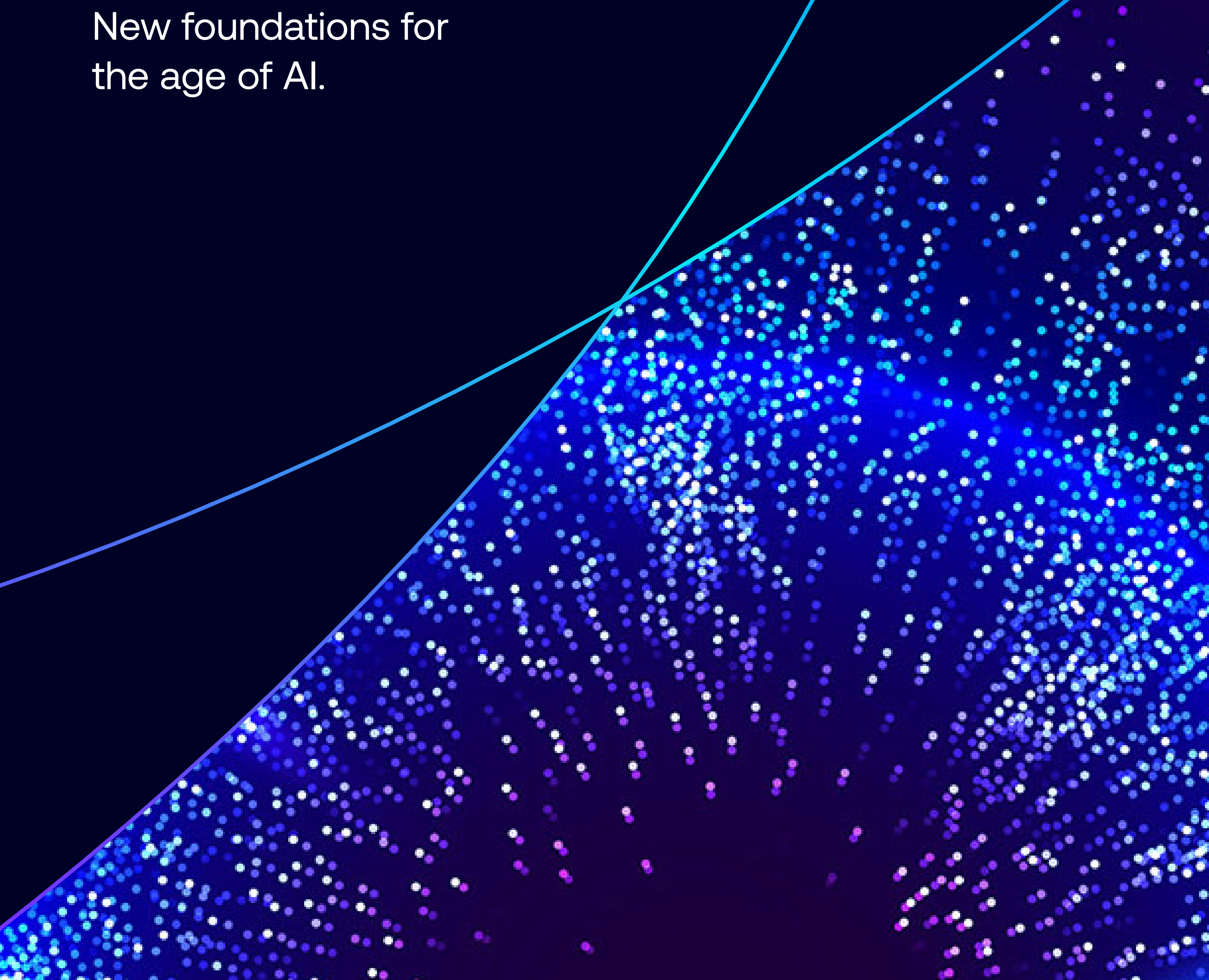
arm

Silicon Reimagined

New foundations for
the age of AI.



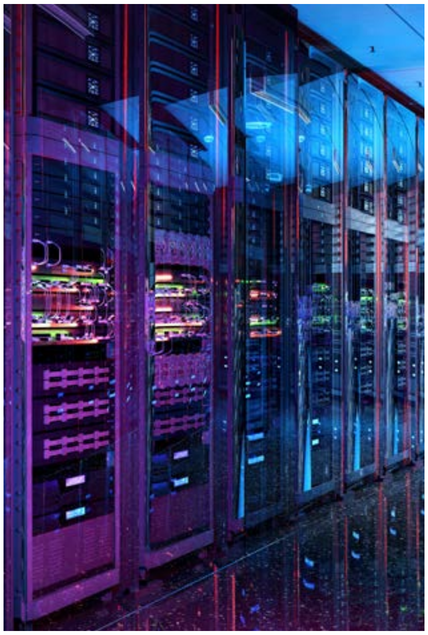
Contents

EXECUTIVE SUMMARY

The semiconductor industry is undergoing a pivotal transformation driven by the rise of artificial intelligence (AI) and the slowing of traditional Moore's Law scaling. This report examines how silicon technologies are evolving to meet the unprecedented computational demands of AI, while addressing critical challenges around power efficiency, security, and reliability.

Arm and industry experts have provided perspectives to give you a sense of where the industry stands today at the dawn of AI, and what challenges and opportunities lie ahead.

EXECUTIVE SUMMARY



KEY FINDINGS

- The traditional approach to semiconductor scaling through Moore's Law **is reaching its physical and economic limits**. The industry is shifting toward innovative alternatives, like custom silicon, compute subsystems (CSS) and chiplets, to deliver continued improvements in performance and efficiency.
- **Power efficiency has become a paramount concern** as AI workloads grow increasingly compute-intensive. Silicon designs are incorporating optimized memory hierarchies, advanced packaging technologies, and sophisticated power management techniques to help reduce energy consumption while maintaining performance.
- **Security threats are evolving alongside AI capabilities**, with the emergence of AI-powered cyber weapons presenting new challenges. The semiconductor industry is responding through multiple layers of hardware and software protection—from cryptographic safeguards embedded in silicon to AI-enhanced security monitoring.
- **Silicon design itself is undergoing radical change**. Chip design and manufacturing – traditionally existing at an arm's-length relationship – are becoming more entwined as new process nodes require deeper collaboration across the ecosystem. Advanced packaging technologies for chiplets are emerging as key drivers of future innovation.
- The software ecosystem **remains critical to unlocking the potential of new silicon architectures**. Support for broad compatibility across AI frameworks while enabling optimizations for custom silicon is essential for widespread adoption.

EXECUTIVE SUMMARY

LOOKING AHEAD, SUCCESS IN SILICON DESIGN WILL INCREASINGLY DEPEND ON:

- Close **collaboration** between IP providers, foundries, and system integrators
- System-level **optimization** across compute, memory, and power delivery
- The standardization of **interfaces** for modular design approaches
- **Specialized architectures** tailored for specific workloads
- Robust **security frameworks** adaptable to emerging threats

The world stands at a breathtaking inflection point with AI, which has moved from experimentation to delivering real business impact. The pace and scale of AI adoption represents one of the most significant technological transformations in recent history.

In previous industry turning points, the promise of a given innovation tended to run ahead of what the ecosystem could deliver. But today, despite some challenges, hardware and software are poised to meet the demands of this AI revolution through unprecedented collaboration across the technology stack. From custom silicon solutions and advanced packaging technologies to sophisticated software frameworks and security measures, the semiconductor industry is evolving in lockstep with AI's computational needs.

This alignment of innovation across hardware and software, coupled with standardization efforts and focus on power efficiency, positions the industry to not just support, but accelerate the next wave of AI advances.

