2

IoT Architecture and Core IoT Modules

The edge computing and IoT ecosphere starts with the simplest of sensors located in the remotest corners of the earth and translates analog physical effects into digital signals (the language of the Internet). Data then takes a complex journey through wired and wireless signals, various protocols, natural interference, and electromagnetic collisions, before arriving in the ether of the Internet. From there, packetized data will traverse various channels arriving at a cloud or large data center. The strength of IoT is not just one signal from one sensor, but the aggregate of hundreds, thousands, potentially millions of sensors, events, and devices.

This chapter starts with a definition of IoT versus machine-to-machine architectures. It also addresses the architect's role in building a scalable, secure, and enterprise IoT architecture. To do that, an architect must be able to speak to the value the design brings to a customer. The architect must also play multiple engineering and product roles in balancing different design choices.

This chapter provides an outline to how the book is organized and how an architect should approach reading the book and performing their role as an architect. The book treats architecture as a holistic exercise involving many systems and domains of engineering. This chapter will highlight:

- **Sensing and power**: We cover the transformation of physical to digital sensing, power systems, and energy storage.
- **Data communication**: We delve into the communication of devices using near-meter, near-kilometer, and extreme-range communication systems and protocols as well as networking and information theory.
- **Edge computing**: Edge devices have multiple roles from routing, to gateways, edge processing and cloud-edge (fog) interconnect. We examine the role of the edge and how to successfully build and partition edge machines. We also look at communication protocols from the edge to the cloud.
- **Compute, analytics and machine learning**: We then examine dataflow through cloud and fog computing, as well as advanced machine learning and complex event processing.
- **Threat and security**: The final content investigates security and the vulnerability of the largest attack surface on earth.

A connected ecosystem

Nearly every major technology company is investing or has invested heavily in IoT and the edge computing space. New markets and technologies have already formed (and some have collapsed or been acquired). Throughout this book, we will touch on nearly every segment in information technology, as they all have a role in IoT.

