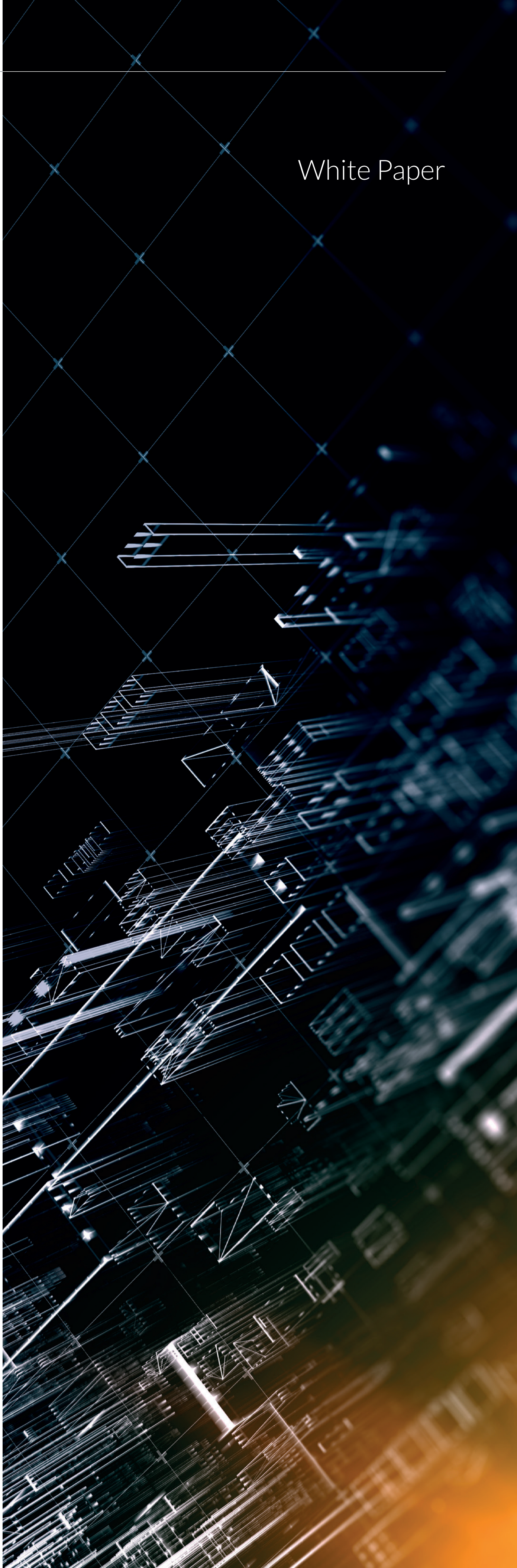


White Paper

Arm Total Compute

Engineering for tomorrow's workloads

Index

2 Introduction

3 Digital Immersion

3.1 Wide-ranging New Use Cases and Experiences

3.2 Meeting Digital Immersion Requirements

4 A Shift in Approach to System Design

5 The Three Pillars of Total Compute

5.1 Compute Performance

5.2 Security

5.3 Developer Access

6 A New Direction for Arm

2 Introduction

Today, each of us carries a computer in our pocket that is millions of times more capable than the guidance systems used to put man on the moon. Smartphones have transformed society, bringing the ability to not only communicate but to learn, share and collaborate to solve problems for billions of people the world over.

The past decade has seen the smartphone grow to represent the pinnacle of mainstream technology. It has enabled the democratization of efficient compute and revolutionized communication, knowledge sharing, entertainment and the management of personal data—all within a lightweight and intuitive form factor.

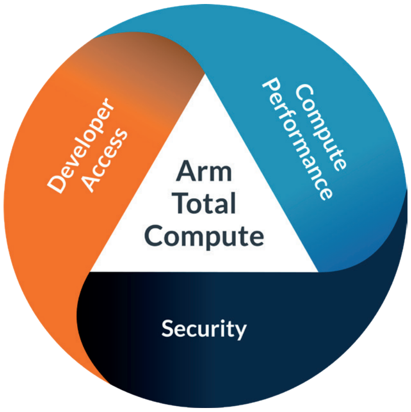
But just as readily as it has defied expectations, the smartphone has raised them. Consumers are embracing new applications and immersive experiences as quickly as they appear. As app designers race to deliver the 'next big thing', the performance-per-watt envelope is being stretched ever closer towards the limits of even the most high-end device capabilities. Many of today's most popular smartphone apps rely on multiple IP components, combining data from the device's camera, accelerometer and other sensors in real-time. The processing of the resulting data can now be done on the device via artificial intelligence (AI) algorithms running on the CPU, GPU or a dedicated neural processing unit (NPU).

The delivery of even the most recreational applications—manipulating selfies in real time, for example—relies on effective interplay between different hardware components within a device. And with technologies such as XR (augmented reality (AR) and virtual reality (VR)) soon to become more mainstream, dramatic increases in overall performance, performance per watt and efficiency will be required to bring new, compute-intensive experiences to life—all within the same limited thermal envelope.

