MIT Technology Review Insights

Produced in partnership with

arm

Heterogeneous compute and a new paradigm for AI at the edge.

The future of AI processing

Preface

"The future of AI processing" is an MIT Technology Review Insights report sponsored by Arm. Drawing on in-depth interviews with senior technology executives, this report explores how heterogeneous compute is enabling new use cases and unlocking AI at the edge now and into the future. Adam Green was the author of the report, Francesca Fanshawe was the editor, and Nicola Crepaldi was the producer. The research is editorially independent, and the views expressed are those of MIT Technology Review Insights.

We would like to thank the following experts for their time and insights:

Ian Bratt, Vice President of Machine Learning Technology and Fellow, Arm

Sy Choudhury, Director, AI Partnerships, Meta

JY Jung, Executive Vice President, Head of SmartThings, Samsung Electronics

Ali Saidi, Senior Principal Engineer, AWS Graviton, Amazon Web Services's Annapurna Labs

CONTENTS

01 Executive summary 4

02 Introduction 5

Fitting the compute to the task 5

Defining heterogeneous compute 5

Moving to the edge 6

Controlling the cost of compute 6

03 Live 7

Wearable AI 7

Enhanced home care 8

AI accelerates 8

Security at the edge 9

04 Work 10

Boosting productivity 10

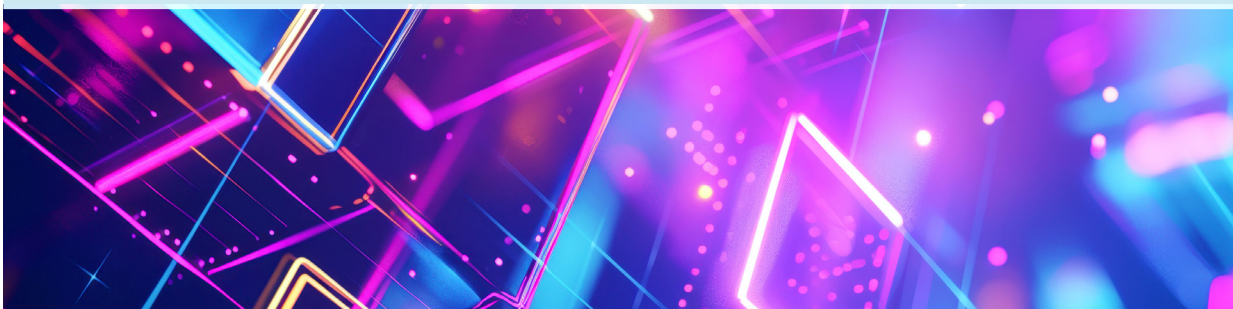
The power of prediction 12

Advancing robotics 12

Evolving workloads 13

05 Play 14

06 Conclusion 15



01 Executive summary

Artificial Intelligence (AI) is emerging in everyday use cases, thanks to advances in foundational models, more powerful chip technology, and abundant data. To become truly embedded and seamless, AI computation must now be distributed – and much of it will take place on device and at the edge.

To support this evolution, computation for running AI workloads must be allocated to the right hardware based on a range of factors, including performance, latency, and power efficiency. Heterogeneous compute enables organizations to allocate workloads dynamically across various computing cores like central processing units (CPUs), graphics processing units (GPUs), neural processing units (NPUs), and other AI accelerators. By assigning workloads to the processors best suited to different purposes, organizations can better balance latency, security, and energy usage in their systems. Key findings from the report are as follows:

- **More AI is moving to inference and the edge.** As AI technology advances, inference – a model's ability to make predictions based on its training – can now be run closer to users and not just in the cloud. This has advanced the deployment of AI to a range of different edge devices, including smartphones, cars, and industrial internet of things (IIoT). Edge processing reduces the reliance on cloud to offer faster response times and enhanced privacy. Going forward, hardware for on-device AI will only improve in areas like memory capacity and energy efficiency.

- **To deliver pervasive AI, organizations are adopting heterogeneous compute.** To commercialize the full panoply of AI use cases, processing and compute must be performed on the right hardware. A heterogeneous approach unlocks a solid, adaptable foundation for the deployment and advancement of AI use cases for everyday life, work, and play. It also allows organizations to prepare for the future of distributed AI in a way that is reliable, efficient, and secure. But there are many trade-offs between cloud and edge computing that require careful consideration based on industry-specific needs.

- **Companies face challenges in managing system complexity and ensuring current architectures can adapt to future needs.** Despite progress in microchip architectures, such as the latest high-performance CPU architectures optimized for AI, software and tooling both need to improve to deliver a compute platform that supports pervasive machine learning, generative AI, and new specializations. Experts stress the importance of developing adaptable architectures that cater to current machine learning demands, while allowing room for technological shifts. The benefits of distributed compute need to outweigh the downsides in terms of complexity across platforms.

02 Introduction

AI has begun to infiltrate everyday life, work, and play in earnest. Wearable health trackers, smart home systems, and hyper-connected cars are becoming common place in our daily lives. Chatbots, co-pilots, and industrial internet of things devices (IIoT) are changing the way we work. Enhanced gaming experiences and AI-powered tools for image, video, and sound are transforming the way we play.

