ARM

White paper

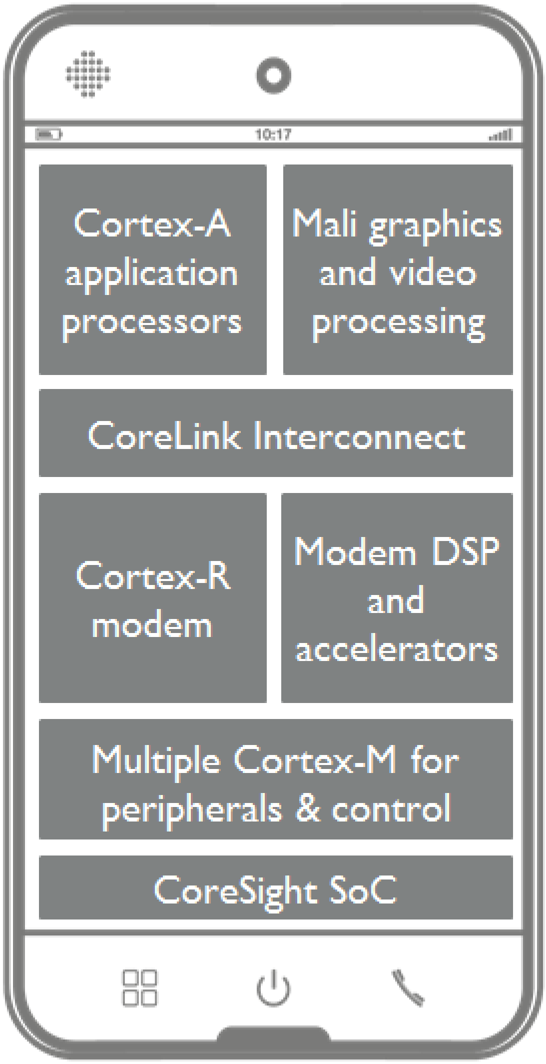
The Route to 5G

The key features of next generation cellular communications and the key technology components required.

David Maidment, Mobile Segment Marketing, ARM

Neil Werdmuller, Senior Product Marketing Manager, ARM

Introduction



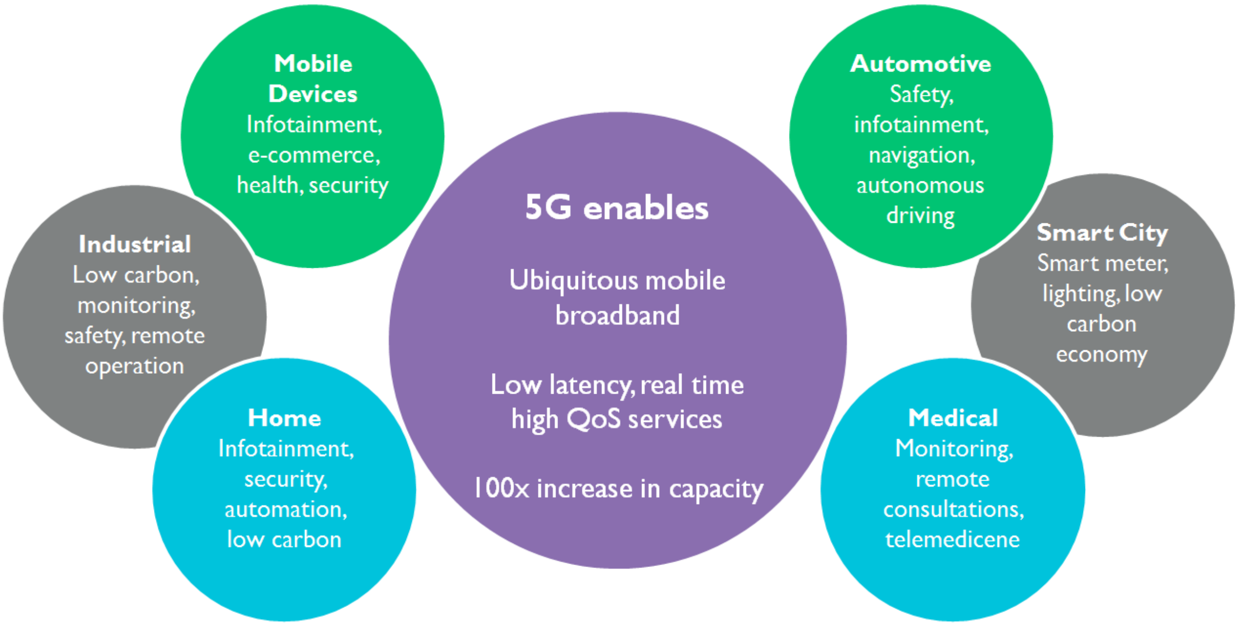
The mobile phone has seen explosive growth over the last two decades to become an essential part of our everyday lives. Right from the outset, ARM® has been at the heart of these devices enabling ever more sophisticated capabilities whilst maintaining a focus on maximising battery life through industry leading power efficiency.