Introducing Total Compute

Each new iteration of Arm IP pushes the limits of performance for power-constrained devices. Yet there is a hard limit: one defined by a device's thermal and power envelope, and we cannot simply continue to push. As consumers seek ever richer and more immersive experiences from their smart devices, the way we design compute systems must change. Arm Total Compute takes a holistic, solution-focused approach to SoC (system on a chip) design, moving beyond individual IP elements to designing and optimizing the system as a whole.

Total Compute is a shift in approach in how Arm thinks about and designs its IP. It will ensure that tomorrow's devices will be fully capable of processing the advanced, complex and demanding multi-domain workloads of the future. Alongside performance improvements, it will enable more robust security than ever before through multi-layered solutions across the system, from the actual device to cloud services, providing end-to-end personal data protections.



And, fundamentally, it will enable developers to access and use all of a device's performance and security features, in order to design and build more sophisticated, efficient and cost-effective solutions across multiple compute platforms that meet future market and device requirements.

It is these three pillars combined—**Compute Performance, Security and Developer Access**—that make Total Compute.

3 Digital Immersion

Today's consumer devices are the hub of everything we do as people, from everyday productivity tasks such as communicating, shopping and banking through to more complex workloads like video streaming, gaming and XR. These experiences are what Arm refers to as 'digital immersion'. They will grow ever richer, safer and more fulfilling as key enabling technologies within the Fifth Wave of Computing—5G, AI and the Internet of things (IoT)—reach maturity.

Digital immersion relies on a relentless drive for compute performance and communication bandwidth, coupled with the responsibility we owe consumers to keep their data secure.
At the same time, we need to deal with physical constraints, such as form factor and battery life.

While the smartphone has defined today's digital immersion experiences, these are being continuously re-defined by the capabilities and needs of other consumer devices.
The industry is creating a wide range of new use cases and experiences for consumers.

3.1 Wide-ranging New Use Cases and Experiences

The drive to enable more complex use cases is affecting a wide variety of devices in different market segments, from laptops and tablets to the smart home and automotive platforms.
For example, the rise of streaming services such as Netflix, Amazon Prime and Disney+ as the primary content delivery mechanism in many homes has resulted in smart TVs that are orders of magnitude more performant than their counterparts a decade ago.

Yet while many of today's smart TVs may pack a powerful Arm processor behind their screens, there are always more complex use cases on the horizon. Next generations of smart TVs are likely to need far greater compute performance as consumers seek greater resolutions (8K and beyond), higher frame rates and entirely new, intensive use cases, such as cloud gaming.

> “The smartphone ecosystem built on Arm technology has grown from about 500 apps in 2008 to 8.9 million apps in 2019.”

Moreover, a number of television manufacturers are using AI to deliver an optimized experience: picture quality enhancements such as upscaling, alongside voice assistants, now require compute systems within smart TVs to be capable of processing AI workloads.

AI capability is already a key requirement in many smart devices. Computational photography, for example, employs machine learning (ML) to identify people, objects or scenes, and is also used for optimizing image quality, and zoom on cameras. It generates depth data from dual cameras and enables real-time manipulation and recognition: whether that is adding background blur or a complete background removal. It is now commonplace in smartphones, but also finding its way into everything from industrial IoT video sensors to video baby monitors.

While AR has seen slow uptake from consumers, the technology is growing in popularity in business sectors, such as retail, design, manufacturing, maintenance, medical training and construction. AR renders realistic 3D objects in real-time, then combines ML algorithms with camera and gyroscope data in order to position these virtual items within the physical environment.

Consumer interest may pique as new consumer wearables for both AR and VR experiences come to market, heralding a new wave of digital immersion. While VR is likely to remain fundamentally a gaming platform, AR has the potential to deliver an always-on experience, adding context to everything we see and do. However, the fundamental challenge for AR and VR is packing the high compute performance and appropriate power supply needed into a compact, lightweight form factor that can be worn comfortably for long periods, such as future smartglasses.

Moreover, smaller wearables, such as smartwatches, present another important opportunity, as they make the transformation from fitness gadgets to fully-fledged healthcare devices.

Compounding this further is a brand-new wave of connectivity: 5G. The increased network speed is delivering existing use cases and experiences, such as web browsing and video streaming, far quicker and more conveniently than ever before and with lower latency. However, as more data and information are being captured on the device, it is adding to the already complex and compute-intensive workloads of the future.

As part of the drive towards greater digital immersion, it is also important to look at the smartphone developer ecosystem. This ecosystem built on Arm technology has grown from about 500 apps when App stores first emerged in 2008 to 8.9 million apps and 204 billion app downloads worldwide in 2019.

All of these applications are contributing to a deluge of diverse, digital data that is doubling every two years.

More active smartphone users than ever means more performance-hungry applications running and more data. Users are also quite discerning: they want a seamless, holistic experience. They want choice in services, and they want those services to go beyond their smartphones and onto a diverse range of devices that meet their personal preference. With increasing knowledge and intimacy of our personal data trails, it becomes easier to target consumers based on their preferences and interests. The current growth of data deluge means greater scope for influencing the decision-making process for every individual, which has wider societal implications. This is why a system-wide security foundation for protecting both data and applications is key, as lack of trust in devices will be an inhibitor for future digital immersion experiences.