The wide availability of powerful AI models and the progress of machine learning in the last few years have enabled this explosion of use cases, and have led to a shift in focus from training to inference: the operational phase of AI, where models apply what they have learned to real-world situations.

Fitting the compute to the task

For AI use cases to reach their full potential – and become truly ubiquitous – organizations need to optimize the deployment and execution of trained models, driving the need for scalable, cost-effective, and low-latency computing solutions. Heterogeneous computing places workloads on the most efficient processing unit for the task at hand, taking into account power efficiency, compatibility, or the number of cores available.

While recent advances in CPUs are enhancing their ability to handle tasks and real-time inference, AI workloads still need specialized hardware such as GPUs, neural processing units (NPUs), and other AI accelerators designed for high-volume parallel computation. "Heterogeneous computing is for performance and efficiency," explains Ian Bratt, vice president of machine learning technology at Arm. "You might have a workload you're running on one component that it's well suited for, and one part might be better suited for a different component."

Training foundation models, for instance, is typically feasible only with GPUs and AI accelerators in the cloud, while real-time inference can be performed on the edge on devices. Foundational model training in data centers relies on the parallel processing power of GPUs and other accelerators, such as tensor processing units (TPUs) or

Defining heterogeneous compute

Heterogeneous computing enables a single system to have multiple computing sub-systems, by using a variety of computing cores or by using processors based on different computer architectures. An early and relatively common form of heterogeneous computing is the combination of CPUs and a GPU, often used for gaming and other graphics-rich applications.

These processors may execute core instructions differently, work in parallel to accelerate compute speed, and minimize the time required to complete a task. By assigning different workloads to processors that are designed for specific purposes or specialized processing, performance and energy efficiency is improved. This is particularly useful in the development of AI and machine learning workloads where vast amounts of data are processed and converted for a seamless user experience.[1]

data processing units (DPUs), while inference on consumer devices, such as smartphones, can be powered by the latest CPU architectures that offer low latency and improved energy efficiency. These architectures bring privacy benefits because they do not need to send data outside of the device. Moreover, their widespread use, ease of programming, versatility, and memory locality enable developers to implement AI workflows on the CPU to reduce latency, especially when inference is integrated with other processing workflows that take place locally on the CPU.[2]

Moving to the edge

Applications like voice assistants, predictive text, real-time translation, and advanced camera processing are increasingly deployed on device with less reliance on cloud processing. Microchip innovations will further optimize processing and applications for both consumers and developers.[3] In many cases, a blend of approaches is optimal. In gaming, physics simulations and rendering engines must interact seamlessly with other game computations to deliver a smooth experience.[4] These workflows require efficient utilization of different microchip architectures both locally and in the cloud. According to Ali Saidi, senior principal engineer for AWS Graviton, part of AWS's Annapurna Labs, some gaming computation needs to be performed in the cloud. "The reason the game servers are there is because they're trying to create a level playing field," Saidi says. "The game server is doing all the physics computation behind the scenes to make sure that no one's trying to cheat."

As more powerful multimodal and language models get smaller, it is becoming possible to do more on device, including inference. In gaming, generative AI agents are autonomous systems, capable of perceiving their environments and performing a variety of roles including non-player character interactions, dynamically creating landscapes, quests, and narratives, and governing the behavior of objects and physics. Early experiments have demonstrated that using on-device models for agents can enhance the gaming experience by reducing latency and overcoming some of the drawbacks found with LLMs, like game-breaking hallucinations, coherence issues, and unnatural tone.[5]

Meta too sees a shift towards smaller AI workloads on device in the future. "Currently, MetaAI's text and image output or generation features are all happening in the data center," says Sy Choudhury, director of AI partnerships at Meta. "But as we look to the near future, we will be leveraging the smaller Llama models for some on-device use cases. [Storage] will determine how quickly we move. Devices are getting capable enough that we have a good path for the memory, but storage is still a key issue."

As technologies advance and use cases proliferate, heterogeneous computing offers the flexibility and scalability to accelerate the dawn of truly ambient and distributed AI. This report explores AI use cases in three domains – everyday life, work, and play – and the cutting-edge systems and architectures that enable them.

Controlling the cost of compute

AI's rapid growth poses sustainability challenges, with US data center energy use doubling since 2018 and projected to reach 12% of the country's electricity consumption by 2028.[6,7] Heterogeneous computing can enhance energy efficiency by allocating AI workloads to the most suitable processors. For instance, offloading data-intensive tasks to specialized accelerators can significantly cut energy use.[8]

Making the cloud as privacy-preserving as smartphones would be orders of magnitude more expensive in terms of energy consumption, according to a Google case study.[9] However, AI model training remains significantly more power-intensive on-device, which highlights the need for a hybrid approach for training and inference to optimize energy efficiency. Heterogeneous compute enables hybrid approaches.

"Everyone wants to do more with less, whether that is energy or cost," says Ali Saidi, senior principal engineer for AWS Graviton. Saidi argues that cost savings from using better hardware can make a material difference to organizations of all sizes. Customers of AWS Graviton, a family of CPUs designed by Amazon's Annapurna Labs, have redeployed savings into other areas: new ideas, software, or proofs of concept. "If you think about startups, using some of these [decisions] end up saving money, and that means they need to raise less capital. They need fewer funding rounds or they have more distance between those funding rounds," Saidi says.

