1

Introduction

The fifth-generation (5G) and the upcoming sixth-generation (6G) wireless networks are poised to be the most significant and transformative technology in the coming decade, disrupting numerous industries ranging from energy, agriculture, manufacturing to transportation, retail, healthcare, entertainment, and financial services. Surprisingly and unbeknownst to most outsiders, one of the key technology enablers is the new computer architecture and software design model, in addition to the wireless communications technology [1, 2].

The main objective of this book is to help university students and professional engineers understand the 5G/6G edge computing architecture and to describe step by step how to unleash the full 5G/6G potential using custom edge computing technologies.

This book is divided into five parts:

- Part 1 (Introduction):
  - Chapters 1 and 2 introduce the concept of edge and custom computing and provide an overview of 5G/6G technologies.
- Part 2 (Theory):
  - Chapters 3 and 4 discuss how to use high-level synthesis (HLS) and coding theory to realize secure edge acceleration with high performance.
- Part 3 (Architecture):
  - Chapters 5 and 6 elaborate the 5G/6G hardware and software acceleration architecture.
- Part 4 (Applications):
  - Chapters 7 and 8 describe a few 5G/6G killer applications with acceleration technology including some practical development strategies.
- Part 5 (Future Roadmap):
  - Chapter 9 discusses the road ahead in the coming decade.

Edge Computing Acceleration: From 5G to 6G and Beyond, First Edition.
Patrick Hung, Hongwei Kan, and Greg Knopf.

1.1 Introducing 5G and Internet of Everything

With 20 times higher bandwidth and 10 times reduction in network latency, fifth-generation (5G) technologies enable many new applications, such as remote surgery and autonomous driving. These incredible applications require both high bandwidth and extremely low latency. To realize these useful applications, we need high-performance communication and computing platforms demanding different kinds of acceleration technologies, known as heterogeneous acceleration architecture (Figure 1.1).

Although the fundamental wireless technology – wireless spectral efficiency – has only increased three times from 4G/LTE to 5G, the area traffic capacity is 100 times larger. This is made possible mainly by building many more 5G cell towers, each covering a smaller area and supporting a wider spectrum. A typical 4G/LTE cell tower may have a range of 10 km, but a typical 5G tower may have a range of 500 m or less.

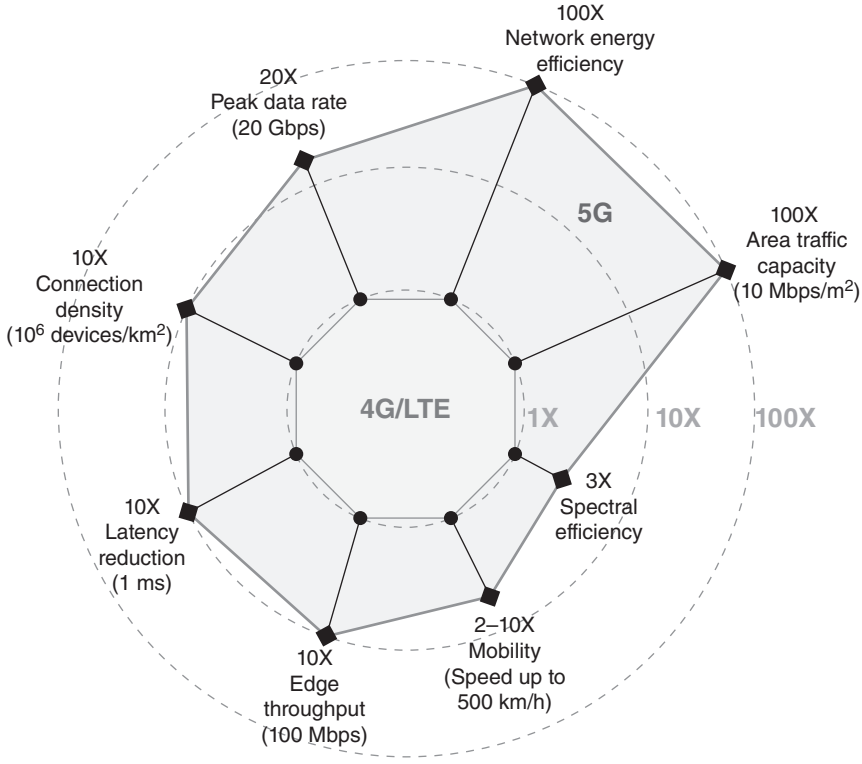

Figure 1.1 Comparison between 4G and 5G features and performance metrics.

The 3rd-Generation Partnership Project (3GPP) is responsible for driving the 5G telecommunications standards, which are detailed in the 3GPP Specifications Release 15 (NR Phase 1) and Release 16 (NR Phase 2) documents. The 3GPP 5G network architecture defines the network entities based on their functions and nature (control and data planes) [3–6].

There are three technical specification groups (TSGs) within the 3GPP:

- **TSG SA** (services and systems aspects) focuses on the overall architectures and services.
- **TSG CT** (core network and terminals) focuses on the core network (CN) architecture and terminal interfaces.
- **TSG RAN** (radio access networks) focuses on the radio transmission and its technical requirements.

The TSG SA defines three different sets of requirements for new 5G usages:

- **Enhanced Mobile Broadband (eMBB):** a new requirement that defines higher data rates, traffic and connection densities, and user mobility.
- **Massive Machine-Type Communications (mMTC):** a new requirement that supports very high traffic densities of devices.
- **Ultra-Reliable Low Latency Communications (URLLC):** a new requirement that provides very low latency and very high communications service availability (Figure 1.2).

Figure 1.2 compares the bandwidth and latency requirements of some 4G and 5G applications. The typical 4G network latency is 10–20 ms, while the minimum 5G network latency can go as low as 1 ms. The maximum 4G network bandwidth is around 100 Mbps, but the 5G network bandwidth can go up to 1 Gbps. Some new mobile applications, such as mobile telepresence, may only require higher network bandwidth. On the other hand, most disruptive applications, such as autonomous driving and smart energy grid, may demand very short network latencies.

Although human beings may not perceive any difference between 5 and 50 ms, this difference is so important in many industrial robotic systems. Indeed, an interconnection of machine systems – called the Internet of Everything (IoE) – can take full advantage of the three new 5G features (eMBB, mMTC, and URLLC) [7, 8] (Figure 1.3).

In 2012, Cisco Systems extended the concept of the Internet of Things (IoT) and coined the term "the Internet of Everything (IoE)," representing a networked connection of people, processes, data, and things. It goes beyond simple machine-to-machine (M2M) communications, forming a network of networks connecting all data, technologies, processes, and people [9–12].

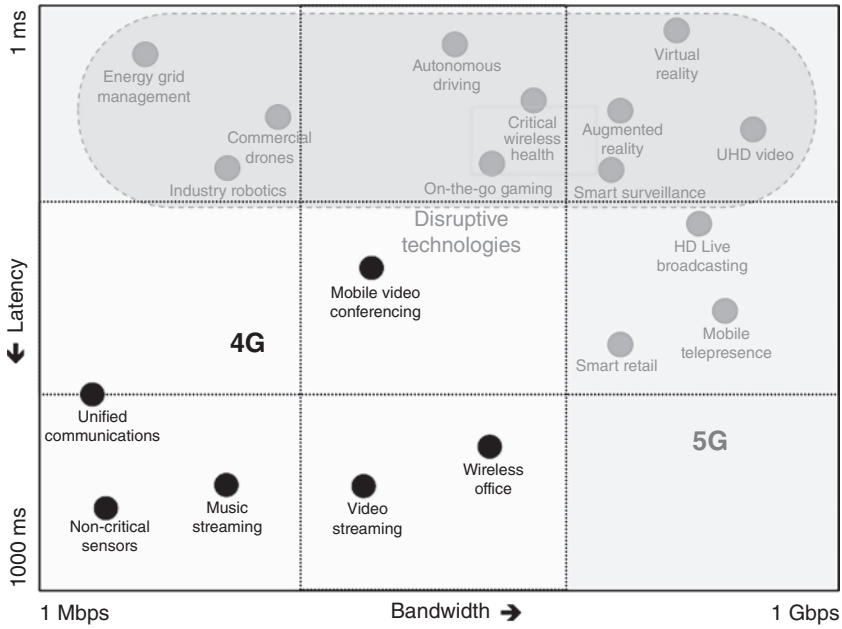

Figure 1.2 Latency and bandwidth requirements of some typical 4G and 5G applications.