3.2 Meeting Digital Immersion Requirements

With Moore's law slowing down, we cannot just keep on optimizing the individual compute components that make up an SoC, such as CPUs, GPUs and NPUs. New workloads are increasingly complex and require an optimized compute system solution to provide the best user experience. Designing compute components in isolation from each other will not be sufficient to enable these new complex workloads. We need to go beyond and take a system-level solution view of the SoC, putting digital immersion use cases and experiences at the heart of the designs. Future compute platforms need to deliver the best user experience as one efficient compute system. For example, Arm is designing technology to enable our Mali-GPUs to work better with our Cortex-A CPUs in a system to deliver a better AAA gaming experience while maximizing battery life.

Each new iteration of Arm IP pushes the limits of performance for power-constrained devices. Yet as AI becomes ubiquitous and begins to enable ever more complex use cases, such as AI camera and computer vision, it will require a far more capable class of device.

However, it is not just the compute IP. Engineers and developers require high-performing security features, and software and tools that are easy to implement and transcend across the entire ecosystem. For security, this is protection across hardware, software, operating systems, applications, and services. Meanwhile, developers need good performance out-of-the box and, when there is a problem, make the way of solving it consistent. This means easy access to the performance and security features, with accessible and high performing software and tools to boost their development process.

As a result, a more consistent, scalable, trusted, secure and stable technology stack is needed across the entire SoC development framework, from hardware to software. This requires a shift in approach to SoC design that is relevant to the entire ecosystem, enabling engineers and developers to leverage solutions-based, heterogeneous compute IP blocks to free them to focus on the system or end-application as a whole.

We must confront this SoC design challenge head-on if we are to enable new, complex use cases and ever-greater digital immersion. If this was a desktop computer, we might simply add a more powerful processor, graphics card, extra cooling systems and a more powerful power supply. Yet within the mobile device sector we do not have this luxury. Given the energy constraints of such small, lightweight form factors, we must adopt a new approach.

At Arm, we call this new approach Total Compute.

4 A Shift in Approach to System Design

To date, Arm has focused on developing and optimizing discrete IP 'components' for our silicon partners to then integrate them into their complex system on a chips (SoCs).

However, continuing to design and optimize discrete compute components in isolation from each other will not be sufficient. We need to adopt a holistic, solution-focused approach to SoC design—moving beyond individual IP elements to optimizing the system as a whole.

Total Compute sees Arm move to delivering complete SoC solutions optimized for specific use cases.

In this way, we can deliver best-in-class performance and efficiency on next-generation devices. Having access to a common underlying architecture makes it easier for our partners and the overall software ecosystem to deliver products to market faster and with performance built around leading-edge digital immersion experiences.

While we are very aware of the significant engineering challenges from SoC design, focusing on AR and the different use cases and experiences it enables is a good example of how Total Compute could work in practice. The CPU is driving performance in a power efficient manner. The GPU is driving the graphics. AI is being used for detection—from the user's location to specific objects and landmarks. Then, we need to bring this IP together to work seamlessly in the system. This is why we need System IP, such as interconnects and controllers, to help build better systems focused on low-power constraints. SoC Designers can then optimize their system solutions and dedicate more time on differentiation, developing new ideas, technical improvements, and solutions for their customers.

Alongside the need for greater performance and developer access, we also need to deliver these experiences securely. Not just on the device itself, but across the entire device ecosystem. Arm is building security into all aspects of our Total Compute offering, from solutions on the device to support throughout the entire device ecosystem. This is taking Arm's commitment to secure endpoint devices (the label Arm applies to physical devices at the edge of the network) one step further, providing 'defense in depth' security for hardware, firmware, software, operating systems, applications, and services.

Strong foundations

While this is a shift in approach in every sense, this is not Arm starting from scratch. We are building upon long-established technologies, such as Arm's power control frameworks and microcontroller-controlled power management schemes, to achieve optimal performance for a given use case and power budget. The investigation includes looking at power delivery for particularly demanding use cases.

> “It is becoming mandatory to optimize at a system level across all IP boundaries (CPU, GPU, NPU, memory, etc.), while removing any bottlenecks in the system.”

Why A Solution-Focused Approach?

- Complex use cases require greater performance, but doing so implies that all the IP work together seamlessly. Simply increasing performance at an IP level is not enough. It is becoming mandatory to optimize at a system level across all IP boundaries (CPU, GPU, NPU, memory, etc.), while removing any bottlenecks in the system. This ensures that the full solution creates a virtuous circle to deliver and sustain the required performance for these complex use cases in the same power envelope. The sum (solution) is greater than its (IP) parts. The SoC itself can become a bottleneck since interconnect and bandwidth requirements between powerful, diverse components of IP can introduce latency issues.
- Amid increasing security and data privacy requirements, diverse IP components in an SoC can lead to engineering challenges when trying to create a consistent security approach with cohesive end-to-end security features.
- It provides more accessible and performant IP and software and tools to developers, enabling a seamless and quicker development process and more immersive applications for consumers.
- For software developers, multiple hardware architectures employing a variety of differently sourced IP components can mean time-consuming and costly software development and encourage a fragmented software ecosystem.