03 Live

From the first home-control computer, the Echo IV, in the 1960s to today's sophisticated home devices, smart technology has become ever more present in our living spaces.[10] The global smart home market is projected to reach nearly $196 billion by 2030, with devices that enhance comfort while reducing energy use.[11]

Edge AI processing allows secure device performance with reduced latency by conducting inference locally, powering a kind of ambient intelligence that could make our living spaces responsive, like thermostats, lighting, and appliances that adjust to user needs. As JY Jung, executive vice president at Samsung explains: "What AI changes is how these [home] devices understand the user's lifestyle and how users interact with them, and then they optimize themselves for those users."

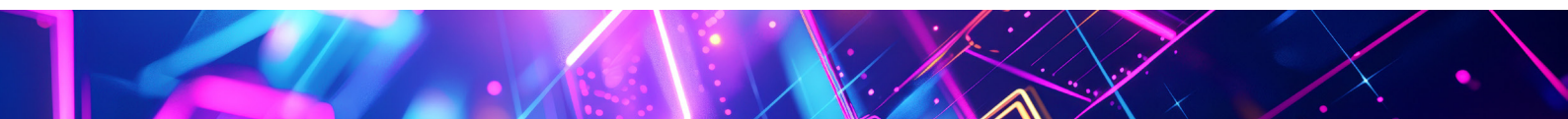
Smart home devices rely on a combination of local and cloud-based processing to manage tasks efficiently. For example, a smart baby monitor might use an on-device AI model for cry detection, while leveraging cloud-based machine learning for long-term sleep pattern predictions. Similarly, intelligent home interfaces use multimodal AI processing to interpret touch, voice, and visual data across different processing units and architectures.

"If you want to take a glance at the air quality at home, it's going to be very lengthy to explain that with voice. So [a visual display] is seen as complementary. Touch interaction, voice-based interaction are complementary modalities, and are needed for machines with AI," Jung explains.

Wearable AI

The Meta smart glasses are a notable step forward in the evolution of everyday AI. Created in collaboration with Ray-Ban, the glasses seamlessly blend multimodal AI into a wearable experience. Real-time AI inference allows users to interact with Meta LLMs for a range of tasks. Users can ask questions, dictate messages, receive navigation assistance, and even get real-time translations of conversations with voice commands. And AI object

A heterogeneous AI system allows developers to optimize performance by using compact processors for on-the-go tasks, while tapping into the power of scalable cloud infrastructure when needed.



recognition can help users to identify landmarks, read text, or gain contextual insights about their surroundings.[12]

Computational workloads are dynamically allocated between edge and cloud processing to balance performance and efficiency. Some features, such as sharing images via voice commands, may require cloud processing,[13] whereas other functions can be handled locally by the glasses' embedded processing units. A heterogeneous AI system allows developers to optimize performance by utilizing the compact processors within such devices for on-the-go tasks, while tapping into the power of scalable cloud infrastructure when needed.

Meta partnerships with platforms like Spotify and Audible highlight the potential of wearable AI. Users can interact with content by asking questions about the music they are listening to or recording and sending messages on WhatsApp using voice commands. With voice, touch, and vision serving as inputs to the AI workflows running behind the scenes, smart wearables hold transformative potential. Another key partnership with the app 'Be My Eyes' aims to support blind and low-vision users by providing real-time assistance. Volunteers can connect with users via point-of-view video calls, which provide the visual perspective of the person making the call to describe surroundings and assist with everyday tasks.[14]

Enhanced home care

In the future, we might see more personal health technology in home environments, especially for older populations and those with chronic diseases to improve disease management, trigger emergency responses, and empower individuals to take control of their health.[15] Remote patient monitoring allows medical professionals to monitor vital signs like heart rate and blood pressure. Wearable biosensors can stream patient data to secure cloud platforms to provide doctors with real-time insights, while allowing patients to move freely.[16] These technologies spot early warning signs and help improve overall health outcomes. AI-powered solutions like Respiro help asthma patients by tracking inhaler use and providing personalized feedback.[17] The Veta Smart Case ensures those with severe allergies always have access to their medication and notifies caregivers if an epinephrine injector is left behind.[18]

As patient monitoring and health tracking becomes part of daily life, heterogeneous compute is crucial in optimizing the balance between on-device privacy and cloud scalability to enable further innovations in compact microchip architectures and AI model performance. Smart health applications will become more powerful as manufacturers and developers collaborate within common standards. The IoT industry is working toward increased device interoperability through standards like Matter to deliver seamless smart device experiences across platforms and device makers, with similar initiatives across the automotive industry.[19]

AI accelerates

AI is powering a new era of car comfort and safety. Use cases include:

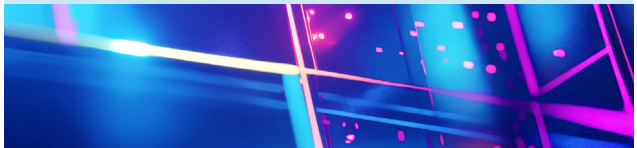
Uber is using NVIDIA's generative AI capabilities to enhance autonomous vehicle safety and scalability.[20]