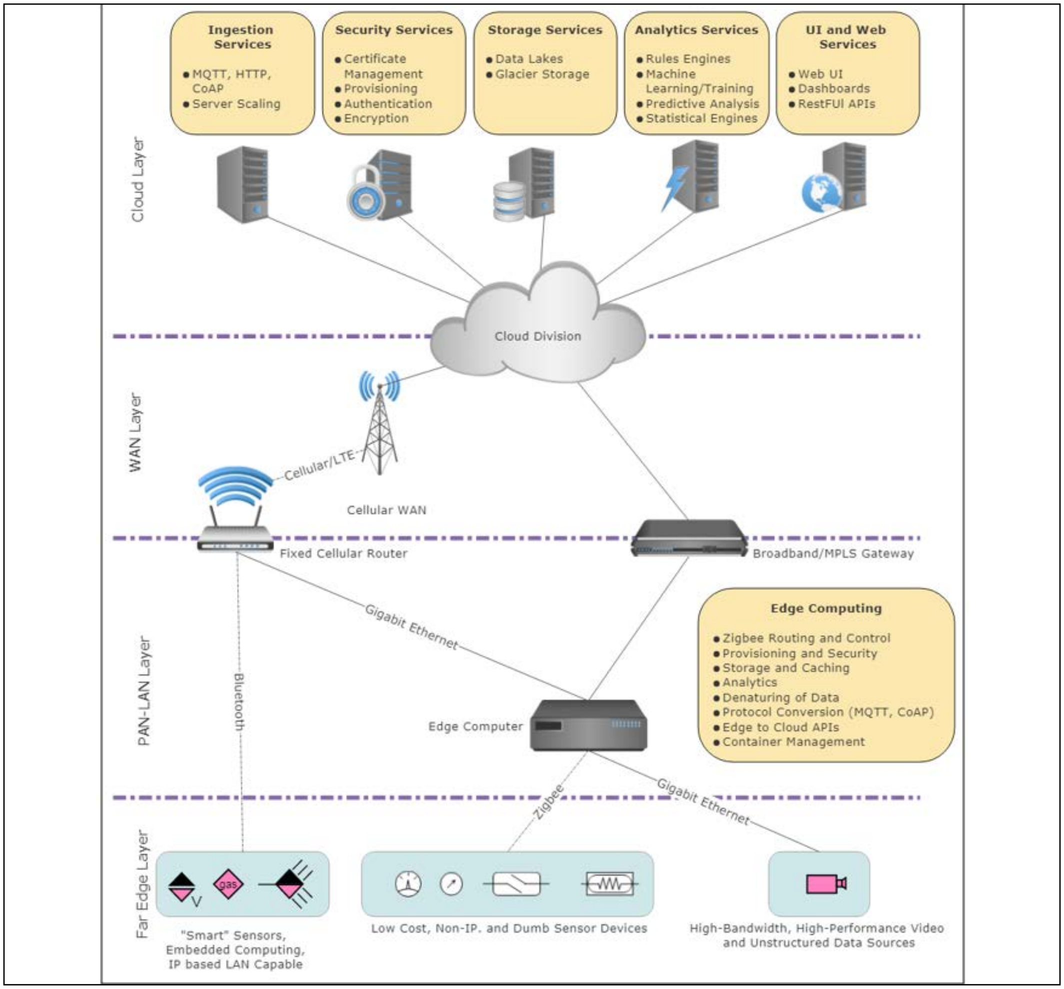

Figure 1: Example of the architectural layers of an IoT/edge computing system. This is one of the many potential configurations that must be considered by the architect. Here we show the sensor-to-cloud routes through direct communication and through edge gateways. We also highlight the functionality provided by the edge compute nodes and the cloud components.

As illustrated in the preceding figure, here are some of the components within an IoT/edge solution that we will study:

- **Sensors, actuators, and physical systems**: Embedded systems, real-time operating systems, energy-harvesting sources, micro-electro-mechanical systems (MEMs).
- **Sensor communication systems**: **Wireless personal area networks** (**WPANs**) reach from 0 cm to 100 m. Low-speed and low-power communication channels, often non-IP based, have a place in sensor communication.
- **Local area networks (LANs)**: Typically, IP-based communication systems such as 802.11 Wi-Fi used for fast radio communication, often in peer-to-peer or star topologies.
- **Aggregators, routers, gateways**: Embedded systems providers, cheapest vendors
- **Wide area networks (WANs)**: Cellular network providers using LTE or Cat M1, satellite network providers, low-power **wide-area network (LPWAN)** providers like Sigfox or LoRa. They typically use Internet transport protocols targeted for IoT and constrained devices like MQTT, CoAP, and even HTTP.
- **Edge computing**: Distributing computing from on-premise data centers and cloud to closer to the sources of data (sensors and systems). This is to remove latency issues, improve response time and real-time systems, manage the lack of connectivity, and build redundancy of a system. We cover processors, DRAM, and storage. We also study module vendors, passive component manufacturers, thin client manufacturers, cellular and wireless radio manufacturers, middleware providers, fog framework providers, edge analytics packages, edge security providers, certificate management systems, WPAN to WAN conversion, routing protocols, and software-defined networking/software-defined perimeters.
- **Cloud**: Infrastructure as a service provider, platform as a service provider, database manufacturers, streaming and batch processing manufacturers, data analytics packages, software as a service provider, data lake providers, and machine learning services.
- **Data analytics**: As the information propagates to the cloud en masse, dealing with volumes data and extracting value is the job of complex event processing, data analytics, and machine learning techniques. We study different edge and cloud analytic techniques from statistical analysis and rules engines to more advanced machine learning techniques.

- **Security**: Tying the entire architecture together is security. End-to-end security from edge hardening, protocol security, to encryption. Security will touch every component from physical sensors to the CPU and digital hardware to the radio communication systems to the communication protocols themselves. Each level needs to ensure security, authenticity, and integrity. There cannot be a weak link in the chain, as the IoT will form the largest attack surface on earth.