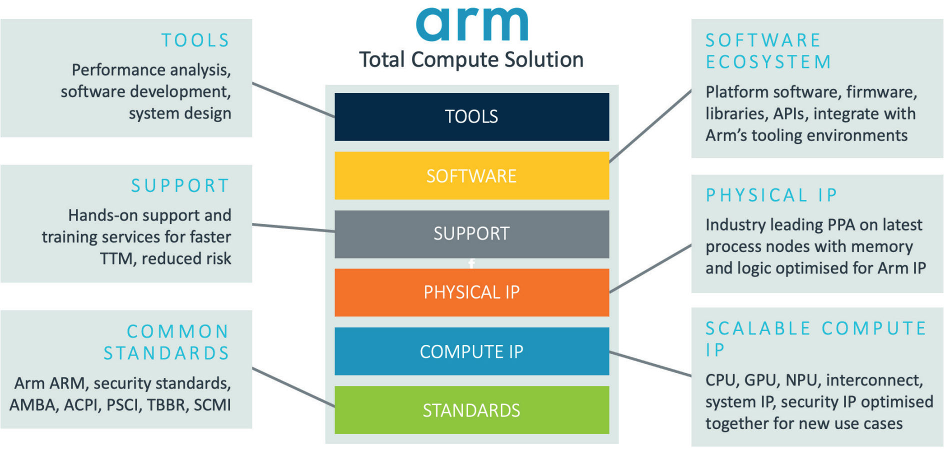
A crucial part of the SoC design before it is created in silicon is physical design. Arm proposes a range of physical IP: among other features, Arm POP IP acts as the bridge between Cortex-A CPU and silicon process technology. It enables industry leading PPA on the latest process nodes, with memory and logic optimized for Arm IP and within a power envelope, adding to the Total Compute compute performance pillar.

POP IP is core-hardening acceleration technology designed to get products to market quickly, lowering technical and project risk. This optimization across IP boundaries underlines how we aim to solve the workloads of tomorrow, as the requirements for more edge compute on devices are changing.

However, as the complexity of devices increases, so does the requirements for security. Achieving secure devices requires a symbiotic relationship between hardware and software where security can no longer be an afterthought. This is built on an expanding portfolio of secure IP and software within a collaborative ecosystem.

Key Components of Total Compute

- Scalable compute IP optimized to work better together to address new use case requirements.
- Common standards, enabling consistent and easily deployable technologies across the ecosystem.
- Software components co-designed alongside hardware that are ready to use out of the box to accelerate the availability of technologies on the devices.
- Tools to support system design, software development and performance optimization.
- Physical IP, including POP components, which leverage Arm expertise in bringing competitive PPA and accelerate the time-to-market of silicon, despite the increasing complexity of process node and Arm IP.
- Support and training to accelerate ramp-up of engineering teams on the latest solutions and technologies, reducing risk and time-to-market.



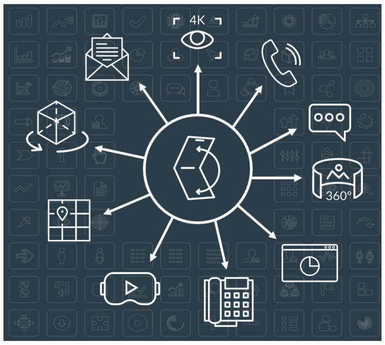
arm
Total Compute Solution
TOOLS
SOFTWARE
SUPPORT
PHYSICAL IP
COMPUTE IP
STANDARDS
TOOLS
Performance analysis, software development, system design
SUPPORT
Hands-on support and training services for faster TTM, reduced risk
COMMON STANDARDS
Arm ARM, security standards, AMBA, ACPI, PSCI, TBBR, SCMI
SOFTWARE ECOSYSTEM
Platform software, firmware, libraries, APIs, integrate with Arm's tooling environments
PHYSICAL IP
Industry leading PPA on latest process nodes with memory and logic optimised for Arm IP
SCALABLE COMPUTE IP
CPU, GPU, NPU, interconnect, system IP, security IP optimised together for new use cases

This optimization across IP boundaries underlines how we aim to solve the workloads of tomorrow, as the requirements for more edge compute on devices are changing.

COMPUTE PERFORMANCE

SECURITY

DEVELOPER ACCESS



5 The Three Pillars of Total Compute

5.1 Compute Performance

A solution-based approach for accelerating performance growth

While Moore's Law butts against a performance ceiling dictated by physical design, the need for exponentially greater performance is increasing. New, more complex workloads require ever more performant hardware within the same limited thermal envelope.

The challenge with integrating diverse components of IP into the SoC is that increasing the active die area can lead to an increase in thermal and power budgets. This is why the emphasis on the entire system is needed, so each IP block is developed with a common underlying architectural approach for performance, efficiency and data exchange.

Total Compute aims to deliver greater performance at best-in-class power efficiency and area cost through adopting a holistic, system-level approach to SoC design. It manages the demands of increased performance and bandwidth with features like finer-grained power management, the use of system caches and, of course, security and privacy implementations.

Moving beyond optimizing individual IP and taking a system-level solution view of the entire SoC means focusing on the use cases and experiences on the next-generation devices and ensuring that the entire system works together seamlessly to provide maximum performance and efficiency.