SIGNPOSTS ON THE ROAD TO AI COMPUTING

By Kevork Kechichian, EVP, Solutions Engineering, Arm

In the past 40 years, silicon technologies have undergone profound evolution and change. This has defined the billions of consumer technology products we use today and powers computing solutions in the datacenter and cloud that are vital for processing in the age of AI.

Silicon continues to evolve, particularly in response to the ongoing evolution of AI, where computing demands have never been greater. Therefore, the industry is embracing new approaches and technologies, like custom silicon, compute subsystems (CSS) and chiplets, that are set to define the next decade of technology innovation.

THE INTRODUCTION OF MOORE'S LAW

The 1980s and 1990s set the scene for the past 30 to 40 years of chip designs. Very large scale integration (VLSI) and ultra large scale integration (ULSI) techniques enabled millions to billions of transistors onto a single chip, which led to the development of increasingly powerful chipsets. These approaches were largely built on Moore's Law, which predicted that the number of transistors would double approximately every two years. This led to exponential increases in computing power and efficiency.

> **"The industry-wide drive for efficient computing in AI will cover all technology touchpoints, from the big datacenters to the edge – the actual devices."**

The push for Moore's Law also led to the continual miniaturization of transistors and other components on chipsets throughout the early 1990s. This fueled the rise of mobile chips, as the first cell phones hit the market.

MOBILE CHIPSETS

Arm's own computing heritage is grounded in power-efficient chip designs for the very first cell phones. The Arm-powered Nokia 6110 GSM cell phone of the mid 1990s was a massive commercial success, which led to semiconductor companies developing chipsets that were increasingly optimized for mobile devices.

As these mobile chipsets evolved, there were continuous improvements in performance and power efficiency, alongside a range of new features. This ultimately led to the emergence of smartphones in the mid-2000s, which drove the development of more powerful and sophisticated mobile chips used today and commonly referred to as systems-on-chips (SoC).

These SoCs are highly integrated and incorporate all or most electronic components onto a single chip, including CPUs, GPUs, modems, image signal processors, memory, I/O interfaces and even AI accelerators, like NPUs. This enables a wide range of applications and services that are used on today's smartphones and other mobile devices.

ADOPTION OF MOBILE SOCS ACROSS DIVERSE TECHNOLOGY MARKETS

These mobile SoCs have transcended to other consumer devices and technology markets, such as PC and laptop and IoT and embedded systems. Even for cloud computing and datacenters that require a host of performance and scalability features, the CPU designs have benefited from the performance-per-watt design ethos of mobile SoCs.

The need to add more computing functionalities and components has led to an industry-wide push for greater performance. However, as Moore's Law reaches its limit due to various physical design, technical, and economic factors, there has been a renewed focus on power-efficient computing alongside performance in modern SoC designs.

THE DRIVE FOR AI-BASED EFFICIENT COMPUTING

This is particularly relevant in the age of AI where computing workloads continue to grow both in size and complexity. However, this is not sustainable in the long term, particularly with rising energy usage and costs.

The industry-wide drive for efficient computing in AI covers all technology touchpoints, from the big datacenters to the edge – the actual devices. Where needed and relevant, being able to process AI workloads at the edge provides a range of benefits to consumers and businesses, including improved bandwidth and data privacy and a faster user experience. But this requires more efficient AI processing for smaller devices where space and power is limited.

> **“In the age of AI, approaches to silicon design continue to evolve, with a renewed focus on efficient computing to manage increasingly complex computing workloads.”**

INDUSTRY-WIDE ADOPTION OF CUSTOM SILICON

A significant part of the push for efficient AI is custom silicon. Whereas many SoCs are general purpose chipsets, custom silicon is designed to meet specific requirements for a particular market, application, or customer.

Almost everyone involved in the semiconductor industry is exploring and investing in custom silicon, particularly the ‘big four’ cloud hyperscalers, who are responsible for nearly half of the $100 billion spent on cloud servers every year.

The AWS Graviton4, which is built on Arm technology, is a purpose-built custom silicon solution that accelerates datacenter and AI workloads and delivers significant performance and energy efficiency improvements.

In 2023, Microsoft announced its first custom silicon solution for the cloud, the Microsoft Azure Cobalt, built on the Arm Neoverse CSS, which is designed to tackle complex computing infrastructure challenges. Most recently, Google Cloud announced a custom silicon design for its Arm Neoverse-based Axion CPU, which is designed for complex server workloads in the datacenter.

However, custom silicon is not just for the big cloud hyperscalers. Smaller companies are creating their own custom silicon solutions to meet a variety of complex computing requirements. Faraday Technology, an emerging fabless technology provider, is developing its own 64-core custom silicon SoC for datacenters and advanced 5G networks, which is supported by Arm and Intel Foundry Services (IFS). Moreover, Rebellions, a South Korean AI chip company, announced the development of a new large-scale AI silicon platform to drive power efficiency for AI workloads.

THE RISE OF CSS AND CHIPLETS

Arm Neoverse CSS provides a time-to-market advantage in that it delivers a validated core compute functionality with configurable memory and I/O interfaces. While the CSS itself maintains software consistency, SoC designers can add customized subsystems around it to create differentiated solutions. Companies like Microsoft, Faraday, and Rebellions demonstrate how CSS enables rapid development of custom silicon, while preserving flexibility for system-level optimization.

Advanced packaging technologies and techniques are another aspect of the recent silicon evolution, with these feeding into the development of chiplets. These allow the stacking and interconnection of multiple semiconductor dice to increase performance, improve efficiency and create modern design possibilities, like die-to-die interfaces and new 2.5D and 3D packaging solutions.

Ideally, instead of designing a new chip, silicon manufacturers can add more chiplets to increase computational power and performance, or even upgrade existing chiplets, to bring new products to market faster. Producing smaller chips also increases yield and reduces waste during the manufacturing process.

While many challenges still exist for the industry to navigate, a growing chiplet marketplace can pave the way for custom silicon that reduces costs and uses existing components to become building blocks for the chip, as well as provide greater commercial differentiation for technology companies – big and small. However, the industry needs to work together to provide new agreements for how to build systems that allow the re-use of efforts to provide commercial benefits for companies. In this regard, standardization is a critical part of the process. Arm is already working across the technology ecosystem to accelerate the chiplet market through common frameworks and industry standards.

ONGOING SILICON EVOLUTION

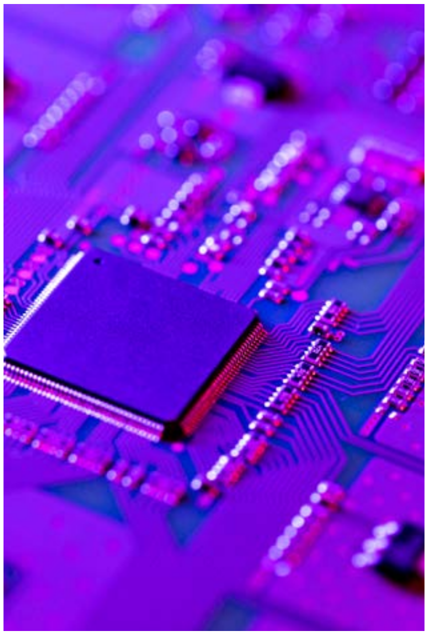
The evolution of silicon in the past 40 years has been one of the most exciting areas of innovation within the tech industry. It's had a profound impact on the world's technologies, enabling them to be more powerful, efficient, and connected, while delivering optimum experiences for users.