This ecosystem will need talents from the body of engineering disciplines, such as:

- Device physics scientists developing new sensor technologies and many-year batteries
- Embedded system engineers working on driving sensors at the edge
- Network engineers capable of working in a personal area network or wide area network, as well as on a software-defined networking
- Data scientists working on novel machine learning schemes at the edge and in the cloud
- DevOps engineers to successfully deploy scalable cloud solutions as a well as *fog* solutions

IoT will also need service vendors such as solution provision firms, system integrators, value-added resellers, and OEMs.

IoT versus machine-to-machine versus SCADA

One common area of confusion in the IoT world is what separates it from the technologies that define **machine to machine** (**M2M**). Before IoT became part of the mainstream vernacular, M2M was the hype. Well before M2M, **SCADA** (**supervisory control and data acquisition**) systems were the mainstream of interconnected machines for factory automation. While these acronyms refer to interconnected devices and may use similar technologies, there are differences. Let's examine these more closely:

- **M2M**: It is a general concept involving an autonomous device communicating directly to another autonomous device. *Autonomous* refers to the ability of the node to instantiate and communicate information with another node without human intervention. The form of communication is left open to the application. It may very well be the case that an M2M device uses no inherent services or topologies for communication. This leaves out typical Internet appliances used regularly for cloud services and storage. An M2M system may communicate over non-IP based channels as well, such as a serial port or custom protocol.

- **IoT**: IoT systems may incorporate some M2M nodes (such as a Bluetooth mesh using non-IP communication), but they aggregate data at an edge router or gateway. An edge appliance like a gateway or router serves as the entry point onto the Internet. Alternatively, some sensors with more substantial computing power can push the Internet networking layers onto the sensor itself. Regardless of where the Internet *on-ramp* exists, the fact that it has a method of tying into the Internet fabric is what defines IoT.
- **SCADA**: This term refers to supervisory control and data acquisition. These are industrial control systems that have been used in factory, facility, infrastructure and manufacturing automation since the 1960s. They typically involve **programmable logic controllers** (**PLCs**) that monitor or controls various sensors and actuators on machinery. SCADA systems are distributed and only recently have been connected to Internet services. This is where Industry 2.0 and the new growth of manufacturing is taking place. These systems use communication protocols such as ModBus, BACNET, and Profibus.

By moving data onto the Internet for sensors, edge processors, and smart devices, the legacy world of cloud services can be applied to the simplest of devices. Before cloud technology and mobile communication became mainstream and cost-effective, simple sensors and embedded computing devices in the field had no good means of communicating data globally in seconds, storing information for perpetuity, and analyzing data to find trends and patterns. As cloud technologies advanced, wireless communication systems became pervasive, new energy devices like lithium-ion became cost-effective, and machine learning models evolved to produce actionable value. This greatly improved the IoT value proposition. Without these technologies coming together when they did, we would still be in an M2M world.

The value of a network and Metcalfe's and Beckstrom's laws

It has been argued that the value of a network is based on Metcalfe's law. Robert Metcalfe in 1980 formulated the concept that the value of any network is proportional to the square of connected *users* of a system. In the case of IoT, "users" may mean sensors or edge devices with some form of communication.

Generally speaking, Metcalfe's law is represented as:

$$V \propto N^2$$

Where:

- V = Value of the network
- N = Number of nodes within the network

A graphical model helps to understand the interpretation as well as the crossover point, where a positive **return on investment** (**ROI**) can be expected:

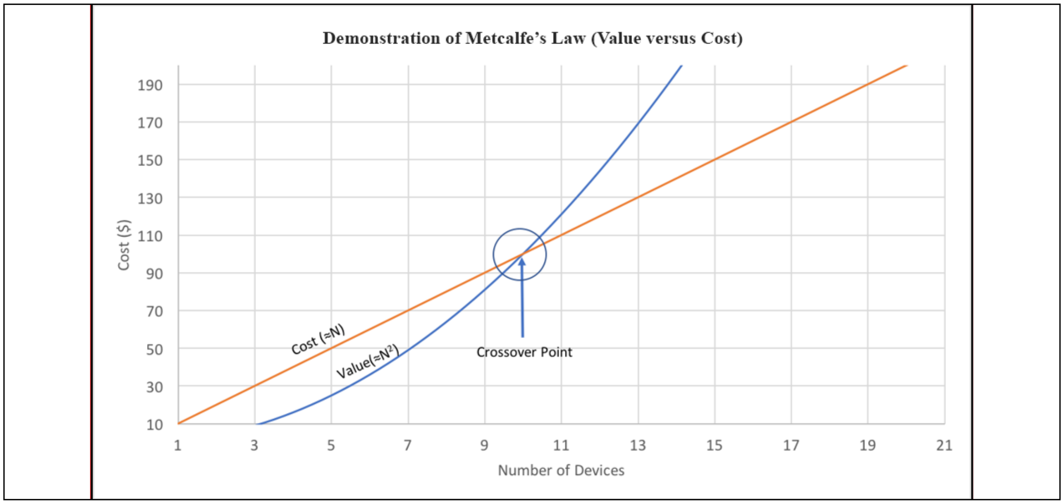

Figure 2: Metcalfe's law: The value of a network is represented as proportional to N^2. The cost of each node is represented as kN where k is an arbitrary constant. In this case, k represents a constant of $10 per IoT edge sensor. The key takeaway is the crossover point occurs rapidly due to the expansion of value and indicates when this IoT deployment achieves a positive ROI.

An example validating Metcalfe's law to the value of blockchains and cryptocurrency trends was recently conducted. We will go much deeper into blockchains in the security chapter.

A recent white paper by Ken Alabi finds that blockchain networks also appear to follow Metcalfe's law, *Electronic Commerce Research and Applications*, Volume 24, C (July 2017), page number 23-29.

Metcalfe's law does not account for service degradation in cases in which service degrades as the number of users and/or data consumption grows, but the network bandwidth does not. Metcalfe's law also doesn't account for various levels of network service, unreliable infrastructure (such as 4G LTE in a moving vehicle), or bad actors affecting the network (for example, denial of service attacks).

To account for these circumstances, we use Beckstrom's law:

$$\sum_{i=1}^{n} V_{i,j} = \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{B_{i,j,k} - C_{i,j,k}}{(1+r_k)^{t_k}}$$

Where:

- $V_{i,j}$: Represents the present value of the network for device *i* on network *j*
- *i*: An individual user or device on the network
- *j*: The network itself
- *k*: A single transaction
- $B_{i,j,k}$: The benefit that value *k* will bring to device *i* on network *j*
- $C_{i,j,k}$: The cost of a transaction *k* to a device *i* on network *j*
- r_k: The discount rate of interest to the time of transaction *k*
- t_k: The elapsed time (in years) to transaction *k*
- *n*: The number of individuals
- *m*: The number of transactions

Beckstrom's law teaches us that to account for the value of a network (for example, an IoT solution), we need to account for all transactions from all devices and sum their value. If the network *j* goes down for whatever reason, what is the cost to the users? This is the impact an IoT network brings and is a more representative real-world attribution of value. The most difficult variable to model in the equation is the benefit of a transaction *B*. While looking at each IoT sensor, the value may be very small and insignificant (for example, a temperature sensor on some machine is lost for an hour). At other times, it can be extremely significant (for example, a water sensor battery died, and a retailer basement is flooded, causing significant inventory damage and insurance adjustments).

An architect's first step in building an IoT solution should be to understand what value they are bringing to what they are designing. In the worst case, an IoT deployment becomes a liability and actually produces negative value for a customer.