So how do we kickstart the development and optimization of a use case driven SoC system? First, it starts with use case selection and analysis at the start of our Total Compute development cycles and then the de-composing of these into key workloads. For example, for security, ML, or compute requirements for specific applications, such as 8K, AAA gaming, high refresh rates, etc., for a given power envelope. The resulting analysis then drives architectural choices across IP compute domains and software frameworks. The workloads enable the analysis of the requirements in terms of UX, platform (performance vs. efficiency for each sub-system items, such as GPU, CPU, memory bandwidth, etc.), and technology (cache, etc.).

A fundamental part of Total Compute system design will be analyzing how interconnecting data and compute is best deployed between the different IP components and compute domains. The integration of data compression between multiple pieces of IP is key to success.

Through evaluation and system analysis, we then optimize for these in hardware and software at the system level, developing powerful yet efficient micro-architectures that support tomorrow's workloads and not yesterday's benchmarks.

Together with the appropriate use of data compression, memory bandwidth usage can be reduced, which in turn can enable more intensive use cases, and reduced power consumption. For example, compression may overcome an external memory bottleneck to enable higher resolution video or higher frame rates. This can lead to power savings for the same workload or can enable such features on a lower tier device with lower external memory bandwidth.

Take the mobile gaming experience as an example. It is very difficult to translate the improvements in synthetic benchmarks for individual IP components into meaningful performance or efficiency improvements once these components are combined together in a system. Through Total Compute-centric system analysis of a high-performance gaming workload, it is possible to identify new metrics for the compute platform.
For example, achieving a certain number of frames per second (FPS) within a specific sustained power budget.

It then becomes possible to look at new capabilities and features needed in each compute component to improve the overall gaming experience. In essence, using a top-down system view to evolve the compute IP components becomes a key emphasis for the Total Compute approach.

In all use cases, system performance improvement is a multi-dimensional metric driven by use cases and their requirements and not a fixed benchmark. New capabilities and performance improvements are assessed within a specific context, be it sustained performance for gaming, flexible and scalable ML capability, or laptop-class performance at smartphone levels of power efficiency. These include:

- Optimal, scalable, heterogeneous compute, laptop class performance at smartphone power efficiency
- Immersive and interactive UX through enhanced imaging
- Sustained performance for gaming
- Powerful, flexible ML capabilities across compute clusters

As the requirements for more edge compute are changing and the complexity of devices increase, so are the requirements for security. Achieving secure devices requires a symbiotic relationship between hardware and software where security is no longer an afterthought. This is built on an expanding portfolio of secure IP and software within a collaborative ecosystem.

5.2 Security

Creating a secure foundation through hardware and software

For the past two decades, Arm has pioneered security in smart devices. Arm TrustZone technology is used in billions of mobile devices worldwide today. This is being expanded further into smart watches, DTVs, connected home devices and now the next generation of laptops.

Many of the most useful applications on these devices rely on our personal data. When our devices 'know' us, they are far more useful—suggesting what words we might type next, identifying our faces for security and reminding us of appointments. However, the more personal information that users offer devices, the greater the opportunity for those looking to exploit it.

Through Total Compute, Arm is looking to address both security and privacy. Only with the appropriate security architecture and robust implementation can manufacturers and application developers address the privacy concerns of consumers, which is key to gaining and maintaining consumer trust:

1. By implementing best in class security foundations from silicon to firmware, device manufacturers can prevent adversaries from exploiting vulnerabilities. This will assure consumers that private data, financial information and other common personal information are kept safe.
2. Consumers are rightly concerned about their data being used without their consent, so we need to protect the processing of their data on the device. This means protecting it in a secure processing environment to minimize data collection and leakage.
3. We also need to consider new emerging security challenges, with the convergence of big data and AI now being deployed everywhere.

Total Compute aims to lay the security foundations to enable our partners to realize the revolutionary benefits of AI everywhere without compromising the privacy rights of individuals. It is key for our solution to offer trustworthy and easily deployable security capabilities, so that our partners can protect the data and identity of users during digital immersion experiences.

As a result, we are rethinking the security architecture, from diverse and expensive security solutions to standardized and scalable solutions, which can address the diversity of issues and needs across all market segments.

