UNDERSTANDING INFRASTRUCTURE EDGE COMPUTING

CONCEPTS, TECHNOLOGIES AND CONSIDERATIONS

ALEX MARCHAM

WILEY

Understanding Infrastructure Edge Computing

Understanding Infrastructure Edge Computing

Concepts, Technologies and Considerations

Alex Marcham

WILEY

This edition first published 2021

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data
Names: Marcham, Alex, author.
Title: Understanding infrastructure edge computing : concepts, technologies
and considerations / Alex Marcham.
Description: Hoboken, NJ, USA : Wiley, 2021. | Includes bibliographical
references and index.
Identifiers: LCCN 2020050691 (print) | LCCN 2020050692 (ebook) | ISBN
9781119763239 (hardback) | ISBN 9781119763246 (adobe pdf) | ISBN
9781119763253 (epub)
Subjects: LCSH: Edge computing.
Classification: LCC QA76.583 .M37 2021 (print) | LCC QA76.583 (ebook) |
DDC 005.75/8–dc23
LC record available at https://lccn.loc.gov/2020050691
LC ebook record available at https://lccn.loc.gov/2020050692

Cover Design: Wiley
Cover Image: © Metamorworks/Shutterstock

Set in 10/13.5pt STIXTwoText by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

To the Fun Police. Careful!

Contents

Preface

How to Use This Book

This book is intended to be read from start to finish in order for the reader to get the most benefit from all of the subject areas which it covers. However, for information on a specific topic, each of the chapters in this book can be read in a relatively standalone manner. There is crossover between chapters in many cases, for example, between a section on the physical redundancy of an edge data centre facility in one chapter and a section describing infrastructure edge computing network level resiliency in another, where if the reader has not read the prior section, some context may be lost.

I hope however you choose to read it that you enjoy reading this book as much as I did when writing it.

About This Book

As with any emerging area of technology, the information presented within this book represents a moment in time and the best practices available at that moment in time. The information here is represented to the best of the author's knowledge and does not favour one vendor over another.

Audience

This book was written for an audience of technologists, decision makers, and engineers in the fields of telecommunications, networking, data centres, and application development and operation who are interested in new emerging areas of technology, such as edge computing, fifth generation (5G), and distributed artificial intelligence (AI).

About the Author

Alex Marcham has been in the networking industry for over a decade working on wireless networks, enterprise networks, telecommunications, and edge computing. He created the terms infrastructure edge and device edge and was the primary author of the Open Glossary of Edge Computing, which is now a Linux Foundation project. When not at work, he can often be seen hiking somewhere remote.

Acknowledgements

This book would not have come to fruition were it not for the help of a few special people.

First, I would like to thank the friends whom I share each day with as we all do our best to keep each other moderately sane from one week to the next. I'll always do my best to listen and help you as you each do for me, and I wish you all the greatest happiness and success in life. That is, unless one of you says that my hair is rubbish again, in which case we will be forced to engage in a cage fight.

Second, thank you to my family. Although we may spend a lot of time apart, physical distance is no match for our combined love of badgers, elephants, and hummingbirds. That said, it is a lot easier to maintain a set of hummingbird feeders than it would be to provide for a load of badgers or a passing herd of elephants, but this is matched by the difficulty of photographing any hummingbird properly.

Third, thanks to the team at Wiley for their insight and support for this project from start to finish. The telepathic portion of this book will be available at a later date, so this will have to do for now.

Finally, thanks to everyone I have spoken to and learned from on the topics of engineering, writing, and life in the past three decades across the world. We are the sum of our choices and experiences.

1

Introduction

Few could have guessed the impact the internet would have on us all at its inception. Today, the internet and the services it provides are essential for billions of people across the world. It is a primary source of communication with friends, family, and our communities; it is the primary way in which we access many essential services, as well as the way that increasing numbers of us go to work, pursue our educational goals, and access sources of entertainment, all on demand.

We did not get to this point by accident. Although the current state of the internet could not have been fully foreseen decades ago, it is due to the continuous efforts of skilled and driven people from across many different disciplines that the modern internet is able to support us as it does today. The story of the internet is not one of a single grand original design; it is one of consistent iteration and ingenuity to adapt to new technical and business challenges which have emerged over the decades.

As they have in the past, new and emerging use cases are driving the evolution of internet and data centre technology. This is resulting in new generations of infrastructure which are reimagining how the internet that we all use on a daily basis should be designed, deployed, and operated as a whole.