Volkswagen has integrated ChatGPT into its cars to enable intuitive voice interactions, and seamless infotainment and vehicle control.[21]

Sony Honda Mobility's AFEELA electric car merges AI with gaming technologies by incorporating Unreal Engine-powered augmented reality navigation and a Microsoft-backed voice assistant for an immersive driving experience.[22]

Lotus is embracing AI for safety tasks like detecting, diagnosing, and responding to driving faults proactively, with Arm technology providing the computing backbone.[23]

Mercedes-Benz is using innovative over-the-air updates in its AI-powered software-defined vehicles that adapt to individual driver preferences like climate control and seat position.[24]



"In the near future, we will be leveraging smaller models for some on-device use cases. Devices are getting capable enough that we have a good path for the memory, but storage is still a key issue."

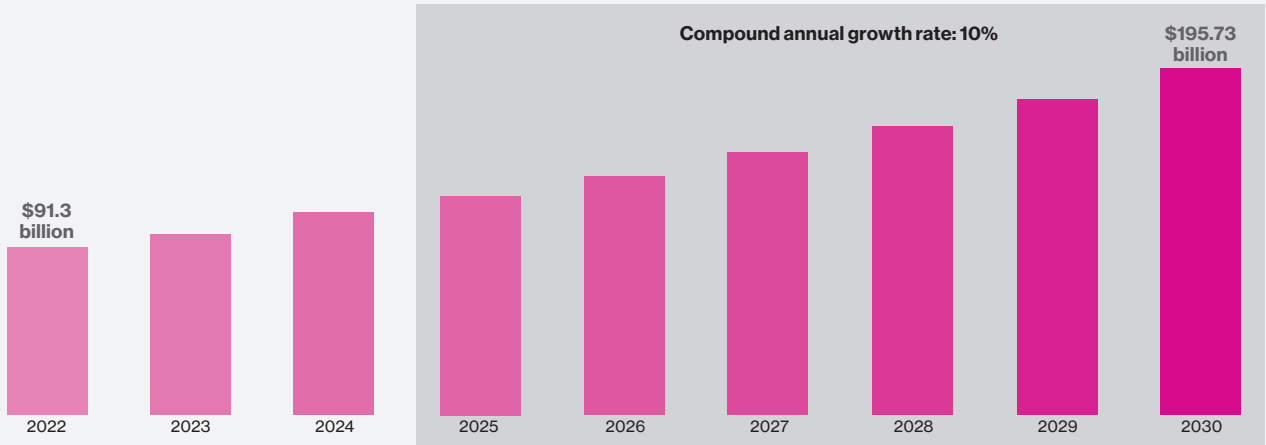
Sy Choudhury, Director of AI Partnerships, Meta

Security at the edge

With 70% of consumers expressing lack of trust in AI product usage, privacy and trust are critical factors for taking AI closer to users.[25] Edge processing helps address these concerns by keeping data local and reducing privacy risks, which is particularly important when processing sensitive information like financial or health data. Security is a key benefit of the tooling and software made available for developers and app makers in a heterogeneous AI system. Developers can allocate different parts of an AI workload to different deployments (i.e., locally versus cloud), and utilize]more of the on-device processing power when needed. For example, a heart disease prediction model can be deployed locally on a fitness tracker to generate predictions without sending private health data like blood pressure readings and sleep patterns to cloud servers.

Going forward, the hardware for running on-device AI processing will only improve. "Any device can be an edge device to process data and run an inference engine," Jung says. "We did a lot of work to optimize the hardware so that we can offer affordable products to our customers. But with these emerging AI technologies and the great benefits that we can deliver to our customers, we are now thinking differently as to how we can add more of the computation and memory in our products so they can process those things from within."

Figure 1: Global smart home market size, 2022 to 2030



Source: Compiled by MIT Technology Review Insights based on data from "Smart Home Market - Global Industry Assessment and Forecast," Vantage Market Research, 2025

04 Work

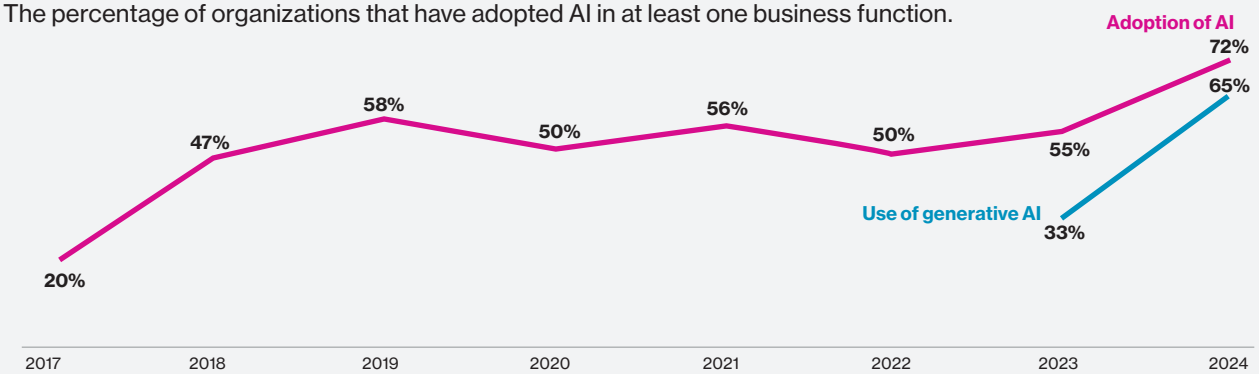
The scale of AI usage at work is hard to calculate, but a 2025 McKinsey survey found that 78% of organizations have integrated AI into at least one business function, up from 72% in 2024 and 55% in 2023.[26] In the near future, self-directed agents will likely increase the scale of AI in the workplace further. "In 2025, I think we're going to see leading LLMs [becoming] much more agentic – meaning self-directed execution of AI agents – and they will become more agents than assistants," Meta's Choudhury predicts.