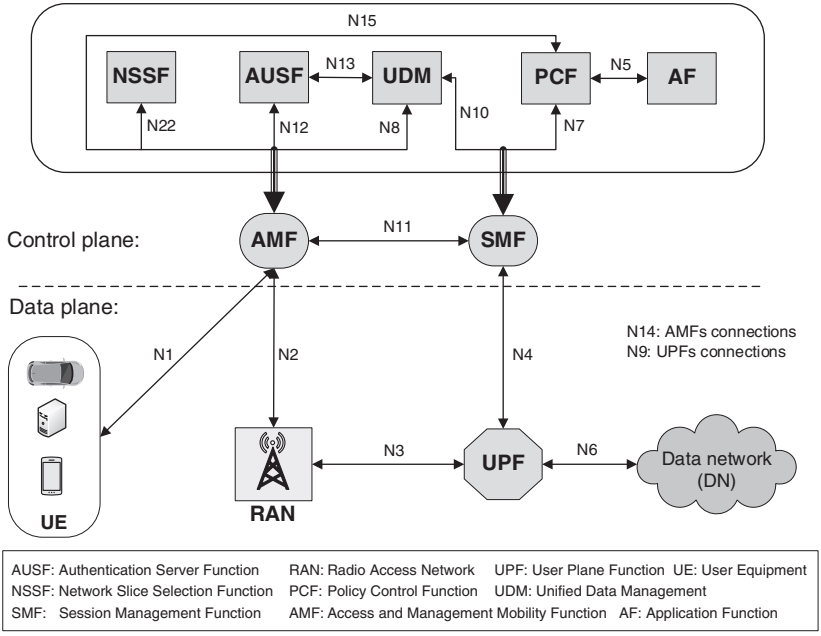

Figure 1.3 3GPP 5G network architecture.

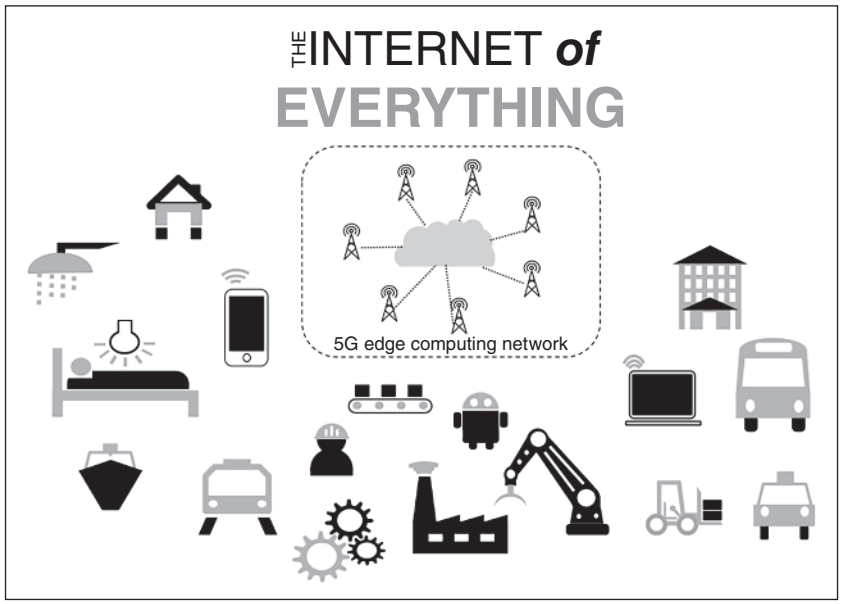

Figure 1.4 Ubiquitous Internet of Everything (IoE) devices.

Today many IoE devices are adopting a tethered communication scheme – using Wi-Fi or Bluetooth to first connect to a smartphone or an access point – to communicate with other IoE devices. Unfortunately, tethered communication is slow, unreliable, and insecure, limiting the performance of these applications [13, 14].

The 5G/6G wireless network enables IoE devices to adopt an untethered communication scheme, thus allowing each device to have a secure channel connecting to a radio access network (RAN) with guaranteed network latency, bandwidth, and security (Figure 1.4).

According to Juniper Research, IoE devices reached 46 billion units by 2021. Figure 1.4 depicts a framework of different connected devices and things. The top IoE applications include smart factories, smart buildings, smart grids, connected gaming and entertainment, smart vehicles, remote healthcare, remote education, and connected marketing and advertisement.

These ubiquitous IoE devices have some essential characteristics:

- **Resource Limitations:** Many IoE devices are limited by size, weight, and power (SWaP) and do not have sufficient resources to process or store information locally. Accordingly, the information is offloaded to a remote server for storage and processing.
- **Network Latency:** Extremely low network latency is needed for time-critical applications. Industrial control systems can only tolerate delays of the order of milliseconds. Autonomous vehicles and virtual reality applications also have

similar latency requirements. The network latency can be excessive if the servers are in a remote cloud.

- **Network Bandwidth:** A large IoE network can generate a huge amount of real-time data. For example, 120,000 CCTV devices transmitting 1080p video to a remote cloud may generate 1 Tbps traffic. If the servers are in a remote cloud, the 1 Tbps network data can cause serious traffic congestion in the Internet backbone.
- **Security and Reliability:** Many critical applications, such as energy grids and smart factories, must be ultra-secure and reliable. Due to the multi-hop nature of the Internet backbone, it is probably impossible to guarantee ultra-security and reliability with a remote cloud server.
- **Reconfigurability:** After deployment, an IoE system oftentimes requires future security patches and performance updates. While it is easier to update a generic software architecture, it is difficult to update a high-performance hardware architecture. This is an important reason why custom computing is desirable for 5G computing architecture.

1.2 Edge Computing Architecture

Although 5G can provide large bandwidths and low latencies, if the other end of the communication is far away, these parameters cannot be guaranteed. In particular, the communication latency is limited by the speed of light. For example, a distance of 4000 km – between Los Angeles and New York – has at least 20 ms delay on fiber optics [15–17].

In reality, the actual latencies, including all electronic and electro-optic delays, are much longer. Table 1.1 shows the typical round-trip latencies from Tokyo to some target cities.

Table 1.1 Round-trip time (RTT) from Tokyo to overseas cities.

Target city	Round-trip time from Tokyo (ms)
Cape Town	360
New York	176
London	218
Los Angeles	128
Paris	236
San Paulo	275
Shanghai	67
Sydney	114

Source: Adapted from http://Wondernetwork.com.

It is important to note that a saving of 20 ms – migrating from a 4G network to a 5G network – is not meaningful if the application involves a network round-trip time of 360 ms. We must keep this in mind when we implement a remote application, such as remote surgery or Internet gaming. The remote server cannot be too far away, or the underlying algorithm must be able to hide the latency from the users.