Distributed artificial intelligence (AI) and machine learning (ML) are set to permanently reshape how many industries, from healthcare and retail to manufacturing and construction, operate due to their ability to enhance the decision-making process and automate difficult tasks with extraordinary speed and precision. City-scale internet of things (IoT) and cyber-physical systems provide machines the means to interact physically with our world in ways that have been impossible or impractical to achieve before, supported by fifth generation (5G) cellular network connectivity and new versions of cloud computing, which are able to support high-bandwidth, low-latency, and real-time use cases.

Understanding Infrastructure Edge Computing: Concepts, Technologies and Considerations,
First Edition. Alex Marcham.

The key element underpinning all of these areas of advancement in both technology and business is infrastructure edge computing. It is one thing to demonstrate a use case in a laboratory environment where everything is a known variable; it is quite another to then operate a commercial service in the real world with all of the messy constraints that introduces, from cost to performance to timescales.

Edge computing is one of the most frequently mentioned emerging technologies, which many believe will make a significant impact on the landscapes of both technology and business during the decade of the 2020s. The concept seems simple: By moving compute resources as close as possible to their end users, theoretically the latency between a user and their application can be reduced, the cost of data transport can be minimised, and these two factors combined will make new use cases practical.

But what really is edge computing, beyond the hype, marketing material, and hyperbole that always accompany any major technological shift? With so many competing definitions of even the most basic elements of the technology, can we succinctly define concepts and terminology which allow us to have a consistent understanding of the challenges we are trying to solve together as an industry?

What are the key factors driving edge computing, and what must a solution provide in order to solve key technical and business challenges? How does edge computing really replace, compete with, or augment cloud computing? What is infrastructure edge computing, and does it stand alongside the traditional regional, national, and on-premises data centre, or does it seek to replace them entirely?

This book aims to answer all of these questions and provide the reader with a solid foundation of knowledge with which to understand how we got to this inflection point and how infrastructure edge computing is a vital component of the next-generation internet – an internet which enables suites of new key use cases that unlock untapped value globally across many different industries.

2

What Is Edge Computing?

2.1 Overview

Before delving into the details and technical underpinnings of infrastructure edge computing, it is necessary to understand some of the history, terminology, and key drivers behind its development, adoption, and usage. This chapter aims to detail some of these factors and provide the reader with a shared base of knowledge to build upon throughout the rest of this book, starting with terminology.

2.2 Defining the Terminology

One of the most challenging aspects of edge computing has been agreeing upon a set of terminology and using it consistently across the many industries to which edge computing is of interest. This is by no means a unique challenge when it comes to emerging technologies, but in the case of edge, it has contributed significantly to confusion between multiple groups and companies who have struggled to reconcile their individual definitions of edge computing so that ultimately a shared view of what the problem to be solved is, in addition to where it is and how to solve it, could emerge and be used.

Part of the challenge in defining edge computing is that by its very nature, the concept of an edge is contextual: An edge is at the boundary of something and often delineates the specific place where two things meet. These two things may be physical, as pieces of hardware; they may be logical, as pieces of software; or they may be more abstract, such as ownership, intent, or a business model.

Understanding Infrastructure Edge Computing: Concepts, Technologies and Considerations,
First Edition. Alex Marcham.

Another part of the challenge has been attempting to compress the many dimensions across which a group or company may be concerned with edge computing into a small number of terms which are general enough and yet able to convey a specific meaning. Although it is appealing to create terms which describe a complex and specific set of dimensions as they relate to edge computing, this is a challenging path to create terminology which is general enough to use outside of that same group because the more dimensions a term or phrase aims to address, the less approachable it becomes.

The key to any set of terminology is consistency, and the way to achieve that even in highly technical discussions is to limit the scope of the concepts which the terminology aims to define. Once the key parameters of the definition are established, a neutral set of terminology can be created which then serves as the basis for additional layers of complexity to be added, promoting adoption and usage.

The Open Glossary of Edge Computing [1], a project arising out of the initial State of the Edge report [2] and co-authored by the author of this book, established a neutral and limited dimension set of terminology for edge computing which has seen adoption across the industry and aims to simplify the discussions around edge computing by using the physical location of infrastructure and devices to delineate which type of edge computing each is able to perform by using the last mile network as the line between them to create a clear point of separation. Additional dimensions such as ownership, a specific business model, or any other concern can then be layered on top of this physical definition.

Along with the State of the Edge itself, the Open Glossary of Edge Computing has been adopted by the Linux Foundation's LF Edge [3] group as an official project and continues to contribute to a shared set of terminology for edge computing to help facilitate clear discussion and shared understanding.

2.3 Where Is the Edge?

As previously described, an edge is itself a contextual entity. By itself, an edge cannot exist; it is the creation of two things at the point at which they interact. This somewhat floaty definition is one part of what has made establishing a concise and clear definition of edge computing difficult, especially when combined with the many different factors and dimensions that edge computing will influence.