The complexity of smartphones is not often fully appreciated, with the average smartphone containing over 10 ARM based processors managing functions such as touch screen, sensor processing, location/positioning, camera, graphics, applications as well as the ever growing plethora of connectivity such as Wi-Fi, Bluetooth and LTE.

With the wireless industry gearing-up to standardise the next generation of mobile broadband devices, or so called '5G', this whitepaper takes a look at what that will mean in terms of technologies, challenges and the use cases that it will ultimately enable. In particular we look at how the ARM Cortex®-R8 real time processor 'under the hood' will enable a new breed of multi-Gigabit devices whilst still keeping power efficiency at the heart of its design.

Fast paced mobile broadband evolution

As mobile broadband continues to evolve we see both new and evolved use cases opening-up. The dawn of 5G will bring a continuation of the always on, always connected world and in turn bring new ways of interacting with it. As well as multi-gigabit services, 5G also promises to deliver on low throughput, energy constrained devices or so called “Massive Machine-Type Communications” (mMTC). We are seeing the emergence of mMTC today in LTE with standards such as LTE Cat-0 and NB-IOT, both of which promise to pave the way to 5G.



Mobile devices are what come immediately to mind when we think 5G. Given the sophistication of our smartphone experience today how is that set to differ with the advent of 5G? A number of the proposals being considered for 5G standardisation are around network efficiencies, mostly looking at how to manage the sheer volume and demand of mobile data. Multi-gigabit services will allow consumers to download digital content near instantaneously and ultra-low latency connections enable services such as virtual and augmented reality and enable new automotive applications.

Beyond traditional mobile we see 5G as a key enabling technology for a plethora of additional services. 5G will bring remote telemedicine to reality allowing physicians and healthcare workers to remotely manage patients via 5G connected devices, a true opportunity to democratise healthcare and wellbeing.

The low carbon economy represents arguably one of the greatest challenges to the developed world over the next decade. The widespread availability of efficient and reliable wireless internet, will help in achieving this low carbon economy, through enabling efficiency and bringing greater levels of control and integration. From managing smart street lighting, remote emissions monitoring, public transport and public information there are countless opportunities for 5G to have an impact on the way we live our daily lives. Even the 5G network architecture itself has a requirement to reduce energy consumption, which as well as reducing OPEX for the mobile operator would also significantly reduce carbon emissions.

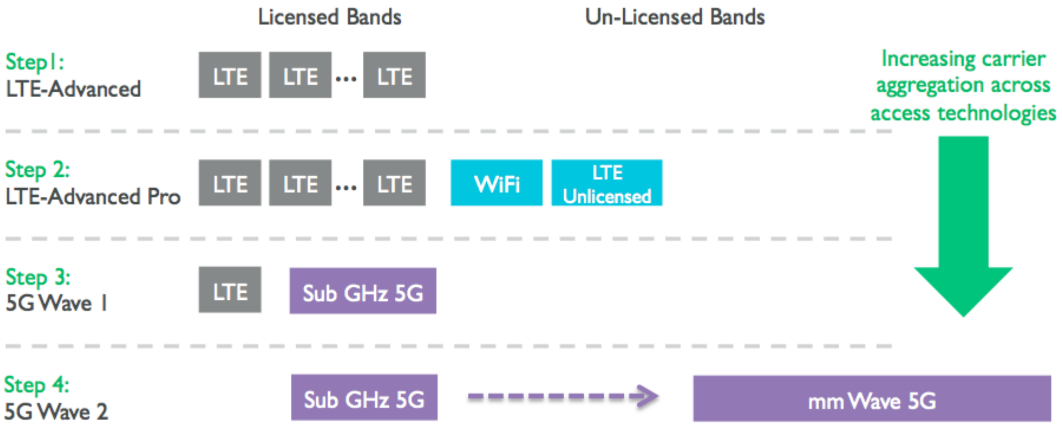
5G is also expected to bring new experiences within the home and when we drive our cars. 5G is seen as the 'beyond mobile internet' technology. 5G's expected capacity and low latency will allow it to be deployed in ways not previously accessible with traditional 4G/Wi-Fi services. As an example, the area of connected car and autonomous driving is seen as a key area that depends heavily on highly reliable low latency wireless connectivity for applications such as safety and collision avoidance, and it is envisioned that this use case will be unlocked with 5G.