As much as possible, the IoE servers should be placed close to the wireless networks to optimize the network security, reliability, bandwidth, and latency. This requirement gives rise to the emerging **edge computing** architectures, where servers are located at the edge of the Internet near the wireless devices. Therefore, edge computing is a distributed computing paradigm or a model where the processing and storage elements are placed close to the sensing device (data source) for improved latency, throughput, security, and several other benefits. The traditional method sends the sensor data to the cloud for processing and storage. As shown in Figure 1.5, some edge computing nodes do not require the cloud because they are independent. The edge computing server is placed together (in the same housing) with the sensing device for many independent edge computing systems (Figure 1.5).

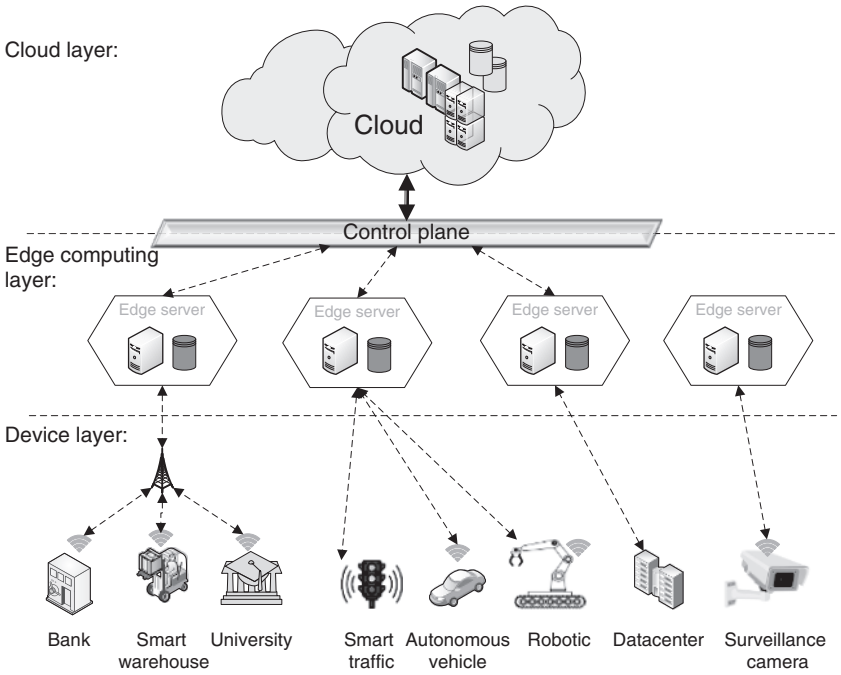

Figure 1.5 Edge versus cloud computing architecture.

1.2.1 Edge Versus Cloud Computing

Cloud computing, which has been a dominant enabler in enterprise IT delivery in the past decades, offers many key advantages [18]:

1) An enterprise can avoid the huge capital expenditure (CAPEX) of building a datacenter.
2) Cloud computing can lower operating expenses (OPEX) by exploiting economies of scale.
3) Cloud computing can minimize electricity costs by strategically locating servers in regions with the lowest electricity and air conditioning costs.

While cloud computing is a good value proposition for many enterprise applications, it is not a viable option for many time-critical IoE applications. Edge computing is a new paradigm in which substantial computing and storage resources are placed in close proximity to mobile devices or sensors [19–21] (Figure 1.6).

There have been many research activities and industry developments related to edge computing in the past decade:

- In early 2013, Nokia and IBM jointly introduced an edge computing platform called the Radio Applications Cloud Server (RACS).
- In 2014, a mobile edge computing standardization effort began under the auspices of the European Telecommunications Standards Institute (ETSI).

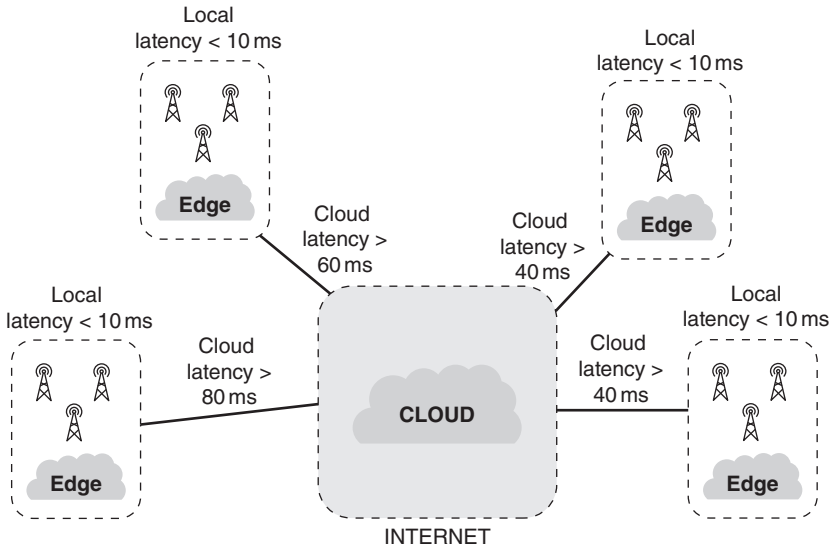

Figure 1.6 Edge versus cloud network latencies.

- In 2015, multiple mobile technology organizations – led by Vodafone, Intel, and Huawei – launched the Open Edge Computing Initiative (OEC).
- In late 2015, Cisco, Microsoft, Intel, Dell, and ARM, in partnership with Princeton University, created the OpenFog Consortium.

The origin of edge computing dates back to the 1990s, when Akamai introduced content delivery networks (CDNs) to accelerate web performance. A CDN uses nodes at the edge close to users to pre-fetch and cache web content. These edge nodes can also perform some content customizations, such as adding location-aware advertisements. The emerging edge computing architectures generalize and extend the CDN concept by leveraging cloud computing infrastructure within a 5G network.

1.2.2 Edge Design Options

Broadly speaking, there are two different architectural options for edge computing:

- **Mobile Edge Computing (MEC):** MEC consists of commercial off-the-shelf (COTS) computers located near the 5G RAN to reduce latency and improve context awareness. It is a standalone architecture that can function without a remote cloud.
- **Fog Computing (FC):** FC is a decentralized computing infrastructure with fog computing nodes (FCNs) placed between the end devices and a cloud. The FCNs are based on heterogeneous network elements including routers, switches, and gateways. Unlike MEC, FC is tightly linked to a cloud and cannot function without a remote cloud.

Both MEC and FC are computing and storage networks, which are built on top of 5G communications networks. Indeed, we are migrating from a simple 4G/LTE communications network to an integrated computing and communications network (Figure 1.7).

Interestingly, the 5G communications network has two similar layers:

- **Control Plane** (CP), also known as user plane function (UPF), consists of generic computers running a network operating system and controlling the network domain. The CP controls and manages the user equipment's access

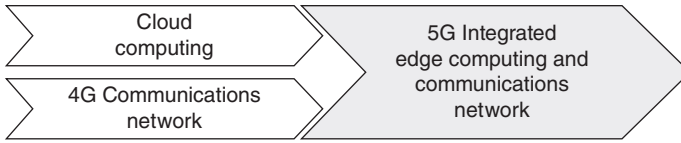

Figure 1.7 Edge computing is driving many 5G applications.

