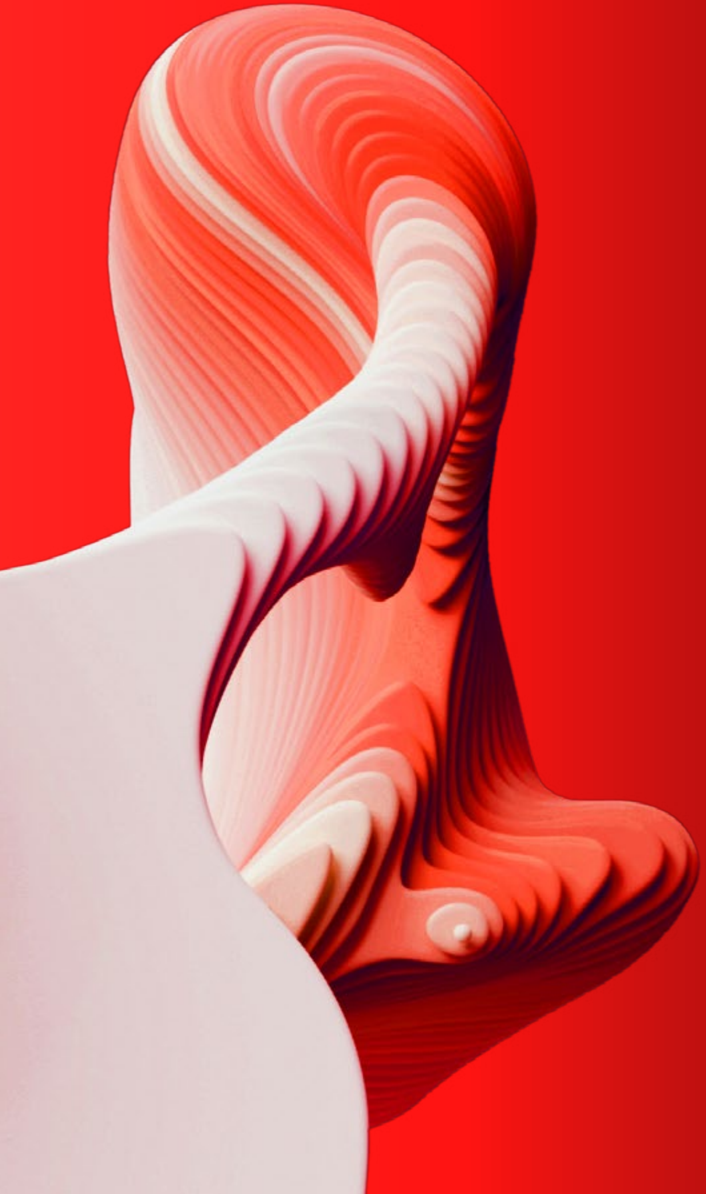
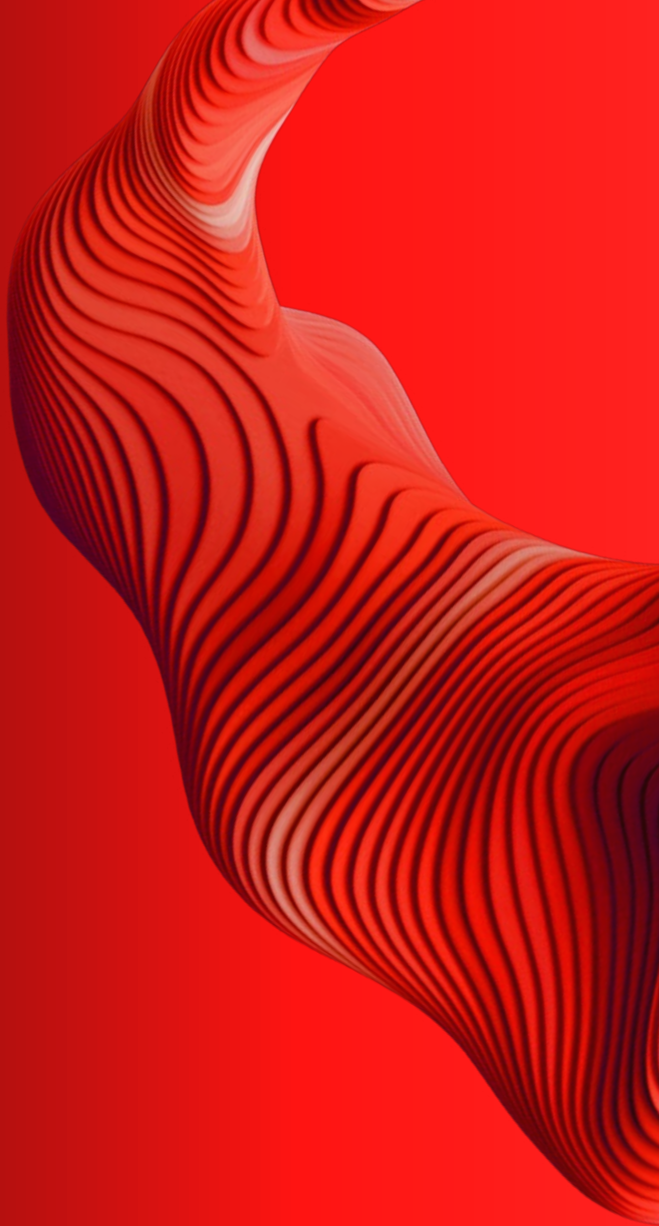


2025 TECH TRENDS REPORT • 18TH EDITION

COMPUTING



FTSG

Future Today Strategy Group's 2025 Tech Trend Report

Our 2025 edition includes 1000 pages, with hundreds of trends published individually in 15 volumes and as one comprehensive report. Download all sections of Future Today Strategy Group's 2025 Tech Trends report at **www.ftsg.com/trends.**

18TH EDITION

2025 TECH TRENDS REPORT

FTSG

ARTIFICIAL INTELLIGENCE

FTSG

WEB3

FTSG

METAVERSE & NEW REALITIES

FTSG

BIOENGINEERING

FTSG

ENERGY & CLIMATE

FTSG

MOBILITY, ROBOTICS, & DRONES

FTSG

COMPUTING

FTSG

BUILT ENVIRONMENT

FTSG

2025 TECH TRENDS REPORT • 18TH EDITION

EXECUTIVE SUMMARY

FTSG

NEWS & INFORMATION

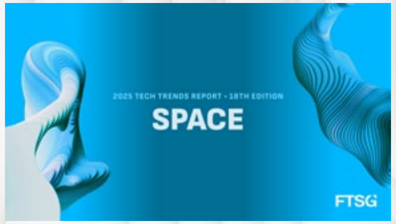
FTSG

HEALTH CARE & MEDICINE

FTSG

FINANCIAL SERVICES & INSURANCE

FTSG



SPACE

FTSG

HOSPITALITY & RESTAURANTS

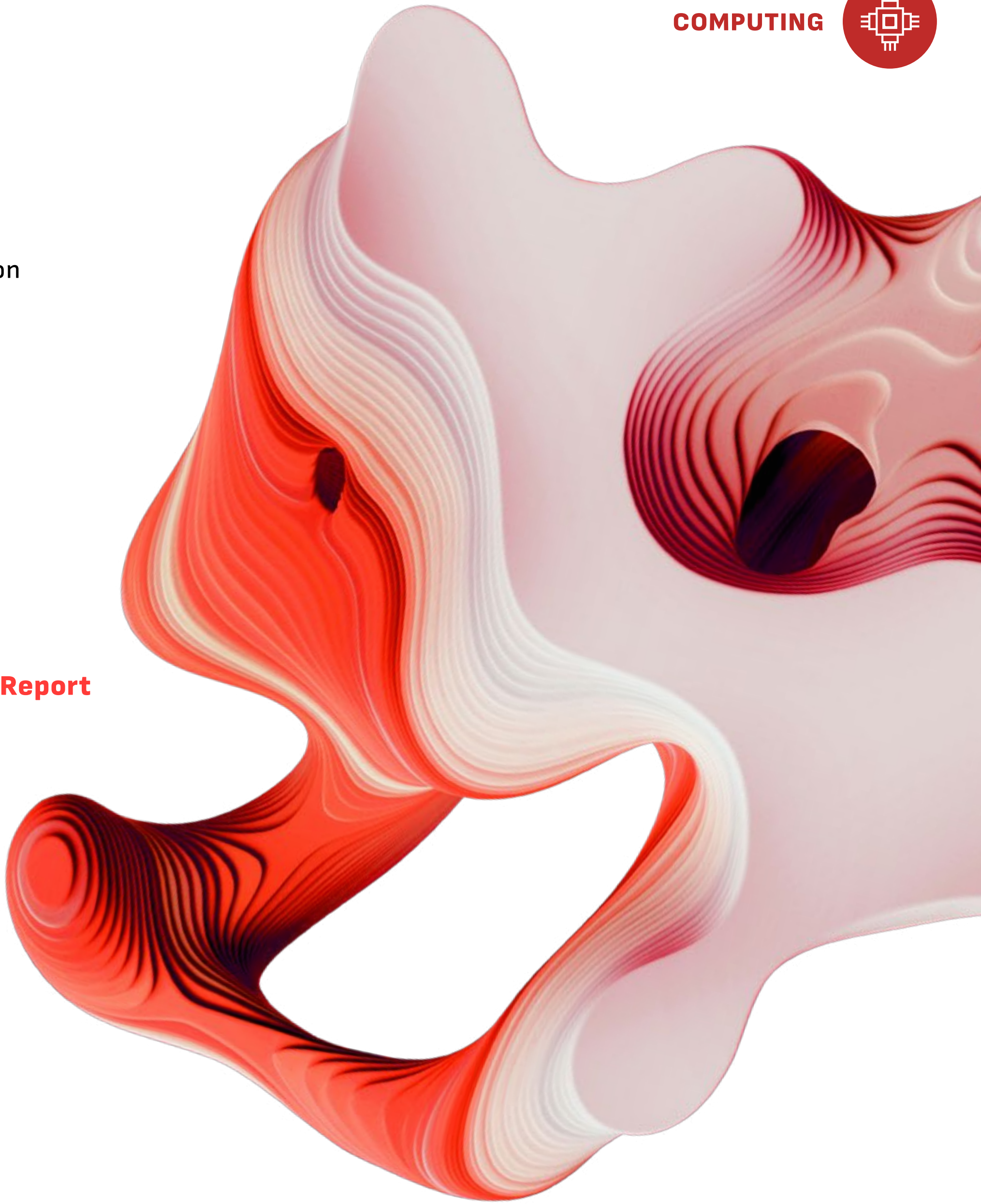
FTSG

SUPPLY CHAIN, LOGISTICS, & MANUFACTURING

FTSG

ENTERTAINMENT

FTSG



Sam Jordan
Technology & Computing Lead

The Next Currency of Power

The next decade will hinge on algorithms and atoms. Behind the headlines about AI model capabilities lies a fundamental challenge: our ability to build computing infrastructure at an unprecedented scale. We need semiconductors, cooling systems, fiber optic networks, power plants, grid infrastructure, and skilled workers. Building computing infrastructure to meet AI demand won't just come down to technology—the core components needed are enmeshed in bureaucracy and politics. Consider the energy required to power data centers. Many view electricity as a zero-sum game, where using more power for AI means less for essential services like hospitals, schools, and home cooling. While the law of conservation of energy holds true–energy cannot be created or destroyed–its form and distribution are not fixed. The real question is not how to split our current power supply, but how to expand it faster. It's about creating more efficient computing architectures and building flexible, resilient power systems. The country that masters these challenges will lead the AI race.

The stakes are higher than economic competition alone. National security depends on computational supremacy. When adversaries can deploy more powerful AI systems faster, they gain advantages in everything from intelligence to weaponry. Building robust computing infrastructure is just as much about maintaining strategic autonomy as it is technological leadership. And this is where quantum computing enters the chat. When Google's 2024 Willow quantum chip completed a problem in five minutes that would take today's supercomputers 10 septillion years to solve, it sparked equal parts excitement and bewilderment. The natural reaction was "Wait, what? Does this prove the multiverse exists? Does Willow's ability to perform calculations across multiple quantum states challenge our understanding of causality?" I'll let physicists tackle those questions. But to be clear, while Willow represents a genuine leap forward in error correction, it hasn't achieved true quantum advantage: the ability to solve practical problems faster and cheaper than classical computers. The first nation to achieve true quantum advantage won't just lead in AI—they'll crack unbreakable codes and solve problems once thought impossible. Every quantum breakthrough, each advance in AI capability, shapes the balance of power. And thus, the race for computing supremacy isn't just about technology or even economics—it's about the future of power itself.

The AI boom drives a global rush to construct massive computing infrastructure.

1

AI's growth fuels US chip manufacturing surge

In response to AI's skyrocketing demand, more than 90 new US chip fab projects were announced by August 2024. Tech giants are investing nearly $450 billion across 28 states to meet AI's power and data demands.

2

US-China chip war escalates with tightened export controls

The two-year feud heats up as the US partners with allies to tighten export controls on advanced semiconductor tech, citing security concerns. This sparks Chinese countermeasures and investment in domestic chip production.

3

Tech giants invest in small modular reactors for AI power needs

OpenAI, Oracle, and SoftBank announce Stargate, a $500 billion project to build AI infrastructure, including data centers and power facilities, across the US. Microsoft, Amazon, and Google are partnering with small modular reactor developers, investing billions in nuclear energy to power their data centers.

4

Google unveils "Willow" quantum chip

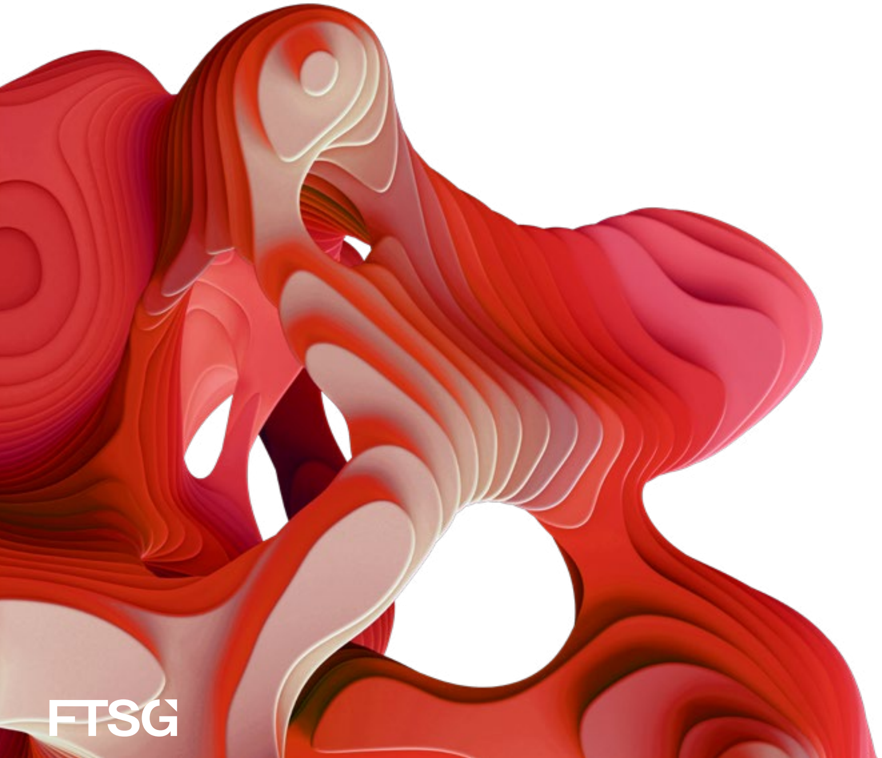
Willow marks a major breakthrough in quantum error correction. By exponentially reducing errors as more qubits are added, Willow paves the way for scalable quantum systems.

5

Embedded intelligence: personal devices get smarter

Intel, Microsoft, Nvidia and others are embedding AI directly into personal devices, boosting power and privacy. On-device AI enables faster, personalized responses without cloud reliance, making PCs and smartphones smarter and more self-sufficient for users.

As AI's compute demands soar, nations vie for supremacy through energy expansion and innovative architectures, reshaping global power dynamics.

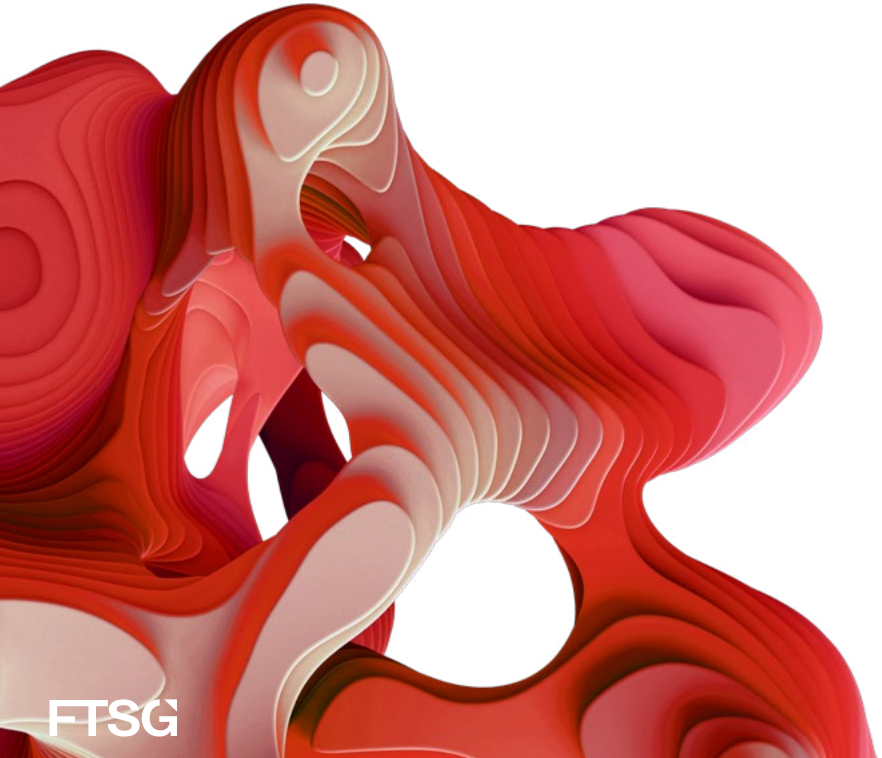


Over the past decade, ever-increasing computing power drove AI advances, with top models doubling their training compute every six months. Yet this relentless scaling is now hitting practical and physical limits: global infrastructure and energy capacity are struggling to keep pace. The Institute for Progress estimates that by 2030, global AI power demand could rise by as much as 130 GW, while U.S. electricity generation is projected to grow by just 30 GW in the same period. Meanwhile, China continues to outpace the U.S. in energy expansion, adding an average of 50 GW per year since 2010.

Bolstered by strong manufacturing and construction, China is also establishing new chip fabrication plants, while Saudi and Emirati sovereign wealth funds funnel resources into massive data centers. Should the most advanced models take shape overseas, the U.S. risks losing oversight of how these technologies are developed and deployed—particularly for dual-use applications that could fall into the wrong hands.

Against this backdrop, DeepSeek's r1 model has sparked fresh worries about American AI supremacy, leading to concerns that the U.S. might be losing its edge. DeepSeek's release highlights two parallel truths about AI progress: bigger models trained with more computing power do deliver performance gains, but major breakthroughs can also stem from more efficient, smarter algorithms. This challenges the assumption that ever-larger hardware investments are the only avenue to AI progress. Still, it would be a mistake to conclude we can abandon compute infrastructure altogether; the future will demand both efficiency improvements and robust hardware.

In response, the U.S. and its allies must consider not only expanding energy production but also rethinking compute architectures. One promising frontier is neuromorphic computing, inspired by the human brain's efficiency in processing information. Another emerging concept, organoid intelligence, relies on clusters of human brain cells to perform computational tasks. These novel approaches mark a shift away from traditional silicon-based architectures toward technologies specifically engineered to handle AI workloads more sustainably.



The potential for quantum computing to enhance AI efficiency is also substantial. Quantum computers leverage quantum mechanical phenomena like superposition and entanglement to perform calculations exponentially faster than classical computers for certain types of problems. This immense processing power could allow AI algorithms to tackle complex tasks that are currently infeasible or too time-consuming with classical computers.

Behind the scenes, quantum computing is making slow and steady progress in reducing noise and improving error correction, but quantum advantage has not yet been demonstrated. Even if it were, there's the question of whether the business models are viable enough for companies to invest in quantum computing, given its extremely capital-intensive nature. Nevertheless, due to geopolitical considerations, we must persist. The first country to achieve quantum advantage would gain significant strategic benefits, such as the ability to break current encryption methods, solve complex optimization problems, accelerate drug discovery, and revolutionize materials science—which could all lead to substantial economic and security advantages.

Computing today is not merely a technical pursuit; it drives global influence and sets up economic and cultural landscapes. The ability to scale computational infrastructure and advance energy-efficient architectures will impact everything from entertainment to human-computer interaction, transforming how we work, create, and live. This convergence of technological capability with geopolitics underscores a defining aspect of our time—where computing power both drives innovation and shapes the contours of our social, economic, and cultural narratives.

AI's surging demands pushed computing to its limits, driving advancements in infrastructure, energy strategies, and new architectural frontiers.

JANUARY 2024

China Announces Chip Fab Expansion

China aggressively expands chip fabs, announcing 18 new semiconductor manufacturing facilities.

MAY 2024

Microsoft and OpenAI Plan 100 MW Data Center

The $3.3 billion Wisconsin facility will feature 100,000 AI accelerators and support advanced AI operations.

OCTOBER 2024

AI Workloads Drive Nuclear Investments

Tech companies announce investments in nuclear plants to meet rising energy demands of large-scale AI.

DECEMBER 2024

Google Announces Willow

The chip completed a benchmark in under five minutes, a task that would take supercomputers 10 septillion years—far exceeding the universe's age.

JANUARY 2025

Deepseek Disrupts Nvidia

Deepseek's open-source LLM outperformed expectations at a fraction of the cost, tanking Nvidia stocks just after it unveiled the Blackwell GeForce RTX 50 Series, the world's most powerful GPU.

« PAST

The future of computing hinges on overcoming energy constraints and navigating geopolitical tensions, as AI's growth outpaces current infrastructure and resources.

FUTURE »

MID-2020s

Protocols for AI Data Centers

Hardware screening, cluster encryption, and real-time monitoring protect IP and secure AI infrastructure.

MID TO LATE 2020s

US Enhances Export Controls

The US and its partners expand chip controls to curb China's AI growth, impacting global markets.

LATE 2020S TO EARLY 2030s

"Invisible Interfaces" in Personal Devices

Devices move toward letting users interact seamlessly through eye tracking, voice, and even on-skin gestures.

EARLY 2030s

Power Bottleneck for Data Centers

As AI training clusters scale, they demand gigawatt-level power, and new energy benchmarks for data centers.

MID-2030s

Nuclear-Powered Data Centers Come Online

As proof that AI can run on reliable, clean energy, more companies adopt nuclear power in tech infrastructure.

Access to energy and compute power could define which organizations lead and which will be left behind.

Cheaper Energy, More Compute, Faster Innovation

The cheaper the energy, the cheaper the compute, making high-demand AI applications more affordable and experimentation more attractive. This shift not only accelerates innovation across industries but also allows companies with cheaper compute to gain a competitive edge in AI-driven markets, expanding the scope of what's possible in fields like health care, autonomous systems, and scientific research.

Compute Scarcity Will Reshape the Market

With rising AI adoption, demand for scalable, specialized compute infrastructure is surging, fueling competition and creating revenue opportunities in AI cloud services and hardware leasing. But as VC funding slows, a scarcity of compute resources could redefine market dynamics, with premium access favoring companies that can afford the cost, intensifying competition in AI infrastructure.

Data Center Security

As companies invest in creating differentiated, advanced AI models, robust data center security becomes essential for protecting valuable IP and maintaining a competitive edge. With the rising threat of cyber espionage, companies must prioritize end-to-end security, including their supply chains, to prevent costly breaches and IP theft. This also strengthens client trust, crucial as data protection becomes a top priority in AI.