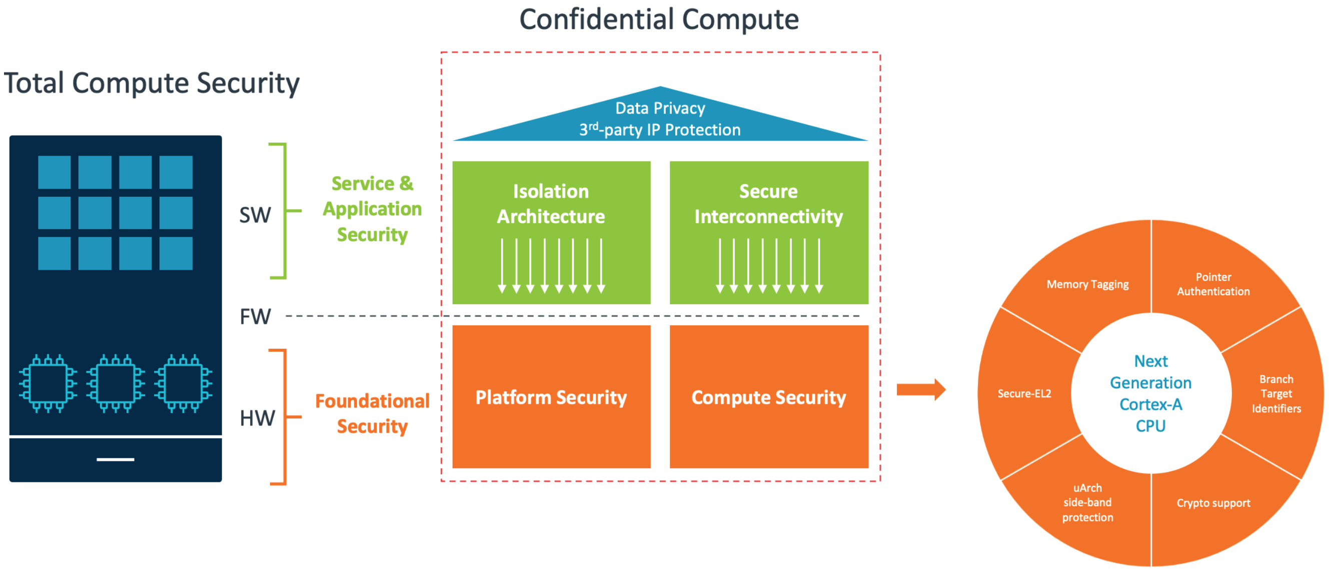
Robust device security is critical to the success of Arm and our partners. Under the hood, consumer devices are composed of hardware components, firmware to run those hardware components, and from hundreds of thousands to millions of lines of software code. We have seen that all of these can be vulnerable to different kinds of hacking attempts and need to be secured. The combinations and implementations are numerous and can be fragmented, making it challenging to do security patches quickly and broadly. It is an endless loop of finding holes and patching holes.

If we are to ask users to trust their personal data to a greater diversity of devices then security must be a priority, not an afterthought. In a 2020 survey by Northstar Research[1] commissioned by Arm, over a third of respondents stated that should a device they use be compromised and their personal data stolen, they would permanently stop using that category of device. A further third said they would switch to a competing brand. As well as meeting the expectation of consumers, we also have an obligation to the industry. Arm is committed to building technology that helps deliver better integrated security, which is easier to implement at the IP level, as well as in firmware, platform software and application levels.

Total Compute takes our commitment to security one step further and is at the foundation of this new approach. It enables us to protect personal data and the applications that handle them through defense in depth' across hardware, software, operating systems, applications and services by ensuring end-to-end system level security protection, built from the ground up. This means we start with the architecture, the base design for all compute. Our objective is to mitigate vulnerabilities in the device before mass production and greatly reduce the attack vectors in consumer devices. In practical terms, this approach resolves into a focus on two key aspects of security:

- Foundational Security for hardware (HW) and firmware (FW)
- Application & Service Security for software (SW)

As the complexity of the next generation of consumer devices continues to increase, security must move beyond standalone solutions that only protect one aspect of a device. We need multi-layered solutions where there is system cooperation, from device hardware all the way through to cloud services that utilize the plethora of personal data on our devices. Crucially, security through Total Compute moves beyond solutions that protect individual components of IP to a holistic solution that protects the system as a whole. This is being achieved through Confidential Computing[2] , the next evolution of secure computing which makes all aspects of the ecosystem less vulnerable to attack. For Arm, Confidential Computing covers everything from the security architecture to application software and all the services that consumers use on their devices, offering defense in depth throughout.


Confidential Compute

Total Compute Security

“Up to 70 percent of bugs are memory safety errors.”

Memory Tagging Extension (MTE)

At a hardware level, Arm enables innovative security features such as Memory Tagging Extension (MTE) to tighten security vulnerabilities that can occur in memory subsystems. This is needed because software vendors report that vulnerabilities due to violations of memory safety account for most of the security issues in their products. In fact, up to 70 percent of bugs are memory safety errors.

MTE makes detecting memory safety violations easier and more efficient across the entire ecosystem. Immediately they can address this class of bugs before silicon vendors send the SoCs to the OEMs. Next, OEMs benefit because MTE further helps them detect additional memory safety bugs before the devices go to mass production. Before hardware availability, tools like HWASAN are available in Android to support code checking. Once in the market, the OSV and application developers can use MTE enabled devices to find their own buffer overflows and heap corruption in their code.

Google has already pledged to support MTE through Android 11 in anticipation of future hardware. On the Android Source website, Google revealed that starting in Android 11, for 64-bit processes, all heap allocations will have an implementation defined tag set in the top byte of the pointer on devices with kernel support for Arm Top-byte Ignore (TBI). TBI is a feature that is available for 64-bit code in all Armv8 AArch64 hardware, which means that the hardware ignores the top byte of a pointer when accessing memory. This is necessary for future hardware with MTE support.

Interoperable standards for device attestation

However, the role of Total Compute in security goes far beyond hardware, so working with external industry consortiums is key. For example, we are working with the Confidential Compute Consortium to establish interoperable standards for device attestation. This is where evidence of the authenticity and trustworthiness of the device and its firmware can be obtained and verified, so that developers can establish trust in the platform before running their application. For software developers, not only does it represent a step change in security, but it will make secure OTA (over-the-air) updates of their applications significantly easier and reduce engineering costs.