IoT and edge architecture

The coverage in this book will span many technologies, disciplines, and levels of expertise. As an architect, one needs to understand the impact that choosing a certain design aspect will have on scalability and other parts of the system. The complexities and relationships of IoT technologies and services are significantly more intercoupled than traditional technologies not only because of the scale but also due to the disparate types of architecture. There is a bewildering number of design choices. For example, as of this writing, we counted over 700 IoT service providers alone offering cloud-based storage, SaaS components, IoT management systems, middleware, IoT security systems, and every form of data analytics one can imagine. Add to that the number of different PAN, LAN, and WAN protocols that are constantly changing and varying by region. Choosing the wrong PAN protocol could lead to poor communications and significantly low signal quality that can only be resolved by adding more nodes to complete a mesh. The role of an architect should ask and provide solutions for problems that span the system as a whole:

- The architect needs to consider interference effects in the LAN and WAN—how will the data get off the edge and on the Internet?
- The architect needs to consider resiliency and how costly the loss of data is. Should resiliency be managed within the lower layers of the stack, or in the protocol itself?
- The architect must also make choices of Internet protocols such as MQTT versus CoAP and AMQP, and how that will work if he or she decides to migrate to another cloud vendor.

Choices also need consideration with regards to where processing should reside. This opens up the notion of edge/fog computing to process data close to its source to solve latency problems, but more importantly to reduce bandwidth and costs of moving data over WANs and clouds. Next, we consider all the choices in analyzing the data collected. Using the wrong analytic engine may result in useless noise or algorithms that are too resource-intensive to run on edge nodes. How will queries from the cloud back to the sensor affect the battery life of the sensor device itself? Add to this litany of choice, and we must layer on security as the IoT deployment we have built is now the largest attack surface in our city. As you can see, the choices are many and have relationships with one another.

There are many choices to consider. When you account for the number of edge computing systems and routers, PAN protocols, WAN protocols, and communication, there are over 1.5 million different combinations of architectures to choose from:

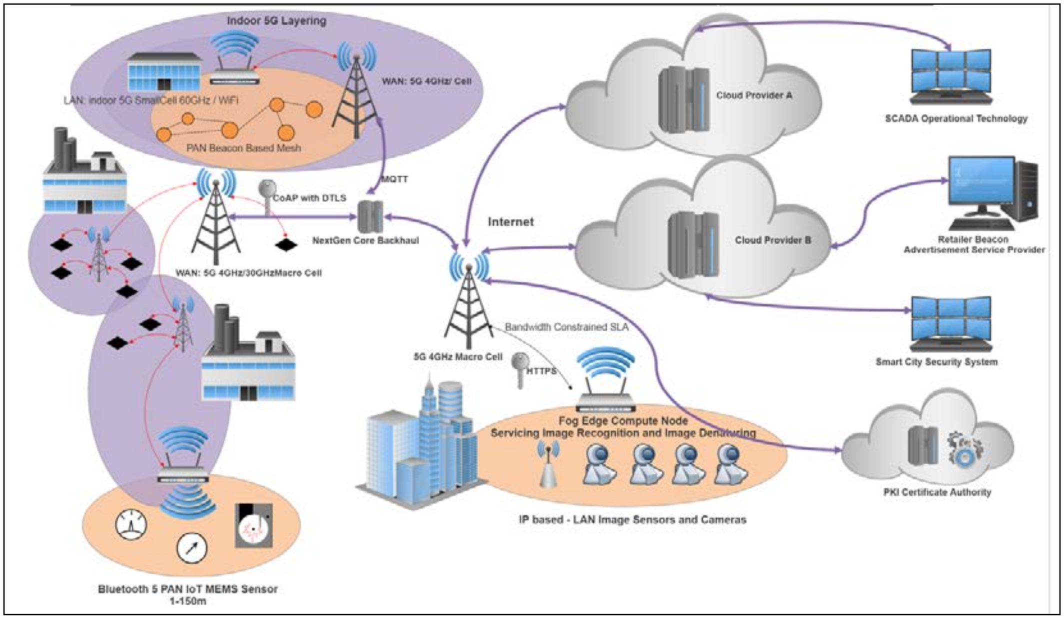
Indoor 5G Layering
WAN: 5G 4GHz/ Cell
LAN: indoor 5G SmallCell 60GHz / WiFi
PAN Beacon Based Mesh
MQTT
CoAP with DTLS
NextGen Core Backhaul
WAN: 5G 4GHz/30GHzMacro Cell
Internet
Cloud Provider A
SCADA Operational Technology
Cloud Provider B
Retailer Beacon
Advertisement Service Provider
Bandwidth Constrained SLA
5G 4GHz Macro Cell
HTTPS
Smart City Security System
Fog Edge Compute Node
Servicing Image Recognition and Image Denaturing
PKI Certificate Authority
IP based - LAN Image Sensors and Cameras
Bluetooth 5 PAN IoT MEMS Sensor
1-150m