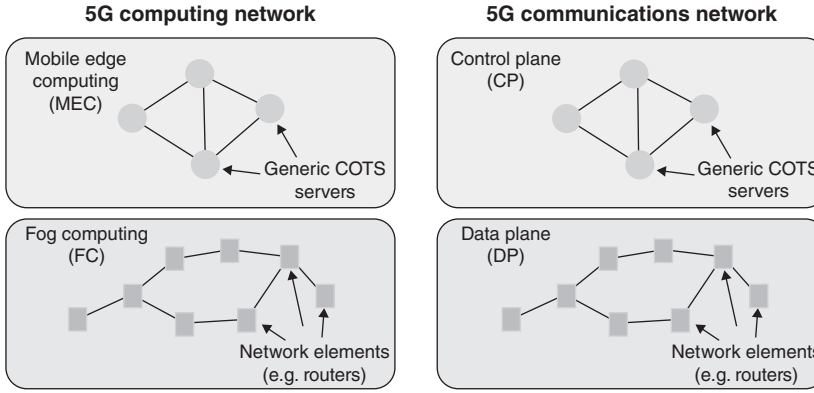

Figure 1.8 5G computing and communications networks.

to the network. It performs the access and mobility function (AMF), session management function (SMF), authentication server functions (AUSF), unified data management (UDM), and policy control functions.

- **Data Plane** (DP) consists of the configurable network devices within the network domain. The data plane interfaces with the radio access network (RAN) and the data network (DN), and then performs the UPF on the received user data. Therefore, the data plane performs tasks such as packet inspection and processing, routing, and ensuring the quality of service (QoS) (Figure 1.8).

In an integrated 5G communications and computing network, MEC and CP can likely share the same servers, while FC and DP can share the same network elements. Nevertheless, each implementation is different, and the mobile network operator (MNO) would determine the network architecture. Table 1.2 compares the two edge computing design options.

1.2.3 Key Benefits of Edge Computing

Edge computing provides the following key benefits:

- **Responsive Services:** Some mobile applications are very sensitive to the long latency in cloud computing. Applications such as autonomous driving, virtual reality (VR), augmented reality (AR), robotics, and energy grids can work better in the edge computing environment.
- **Network Scalability:** Edge computing can reduce massive ingress traffic to the cloud, thus improving network scalability. It processes massive raw data locally and can cut up to 99.9% of cloud traffic.

Table 1.2 Comparison of 5G private network deployment architectures.

Deployment architecture	RAN	DP	CP	MP	Perf. grade	Flexibility grade	Security grade	Investment efforts	Management simplicity	Resource efficiency
DA-A	On-prem	On-prem	On-prem	On-prem	Best	Best	Best	Worst	Worst	Worst
DA-B	On-prem	On-prem	On-prem	Remote	Best	Best	Medium	Worst	Medium	Medium
DA-C	On-prem	On-prem	Remote	Remote	Medium	Medium	Medium	Medium	Best	Best
DA-D	On-prem	Remote	Remote	Remote	Worst	Worst	Worst	Best	Best	Best

The table compares four different 5G private network deployment architectures (DA-A, DA-B, DA-C, and DA-D). DA-A is the most expensive and most flexible architecture, and DA-D is the cheapest and the least flexible architecture. DA-B and DA-C are between these two extremes.

- **Masking Cloud Outages:** Edge servers are independent and can function even without a remote cloud. When cloud outages are short, mobile devices may not even be aware of the outages.
- **Distributed Security:** Instead of storing all data in a centralized server, critical data are distributed in multiple edge computing servers. This creates a robust distributed security architecture.

1.3 Custom Computing

The 5G technologies have presented many market opportunities and challenges at the same time. Here are the seven most critical challenges (Figure 1.9).

We have discussed about the edge computing architecture in the last several sections. Although the standard edge computing architecture, using software-defined networking (SDN) and network function virtualization (NFV) [22, 23], can be cost-effective, scalable, flexible, and reliable, it may not be able to meet the stringent latency and bandwidth requirements. Thus, custom edge computing architecture can be used to dramatically speed up the 5G performance while maintaining the reconfigurability [24, 25].

1.3.1 Introduction to Custom Computing

Custom computing systems are special computer systems designed for specific applications, such as signal processing or database operations, and can be reconfigurable. These systems are desirable when the general-purpose system has undesirable features such as high hardware resource consumption, high latency, low throughput, and high-power consumption. Custom computing systems need to meet some SWaP requirements, usually employ special acceleration hardware elements, including:

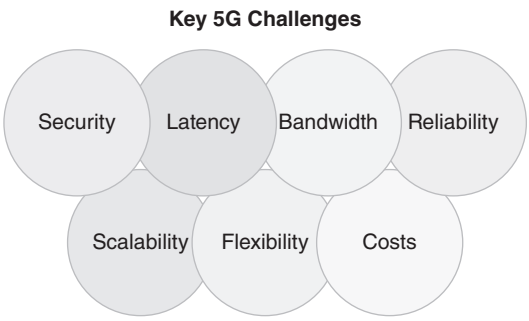

Figure 1.9 Key challenges to fully realize 5G potentials.

- **Field-Programmable Gate Arrays** (FPGA) chips, such as Xilinx Virtex-7.
- **Graphic Processing Unit** (GPU) chips, such as NVIDIA Titan V GPU.
- **System-on-Chip** (SOC), such as Google Tensor Processing Unit (TPU).

Custom computing is gaining popularity among all major cloud providers, thanks to the emergence of machine learning and big data applications. For example,

- Amazon EC2 F1 Instances and Alibaba ECS F3 Instances use FPGA-based PCI Express cards running on generic servers to enable the delivery of custom FPGA hardware acceleration.
- Amazon EC2 P2/P3/G3/G4 Instances and Alibaba EGS Instances use GPU-based PCI Express cards running on generic servers to enable the delivery of custom GPU acceleration.

By running the applications in hardware instead of software, custom computing can be used to improve their computational performance and reduce power consumption and costs. Unlike software algorithms, hardware implementations in FPGA and GPU are less susceptible to cyberattacks. As a result, custom computing architectures are ideal for implementing security elements, such as network firewalls and intrusion detection systems (IDSs).

1.3.2 5G/6G Security Concerns

Among the seven major challenges in the 5G era, security considerations are arguably the most critical concerns, receiving the most media attention in the last few years. Specifically, the 5G stakeholders are worried about the potential cyberattacks from foreign agents and malicious hackers [26–33].

Thanks to many emerging applications, the 5G networks are susceptible to many new threats. The common attack types include:

- **Denial of Service (DoS):** Attackers can temporarily disable the network services.
- **Network Message Modification:** Attackers can temporarily take over the network.
- **Data Corruption:** Attackers can erase important network data.
- **Eavesdropping:** Attackers can intercept confidential data from a 5G network.
- **Fraud:** Attackers can use network services at the expense of another subscriber.

In general, these cyberattacks may be originated from:

- Fake base stations.
- Rogue IoT devices, rogue base stations, and rogue network elements.
- Roaming networks.
- The Internet (Figure 1.10).

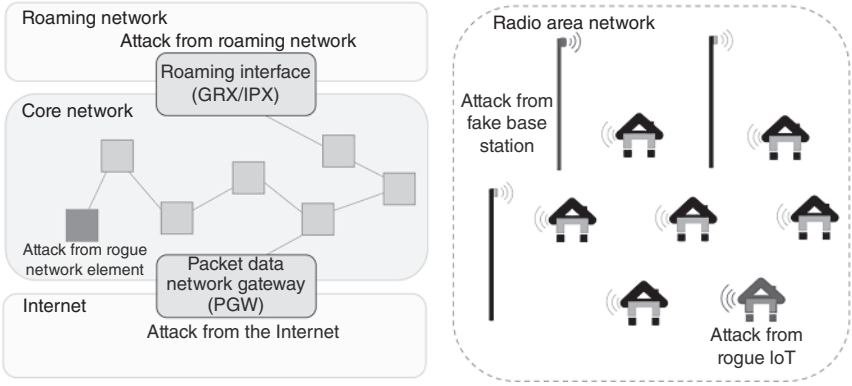

Figure 1.10 Many different ways to attack 5G network.