Seamless Ecosystems Boost Brand Loyalty

As users increasingly rely on wearables and interconnected devices, one of two outcomes may emerge: They could become locked into single-brand ecosystems that offer seamless integration, enhancing customer lifetime value. Alternatively, users might demand interoperability across brands, expecting devices to work fluidly in any ecosystem. This push could drive new standards for compatibility, challenging brands to adapt.

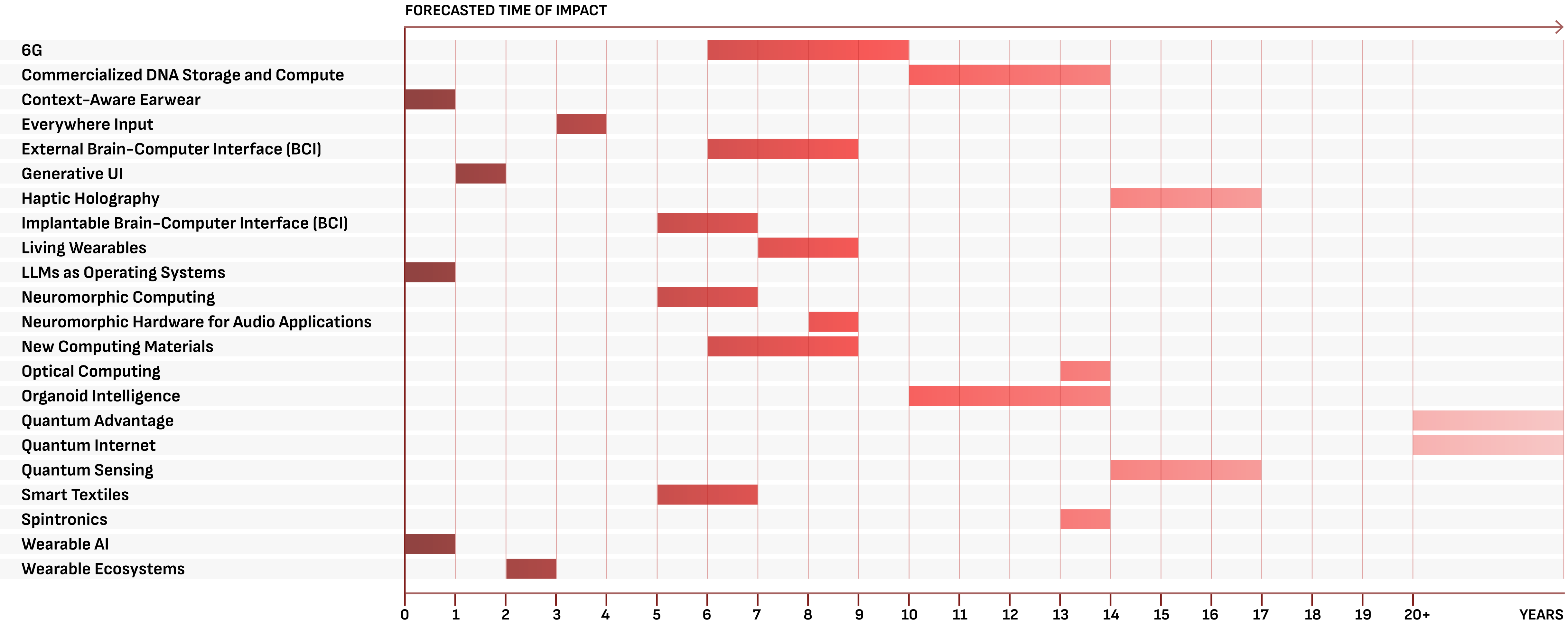
Cutting Costs with Distributed and Edge Computing

AI-enabled PCs and wearable ecosystems use edge computing to process data locally, reducing cloud dependency and transmission costs while enabling faster, real-time responses. This shift lowers cloud expenses and enhances the user experience, especially in sectors like gaming, health care, and retail, where low latency is essential.

Cross-Disciplinary Talent for Novel Architectures

As computing merges with biology, neurology, and physics, companies will need cross-disciplinary talent; recruitment and training will favor hybrid skills and interdisciplinary collaboration. Finding employees with computational and biological sciences experience may drive partnerships with academic institutions and require new recruitment strategies and in-house training to build teams adept at meeting complex needs.

AI-driven interfaces and wearables transform operations short-term, while advanced computing and quantum technologies redefine the long-term tech landscape.



The battle for AI computing shapes two paths to power: nations race for strategic control, while individuals compete through AI access and literacy.

BIOLOGICAL COMPUTING

Bio compute may start in research but could expand as AI demands strain traditional architectures, though entry costs will remain high. Ethical and regulatory issues will likely arise, especially in regions where biology is politicized.

COMPUTE INFRASTRUCTURE BUILD-OUT

Countries are rapidly expanding AI infrastructure, creating opportunities for companies that provide fiber optic cables, concrete, and energy. However, regulatory hurdles could slow progress, while China's centralized approach may give it a speed advantage.

INFRASTRUCTURE POLITICIZATION

Expect political debates to intensify as computing infrastructure strains resources, sparking environmental concerns. But because this same infrastructure enables AI-driven climate solutions, stakeholders will have to weigh immediate costs versus future benefits.

QUANTUM ADVANTAGE

Quantum computing hasn't shown an advantage over traditional compute yet, and even if it does, high costs may slow business adoption. But quantum advantage could disrupt geopolitics by breaking encryption, sparking an arms race for quantum security.

AUDIO-CENTRIC FORM FACTORS

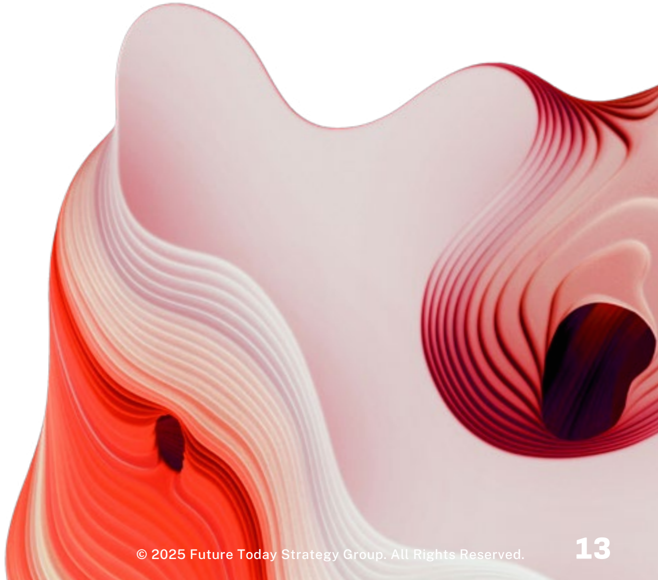
Multimodal AI can process text, images, and audio to enable more intuitive interactions. With audio-centric applications rising, form factors may evolve, embedding compute power in diverse spaces while reshaping security for voice-driven devices.

EDGE AI

Edge AI processes data where it's generated, reducing latency while enhancing privacy. Expect edge devices to become more powerful and efficient, supporting real-time applications in areas like health care and augmented reality.

WEARABLE AI ECOSYSTEM

Wearable AI ecosystems could distribute processing across devices, enhancing efficiency. Expect smartwatches, earbuds, and fitness bands to share compute and combine sensor data, providing richer insights into health and activity.



These individuals are rewriting how we'll work, think, and create with computers.

- **Dr. Taner Esat, researcher at Forschungszentrum Jülich in Germany,** for his significant contributions to the field of quantum sensing.
- **Dr. Yoel Fink, researcher and professor at the Massachusetts Institute of Technology,** for his work on multifunctional fibers and fiber assemblies, including developing rechargeable lithium-ion batteries in fiber form and multifunctional fibers for in vivo photopharmacology.
- **Dr. Ali Heydari, director of data center cooling and infrastructure at Nvidia,** for his work on advanced liquid-cooling systems.
- **Dr. Tom Harty, co-founder and chief technical officer at Oxford Ionics,** for his work on developing a quantum chip.
- **Dr. Seung-Woo Lee, researcher at the Quantum Technology Research Center at the Korea Institute of Science and Technology,** for his quantum error correction technology.
- **Michael Intrator, CEO and co-founder of CoreWeave, a specialized cloud computing company focused on GPU-accelerated workloads,** for his leadership in building out the computing infrastructure needed to fuel the AI boom.
- **Emre Ozer, senior director of processor development at Pragmatic,** a flexible chip manufacturer in Cambridge, England, for developing the first-ever flexible programmable chip not made of silicon.
- **Dr. Paolo Pintus, assistant professor at the University of Cagliari, Italy** for his significant contributions to the development of integrated photonic devices, especially in silicon photonics.
- **Brian Potter, senior infrastructure fellow at the Institute for Progress,** for his analysis on the technology and economics of the compute infrastructure build-out.
- **Dr. Shreyas Sen, Elmore Associate Professor at Purdue University,** for his work on designing AI chips inspired by the human nervous system to improve efficiency.
- **Jordan Schneider, founder of the "ChinaTalk" podcast,** for his writings and analysis on the geopolitics and national security of compute infrastructure.
- **Mike Davies, head of Intel's Neuromorphic Computing Lab,** for his leadership in developing neuromorphic computing technology.

The build out of AI computing infrastructure offers opportunities and global influence to its developers....

OPPORTUNITIES

New Export Players Emerge

Countries focused on developing a robust and scalable computing infrastructure could lead in exporting high-quality AI technologies, services, and intellectual property to emerging markets, strengthening economic influence.

Wearables Get Lighter and More Powerful

By offloading processing to interconnected ecosystems, companies can create lightweight wearables that don't sacrifice performance. These more comfortable devices could reshape industries like entertainment and health.

Cutting Edge Computing Improves Efficiency

Novel computing architectures like neuromorphic computing can reduce data center operational costs by cutting energy and cooling needs, offering a way to handle computationally intensive applications while meeting sustainability goals.

Personalized Health Gets Real

Biometric and body-interface technology allows companies to offer new, secure health monitoring products. Personalized health-tracking devices could combine fitness monitoring with secure logins, creating products that serve dual purposes.

...but the race to build it exposes critical vulnerabilities in energy, security, and privacy.

THREATS

Cyber Breaches Threaten AI Clusters

High-value AI clusters are prime targets for sophisticated cyberthreats, making cybersecurity investment essential. The stakes are especially high as thieves could use compromised AI clusters to steal proprietary models, impacting industries globally.

Energy Infrastructure Needs A Boost

US energy infrastructure lags in AI demands due to long build times, permitting, and supply chain issues. Meanwhile, China's centralized approach avoids regulatory hurdles and relies on nonrenewable energy, giving it a competitive edge.

US Supply Chain Faces Vulnerabilities

The US tech supply chain is particularly vulnerable in advanced semiconductor access, and disruptions in this supply chain could heavily impact American tech companies dependent on these critical components.

Data Privacy Risks Rise

As context-aware devices integrate sensors (cameras, microphones, biometrics), they're always monitoring environments. Such technologies can inadvertently capture sensitive data, leading to potential misuse by unauthorized entities.

The computing build-out creates opportunities but demands strategic planning.

Participate in the computing infrastructure expansion—it's not just for AI-focused hardware companies. This build-out needs advanced networking, security systems, supply chain tracking, and sustainable energy to connect and support data centers. Opportunities abound across industries as each component is vital for scaling AI, mitigating bottlenecks, and ensuring a secure, efficient infrastructure ecosystem.

Prepare employees to understand which products justify intensive compute use. As compute costs rise, companies will need to prioritize profitable applications over open experimentation, and all employees will need to be trained to embrace a strategic approach that maximizes resource value while balancing innovation and cost-efficiency.

Diversify component sourcing to reduce dependency and risks tied to a single country or region. Establish relationships with multiple suppliers and partners to distribute production and adapt smoothly to shifting geopolitical conditions.

Leverage AI at the edge to optimize data flow and reduce cloud storage and bandwidth costs. Edge devices can preprocess and filter data, sending only essential insights to the cloud. This reduces bandwidth, lowers storage needs, and keeps critical functions close to the source, creating a cost-effective and responsive data ecosystem through seamless cloud-edge integration.

While biological computing may seem niche, it has the potential to disrupt various industries, from health care to data storage. Companies unprepared for this shift may face competitive disadvantages. With energy-efficient advantages over traditional electronics, biological computing offers a sustainable alternative for companies focused on reducing energy use and environmental impact.

For young people, this can feel like a daunting time with many jobs at risk of disappearing in the future. To those reading this: Focus on building adaptable skills. The computing infrastructure build-out could be an ideal opportunity to gain practical experience and develop highly transferable skills.

Technology Deployment
Capital Expenditure
Talent Development
Regulatory Influence

Important terms to know before reading.

AI HYPERSCALERS
Major cloud providers with extensive infrastructure and resources that allow users to run AI applications at massive scale. These hyperscalers can support highly intensive AI workloads, offering the scalable computing power, storage, and specialized tools needed to deploy and manage large AI applications.

AI TRAINING CLUSTERS
High-performance computing setups that aggregate computational resources specifically to handle the intensive demands of AI model training. These clusters typically consist of multiple GPUs or specialized hardware optimized for handling large-scale data and complex algorithms in AI training.

BIOCOMPUTER
A computer that uses biological molecules like DNA and cells to store and process information.

BRAIN-COMPUTER INTERFACE (BCI)
A direct interface between the brain and computer that can enable control and communication by thought alone, with potential to help people with disabilities as well as elucidate cognition.

CENTRAL PROCESSING UNIT (CPU)
The key computer component that performs the computations, makes decisions on data, and tells the other components what to do. You can think of it as the computer's mission control center.

CLASSICAL COMPUTER
The standard binary digital computer that manipulates zeros and ones to store data and perform computations sequentially using hardware chips and switches.

EXASCALE COMPUTERS
Supercomputers capable of performing over 1 exaFLOPS, which is a quintillion calculations per second.

FAULT TOLERANCE
The ability of a quantum system to operate reliably despite errors and noise.

FORM FACTOR
The overall physical attributes and dimensions of a device according to standard specifications or for particular use cases. It impacts the usability and compatibility of hardware.

GRAPHICS PROCESSING UNIT (GPU)
A specialized circuit designed to rapidly process and manipulate computer graphics and image data.

HYBRID CLASSICAL-QUANTUM
A computational architecture that combines both classical computers and quantum computers to exploit the complementary strengths of each.

NEUROMORPHIC COMPUTING
Computer architectures that are inspired by the biological brain's structure and function.

OPEN SOURCE
Computer software or other products with source code that anyone can inspect, modify, and enhance.

ORGANOID INTELLIGENCE
A new scientific field of study that aims to actualize biological computing by utilizing 3D cultures of human brain cells and brain-machine interfaces.

PERVASIVE (UBIQUITOUS) COMPUTING
Aims to seamlessly integrate computer hardware and software into all objects and activities, creating an always-available, helpful computing environment.

Q-DAY
The hypothetical point in the future when a fully operational quantum computer capable of running practical quantum algorithms finally becomes available.

QUANTUM ADVANTAGE
Also known as quantum supremacy, refers to the potential capability of quantum computers to solve certain problems that are intractable for classical computers in practical time frames.

QUANTUM-AS-A-SERVICE
The provision of quantum computing resources on demand as a cloud service.

QUANTUM COMPUTER
A type of computer that utilizes quantum mechanical phenomena like superposition and entanglement to perform computations. Unlike classical computers, which operate on binary bits (0 or 1), these computers run on quantum bits, or qubits, representing a 0, 1, or a quantum superposition of both states at the same time. Since they consider multiple possibilities simultaneously, they can potentially be much faster at some types of problems than classical computers.

QUANTUM ENTANGLEMENT
A phenomenon in which two or more quantum particles are intrinsically linked to each other in such a way that the state of one particle cannot be described independently of the others, even when separated by a large distance.

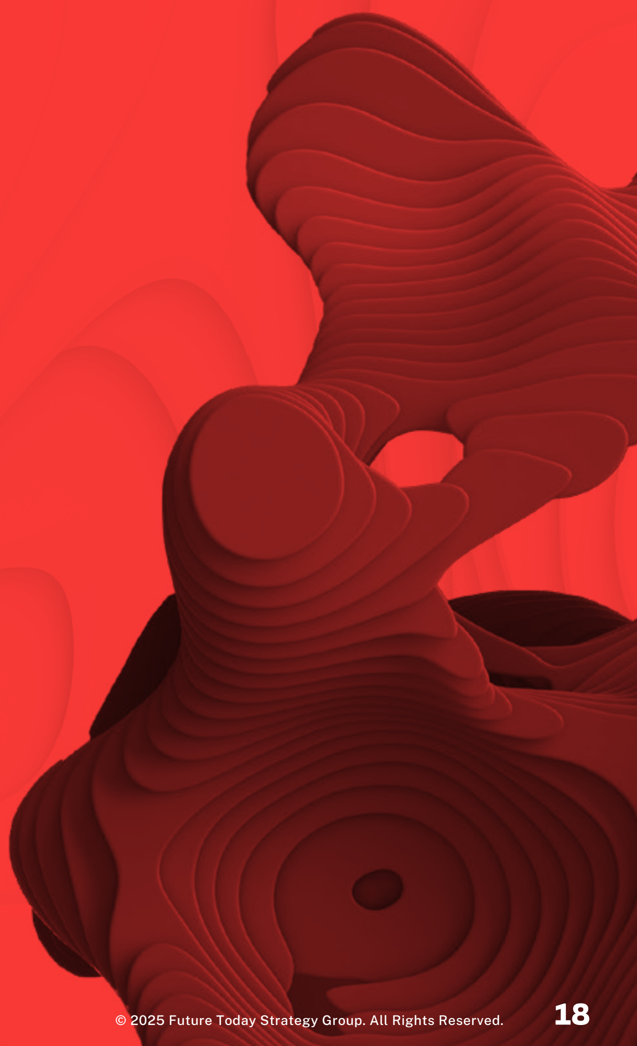
QUANTUM INTERNET
A hypothetical global quantum communication network that connects quantum processors using quantum entanglement and teleportation.

QUANTUM SUPERPOSITION
Allowing a quantum system to exist in multiple possible states at the same time until it is measured. The quantum parallel processing enabled by superposition is fundamental to achieving speedups and novel applications using quantum computers.

QUANTUM SUPREMACY
This refers to a quantum computer performing a task that no classical computer can match, regardless of its practical usefulness.

QUBIT
The basic unit of information in quantum computing. Unlike classical bits, qubits can be in a superposition of 0 and 1 simultaneously. The superposition, entanglement, and interference properties of qubits are what allow quantum algorithms to efficiently solve certain problems that are believed to be intractable on classical computers.

RISC-V
An open-source instruction set architecture based on established reduced instruction set computer (RISC) principles.



COMPUTING TRENDS

CHIPS

What are the different types of AI chips?

GPUs (Graphics Processing Units)

Originally designed for rendering graphics for gaming applications, GPUs have become indispensable in AI workloads due to their ability to handle parallel processing efficiently. GPUs are highly versatile and well-suited for the high computational demands of training AI models. Their ability to perform thousands of parallel operations simultaneously makes them particularly effective for matrix-heavy tasks like deep learning. In the AI field, GPUs are widely used in both training and inference, though they are more commonly associated with training large-scale neural networks. For example, Nvidia GPUs, such as the A100 and V100, are a standard in AI research and commercial applications due to their performance. Other notable manufacturers include AMD and Intel.

TPUs (Tensor Processing Units)

Developed by Google, TPUs are designed specifically to accelerate machine learning workloads, especially those involving deep learning models that use TensorFlow. TPUs are highly optimized for matrix operations and are tailored for deep learning tasks, making them efficient for both training and inference. TPUs excel at executing operations like matrix multiplications, which are central to neural networks. These processors are often used for training large models and handling inference tasks in Google Cloud's AI services. TPUs have powered significant projects like AlphaGo and large-scale models such as GPT. Google remains the sole manufacturer of TPUs, with the TPU v4 being the most recent version available for use.

FPGAs (Field Programmable Gate Arrays)

FPGAs are reconfigurable chips that can be programmed after manufacturing, allowing a unique blend of flexibility and performance. These chips are highly customizable, enabling developers to adapt the hardware for specific AI models and tasks. FPGAs are especially valuable in low-latency applications, where custom processing pipelines can significantly accelerate inference. They are primarily used in inference tasks, particularly in edge computing or environments that demand high power efficiency and fast execution times. Microsoft, for instance, utilizes FPGAs in its Azure AI infrastructure to handle inference workloads. Leading manufacturers in the FPGA space include Xilinx, now part of AMD, and Intel, with its Arria and Stratix series.

ASICs (Application-Specific Integrated Circuits)

ASICs are custom-built chips designed for a specific application, offering unparalleled performance and energy efficiency for that particular task. Since they are built for a single purpose, ASICs are extremely efficient, making them ideal for inference tasks where speed and power consumption are critical. These chips are primarily used in large-scale inference deployments, such as in data centers or in specialized hardware products like smartphones. For example, Apple's Neural Engine handles on-device AI tasks like facial recognition, while Google's Edge TPU powers inference in low-power environments. Tesla's Dojo D1 is another ASIC designed for high-performance computing in self-driving applications. These manufacturers create chips finely tuned for specific AI tasks, offering performance that general-purpose processors can't match.

NPUs (Neural Processing Units)

NPUs are specialized chips designed to handle AI tasks, with a focus on deep learning and neural network operations. Their architecture is optimized to efficiently process the specific types of calculations AI models require, such as matrix multiplications in convolutional layers. NPUs excel in performing these operations quickly and efficiently, making them highly suitable for both inference and, to a lesser extent, training. They are most commonly used for inference in mobile devices, edge computing hardware, and other low-power environments where energy efficiency is crucial. NPUs are increasingly found in smartphones, IoT devices, and automotive applications. Examples of NPU-based systems include Huawei's Kirin NPU, Qualcomm's Hexagon DSP with AI Engine, and Apple's Neural Engine.

RISC-V AI Chips

RISC-V is an open-source CPU architecture that has gained attention for AI workloads due to its flexibility and scalability. Unlike proprietary chip architectures, RISC-V allows developers to modify and optimize the architecture for specific AI tasks, which can be particularly advantageous for companies or applications that require custom hardware solutions. This flexibility has made RISC-V chips appealing for both training and inference tasks, although they are more commonly found in experimental AI systems or edge devices where adaptability is key. SiFive is one of the leading companies developing RISC-V-based AI chips, with other companies exploring custom designs for specialized AI tasks based on this architecture.

What is the difference between training vs. inference chips?

Training

Training a machine learning model involves feeding massive amounts of data into the model and adjusting its parameters over time. This process is compute-intensive, often requiring high parallelism and flexibility, which is why GPUs are commonly used for training. TPUs are also used for training, particularly in Google's ecosystem.

Inference

This is the phase when the trained model makes predictions. Inference workloads tend to focus on low-latency and power efficiency, particularly in edge or real-time applications. FPGAs and ASICs excel here because they can be optimized for fast and efficient inference tasks, while GPUs and TPUs can also be used for inference but are less power-efficient.

Why different chips for different applications?

Training and inference have different computational demands, and applications can range from cloud-based AI model training to real-time inference on mobile devices. This variability leads to the following considerations:

1. **Flexibility vs. specialization**
   GPUs offer flexibility, making them suitable for a wide range of tasks, while ASICs provide performance for specific, repetitive tasks.
2. **Cost and efficiency**
   ASICs and FPGAs are cost-efficient for large-scale, repetitive tasks like inference, whereas the general-purpose nature of GPUs can lead to higher operational costs.
3. **Scalability**
   TPUs are designed to scale well in cloud environments, where massive amounts of data can be processed. FPGAs are more suited for small-scale, specialized environments where low-latency and power efficiency are essential.

The choice of AI chip depends heavily on the specific needs of the application, including whether you're training a model or using it for inference, the power and latency requirements, and the environment in which the chip will operate (cloud vs. edge).