Enabling the next generation devices

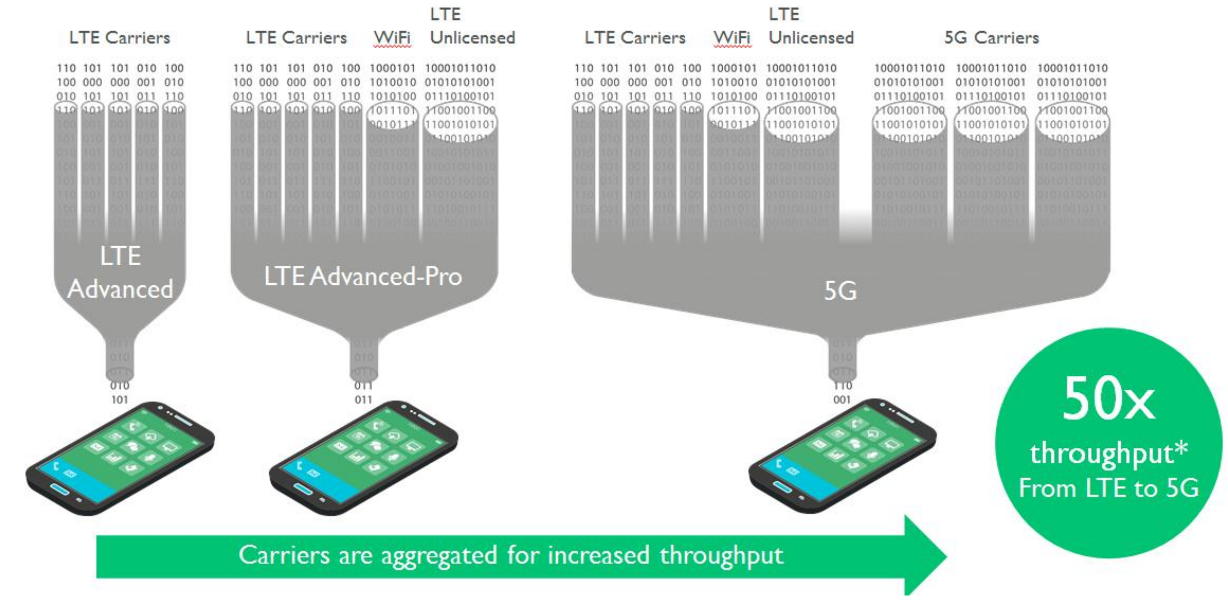
Spectrum is a precious resource and as the demand for mobile services has exploded over the last decade so has the demand for radio spectrum. Traditionally we think of radio spectrum as being divided into 'blocks' or 'carriers' that can be applied to different purposes such as TV, Wi-Fi, Bluetooth or mobile phones. The allocation of spectrum to various purposes is managed regionally by regulators such as FCC in the US and OFCOM in the UK.

In the early days of mobile phones it was relatively simple, blocks of spectrum were allocated (often by auction) for the purpose of providing predominantly voice based services which in their very nature consumed very small amounts of spectrum. During the last decade, and following the birth of the smartphone, that situation changed and spectrum was more and more used for mobile broadband services. Generally speaking, the higher the throughput delivered to the user the more spectrum that is used to deliver that service. If you also multiply that by the number of users you see very quickly that demand outstrips supply for mobile data, spectrum allocation in the traditional sense doesn't keep-up.

With that in mind how as an industry do we keep up in terms of delivering the mobile broadband experience and what does that mean for handset technology challenges over the next decade?