In the age of AI, approaches to silicon design continue to evolve, with a renewed focus on efficient computing to manage increasingly complex computing workloads. This is reflected in the move to purpose-built chipsets through custom silicon, CSS, and innovative approaches like chiplets that optimize power and performance area for silicon designs.

ARM'S ROLE

With a 35-year heritage in SoC designs, Arm has been at the forefront of this silicon evolution. From the very first mobile chipsets to SoC designs that are used across diverse technology markets, to new custom silicon approaches being embraced by technology leaders. As silicon continues to evolve, Arm and our partners from the semiconductor industry and beyond continue to play a pivotal role in defining the technologies of the future.

ENERGY EFFICIENCY AND SUSTAINABILITY IN AI COMPUTING



By Nick Horne, VP, ML Engineering, Arm

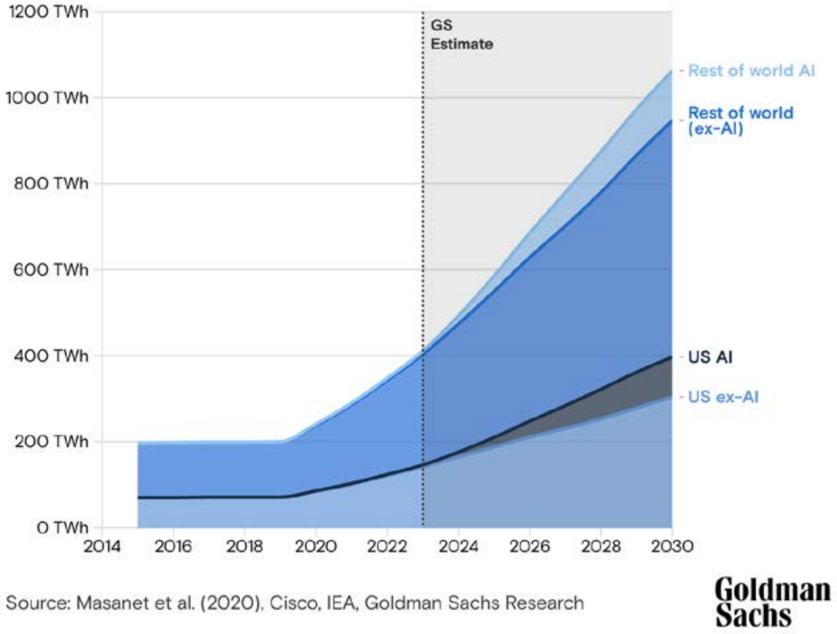
AI is driving a computational revolution, pushing the boundaries of silicon designs and engineering. However, the demands of AI workloads – training massive models and running complex inference tasks – are creating significant challenges for power efficiency, scalability, and cost effectiveness.

ENERGY CHALLENGES IN AI AND SILICON SOLUTIONS

AI workloads are incredibly compute intensive, requiring immense power and energy demands for operations that are only likely to continue to grow in the future. From a silicon design perspective, there are two activities that consume the bulk of the power – computation and data movement. On top of this, cooling is needed to extract the heat generated by these operations.

- **Computation**: AI relies on massive multiply-accumulate operations, which require power-efficient compute structures in the silicon.
- **Data movement**: At the same time, this computation does not happen in the same place. In many instances, certain AI outputs must be processed in another operation or computing component, which requires data movement and introduces additional energy and power overhead. Therefore, optimizing communication between compute components is critical to efficient AI in silicon, with these components existing within the same chip or in another blade or rack entirely.
- **Cooling**: Packing compute and memory units closely to minimize latency and power loss creates heat dissipation challenges. As a result, efficient cooling solutions are essential to manage energy density and ensure hardware reliability.

Data Center Power Demand



Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Research

Source: Goldman Sachs, AI is poised to drive 160% increase in datacenter power demand

To reduce energy costs, silicon designs are integrating optimized memory hierarchies and co-designed communication mechanisms. These solutions minimize data transfer and also minimize the power for the remaining data transfer, leveraging techniques like chip stacking, high-bandwidth memory, and advanced interconnects. At the same time, AI frameworks and algorithms are being increasingly fine-tuned for performance-per-watt and performance-per-cost metrics, which balance compute power with economic sustainability.

> “A fight for ‘AI supremacy’ worldwide (is) leading to significant investment into training an increasing number of models that continue to grow. However, this ‘brute force’ approach is economically unsustainable, so there will be a demand for smarter, more power-efficient silicon solutions.”

GROWING COMPLEXITY IN AI COMPUTING

Currently, there is a fight for "AI supremacy" worldwide, with this leading to significant investment in training an increasing number of models. This requires vast compute resources and months-long training on thousands of devices.

However, this "brute force" approach is economically unsustainable, so there will be a demand for smarter, more power-efficient silicon solutions. At the same time, each failure from these significant AI investments wastes compute and, as a result, power and energy. Minimizing this wasted power requires significant work across all parts of the AI stack, from the silicon to the frameworks.

Fortunately, there have been improvements with more power-efficient silicon designs and the development of smaller, more efficient AI models. Current advances in datatype optimizations, like an industry-wide push to FP4 (4-bit floating point) flexible architectures with new instruction sets and features, are providing incremental gains that are contributing toward increasingly efficient AI. Moreover, the push towards stacking die and 3D packaging technologies is leading to more power-efficient approaches to silicon, like chiplets.

Meanwhile, the industry has proven that smaller, more efficient AI models are possible. These are ideal for generative AI workloads on smaller devices, offering a reduced footprint, higher levels of accuracy, and improved portability. Looking ahead, the future of AI is likely to lie in specialized models that are tailored for specific tasks, such as medical diagnostics or localized speech recognition.

AI AND SILICON: A SYMBIOTIC EVOLUTION

A question often posed is whether AI is influencing silicon or whether the opposite is true. In reality, it is a bit of both. The development of AI and silicon is a reciprocal process where both domains influence each other.

There are already instances where AI defines silicon designs, with data scientists demanding hardware capable of supporting new datatypes and model architectures to drive AI innovation. Therefore, silicon designs are being constantly adapted to meet these needs, as hardware limitations can constrain the growth of AI. However, the big challenge is that hardware takes years to tape-out, whereas in the fast-moving world of AI, new models are being released on a regular cadence, such as the multiple releases of new ChatGPT and Llama models.

At the same time, innovations in the underlying architecture, like new instruction sets and features, can enable significant AI innovation opportunities. For example, Arm regularly releases new features, like SVE2 and Scalable Matrix Extension (SME), that are built-in to the Arm architecture, adding new AI capabilities for the ecosystem. It must also be recognized that today's hardware has adapted to become very capable of training transformer-style models, which form the basis of generative AI.

The ideal scenario is one where AI and silicon lead to a holistic codesign approach, where hardware and algorithms are developed in tandem to achieve optimal performance and efficiency.

THE ROLE OF EDGE AND DATACENTER PROCESSING

Now and in the future, AI requires a mix of processing in the cloud and at the edge. Both have a role to play in the drive toward more efficient AI. In fact, edge processing will augment with datacenter processing, handling inference tasks in a more power-efficient manner. By distributing workloads intelligently, overall system efficiency can be improved.

As computing capabilities grow on devices, AI processing at the edge is becoming more common. This is being driven by silicon designs that can run a variety of AI workloads in power-constrained environments, like mobile devices. AI at the edge provides power and cost savings, as it removes the need to transfer data back and forth between devices and the cloud. This offers enhanced privacy by keeping personal data local, while lowering latency that is crucial for faster AI-based experiences.

The move to AI at the edge is being supported by a range of devices, from mobile to IoT, that are adopting smaller, more power-efficient AI models. For example, Arm's optimization work with Meta means its Llama 3.2 large language models (LLMs) are running on Arm-powered mobile devices at faster speeds than ever before.