Figure 3: IoT design choices: The full spectrum of various levels of IoT architecture from the sensor to the cloud and back.

Role of an architect

The term **architect** is often used in technical disciplines. There are software architects, system architects, and solution architects. Even within specific domains, such as computer science and software engineering, you may see people with the title SaaS architect, cloud architect, data science architect, and so on. These are individuals who are recognized experts with tangible skills and experience in a domain. These types of specialized vertical domains cross several horizontal technologies. In this book, we are targeting the IoT architect.

This is a horizontal role, meaning it will touch a number of these domains and bring them together for a usable, secure, and scalable system.

> We will go as deep as necessary to understand an entire IoT system and bring a system together. At times, we will go into pure theory, such as information and communication theory. Other times, we will brush on topics that are on the periphery of IoT systems or are rooted in other technologies. By reading and referencing this book, the architect will have a go-to guide on different aspects of IoT that are all needed to build a successful system.

Whether you are disciplined in electrical engineering or computer science, or have domain expertise in cloud architectures, this material will help you understand a holistic system—which should be, by definition, part of the role of an architect.

This book is also intended for geographically global and massive scaling. It is one thing to build a proof of concept with one or two endpoint devices. It is by far a different challenge to build an IoT solution that stretches multiple continents, different service providers, and thousands of endpoints. While every topic can be used for hobbyist and maker movements, this is intended to scale to global enterprise systems on the order of thousands to millions of edge devices.

The architect will ask questions for the full stack of connected systems. He or she will be aware of how optimizing for one solution may in fact deliver a less than desirable effect in another part of the system.

For example:

- Will the system scale and to what capacity? This will affect decisions on wide area networking, edge-to-cloud protocols, and middleware-to-cloud provisioning systems.
- How will the system perform with loss of connectivity? This will impact the choices of edge systems, storage components, 5G service profiles, and network protocols.
- How will the cloud manage and provision edge devices? This will affect decisions on edge middleware and fog components, and security services.
- How will my customer's solution work in a noisy RF environment? This will affect decisions on PAN communications and edge components.
- How will software be updated on a sensor? This will affect decisions around security protocols, edge hardware and storage, PAN network protocols, middleware systems, sensor costs and resources, and cloud provisioning layers.

- What data will be useful to improving my customer's performance? This will affect decisions on what analytics tools to use, where to analyze the data, how data will be secured and denatured, and edge/cloud partitioning.
- How will devices, transactions, and communication be secured from end to end?

Part 1 – Sensing and power

An IoT transaction starts or ends with an event: a simple motion, a temperature change, perhaps an actuator moving on a lock. Unlike many IT devices in existence, IoT in a large part is about a physical action or event. It responds to affect a real-world attribute. Sometimes this involves considerable data being generated from a single sensor, such as auditory sensing for preventative maintenance of machinery. Other times, it's a single bit of data indicating vital health data from a patient. Whatever the case may be, sensing systems have evolved and made use of Moore's law in scaling to sub-nanometer sizes and significantly reduced costs. Part 1 explores the depths of MEMs, sensing, and other forms of low-cost edge devices from a physical and electrical point of view. The part also details the necessary power and energy systems to drive these edge machines. We can't take power for granted at the edge. Collections of billions of small sensors will still require a massive amount of energy in total to power. We will revisit power throughout this book, and how innocuous changes in the cloud can severely impact the overall power architecture of a system.