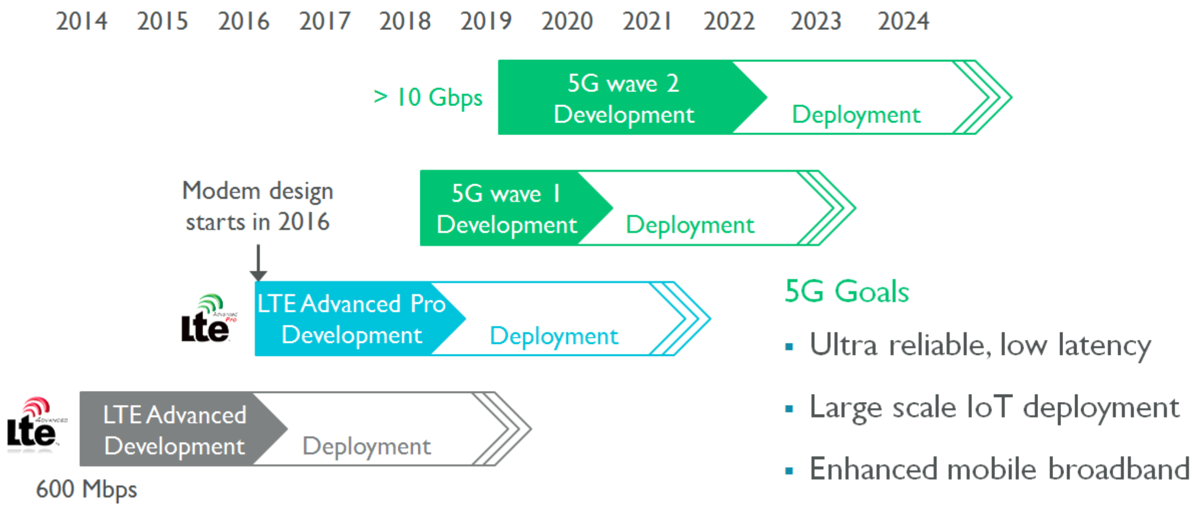


Carrier aggregation coupled with advanced antenna techniques called MIMO are used to help alleviate the pressure and deliver more efficient delivery of services. The move from 3G systems to LTE has seen ever increasing data rates and although these in part have come about through more sophisticated modulation and MIMO techniques, in the main the throughputs have been fuelled by advancements in carrier aggregation allowing more efficient use of fragmented spectrum.



Licensed bands are parts of the spectrum that have particular limitations of use imposed on them; for example a block of spectrum may be restricted to mobile services use and allocated to a particular cellular carrier. The advantage of a licensed band is that the carrier has complete control of that part of the spectrum, so can manage the quality of service (QoS) and provision accordingly. The limitation of licensed band is that it is an ever more precious resource unable to meet the ever growing demands of mobile data and the rapidly growing number of subscribers. To counter this, operators are increasingly looking at how they can use unlicensed bands in combination with their existing licensed band services. Increasingly we are seeing a drive to include unlicensed bands in carrier aggregation whereby a device can simultaneously use both licensed bands (often as a control channel) with unlicensed offload e.g. to technologies such as Wi-Fi as well as emerging LTE-Unlicensed technology. Many of the remaining advances in the LTE standard within 3GPP are focused on the management of these unlicensed offload techniques.

5G promises a lot, but as a standard is not yet even defined. As we have seen in the previous section, if it is done well then 5G will unlock the next 20 years of digital services bringing new and enhanced use cases to our everyday lives.



The heart of 5G will bring new modulation and ever more complex MIMO techniques to maximise the efficiency of the precious spectrum resource and to offer up to 50x increase in throughput compared to early LTE capabilities. The notion of 5G is also to cover a very wide range of frequency bands, far beyond what we are seeing in LTE today. The rational for this is to harmonise access technologies across all of the bands and to maximise efficiency whilst working to build increased capacity for the next generation of services. From sub GHz bands that will provide wide area services, through to local area GHz bands used widely today by Wi-Fi, we will see 5G in a broad range of deployment scenarios. Extending further, 5G promises to open as yet underutilised mm wave bands beyond 30GHz. These bands offer the ability to deliver the headline multi-Gbps throughput rates often associated with 5G. The drawback of mm wave bands is one of coverage: we can only expect devices to work within "line of sight" and a few tens of meters of the base station, which in itself will bring deployment challenges.