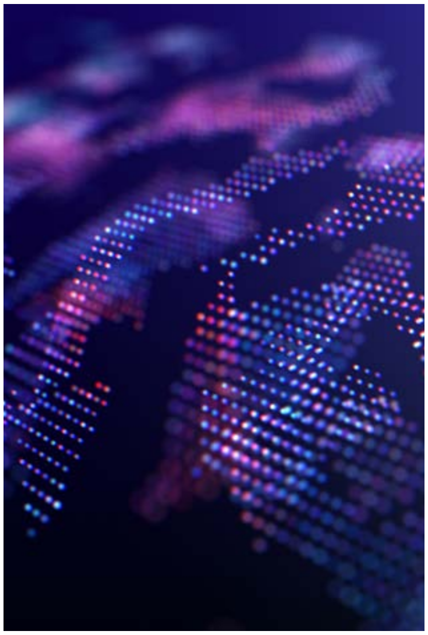
Efficient AI is indispensable to datacenters, especially due to the large-scale AI workloads that are being processed in these environments. The industry can never move away from the fact that large-scale AI training needs to happen in the cloud because the performance requirements for these workloads are simply too high. Therefore, processing in the datacenter needs to be as efficient as possible.

In the datacenter, AI is being leveraged to enhance resource allocation, cooling energy usage, helping to ensure more sustainable operations. At the same time, AI algorithms are being used to optimize operations by balancing power usage and performance when managing datacenter workloads. AI is also being leveraged for system monitoring to reduce failures and downtime.

A BALANCED FUTURE

Efficient AI for silicon represents a convergence of challenges and opportunities. Innovations in hardware, codesigned with software, are paving the way for sustainable, scalable AI solutions. As edge and datacenter processing complement each other, the future of silicon is poised to unlock the full potential of AI, while addressing the pressing need for power-efficiency. The path forward lies in balancing power, performance, and economic viability, alongside hardware architectures that are adaptable to ongoing changes across the wider AI and technology landscape. This can help ensure that the AI revolution on silicon is as sustainable as it is transformative.

SILICON UNDER SIEGE: DEFENDING CHIPS IN THE AGE OF AI



By Rob Coombs, Director, Business Development, Architecture & Technology, Arm

When it comes to AI technology, the bad guys seem to be working just as hard as the good guys.

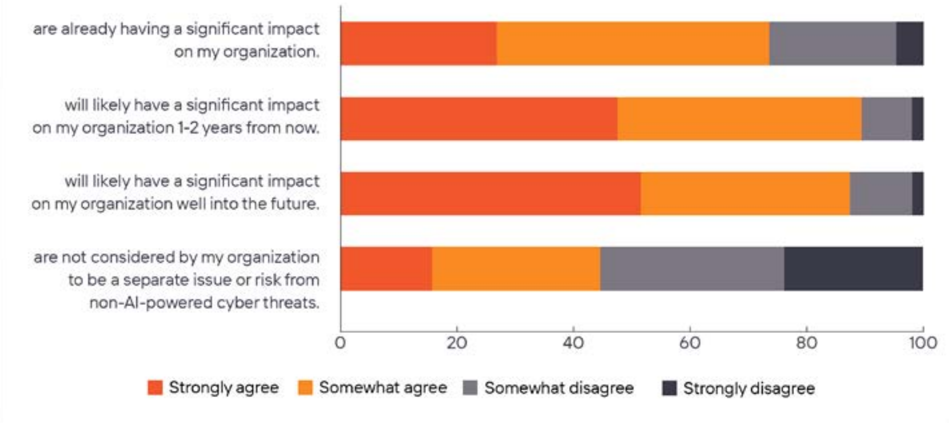
"We've reached a point where fully autonomous, agentic AI cyber weapons can evolve themselves, identify vulnerabilities, and execute sophisticated attacks," says Matt Griffin, founder of the 311 Institute. "In one instance, an AI weapon daisy-chained zero-day exploits to breach military-grade systems in just two minutes. The implications for cybersecurity are staggering. It's no longer just about defending against traditional malware but preparing for a future where AI itself becomes the attacker."

Nearly 74% of survey respondents to a 2024 DarkTrace survey reported seeing AI-powered cyber threats significantly impact their organizations, while 60% of respondents fear their organizations are not adequately prepared to defend against AI-powered threats and attacks. In an industry where "old-school" back-door and side-channel attacks still chill the spine, AI security threats are another grim possibility altogether.

> **"It's no longer just about defending against traditional malware but preparing for a future where AI itself becomes the attacker."**

While AI-powered cyber weapons are an emerging threat frontier, the semiconductor industry is actively evolving its defense capabilities through multiple layers of hardware and software protection. From cryptographic safeguards embedded directly in silicon to AI-enhanced security monitoring, modern SoC architectures are being fortified against both traditional and next-generation attacks. The challenge lies not just in defending against known threats, but in architecting security features that can adapt to and counter the increasing sophistication of AI-driven attacks, while maintaining system performance and reliability. This has led to several key developments in how security is integrated into modern semiconductor designs.

AI-Powered Cyber Threats:



Source: DarkTrace, 2024 State of AI Cyber Security Report

INTEGRATION OF SECURITY FEATURES IN MODERN SOCS

Modern SoCs have evolved far beyond simple processing units, incorporating sophisticated security mechanisms that work in concert to create multiple layers of defense against both current and emerging threats.

The first line of defense begins with robust cryptographic protections built directly into the silicon. Through hardware-accelerated encryption and secure boot processes, SoCs create a chain of trust that prevents unauthorized code from executing on the system. This protection extends throughout the device’s lifecycle through secure update mechanisms that adhere to rigorous standards like ISO 24089, ensuring that every software update is authenticated and tamper-proof.

As AI models become more prevalent in everyday computing, protecting their integrity has become paramount. This has given rise to confidential computing architectures that create secure enclaves for sensitive AI operations, even in potentially untrusted environments. Memory vulnerabilities, long a favorite target of attackers, are now being addressed through innovations like memory tagging extension in the Armv9 architecture, making it significantly harder for malicious actors to exploit memory-based attacks.

The concept of trust in modern SoCs is anchored by a hardware root of trust (RoT), which acts as a silicon-based source of truth for the entire system. This foundation enables robust authentication and attestation capabilities, verifying the integrity of every component and process from the moment the device powers on.

Perhaps most intriguingly, AI itself is being wielded as a powerful tool for enhancing security. Through network-based monitoring and advanced code analysis, AI-driven technologies can now detect suspicious behaviors and identify vulnerabilities at a scale and speed impossible for human operators. This creates a powerful feedback loop where AI not only powers new applications but also helps protect the very systems it runs on.

RELIABILITY CONSIDERATIONS FOR AI WORKLOADS

The meteoric rise of AI has introduced unprecedented demands on silicon architectures. Today's AI systems must not only handle massive computational loads but also maintain unwavering reliability – a challenge that pushes semiconductor design to new frontiers of innovation.

At the heart of this challenge lies the dynamic nature of AI itself. Unlike traditional software, AI models experience what researchers call "concept drift" and "model decay" over time, where their performance gradually deteriorates as real-world conditions evolve beyond their training data. This necessitates periodic retraining and updates, requiring silicon architectures with sophisticated fault-tolerant designs that can maintain stability even as workloads shift and evolve.

The sheer scale of AI computations makes error handling particularly critical. Even minor data corruptions can cascade into significant problems, which is why modern designs incorporate robust hardware-level error detection and correction mechanisms. Error-correcting codes (ECC) work tirelessly across memory and communication subsystems, helping to ensure data integrity at every step of the AI computation process.