Part 2 – Data communication

A significant portion of this book surrounds connectivity and networking. There are countless other sources that dive deep into application development, predictive analytics, and machine learning. This book too will cover those topics, but an equal amount of emphasis is given to data communications. The IoT wouldn't exist without significant technologies to move data from the remotest and most hostile environment to the largest data centers at Google, Amazon, Microsoft, and IBM. The acronym IoT contains the word *Internet*, and because of that, we need to dive deep into networking, communications, and even signal theory. The starting point for IoT isn't sensors or the application; it's about connectivity, as we will see throughout this book. A successful architect will understand the constraints of Internetworking from a sensor to a WAN and back again.

This communication and networking section starts with theory and mathematical foundations of communication and information. Preliminary tools and models are needed by a successful architect not only to understand why certain protocols are constrained, but also to design future systems that scale successfully at IoT levels.

These tools include wireless radio dynamics like range and power analysis, signal-to-noise ratio, path loss, and interference. Part 2 also details foundations of information theory and constraints that affect overall capacity and quality of data. The foundations of Shannon's law will be explored. The wireless spectrum is also finite and shared, so an architect deploying a massive IoT system will need to understand how the spectrum is allocated and governed.

Theory and models explored in this part will be reused in other parts of the book.

Data communication and networking will then build up from the near-range and near-meter communication systems known as **personal area networks** (**PANs**), typically using non-Internet protocol messages. The chapter on PAN will include the new Bluetooth 5 protocol and mesh, as well as Zigbee and Z-Wave in depth. These represent the plurality of all IoT wireless communication systems. Next, we explore wireless local area networks and IP-based communication systems including the vast array of IEEE 802.11 Wi-Fi systems, thread, and 6LoWPAN. The chapter also investigates new Wi-Fi standards such as 802.11p for in-vehicle communication.

The part concludes with long-range communication using cellular (4G LTE) standards, and dives deep into the understanding and infrastructure to support 4G LTE and new standards dedicated to IoT and machine-to-machine communication, such as Cat-1 and Cat-NB. The last chapter also covers the 5G standard and publicly licensed cellular (MulteFire) to prepare the architect for future long-range transmissions where every device is connected in some capacity. A proprietary protocol like LoRaWAN and Sigfox are also explored to understand the differences between architectures.

Part 3 – Edge computing

Edge computing brings nontraditional computing power close to the sources of data. While embedded systems have existed in devices for the last 40 years, edge computing is more than a simple 8-bit microcontroller or analog-to-digital converter circuit used to display temperature. Edge computing attempts to solve critical problems as the number of connected objects and the complexity of use cases grows in the industry. For example, in IoT areas we need the following:

- Accumulate data from several sensors and provide an entry point to the Internet.
- Resolve critical real-time responses for safety-critical situations like remote surgery or automated driving.

- Solutions that can manage an overwhelming amount of processing of unstructured data like video data or even streaming of video to save on costs of transporting the data over wireless carriers and cloud providers.

Edge computing also comes in layers as we will examine with 5G infrastructure, multiaccess edge computing, and fog computing.

We will closely examine the hardware, operating systems, mechanics, and power that an architect must consider for different edge systems. For example, an architect may need a system that delivers on a constraining cost and power requirement but may forgo some processing ability. Other designs may need to be extremely resilient as the edge computer may be in a very remote region and essentially need to manage itself.