So what does this all mean for the future of the smartphone, and in particular the modem baseband processing? Looking back at the trends, we see three main themes that SoC designers are faced with in meeting these new requirements:

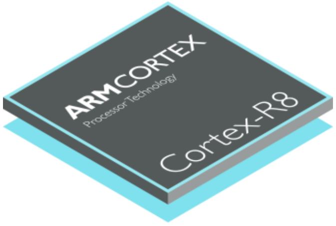
- Data rates continue to rise, we can expect to see Gigabit services based on LTE in the near future and throughputs perhaps as high as 10 to 20 Gbps in 5G

- Massive increase in carrier aggregation. Throughput and network capacity is ultimately served by ever more sophisticated carrier aggregation. This aggregation puts a high processing complexity in the handset modem processor as it serves multiple radio access bearers in parallel. This theme will continue and be at the heart of evolved LTE technologies (LTE-Advanced Pro) and 5G

- Continued drive to energy efficiency and maximising battery life in handsets. As new access technologies are introduced they cannot compromise the user experience and as such handset modems need energy efficiency to be at the heart of their design

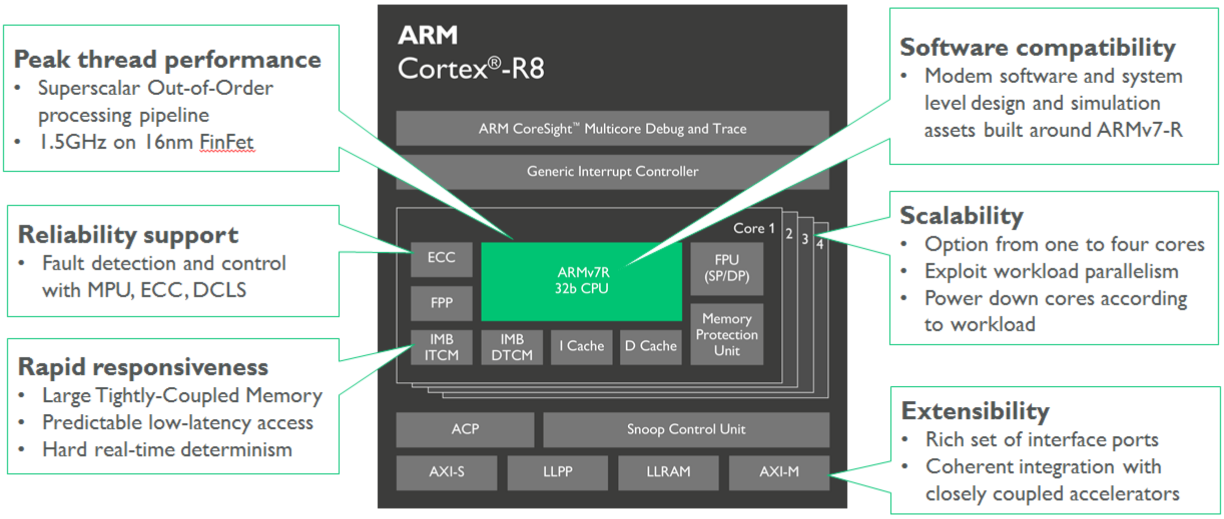
As we will read in later sections of this paper, the ARM Cortex-R8 processor is uniquely placed to allow designers to meet these demands whilst maintaining backwards compatibility and legacy support for previous generation technologies such as 3G and LTE which remain as mandatory in today's multi-mode devices.

Introducing Cortex-R8, the heart of next generation mobile broadband.



The Cortex-R8 processor is the highest performance embedded real-time processor ever from ARM. It builds on proven technology but extends the performance to new even higher levels. The processor pipeline adopts a number of ARM technologies developed for the highest performance applications processors, but adapts them to offer the highest performance while meeting the requirements for hard real-time.

Hard real-time means that the processor can switch to handle a new critical event very rapidly, and with a known (deterministic) worst case delay. This worst case delay is typically just a few nano-seconds and enables interrupts from other parts of the system to be detected and acted upon very quickly. Layer 1 control tasks in LTE-Advanced Pro and 5G modems will handle multiple carriers and very high data rates. A processor must therefore run at a very high clock frequency and be able to very rapidly switch between tasks and handle incoming events. Unlicensed carriers, such as WiFi, provide much higher data and packet rates than LTE and combining and controlling these diverse carriers requires a specialised processor. The Cortex-R8 with an eleven stage pipeline can be clocked very rapidly to deliver the performance needed. The pipeline is 'out-of-order', which means that processing can continue even when some instructions are waiting for data from the slower external memory systems – this greatly reduces pipeline 'stalls' and delivers the best possible performance.