Furthermore, we are working to provide developers with a secure and safe space to play by protecting the user’s personal data and their business assets. If we are to ask users to trust their personal data to a greater diversity of devices then security must be a priority, not an afterthought. Security protections for data that is being stored or transferred are well understood and widely implemented through encryption and protocol design. However, when the data is being processed or in use, the security protections are less understood and therefore could be at a greater risk. Total Compute provides a secure and accessible platform for developers to protect both their applications and the personaldata that the applications use.

These approaches will ensure Arm and its partners meet future and emerging needs for protecting consumer data, while still allowing data analytics and insight. Digital immersion use cases often require multiple parties to combine, analyze and learn from sensitive data without exposing the data or ML algorithms to third parties. Often referred to as multi-party computing, decentralized learning, or privacy-preserving analytics, the security challenge being addressed by Total Compute is providing a Confidential Computing environment to perform computations in collaboration with third parties, while minimizing the risk of data leakage and preventing data theft. These are some of the emerging developer needs, so they can create new immersive experiences enabled by the security foundation in Total Compute.

5.3 Developer Access

Making more performant software and tools more accessible to developers

There are 23 million developers in the world[3], all focused on delivering the best experiences. Arm already invests heavily in software and tools to assist developers, ensuring that everything runs better, faster, more efficiently, more consistently and more securely on Arm technologies across the entire compute system. These allow developers to explore, select and check IP choices, optimize and analyze performance. Total Compute takes this strong commitment even further.

The final pillar of Total Compute is all about developer access. The idea here is that the two previous pillars are not going to come to fruition unless we provide the tools, insight and developer enablement to unlock the capabilities built into the platform.

As an example, it is good to take a look at the Android ecosystem. The diversity of the Android ecosystem is its greatest strength, but also a significant challenge for anyone who wants to deliver a consistently great experience with broad reach. As of mid-2020, there were 440 different Android devices with published Geekbench 5 scores, which are a lot of different configurations, form factors and performance points that developers have to consider. So when it comes to performance and efficiency there are two messages from developers that we hear loud-and-clear:

1. Give us great performance out of the box on as many devices as possible.
2. Make the process of getting the best possible performance as simple and consistent as possible.

Developers within the Arm ecosystem have traditionally needed to learn different tools for each of the various components on an SoC. For those developing for multiple SoCs, each with different combinations of IP, this can be a complicated, time-consuming and costly process. Furthermore, any SoC that combines new and legacy IP components stands to introduce further issues in code fragmentation.

Total Compute will provide developers with frameworks for programming, debugging and analyzing the SoC holistically.

Better tools, better support

There are two key approaches for improving developer access through Total Compute. The first is better Arm tools. The second is better support for Arm products in the toolchain for developers to use. The overall aim is to ease deployment across platforms with the latest security features and unlock full performance by enabling pre-shipment testing, in-field capabilities and faster debug.

Arm already invests heavily in software and tools to assist developers. These allow developers to explore, select and check IP choices, and optimize and analyze performance. Total Compute takes this commitment even further by making Arm technology more accessible, so developers do not need to hunt for tools optimized for each architecture or various, multiple discrete IP blocks they are using. For those developing for multiple SoCs with different sets of IP, this can be a complicated, time-consuming and costly process. Any SoC that combines new and legacy IP blocks stands to introduce further issues in code fragmentation.

Defragmentation

The open approach of Trusted Firmware provides SoC developers and OEMs with a reference trusted code base complying with the relevant Arm specifications. Its adoption by as many companies as possible will mean that defragmenting the software ecosystem through Total Compute ensures that operating systems, runtimes, and application platforms all work best on Arm. This leads to a quicker time-to-market for partners and more of a focus on improving the security and performance of the entire system, which adds further value to our partners.

Tooling and support

Arm's tooling and additional support helps developers not only understand how different workloads function across different IP in order to identify where bottlenecks occur within the system, but also how to implement their solutions efficiently. Through Total Compute, developers will be able to conduct full-system performance analysis for workloads ranging from productivity to gaming and ML via a range of tools through two suites: Arm Development Studio and Arm Mobile Studio. This will improve the usability and simplicity of our software and tools for developers when they are working on their applications, ensuring that this work can be done in a seamless, secure and reliable way. Arm Development Studio is a suite of tools for interrogating hardware counters and further optimizing systems for complex and graphics-based applications. Performance analysis provides further evidence of developer access. This allows developers to understand how workloads function across the different technologies and identifies where the bottlenecks occur within the system. Further supporting this is Performance Advisor, part of Arm Mobile Studio. This is a new Arm tool that generates easy-to-read performance analysis reports which serve a broad range of developers and artists in the mobile gaming market with a simplified visual representation of graphics issues and bottlenecks. The reports are based on the rich technical performance data gathered from platforms with Arm Cortex CPUs, Mali GPUs and Ethos NPUs. The professional edition of Mobile Studio includes a Continuous Integration feature enabling automated performance analysis and hourly regression tests across multiple devices.

Arm also offers software frameworks and compute libraries to improve performance across different compute domains. Arm NN is a great example of one such software framework that provides ML performance benefits across all our IP. It is being increasingly used by developers looking to make ML improvements across their applications.