Boosting productivity

Meta originally leveraged in-house AI for internal needs but is now offering those tools to enterprises. "We were the first to use our own language models to build out our intranet search," Choudhury recalls. "We have an internal tool now that you can ask, 'what's the policy of this or that?' It will not only bring up the relevant pages, but also summarize all the relevant information." By offering Llama to the public, Meta – which plans to spend as much as $65 billion on AI this year[27] – is enabling developers and businesses, including AT&T, Zoom, and Goldman Sachs, to customize applications.[28]

Meta's partnerships with applications like Microsoft 365 allow users to access productivity tools like MS Excel and PowerPoint in a virtual reality environment.[29] And Meta Quest headsets could transform workplace collaboration further by integrating AI tools and mixed-reality features that enable immersive virtual meetings and dynamic 3D environments. AI-driven features, like real-time translation and photorealistic avatars, can help enhance communication by reducing language barriers and providing realistic virtual representations. These position Meta Quest as a tool for the remote workplace to enhance engagement and collaboration among employees.[30]

Figure 2: Growth of AI in the workplace
The percentage of organizations that have adopted AI in at least one business function.



Source: Compiled by MIT Technology Review Insights with data from "The state of AI: How organizations are rewiring to capture value," McKinsey, 2025

Arm

As AI continues to reshape the modern workplace, enterprises are seeking ways to harness its potential, while balancing cost, performance, power efficiency, and security. Findings from Arm's **AI Readiness Index** report reveal that 82% of businesses worldwide are already adopting AI, with 80% prioritizing AI investments for operational efficiencies.

Arm is at the heart of this business transformation, empowering enterprises to unlock AI's full value – whether it's in the cloud or at the edge.

Boosting productivity from cloud to edge

Arm-based cloud and data center computing solutions are enabling businesses to implement AI more cost-effectively and with greater power efficiency. Developers are also supported extensively, with tools like the Arm extension for **GitHub Copilot** accelerating software development on Arm-based server platforms. Meanwhile, at the edge, Arm-based laptops provide a productivity boost for users with multi-day battery life and high performance for the most used business applications.

Driving innovation in industrial IoT and robotics

In industrial IoT and robotics, the Arm compute platform is powering real-world applications and solutions that can sense, adapt, and act autonomously in complex – sometimes potentially hazardous – environments. For example, the AI in Under Control Robotics' (UCR) first humanoid robot prototype, **Moby**, runs on NVIDIA's Jetson Orin platform using Arm Cortex-A processors, making it far more power-efficient and cost-effective than traditional robotic systems.

Transforming business innovation

As AI adoption across enterprises continues to evolve, Arm is collaborating with OpenAI and SoftBank Group on advanced enterprise AI called "**Cristal intelligence**." This features the latest and most advanced models developed by OpenAI, including new AI agents that execute tasks independently to drive innovation and boost productivity across businesses. These are already automating everyday tasks to help transform existing management and operational practices.

Arm, OpenAI, and SoftBank Group have a shared vision to enable these AI agents at scale to make every worker more effective and productive, while empowering them to solve ever more complex problems. Looking ahead, Cristal intelligence's AI agents will lay the groundwork for even more advanced systems that can learn and adapt to any company's needs.

The foundation for AI

Arm is working closely across its industry-leading ecosystem to unleash the full potential of **AI for enterprises**. From power-efficient solutions in the data center and cloud to advanced computing and performance at the edge, the Arm compute platform provides the foundation for power-efficient, scalable and secure AI that is revolutionizing how enterprises innovate and operate globally.

Figure 3: Edge AI market size, 2023 to 2034

The global edge market size is anticipated to grow at a compound annual growth rate of around 21% (figures in $ billion).

Source: Compiled by MIT Technology Review Insights based on data from "Edge AI market size, share and trends 2024 to 2034," Precedence Research, 2025

The power of prediction

Edge AI helps to improve prediction and response times in an industrial setting, lowering the likelihood of operational failure and machine downtime. By analyzing data collected by sensors locally – machine noise, temperature, and vibration, for example – systems can avoid the potential delay caused by sending data to distant servers and act immediately to resolve an issue. Computer vision systems can also detect defects in products before they are advanced to the next phase, and can guide autonomous systems that operate in unpredictable and dynamic environments like handling and packaging.