Cortex-R8 also enhances the Tightly Coupled Memory (TCM) to allow more code and data to be held in fast memories, so that there is no delay to access key routines and data. Unlike caches that are managed by the processor, TCMs are managed by the developer so critical instructions and data structures are always immediately available. In modems there are some very critical real-time routines while other routines are less critical and can run in the background.

Cortex-R8 enables up to four processors to be integrated in a single coherent cluster. For modems these processors are typically run in an asymmetric processing mode for the best efficiency, and also to enable processors to be powered down when the mobile phone is in idle mode and only power up the additional processors as throughput rises, greatly extending battery life. This configurability also enables modem developers to create different modems addressing different LTE Categories with a single investment in software but scalable performance.

Cortex-R8 implements a wide and flexible choice of interface ports to the rest of the modem system. Dedicated ports to control external hardware and accelerators provide the lowest latency control to deliver the best possible performance in these complex systems.

However, Cortex-R8 is not just for modem designs. This same industry leading performance is ideally suited for Enterprise storage products, both HDD and SSD and any other embedded real-time platforms where scalable performance is required. Cortex-R8 implements new error detection, correction and containment schemes to ensure the best possible reliability.

Delivering the next generation mobile broadband experience

As well as a new optimised and highly efficient air interface for 5G, the modified network that will support the 5G infrastructure requirements must provide the ability to create simplified management and orchestration layers that abstract underlying hardware and software complexity.

A mix of different equipment will be needed for the successful deployment of new 5G infrastructure. Depending on geographies, technologies like Cloud RAN, distributed content delivery, scalable control networks and adaptive antenna arrays will likely be needed. Cloud RAN, as an example, is hugely disruptive as a new technology and is provided where multiple base station elements and the associated control network are collapsed together into the "cloud".

In order to meet these new platform requirements across Cloud RAN, distributed content delivery and scalable control networks, significant progress has been made by utilising some key new enabling technologies:

- **Software defined networking (SDN)** is a new way of providing scalable connectivity in the network and simplifying the old legacy network. SDN is a set of standards developed initially by the Open Networking Foundation that provide an abstraction layer of network functionality by separating the control plane from the data plane. Network management and operations can be centralized instead of being spread out among different network layers and boxes. Centralized control via a simplified abstraction software layer has led to benefits such as lower operational expenditures (OpEx) and more automation, control, flexibility, agility, and application innovation. SDN will alter how equipment is connected to the network infrastructure and connectivity between Access and Aggregation nodes will evolve accordingly.

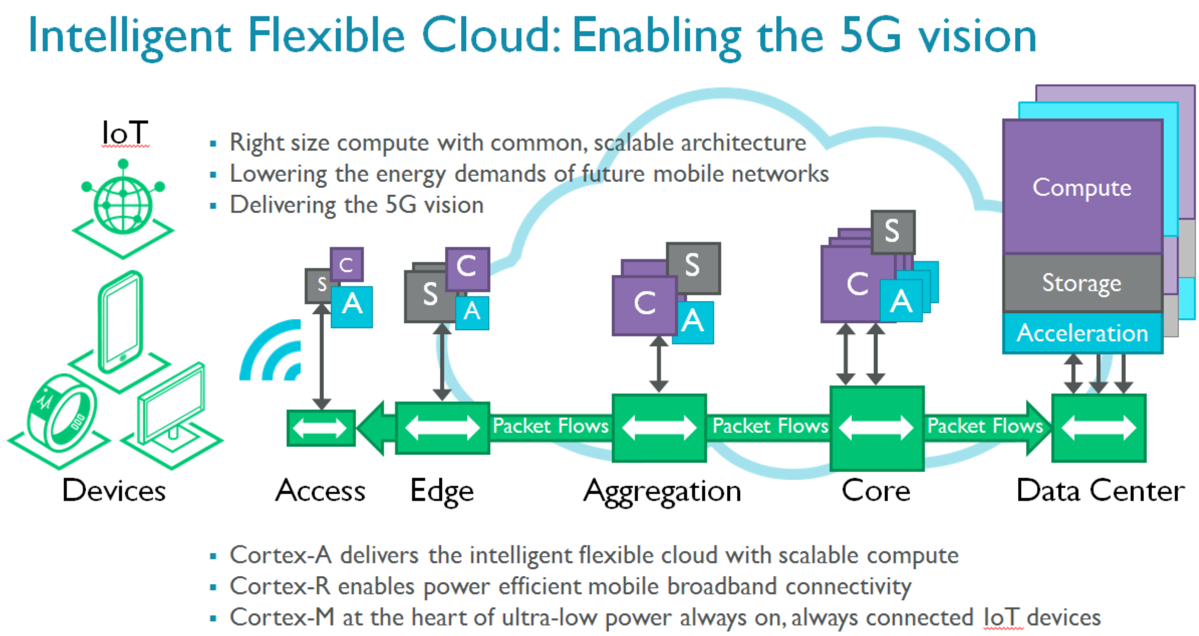
- **Network Function Virtualization (NFV)** allows legacy functions to be moved from proprietary hardware appliances to more standardized servers, switches, and storage elements. As these new functions are implemented in software, they can be easily located on platforms in data centers, network nodes, or customer premises to take advantage of global network efficiencies. NFV benefits therefore include lower capital expenditures (CapEx) and OpEx through less reliance on proprietary and single-purpose hardware. Using NFV can accelerate time-to-market for services due to faster configuration, testing, and integration. In order to support many of the latency sensitive 5G functions and end use cases, the NFV implementations must be paired with optimized network offload capability and paired with technologies like Mobile Edge Compute that move these virtualised network functions as close as possible to the edge of the access network to avoid excessive transitions across the network infrastructure.