As AI models continue to grow exponentially in size and complexity – with some models now reaching trillions of parameters – the need for scalable compute architectures has become paramount. The industry's response has been innovative: modular chiplet designs and specialized AI accelerators are increasingly being integrated within SoCs, offering the flexibility to scale computing power while maintaining reliability.

However, this increasing sophistication brings its own security challenges. While AI enhances our defensive capabilities, it also provides new attack vectors for malicious actors. From targeted financial fraud to sophisticated manipulation of biometric authentication systems, the threats are evolving rapidly. The response has been to double down on hardware-based security, building systems with immutable roots of trust and standardized security methodologies that can withstand even AI-powered attacks.

IMPLICATIONS OF EDGE PROCESSING FOR DATA PRIVACY

As AI moves from centralized datacenters to the outer edges of our networks, a new paradigm in data privacy and security is emerging. Edge computing represents more than just a technological shift: It fundamentally changes how we think about data protection, bringing both promising solutions and novel challenges to the forefront of semiconductor design

The core advantage of edge processing lies in its locality. By processing data where it's generated – whether in smartphones, software-defined vehicles (SDVs), or industrial IoT – edge computing dramatically reduces the need to transmit sensitive information across networks. This localization naturally enhances privacy and security, creating a smaller attack surface for potential data breaches.

However, protecting data at the edge requires sophisticated hardware solutions. Modern SoCs incorporate secure enclaves and trusted execution environments (TEEs), creating fortified digital vaults within the chip itself. These isolated environments help ensure that sensitive AI operations, particularly inference processing, remain shielded from unauthorized access or manipulation.

The protection of AI models themselves presents another critical challenge. These models, often representing millions of dollars in research and development, must be secured against both theft and tampering. This has led to the development of comprehensive security protocols that help protect models during deployment, storage, and runtime operations, helping to ensure their integrity throughout the entire lifecycle.

And then there's the question of data authenticity, which is crucial. In an edge computing environment, ensuring the integrity of input data becomes paramount. Consider processing real-time video feeds in autonomous driving solutions - any compromise in the authenticity of this data could have catastrophic consequences. This reality has driven the development of sophisticated data provenance systems that verify and protect input integrity.

The heterogeneous nature of edge computing adds another layer of complexity. With devices employing diverse computing architectures, securing application deployment becomes increasingly challenging. The industry has responded with advanced security measures, including privileged access controls and secure memory management units (MMUs), creating a unified security framework across this diverse ecosystem.

SECURITY STANDARDS FOR CUSTOM SILICON

The rise of specialized computing has driven unprecedented demand for custom silicon solutions, each tailored to specific applications. While this customization enables remarkable performance optimization, it also necessitates robust security frameworks to ensure these specialized chips maintain stringent protection standards.

Leading this standardization effort is the Platform Security Architecture (PSA) Certified framework, which has emerged as the gold standard for secure silicon design. By establishing comprehensive requirements for secure boot processes, cryptographic services, and update protocols, PSA Certification provides manufacturers with a clear roadmap for building security into the very foundation of their custom silicon solutions

Complementing this framework, the Security Evaluation Standard for IoT Platforms (SESIP) offers a structured methodology for assessing security in custom chips. This standard delves deep into crucial security aspects, from secure initialization procedures to the thorough purging of sensitive data, ensuring that custom silicon maintains its integrity throughout its operational lifecycle.

The regulatory landscape adds another critical dimension to custom silicon security. Standards such as UNECE R155 for vehicle cybersecurity and ISO/SAE 21434 for automotive systems reflect the growing recognition that custom silicon must meet rigorous industry-specific requirements. These regulations are not just checkboxes; they are evolving frameworks that help manufacturers stay ahead of emerging threats while ensuring their solutions remain compliant across global markets.

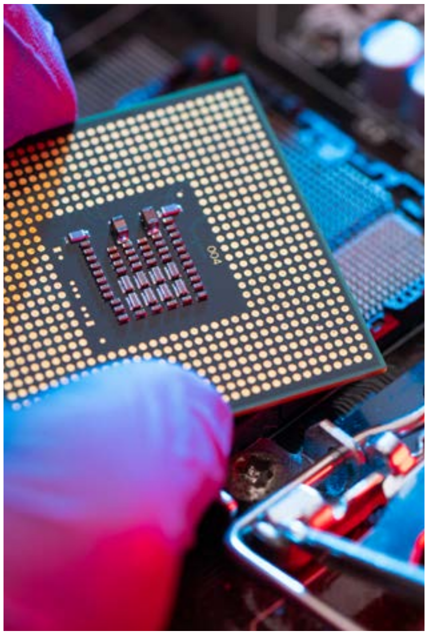
UNLOCKING THE POTENTIAL OF SILICON WITH SECURITY

As we look to the future of silicon, the dual imperatives of security and reliability will shape the next generation of semiconductor innovations.

From integrating advanced security features in SoCs to addressing the unique challenges posed by AI workloads and edge processing, the industry must prioritize robust frameworks and standards. By adopting comprehensive strategies, the silicon ecosystem can not only mitigate risks but also unlock transformative potential across sectors.

CHAPTER 4

THE GREAT REWIRING: INSIDE THE PROFOUND TRANSFORMATION OF CHIP DESIGN IN THE AGE OF AI



By Rainer Herberholz, Senior Physical Architect & Fellow, Arm

The only thing constant in life is change. And if Heraclitus were alive, he would be stunned at how much faster chip design is changing today with the rise of AI workloads and the slowing of semiconductor scaling.

For decades, the semiconductor industry marched steadily forward to the drumbeat of Moore's Law – the observation that the number of transistors on a chip doubles approximately every two years, while cost, performance, and power of logic gates all improved. This predictable cadence allowed chip designers and manufacturers to operate almost independently, with minimal need to coordinate deeply on implementation details. Each new process node brought reliable improvements in density, performance, and power efficiency that designers could count on.

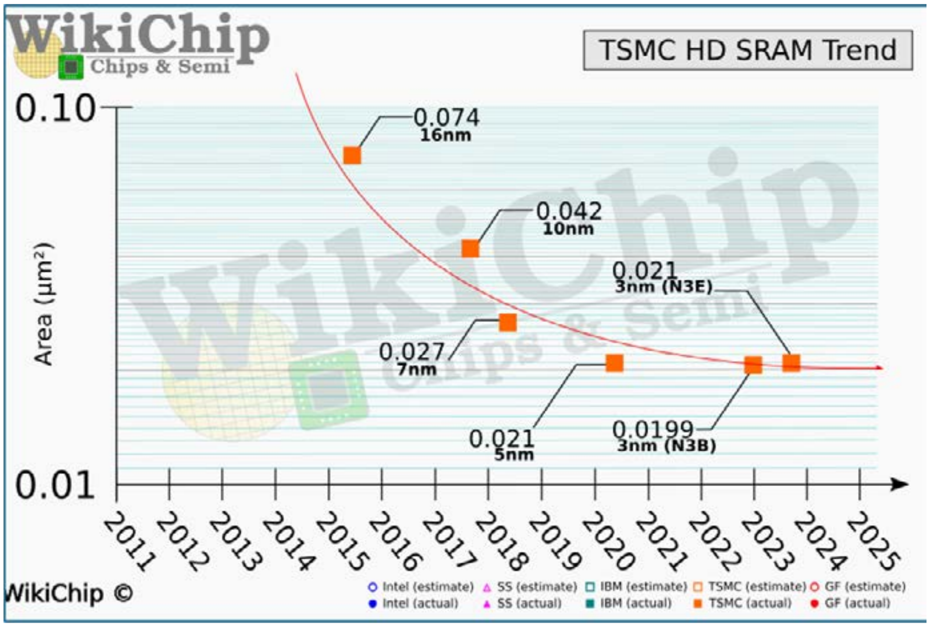
> "The trend toward chiplet-based designs is not actually about making things smaller. In fact, total system sizes continue to grow as gate counts increase faster than what scaling alone could support."