Data gathered and processed at the edge can also enhance supply chains, helping manufacturers manage material flows more efficiently. Companies like Siemens and Spinnova use AI and digital twins to optimize resources and reduce emissions by streamlining sustainable fiber production.[31] SceneScript, a novel method developed by Meta for reconstructing 3D environments, uses inference to predict structural elements, such as doors and windows, in a spatial environment.[32] These could boost capabilities in navigation, interior design visualization, and even help virtual assistants understand spatial queries about their surroundings.

Advancing robotics

The intersection of AI and robotics, which NVIDIA chief executive officer Jensen Huang has earmarked as the next chapter in the AI revolution, continues to accelerate. NASA Jet Propulsion Laboratory has developed NeBula, an AI system that enables robotic teams, including Boston Dynamics Spot, to autonomously explore Martian-like caves. These robots work together, mapping unknown terrain, detecting scientifically significant features, and even taking samples for analysis. Their ability to traverse rugged landscapes makes them ideal for future planetary exploration, search-and-rescue missions, and navigating hazardous environments.[33]

"Any device can be an edge device to process data and run an inference engine. With these emerging AI technologies, we are now thinking differently as to how we can add more of the computation and memory in our products so they can process those things from within."

JY Jung, Executive Vice President, Samsung

Back on earth, the agriculture sector is hopeful that robotic and autonomous systems can help improve yield and lower environmental impact. Drones, autonomous vehicles, and an array of sensors feed data into machine learning models that can help farmers monitor crop health, detect pests and diseases, and apply fertilizers and pesticides with pinpoint accuracy. Machinery-maker John Deere, for instance, is integrating AI into autonomous tractors and smart sprayers to optimize farming operations. The vehicles use 360-degree camera arrays and machine learning to navigate complex environments.[34]

Evolving workloads

Whether on Mars, on the farm, or on an avatar-rich conference call, the interaction of robots, autonomous vehicles, drones, cameras, and sensors must work like an orchestra, with efficient workload allocation across various systems and architectures. "The workload of the future is changing, and to enable that the underlying compute platform has to change," Arm's Bratt explains.

Despite progress in microchip architectures, such as the latest high-performance CPU architectures optimized for AI, both software and tooling need to improve. "For many years, we've had a roadmap of optimizations to our CPU to increase machine learning performance. We've [also] got a lot of work around optimizing software to enable that future compute platform where machine learning and generative AI is pervasive in everything," Bratt explains.

Coordination is a key challenge while organizations evolve their architecture, especially to unlock tools like AI agents. "Organizations are dependent on different third parties for aspects like security monitoring. They need those people to come along for the ride and people don't think about that until they smash into it," AWS's Saidi observes. "They didn't realize they had this agent and it is mandated on all these machines until the security team came knocking."

"Organizations are dependent on different third parties for aspects like security monitoring. They need those parties to come along for the ride, but people don't think about it until the security team comes knocking."

Ali Saidi, Senior Principal Engineer, AWS Graviton, Amazon Web Services

05 Play

AI is transforming entertainment, enhancing the creative process, and opening new modes and mediums of consumption. Zedd, a record producer, has used AI to create compositions that he believes might otherwise be impossible to make.[35] English musician Imogen Heap has developed an AI twin to enhance her creative process and engage with fans.[36] The artist Grimes has encouraged fans to create AI-generated tracks using her voice.[37] In the visual arts, AI allows creators to explore and blend styles, pushing the boundaries of traditional techniques. Digital artist Refik Anadol uses AI to transform vast datasets into immersive and evolving sculptures for interactive art experiences.[38]

AI can also broaden access to advanced artistic tools, allowing enthusiasts to make high-quality content. Adobe, for example, offers Adobe Firefly, a multimodal generative AI platform replete with models that allow users to create images, videos, and audio content from simple text prompts, streamlining the creative process and making sophisticated design capabilities accessible to a broader audience.[39]

In gaming, legislation such as the EU's Digital Markets Act and the UK's Digital Markets, Competition, and Consumers Act are supporting small-scale developers to distribute games more freely and give players greater control over their data.[40] AI-driven player protection measures are enhancing responsible gaming, with machine learning detecting problematic behavior and enabling early interventions. These shifts promote fairness and create a more transparent and inclusive gaming ecosystem.[41]

Across all use cases, the key is to allocate different compute workloads intelligently. Heterogeneous compute can help entertainment brands deliver rich experiences with minimal latency and more personalized content. For instance, Netflix and Spotify are exploring models and on-device processing to enhance content quality and reduce device playback issues with compression and encoding techniques.[43] But cloud processing also remains crucial for many use cases. "I think you're going to keep seeing training being [done] in data centers because of the size of the models," AWS's Saidi explains. "For a certain set of inferencing, you can't take a 300 billion parameter model and put it on my phone."

Across all use cases, the key is to allocate compute workloads intelligently. Heterogeneous compute can help entertainment brands deliver rich experiences with minimal latency and more personalized content.

06 Conclusion

As AI continues to expand across industries, the challenge is no longer just about deploying AI – it is about optimizing its performance, security, and cost efficiency. Heterogeneous computing has emerged as a critical enabler, allowing organizations to dynamically allocate workloads to best suit use case-specific needs and priorities.