- **Distributed intelligence. By supporting more Distributed Intelligence in the network , fundamental decision points can be distributed over available resource already deployed in the cloud.** Workload optimized hardware and software are used to enable network, storage, and compute functions in distributed points across the network. Workload-optimized hardware based on highly integrated SoCs with heterogeneous processing capabilities makes it possible to add intelligence anywhere in the network – even scaling down to the most power and form factor constrained locations. A common software platform enables developers and IT users to more rapidly deploy services.

- **Storage** As the network and services evolve in 5G, so we will also see the integration of storage directly within the infrastructure network. Whereas traditionally we associate storage 'in the cloud' within data centers, increasingly we see migration of storage to all nodes within the network. At the core of 5G requirements we see high bandwidth, low latency services. These requirements not only impact the air interface radio aspects, but also push back through the whole network. Distributing storage and pushing intelligence to the edge helps to meet these objectives by minimizing round trip delays and delivering services and intelligence where it is needed.

These technology standards and architectures are part of the foundation of the next generation infrastructure network – or “intelligent flexible cloud”. It is flexible because it can easily and quickly address diverse network requirements and augment the air interface specific challenges of 5G. It is intelligent because it leverages business, customer, and network data to enhance existing services and as the basis for the creation of new services that will be highly innovative and competitive.

ARM and its partners provide a common ARM-based processing platform to host the diverse needs of delivering the route to 5G.



Why use ARM Cortex-A series processors in the Network Infrastructure?

ARM is delivering processors and interconnect IP to meet the needs of network infrastructure, with future needs directly driving ARM’s roadmap. Key to this delivery are a variety of Cortex A processor cores and Cache Coherent interconnect such as the Cortex-A72, Cortex-A53 and the CCN family of interconnect products.

New SoC platforms for providing a mix of heterogeneous CPU, DSP and function-specific accelerator cores are essential for meeting the throughput requirements, latency and flexibility for 5G deployments. More and more capabilities will be integrated onto single SoCs, and these will typically process multiple traffic types, including data channel payload, control plane traffic, front-end processing, and user scheduling.

With this trend towards integration and higher performance SoCs, there will be a mix of processing elements to support bursty high-speed traffic payloads and latency-sensitive traffic via processor cores and smart signal processing elements.

Network infrastructure applications blend different levels of three distinct functions: control plane processing, packet or backhaul processing, and event or traffic scheduling over the several clusters of cores available on any particular SoC device.

5G basestation equipment will have a very different mix of these functions from 5G core equipment. Designers have to determine the best mix of processor capability to handle the processing required, selecting from the available technology choices to deliver their designs in the time required to match market opportunity.

Control Plane: Control-plane functions require the maximum amount of processing per packet, typically involving tens of thousands of instructions per packet, usually allocated in a 'run to completion' mode. Out-of-order and multi-stage pipelines can be utilized very effectively.

High-performance cores having virtualization capability can meet the needs of the control plane, the content delivery network (CDN), and other functions that require high single-thread performance. Applications running in the control plane include NFV, CDN for cloud and edge networks, and potentially new remote access technology requiring significantly more performance (such as 5G for example).