A DoS attack on a 5G/6G network is not only an inconvenient event but can also wreak havoc in an autonomous driving vehicle or a smart manufacturing factory. Accordingly, it is important to build a reliable 5G/6G network that can:

- Prevent cyberattacks.
- Detect cyberattacks quickly.
- Limit detrimental consequences of cyberattacks.

To prevent cyberattacks, it is useful to use custom computing technology to implement 5G high-performance security firewalls:

- **GTP firewalls** defend the network against GPRS tunneling protocol (GTP) based attacks. These cyberattacks may be initiated from RAN or roaming networks.
- **SGi/Gi firewalls** defend the network against TCP/IP attacks from the Internet. These attacks may exploit vulnerabilities in SMTP, NFS, FTP, Telnet, NTP, and HTTP.
- **Diameter firewalls** defend the network against diameter protocol-based attacks. Diameter protocol contains subscriber information, similar to the traditional SS7 protocol.

To detect cyberattacks, most enterprise networks today may already have IDSs. For example, Splunk IDS uses log data and history to identify network anomalies. However, Splunk IDS typically does not process the log data in real time. Accordingly, it is useful to use custom computing technology to implement 5G real-time IDSs.

To limit the detrimental consequences of a successful cyberattack, it is important not to use a centralized authentication and encryption server. Instead, it is useful to implement a distributed authentication and encryption architectures. These

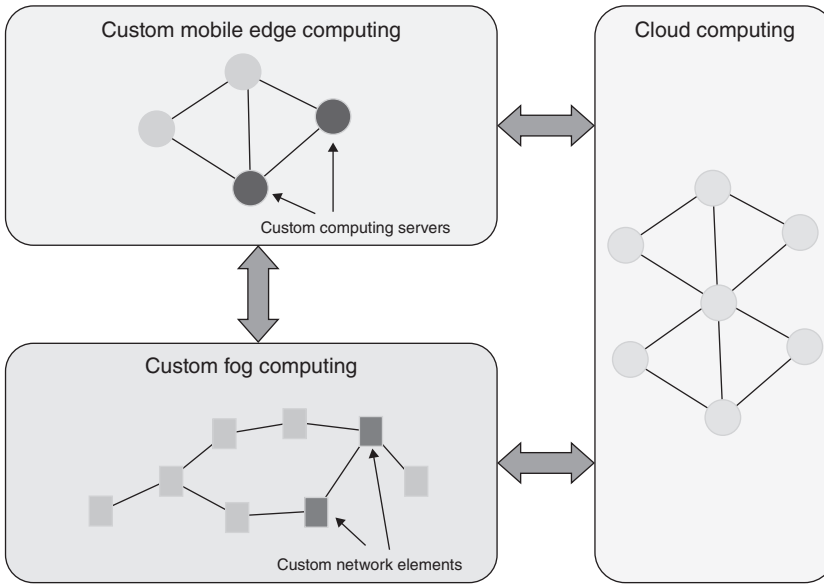

Figure 1.11 Employment of custom computing in mobile edge computing and fog computing. In both cases, custom computing can reduce network latency and power dissipation and enhance network bandwidth and digital security.

architectures can take advantage of custom computing technologies to implement complicated security algorithms.

1.3.3 Custom Edge Computing Cards

Custom edge computing systems can be employed in both mobile edge computing (control plane) and fog computing (user/data plane) (Figure 1.11).

In custom mobile edge computing architectures, one or more acceleration cards are added to COTS servers. These acceleration cards run the original software tasks in hardware, improving the performance by two or three orders of magnitude. At the same time, the power consumption can also be dramatically reduced (Figure 1.12).

In custom fog computing networks, special SOCs and FPGAs are added to the proprietary network computing elements. Oftentimes, the fog computing networks consist of proprietary embedded systems running non-standard operating systems. An emerging architecture is called the "Smart Network Interface" (or SmartNIC). Unlike a standard network interface card, a SmartNIC card may be FPGA based or SOC based and can be reconfigurable. A key feature of SmartNIC is minimizing the network latency (Figure 1.13).

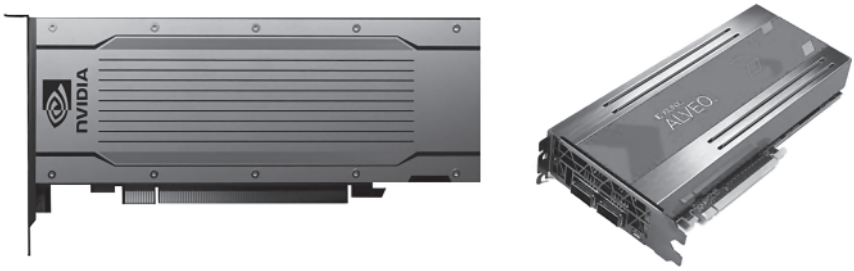

Figure 1.12 An NVIDIA Tesla T4 PCIe card and a Xilinx Alveo U200 PCIe card. In a 5G MEC or CP network, these acceleration cards can be used to increase performance and reduce overall power consumption.

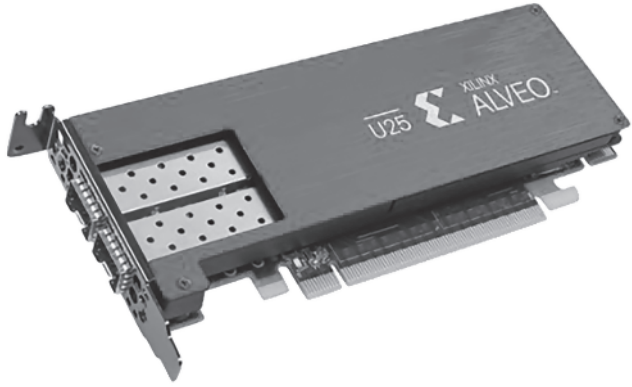

Figure 1.13 Xilinx Alveo U25 SmartNIC acceleration card encompasses network, storage, and compute functions, providing a comprehensive SmartNIC platform.

1.4 Deployment Considerations

According to MTN Consulting, the revenues of MNOs have remained stagnant in the last few years with a 1.6% decline in 2019 even before the onset of the COVID-19 pandemic. On the other hand, building a 5G base station was three times more expensive than building a 4G base station, costing somewhere between US$30K and US$70K in 2019. To make the situation even worse, a typical 5G base station consumes twice or more the power of a 4G base station [34–36].

Construction and operation costs are the most critical barriers to the deployment of 5G network infrastructures. The conundrum is how we can unleash the full 5G potential without breaking the bank or taking excessive risks.

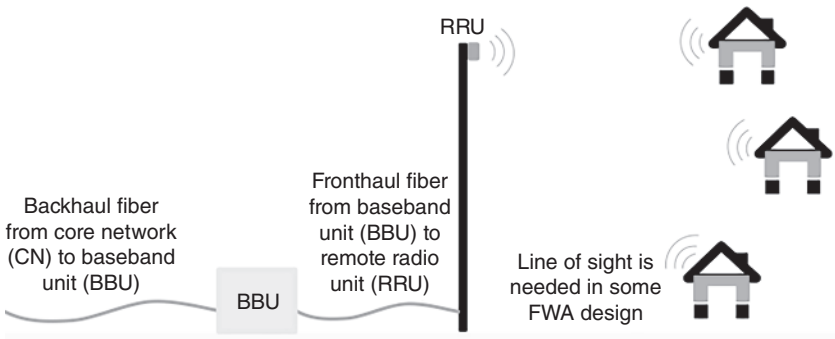

Figure 1.14 5G fixed wireless access (FWA) application.