When building their systems, businesses and developers must carefully consider the trade-offs between cloud and edge computing. One consideration is complexity. The benefits of distributed compute must outweigh the potential complications across platforms, says Bratt. "Developers want to operate one program and have it work in many systems. But the heterogeneity of each system is unique because they are coming from different providers and vendors... There are engineering realities that make this tough to manifest."

It is also critical to leave flexibility for future innovations. "The engineering trade-off that all systems designers are dealing with now is how do we specialize for machine learning workloads but not in a way that we're shooting behind the puck. Because then the workload changes and now you've wasted all this effort in your silicon and you're not performant on that new workload," Bratt says. Looking ahead, organizations must strike a balance between specialization and adaptability to ensure their AI systems remain efficient, scalable, and prepared to support the coming wave of AI products and services.

> "The workload of the future is changing. To enable that, the underlying compute platform has to change."
>
> Ian Bratt, Vice President of Machine Learning Technology, Arm

Endnotes

1. “Glossary: Heterogenous Compute”, Arm, 2025, https://www.arm.com/glossary/heterogenous-compute.
2. “A Comprehensive Guide to Understanding AI Inference on the CPU”, Arm, 2024, https://www.arm.com/resources/ebook/cpu-inference.
3. “The CPU Architecture for the Future of AI”, Arm, https://www.arm.com/architecture/cpu/a-profile/armv9.
4. Lucy O’Brien, “How Ubisoft’s New Generative AI Prototype Changes the Narrative for NPCs,” Ubisoft, March 19, 2024, https://news.ubisoft.com/en-us/article/5qXdxhshJBXoanFZApdG3L/how-ubisofts-new-generative-ai-prototype-changes-the-narrative-for-npc.
5. Nathan Yu, “NVIDIA & Inworld AI demo on-device capabilities at Computex,” June 2, 2024, https://inworld.ai/blog/nvidia-inworld-ai-demo-on-device-capabilities#.
6. Arman Shehabi et al., “2024 United States Data Center Energy Usage Report,”, Berkeley Lab, December 2024, https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf.
7. Laila Kearney, “US data-center power use could nearly triple by 2028, DOE-backed report,” Reuters, December 20, 2024, https://www.reuters.com/business/energy/us-data-center-power-use-could-nearly-triple-by-2028-doe-backed-report-says-2024-12-20/.
8. Francis Wang et al., “Architecture-Level Energy Estimation for Heterogeneous Computing Systems,” IEEE, April 28, 2021, https://ieeexplore.ieee.org/document/9408176.
9. David Patterson et al., “Energy and Emissions of Machine Learning on Smartphones vs. the Cloud,” Communications of the ACM, January 26, 2024, https://cacm.acm.org/research/energy-and-emissions-of-machine-learning-on-smartphones-vs-the-cloud/.
10. Dag Spicer, “The ECHO IV Home Computer: 50 Years Later,” Computer History Museum, May 31, 2016, https://computerhistory.org/blog/the-echo-iv-home-computer-50-years-later/.
11. “Smart Home Market – Global Industry Assessment and Forecast,” Vantage Market Research, September, 2024, https://www.vantagemarketresearch.com/industry-report/smart-home-market-1063.
12. Ray-Ban Meta AI glasses, https://www.meta.com/ai-glasses/.
13. Learn more about cloud media on Ray-Ban Meta glasses, https://www.meta.com/help/ai-glasses/734190441863923/.
14. Ray-Ban Meta AI Glasses Are Getting New AI Features and More Partner Integrations, September 25, 2024, https://about.fb.com/news/2024/09/ray-ban-meta-glasses-new-ai-features-and-partner-integrations/.
15. “Healthcare: Improve Healthcare Insights and Outcomes on Arm,” Arm, https://www.arm.com/markets/healthcare.
16. “LifeSignals,” Arm, https://www.arm.com/company/success-library/arm-designs/medical-wearable.
17. “Life-Changing Tech: Get to Know Respiro,” Arm, https://www.arm.com/company/success-library/made-possible/life-changing-tech.
18. “Veta Smart Case,” Arm, https://www.arm.com/company/success-library/arm-designs/allergy-management.
19. Simon Hill, “Here’s What the ‘Matter’ Smart Home Standard Is All About,” Wired, June 9, 2024, https://www.wired.com/story/what-is-matter/.
20. “Uber Teams Up with NVIDIA to Accelerate Autonomous Mobility,” Uber Investor, January 6, 2025, https://investor.uber.com/news-events/news/press-release-details/2025/Uber-Teams-Up-with-NVIDIA-to-Accelerate-Autonomous-Mobility/.
21. “World premiere at CES: Volkswagen integrates ChatGPT into its vehicles,” Volkswagen, January 8, 2024, https://www.volkswagen-newsroom.com/en/press-releases/world-premiere-at-ces-volkswagen-integrates-chatgpt-into-its-vehicles-18048.
22. Asuka Kawanabe, “Future Voice Vol.4 Mobility’s Future: A Platform for Entertainment and Self Expression,”, Afeela, April 24, 2024, https://www.shm-afeela.com/en/stories/2024-04-24/.
23. “Lotus: Eletre and Emeya,” Arm, https://www.arm.com/company/success-library/made-possible/lotus.
24. “The Road Ahead with Mercedes-Benz,” Arm, https://www.arm.com/company/success-library/made-possible/mercedes-benz.
25. Michelle Faverio, “Key findings about Americans and data privacy,” Pew Research Center, October 18, 2023, https://www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy/.
26. “The state of AI in 2025: Gen AI adoption spikes and starts to generate value,” McKinsey & Company, March 12, 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.
27. Daniel Howley, “Meta to spend as much as $65 billion on AI in 2025,” Yahoo Finance, January 24, 2025, https://finance.yahoo.com/news/meta-to-spend-as-much-as-65-billion-on-ai-in-2025-155259000.html.
28. “Unlocking AI Technologies with Meta,” Arm, https://www.arm.com/company/success-library/made-possible/meta.
29. Jeff Teper, “Microsoft and Meta partner to deliver immersive experiences for the future of work and play,” Microsoft, October 11, 2022, https://blogs.microsoft.com/blog/2022/10/11/microsoft-and-meta-partner-to-deliver-immersive-experiences-for-the-future-of-work-and-play/.
30. “Meta Horizon Workrooms,” Meta, https://forwork.meta.com/horizon-workrooms/.
31. “Siemens and Spinnova drive innovation and sustainability in textile manufacturing,” January 7, 2025, https://press.siemens.com/global/en/pressrelease/siemens-and-spinnova-power-innovation-and-sustainability-textile-manufacturing.
32. “Introducing SceneScript, a novel approach for 3D scene reconstruction,” Meta, March 20, 2024, https://ai.meta.com/blog/scenescript-3d-scene-reconstruction-reality-labs-research/.
33. “NASA JPL: Search For Life,” Boston Dynamics, https://bostondynamics.com/case-studies/nasa-jpl-search-for-life/.
34. “John Deere Reveals New Autonomous Machines & Technology at CES 2025,” January 6, 2025 https://www.deere.com/en/news/all-news/autonomous-9RX/.
35. Jack Irvin, “Zedd Says He ‘Would Have Never Been Able to Make the Music I Make If It Wasn’t for AI’,” People Magazine, August 23, 2024, https://people.com/zedd-telos-album-intel-lucky-ai-experience-exclusive-8700314.
36. Katie Hawthorne, “‘I’m empowering my song to go and make love with different people’: Imogen Heap on how her AI twin will rewrite pop,” The Guardian, October 16, 2024, https://www.theguardian.com/music/2024/oct/16/im-empowering-my-song-to-go-and-make-love-with-different-people-imogen-heap-on-how-her-ai-twin-will-rewrite-pop.
37. Simmone Shah, “Grimes Celebrates Trailblazers Creating ‘Magic’ With AI,” Time, February 10, 2025, https://time.com/7212518/grimes-ai-time100-impact-awards-dubai/.
38. Simmone Shah, “Refik Anadol Sees Artistic Possibilities in Data,” Time, February 10, 2025, https://time.com/7212509/refik-anadol-artistic-possibilities-in-data/.
39. “Adobe Firefly,” Adobe, https://www.adobe.com/products/firefly.html.
40. Will Deller, “Horizon Scan 2025: Key Trends in Games & Esports,” Bird&Bird, February 11, 2025, https://www.twobirds.com/en/insights/2025/global/horizon-scan-2025-key-trends-in-games-and-esports.
41. “AI-powered responsible gaming transforms player protection,” SiGMA, https://sigma.world/news/ai-powered-responsible-gaming-transforms-player-protection/.
42. Christos G. Bampis, Li- Heng Chen, and Zhi Li, “For your eyes only: improving Netflix video quality with neural networks,”, Netflix Tech Blog, November 14, 2022, https://netflixtechblog.com/for-your-eyes-only-improving-netflix-video-quality-with-neural-networks-5b8d032da09c.