This book will focus on the accepted definition from the Open Glossary of Edge Computing which uses the physical and role-based separation provided by using the last mile network as a line of demarcation between the infrastructure edge and device edge to provide separation and clarity.

2.3.1 A Tale of Many Edges

Although there are many potential edges, for the purposes of this book and to the most general definition of edge computing, the edge that is of the greatest importance is the last mile network.

The last mile network is the clearest point of physical separation between end user devices and the data centre infrastructure which supports them. In this context, the last mile network refers to the transmission medium and communications equipment which connects a user device to the network of a network operator who is providing wide area network (WAN) or metropolitan area network (MAN) service to one or more user devices, whether large or small, fixed position or mobile.

Examples of last mile networks include cellular networks, where the transmission medium is radio spectrum and the communications equipment used includes radio transceiver equipment, towers, and antennas. Wired networks such as those using cable, fibre, or digital subscriber line (DSL) are also examples of last mile networks which use a copper or fibre-based transmission medium. The specific type of last mile network used is irrelevant here for the terminology of edge computing.

This definition cannot capture all of the potential nuance which may exist; for example, in the case of an on-premises data centre which is physically located on the device side of the last mile network, the owner of that data centre may regard it as infrastructure rather than as a device itself. However, a different definition and accompanying set of terminology offering equal clarity without introducing unnecessary dimensions into the equation has not been established within the industry, and so this book will continue to use the infrastructure edge and device edge, separated by a last mile network.

Fundamentally, if everything can be recast as an example of edge computing, then nothing is truly an example of edge computing. It is similar to referring to a horse and cart as a car because both of them consist of a place to sit, four wheels, and an entity that pulls the cart forward. This is important to note with both the infrastructure edge and the device edge. In the case of the former, an existing data centre which exists a significant distance away from its end users should not be referred to as an example of edge computing. If, however, that same data centre is located within an acceptable distance from its end users and it satisfies their needs, an argument can be made for it to be so.

Similarly, if a device edge entity, such as a smartphone which already had significant local compute capabilities is now referred to as an edge computing device yet does not participate in any device-to-device ad hoc resource allocation and utilisation, this is a somewhat disingenuous application of the term edge computing. However, where there was once a dumb device or no device at all which is now being augmented or replaced with some local compute, storage, and network resources, this can be reasonably argued to be an example of device edge computing, even if limited in capability.

Although "edge washing" of this type is not unique to edge computing as similar processes occur for most technological changes for a period of time, due to the difficulties previously mentioned in the industry arriving at a single set of terminology around edge computing, this can be challenging to identify. This identification challenge can be addressed by using the framework described in the next section.

2.3.2 Infrastructure Edge

The infrastructure edge refers to the collection of edge data centre infrastructure which is located on the infrastructure side of the last mile network. These facilities typically take the form of micro-modular data centres (MMDCs) which are deployed as close as possible to the last mile network and, therefore, as close as possible to the users of that network who are located on the device edge. Throughout this book, these MMDCs will typically be referred to as infrastructure edge data centres (IEDCs), whereas their larger cousins will be referred to as regional or national data centres (RNDCs).

The primary aim of edge computing is to extend compute resources to locations where they are as close as possible to their end users in order to provide enhanced performance and improvements in economics related to large-scale data transport. The success of cloud computing in reshaping how compute resources are organised, allocated, and consumed over the past decade has driven the use of infrastructure edge computing as the primary method to achieve this goal; the infrastructure edge is where data centre facilities are located which support this usage model, unlike at the device edge.

Although it is typically deployed in a small number of large data centres today, the cloud itself is not a physical place. It is a logical entity which is able to utilise compute, storage, and network resources that are distributed across a variety of locations as long as those locations are capable of supporting the type of elastic resource allocation as their hyperscale data centre counterparts. The limited scale of an MMDC compared to a traditional hyperscale facility, where the MMDC represents only a small fraction of the total capacity of that larger facility, can be offset by the deployment of several MMDC facilities across an area with the allocation of only a physically local subset of users to each facility (see Figure 2.1).