What is the best way to deploy 5G? It depends on:

- Existing infrastructure.
- Use cases.
- Business models.

For example, some 5G equipment providers, including Ericsson and Huawei, actively promote the 5G fixed wireless access (FWA) market. Here, the household uses a 5G wireless router to provide Wi-Fi and LAN used within the house (Figure 1.14).

In this use case, there are two interesting points:

- First, the quality and reliability of the FWA network will depend on both the 5G wireless network and the fiber optic networks. Much of the 5G capital expenses are used to build the fiber optic infrastructure.
- Second, FWA is a good value proposition for many rural and sub-urban areas. However, FWA may not work well in some urban areas where fiber-to-the-premise (FTTP) is already commonplace. In urban areas, it makes sense to focus on other mobile applications.

In this section, we examine various 5G network deployment options and how they can affect the 5G network architecture.

1.4.1 5G/6G Cell Architecture

While the cell radius of a 4G network cell tower – called macrocell – varies between 2 and 40 km, a small cell may have a radius of less than 500 m [37–39]. A smaller 5G network cell can support higher traffic densities by taking advantage of:

- Higher frequency spectrums, which have shorter transmission ranges.
- More effective spectrum reuse across different cells.

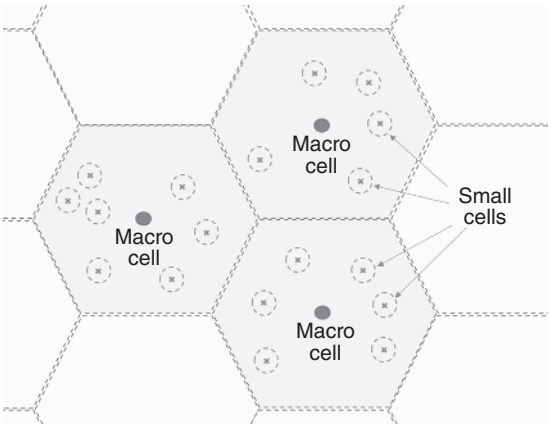

Figure 1.15 Small cells versus macrocells.

According to McKinsey's research, sites with traffic density above 0.5 petabytes per square kilometer per year will require a cell radius of less than 200 m. The traffic densities of many major cities are projected to reach 1–2 petabytes per square kilometer per year.

To leverage the existing infrastructure, most MNOs adopt a two-tier approach by [40, 41]:

- Upgrading the macrocells to support massive multi-input multi-output (mMIMO) protocols, with multiple antenna and beamforming technologies.
- Building small cells, as needed, within each macrocell (Figure 1.15).

These two approaches are complementary and somewhat independent. Several research papers discuss optimizing network resources. In reality, an MNO may decide a certain deployment option based on business considerations only.

There are different types of small cells:

- **Microcells:** with a radius of less than 2 km – are usually used to provide good wireless coverage within an in-building area, such as a train station or a shopping mall.
- **Picocells:** with a radius of less than 200 m – are usually used to cover a small area for a temporary event, such as a sporting event.
- **Femtocells:** with a radius of less than 10 m – are usually installed at home or in office.

Femtocells are often installed and maintained by end users, whereas microcells and picocells may be installed and maintained by either end users or MNOs. In all three cases, a 5G NR small cell must comply with 3GPP TS 38 series NG-RAN gNB specifications.

Here are some characteristics of a small cell:

- A small cell is connected to its 5G core network (5GC) by an NG interface, and the small cells are interconnected to each other by an Xn interface.
- Unlike a macrocell, a small cell may not support mMIMO with more than 16TX and 16RX.
- A small cell may support only 5G NR standards or both 4G/LTE and 5G NR standards.

1.4.2 5G/6G Private Network

In the previous section, we have discussed macrocells and small cells. Specifically, end users may install, maintain, and operate small cells. In this section, we will examine this specific market segment, called 5G private network.

There are a growing number of private 5G networks – aka non-public 5G networks – as we migrate from 4G/LTE to 5G. A private 5G network uses the same technology as public 5G networks but it allows an enterprise to manage its own 5G network to serve a limited geographic area with optimized services using dedicated equipment [42, 43].

A key component of a private 5G network is to provide:

- **Lower Latency:** In smart manufacturing, a private 5G network could be tailored for application delivery prioritization and to ensure predictable latency in all network traffic.
- **Higher Security:** A private 5G network could provide cellular-grade security, which can be used to keep sensitive data encrypted, authenticated, protected, and local to the premises.
- **Wider Coverage:** Many remote areas are not sufficiently covered by public cellular infrastructure. A private 5G network could be installed to provide wide coverage in agricultural fields, container ports, warehouses, or deep inside buildings.

Private 5G network use cases include:

- A smart factory can take advantage of a private 5G network to support services, such as predictive diagnostics, workforce management, machine automation, and factory safety management.
- A complex warehouse that fulfills pick-and-pack customer orders using robots and control software. A private 5G network is useful to help meet strict performance requirements needed to control countless fast-moving robots from a single base station.
- A large mine can use a private 5G network to support a complex heterogeneous system, comprising in-pit CCTV monitoring, intelligent earth sensing,

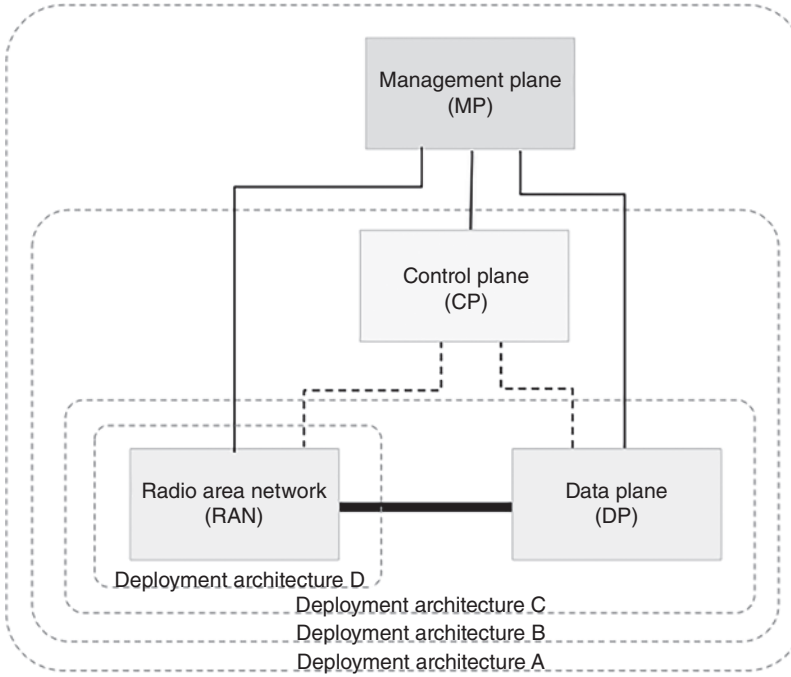

Figure 1.16 Deployment architectures A, B, C, and D.

telemetry, and many other monitoring services. This can help improve the reliability and safety of the mine.

There are many design alternatives in a private 5G network. Generally speaking, there are two main design aspects: deployment architecture and operation model [44–46]. Each design aspect has a few design options.