About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of *MIT Technology Review,* the world's longest-running technology magazine, backed by the world's foremost technology institution – producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the US and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. This content was researched, designed, and written entirely by human writers, editors, analysts, and illustrators. This includes the writing of surveys and collection of data for surveys. AI tools that may have been used were limited to secondary production processes that passed thorough human review.

About Arm

Arm is the industry's highest-performing and most power-efficient compute platform with unmatched scale that touches 100 percent of the connected global population. To meet the insatiable demand for compute, Arm delivers advanced technology solutions that allow the world's leading companies to unleash the unprecedented experiences and capabilities of AI.

The Arm architecture provides the ideal foundation for advanced workloads, with our flexible compute platform offering the heterogeneous computing capabilities to run AI on the CPU or accelerator technologies. Our CPU designs can process a broad range of the AI inference workloads most used across billions of devices, from sensors and smartphones to the cloud and datacenter. Integrating Arm's CPU designs with our leading-edge GPUs offers acceleration capabilities for a broad set of on-device AI use cases. Moreover, the Arm Ethos-NPU runs alongside Cortex-A based systems to deliver accelerated performance for edge AI workloads.

The Arm compute platform also provides our industry partners with the flexibility to create their own customized, differentiated silicon solutions for AI. Together with the world's largest computing ecosystem and 20 million software developers, we are building the future of AI on Arm.

arm

Illustrations
Art by Adobe Stock.

MIT Technology Review Insights

www.technologyreview.com

insights@technologyreview.com