But that era is fast drawing to a close. As we push up against the physical limits of semiconductor scaling, we find ourselves at an inflection point that demands fundamental changes in how chips are designed and manufactured.

THE END OF TRADITIONAL SCALING

The first cracks in the foundation appeared around the 28 nm node – when planar technologies ran into scaling problems – but the real challenges have emerged at 7 nm and beyond. As we move into the gate-all-around era at 2 nm and below, the benefits of traditional scaling are becoming increasingly difficult and expensive to achieve. The criteria for what constitute an advantageous next process node – and how to make it work for specific designs – have become far more complex and interconnected.

Today, it is clear that traditional scaling – the steady shrinking of transistor pitches that powered decades of progress – has not just slowed; it has effectively ended. This reality is starkly illustrated in SRAM scaling, where there has been no meaningful cell area reduction from 5 nm to 2 nm nodes, though power and performance improvements continue.



Source: WikiChip, Chip area scaling expected to be limited now by SRAM, particularly those chips emphasizing AI.

What remains of "scaling" now comes primarily through structural innovations known as "scaling boosters" – architectural changes that improve density through means other than simple pitch reduction. These include continuous active/OD (COD), contact overactive gate (COAG), gate-all-around (GAA) transistors and backside power delivery.

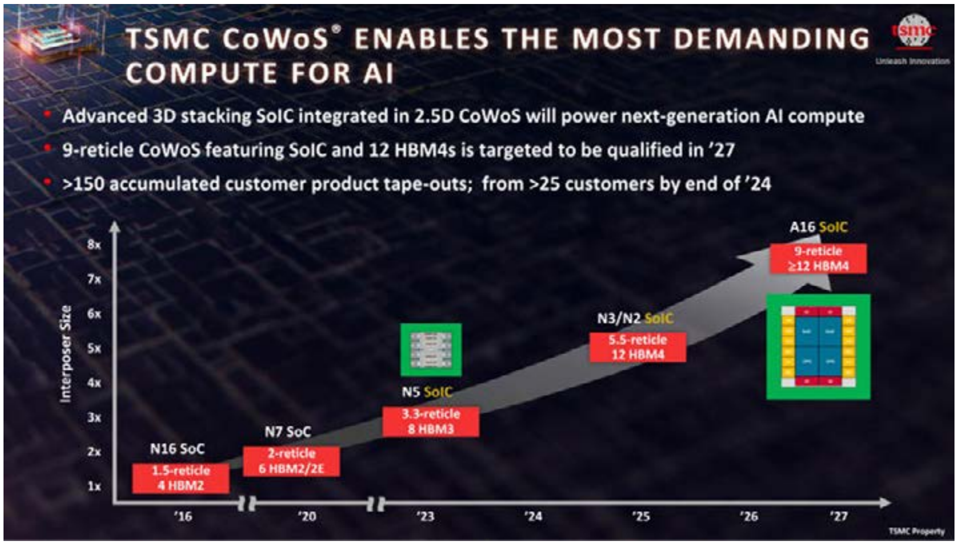
The next major advance may come from complementary FETs (CFETs), also known as monolithic 3D, though its commercial timeline remains uncertain. More revolutionary advances, such as new channel materials like MOS2, lie even further in the future.

This new reality has implications for on-die memory. While logic can still benefit from various scaling boosters, SRAM won't see major density improvements until CFET technology matures. This means on-die memory is becoming significantly more expensive per bit compared to logic – a crucial consideration for compute architectures.

This has profound implications for the semiconductor ecosystem. Companies can no longer create silicon IP without deeply understanding how that IP will be implemented in actual chips. The historical separation between design and manufacturing is dissolving, replaced by necessary collaboration between domains that previously operated independently.

THE RISE OF CHIPLETS AND ADVANCED PACKAGING

As traditional scaling ends, advanced packaging has emerged as perhaps the true successor to Moore's Law – though one that comes with its own set of profound constraints. The trend toward chiplet-based designs is not actually about making things smaller. In fact, total system sizes continue to grow as gate counts increase faster than what scaling alone could support. This evolution began in the 2010s when TSMC introduced CoWoS (chip on wafer on substrate) to allow systems to scale beyond the size limitations of fab lithography equipment, with 9x scaling now on the roadmap.



TSMC CoWoS® ENABLES THE MOST DEMANDING COMPUTE FOR AI

- Advanced 3D stacking SoIC integrated in 2.5D CoWoS will power next-generation AI compute
- 9-reticle CoWoS featuring SoIC and 12 HBM4s is targeted to be qualified in '27
- >150 accumulated customer product tape-outs; from >25 customers by end of '24

Source: TSMC 'Super Carrier' CoWoS interposer gets bigger, enabling massive AI chips to reach 9-reticle sizes with 12 HBM4 stacks

Partitioning systems into chiplets and integrating them side-by-side or 3D-stacked adds new opportunities. These include:

- Greater flexibility to create different designs for different use cases.
- The ability to optimize different components for specific functions.
- Improved yield by manufacturing smaller dies.
- Greater potential for reuse across different products.

However, these benefits come with significant challenges. Power delivery becomes more complex when components are spread across multiple dies. 3D-stacking increases the power-density and thus results in challenges for power delivery, as well as thermal management. The interfaces between chiplets introduce new design considerations around latency, as well as power-control and efficiency.

THE MEMORY CHALLENGE

As memory scaling slows on-die, new forms of compute, like advanced

AI workloads, emerge with a massively increased demand for memory bandwidth. However, traditional architectures struggle to deliver this compute efficiently. This is because the physical separation between compute and memory elements creates bottlenecks that limit performance and consume power.

This has sparked innovation in memory architectures, including new approaches to high-bandwidth memory (HBM) integration. The logic die underneath HBM stacks is evolving to support more sophisticated processing capabilities, enabling near-memory computing architectures that help reduce the time and power required to move data between memory and processors. This represents a significant shift from treating memory as a commodity component to viewing it as an integral part of the compute architecture.

POWER: THE NEW CURRENCY

As chips become more complex and AI workloads more demanding, power management has emerged as perhaps the most critical challenge facing designers. Training a single large AI model can consume enormous amounts of electricity. For example, in the United States alone, datacenter electricity consumption was 2.5% of the U.S. total (~130 TWh) in 2022 and is expected to triple to 7.5% (~390 TWh) by 2030, according to the Boston Consulting Group. That's the equivalent of the electricity used by about 40 million U.S. houses – almost a third of the total homes in the U.S.

This reality is driving innovation in several areas:

- More sophisticated power domain management
- New approaches to voltage regulation and distribution
- Enhanced thermal management techniques
- Optimization of compute resources based on workload characteristics

The challenge is particularly acute in chiplet-based designs, where power delivery must be coordinated across multiple dies, while staying within strict thermal limits. The industry is working toward standardization of power architectures for reusable chiplets, but significant work remains.

AI'S IMPACT ON CHIP DESIGN

The rise of AI workloads is fundamentally altering how chips are designed. These workloads demand different architectures optimized for massive parallel processing and memory bandwidth. This has led to:

- New specialized accelerator architectures
- Innovation in memory subsystems
- Enhanced focus on power efficiency
- Evolution in packaging and integration approaches

Moreover, AI is becoming an integral part of the chip design process itself. Machine learning (ML) techniques are being applied to various aspects of design, from place-and-route to power optimization. This creates an interesting feedback loop where AI is helping to design the very chips that will run AI workloads.