CHIPS

The US-China Chip War

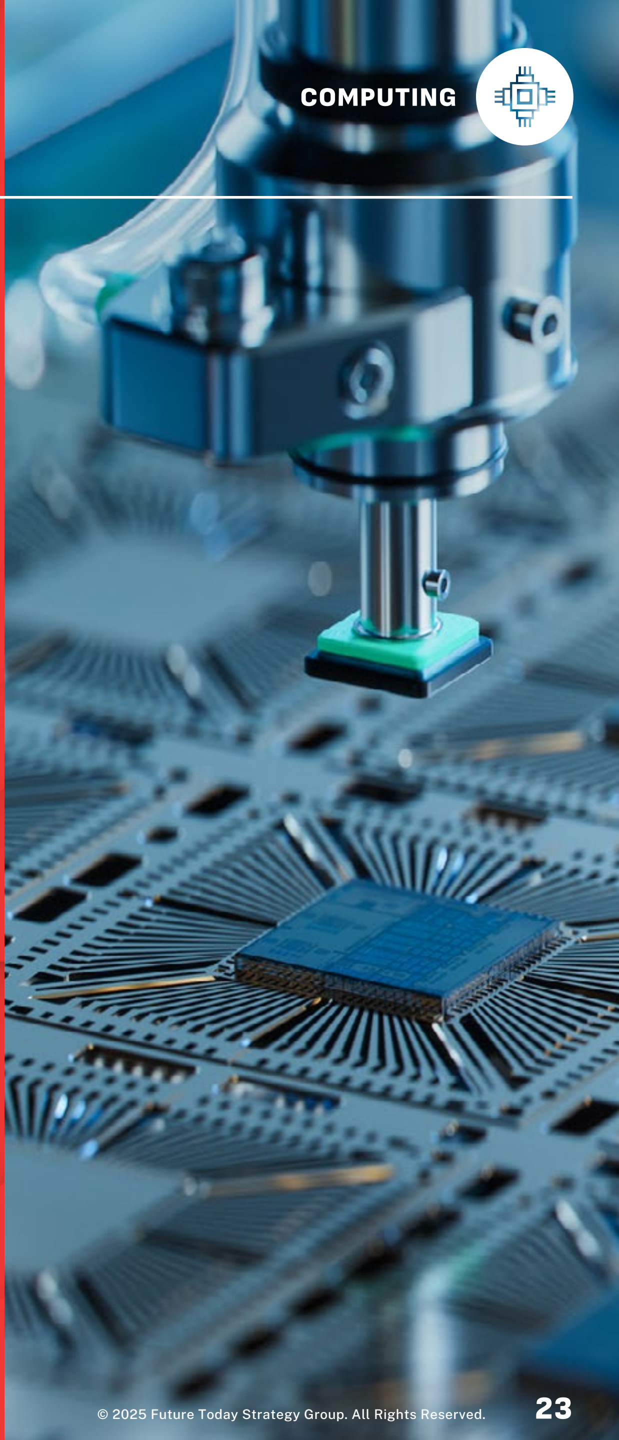
The ongoing US-China chip war has intensified, with the US Department of Commerce citing national security and foreign policy concerns to justify multiple rounds of export controls to limit China's development in advanced semiconductor technology. To strengthen its position, the US has been building a coalition with allies, notably striking a multilateral trade agreement with the Netherlands and Japan in January 2023 to restrict sales of advanced lithography equipment to China. The Netherlands, home to leading chipmaking equipment manufacturer ASML, is a crucial player. Despite these sanctions, older deep ultraviolet scanners have allowed China to produce 7-nanometer chips via its leading chip manufacturer, Semiconductor Manufacturing International Corp. (SMIC), although it remains behind global leaders such as Taiwan's TSMC.

The consequences of these controls have been felt both in China and the US. US-based chip giant Nvidia's sales in China and Hong Kong fell by 20% as of February 2023. China has responded to these restrictions with countermeasures, including banning the use of US-made Micron Technology chips in critical infrastructure projects, restricting exports of rare earth elements vital for chip manufacturing, and heavily investing in domestic semiconductor production capabilities. Beijing has poured more than $150 billion into its chip industry, including a $47 billion investment fund announced in May 2023. SMIC has become China's de facto national semiconductor champion, though it still lags behind global rivals in producing the most advanced chips.

Chips Onshoring

The US has ramped up efforts to onshore semiconductor manufacturing, spurred by global supply chain disruptions and national security concerns. The 2022 CHIPS and Science Act was a significant turning point: It earmarked $52 billion in grants and research funding for US-based semiconductor manufacturing, as well as an investment tax credit for chip manufacturers that establish or expand US operations. This has sparked a wave of private sector investments: More than $395 billion has been announced since its passage, and the US is on track to produce nearly 30% of the world's leading-edge chips by 2032, exceeding initial projections. TSMC announced that its first Arizona fab, which was previously delayed, has achieved early production yields that surpass comparable fabs in Taiwan—signaling early onshoring success.

The US isn't alone in these efforts. China was an early mover in trying to reduce dependence on foreign-made chips; the country launched the China Integrated Circuit Industry Investment Fund, commonly known as the "Big Fund," in 2014 to foster domestic chip production and achieve technological self-sufficiency. The European Union has rolled out the European Chips Act, which aims to attract 43 billion euros in public and private investment to boost semiconductor research, development, and manufacturing, including mega



CHIPS

fabs. And several countries in Southeast Asia and India are also working to expand their semiconductor capabilities. Malaysia is actively expanding into semiconductor manufacturing and design, building on its strengths in packaging and testing. Singapore has the most complete semiconductor supply chain in Southeast Asia and is the only country in the region with foundry manufacturing. While already major players, Japan and South Korea are also increasing investments.

AI Chips

Nvidia currently holds a commanding presence in the AI chip market, capturing between 70% and 95% of the market share. Its flagship AI accelerators, such as the H100, and proprietary CUDA software have enabled the company to triple its year-over-year sales for three consecutive quarters, driven by unprecedented demand for AI processors. The company's aggressive strategy, which includes releasing a new AI chip architecture annually, aims to deepen its dominance, making it difficult for competitors to catch up. However, despite its market lead, the company faces growing competition and risks.

The potential for disruption looms as tech giants like Google, Microsoft, Meta, Intel, and Amazon develop their own AI chip solutions. Google's Tensor Processing Units (TPUs), in use since 2015, have evolved into powerful chips, such as the newly released Trillium, which powers its Gemini and Imagen models. Similarly, Microsoft has begun incorporating AI chips from AMD, whose Instinct MI300X is seen as a potential rival to Nvidia's GPUs. The MI325X, expected to debut by year's end, features more than 150 billion transistors and 288 gigabytes of high-bandwidth memory, but its performance in real-world applications is yet to be determined. Intel, too, is attempting to carve out space in the AI market: The company recently launched its third-generation AI accelerator, Gaudi 3, which it claims is more cost-effective and efficient than Nvidia's H100 in running inference tasks. Another AI chip, Lunar Lake, launched last fall. Intel's AI efforts highlight the intensifying competition in AI chips, with price and efficiency becoming key factors. Startups like Cerebras Systems are also entering the fray, marketing its CS-3 system as a viable competitor to Nvidia. The field of AI chips is rapidly evolving, and while Nvidia currently leads, the market is far from static.

Chips for Inference

Once AI models are trained, they require significant compute power to deploy or "infer" answers for users, a process called inference. While model training builds the foundation, 99% of the compute in a model's lifecycle is dedicated to inference, running the model repeatedly to respond to real-time queries. This means efficiency gains during inference are crucial to manage compute costs effectively. For instance, increasing training compute can reduce inference demands, potentially lowering costs by up to 80%. But inference workloads in data centers differ greatly from training tasks: Since inference doesn't require as much raw processing power, data centers can use older chips effectively, focusing instead on memory bandwidth—the ability of chips to access and manage data quickly, which is the primary performance constraint in inference. Unlike training, which can run on large clusters of accelerators, inference is often managed in smaller batches and can operate efficiently with only tens of accelerators instead of thousands. This compact setup also enables data centers to allocate resources flexibly, using space for various hardware types and workloads.

Another key difference is the proximity requirement for inference: While training can be done remotely, inference needs low-latency connections close to end users, often necessitating data centers near backbone fiber networks for fast responses. In response, companies like Nvidia and AMD are expanding their GPU offerings, including the Nvidia-dominant A100 and H100 GPUs and AMD's Instinct MI300 series, optimized for both training and inference. Meanwhile,

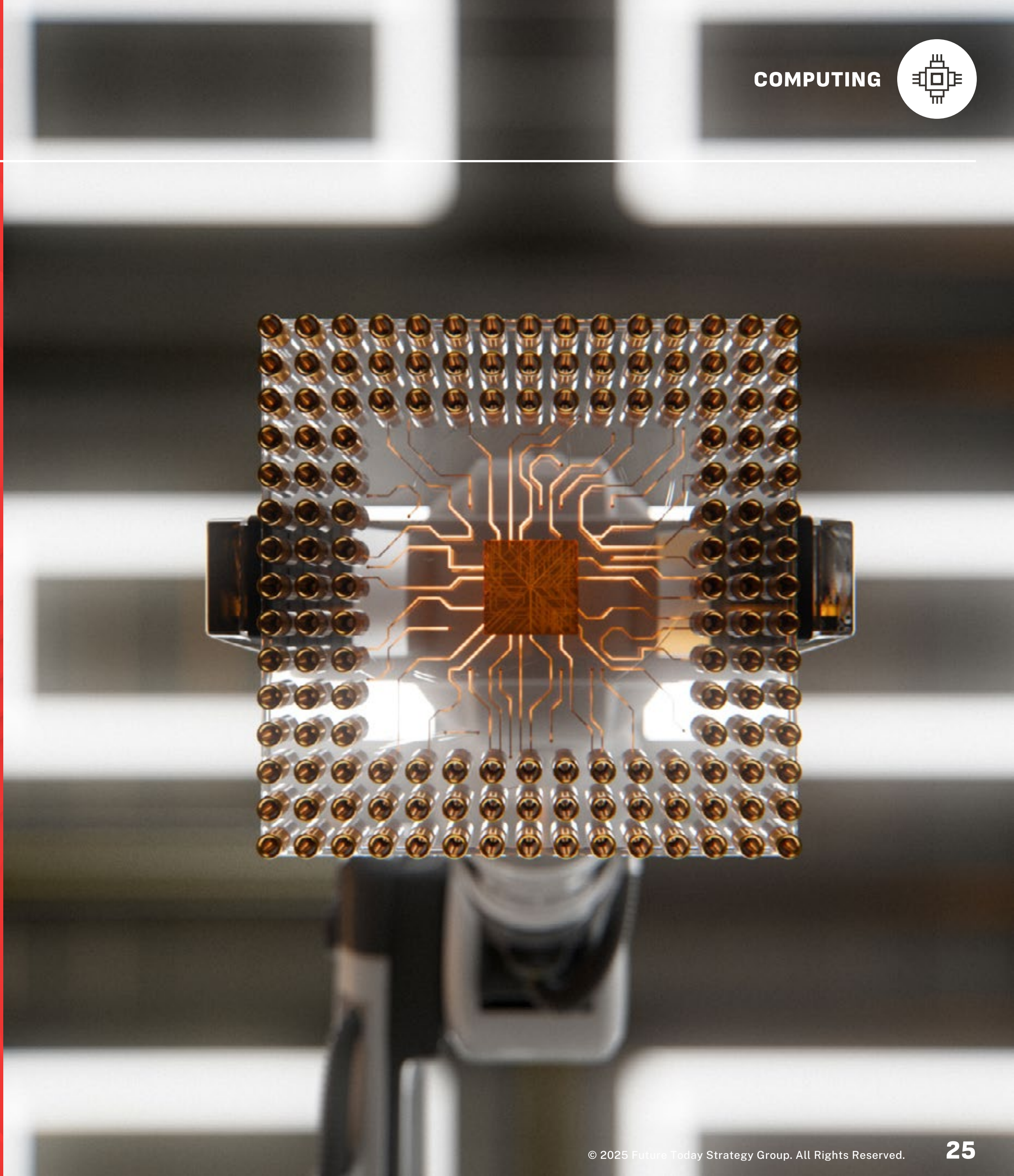
CHIPS

Qualcomm, AWS, Meta, and specialized AI chip startups like Groq and Untether AI are developing inference-specific chips, recognizing that efficient, lightweight inference is essential for scalable AI deployment across industries.

AI-Assisted Chip Design

AI is playing an increasingly vital role in chip design, with major tech companies like Nvidia leveraging it to enhance the process. Chip design requires the precise placement of billions of transistors, which significantly affects cost, performance, and power efficiency. Traditionally, designing these chips has been a highly complex and labor-intensive task. Now, AI techniques like reinforcement learning are being used to optimize chips for power, performance, and area, creating more efficient designs than those made by humans alone. Companies such as Nvidia, Intel, AMD, IBM, Google, and Apple are integrating AI into various stages of the chip design process, from conceptual design and transistor modeling to simulation, verification, and testing. AI has been particularly effective in simplifying processes like clock tree synthesis, a critical step in distributing signals across a chip. Intel, for instance, employed AI to help design its Meteor Lake processors, which consist of innovative chiplets stacked into a single package. These processors feature a dedicated neural processing unit (NPU) alongside the CPU and GPU, optimizing them for AI tasks such as image recognition, video processing, and natural language processing.

AI's impact on chip design extends throughout the entire product lifecycle. The collaboration between chip designers, electronic design automation (EDA) tool providers, and foundries is key to advancing AI-driven chip design. A notable example is the partnership between Synopsys and TSMC, which are working together to develop AI-assisted EDA solutions for advanced chip process nodes. These efforts underscore AI's transformative potential, enabling more powerful and efficient chips that can better handle the growing demands of modern AI applications.



CHIPS

Counterfeit Chip Detection

Counterfeit chips are a growing problem, costing US semiconductor companies more than $7.5 billion annually and posing serious risks to both industry and consumer safety. The shortage of new semiconductor chips has led to an increase in counterfeit chips, which can result in malfunctions or even security vulnerabilities. There are estimates that as many as 15% of all spare and replacement semiconductors purchased by the Pentagon are counterfeit. To combat this, researchers are developing advanced technologies to verify the authenticity of semiconductor components. One such breakthrough is RAPTOR (residual attention-based processing of tampered optical responses), an AI-powered detection method developed by Purdue University researchers that uses deep learning to analyze light patterns from gold nanoparticles embedded in semiconductor chips or their packaging. These nanoparticles scatter light in unique ways, creating specific patterns that can be recorded and stored for later authentication. The RAPTOR technique involves taking microscopic images of the nanoparticle scattering patterns, allowing for high-contrast imaging even though the materials are mostly transparent to light. The AI model then distinguishes between natural degradation of the chip over time and intentional tampering by analyzing these patterns. The model is trained to detect adversarial attempts to replace nanoparticles, which could potentially hide tampering efforts. This technology can detect tampering with 97.6% accuracy, outperforming traditional detection methods by up to 40%.

Flexible, Programmable Chips

A flexible, programmable chip is a type of microprocessor that combines the processing capabilities of traditional chips with a bendable or foldable structure, enabling it to operate on flexible surfaces. Unlike conventional rigid silicon-based chips, these flexible chips are made from bendable and ultrathin alternative materials. Because of this design, the chip can conform to non-flat surfaces and endure physical stress without breaking or losing functionality. Programmability is key to these chips, as it allows developers to write and execute specific instructions so the chip can perform complex tasks like data processing and machine learning operations.

The Flex-RV microprocessor exemplifies this advancement in flexible computing. Developed by Pragmatic Semiconductor, this 32-bit processor runs on an open-source RISC-V architecture and is built with non-silicon materials, using indium gallium zinc oxide transistors on a polyimide substrate. Designed to operate while bent, Flex-RV offers a durable, energy-efficient solution capable of executing machine learning tasks while consuming less than 6 MW at 60 kHz. Robustness tests show that Flex-RV maintains consistent performance, even under bending, with only minor performance variation. The Flex-RV microprocessor is groundbreaking because it brings computing power to flexible, affordable applications that rigid silicon chips cannot support, such as wearable health monitors, smart packaging sensors to track produce freshness, and adaptable components for soft robotics.

Advancing Data Processing with Photonics

Silicon-photonic (SiPh) chips are circuits that combine traditional silicon electronics with optical (light-based) components, allowing them to process data using light rather than electricity. This approach enables much faster data speeds and lower energy use, making SiPh chips ideal for tasks that require processing large volumes of data efficiently. In February 2024, researchers at the University of Pennsylvania announced development of one of these types of chips to perform mathematical computations (vector-matrix multiplication) for training AI. Beyond faster speeds and lower energy use, this chip offers enhanced privacy: By performing many computations simultaneously, it eliminates the need to store sensitive data in a computer's working memory. This design makes a computer powered by this technology nearly impossible to hack. Chinese

CHIPS

researchers have developed a photonic chip called the OPCA chip, which operates entirely with light and processes 100 billion pixels in just 6 nanoseconds. This speed is ideal for edge AI applications like autonomous driving and robotics, enabling rapid, energy-efficient image processing directly on the device.

RISC-V

RISC-V, first released in 2010 by UC Berkeley's Parallel Computing Lab, is an open-source, modular instruction set architecture that allows for flexible and cost-effective chip design. This openness enables companies to innovate without the licensing fees associated with proprietary architectures like x86 and Arm. However, RISC-V faces notable trade-offs, particularly in performance. As of 2024, RISC-V processors still lag behind top-tier Arm and x86 CPUs in raw speed, with benchmarks consistently placing even the fastest RISC-V chips well behind their competitors. To address these challenges, several initiatives are underway. The Berkeley SonicBOOM project is advancing experimental cores with features like out-of-order execution to enhance efficiency. Companies like Tenstorrent, led by renowned chip designer Jim Keller, are also investing in high-performance RISC-V solutions to close the gap.

Adoption of RISC-V is steadily growing, particularly in data centers. Startups such as Ventana Micro Systems and Tenstorrent are leveraging its flexibility to create innovative server solutions. Major industry players have also shown strong commitment to RISC-V. Samsung has established a Silicon Valley R&D lab dedicated to RISC-V development, while Alibaba's C910 server chip, based on RISC-V, is already powering cloud servers for Scaleway in France. Alibaba plans to release an even more advanced server chip soon, signaling confidence in RISC-V's future. Nvidia has embraced the architecture on a massive scale, with projections indicating that around 1 billion RISC-V cores will be shipped in its 2024 product lineup alone. This adoption highlights RISC-V's versatility and its growing influence across the technology landscape.

Vertical Integration

Companies are increasingly adopting vertical integration strategies, aiming to control the entire development ecosystem—from hardware to large language models (LLMs). Nvidia, in particular, stands out for its aggressive pursuit of vertical integration, which has positioned the company as a dominant player in the AI and high-performance computing markets. The company has created a comprehensive, vertically integrated ecosystem that includes its proprietary CUDA software stack, which has become the de facto standard for AI development. By controlling the entire stack—ranging from chips to software frameworks and services—Nvidia sets a high barrier for competitors. Its dominance in both hardware and software has created significant lock-in effects, as developers and enterprises become reliant on CUDA for seamless AI deployment. For companies to compete with Nvidia, they need to not only match its advanced hardware capabilities but also offer an equivalent software ecosystem, a daunting challenge in a rapidly evolving field.

While vertical integration is currently in the spotlight due to AI, it's not a novel concept. MBA programs have long taught this strategy, and tech giants like IBM have employed it for decades. IBM's history of vertical integration in computing predates the current AI boom—for decades, the company has developed, manufactured, and controlled multiple layers of the computing stack, from hardware to software. This long-standing approach has allowed IBM to remain a key player in the tech industry, illustrating how vertical integration can provide a sustained competitive advantage across generations.

New Materials to Power Advanced Computing

Recent advances in semiconductor materials are driving breakthroughs in computing, electronics, and bioengineering. Moving beyond traditional silicon, researchers

CHIPS

are exploring topological semimetals and other quantum materials that offer higher power output with lower energy use. Topological semimetals, for example, exhibit unique electron behavior that supports spintronics—a technology that utilizes electron spin rather than charge for data storage and processing, enabling faster and more efficient devices. Two recent examples: University of Minnesota researchers created a thin-film topological semimetal, and a team in Korea created a new p-type semiconductor alloy of selenium and tellurium (Se-Te). The latter material, which can be deposited at room temperature, shows improved mobility and a higher on/off current ratio, outperforming existing transistors. Such advancements are expected to benefit next-generation displays, including high-refresh-rate OLED TVs and extended reality devices.

In chip manufacturing, where creating smaller, faster processors is always a goal, Applied Materials introduced two new materials for use in ultra-tiny 2nm (nanometer) chips. For perspective, a nanometer is about 50,000 times thinner than a human hair. These materials include a ruthenium-cobalt metal coating that reduces thickness, allowing electric current to pass through with less resistance, making chips faster and more efficient.

Meanwhile, scientists at the University of Chicago's Pritzker School of Molecular Engineering developed a hydrogel semiconductor, which combines the electrical properties of semiconductors with the flexibility of hydrogels, which are soft stretchy materials. This soft semiconductor could be used in flexible medical devices implanted in the body, providing better comfort and adaptability to biological tissues.

Optical Computing

Optical computing processes data with light (photons) rather than electricity, which could lead to faster, more energy-efficient computing systems. An international team, including scientists from the University of Pittsburgh, has developed a photonic memory platform using a magneto-optical material called cerium-substituted yttrium iron garnet (Ce) on silicon—a setup that stores and processes data in light rather than electronics. By applying a magnetic field, the researchers can adjust light speed within micro-ring resonators, giving them control over the light's direction and speed. This technology achieves both high switching speed and energy efficiency, with durability reaching 2.4 billion cycles, making it suitable for demanding photonic memory applications.

At Purdue University, scientists achieved a quantum photonic breakthrough by trapping cesium atoms on photonic circuits to act like transistors for light, controlling the flow of photons. By using lasers to freeze cesium atoms near absolute zero, the atoms can interact precisely with circulating light in a microring resonator and the atoms can control photon movement. This atomic-level control over light flow brings optical computing closer by enabling logic and data storage in photonic circuits. In another advancement, researchers at the Swiss Federal Institute of Technology in Lausanne developed a cost-effective, efficient photonic integrated circuit (PIC) based on lithium tantalate. Unlike silicon, which struggles with high-speed data, lithium tantalate performs well in optical systems and is easier to produce at scale. The team's new PICs achieve minimal optical loss, supporting high-speed operations at telecom frequencies. This breakthrough promises scalable, affordable photonic systems for optical computing, communication, and lidar, making high-speed, low-energy optical technologies more accessible. As a whole, these advancements move optical computing from theoretical to practical, scalable applications.

COMPUTING INFRASTRUCTURE FOR AI

COMPUTING INFRASTRUCTURE FOR AI

Scaling AI Clusters

The computing requirements for training leading-edge AI models have surged, with computational demand increasing fivefold annually from 2010 to 2024. This escalation calls for highly advanced AI training clusters—specialized networks of interconnected systems designed to support the immense processing loads required by modern machine learning models. These clusters primarily consist of numerous GPUs or TPUs working in tandem to meet the rigorous needs of AI training, and their sizes are expanding rapidly to keep pace with these demands.

For the US to remain competitive in AI development, projections indicate that by 2030, the most advanced AI clusters may require around 5 GW of power and incorporate up to 1 million accelerators. Currently, the largest clusters contain approximately 100,000 GPUs and demand tens of megawatts. By comparison, OpenAI's GPT-4 reportedly needed 25,000 GPUs, drawing 30MW of power, enough to power 25,000 US homes. Future expansions by companies like Microsoft and OpenAI aim to build clusters with even greater capabilities, with plans underway for facilities requiring up to 5 GW, or 150 times today's typical power usage.

The infrastructure push includes investments by key AI computing companies. The Stargate Project is an initiative to build a network of AI-dedicated data centers. OpenAI and its partners have announced plans to invest $100 billion to build a supercluster in Texas as part of a larger $500 billion computing infrastructure expansion project. CoreWeave recently secured $7.5 billion in financing to expand its data center operations, aiming to double its capacity by the end of 2024. Similarly, Microsoft and OpenAI's planned cluster in Wisconsin, set to launch in 2026, will feature around 100,000 of Nvidia's latest accelerators, consuming more than 100MW. The next stage, expected by 2028, envisions a supercomputer requiring close to 5 GW, indicating the scale of infrastructure needed to sustain AI training growth. Building this infrastructure at an unprecedented scale will be essential to meet AI's increasing computational demands and to support continued advancement in AI capabilities globally.