5G network architecture encompasses four main blocks:

- **Management Plane** (MP), aka management and orchestration layer, consists of network software managing a network domain.
- **Control Plane** (CP) consists of generic computers running a network operating system and controlling the network domain.
- **Data Plane** (DP), also known as the User Plane Function (UPF), consists of the configurable network devices within the network domain.
- Radio Access Network (RAN) consists of legacy and 5G NR (New Radio) interfaces supporting mMIMO and beamforming technologies (Figure 1.16).

In the most flexible private 5G network architecture (Deployment Architecture A), the above four blocks are all on-premise. While this deployment architecture is very secure and provides the best performance, it is also the most expensive. On the

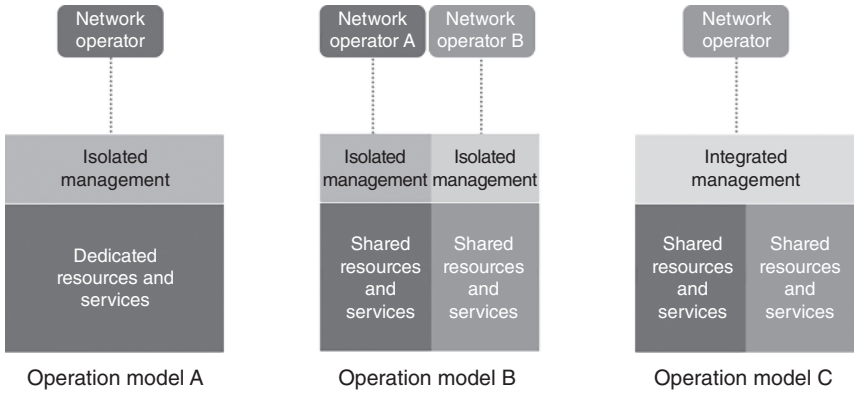

Figure 1.17 Different 5G operation models.

other extreme (Deployment Architecture D), only the RAN block is on-premise. The security level is the worst but the investment is the minimum (Table 1.2).

A private 5G network may adopt a certain operation model, independent from the deployment architecture discussed previously.

- First, the network infrastructure resources may be owned by a single user and used exclusively for a particular vertical application or are shared by multiple users and used for multiple vertical applications.
- Second, the network may be operated and managed independently from all other applications or by a third-party MNO that manages the other traffic at the same time (Figure 1.17 and Table 1.3).

1.4.3 Infrastructure Sharing

In the previous section, we have examined how a single private enterprise can launch its 5G wireless deployment using different private 5G/6G network architectures [42]. In this section, we will look at the 5G/6G network from a more global perspective.

The University of Cambridge and the Technical University of Madrid recently conducted research investigating various 5G deployment strategies in Britain. Specifically, the researchers focused on three questions:

- Should the MNOs increase or reduce the capital investment?
- Should the MNOs maintain the present infrastructure sharing?
- What is the reasonable 5G coverage expectation?

There are four MNOs sharing two macrocell networks in Britain:

- O_2/Telefonica and Vodafone are sharing one macrocell network.
- EE and Hutchison Three are sharing another macrocell network.

Table 1.3 Comparison of three 5G operation models.

Operation model	Operation and management	Network ownership	Usage simplicity	Service continuity	Liability protection	Security grade	Lock-in protection	Deployment flexibility
OM-A	Isolated	Dedicated	Worst	Worst	Worst	Best	Best	Best
OM-B	Isolated	Shared	Medium	Medium	Medium	Medium	Medium	Medium
OM-C	Integrated	Shared	Best	Best	Best	Worst	Worst	Worst

The table compares three different 5G private network operation models (OM-A, OM-B, and OM-C). OM-A is the most secure operating model; however, it requires the user to pay for its infrastructure and also to manage its network. This task may be too daunting for most users. OM-C is the cheapest and simplest operation model at the expense of less flexibility and security. OM-B is a compromise between OM-A and OM-C.

Table 1.4 Some CAPEX estimation for 5G deployment.

CAPEX model	CAPEX per year	CAPEX over 10 years
−25% model	£1.50 B	£15.0 B
−12.5% model	£1.75 B	£17.5 B
Baseline model	£2.00 B	£20.0 B
+12.5% model	£2.25 B	£22.5 B
+25% model	£2.50 B	£25.0 B

Together, the four MNOs are spending between £1.5 billion and £2.5 billion on capital expenses each year. For 5G deployment, the capital investment will primarily be spent on macrocell upgrades, small cell deployments, fiber backhaul, and core network upgrades. Using an average £2.0 billion spending per year as the baseline, the researchers projected the 5G coverage over the following years (Table 1.4 and Figure 1.18).

Based on the assumptions in this research, a large 5G investment is most effective in the first two years. Subsequently, the differences in between £1.5 billion per year and £2.5 billion per year are getting smaller and smaller, suggesting that an over-investment beyond the first two years has poor returns on investment (ROI).

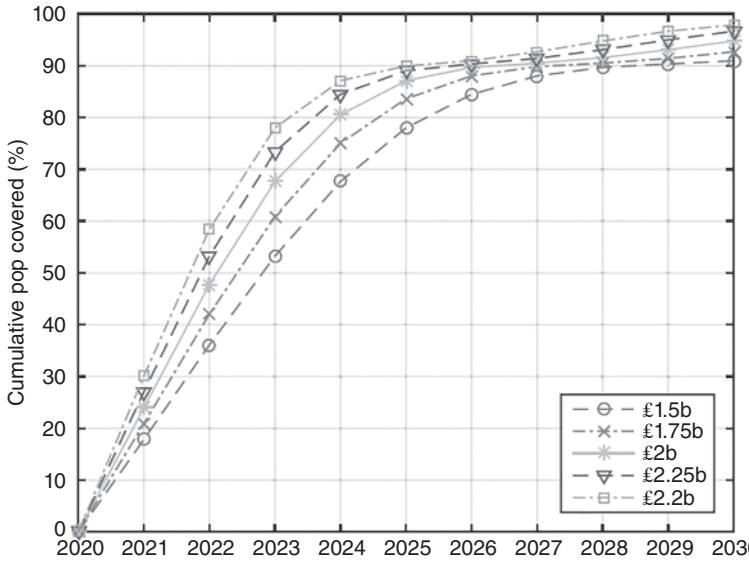

Figure 1.18 CAPEX requirements for various coverage models.

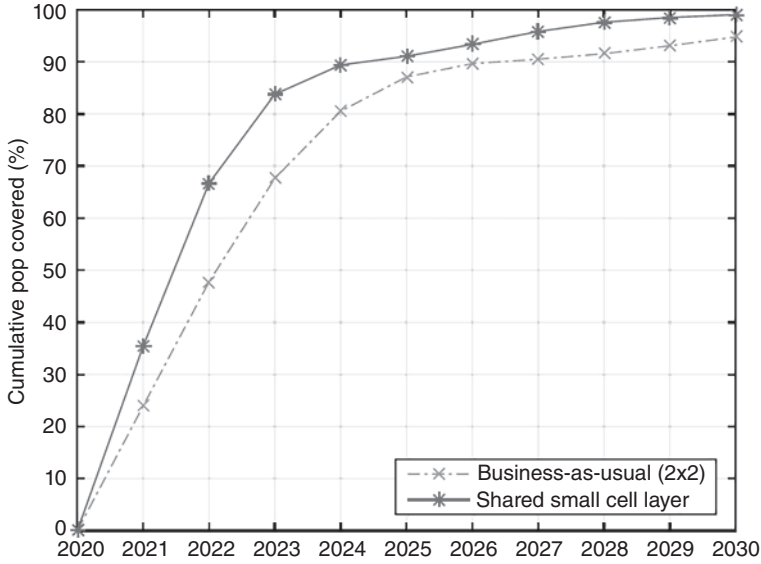

Figure 1.19 Optimizing coverage using shared small cell layer.