Ecosystem

The sheer breadth and scale of the Arm ecosystem is enormous. As a result, our work with partners in the ecosystem cover a wide range of activities, from supporting silicon design to optimizing games. The developer community is a significant part of this ecosystem support. Total Compute ensures that Arm's IP incorporates high quality tools and compliance suites for ecosystem enablement at every stage in the development cycle. For example, at the earliest stages in system design, our partners can manage their virtual prototyping using Arm Fast Models[4] , often in conjunction with EDA verification and validation tools. Fast Models facilitate the seamless co-development of hardware with software before the availability of new silicon, accelerating the time to market for system supporting software and ensuring a fast track for market adoption.

“All new apps published on Google Play Store are required to support 64-bit architectures.”

In many instances, the availability of critical tooling is provided in collaboration with various partners in the ecosystem. One example of this in action is through Arm's partnership with Unity. This integrates Arm's performance analysis capabilities with Unity tools to improve game efficiency and make workloads more accessible to developers on the game engine. We are also working with Unity on Burst compilers, which enables developers who are building projects with Unity to take advantage of the Arm Neon instruction set. This improves the performance of Unity projects supported by Arm architecture on Android devices.

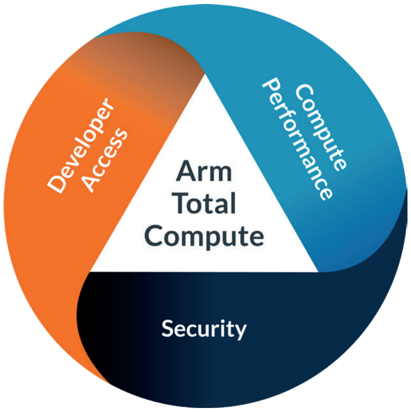
For Unity developers worldwide, the benefit will be the ability to bring more efficient applications to market quicker and for the broadest consumer set. For consumers, this means having access to higher-performing applications, delivering more complex digital immersion experiences.

We are also bringing our expertise to other tools from partners in the ecosystem. Besides our work with Unity, Arm is a launch partner of Google's new Android GPU Inspector (AGI). This is another tool designed to help game studios squeeze maximum performance from GPUs for more immersive mobile games on Android. With AGI, graphics programmers can visualize the GPU utilization in mobile devices that run on Arm Mali GPUs and focus their efforts where it matters the most.

Finally, we are continuing developer engagement and support around 64-bit. As of August 1, 2019, all new apps published on the Google Play Store were required to support 64-bit architectures.

For developers, the benefits of 64-bit are substantial. Migrating applications to 64-bit provides performance gains (up to 20 percent for some workloads) and additional security protections. These benefits directly affect the user experience. For example, through our partnership with Unity, we were able to run analysis across a range of content on Unity 2018 and saw an overall frame rate uplift ranging from 9.5 to 16.7 percent on 64-bit applications. From 2022 onwards, all Arm big-core CPUs will be 64-bit only.

6 A New Direction for Arm



Total Compute is the technology foundation to deliver advanced and secure digital immersion experiences on the next generation of devices. It is the next evolution in how we develop Arm IP and supply the most value to our partners. It delivers consistent solutions with components that have been co-designed according to the three interconnected Total Compute pillars.

This holistic solution aligns with growing partner demand to approach SoC design from a system-wide perspective. It addresses the current real-world compute challenges that exist today, but also ensures that our partners are ready for the next wave of digital immersion.

Future consumer devices will require greater performance to deliver complex digital immersion workloads at a better efficiency. And this performance will have to be delivered on current and future form factors, which all have area and thermal constraints. At the same time, there needs to be improved security, offering 'defense in depth' across all aspects of the compute system. Finally, all these performance and security benefits need to be accessible to developers, so they can create applications that provide these digital immersion experiences.

Total Compute provides the assurance that whatever device is developed—be it a smartphone, smart home device, laptop, tablet, XR headset or an entirely new device category yet to appear in the market—it will be built on a robust, performant and secure framework that is optimized to deliver the best-in-class experience.

The future of compute promises to be incredibly exciting, with digital immersion continuing to accelerate across all consumer devices, delivering immersive experiences. Through the three pillars of compute performance, security and developer access, Total Compute will enable a quicker, safer, more advanced and more seamless compute future that partners and consumers are demanding.

To find out how Total Compute can help you innovate quickly and get to market, visit www.arm.com/totalcompute.

Footnotes

[1] 2020 survey by Northstar Research: https://www.arm.com/blogs/blueprint/read-arm-2020-global-ai-survey

[2] Confidential Computing Consortium: https://confidentialcomputing.io/

[3] Source: Evans Data Corporation, 2018

[4] Fast Models are functionally accurate programmer's view models of Arm CPU and System IP so you can develop software targeting the latest Arm IP well before hardware implementations are available. They also add value as easily deployable and automatable targets for continuous integration and validation.

All brand names or product names are the property of their respective holders. Neither the whole nor any part of the information contained in, or the product described in, this document may be adapted or reproduced in any material form except with the prior written permission of the copyright holder. The product described in this document is subject to continuous developments and improvements. All particulars of the product and its use contained in this document are given in good faith. All warranties implied or expressed, including but not limited to implied warranties of satisfactory quality or fitness for purpose are excluded. This document is intended only to provide information to the reader about the product. To the extent permitted by local laws Arm shall not be liable for any loss or damage arising from the use of any information in this document or any error or omission in such information.