2.3.3 Device Edge

The device edge refers to the collection of devices which are located on the device side of the last mile network. Common examples of these entities include smartphones, tablets, home computers, and game consoles; it also includes autonomous

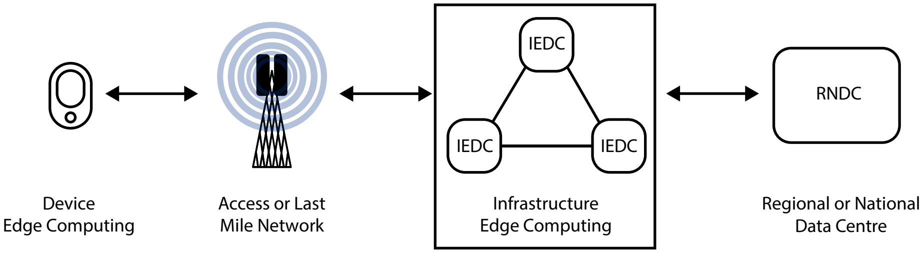

Figure 2.1 Infrastructure edge computing in context.

vehicles, industrial robotics systems, and devices that function as smart locks, water sensors, or connected thermostats or that can provide many other internet of things (IoT) functionalities. Whether or not a device is part of the device edge is not driven by the size, cost, or computational capabilities of that device but on which side of the last mile network that it operates. This functional division clarifies the basic architecture of an edge computing system and allows several more dimensions such as ownership, device capability, or other factors to be built on top.

These devices may communicate directly with the infrastructure edge using the last mile network or may use an intermediary device on the device edge such as a gateway to do so. An example of each type of device is a smartphone that has an integrated Long-Term Evolution (LTE) modem and so is able to communicate directly with the LTE last mile network itself, and a device which instead has only local range Wi-Fi network connectivity that is used to connect to a gateway which itself has last mile network access.

In comparison to infrastructure edge computing, many devices on the device edge are powered by batteries and subject to other power constraints due to their limited size or mobile nature. It would be possible to design cooperative processing scenarios using only device edge resources in which a device can utilise compute, storage, or network resources from neighbouring devices in an ad hoc fashion; however, for the vast majority of use cases and users, these approaches have proven to be unpopular at best with users not wishing to sacrifice their own limited battery power and processing resources to participate in such a scheme at a large scale outside of outliers such as Folding@home, a distributed computing project that is focused on using a network of mains powered computers, not mobile devices. Bearing this in mind, the need for access to dense compute resources in locations as close as possible to their users is provided to users at the device edge by the infrastructure edge (see Figure 2.2).

Although this book is primarily focused on infrastructure edge computing, topics related to device edge computing will be discussed as appropriate, especially as they relate to the interaction that exists between these two key halves of the edge computing ecosystem and their interoperation.

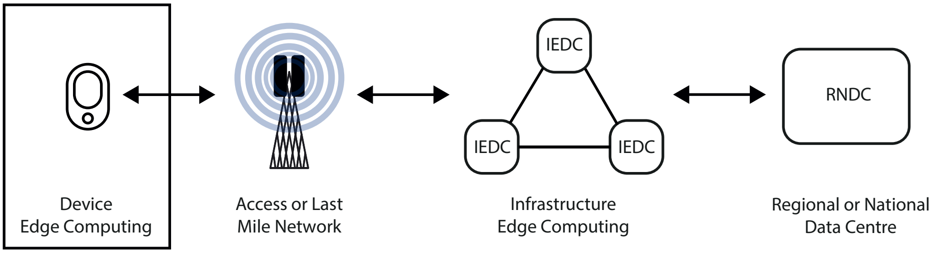

Figure 2.2 Device edge computing in context.

2.4 A Brief History

As with many technologies, upon close inspection, infrastructure edge computing represents an evolution more than the radical revolution that it may initially appear to be. This does not make it any less significant or impactful; it merely allows us to contextualise infrastructure edge computing within the broader trends which over time have driven much of the development of internet and data centre infrastructure since their inception. This progression lets us understand infrastructure edge computing not as the wild anomaly which it has been portrayed as in the past but as the clear progression of an ongoing theme in network design which has been present for decades and driven by the need to solve both key technical and business challenges using simple and proven principles.

2.4.1 Third Act of the Internet

One framework for understanding the technological progression which has brought us to the point of infrastructure edge computing is the three acts of the internet. This structure distils the evolution of the internet since its inception into three distinct phases, which culminate in the third act of the internet, a state which is driven by new use cases and enabled by infrastructure edge computing.

2.4.1.1 The First Act of the Internet

During the 1970s and 1980s, as the internet began to be available for academic and public use, the types of services it was able to support were basic compared to those which would emerge in the 1990s. Text-based applications such as bulletin board systems (BBS) and early examples of email represented some of the most complex use cases of the system. With no real-time element and a simple range of content, the level of centralisation was sufficient to support the small userbase.

It may seem obvious to us in hindsight that the internet would achieve the explosive growth that it has over its lifetime in terms of every possible characteristic from number of users to the volume of data that each individual user would transmit on