Nuclear-Powered Data Centers

The resurgence of nuclear power is gaining momentum as tech giants like Microsoft and Amazon turn to nuclear energy to power their vast data centers. Microsoft has signed a 20-year power purchase agreement with Constellation Energy to buy all the electricity generated by the Three Mile Island nuclear plant's Unit 1, which is set to reopen as the Crane Clean Energy Center by 2028. This marks a remarkable revival of a site known for the 1979 Unit 2 partial meltdown, the worst commercial nuclear accident in US history. AWS recently acquired a data center next to the Susquehanna Steam Electric Station in Pennsylvania. The proximity to nuclear power allows AWS to directly tap into the 2.5 GW of energy produced by the plant, cutting out reliance on the grid and ensuring a steady, clean energy supply. This method of colocating data centers next to nuclear plants is gaining popularity as a way to secure reliable power while avoiding grid-related fees. Proposals for similar setups are surfacing in New Jersey, Texas, Ohio, and other regions. Looking further ahead, ventures like Helion Energy aim to make nuclear fusion a reality. Supported by investors like OpenAI CEO Sam Altman, Helion plans to launch the world's first fusion power plant by 2028. If successful, this breakthrough could provide Microsoft with fusion-generated energy, offering an even more advanced clean power source for the company's future data centers.

Fortifying the Data Center

Securing AI data centers is emerging as a major trend among tech companies as cyberthreats grow more sophisticated and targeted. With AI systems increasingly deployed for critical applications, companies like OpenAI, Microsoft, and Google are focusing intensively on protecting their in-

COMPUTING INFRASTRUCTURE FOR AI

frastructure. Recent high-profile breaches highlight the risks: In early 2023, a hacker accessed OpenAI's internal messaging systems to steal sensitive design information, and more recently, the Chinese espionage group Diplomatic Specter launched a spear-phishing campaign against OpenAI employees in an attempt to exfiltrate proprietary data. This incident underscores the escalating vulnerability of AI facilities, pushing companies to prioritize robust security practices to safeguard their valuable intellectual property.

This shift toward stronger data center security involves securing the entire ecosystem, from the supply chain to internal networks. Companies now face a new imperative: ensuring that all hardware entering their facilities—from chips to network devices—is screened to detect tampering. Enhanced protocols for network and storage security have also become critical. This includes encryption of all data transfers within the cluster, strict identity verification for every device connected to sensitive networks, and real-time monitoring of network links, especially fiber connections between data centers, to detect any unauthorized access.

Fortifying data center security is not just about protecting corporate assets but also about retaining a competitive edge. By investing in security, companies can mitigate the risk of adversaries stealing AI models that would otherwise require massive investments to develop. As competition grows and AI infrastructure becomes even more critical, data center security will be key to advancing these technologies safely and securely.

Automated Data Centers

In 2024, Nvidia filed a patent for "intelligent components of a data center," detailing a system where autonomous robots manage server racks and components. The patent describes robotic systems that handle server rack manipulation and maintenance, controlled by a central monitoring system. This automation addresses the growing challenges of manually managing expanding data centers, by enabling dynamic rack reconfiguration and real-time cooling adjustments based on computational demands and environmental conditions.

Industry experts see this as strategic positioning for Nvidia. "This reinforces Nvidia's leadership in data center innovation," notes Trevor Morgan, OpenDrives' senior vice president of operations. While Nvidia may not build data centers directly, the patent could generate substantial licensing revenue. Gartner analysts project that advanced robotics could enhance data center operational efficiency by 30% through continuous monitoring and precise adjustments.

The industry is already embracing robotics. Novva Data Centers deploys modified Boston Dynamics Spot robots (dubbed WIRE—Wes' Industrious Robot Employee) for autonomous facility patrols, temperature monitoring, and security verification. Similarly, DE-CIX's "Patchy McPatchbot" handles automated fiber optic connections in distribution frames, demonstrating robotics' diverse applications in modern data centers.

Specialized AI Cloud

A specialized AI cloud is a cloud infrastructure designed specifically to meet the demands of AI and machine learning workloads. Unlike general-purpose cloud platforms, specialized AI clouds are built to handle the complex computational needs and large datasets associated with AI applications. These clouds utilize high-performance GPUs like Nvidia's A100 or H100, and sometimes TPUs, which are great for tasks like neural network training. These platforms often come with prebuilt environments that include popular machine learning frameworks like TensorFlow, PyTorch, and JAX, reducing setup time for developers. Specialized AI clouds like Google's Vertex AI, AWS SageMaker, and Azure AI also offer managed services, which provide end-to-end support for building, training, deploying, and monitoring AI models without requiring deep expertise in cloud infrastructure. These

COMPUTING INFRASTRUCTURE FOR AI

platforms typically feature tools to streamline machine learning operations, including automated model versioning, testing, and retraining, ensuring smooth transitions from development to production.

AI on the Mainframe

Mainframes, IBM's legacy systems, quietly underpin much of the financial world, processing 70% of all credit card transactions. Known for their unmatched ability to handle high transaction volumes with top-tier security, mainframes remain vital in industries that demand reliability and speed. Despite attempts over the past decade to migrate workloads from mainframes to the cloud, many companies have struggled. Not all workloads are suited for the cloud—some run more efficiently and cost-effectively on mainframes, while certain applications are simply too difficult to move. But AI may be the key to revitalizing the mainframe. Modern mainframes are now being equipped with AI-specific hardware and software to handle various AI workloads. IBM's current z16 mainframe already integrates AI for running machine learning models, and the upcoming Z mainframe, set for release in 2025, will feature the Telum II processor and Spyre AI Accelerator Card, designed to support large language models (LLMs) and advanced machine learning. These advancements will power applications like fraud detection, real-time transaction analysis, and AIOps for system monitoring and anomaly detection.

One of the reasons companies considered moving off mainframes was the shrinking pool of COBOL programmers as they near retirement age. However, AI's evolution in programming—allowing developers to code in natural language without needing deep expertise—could shift this trend. Instead of training new cohorts in COBOL, companies may find it more practical to have them utilize AI to translate their work and continue using the platform. Additionally, running machine learning models and LLMs on a mainframe offers a security advantage. Since data doesn't need to be distributed across disparate systems for AI processing, enterprises can maintain stricter control over data governance and security, reducing the risk of breaches.

Data Center Cities

Data centers play a key role in the development and training of AI language models, as they provide the massive amounts of electricity, storage, and cloud computing power needed to support their operations. Data centers are the backbone of the AI boom and, as such, cities around the world are positioning themselves as hubs for this critical digital infrastructure. These cities are tapping into the upside potential by expanding their data center capacity, attracting substantial investments and driving economic growth, from Europe to Latin America and Southeast Asia.

Ireland, with more than 80 data centers, sees these facilities as essential to its tech hub vision, though the rapid growth of AI workloads is straining the national grid with unprecedented energy demand. Porto Alegre, Brazil, launched Scala AI City in September 2024, starting with a $500 million investment that could expand to $90 billion, making it Latin America's largest AI-focused data center. Johor Bahru, Malaysia, is the fastest-growing data center market in Southeast Asia, boasting 1.6GW of capacity. In the UK, Northumberland will host one of Europe's largest AI data centers, backed by a $13.3 billion Blackstone investment. In the US, northern Virginia remains the "data center capital of the world," housing key facilities for AWS, Google Cloud, and Microsoft Azure. Other US cities seeing expansion include San Antonio, with a $482.6 million Microsoft investment; Racine, Wisconsin, where Microsoft is building a new AI data center; central Ohio, where AWS is adding $7.8 billion to the $10.3 billion it has invested since 2015; and Memphis, Tennessee, where xAI has made its home.

Hyperscale Water Usage

Hyperscale data centers, which power cloud computing and internet services, use vast amounts of water to cool their servers. These centers can consume 3 million to

COMPUTING INFRASTRUCTURE FOR AI

5 million gallons of water per day—about as much as a city with 30,000 to 40,000 people would use. AI demands even more power and water, and as more hyperscale data centers shift their workloads to AI, their resource consumption continues to rise. This growing water demand has raised concerns, especially in areas like Virginia, where data center water usage jumped from 1.13 billion gallons in 2019 to 1.85 billion gallons in 2023. As a result, major companies such as Amazon, Google, and Microsoft have started to report their water usage after years of keeping it private, and industries have implemented a new standard, Water Usage Effectiveness (WUE), to measure how efficiently these centers use water.

In Memphis, xAI's data center plans to draw more than 1 million gallons of water daily from the Memphis Aquifer to cool its servers. Similarly, Google's data center in The Dalles, Oregon, has nearly tripled its water use over the past five years and now consumes more than a quarter of the city's total water supply. According to Google's 2023 Environmental Sustainability Report, the company's data centers and offices used 5.6 billion gallons of water in 2022, a 20% increase from 2021. Though these numbers are significant, it's worth contextualizing that they still make up a comparatively small amount of the world's total daily water consumption, which is largely dominated by agricultural use. However, as the demand for cloud services fed by thirsty AI grows, so will the pressure on data centers to balance their operations with sustainable water practices.

Advanced Cooling Technologies

As hyperscale data centers continue to consume large amounts of water and AI workloads grow more intensive, traditional air-based cooling methods will likely become insufficient. In response, the industry is shifting toward advanced cooling technologies: Methods like direct-to-chip cooling and submersion cooling—where servers are immersed in nonconductive fluids—are becoming more efficient options.

The challenge is in using water efficiently. One solution comes from the US Department of Energy, which launched the COOLERCHIPS program in 2022, allocating $82 million to develop energy-efficient cooling technologies. The goal is to reduce cooling energy use to below 5% of a data center's total IT load, regardless of system density or location.

The COOLERCHIPS program has spurred collaboration between leading companies and research institutions, including HP, IBM, Nvidia, the US National Renewable Energy Laboratory, UC Davis, University of Florida, and University of Illinois. Some projects aim to dissipate up to 40 watts of heat per square centimeter, a significant challenge for modern data centers. Companies are bringing new liquid-cooling technologies to market. CoolIT, a leader in liquid-cooling for AI and high-performance computing, recently announced a new line of products tailored to support Nvidia's Blackwell AI platform. These include direct liquid-cooled servers and coolant distribution units designed to meet Blackwell's performance demands. In October 2024, Hewlett Packard Enterprise announced the industry's first 100% fanless direct liquid-cooling system for large-scale AI deployments. This system reduces the power needed for cooling each server blade by 37%, cutting down on utility costs, carbon emissions, and noise in data centers.

SCENARIO YEAR 2030

TWO FUTURES: THE COMPUTE CHALLENGE

The US stands at a pivotal moment in AI development: scaling compute infrastructure. With AI's vast appetite for processing power, the future of innovation, economic strength, and global influence hinges on our ability to expand the compute capacity needed to drive these advancements. Success means an era of abundant, affordable compute, fueling scientific breakthroughs, democratizing creativity, and accelerating medical progress. But if we hit a bottleneck, compute power will become scarce and fiercely contested, reserved for only the highest-paying applications, stifling innovation and widening the global gap in AI advancement.

1 THE AGE OF ABUNDANT COMPUTE

Noah sits in his home studio, directing his seventh feature film this year. No cameras, no crew, no million-dollar budget—just his imagination and an AI system running on cheap, abundant compute. He speaks into his microphone: “Adjust the lighting in scene four, make it more noir,” and then he watches as the AI instantly reprocesses the entire sequence, transforming his sci-fi thriller’s atmosphere in real time.

Across town, Dr. Williams runs a protein-folding simulation for the thousandth time today. Five years ago, such computational intensity would have cost millions of dollars. Now, thanks to breakthrough computing architectures that made AI processing as cheap as electricity, her small biotech startup simulates millions of molecular interactions daily. Last week, they discovered a promising treatment for Alzheimer’s. Three other labs confirmed it within days—the kind of rapid verification that only becomes possible when compute constraints disappear.

The great compute scaling of the mid to late 2020s changed everything. Abundant nuclear and geothermal energy supply and AI-optimized chips pushed capacity well beyond the critical 130 GW growth threshold. But it wasn’t just about adding power—new architectures made AI computation exponentially more efficient, creating a surplus of computational capability.

Now, high school students run complex climate models for science projects. Independent artists generate feature-length animations overnight. Small research labs simulate decades of climate change scenarios. The democratization of compute power has unleashed a creative and scientific renaissance. Every day, somewhere in the world, someone with a good idea can test it, simulate it, and bring it to life without worrying about computational costs.

2 HITTING THE COMPUTE BOTTLENECK

Noah sits dejectedly in his empty home studio. “Insufficient compute resources available in your tier.” He speaks to his computer, “Just adjust the lighting in scene four.” Nothing happens. Five years ago, he could have processed this in seconds. Now, independent creators like him can barely afford the most basic AI operations.

The world has hit a hard cap on compute scaling. Regulatory pressures and bureaucracy have strangled AI infrastructure expansion, turning computational power into a luxury commodity. Large companies devote their limited resources only to profitable applications, abandoning experimentation. Even text-based AI models strain budgets, while multimodal processing—combining images, video, and text—has become prohibitively expensive. Basic research has ground to a halt, with tasks like protein simulations now requiring funding on par with particle accelerator experiments. Innovation across medicine, entertainment, and science has slowed to a crawl.

The effects ripple through society. High schools have abandoned their AI curriculum—the computational costs of running even basic training exercises exceed most educational budgets. Medical researchers share limited computing time on aging infrastructure, often waiting months to test new hypotheses. What was once a revolutionary tool for scientific discovery has become another scarce resource, carefully rationed and jealously guarded.

Most concerning is the shift in AI alignment. With China dominating the compute landscape, one country’s values and priorities increasingly shape AI development. US companies, struggling with limited resources, can’t compete with the scale and scope of Chinese AI models. The dream of democratized AI has given way to a harsh reality: In the compute-constrained world of 2030, innovation follows the path of least computational resistance, and that path increasingly leads east.

PERSONAL COMPUTING

PERSONAL COMPUTING

Us as Input

We are increasingly merging with the technology we interact with, transforming into "invisible" accessories that seamlessly integrate with our devices. Traditional input methods like keyboards and mice are being replaced by more natural and intuitive modes of interaction, such as facial recognition, body movements, and eye tracking. Biometric data, including facial recognition and fingerprint scanning, is now commonly used for security, simplifying how we log in and authenticate our identities. Technologies that were once considered niche or assistive, like eye tracking, are now part of mainstream devices, allowing users to navigate screens with just their eyes. Voice input is also becoming a dominant method for communication with our devices, slowly replacing the keyboard for tasks like typing. Even skin input is on the horizon: Researchers at Cornell University have developed on-skin devices that enable direct interaction with technology. One such device, SkinPaper, uses silicone-treated washi paper that conforms to the body, allowing for simple on-skin interactions. This technology has broad potential applications, ranging from health monitoring to assistive technologies for individuals with disabilities. In a separate development, a team from Seoul National University and Stanford University has created a spray-on smart skin. Made from nanowires embedded in a polyurethane coating, this mesh adapts to the natural contours and wrinkles of the skin. It uses AI to recognize hand gestures and typing without the need for external devices like cameras or gloves, offering a more seamless way to control and communicate with devices. These advancements signify a shift in the way we interface with technology. By incorporating our bodies into the interface, these technologies open new possibilities for personal safety, health monitoring, and accessibility, all while making our interactions more intuitive and fluid.

AI embedded PCs

AI-powered PCs are emerging as the next evolution in personal computing, with companies like Intel, Microsoft, Qualcomm, Nvidia, and AMD at the forefront of integrating neural processing units (NPUs) directly into devices. This signifies a shift toward standalone in-device AI capabilities, allowing users to interact with advanced AI features without needing constant internet or cloud connectivity. The inclusion of NPUs in PCs will enable real-time processing of AI tasks, such as natural language understanding, computer vision, and machine learning, all within the device itself. These AI PCs are designed to enhance the user experience, not just by improving processing speed but by making interactions more intuitive and personalized. With an AI-powered PC, you could speak commands like, "Open last week's report," and the device would process and execute the request instantly without needing internet access. These PCs could also adapt to user habits, such as adjusting settings for comfort during long work hours, making interactions smoother and more responsive in real time.

In January 2025, NVIDIA announced Project DIGITS, a desktop-sized personal AI supercomputer, aimed at AI researchers, data scientists and students. There are also rumors of an upcoming NVIDIA AI PC processor called N1, which has not been confirmed at the time of this writing. Microsoft's newly redesigned Copilot+ offers another example of how these AI-driven systems can integrate voice and vision capabilities. With Copilot Voice and Vision, users can interact with their PC in more natural ways, such as speaking commands or having the system recognize and respond to what's displayed on the screen. Intel has plans to produce 40 million AI-enabled CPUs in 2024, increasing to 60 million in 2025. Meanwhile, Qualcomm and AMD are also developing AI PCs with NPUs focused on energy-efficient AI processing, essential for real-time interactions on portable devices. At Computex 2024, Nvidia hinted at new RTX AI PCs, potentially previewing upcoming Copilot+ systems from Asus and MSI. These systems are expected to feature Nvidia's graphics cards alongside AMD's latest Strix CPUs.

PERSONAL COMPUTING

Wearable Ecosystems

One of the biggest challenges with AR glasses is that integrating powerful computing directly into the glasses often results in a bulky, unattractive, and heavy design. This compromises both their aesthetic appeal and comfort, making widespread consumer adoption less likely. A potential solution lies in distributing the computing tasks across a network of interconnected devices. Instead of relying solely on the glasses to handle all functions, a system of complementary devices—a wearable ecosystem—could work together to share processing and sensory tasks.

By distributing the processing load and combining data from various sensors across multiple devices, users can enjoy a more sophisticated and efficient experience. The devices can share information on environmental factors or the user's activities, resulting in richer insights that wouldn't be possible with just one device. This modular approach also helps keep the glasses lightweight and more wearable, as they no longer need to house all the technology themselves. A practical example of these wearable ecosystems is Meta's Orion smart glasses. The Orion system consists of three key components: the AR glasses, a "neural wristband" for control, and a wireless compute puck that looks similar to a large phone battery pack. The glasses function independently of a smartphone or computer, but the puck is crucial for their operation. If the glasses and puck are separated by more than about 12 feet, the glasses become nonfunctional, illustrating how tightly integrated the components are in this wearable ecosystem.

Generative UI

Generative UI refers to the use of generative AI to automatically create, adapt, or personalize a user interface (UI) based on real-time inputs and context. This approach leverages context-aware computing to dynamically adjust the UI, making it more intuitive, responsive, and aligned with the user's specific needs. The interface could change based on factors like the user's environment, location, or activity, creating a seamless and efficient experience.

However, the true potential of generative UI goes beyond simple context-based adaptations; it reimagines the way we interact with applications altogether. Why rely on a static fitness app when a generative AI system can pull relevant data from multiple sources and create a personalized interface tailored just for you? Imagine a world where no two people have the same interface—where design and layout evolve based on individual preferences. One user might see structured menus, while another might navigate through visually rich, clickable images, depending on how they prefer to interact with information.

Now, imagine instead of switching between websites to shop for running shoes, you simply ask or type your query into a central system that generates the perfect interface, drawing from different sources instantly. Even more radically, the need for explicit, visible interfaces could disappear altogether. As multimodal AI models advance, they will understand and respond to natural language, gestures, and other inputs, allowing for seamless, interface-less interactions where users no longer need to rely on predefined screens or layouts. This opens up a future of fluid, personalized human-AI interaction unconstrained by traditional UI norms.

Accessibility Tech Goes Mainstream

Accessibility technologies, once designed specifically for people with disabilities, are increasingly being adapted for mainstream use, leading to innovations that benefit a much wider audience. Devices like OrCam MyEye, originally created to help the visually impaired by reading text and recognizing faces, now serve non-accessible applications such as hands-free navigation and industrial information access. Eye-tracking technology, developed to help those with mobility issues navigate a computer screen, is being integrated into mainstream applications, enabling anyone to navigate digital interfaces without physical touch. Advanced hearing aids

PERSONAL COMPUTING

have evolved beyond simple sound amplification: Modern versions of these devices feature AI-powered sound processing, which helps improve speech recognition in noisy environments, making them desirable to a broader audience. There are also tactile screen technologies originally designed to allow blind users to read Braille digitally, but that now have the potential to be adapted for more widespread use, such as providing tactile and haptic feedback on touchscreen devices. Accessibility technology is no longer limited to a niche market—it can be adapted to help all people see, hear, and communicate more effectively, broadening their impact far beyond their original purpose.

Wearable Intelligence

Wearable AI refers to devices that integrate artificial intelligence into everyday accessories, enabling more intuitive and personalized interactions. These devices, such as smartwatches, glasses, and rings, use AI to process data and assist users in real time. One recent example is Humane AI's AI Pin, a small device that attaches to clothing and functions as a personal assistant. The pin, launched in 2024, supports voice commands and hand gestures, projects information using a laser display instead of a screen, and includes a camera for photos and visual recognition. However, the Humane AI Pin has faced criticism and a lackluster consumer enthusiasm for inconsistent AI responses, slow processing, and limited functionality. And this concept isn't new; Microsoft Research introduced Skinput in the 2010s, which, like the Humane AI pin, used the skin as an interface but didn't make it to market due to fidelity issues. Other AI wearables, like the Samsung Galaxy Ring, Meta's smart glasses with Ray-Ban, and AI-powered watches, are exploring different ways to incorporate AI into daily accessories while addressing issues like power consumption and internet dependency.

High power consumption is a major energy efficiency challenge for wearable AI. Purdue University researcher Shreyas Sen is developing solutions inspired by the human nervous system to overcome these limitations. His research focuses on chips that enable AI wearables to function offline, reducing energy needs. Sen's team is creating "in-sensor analytics" chips that interpret only necessary data for specific tasks. These chips use electro-quasistatic signals, transmitting information 100 times more efficiently than Bluetooth or Wi-Fi. By creating an artificial "peripheral nervous system," Sen envisions AI wearables performing complex tasks without needing frequent charging or constant internet access.

Flexible Displays

Flexible displays have been an exciting prospect for years, yet widespread consumer adoption has been slow. But 2024 may have been a turning point. One recent example is Samsung's "flex in and out" display technology, increasing the durability and usability of foldable screens. Samsung's US patent "Electronic Device Including Flexible Display" outlines a foldable device with an advanced hinge system, which dynamically adapts to various folding angles to improve both flexibility and audio performance. Samsung has also incorporated a photocurable resin in the adhesive layers, boosting the display's resilience to repeated folding—an essential step in addressing durability issues that have previously hindered flexible screens.

Developing highly flexible displays also depends heavily on new material innovations. Scientists are focusing on 2D materials like graphene, hexagonal boron nitride (h-BN), and transition metal dichalcogenides (TMDs), each offering unique benefits for flexibility and durability. Graphene's excellent conductivity and flexibility, TMDs' light-emitting properties, and h-BN's stability and insulation potential make them ideal candidates for flexible OLED displays. Companies are incorporating these into components such as thin-film transistors, transparent electrodes, and protective layers. But significant challenges remain for 2D materials, particularly in achieving

PERSONAL COMPUTING

uniform quality, scalability, and stability for mass production. Ongoing research aims to refine these processes to produce durable, scalable displays for smartphones, wearables, and foldable laptops. If successful, 2D materials could reshape the industry, enabling high-quality, flexible displays for a new generation of consumer electronics.

Smart Textiles

Smart textiles, powered by multimaterial fibers, are transforming how we think about clothing and wearables, enabling fabrics to become dynamic, responsive, and even intelligent. These fibers—long, thin strands engineered to conduct electricity, emit light, and sense environmental changes—are making it possible to integrate electronic functionality directly into fabrics. By weaving such fibers into garments, researchers are developing textiles that can monitor health, track movement, or even change color or shape in response to stimuli. Researchers recently developed a soft, flexible fiber that interacts with the human body to perform wireless energy transfer and sensory processing without requiring rigid chips or batteries. These fibers can light up in response to touch and offer various forms of feedback, making them a versatile component for digital interactions. Such innovations highlight the potential for creating chipless textile electronics that could redefine wearable tech by eliminating the need for bulky hardware.

Another significant challenge in smart textile development has been the scaling of sensor integration, particularly when it comes to making reliable connections between rigid electronic components and soft fabrics. However, recent research into distributed sensing along fibers has made substantial progress in overcoming this obstacle. A prototype garment using helical auxetic yarn sensors demonstrates how strain can be measured along multiple regions of a single fiber, accurately monitoring joint movements with minimal error. This represents a breakthrough in making smart clothing capable of tracking complex biomechanical data while remaining comfortable and unobtrusive.