As we look to the future, several trends are clear:

- **Increased collaboration**: The complexity of modern chip design demands closer cooperation between IP providers, foundries, packaging houses, and system integrators. No single company can go it alone.
- **Systems focus**: Success requires thinking beyond individual components to optimize entire systems, including compute, memory, power delivery, and thermal management.
- **Standardization needs**: The industry must develop new standards for chiplet interfaces, power delivery and thermal management to enable a

truly modular approach to chip design.

- Power-aware design: Every aspect of chip design must be viewed through the lens of power efficiency, from architecture to implementation.
- Specialization: Different workloads will increasingly demand specialized architectures, leading to greater diversity in chip designs.

THE ROLE OF INNOVATION

While traditional scaling may be slowing, innovation in the semiconductor industry is accelerating. New materials, architectures integration techniques are being explored, as we mentioned earlier. The industry is also seeing the emergence of new players focused on specialized solutions for AI and other demanding workloads.

> “Companies can no longer create silicon IP without deeply understanding how that IP will be implemented in actual chips.”

Success in this new era requires embracing complexity, while finding ways to manage it effectively. This means developing new tools and methodologies, fostering collaboration across the ecosystem, and maintaining a relentless focus on system-level optimization.

LOOKING AHEAD

The semiconductor industry stands at a crossroads. The reliable scaling that drove progress for decades is giving way to a more complex landscape of trade-offs and choices. This creates both challenges and opportunities for innovation.

The future of chip design will be characterized by:

- Greater integration of different technologies and approaches

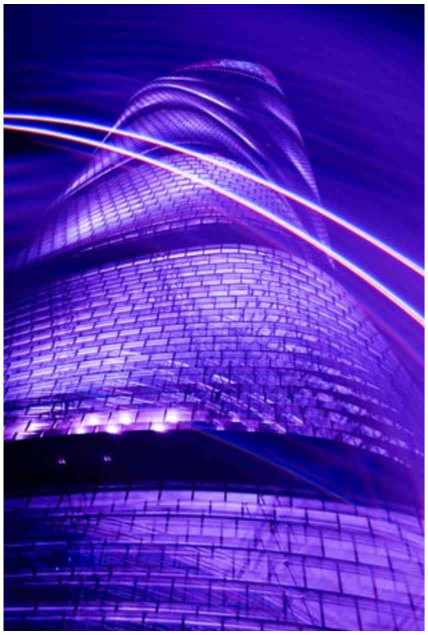
- More sophisticated power management techniques
- Closer collaboration across the ecosystem
- Innovation in memory architectures and integration
- Specialized solutions for AI and other demanding workloads

The companies that succeed will be those that can navigate this complexity while delivering solutions that meet the ever-increasing demands of modern computing workloads. The era of treating chip design as a relatively straightforward process of implementing standardized components is over. In its place, we are seeing the emergence of more nuanced and sophisticated approaches that demand creativity, collaboration, and systems-level thinking.

The only certainty is that the pace of change will continue to accelerate. The future belongs to those who can embrace this change while finding ways to manage its complexity effectively. That probably would not surprise Heraclitus.

CHAPTER 5

ARCHITECTURAL AWAKENING: HOW ARM’S HETEROGENEOUS COMPUTING IS TRANSFORMING AI



By Ben Bajarin, CEO and Principal Analyst, Creative Strategies

The AI revolution is reshaping the computing landscape, with profound implications for processor architecture and datacenter infrastructure. As AI systems grow increasingly sophisticated—from LLMs to advanced reasoning agents—they are driving unprecedented demands for computational power. This evolution is particularly significant for CPU designers and companies like Arm that find themselves at the intersection of two critical trends: the need for specialized AI acceleration and the fundamental requirement for powerful host processors to support these new workloads.

The relationship between AI development and CPU architecture has become more intricate and consequential than ever before, creating both challenges and opportunities in the quest to build more capable and efficient computing systems.

THE RISE OF ADVANCED AI

The rise of AI, particularly generative AI and LLMs, represents a significant catalyst for Arm’s growth in the datacenter. These workloads are computationally demanding and require specialized hardware to operate efficiently. We know that a truly excellent host CPU is necessary to get the full compute capability of any AI accelerator, whether that accelerator is a GPU or a custom designed one like Google’s TPUs, Microsoft Maia, or AWS Tranium and Inferentia.

While we are still only in the very early stages of AI development and mass deployment, we are constantly reminded of how much can change in short amounts of time. It’s only been a few years since ChatGPT was launched into the world and the underlying foundational models which were developed since then have used a range of workloads that balance between the CPU and the GPU. However, it appears we are on the cusp of an entirely new foundational model approach that will only deepen the complexity of AI inference and require even more purpose-built CPU architectures.

> "We are on the cusp of an entirely new foundational model approach that will only deepen the complexity of AI inference and require even more purpose-built CPU architectures."

These new advanced AI models are now capable of producing chains of thought before responding to queries, enabling more thoughtful and task-oriented inference. This "advanced reasoning" approach, demonstrated in OpenAI's 01 model, allows models to spend more time processing before generating outputs, resulting in improved performance. However, this comes at the cost of significantly increased computational demands, with some estimates suggesting up to 10 times higher inference compute costs compared to previous models.

The long-term vision involves enterprises employing numerous "agents" capable of handling complex tasks, necessitating both more sophisticated training and increasingly advanced inference capabilities. This trend extends to previously "solved" domains like computer vision (CV), where achieving human-level understanding and interpretation now requires much more intricate computational processes. The overall trajectory points toward AI systems that can engage in more nuanced, multistep reasoning, but at the expense of substantially higher computational requirements.

THE ROLE OF ARM

The Arm CPU development model, with its focus on power efficiency, scalability, and customization, is well-suited to provide the foundational technology needed to address these demands. Arm-based processors are well-equipped with highly efficient modern data types and vector operations that give them significant capabilities for performing inference in their own right – where latency or other concerns prohibit or make offload inefficient.

Arm technology is also unique in enabling adopters to explore higher-bandwidth and extremely low latency integration to GPU or NPU accelerators. Any accelerator will always still require CPU support for the surrounding business application, orchestration, and the various pre- and post-inference stages of the complete ML stack.

AI-DRIVEN DEMAND

AI workloads are changing the shape of the compute demands across multiple compute blocks. We firmly believe that these workloads will best be served by a symbiotic approach where the CPU, GPU, accelerator technologies, networking, and more, are all working in cohort. The flexibility of the Arm compute platform enables the following:

1. **Heterogeneous compute**: Arm-based CPUs are proving to be excellent companion processors for AI accelerators like GPUs and TPUs, managing data flow and general-purpose compute tasks efficiently and playing a critical role in dealing with any bottlenecks in the workflow.
2. **Inferencing efficiency**: While training large AI models often relies on high-performance GPUs, Arm power-efficient processors are well-suited for inferencing tasks, both at the edge and in datacenters.
3. **Scalability**: The Arm architecture enables seamless integration of CPUs, GPUs, and specialized accelerators, critical for optimized AI systems.

RAPID ADAPTATION TO AI REQUIREMENTS

Today, processor architecture plays a crucial role in determining the efficiency and effectiveness of AI systems. Arm has emerged as a key player in this space, offering a compelling combination of innovation, customization, and power efficiency. Its approach addresses three critical aspects of modern AI computing:

1. **Consistent innovation**: Arm's regular release of new CPU architectures

and supporting features, and focus on enabling custom silicon trends aligns with the evolving requirements of AI workloads.

