intelligence and electricity: A system dynamics approach*, Schneider Electric December 2024, Remi Paccou, Fons Wijnhoven.
36. *Energy and AI*, International Energy Agency, April 2025.
37. *Integrated solutions for daylighting and electric lighting*, Energy and Buildings Volume 277, December 2022, Jan de Boer, Eleanor S. Lee, Niko Gentile, Werner Osterhaus.