PERSONAL COMPUTING

Haptic Holography

Haptic holography merges holography with haptic feedback, enabling users to see, feel, and interact with 3D virtual objects in midair using their bare hands. This approach lets users manipulate virtual objects without physical devices, using ultrasound waves to create the feeling of holding and touching a projected object. Current research in this area includes the University of Glasgow's "aerohaptics" system, which uses controlled air jets to simulate touch sensations, enabling interaction with holographic objects without physical devices. Meanwhile, researchers at UC Santa Barbara have uncovered a new phenomenon fundamental to developing holographic haptic displays, by utilizing focused ultrasound waves in air to create tactile feedback. However, the current technology faces limitations. Advanced imaging and simulations reveal that these haptic feedback systems generate shock waves in the skin, reducing the precision of touch sensations. To advance holographic touch technology, scientists must develop new methods to either mitigate or harness these shock wave effects. As the field progresses, haptic holography has the potential to transform various industries, including virtual and augmented reality, remote surgery and medical training, industrial design and prototyping, and education and interactive learning.

Context-Aware Earwear

Context-aware audio interaction integrates both audio and visual inputs to create smarter, more intuitive ways for users to interact with their devices. Voice is becoming the primary interface for many AI-driven systems, thanks to advancements in natural language processing, allowing for natural communication. However, these interactions require a deeper understanding of context to be truly effective. For example, a request made in a private setting might need a discreet response through earphones, while the same command in a public space could be handled differently. Combined visual and audio input is becoming essential to provide this context, enhancing how devices understand and respond to their users. One way is by integrating cameras into earbuds for various uses, like Meta's "Camerabuds," which include AI-driven object recognition and real-time translation capabilities. Apple is reportedly working on similar technology with additional health monitoring features, showing the potential for multifunctional, context-aware devices. Huawei is also advancing this trend by filing a patent for earbuds with cameras designed for pedestrian safety. These devices would use AI to detect approaching objects or intersections, alerting the wearer to potential hazards—a safety feature that could be especially useful for people navigating busy urban environments. At the University of Washington, researchers have developed an AI system that lets users "enroll" a speaker simply by looking at the person briefly while wearing camera-equipped earphones. The system isolates and plays back only the enrolled speaker's voice, even in noisy environments. This combination of audio and visual inputs allows for real-time, focused listening, improving the user experience in dynamic settings like busy streets or crowded events.

Spatial Audio and 3D soundscaping

Spatial audio and 3D soundscaping involve creating a three-dimensional sound environment that surrounds the listener, enhancing the sense of immersion. This technology positions audio sources in virtual space, making it seem as if sounds are coming from different directions and distances. Using advanced algorithms and binaural audio techniques, spatial audio mimics real-world sound behavior, making it ideal for gaming, virtual reality, and other immersive experiences. A core concept in this technology is the head-related transfer function, which explains how humans perceive the direction of sounds based on differences in timing and frequency as they reach each ear. By digitally replicating these auditory cues, engineers can create a sense of depth and directionality, making the listener feel as if sounds are moving around them. Directional speakers are of-

PERSONAL COMPUTING

ten used to further enhance spatial audio, enabling precise sound targeting in specific areas or to individual listeners. These speakers contribute to dynamic sound environments where sound sources can appear to move around the listener, adding to the immersive experience. Technologies like Dante (Digital Audio Network Through Ethernet) help manage and route audio signals efficiently in complex systems, making it easier to control and direct sound within 3D soundscapes. There are several software platforms that also assist in creating and processing 3D audio. For example, Waves Nx simulates three-dimensional sound through stereo headphones, allowing users to experience immersive soundscapes even in everyday audio settings. As a result, spatial audio and 3D soundscaping have become essential tools in delivering more engaging and realistic sound experiences across various fields.

Neuromorphic Hardware for Audio Applications

Neuromorphic hardware for audio applications refers to specialized hardware that mimics the neural structure of the brain, enabling energy-efficient, always-on audio processing. This technology is ideal for battery-powered devices requiring low power consumption, real-time processing, and on-device computation, offering significant advantages over traditional audio processing methods. By processing audio data locally, neuromorphic hardware enhances privacy and reduces latency, making it perfect for edge devices like smart speakers, earbuds, and security systems. Additionally, neuromorphic hardware excels at pattern recognition, enabling efficient tasks such as keyword spotting, environmental sound classification, and acoustic event detection. This makes it an excellent fit for applications in voice-activated devices, industrial monitoring through vibration analysis, and underwater acoustic detection. A leading example of neuromorphic audio hardware is SynSense's Xylo Audio platform. Built on the Xylo neuromorphic inference core, this ultra-low-power platform supports a variety of audio applications, from in-home event monitoring to underwater acoustic detection. It enables intelligent audio-driven features in battery-powered edge devices, offering commercial and research partners new opportunities to integrate advanced audio capabilities into their products.

SCENARIO YEAR 2032

HAPTIC HOLOGRAPHY

By 2032, passwords have evolved far beyond simple text or biometrics. Now, unlocking devices requires a completely unique, tactile interaction with invisible, holographic objects—an experience that feels astonishingly real but is visible only to the senses. Haptic holography combines 3D holograms with touch feedback, letting users feel virtual objects as if they're real. Using technologies like ultrasound and air jets, it creates sensations like pressing, holding, or manipulating items in midair. Haptic holography has changed device authentication: Not only do users need to "feel" the right object, but they must also perform a specific, predetermined action with it.

The air before Sarah's phone flickers slightly, and though no image appears, she reaches out, instinctively feeling for something unseen. A burst of precisely calibrated sound waves and air pressure creates a sensation in her palm: her grandpa's watch. Sarah had spent countless hours running her fingers over its band and its smooth face. Now this virtual version feels exactly the same—the slight chip on the face from when Grandpa nicked it with a bowling ball, the loose link from the years of taking it off at night and putting it back on again in the morning.

The security sequence requires her to manipulate this phantom watch in ways only she would know—she "winds" the small knob on the side of the haptic hologram watch two times clockwise and once more counterclockwise.

Others might see Sarah standing there with her fingers seemingly dancing through empty air, but she's working through a sophisticated authentication cadence that's unique to her muscle memory. Even if someone recorded her movements, they couldn't replicate the exact pressure she applies to the haptic hologram, the precise angles of manipulation, the specific timing between actions—all deeply encoded in her nervous system through years of handling the real watch.

The beauty of the system lies in its invisibility and intimacy. These haptic holograms respond only to her touch, creating a personal choreography of pressure points and motion that's impossible to forge. It's not just what Sarah does—it's how she does it, the emotional memory built into each gesture, the unique way her fingers remember this treasured object.

BIOLOGICAL COMPUTING

BIOLOGICAL COMPUTING

Neuromorphic Computing

The most energy-efficient computer in existence today is the human brain. Unlike traditional von Neumann architecture, which requires constant data transfer between memory and processor—consuming both time and energy—the brain stores and processes information simultaneously, with remarkable efficiency. This has inspired researchers to rethink computing architectures, leading to the development of neuromorphic computing, which is an umbrella term for computers that aim to mimic the brain's structure and functionality. In 2024, Intel announced the Hala Point system, the world's largest neuromorphic computer. The parallel-processing system is powered by Intel's Loihi 2 processor, enabling 1.15 billion artificial neurons that will be used to address scientific challenges in areas like device physics and informatics, solving problems where traditional AI hardware struggles, particularly at scale.

Sydney-based startup BrainChip recently unveiled the Akida Pico, a neuromorphic chip designed specifically for devices at the extreme edge, such as mobile phones and wearables, where power limitations are critical. Additionally, German startup SpiNNcloud Systems introduced a hybrid computing platform that integrates traditional AI accelerators with neuromorphic computing. The systems, available in various configurations, include a flagship model capable of simulating 10 billion neurons—approximately one-tenth the capacity of the human brain—bringing brain-inspired computing closer to practical applications.

Organoid Intelligence

In 2023, researchers at Johns Hopkins University outlined a vision for biocomputers powered by human brain cells in a paper published in Frontiers in Science. Led by Thomas Hartung, the team presented a roadmap for "organoid intelligence," aiming to develop biological computing using 3D cultures of human brain cells. These tiny 3D organoids, no larger than a pen tip, contain neurons and circuitry capable of supporting basic functions like learning and memory. Unlike neuromorphic computers that mimic brain function using silicon, organoid intelligence utilizes actual human biology for computing operations. Although traditional computers can process calculations much faster than humans, human brains excel at complex decision-making tasks, such as distinguishing between a dog and a cat. Implementing AI on organoids could be crucial for achieving human-like complex decision-making abilities in machines.

This technology is already being adopted for commercial and research purposes. A Swiss company, FinalSpark, offers cloud-accessible computing services using real organoids through its Neuroplatform. The system, which has been functioning continuously for more than four years, is comprised of 16 organoids, maintained in microfluidic incubators at body temperature (37°C). These organoids can transmit and receive electrical signals, learn, and perform tasks through electrical stimulation. Researchers can employ chemical stimulation, such as dopamine release, to "reward" the organoids. Designed for biocomputing research and wetware computing experiments, the platform includes a Python-based API to interact with these organoids. FinalSpark asserts that its system could be up to a million times more energy-efficient than conventional computers.

DNA Storage and Compute

Researchers are exploring DNA computing as an alternative storage medium to resource-intensive data centers, due to DNA's immense storage capacity and long-term stability. The primary challenge has been efficiently storing, retrieving, and computing data using DNA, where information is encoded in nucleic acids rather than binary code, which has hindered the full realization of DNA's potential as a storage medium. Until recently, DNA technology was viewed as useful mainly for long-term data storage, with limited potential for performing all the functions of electronic computing in a repeatable, programma-

BIOLOGICAL COMPUTING

ble manner. This has changed with recent research from North Carolina State University and Johns Hopkins University: Researchers have developed a DNA-based system capable of performing a full range of data storage and computing operations. The breakthrough system can store, retrieve, compute, erase, and rewrite data using DNA as the medium, demonstrating that DNA technologies can now rival traditional electronic systems in versatility. This development marks a significant step forward in DNA computing, as previous technologies could only complete some, but not all, of these operations. It shows that DNA-based computing and data storage may one day provide a viable and efficient alternative to electronic devices, especially for long-term and large-scale data management.

Living Wearables

A "living wearable" is a device that blends living biological elements with traditional wearable technology, often using synthetic biology. These devices combine the features of regular wearables with the unique abilities of living organisms, allowing the wearable to interact with the human body and environment in ways that traditional wearables cannot. By integrating living biological components—like engineered bacteria or cell-free synthetic circuits—into flexible substrates and textiles, these devices can monitor physiological and environmental changes in real time, broadening applications for health care, environmental safety, and even national security. One major benefit is noninvasive health monitoring, where biosensors within these wearables can detect specific markers or pathogens, including disease indicators or environmental toxins, by using synthetic biology techniques like engineered biological circuits. For example, a face mask with embedded CRISPR technology has been developed to detect SARS-CoV-2 by merely pressing a button. This capability, offering laboratory-level detection at room temperature without complex procedures, demonstrates how living wearables can provide immediate, precise diagnostic capabilities.

Living wearables also open the door to self-sustaining systems for prolonged use, which is especially relevant in settings where traditional power sources are limited, such as space exploration. Photosynthetic wearables, for instance, use cyanobacteria to produce oxygen, potentially supporting life in closed environments or enabling sustainable energy production for bio-powered devices. Additionally, skin patches that incorporate living cells could not only monitor health markers but also release therapeutic compounds in response to physiological needs.

Implantable BCI

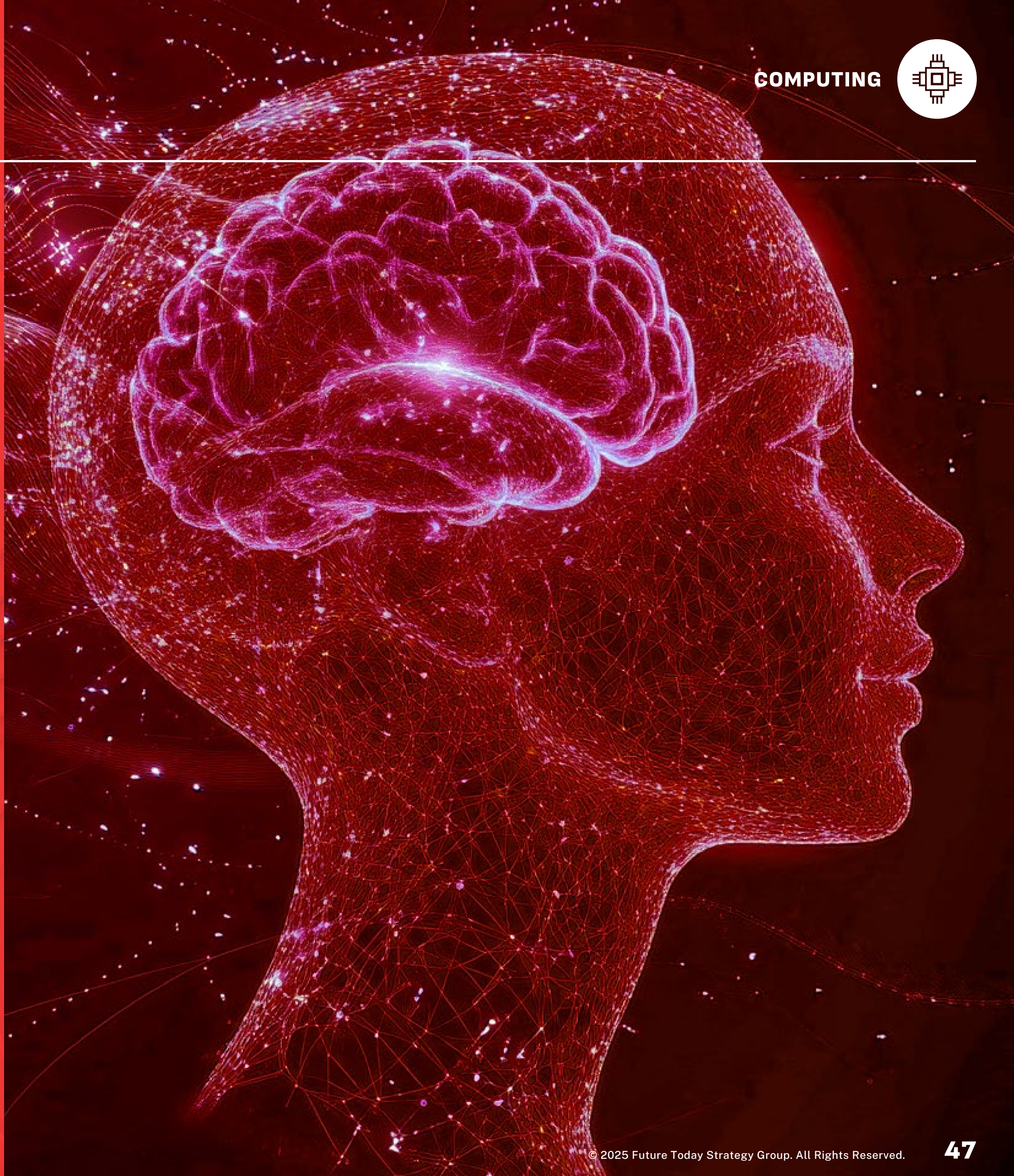
Recent developments in implantable brain-computer interfaces (BCIs) have created new possibilities for enhancing human-machine interaction, particularly in the field of health care. In 2024, Neuralink successfully implanted its BCI device, dubbed "Telepathy," in a second patient, Alex, following its first test on Noland Arbaugh, a man paralyzed from the neck down. This new implant enabled Alex to break world records in BCI cursor control. Despite significant advances, Neuralink encountered early technical challenges and continues to face skepticism. In Arbaugh's case, 85% of the device's flexible threads retracted from his brain within a month after implantation, which affected the overall performance of the device. Another significant milestone comes from UC Davis Health, where researchers developed a BCI capable of translating brain signals into speech with 97% accuracy. This technology was successfully implanted in a patient with ALS, enabling near-fluent communication. Also significant: Synchron implanted an ALS patient with its BCI technology, and the patient was then able to interact with Amazon's Alexa device to control his environment. These advancements demonstrate the profound potential to transform the quality of life for individuals with neurodegenerative diseases or severe paralysis, offering them innovative methods to communicate, interact, and engage with the world. Beyond accessibility, BCIs could revolutionize human-machine interaction

BIOLOGICAL COMPUTING

by drastically increasing the bandwidth between the brain and digital systems, potentially by a factor of 1,000 or more. This technology holds the promise not only of enhancing cognitive capabilities but also enabling ultra-fast, seamless communication between people, pushing the boundaries of how we connect and collaborate.

External BCI

While brain-computer interfaces (BCIs) have the potential to drastically improve mobility and communication for people with disabilities, most consumers are unlikely to embrace implanted devices in the near future. This hesitation suggests that noninvasive BCIs will dominate the commercial market in the short term. Noninvasive BCIs are appealing for being cost-effective, safe, and accessible to a broad range of users. However, because these systems record signals from the scalp rather than directly from the brain, their signal quality can be limited. But recent breakthroughs are addressing these limitations. Researchers at Carnegie Mellon University demonstrated a novel approach to noninvasive BCIs by using focused ultrasound stimulation in combination with a wearable electroencephalogram (EEG) device. Their study, involving 25 human subjects, is the first of its kind to integrate focused ultrasound stimulation for bidirectional BCIs—meaning it can both encode and decode brain waves using machine learning. The technique targets specific neural circuits, enhancing both signal quality and the overall performance of noninvasive BCIs. This work could significantly boost the capabilities of external BCIs, making them more viable for broader applications. Last year saw a growth in noninvasive wearable BCI technology investment, with companies exploring EEGs, ultrasounds, and magnetic stimulation. Magnus Medical, for example, has developed transcranial magnetic stimulation to treat major depressive disorder. While these wearable BCIs provide less insight into the deeper brain regions compared to implanted devices, their ease of use and noninvasive nature make them a promising avenue for wider adoption in both health care and consumer markets.



SCENARIO YEAR 2040

LIVING TACTICAL

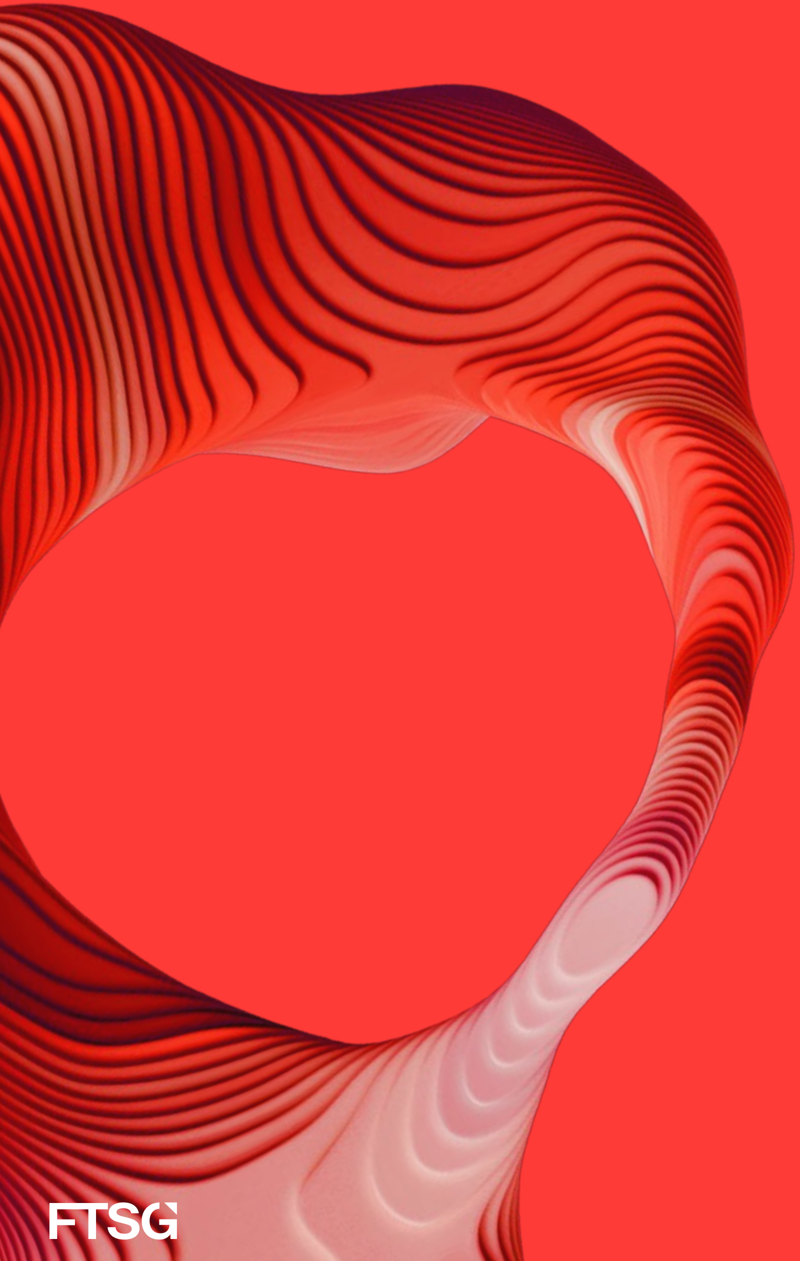
Through the ruins of a darkened city, US troops move like shadows. Each soldier wears a neural helmet—a living wearable that meshes with their body, monitoring vitals and enhancing senses. These bio-integrated devices link to the External Sensory Network, connecting them to silent drones patrolling overhead. The swarm maps the battlefield in real-time, feeding environmental data directly to the soldiers. To the enemy, they are invisible; to each other, they are omniscient.

A gunshot cracks the night. Within seconds, the drones pinpoint the shooter through audio triangulation. The helmets respond, releasing precise doses of performance enhancers to steady hands and sharpen focus. If chemical weapons appear, the neural interface stands ready to inject countermeasures directly into bloodstreams. The soldiers adjust their approach wordlessly, human and machine moving as one.

The drones unleash another tactic: They project phantom sounds across the battlefield—helicopter rotors, tank treads, marching boots. The enemy scatters, chasing ghosts. The helmets pick up on rising hostile heart rates as confusion spreads.

Silent, calculating, alive, the helmets make their soldiers both human and something more. Each step forward is guided by invisible maps and bio-data; every breath monitored, every heartbeat measured. The troops close in on their objective as the enemy falls further into the illusions, fearing threats that aren't real, trusting senses that betray them.

In this new warfare, victory belongs to those who master the symphony of combat. Every footstep tells a story, every echoed command reveals a position, and sound itself becomes both weapon and shield. The soldiers advance through their acoustic realm, invisible in the darkness but seeing everything through the perfect clarity of sound.



QUANTUM COMPUTING

> **Significant adoption of quantum computers is around the corner. Quantum is one of the biggest, most important technological races of our generation.**

Itamar Sivan, Co-founder & CEO of Quantum Machines

QUANTUM COMPUTING

Quantum Advantage

When Google unveiled Willow, many overstated its capabilities. While Willow achieved key breakthroughs—such as surpassing the threshold for quantum error correction and improving coherence times—it did not demonstrate quantum advantage. Quantum advantage refers to the theoretical point where quantum computers can outperform classical computers in solving specific tasks, such as optimization, simulations, or computations. When achieved, it will mark a major milestone in quantum computing; at that point, these machines could tackle problems that are currently beyond the reach of even the most powerful classical supercomputers. But the big hurdles in the way include developing quantum hardware and algorithms that surpass classical methods in both speed and accuracy.

For now, we are in an intermediate stage that is often called quantum utility, where quantum systems begin to outperform classical computers on tasks related to quantum mechanics itself, such as simulating quantum circuits. And while quantum advantage has yet to be conclusively demonstrated on a real-world problem, several significant advancements have brought us closer to this goal. For example, Fujitsu and Osaka University's Center for Quantum Information and Quantum Biology achieved a major milestone by demonstrating that a quantum computer could estimate material energy in just 10 hours—a task that would take a classical computer five years. This was achieved using only 60,000 qubits, fewer than previously believed necessary for fault-tolerant quantum computations. Additionally, companies like Kipu Quantum and Pasqal are advancing quantum optimization and scalability. Kipu Quantum tested the largest quantum optimization problem on a 156-qubit processor, while Pasqal successfully loaded over 1,000 neutral atoms in a single shot, a critical step toward scalable quantum processors and eventual quantum advantage.