2. **Customization potential**: As AI models grow in complexity and scale, Arm's flexibility allows for the creation of specialized solutions tailored to specific AI tasks.
3. **Power efficiency**: The power efficiency of Arm-based processors becomes increasingly valuable for managing the total cost of ownership of large-scale AI deployments.

THE DAWN OF A NEW ERA IN DATA CENTER COMPUTING

The confluence of evolving workloads, rapid innovation, and the demands of AI are creating a perfect storm for Arm's continued role in the datacenter, as well as the reliance on custom semiconductor solutions based on Arm from the largest hyperscalers in the world like Microsoft, Amazon, and Google. While x86 processors will undoubtedly continue to play a significant role, the trend towards Arm-based solutions is accelerating.

The maturation of the Arm software ecosystem represents a critical inflection point, removing one of the primary historical barriers to adoption. This software parity, combined with Arm's inherent advantages in power efficiency and customization potential, suggests that we are at the beginning of a new era in computing.

Arm's journey from mobile devices to the heart of the datacenter appears to be reaching a critical milestone, heralding a new era of diverse, power-efficient, highly customized computing solutions for the evolving demands of the digital age. This is leading to growing support from major players like Amazon, Microsoft, Google, and NVIDIA, who are all contributing to Arm's potential to capture a substantial portion of the datacenter market in the coming years. As the debate shifts from "x86 versus Arm" to a more nuanced discussion of total cost of ownership and performance efficiency, Arm is well-positioned to compete on an increasingly level playing field in the age of AI.

**Ben Bajarin is CEO and Principal Analyst with Creative Strategies, a technology analysis firm. This chapter is adapted from a white paper published by Creative Strategies and commissioned by Arm.*

CRACKING THE CODE: SOFTWARE'S SILICON CHALLENGE

By Mark Hambleton, SVP, Software, Arm

The silicon ecosystem for AI is in a constant state of evolution, shaped by the demands of the software ecosystem and rapid development of AI frameworks. At Arm, the focus has been on supporting silicon solutions that provide broad compatibility across all AI frameworks. This approach helps ensure that the Arm compute platform integrates seamlessly with diverse AI tools, enabling developers to leverage the performance and efficiency of Arm-based silicon, while maintaining flexibility in their software choices. This strategy reflects Arm's commitment to building universal, future-ready solutions that support the entire AI ecosystem.

At the same time, porting AI models to custom hardware can also be costly and time consuming. Developers frequently face the prospect of manually adapting software to fit specific hardware, further escalating costs and delaying deployment. This has led to a preference for targeting CPUs, given their widespread adoption and suitability for most AI inference workloads. CPUs also offer developers the consistency they value, avoiding the fragmentation and inefficiencies associated with bespoke hardware solutions.

> **"The future of AI development relies on the synergy between software and hardware."**

THE IMPORTANCE OF INTEROPERABILITY

Interoperability across AI frameworks is a critical concern for developers. Embedded and IoT devices, particularly those designed for edge AI inference, often need to function across multiple hardware platforms. This is why developers frequently default to CPU back-ends, as their ubiquity helps ensure broader compatibility.

Moreover, the lack of standardized practices within AI development continues to hinder innovation. Many AI models remain proprietary, limiting developers’ ability to address performance issues and slowing progress. Open standards play a vital role in overcoming these barriers, enabling developers to transition seamlessly between platforms. Initiatives like TensorFlow, PyTorch, and ONNX, alongside standards-based APIs, have helped standardize aspects of AI development, allowing data scientists to focus on training, quantization, and deployment that add value to the ongoing innovation of models.

EMERGENCE OF NEW DATA FORMATS

Over the past decade, the evolution of AI has introduced new data formats, transitioning from integer to floating-point representations and, more recently, to smaller floating-point formats. This evolution poses challenges for hardware, as many devices – particularly smartphones – continue to rely on integer formats supported by their NPUs. Keeping pace with such shifts requires hardware to adapt continually, underlining the importance of software compatibility and support.

CRITICAL DEVELOPER TOOLS FOR AI

Developers prioritize tools that provide insights into system behavior, particularly those focused on performance analysis and debugging. Identifying bottlenecks and optimizing models are essential steps in achieving efficient AI deployments.

Cloud-based development environments are also transforming the landscape, offering access to extensive computing resources necessary for training large-scale models. While AI inference often happens at the edge, cloud-based training has become indispensable for managing the computational demands of modern AI workloads.

RECOMMENDATIONS FOR SUPPORTING AI DEVELOPMENT

To foster a thriving developer ecosystem, companies must prioritize enabling hardware through robust back-end support and embracing ever-evolving AI frameworks. Some recommendations include:

- **Leveraging common tools**: Adopting widely used tools helps to simplify development and reduces fragmentation.
- **Providing pre-built back-ends**: Ensuring out-of-the-box compatibility for custom silicon helps accelerate adoption.
- **Maintaining upstream contributions**: Actively contributing to open-source frameworks helps ensure compatibility and avoids stagnation.
- **Keeping up with the frameworks**: As AI frameworks evolve rapidly, staying up to date is essential to remain competitive.

LOOKING AHEAD

The future of AI development relies on the synergy between software and hardware. By embracing open standards, prioritizing interoperability, and providing robust tools for developers, the industry can help accelerate the pace of innovation and ensure that silicon and AI frameworks deliver on their full potential. The next wave of AI silicon solutions will depend on a unified, developer-centric ecosystem that values consistency, accessibility, and forward-thinking adaptability.

THE FUTURE OF SILICON: WHERE INNOVATION MEETS AI

By Kevork Kechichian, EVP, Solutions Engineering, Arm

The semiconductor industry stands at a remarkable time in its evolution. Traditional Moore's Law scaling is giving way to new approaches like custom silicon, CSS, and chiplets. Meanwhile, AI workloads are placing unprecedented demands on computational power and efficiency.

In this report, we've highlighted several key trends shaping our industry's future. First, the push toward more specialized architectures tailored for specific workloads, particularly in AI. Second, the critical importance of power efficiency as computation becomes more intensive. Third, the evolution of security to address emerging AI-powered threats. And fourth, the essential role of software ecosystems in unlocking silicon's potential.

The challenges ahead are significant, but it's why we do the work necessary to confront them. Power delivery and thermal management grow more complex as we move to advanced packaging technologies. Memory bandwidth must scale to meet AI's voracious appetite for data. Security threats continue to evolve alongside AI capabilities.

Yet these challenges are driving unprecedented innovation across the ecosystem. We're seeing closer collaboration between IP providers, foundries, and system integrators. New approaches to memory architecture and power management are emerging. Standards are evolving to enable truly modular chip design.

> **"The future of computing – and particularly AI – depends on our ability to continue pushing the boundaries of what's possible in silicon."**

Most importantly, we're witnessing a fundamental shift in how chips are designed and manufactured. The historical separation between design and manufacturing is dissolving as new process nodes require deeper collaboration. This new era demands creativity, systems-level thinking, and an unwavering focus on efficiency. We think the ecosystem is more than up to the task.

The industry's response to these challenges will define computing's future. Success requires embracing complexity while finding ways to manage it effectively. It demands new tools, methodologies, and unprecedented cooperation across the ecosystem.

At Arm, we recognize the critical role that efficient and scalable computing architectures play in realizing AI's potential. Through continued collaboration with partners across the semiconductor ecosystem, we're focused on addressing the fundamental challenges of power efficiency, security, and performance that will shape the future of computing.

The semiconductor industry's ability to innovate and adapt to AI's demands is crucial in the years ahead. By working together as an ecosystem, we can create the technological foundation needed to unlock AI's transformative capabilities, while managing its computational costs and complexity.