Second, the researchers investigated how the four MNOs can share the small cell infrastructure. This is made possible using the latest SDN/NFV technologies. The existing approach – with two independent small cell infrastructures – is too costly and ineffective. A single small cell infrastructure can quickly ramp up coverage in the first few years. More importantly, a single small cell infrastructure can reduce the deployment time by three or more years, when the coverage exceeds 90% (Figure 1.19).

Finally, the research suggests that providing 30 or 50 Mbps coverage for everyone in urban, suburban, and rural areas is too expensive. While it takes only £6 billion for the coverage of first 90% of the population, the remaining 10% will cost an additional £12 billion. A more sensible goal is to provide 50 Mbps for 90% of the population and 10 Mbps for the remaining 10% located in rural areas.

References

1 Wang, J., Jin, A., Shi, D. et al. (2017). Spectral efficiency improvement with 5G technologies: results from field tests. *IEEE Journal on Selected Areas in Communications* 35 (8): 1867–1875.

2 Holma, H., Toskala, A., and Nakamura, T. (2020). *5G Technology 3GPP New Radio*. Wiley.

3 Release description; Release 15. *3GPP Specification 21.915*, Alain Sultan, 2019.
4 Release description; Release 16. *3GPP Specification 21.916*, Alain Sultan, work-in-progress.
5 View on 5G architecture. Version 3.0. *5GPPP Architecture Working Group*, 4 February 2020. DOI 10.5281/zenodo.3265031.
6 5G spectrum access at 26 GHz and update on bands above 30 GHz. *Ofcom (UK)*, July 2017.
7 A. Sengupta, "Cellular terrestrial broadcast – physical layer evolution from 3GPP release 9 to release 16." *IEEE Transactions on Broadcasting*, Vol. 66, Issue 2, June 2020, pp. 459–470.
8 Pätzold, M. (2019). 5G is coming around the corner. *IEEE Vehicular Technology Magazine* 14 (1): 4–10.
9 Cisco Annual Internet Report (2018–2023) White Paper. *Cisco*, 9 March 2020.
10 Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022. *Cisco*, February 2019.
11 Bradley, J., Loucks, J., Macaulay, J., and Noronha, A. (2013). Internet of Everything (IoE) Value Index. *Cisco*.
12 Target networks in 5G era, embracing mobile network 2020s. *Huawei,* 2017.
13 Albrecht, K. and Michael, K. (2013). Connected: to everyone and everything. *IEEE Technology and Society Magazine*, Winter.
14 Raj, A. (2018). Internet of everything: a survey based on architecture, issues and challenges. *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2–4 November 2018, Gorakhpur India, pp. 1–6. IEEE.
15 Satyanarayanan, M. (2017). The emergency of edge computing. *IEEE Computer* 50 (1): 30–39.
16 Dolui, K. and Datta, S.K. (2017). Comparison of fog computing, cloudlet and mobile edge computing. *IEEE GIoTS* 2017, 6–9 June 2017, CICG Geneva Switzerland.
17 Bright, J. (2019). Driving new business opportunities with edge computing and 5G. *Ovum*.
18 Edge computing made simple. Release 10.0. *Small Cell Forum*, December 2017.
19 Baccarelli, E., Naranjo, P.G.V., Scarpiniti, M. et al. (2017). Fog of everything: energy-efficient networked computing architectures, research challenges, and a case study. *IEEE Access* (5): 9882–9910.
20 Aazam, M. and Huh, E.N. (2016). Fog computing: the cloud-IoT/IoE middleware paradigm. *IEEE Potentials* 35 (3): 40–44.
21 Abdelwahab, S. (2014). Enabling smart cloud services through remote sensing: an internet of everything enabler. *IEEE Internet of Things Journal* 1 (3): 276–288.

22 Kreutz, D., Ramos, F.M.V., Verissimo, P.E. et al. (2015). Software-defined networking: a comprehensive survey. *Proceedings of the IEEE* 103 (1): 14–76.

23 Mijumbi, R., Serrat, J., Gorricho, J.L. et al. (2016). Network function virtualization: state-of-the-art and research challenges. *IEEE Communications Surveys & Tutorials* 18 (1): 236–262.

24 Todman, T.J., Constantinides, G.A., Wilton, S.J. et al. (2005). Reconfigurable computing: architectures and design methods. *IEE Proceedings-Computers and Digital Techniques* 152 (2): 193–207.

25 Sutton, A. (2018). 5G network architecture, design and optimisation. *IET 5G Conference*, 24 January 2018, London UK. https://www.academia.edu/35765795/5G_Network_Architecture_Design_and_Optimisation. Accessed online from Academia on 1 September 2024.

26 The Evolution of Security in 5G. 5G Americas' Board of Governors, July 2019.

27 Surridge, M., Correndo, G., Meacham, K., et al. (2018). Trust modeling in 5G mobile networks. *SecSoN '18: Proceedings of the 2018 Workshop on Security in Softwarized Networks: Prospects and Challenges*, 24 August 2018, Budapest Hungary.

28 Arfaoui, G., Bisson, P., Blom, R. et al. (2018). A security architecture for 5G networks. *IEEE Access* 6: 22466–22479.

29 Benzaïd, C. and Taleb, T. (2020). ZSM security: threat surface and best practice. *IEEE Network* 34 (3): 124–133.

30 M. Liyanage, Ahmed, I., Ylianttila, M., et al. (2015). Security for future software defined mobile network. *Paper presented at the 9th International Conference on NGMAST*, 9–11 September 2015, University of Cambridge UK.

31 Liyanage, M., Ahmad, I., Abro, A.B. et al. (2018). *A Comprehensive Guide to 5G Security*. Wiley.

32 Wheeler, T. (2019). If 5G is so important, why isn't it secure? *The New York Times* January 2019, USA.

33 Adam Janofsky, "The good news about 5G security." *The Wall Street Journal*, 10 2019, USA.

34 Lavallée, B. (2016). 5G wireless needs fiber, and lots of it. *Ciena*, November 2016.

35 Grijpink, F., Ménard, A., Sigurdsson, H., and Vucevic, N. (2018). *The Road to 5G: The Inevitable Growth of Infrastructure Cost*. McKinsey & Company.

36 Duncan Stewart and Paul Lee. (2019). Consumers, operators gear up for 5G. *Deloitte*, April 2019.

37 Deployment issues for enterprise small cells. Release 10.0. *Small Cell Forum*, December 2017.

38 Small cell siting challenges and recommendations. Release 10.0. *Small Cell Forum*, August 2018.

39 Walker, M. (2020). Operators facing power cost crunch. *MTN Consulting*.

40 5G small cell architecture and product definitions. *Small Cell Forum*, July 2020.
41 URLLC and network slicing in 5G enterprise small cell networks. Release 10.0. *Small Cell Forum*, February 2018.
42 Private LTE/5G networks: a primer for developers. *Qualcomm*, 2019.
43 Rostami, A. Private 5G networks for vertical industries: deployment and operation models. *2019 IEEE 2nd 5G World Forum*, 30 September–2 October 2019, Dresden Germany.
44 The Office of Communications Annual Report and Accounts (2019/2020). *Ofcom (UK)*, July 2020.
45 Oughton, E. and Frias, Z. (2018). The cost, coverage and rollout implications of 5G infrastructure in Britain. *Telecommunications Policy* 42: 636–652.
46 Romero-Gázquez, J., Moreno-Muro, F.J., Garrich, M. et al. (2019). A use case of shared 5G backhaul segment planning in an urban area. *Paper presented at ICTON* 2019, 9–13 July 2019, Angers France.