Global Quantum Competition

The global quantum competition is intensifying as major powers race to dominate the field, seeking both strategic and economic advantages. The stakes are high: the first country to harness quantum computing at scale could potentially crack existing encryption methods, disrupt secure communications, and develop highly precise quantum sensors, reshaping military, cybersecurity, and economic landscapes. China and the US are the main players in this race, each with distinct strengths. China excels in quantum communication, having built the world's longest quantum key distribution network between Beijing and Shanghai and launched the Micius satellite for secure quantum communication. However the US still holds the advantage in quantum computing hardware and practical applications. Patent filings tell a similar story: China dominates in domestic quantum communication and sensing patents, while the US leads in quantum computing patents. Other nations are also making moves: Saudi Arabia is set to deploy its first quantum computer in collaboration with French company Pasqal, aiming to install a 200-qubit machine by 2025. Japan is investing heavily as well, with plans to develop a 10,000-qubit quantum computer in partnership with IBM, supported by a $31.7 million government investment. As the quantum landscape evolves, the stakes are high: The first country to achieve quantum breakthroughs will not only gain a military and strategic advantage but also stand to capture a market projected to be worth $1 trillion by 2035.

Hybrid Quantum-Classical

Hybrid quantum-classical computing is emerging as a solution to overcome the current limitations of quantum technology. Instead of replacing classical computing, quantum computers are being integrated with classical systems to solve complex problems more efficiently than either could on its own, and allowing for computations that were previously unachievable. This partnership aims to harness the strengths of both technologies to push beyond

QUANTUM COMPUTING

the computational limits of traditional systems. IBM calls this concept "quantum-centric supercomputing," and several leading supercomputing centers, such as Germany's Jupiter, Japan's Fugaku, and Poland's PSNC, are preparing to incorporate quantum-computing hardware into their systems. IBM's Heron system, with 156 qubits and 5,000 gates before errors occur, is an early step in this direction. By 2025, IBM plans to introduce its Flamingo system, which will connect seven quantum chips, expanding the total qubits to more than 1,000. Such advancements are geared toward optimizing quantum systems for parallel workloads alongside classical computing systems, making the best use of each technology's strengths. Hybrid quantum-classical systems are expected to play a critical role in advancing fields like drug discovery, materials science, and cryptography by running computations that exceed the capabilities of even the most powerful classical supercomputers.

Circuit Knitting

One of the key developments in the space of hybrid quantum-classical is a technique called "circuit knitting." This method divides a single quantum problem into multiple smaller quantum problems that can be processed in parallel by quantum processors. Classical computers then "knit" the results of these smaller quantum calculations together to produce the final solution. By incorporating classical and quantum processes in tandem, the system efficiently solves complex quantum circuits despite the limitations of current quantum hardware. This addresses the ongoing issue of qubit scarcity, which remains a significant barrier to the widespread use of quantum computers: Hybrid approaches like circuit knitting help by breaking down large quantum circuits into manageable subcircuits that fit onto smaller quantum devices, though at the cost of added simulation overhead. Circuit knitting also explores the use of classical communication to improve the efficiency of these local quantum computations.

Scaling the Qubits

Quantum computers, while promising, have not yet demonstrated practical advantages over classical supercomputers. The main issue is the limited number of qubits and their vulnerability to environmental noise, which easily disrupts their quantum state. This noise causes computational errors, forcing researchers to dedicate a large number of qubits solely to error correction. Consequently, a substantial increase in the number of qubits is needed before quantum computers can become truly useful for real-world applications. Leading tech companies are racing to achieve this goal. IBM, for instance, has announced plans to build a 100,000-qubit machine within the next 10 years. The vision is to combine the power of quantum systems with classical supercomputers to drive breakthroughs in areas such as drug discovery, fertilizer production, and battery performance. Similarly, Google has set an ambitious target of scaling up to a million qubits by the end of the decade. However, due to error-correction needs, only about 10,000 of those qubits would be available for actual computation. IonQ, a Maryland-based company, aims to achieve 1,024 "logical qubits" by 2028, with each logical qubit being constructed from 13 physical qubits for robust error correction.

It should be noted that focusing solely on the number of physical qubits can be misleading. The way qubits are constructed greatly influences their performance, particularly in terms of their resistance to noise and their operational stability. For this reason, companies often measure quantum performance through metrics such as "quantum volume" and "algorithmic qubits." These metrics provide a more accurate representation of how well a quantum computer can perform meaningful computations, considering both the quality and scale of the qubit system.

Quantum Error Mitigation

Quantum error mitigation refers to techniques that reduce the impact of errors in quantum computations without completely eliminating them. These methods are essential for current quantum devices, often

QUANTUM COMPUTING

called noisy intermediate-scale quantum devices, which lack the hardware needed for full error correction. The goal is to improve the accuracy of results despite the presence of noise; instead of relying on hardware changes, error mitigation is done at the software level, after a quantum circuit has been executed. IBM has demonstrated a practical approach to error mitigation: Its method involves running quantum programs and analyzing how noise affects the outputs. From this analysis, a noise model is created to simulate the system's imperfections, and then classical computing techniques are applied to estimate what the results would have been without the noise. This process allows for more accurate outputs from noisy quantum circuits. Error mitigation techniques like these are crucial for making near-term quantum devices more reliable, bridging the gap until full error correction becomes feasible.

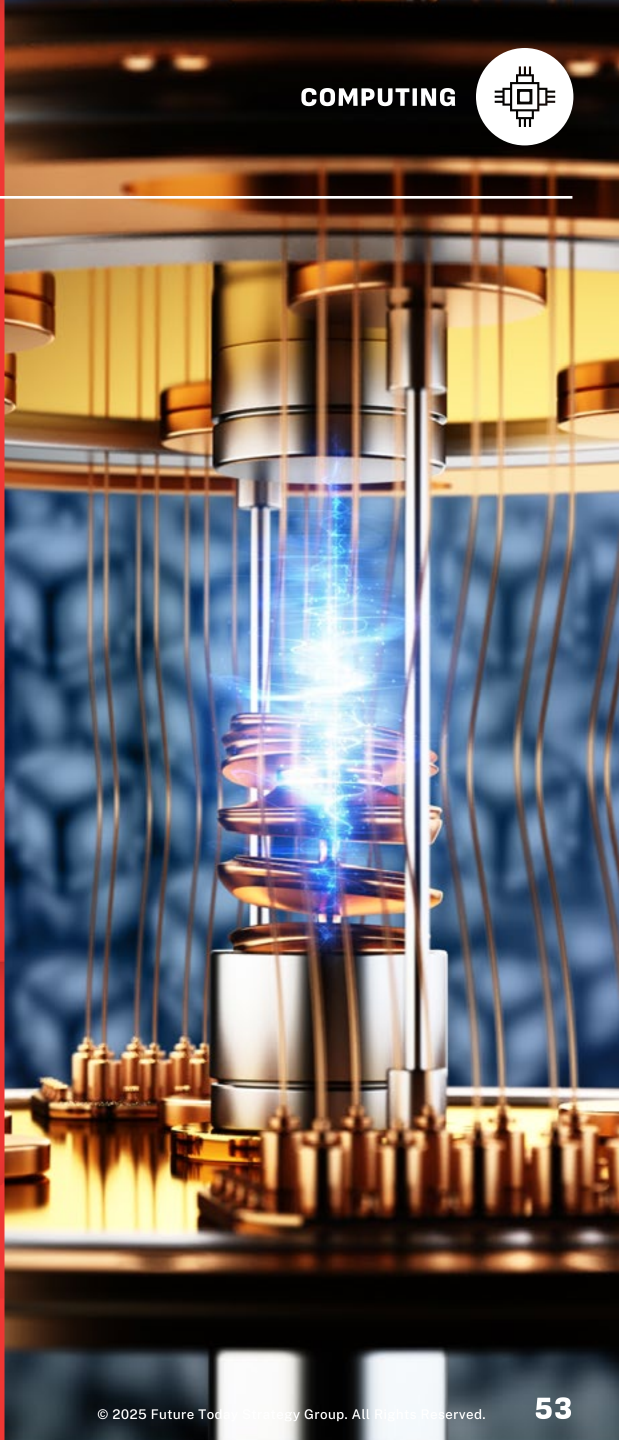
Quantum Error Correction

Unlike error mitigation, which only handles errors after the fact, quantum error correction (QEC) is a critical method that detects and corrects noise during the computation process. This allows quantum computers to run more reliably by fixing mistakes in real time. QEC is more versatile because it doesn't require a specific noise model to work, and it scales effectively even as quantum circuits become more complex. However, it comes with a significant cost—QEC requires many more qubits, additional connections between them, and more operations. For each qubit involved in a calculation, many more qubits are needed for error correction.

Google's Willow chip, announced in December 2024, reached a key milestone by achieving below-threshold error correction, a critical step toward scalable quantum computing. It demonstrated that error rates halve exponentially as more physical qubits are added, enabling faster error correction than error accumulation. This breakthrough paves the way for reliable, large-scale quantum systems capable of performing complex computations, with real-time error correction ensuring integrity throughout operations. Researchers elsewhere also made progress. At the Korea Institute of Science and Technology, they have developed world-class QEC technology with a fault-tolerant architecture that achieves a photon loss threshold of up to 14%, the highest currently known. Their method is also more resource-efficient compared to other techniques, using less compute power and energy to achieve the same level of photon consumption.

Quantum Noise Reduction

Quantum noise reduction aims to prevent errors in quantum computations by improving the underlying hardware and control systems. Unlike error correction, which addresses errors after they occur, noise reduction works at the hardware level to reduce the likelihood of errors arising in the first place. This approach includes techniques like dynamic decoupling and



QUANTUM COMPUTING

improved qubit designs, which enhance the stability of qubits and minimize error rates. By focusing on reducing the occurrence of errors, noise reduction can be combined with other error-handling methods, such as error correction or error mitigation, to further improve quantum computation accuracy.

Google Quantum AI has made significant progress in this area by developing a quantum memory system that operates with error rates below the critical threshold needed for effective quantum error correction. Their system's logical qubit outperformed its best physical qubit by more than double, highlighting the enhanced stability and reliability achieved through advanced error suppression techniques. This breakthrough demonstrates the potential for scaling up quantum computing systems while maintaining low error rates.

Researchers at the University of Trieste are also working on a quantum noise reduction method that involves linking the main quantum system to an extra, adjustable system. By carefully designing how these systems interact, they can create a special state that protects the quantum information, making it more stable and less likely to be affected by errors. This approach helps improve key quantum states, like NOON states, where particles, like photons, exist in a superposition of being entirely in one place or another at the same time. These states are key to making precise quantum measurements and advancing quantum computing, but they are fragile, so researchers are developing methods to protect them from errors and environmental noise. The technique could be applied using certain types of interactions between particles, and it shows the range of creative solutions being developed to reduce noise and make quantum systems more reliable.

Quantum Sensing

Quantum sensing leverages the unique properties of quantum systems, such as entanglement, quantum interference, and quantum state squeezing, to achieve highly precise measurements. This enables the detection of extremely small changes in physical quantities, like electric and magnetic fields, down to the atomic level. Such precision opens the door to a wide range of applications, and these quantum sensors are already being used to advance biomedical research. For example, wearable sensor helmets are being developed to record brain activity with unprecedented precision.

A significant advancement in quantum sensing comes from an international team at Forschungszentrum Jülich and the IBS Center for Quantum Nanoscience. They have developed a quantum sensor capable of detecting tiny magnetic fields at the atomic scale. This sensor, based on a single molecule, functions like an MRI for quantum materials, providing an unparalleled level of spatial resolution. Unlike traditional sensors that depend on defects in crystals to detect fields, this new tool uses a more direct approach, allowing scientists to study electric and magnetic properties with greater precision. In another advancement, physicists at Harvard have made progress with "spin squeezing," a quantum technique that enhances the precision of quantum sensors by reducing uncertainty in measurements. Traditionally, spin squeezing required complex interactions between many atoms, but the Harvard team discovered a method to achieve this using natural magnetism, such as that found in everyday fridge magnets. This could lead to more portable and practical quantum sensors, with potential applications in biomedical imaging and atomic clocks, further advancing the field of quantum sensing.

Encryption Breaking Quantum

Recent developments in quantum computing have sparked concern over the potential impact on global cybersecurity. Chinese researchers recently claimed to have used D-Wave's quantum annealing systems to crack widely used encryption methods like RSA, sparking headlines about the end of modern cryptography. Published in the Chinese Journal of Computers, the paper outlined their approach to breaking

QUANTUM COMPUTING

encryption, suggesting quantum computers could accelerate attacks on symmetric encryption systems as well. However, these claims were significantly overhyped. While there have been advancements in using quantum computing for cryptographic attacks, this progress is incremental rather than revolutionary and doesn't yet pose an immediate, large-scale threat to current encryption systems. Still, the potential consequences of a true breakthrough in quantum code-breaking would be huge, making even these incremental advancements noteworthy.

Meanwhile, researchers are continuing to improve quantum algorithms with the goal of making quantum code-breaking more feasible. At MIT, a team has proposed an optimized version of Shor's algorithm, which is specifically designed for breaking encryption. Introduced in 1994, the algorithm demonstrated that quantum computers could, in theory, break RSA encryption far faster than classical computers. However, the challenge remains in building quantum computers large and stable enough to run this algorithm effectively. MIT's new approach improves upon previous designs by integrating a faster algorithm from NYU with more efficient memory use. This makes the quantum circuit smaller and less noisy, bringing the practical application of quantum computers for encryption-breaking one step closer. While this improvement is promising, it will still take time to determine whether these advancements can truly challenge today's encryption methods.

Post-Quantum Cryptography

While white hat teams attempt to break encryption, other researchers are working on developing quantum-safe cryptography to replace the widely-used encryption methods that secure much of today's digital information. Traditional encryption, such as RSA or ECC (Elliptic Curve Cryptography), relies on the difficulty of factoring large numbers or solving discrete logarithms—problems that quantum computers could solve exponentially faster using algorithms like Shor's. This means that sensitive data, from financial transactions to government communications, could become vulnerable to attacks. In August 2024, the US National Institute of Standards and Technology approved three post-quantum cryptography algorithms for mainstream development. These algorithms are designed to secure a broad range of data, from private emails to online transactions, against the potential threat posed by quantum computers, which could break current encryption methods within the next decade. IBM, which developed two of the approved algorithms, is now integrating these solutions into its products, including IBM z16 and IBM Cloud, as part of its broader Quantum Safe roadmap. IBM has also introduced Quantum Safe Transformation Services to assist clients in transitioning to quantum-safe technologies. In addition, the company has developed the Cryptography Bill of Materials, a standard for managing cryptographic assets across software and systems, which aims to enhance the security of digital infrastructure.

While quantum computers are not yet capable of breaking encryption, the data being encrypted today could be harvested and stored by adversaries, only to be decrypted once quantum technology matures. This threat is known as "harvest now, decrypt later." Without post-quantum cryptography, future advances in quantum computing could expose confidential information retroactively.

The Quantum Internet

The quantum internet is a proposed network that would link quantum devices to exchange quantum information. While a global quantum internet is still theoretical, significant progress is being made toward its development. In 2024, physicists at Harvard University demonstrated the longest quantum network using existing fiber-optic cables in the Boston area, successfully linking two quantum memory nodes 22 miles apart. The technology overcomes a significant issue in quantum communication: Traditional materials for boosting signal strength over long distances don't

QUANTUM COMPUTING

work with quantum information. Instead, the team used specialized diamond nodes to capture, store, and correct quantum data as it was transmitted. Meanwhile, a research team in China has advanced quantum internet by creating a multi-node quantum network across a metropolitan area. Previously, quantum networks were limited to two-node setups in lab environments, which are far from practical for large-scale quantum communication. By establishing entanglement across three nodes in a metropolitan area, the research team demonstrated that it's possible to expand quantum networks to more nodes and greater distances, which is crucial for building a functional quantum internet.

A major challenge in developing a quantum internet is transmitting quantum information over long distances while maintaining entanglement. Researchers from the University of California, Santa Barbara addressed this by improving photon emitters, which carry quantum data. Photons are ideal for this task because they interact weakly with their environment, but generating them efficiently at telecom wavelengths for fiber-optic networks has been difficult. The team found that atomic vibrations in materials reduce photon emission efficiency at these wavelengths. By optimizing materials and engineering vibrational properties, they improved emitter performance.

Quantum Software

Quantum software is a critical component of the quantum computing ecosystem, enabling the development, testing, and execution of quantum algorithms on quantum hardware and simulators. It includes quantum algorithms—like Shor's algorithm for factoring and Grover's algorithm for search tasks—which leverage quantum computing's unique properties like superposition and entanglement. Quantum programming languages like IBM's Qiskit, Google's Cirq, and Microsoft's Q# are used to develop these algorithms and run them on quantum processors. One notable application of quantum software is its use in biotechnology. Researchers at Moderna and IBM used Qiskit to predict mRNA secondary structures—a task crucial for designing RNA-based therapies. Additionally, quantum simulators and emulators are important tools for testing quantum algorithms in classical environments. Simulators like IBM's Matrix-Product State provide a quantum-like environment for testing algorithms without needing direct access to quantum hardware. These tools are essential for refining and optimizing quantum algorithms before they are executed on real quantum devices. Quantum software, though in its early stages, is rapidly advancing fields like cryptography, optimization, and drug discovery, and will play a central role in the broader adoption and application of quantum computing.

Quantum Microprocessor Chips

Researchers at Hong Kong Polytechnic University have developed a quantum microprocessor chip for molecular spectroscopy simulations, especially for large, complex molecules. Classical computers

> **This year, quantum computing will reach a pivotal moment as error correction breakthroughs and scalable hardware architectures converge to enable the first commercially viable systems that could revolutionize fields from drug discovery to financial modeling.**
>
> Amy Webb
> CEO • Future Today Strategy Group

QUANTUM COMPUTING

struggle with the computational demands of these tasks, but quantum computers excel by using quantum phenomena like superposition and entanglement. This chip, featuring a 16-qubit system, integrates both hardware and software, including a photonic network and quantum light sources, enabling more accurate simulations. Applications of this technology include drug discovery, materials design, and advanced quantum chemistry, where classical methods fall short.

Meanwhile, Oxford Ionics, a spinoff from the University of Oxford, has also made remarkable progress in the quantum chip field. Its new high-performance quantum chip has set new records for quantum computing without relying on error correction. This development is particularly significant as the chip can be manufactured using existing semiconductor fabrication processes, making it more scalable and commercially viable. Oxford Ionics anticipates that a fully functional and practical quantum computer could be available within the next three years. These advancements in quantum microprocessor chips, both for molecular simulations and broader applications, are accelerating the timeline for real-world quantum computing use, paving the way for breakthroughs in numerous industries.

Quantum Cooling

Quantum cooling is the process of reducing temperatures to near absolute zero to enable quantum mechanical systems, such as quantum computers, to operate effectively. Achieving these extremely low temperatures is essential because quantum computing components, like qubits, are highly sensitive to thermal noise, which disrupts their delicate quantum states. Without sufficient cooling, qubits cannot maintain coherence, making reliable quantum computation nearly impossible.

To address this challenge, engineers have developed a device that converts heat into electrical voltage at temperatures colder than outer space. This device advances thermoelectric technology and enables more efficient cooling for quantum systems. It uses the Nernst effect, where a magnetic field applied perpendicular to a temperature gradient generates voltage. The device's two-dimensional design allows researchers to control efficiency electrically, making it more adaptable and effective for quantum applications. Additionally, NIST has achieved a breakthrough in refining pulse tube refrigerators (PTRs), essential tools for reaching the ultra-low temperatures required in quantum research. PTRs, a staple for more than 40 years, are known for their high energy demands and operational costs. NIST's innovative redesign incorporates a specialized valve that contracts as temperatures decrease, significantly reducing helium waste and enhancing efficiency. This advancement allows PTRs to reach near-absolute zero temperatures up to 3.5 times faster than previous models, cutting both setup time and energy use. By making low-temperature environments more readily available and cost-effective, this improvement could accelerate experimental progress, reducing barriers to further advancements in quantum computing.

Quantum Machine Learning

Quantum machine learning (QML) combines quantum computing with machine learning to speed up and enhance algorithm performance using principles of quantum mechanics. By applying quantum properties like superposition and entanglement, QML aims to improve traditional machine learning techniques, particularly for large-scale data processing and complex pattern analysis. It has promising applications, including optimizing complex algorithms, speeding up classification tasks, enabling high-dimensional data analysis, improving feature selection and dimensionality reduction, generating realistic data models, and uncovering intricate patterns in data. QML also shows potential for advancing solutions in quantum chemistry problems.

Researchers are developing fully quantum algorithms that depend on large-scale quantum computations to manage complex machine learning tasks. These algorithms

QUANTUM COMPUTING

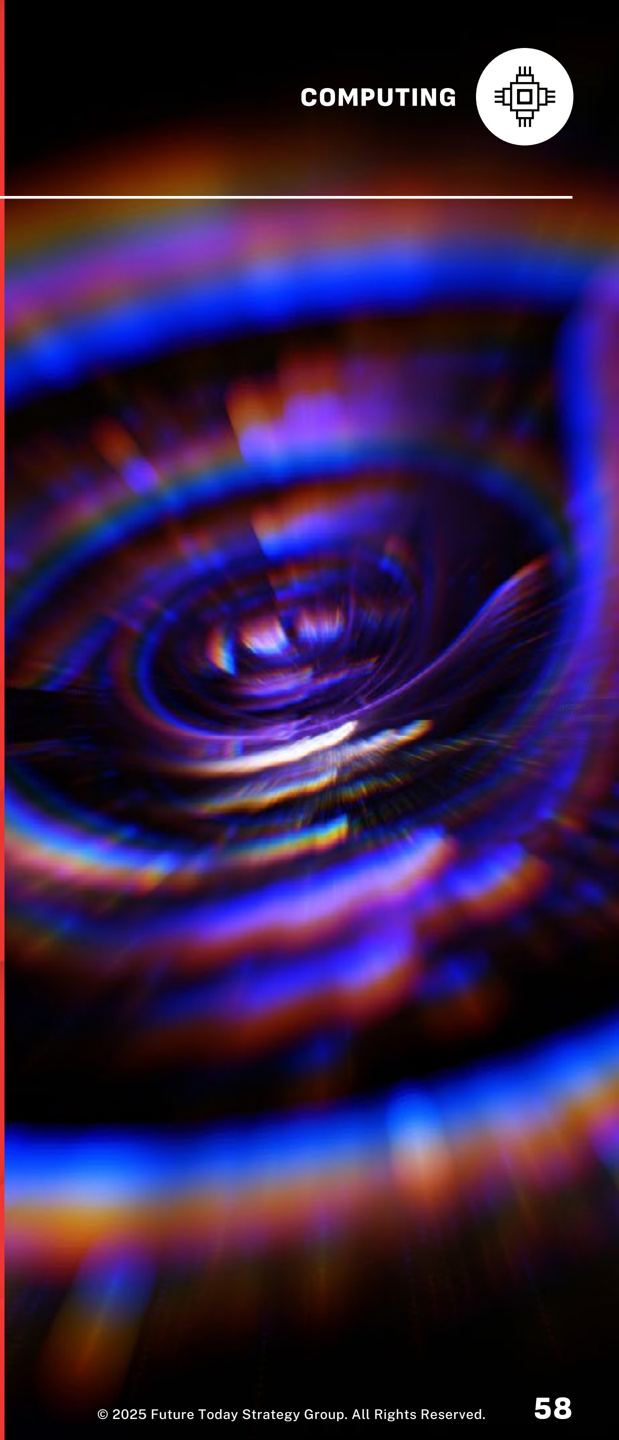
could offer exponential speed improvements over classical methods, especially for large datasets and intricate models. Key advancements include quantum neural networks, which are inspired by classical neural networks but operate with quantum mechanics for faster training and inference, and quantum support vector machines, which use quantum kernels to better separate complex data patterns. As QML evolves, it could lead to major advancements in data processing and analysis across various scientific and technological fields.