To bridge data from sensors to the Internet, two technologies are needed: gateway routers and supporting IP-based protocols designed for efficiency. This part explores the role of router technologies at the edge for bridging sensors on a PAN network to the Internet. The role of the router is especially important in securing, managing, and steering data. Edge routers orchestrate and monitor underlying mesh networks and balance and level data quality. The privatization and security of data is also critical. Part 3 will explore the router role in creating virtual private networks, virtual LANs, and software-defined wide area networks. There literally may be thousands of nodes serviced by a single edge router, and in a sense, it serves as an extension to the cloud, as we will see in the *Chapter 11*, *Cloud and Fog Topologies*.

This part continues with the protocols used in IoT communication between nodes, routers, and clouds. The IoT has given way to new protocols rather than the legacy HTTP and SNMP types of messaging used for decades. IoT data needs efficient, power-aware, and low-latency protocols that can be easily steered and secured in and out of the cloud. This part explores protocols such as the pervasive MQTT, as well as AMPQ and CoAP. Examples are given to illustrate their use and efficiency.

Part 4 – Compute, analytics, and machine learning

At this point, we must consider what to do with the data streaming in from edge nodes into a cloud service. First, we begin by talking about the aspects of cloud architectures such as SaaS, IaaS, and PaaS systems. An architect needs to understand the data flow and typical design of cloud services (what they are and how they are used). We use OpenStack as a model of cloud design and explore the various components from ingestor engines to data lakes to analytics engines.

Understanding the constraints of cloud architectures is also important to make a good judgment on how a system will deploy and scale. An architect must also understand how latency can affect an IoT system. Alternatively, not everything belongs in the cloud. There is a measurable cost in moving all IoT data to a cloud versus processing it at the edge (edge processing) or extending cloud services downward into an edge computing device (fog computing). This part dives deep into new standards of fog computing such as the OpenFog architecture.

Data that has been transformed from a physical analog event to a digital signal may have actionable consequences. This is where the analytics and rules engines of the IoT come in to play. The level of sophistication for an IoT deployment is dependent on the solution being architected. In some situations, a simple rules engine looking for anomalous temperature extremes can easily be deployed on an edge router monitoring several sensors. In other situations, a massive amount of structured and unstructured data may be streaming in real time to a cloud-based data lake, and require both fast processing for predictive analytics and long-range forecasting using advanced machine learning models, such as recurrent neural networks in a time-correlated signal analysis package. This part details the uses and constraints of analytics from complex event processors to Bayesian networks to the inference and training of neural networks.

Part 5 – Threat and security in IoT

We conclude the book with a survey of IoT compromises and attacks. In many cases, IoT systems will not be secured in a home, or in a company. They will be in public, in very remote areas, in moving vehicles, or even inside a person. The IoT represents the single biggest attack surface for any type of cyberattack. We have seen countless academic hacks, well-organized cyber assaults, and even nation-state security breaches with IoT devices being the target. Part 5 will detail several aspects of such breaches and the types of remediation any architect must consider when making a consumer or enterprise IoT deployment a good citizen of the Internet. We explore the proposed congressional act to secure the IoT and understand the motivation and impact of such a government mandate.

This part will checklist the typical security provisions needed for IoT, or any network component. Details of new technologies such as blockchains and software-defined perimeters will also be explored to provide insight into future technologies that will be needed to secure the IoT.

Summary

This book will bridge the spectrum of technologies that comprise edge computing and the IoT. In this chapter, we summarized the domains and topics covered in the book. An architect must be cognizant of the interactions between these disparate engineering disciplines to build a system that is scalable, robust, and optimized. An architect will also be called upon to provide supporting evidence that the IoT system provides a value to the end user or the customer. Here, we learned about the application of Metcalfe's and Beckstrom's laws as tools for supporting an IoT deployment.

In the next chapters, we will learn about communication from sensors and edge nodes to the Internet and cloud. First, we will examine the fundamental theory behind radio signals and systems and their constraints and limits, and then we will dive into near-range and long-range wireless communication.