Data Plane:

The edge of the network sees data rates in the range of hundreds of Mbps or perhaps a Gbps; the access/cloud segment experiences data from one to the tens of Gbps; and the core processes from twenty to hundreds of Gbps. Unlike the control plane, the challenge here is in handling bursts of backhauled traffic, processing the headers, and placing the data into buffers without dropping any packets.

This involves a completely different blend of processing. Many data-plane designs use dedicated DSPs for this capability, and they interface these data-plane processors to the SoC system through the ARM AMBA® interconnect. The DSP offers a dedicated and optimized instruction set for data plane processing and offloads the CPU from power-hungry and computation-intensive functions.

Instead of the tens of thousands of instructions per packet used for control processing, packet processing may use only hundreds of instructions per packet. Access to cache (instruction, data, L2, and L3) and external memory are also different for data packet processing.

A key distinction can be made between data and control processing. ARM uses the terms 'stateless' and 'stateful' to distinguish between them. Stateless processing uses a sea of small cores to handle streams of packets coming into the system-on-chip (SoC). Each core executes in a 'run to completion mode' to classify headers and dump packets into memory. Each packet is handled independently; the core knows nothing about any prior packet. The number of cores and the size of the interconnect scale simply according to the speed of the interface. Stateful processing, by contrast is used in a higher level of decision making, where the history of packets matters. This is where flows and sessions can be managed, typically in the control plane.

Scheduling:

Another challenge in 5G systems is orthogonal to the first two. For user access scheduling, where users have to be scheduled according to air interface bandwidth available, latency is key. In the case of LTE for example, where there may be hundreds of subscribers on the air interface to be scheduled into their own timeslot; all of this has to be calculated over potentially multiple cores in accordance with the time constraints of the 5G standard: likely to be less than 0.5ms. This involves a lot of priority calculation, scheduling of receive and transmit tasks, and signaling to and from the DSPs, processors and memory. So, the ability to use multiple cores in a heterogeneous architecture and then to switch contexts between them is essential.

Technology Requirements:

As data consumption on smart connected devices spirals up, fuelled by new air interface technology like 5G, the burden on system designers to deliver more performance using the same power and equipment footprint is driving new designs. ARM have been developing IP to support higher-performance, multicore-capable processors. Coherent interconnects, and optimized performance-enhancing physical and logical IP all serve to support these very flexible heterogeneous architectures that will be key to ensuring the performance requirements for 5G are met. New ARM cores like the Cortex-A72 and Cortex-A53 have enabled performance/watt and performance scalability targets to be met for next generation SOC designs. In addition, in an R&D budget-challenged world, an industry-standard instruction set architecture (ISA) with a well-supported software and tools ecosystem enables SoC design managers to deliver their products to market faster and to conserve R&D dollars for developing value-added and differentiated application-specific features.

Summary

For over 20 years, ARM has been at the heart of the mobile revolution. From early 2G handsets through 3G and into LTE there have been over 20 billion handsets shipped with ARM at the heart of the cellular modem. ARM based modems enable the smartphones that we depend upon in our everyday lives.

As LTE comes to maturity we have delivered on the reality of connecting all aspects of our digital lives. Beyond information based services such as email, news and social media, we are seeing a drive to more sophisticated use cases that touch all aspects of our daily lives such as health, wellbeing, medical and beyond. Even the form factor of these devices is starting to change and break the constraints of the traditional smartphone; new applications such as wearables integrate seamlessly in our daily routine.

Looking ahead to the next decade of mobile, what can we expect to see? More demand for services brings a greater need for capacity in the networks that serve us. Making more efficient use of radio spectrum is key and this has spawned the industry research into 5G radio systems. Enabling more capacity in our mobile networks to serve not just more users, but more 'things' or objects or devices as our world gets connected in the greater IoT umbrella. Delivering on a low carbon economy through more efficient resource usage and monitoring, democratising healthcare through remote medicine or enabling the connected car are just a few of the areas due to benefit from the continued mobile revolution.

As we build in these access technologies to deliver these benefits, so to do we need to deliver advanced processing capabilities 'under the hood' in the future generation of mobile devices. The ARM Cortex-R8 processor is designed to deliver on that promise through energy efficient real time processing, allowing device vendors and OEMs to realise the potential of 5G.