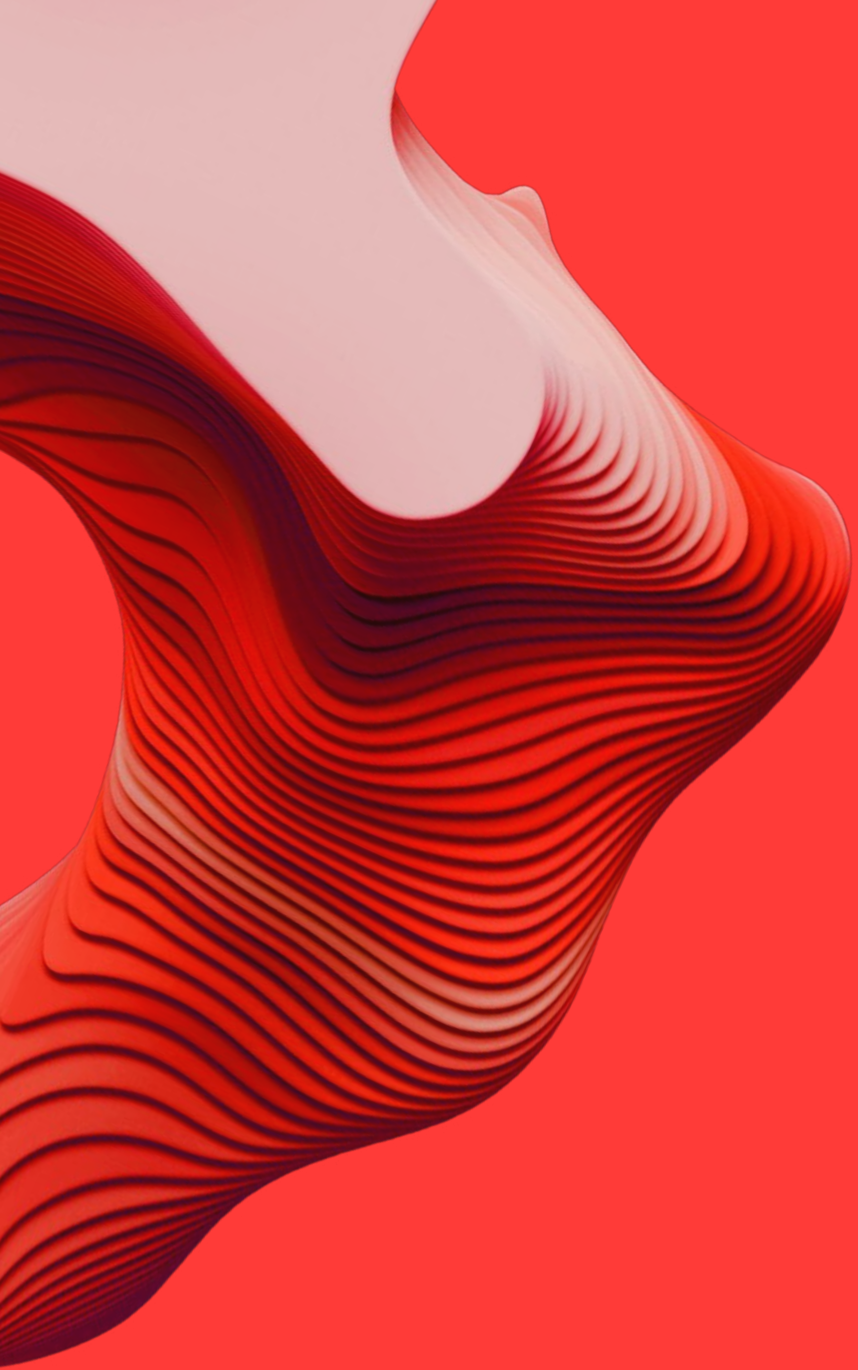
Spintronics

Traditional electronics rely on electron movement (charge) alone to store and process information. But in the emerging field of spintronics, short for spin transport electronics, scientists harness the quantum properties of an electron's charge and spin, offering a pathway to reduce energy consumption and increase processing speed. Recent advances at BESSY II, a joint German-Spanish research facility, demonstrated enhanced quantum effects in layered structures of graphene, cobalt, and iridium that stabilize "spin textures"—organized electron spin patterns essential for spintronic functionality.

A groundbreaking development in late 2024 at the Fraunhofer Institute for Applied Solid State Physics demonstrated spin-based computing using diamond quantum systems. The SPINNING project created qubit registers using color centers in diamond's crystal lattice, achieving entanglement between two six-qubit registers across 20 meters. This achievement delivered impressive results with coherence times over 10 milliseconds and error rates below 0.5%, marking a significant advance in quantum computing using electron spins.

The potential of spintronics is particularly relevant for applications requiring high processing power with minimal energy loss, such as data centers, artificial intelligence, and portable electronics. Spintronics could improve these areas by reducing battery drain and energy costs, a major advantage over traditional semiconductor-based electronics. The University of Utah recently achieved a breakthrough by turning ordinary LEDs into spintronic devices at room temperature, eliminating the need for magnetic fields. This demonstrates how existing electronics can be converted into spintronic devices, paving the way for energy-efficient, high-speed electronics.





NETWORKING

NETWORKING

6G

In 2018, 5G began rolling out. By 2022 and 2023, it expanded globally, albeit more slowly than anticipated. Now in 2025, 6G is in early development with potential market introduction in the early 2030s. Featuring extremely low latency, 6G aims to enable deeply immersive experiences across various applications. A major initiative is Nvidia's 6G Research Cloud, combining generative AI and Omniverse tools to simulate and test next-generation wireless networks. This platform incorporates Nvidia's neural radio framework, enhanced radio access networks, and digital twins, helping developers experiment with and define 6G features. Research is pushing 6G forward, notably in terahertz communications, which allow transmission rates potentially reaching terabits per second. For instance, a University of Adelaide team has developed a polarization multiplexer, doubling communication capacity while minimizing data loss and enabling cost-effective mass production. In parallel, researchers at the University of Glasgow have introduced an advanced antenna leveraging metamaterials; it's designed for high-speed performance in the 60 GHz mmWave band, a critical frequency for 6G's envisioned ultra-fast data transfer and robust connectivity. This research will not only lead to higher-speed, more reliable communication but also supports new applications for 6G in areas like sensing and imaging.

AI-RAN

AI-RAN, short for AI-driven Radio Access Networks, integrates artificial intelligence directly into mobile network infrastructure to optimize performance, enhance reliability, and reduce power consumption. By embedding AI algorithms across core functions—from the physical network layer to resource management—AI-RAN facilitates the development of AI-native networks that are self-organizing, self-optimizing, and self-managing. This intelligent infrastructure can dynamically adjust network parameters, boosting efficiency while supporting high-performance, low-latency applications. Launched in February 2024, the AI-RAN Alliance includes T-Mobile, Nvidia, Ericsson, Nokia, SoftBank, AWS, Arm, DeepSig, Microsoft, and Samsung Electronics, among others, aiming to unify advancements in AI-RAN technologies. SoftBank, for instance, has collaborated with Nvidia to establish a research lab in Santa Clara, California, focused on AI-RAN's potential at the network edge, also known as multi-access edge computing (MEC). By deploying AI processing near devices, MEC can handle AI tasks with minimal latency, a capability essential for real-time applications, remote operations, and highly interactive experiences. This collaborative approach among telecoms, cloud providers, and AI companies under the AI-RAN Alliance represents a significant step forward in transforming mobile networks, enabling them to intelligently adapt and evolve with minimal manual intervention.

NETWORKING

Photonic-Enhanced Wireless Communication

Photonic-enhanced wireless communication combines photonics (light-based technology) with traditional high-frequency radio waves to boost wireless data transmission. Traditional wireless networks are often limited by bandwidth and face congestion, especially as data demands increase. By using millimeter-wave photonics alongside high-speed electronics, photonic-enhanced wireless expands the available bandwidth, enabling data transfer across a much broader range of frequencies, typically from 5 GHz up to 150 GHz. This wider frequency range means data can move faster, and at much higher capacities, by accessing less crowded, higher-frequency channels. Researchers at University College London recently set a record by transmitting data at 938 gigabits per second across this wide spectrum, marking a speed over 9,000 times faster than typical 5G download rates. This breakthrough leverages the combined strengths of radio and light-based technologies, supporting high-speed data flow ideal for applications in future high-speed communications, mobile networks, and more. This approach may be commercially viable within three to five years, potentially transforming how networks handle data-intensive applications.

AI at the Edge

In the ideal state of edge AI, intelligent devices like wearables, smart homes, and industrial sensors would process data independently, right at the point of data collection. This would allow them to respond instantly to changes, keep user data private, and operate continuously without needing cloud connectivity. Imagine a wearable health monitor detecting an irregular heartbeat in real time or a smart home adjusting lighting based on activity—each action would be immediate, efficient, and private. However, reaching this ideal state presents challenges. Artificial neural networks (ANNs), which power much of AI, are computationally intensive, requiring significant power and processing capabilities. IoT devices, inherently small and low-power, struggle to meet these demands; they often lack the processing speed, memory, and battery life to run complex algorithms. As a result, many edge devices still rely on cloud servers to handle intensive tasks, a habit that creates latency, consumes more power, and can compromise data privacy.

A recent breakthrough by researchers at Tokyo University of Science could help overcome these limitations: a new training algorithm for a specific type of ANN, called a binarized neural network (BNN), designed to run efficiently on IoT devices. By implementing this algorithm within a computing-in-memory architecture, they have enabled devices to perform AI tasks more effectively with far less power. This advance could unlock a future where wearables, smart homes, and other IoT devices operate with full AI capabilities at the edge, minimizing the need for cloud dependency, reducing energy consumption, and improving response times—bringing us closer to the ideal state of edge AI.

Satellite Internet

Starlink has revolutionized satellite internet by making high-speed, low-latency internet accessible in remote and underserved regions worldwide. Unlike traditional satellite providers that use geostationary satellites orbiting 22,000 miles above Earth, Starlink operates a constellation of low-Earth orbit (LEO) satellites positioned about 350 miles up. This proximity reduces latency and dramatically improves connection speeds, making satellite internet viable for applications previously limited to terrestrial networks. At the time of this writing, SpaceX has deployed more than 5,000 Starlink satellites, with plans to expand this constellation by thousands more. This network, now available in more than 50 countries, has had a profound impact on areas lacking traditional infrastructure, offering new possibilities for education, health care, and economic development. With a special focus on expanding service in Africa, Starlink is reshaping how and where people can access the internet without requiring government-funded infrastruc-

NETWORKING

ture projects. SpaceX has even introduced a more compact version of its equipment, the Starlink Mini, which makes the service more accessible and easier to deploy. While Starlink leads the way, some competition is emerging. Amazon's Project Kuiper aims to launch more than 3,000 LEO satellites to establish its own global internet network. Similarly, China is developing the GuoWang constellation to create a self-reliant satellite internet option and compete internationally. LEO satellite constellations are disruptive because they eliminate the need for traditional infrastructure. Unlike cable and fiber providers, which rely on extensive physical networks, Starlink's LEO satellites deliver internet directly from space, reaching remote and underserved areas. This approach is especially valuable in regions where building infrastructure is prohibitively expensive or logistically difficult.

Connected PCs

Why do we still rely on Wi-Fi to connect our PCs to the internet, or turn to mobile hotspots when Wi-Fi isn't available? Apple is aiming to change that. In early 2025, the company plans to debut its custom-built 5G modem chip in devices like the iPhone SE, a budget iPad, and the iPhone 17 "Air." This rollout will serve as a testing ground before integrating the chip into flagship products over the next three years, potentially including Macs. A Mac equipped with a 5G chip would be able to connect directly to cellular networks, eliminating the need for Wi-Fi or mobile hotspots. Some PCs, like the Lenovo ThinkPad X1 Nano, HP Elite Dragonfly G3, Dell Latitude 9430, and Microsoft Surface Pro X, already support cellular connectivity. This technology offers clear advantages: internet access in most areas with cellular coverage and greater security, as cellular networks are typically safer than public Wi-Fi. However, the rise of satellite internet raises questions about the future of cellular connectivity. If satellite-based internet becomes ubiquitous and reliable, will cellular capabilities still be necessary for connecting our devices?

SCENARIO YEAR 2032

JOHN DEERE, THE TELECOM PROVIDER

By 2032, John Deere is the digital lifeline of rural America. What began as a push for connected farm equipment in the 2020s has transformed Deere from an agricultural machinery giant into a powerhouse of rural telecom and digital infrastructure. Early partnerships with SpaceX's Starlink laid the groundwork, as Deere brought satellite internet to remote farms, enabling game-changing tools like autonomous tractors, ExactShot precision planting, and See and Spray weed detection. These innovations let farmers run data-driven operations even where cellular networks fell short, and soon, Deere realized its future was far beyond tractors.

Today, John Deere isn't just a farm machinery brand—it's a proprietary satellite and 6G network delivering high-speed, low-latency internet across rural landscapes. DeereLink powers Farm Hubs, localized small data centers that turn farming into a fully digital service. These small hubs analyze environmental factors, store rich farming insights, and manage regional agricultural data—close to the farm itself—supporting Deere's farming-as-a-service (FaaS) packages. Farmers can subscribe to modular, autonomous machines that handle planting, watering, and harvesting, all guided by DeereLink's intelligence network.

But Deere's impact reaches far beyond the fields. The Rural Grid extends its connectivity to entire communities, linking schools, health care centers, and municipal offices. Farm Hubs double as community compute centers, providing internet access, immersive remote education, and AI-driven telehealth, making it possible for rural families to access resources once reserved for urban areas.

By 2032, John Deere has redefined its mission. No longer a manufacturer of heavy machinery, it has become the invisible backbone of rural connectivity—a digital infrastructure titan linking America's countryside to cutting-edge computational power and data insights, ensuring rural regions are as connected and empowered as any metropolis.

We extend our gratitude to the NYU Stern MBA students from the Fall 2024 Strategic Foresight class for their exceptional contributions in crafting this scenario.

AUTHORS & CONTRIBUTORS

Sam Jordan

Technology & Computing Lead

Sam Jordan is a Senior Manager and the Technology and Computing Lead at FTSG. Her research focuses on the future of computing, spanning large-scale systems, personal devices, AI, and telecommunications. She also covers the space industry, analyzing advancements in satellite technology, communications infrastructure, and emerging aerospace innovations. She has worked with some of the world's largest technology companies to advance human-computer interaction, develop AI strategies, and drive innovation in device evolution.

Before joining FTSG, Sam was the CEO and co-founder of TrovBase, a secure platform for data discovery and analysis sharing. She also worked at IBM, where she helped large enterprises modernize their IT infrastructure, specializing in mainframes and integrating modern software and methodologies into legacy systems.

Sam currently serves as a coach in the Strategic Foresight MBA Course at NYU Stern School of Business and is an Emergent Ventures Fellow at the Mercatus Center. She holds a B.S. in Economics and Data Analysis from George Mason University and an MBA from NYU's Stern School of Business.

Chief Executive Officer
Amy Webb

Managing Director
Melanie Subin

Director of Marketing & Comms.
Victoria Chaitoff

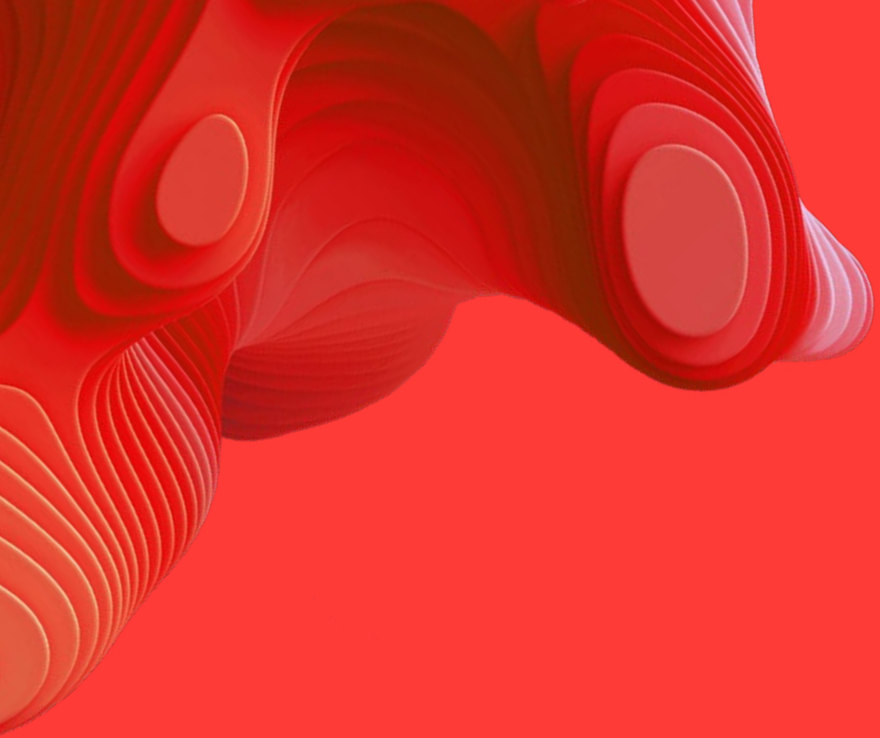
Creative Director
Emily Caufield

Editor
Erica Peterson

Copy Editor
Sarah Johnson

Andrew Hornstra
AI Solutions Architect,
Contributor

FTSG



SELECTED SOURCES

"2024 Is the Year for Brain-Computer Interfaces." Technology Networks, July 11, 2024. https://www.technologynetworks.com/neuroscience/blog/2024-is-the-year-for-brain-computer-interfaces-388563.

Abbas, Aumber, et al. "Flexible and Stretchable Electronics for Wearable Health Monitoring: Recent Advances and Future Challenges." Nano Today 45 (2024): 101111. https://doi.org/10.1016/j.nantod.2024.101111.

Aikawa, Akira. "Japan's Riken Plans Quantum Link to Supercomputer Fugaku." Nikkei Asia, May 20, 2024. https://asia.nikkei.com/Business/Technology/Japan-s-Riken-plans-quantum-link-to-supercomputer-Fugaku.

Albert, Kayla. "Inspired by the Human Body, Engineer Designs Chips That Could Make Wearable AI More Energy Efficient." Purdue University Newsroom, August 20, 2024. https://www.purdue.edu/newsroom/2024/Q3/inspired-by-the-human-body-engineer-designs-chips-that-could-make-wearable-ai-more-energy-efficient/.

Alevras, Dimitris, et al. "mRNA Secondary Structure Prediction Using Utility-Scale Quantum Computers." ArXiv, May 30, 2024. https://arxiv.org/abs/2405.20328.

Amos, Zac. "'Harvest Now, Decrypt Later': Why Hackers Are Waiting for Quantum Computing." VentureBeat, September 21, 2024. https://venturebeat.com/security/harvest-now-decrypt-later-why-hackers-are-waiting-for-quantum-computing/.

Anselmi, Fabio, et al. "Inverse Design of Ancillary System for Quantum Noise Cancellation." ArXiv, August 8, 2024. https://arxiv.org/abs/2408.04418.

"Aramco Partners with Pasqal to Deploy the First Quantum Computer in Saudi Arabia." Energy Connects, May 20, 2024. https://www.energyconnects.com/news/technology/2024/may/aramco-partners-with-pasqual-to-deploy-the-first-quantum-computer-in-saudi-arabia/.

"ASML Risks Losing Chinese Market Permanently If It Complies with US Restrictions." Global Times, September 1, 2024. www.globaltimes.cn/page/202409/1319035.shtml.

Banerjee, Shounak. "How Software Built Nvidia's $2.97T Hardware Empire." Command AI, July 26, 2024. www.command.ai/blog/nvidia-product-case-study/.

Barrowclough, Nicholas. "Transforming Data Centre Cooling for a Sustainable Future." Data Centre Magazine, November 16, 2023. https://datacentremagazine.com/articles/transforming-data-centre-cooling-for-a-sustainable-future.

Block, Maxwell, et al. "Scalable Spin Squeezing from Finite-Temperature Easy-Plane Magnetism." Nature Physics 20, no. 3 (2024): 319–323, https://doi.org/10.1038/s41567-024-02562-5.

Burns, Joe. "For Data Center Cooling Project, Advanced Cooling Technologies Gets $1.1M DOE Grant." Utility Dive, October 7, 2024. https://www.utilitydive.com/news/department-of-energy-advanced-data-center-computing-cooling-energy-funding-grant/729049/.

Butts, Dylan. "Malaysia Emerges as Asian Data Center Powerhouse Amid Booming Demand." CNBC, June 17, 2024. https://www.cnbc.com/2024/06/17/malaysia-emerges-as-asian-data-center-powerhouse-amid-booming-demand.html.

Butts, Dylan. "U.S. Imposes New Export Controls on China over Quantum Computing and AI." CNBC, September 6, 2024. www.cnbc.com/2024/09/06/us-china-quantum-chip-related-export-controls.html.

Cai, Zhenyu, et al. "Engineered Living Materials: Prospects and Challenges for Using Biological Systems to Direct the Assembly of Smart Materials." Reviews of Modern Physics 95, no. 4 (2023): 045005. https://doi.org/10.1103/RevModPhys.95.045005.

Carrier, Frédérique. "The Chip Industry's Reshoring Revolution." RBC Wealth Management, November 14, 2023. www.rbcwealthmanagement.com/en-us/insights/the-chip-industrys-reshoring-revolution.

Chen, Sophia. "Google Says It's Made a Quantum Computing Breakthrough That Reduces Errors." MIT Technology Review, September 11, 2024. https://www.technologyreview.com/2024/09/11/1103828/google-says-its-made-a-quantum-computing-breakthrough-that-reduces-errors/.

Chernicoff, David. "Brazil's Scala Data Centers Secures LATAM Dominance, Plans to Develop 'AI City.'" Data Center Frontier, September 19, 2024. https://www.datacenterfrontier.com/site-selection/article/55140968/brazils-scala-data-centers-secures-latam-dominance-plans-to-develop-ai-city.

Chiang, Helen. "Chips and Challenges: Southeast Asia and India's Semiconductor Manufacturing Crossroads." IDC Blog, September 30, 2024. blogs.idc.com/2024/09/30/chips-and-challenges-southeast-asia-and-indias-semiconductor-manufacturing-crossroads/.

Chiang, Sheila. "Intel Unveils New AI Chips as It Seeks to Reclaim Market Share from Nvidia and AMD." CNBC, June 4, 2024. www.cnbc.com/2024/06/04/intel-unveils-new-ai-chips-as-it-seeks-to-take-on-nvidia-and-amd.html.

Choucair, Cierra. “KIST Develops Hybrid Quantum Error Correction, Merging the Power of DV and CV Qubits.” The Quantum Insider, October 16, 2024. https://thequantuminsider.com/2024/10/16/kist-develops-hybrid-quantum-error-correction-merging-the-power-of-dv-and-cv-qubits/.

Cohen, Ariel. “China’s Massive Barrage in the Chip Battle.” Forbes, May 31, 2024. www.forbes.com/sites/arielcohen/2024/05/31/chinas-massive-barrage-in-the-chip-battle/.

“COOLERCHIPS.” ARPA-E, U.S. Department of Energy, September 22, 2022. https://arpa-e.energy.gov/technologies/programs/coolerchips.

“CoolIT Systems Joins Nvidia Partner Network, Unveils AI-Centric Liquid-Cooling Solutions.” HPCwire, October 15, 2024. https://www.hpcwire.com/off-the-wire/coolit-systems-joins-nvidia-partner-network-unveils-ai-centric-liquid-cooling-solutions/.

Dai, Yahao, et al. “Soft Hydrogel Semiconductors with Augmented Electrochemical Performance.” Science 386, no. 4321 (October 25, 2024): 431–439. https://www.science.org/doi/10.1126/science.adp9314.

David, Emilia. “Intel Plans to Be Inside 100 Million AI PCs by Next Year.” The Verge, February 27, 2024. https://www.theverge.com/2024/2/27/24084792/intel-plans-to-be-inside-100-million-ai-pcs-by-next-year.

Ding, Wenxin, et al. “Anomalous Hall Effect and Quantum Criticality in Geometrically Frustrated Heavy Fermion Metals.” Physical Review Letters 133, no. 10 (September 6, 2024). link.aps.org/doi/10.1103/PhysRevLett.133.106504.

Emerson, Sarah. “How Elon Musk Muzzled Government Employees from Talking about xAI’s New Supercomputer.” Forbes, October 11, 2024. https://www.forbes.com/sites/sarahemerson/2024/10/11/how-elon-musk-muzzled-government-employees-from-talking-about-xais-new-supercomputer/.

“Fact Sheet: President Biden to Highlight $3.3 Billion Investment in Racine, Wisconsin, and How His Investing in America Agenda Is Driving Economic Comebacks in Communities Across the Country.” The White House, May 8, 2024. https://www.whitehouse.gov/briefing-room/statements-releases/2024/05/08/fact-sheet-president-biden-to-highlight-3-3-billion-investment-in-racine-wisconsin-and-how-his-investing-in-america-agenda-is-driving-economic-comebacks-in-communities-across-the-country/.

“Fact Sheet: Two Years After the CHIPS and Science Act, Biden-Harris Administration Celebrates Historic Achievements in Bringing Semiconductor Supply Chains Home, Creating Jobs, Supporting Innovation, and Protecting National Security.” The White House, August 9, 2024. www.whitehouse.gov/briefing-room/statements-releases/2024/08/09/fact-sheet-two-years-after-the-chips-and-science-act-biden.

Fist, Tim, and Arnab Datta. “How to Build the Future of AI in the United States.” Institute for Progress, October 23, 2024. https://ifp.org/future-of-ai-compute/.

Fitch, Asa. “CoreWeave Raises $7.5 Billion in Debt for AI Computing Push.” The Wall Street Journal, May 17, 2024. www.wsj.com/tech/ai/coreweave-raises-7-5-billion-in-debt-for-ai-computing-push-99fd2241.

Fraunhofer Institute for Applied Solid State Physics. “Major Development Successes in Diamond Spin Photon Quantum Computers.” ScienceDaily, October 28, 2024. www.sciencedaily.com/releases/2024/10/241028132358.htm.

Fujiwara, Yuya, and Takayuki Kawahara. “TGBNN: Training Algorithm of Binarized Neural Network with Ternary Gradients for MRAM-Based Computing-in-Memory Architecture.” IEEE Access 12, October 8, 2024: 150962–150974. https://ieeexplore.ieee.org/document/10707281.

Gambetta, Jay. “IBM’s Big Bet on the Quantum-Centric Supercomputer.” IEEE Spectrum, August 21, 2024. https://spectrum.ieee.org/ibm-quantum-computer-2668978269.

Gardizy, Anissa, and Amir Efrati. “Microsoft and OpenAI Plot $100 Billion Stargate AI Supercomputer.” The Information, March 29, 2024. www.theinformation.com/articles/microsoft-and-openai-plot-100-billion-stargate-ai-supercomputer.

Genkina, Dina. “Brain-Inspired Computer Approaches Brain-Like Size.” IEEE Spectrum, May 8, 2024. https://spectrum.ieee.org/neuromorphic-computing-spinnaker2.

Gentinetta, Gian, et al. “Overhead-Constrained Circuit Knitting for Variational Quantum Dynamics.” Quantum 8 (March 21, 2024): 1296. https://doi.org/10.22331/q-2024-03-21-1296.

Greengard, Samuel. “AI Reinvents Chip Design.” Communications of the ACM, August 22, 2024. https://cacm.acm.org/news/ai-reinvents-chip-design/.

Gudge, Ethan. “‘Pivotal’ Quantum Computing Chip Unveiled.” BBC News, July 16, 2024. https://www.bbc.com/news/articles/cgxq789931go.

Hannigan, Brett C., et al. “Distributed Sensing Along Fibers for Smart Clothing.” Science Advances 10, no. 12 (March 20, 2024). https://doi.org/10.1126/sciadv.adj9708.

Hautzinger, Matthew P., et al. “Room-Temperature Spin Injection Across a Chiral Perovskite/III–V Interface.” Nature 631 (2024): 307–312. https://doi.org/10.1038/s41586-024-07560-4.

Heath, Alex. “Hands-on with Orion, Meta’s First Pair of AR Glasses.” The Verge, September 25, 2024. https://www.theverge.com/24253908/meta-orion-ar-glasses-demo-mark-zuckerberg-interview.

“IBM-Developed Algorithms Announced as NIST’s First Published Post-Quantum Cryptography Standards.” IBM Newsroom, August 13, 2024. https://newsroom.ibm.com/2024-08-13-ibm-developed-algorithms-announced-as-worlds-first-post-quantum-cryptography-standards.

Intel Corp. “Intel Unleashes Enterprise AI with Gaudi 3, AI Open Systems Strategy and New Customer Wins.” Intel Newsroom, April 9, 2024. www.intel.com/content/www/us/en/newsroom/news/vision-2024-enterprise-ai-gaudi-3-open-systems-strategy.html.

Jabbar, A., et al. “60 GHz Programmable Dynamic Metasurface Antenna (DMA) for Next-Generation Communication, Sensing, and Imaging Applications: From Concept to Prototype.” IEEE Open Journal of Antennas and Propagation 5, no. 3 (June 2024): 705–726. https://doi.org/10.1109/OJAP.2024.3386452.

Jacobi, Christian, and Elpida Tzortzatos. “New Telum II Processor and IBM Spyre Accelerator: Expanding AI on IBM Z.” IBM Blog, August 26, 2024. www.ibm.com/blog/announcement/telum-ii/.

Juliussen, Egil. “Nvidia GTC 2024: Why Nvidia Dominates AI.” EE Times Europe, June 10, 2024. www.eetimes.eu/nvidia-gtc-2024-why-nvidia-dominates-ai/.

Kim, Youngseok, et al. “Scalable Error Mitigation for Noisy Quantum Circuits Produces Competitive Advantage.” Nature Physics 19, no. 3 (2023): 319–323. https://doi.org/10.1038/s41567-022-01914-3.

Kim, Youngseok, et al. “Scalable Error Mitigation for Noisy Quantum Circuits Produces Competitive Advantage.” Nature Physics 19, no. 3 (2023): 319–323. https://doi.org/10.1038/s41567-022-01914-3.

Knaut, C., et al. “Entanglement of Nanophotonic Quantum Memory Nodes in a Telecom Network.” Nature 629, no. 16 (2024): 573–578. https://doi.org/10.1038/s41586-024-07252-z.

Kosnoff, Joshua, et al. “Transcranial Focused Ultrasound to V5 Enhances Human Visual Motion Perception.” Nature Communications 15, no. 1 (2024): 48576. https://doi.org/10.1038/s41467-024-48576-8.

Lee, Jaehak, et al. “Fault-Tolerant Quantum Computation by Hybrid Qubits with Bosonic Cat Code and Single Photons.” PRX Quantum 5, no. 3 (August 2, 2024): 030322. https://doi.org/10.1103/PRXQuantum.5.030322.

Leswing, Kif. “Nvidia Dominates the AI Chip Market, but There’s Rising Competition.” CNBC, June 2, 2024. www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html.

Lin, Kevin N., et al. “A Primordial DNA Store and Compute Engine.” Nature Nanotechnology 19, no. 11 (2024): 1654–1664. https://doi.org/10.1038/s41565-024-01771-6.

Liu, Jian-Long, et al. “Creation of Memory—Memory Entanglement in a Metropolitan Quantum Network.” Nature 629 (May 15, 2024): 579–585. https://doi.org/10.1038/s41586-024-07308-0.

Lopez, John. “China Introduces Photonic Chip Capable of Processing Hundred Billion Pixels in Just 6 Nanoseconds.” Tech Times, June 13, 2024. www.techtimes.com/articles/305663/20240613/china-introduces-photonic-chip-capable-processing-hundred-billion-pixels-6-nanoseconds.htm.

Martinez Jr, Alberto. “The Impact of 3D Audio Technology in Music Production.” Flourish & Prosper Music Group, March 4, 2024. https://flourishprosper.net/music-resources/the-impact-of-3d-audio-technology-in-music-production/.

Metz, Cade. “A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too.” The New York Times, July 4, 2024. www.nytimes.com/2024/07/04/technology/openai-hack.html.

Moorhead, Patrick. “Applied Engineers New Materials for AI and the Angstrom Era of Chips.” Forbes, August 5, 2024. www.forbes.com/sites/patrickmoorhead/2024/08/05/applied-engineers-new-materials-for-ai-and-the-angstrom-era-of-chips/.

Mozur, Paul, Ana Swanson, and John Liu. “China’s Top Chip Maker Defies U.S. Sanctions, but for How Long?” The New York Times, September 16, 2024. www.nytimes.com/2024/09/16/technology/smic-china-us-trade-war.html.

Muñiz Cano, Beatriz, et al. “Rashba-like Spin Textures in Graphene Promoted by Ferromagnet-Mediated Electronic Hybridization with a Heavy Metal.” ACS Nano 18, no. 24 (2024): 15716–15728. https://doi.org/10.1021/acsnano.4c02154.

Myers, Andrew. “Spray-on Smart Skin Uses AI to Rapidly Understand Hand Tasks.” Stanford University Department of Chemical Engineering, December 5, 2024. https://cheme.stanford.edu/spray-smart-skin-uses-ai-rapidly-understand-hand-tasks.

Naddaf, Miryam, and Liam Drew. “Second Brain Implant by Elon Musk’s Neuralink: Will It Fare Better Than the First?” Nature, August 6, 2024. https://doi.org/10.1038/d41586-024-02368-8.

National Research Council of Science and Technology. “Researchers Develop New P-Type Semiconductor Materials for Next-Gen Displays.” Tech Xplore, September 13, 2024. techxplore.com/news/2024-09-p-semiconductor-materials-gen-displays.html.

Nevo, Sella, et al. “Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models.” RAND Corp., May 2024. www.rand.org/pubs/research_reports/RRA2849-1.html.

Nguyen, Peter Q., et al. “Population Genomic Analysis of Aegilops Tauschii Identifies Targets for Bread Wheat Improvement.” Nature Biotechnology 40, no. 3 (2022): 422–431. https://doi.org/10.1038/s41587-021-00950-3.

Nikkhah, Vahid, et al. “Inverse-Designed Low-Index Metastructures on a Silicon Photonic Platform for Vector–Matrix Multiplications.” Nature Photonics 18, no. 5 (2024): 402–403. www.nature.com/articles/s41566-024-01394-2.

Nvidia Corp. “Nvidia Announces Financial Results for Fourth Quarter and Fiscal 2023.” Nvidia Investor Relations, February 22, 2023. investor.nvidia.com/news/press-release-details/2023/NVIDIA-Announces-Financial-Results-for-Fourth-Quarter-and-Fiscal-2023/default.aspx.

Omaar, Hodan, and Martin Makaryan. “How Innovative Is China in Quantum?” Information Technology and Innovation Foundation, September 9, 2024. https://itif.org/publications/2024/09/09/how-innovative-is-china-in-quantum/.

OpenAI. “Influence and Cyber Operations: An Update.” OpenAI, October 2024. cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.

OpenAI. “Reimagining Secure Infrastructure for Advanced AI.” OpenAI, May 3, 2024. openai.com/index/reimagining-secure-infrastructure-for-advanced-ai/.

Pasquale, Gabriele, et al. “Electrically Tunable Giant Nernst Effect in Two-Dimensional van der Waals Heterostructures.” Nature Nanotechnology 19, July 2, 2024: 941–947. https://doi.org/10.1038/s41565-024-01717-y.

Pintas, Paolo, et al. “Robust Negative Longitudinal Magnetoresistance and Spin—Orbit Torque in Sputtered Pt▯Sn and Pt▯Sn▯Fe▯▯▯ Topological Semimetal.” Nature Communications 14, no. 4151, July 12, 2023. www.nature.com/articles/s41467-023-39408-2.

Piveteau, Christophe, and David Sutter. “Circuit Knitting with Classical Communication.” IEEE Transactions on Information Theory, 2023. https://doi.org/10.1109/TIT.2023.3310797.

Reardon, Gregory, et al. “Shear Shock Waves Mediate Haptic Holography via Focused Ultrasound.” Science Advances 9, no. 9 (March 1, 2023). https://doi.org/10.1126/sciadv.adf2037.

Rogoway, Mike. “Google’s Water Use Is Soaring in The Dalles, Records Show, with Two More Data Centers to Come.” The Oregonian, December 6, 2022. https://www.oregonlive.com/silicon-forest/2022/12/googles-water-use-is-soaring-in-the-dalles-records-show-with-two-more-data-centers-to-come.html.

Rooney, Paula. “AI on the Mainframe? IBM May Be Onto Something.” CIO, October 3, 2024. www.cio.com/article/3544720/ai-on-the-mainframe-ibm-may-be-onto-something.html.

Rubio-Licht, Nat. “Nvidia May Take Humans out of the Data Center Equation.” The Daily Upside, November 21, 2024. www.thedailyupside.com/technology/artificial-intelligence/nvidia-may-take-humans-out-of-the-data-center-equation/.

Seitinger, Susanne. “AWS Recognized as a First-Time Leader in the 2024 Gartner Magic Quadrant for Data Science and Machine Learning Platforms.” AWS Machine Learning Blog, October 1, 2024. aws.amazon.com/blogs/machine-learning/aws-recognized-as-a-first-time-leader-in-the-2024-gartner-magic-quadrant-for-data-science-and-machine-learning-platforms/.

Semiconductor Industry Association. “Winning the Battle Against Counterfeit Semiconductor Products.” June 2018. www.semiconductors.org/wp-content/uploads/2018/06/ACTF-Whitepaper-Counterfeit-One-Pager-Final.pdf.

Sharma, Deepak, et al. “Linear Symmetric Self-Selecting 14-Bit Kinetic Molecular Memristors.” Nature 633 (2024): 560–566. www.nature.com/articles/s41586-024-07902-2.

Shehzad, Ibrahim, et al. “Automated Cut Finding and Circuit Knitting on Large Quantum Circuits.” QCE 2024, September 15, 2024. https://research.ibm.com/publications/automated-cut-finding-and-circuit-knitting-on-large-quantum-circuits.

Sinanan-Singh, Jasmine, et al. “Single-Shot Quantum Signal Processing Interferometry.” Quantum 8, July 30, 2024: 1427. https://doi.org/10.22331/q-2024-07-30-1427.

“SkinKit: Construction Kit for On-Skin Interface Prototyping.” Hybrid Body Lab, Cornell University. https://www.hybridbody.human.cornell.edu/skinkit.

“Skinput: Appropriating the Body as an Input Surface.” Microsoft Research. https://www.microsoft.com/en-us/research/project/skinput-appropriating-the-body-as-an-input-surface/.

Smith, Matthew S. “Challengers Are Coming for Nvidia’s Crown.” IEEE Spectrum, September 16, 2024, spectrum.ieee.org/nvidia-ai.

Swayne, Matt. “Kipu Quantum Team Says New Quantum Algorithm Outshines Existing Techniques.” The Quantum Insider, September 9, 2024. https://thequantuminsider.com/2024/09/09/kipu-quantum-team-says-new-quantum-algorithm-outshines-existing-techniques/.

Swayne, Matt. “Moderna, IBM Quantum Researchers Use Quantum Computers for Critical Step in RNA-Based Therapeutic Design.” The Quantum Insider, June 3, 2024. https://thequantuminsider.com/2024/06/03/moderna-ibm-quantum-researchers-use-quantum-computers-for-critical-step-in-rna-based-therapeutic-design/.

Trueman, Charlotte. “TSMC Says Arizona Fab Is Now Ahead of Schedule; Signs Semiconductor Talent Agreement with Kyushu University.” DatacenterDynamics, April 2, 2024. www.datacenterdynamics.com/en/news/tsmc-updates-arizona-fab-production-timeline-signs-semiconductor-talent-agreement-with-kyushu-university/.

Turiansky, Mark E., et al. “Rational Design of Efficient Defect-Based Quantum Emitters.” APL Photonics 9, no. 6 (2024): 066117. https://doi.org/10.1063/5.0203366.

Turney, Drew. “Noise-Canceling Headphones Can Use AI to ‘Lock On’ to Somebody When They Speak and Drown Out All Other Noises.” Live Science, June 10, 2024. https://www.livescience.com/technology/artificial-intelligence/noise-canceling-headphones-can-use-ai-to-lock-on-to-somebody-when-they-speak-and-drown-out-all-other-noises.

Tyler, Neil. “SynSense Advances Neuromorphic Audio Processing with Xylo Audio 3 Tapeout.” New Electronics, July 10, 2023. https://www.newelectronics.co.uk/content/news/synsense-advances-neuromorphic-audio-processing-with-xylo-audio-3-tapeout.

U.S. Patent No. 9,557,771. Filed August 6, 2015, and issued January 31, 2017. https://patentimages.storage.googleapis.com/8d/96/b8/cc8b51367a3fd6/US9557771.pdf.

University of Adelaide. “Silicon Chip Propels 6G Communications Forward.” Newswise, August 28, 2024. www.newswise.com/articles/silicon-chip-propels-6g-communications-forward/.

Wang, Chengli, et al. “Lithium Tantalate Photonic Integrated Circuits for Volume Manufacturing.” Nature 629 (2024): 784–790. www.nature.com/articles/s41586-024-07369-1.

Welch, Nicholas. “Litho World & Commerce: Lost in Translation?” ChinaTalk, November 1, 2023. www.chinatalk.media/p/litho-world-and-commerce-lost-in.

Wiggers, Kyle. “Demand for AI Is Driving Data Center Water Consumption Sky High.” TechCrunch, August 19, 2024. https://techcrunch.com/2024/08/19/demand-for-ai-is-driving-data-center-water-consumption-sky-high/.

Williams, Rhiannon. “The Download: A Quantum Breakthrough and the Internet Archive Ruling.” MIT Technology Review, September 11, 2024. https://www.technologyreview.com/2024/09/11/1103848/the-download-a-quantum-breakthrough-and-the-internet-archive-ruling/.

Wilson, Blake, et al. “Authentication through Residual Attention-Based Processing of Tampered Optical Responses.” Advanced Photonics 6, no. 5 (2024), 10.1117/1.AP.6.5.056002.

Yang, Weifeng, et al. “Single Body-Coupled Fiber Enables Chipless Textile Electronics.” Science 384, no. 6641 (2024): 74. https://doi.org/10.1126/science.adk3755.

Yehya, Nadine A. “New Brain-Computer Interface Allows Man with ALS to ‘Speak’ Again.” UC Davis Health, August 14, 2024. https://health.ucdavis.edu/news/headlines/new-brain-computer-interface-allows-man-with-als-to-speak-again/2024/08.

Zafar, Ramish. “ASML Will Now Require Dutch License to Ship Older Chip Machines to China.” Wccftech, September 6, 2024. wccftech.com/asml-will-now-require-dutch-license-to-ship-older-chip-machines-to-china/.

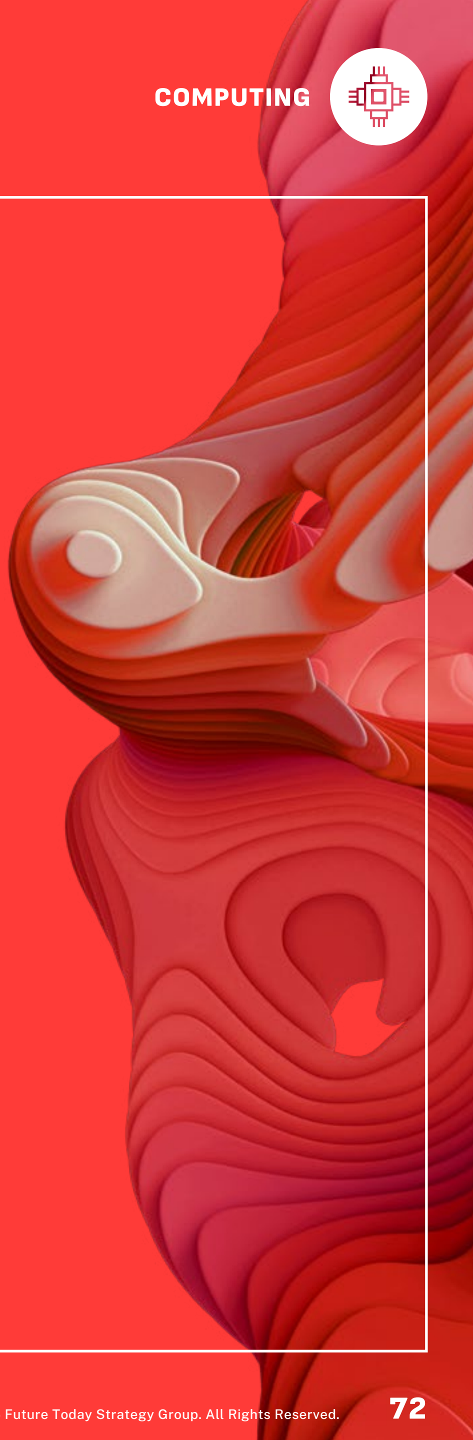
Zhang, Delin, et al. “Robust Negative Longitudinal Magnetoresistance and Spin—Orbit Torque in Sputtered Pt▯Sn and Pt▯Sn▯Fe▯▯▯ Topological Semimetal.” Nature Communications 14, no. 4151 (July 12, 2023). www.nature.com/articles/s41467-023-39408-2.

Zhou, Xinchao, et al. “Trapped Atoms and Superradiance on an Integrated Nanophotonic Microring Circuit.” Physical Review X 14, no. 3 (July 9, 2024). link.aps.org/doi/10.1103/PhysRevX.14.031004.

Zhou, Zichuan, et al. “938 Gb/s, 5–150 GHz Ultra-Wideband Transmission Over the Air Using Combined Electronic and Photonic-Assisted Signal Generation.” IEEE Transactions on Microwave Theory and Techniques 42, no. 20 (August 2024): 1234–1245. https://ieeexplore.ieee.org/document/10643251.

Zhu, Hui Hui, et al. “Large-Scale Photonic Network with Squeezed Vacuum States for Molecular Vibronic Spectroscopy.” Nature Communications 15, no. 6057 (July 18, 2024). https://doi.org/10.1038/s41467-024-50060-2.

ABOUT FUTURE TODAY STRATEGY GROUP



ABOUT US

Future Today Strategy Group is a consulting firm specializing in strategic foresight, a data-driven practice for developing plausible future scenarios to inform today's decisions. As organizations across the globe grapple with an increasingly volatile and uncertain business climate, FTSG provides clarity through long-term strategic planning. Its team of subject matter experts combines best-in-class trends and technology research with actionable strategies to generate business impact. In the two decades since its founding, FTSG has become the preeminent foresight advisory to Fortune 500 companies, world governments, and other major organizations—empowering leaders to make better decisions about the future, today.

CONTACT US

For an introductory conversation to learn how FTSG can assist your organization with its strategic planning and foresight needs, please contact:

inquiries@ftsg.com

ftsg.com

+1 267 342 4300

SOCIAL

Linkedin

@Future-Today-Strategy-Group

The names of companies, services, and products mentioned in this report are not necessarily intended as endorsements by FTSG or this report's authors.

FTSG's 2025 Tech Trends Report relies on data, analysis, and modeling from a number of sources, which includes sources within public and private companies, securities filings, patents, academic research, government agencies, market research firms, conference presentations and papers, and news media stories. Additionally, this report draws from FTSG's previous reports and newsletters. FTSG's reports are occasionally updated on the FTSG website.

FTSG advises hundreds of companies and organizations, some of which are referenced in this report. FTSG does not own any equity position in any of the entities listed in this presentation.

Any trademarks or service marks used in this report are the marks of their respective owners, who do not endorse the statements in this report. All rights in marks are reserved by their respective owners. We disclaim any and all warranties, expressed or implied, with respect to this report.

This report and its contents are proprietary intellectual property of Future Today Strategy Group (FTSG). While internal sharing within organizations is permitted, no part of this report may be modified, published, or commercially distributed without the prior written permission of Future Today Strategy Group.

When citing or referencing this report, please use the following attribution: "Future Today Strategy Group 2025 Tech Trends Report" with appropriate reference to FTSG as the source.

For permission requests regarding commercial use, please contact: inquiries@ftsg.com

FTSG