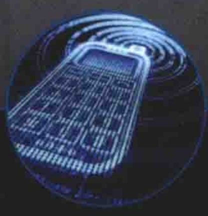
Wiley Series on Information and Communication Technology
Russell T. Hsing, Vincent K. N. Lau, and Mung Chiang, Series Editors

Fog for 5G and IoT



Edited by
Mung Chiang • Bharath Balasubramanian • Flavio Bonomi

WILEY

FOG FOR 5G AND IoT

Edited by

Mung Chiang
Arthur LeGrand Doty Professor of Electrical Engineering, Princeton University, Princeton, NJ, USA

Bharath Balasubramanian
Senior Inventive Scientist, ATT Labs Research, Bedminster, NJ, USA

Flavio Bonomi
Founder and CEO, Nebbiolo Technologies, Milpitas, CA, USA

WILEY

This edition first published 2017

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

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.

Library of Congress Cataloging-in-Publication Data

Names: Chiang, Mung, editor. | Balasubramanian, Bharath, editor. | Bonomi, Flavio, editor.
Title: Fog for 5G and IoT / edited by Mung Chiang, Bharath Balasubramanian, Flavio Bonomi.
Description: Hoboken, NJ, USA : John Wiley & Sons Inc., 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2016042091| ISBN 9781119187134 (cloth) | ISBN 9781119187172 (epub) | ISBN 9781119187158 (epdf)
Subjects: LCSH: Electronic data processing–Distributed processing. | Distributed shared memory. | Storage area networks (Computer networks) | Mobile computing. | Internet of things. | Cloud computing.
Classification: LCC QA76.9.D5 F636 2017 | DDC 004.67/82–dc23
LC record available at https://lccn.loc.gov/2016042091

Cover image: Cultura/Seb Oliver/Gettyimages
Cover design by Wiley

Set in 10/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

FOG FOR 5G AND IoT

WILEY SERIES ON INFORMATION AND COMMUNICATION TECHNOLOGY

Series Editors: T. Russell Hsing, Vincent K. N. Lau, and Mung Chiang

A complete list of the titles in this series appears at the end of this volume.

CONTRIBUTORS

MOSTAFA AMMAR, School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

HELDER ANTUNES, Corporate Strategic Innovations Group, Cisco Systems, Inc., San Jose, CA, USA

A. SALMAN AVESTIMEHR, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

BHARATH BALASUBRAMANIAN, ATT Labs Research, Bedminster, NJ, USA

SUMAN BANERJEE, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

FLAVIO BONOMI, Nebbiolo Technologies, Inc., Milpitas, CA, USA

S.-H. GARY CHAN, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

XU CHEN, School of ECEE, Arizona State University, Tempe, AZ, USA

MUNG CHIANG, EDGE Labs; Department of Electrical Engineering, Princeton University, Princeton, NJ, USA

SANGTAE HA, Department of Computer Science, University of Colorado at Boulder, Boulder, CO, USA

KARIM HABAK, School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

ROBERT J. HALL, AT&T Labs Research, Bedminster, NJ, USA

KHALED A. HARRAS, Computer Science Department, School of Computer Science, Carnegie Mellon University, Doha, Qatar

CARLEE JOE-WONG, Electrical and Computer Engineering, Carnegie Mellon University, Silicon Valley, CA, USA

STEVEN Y. KO, University at Buffalo, The State University of New York, Buffalo, NY, USA

PENG LIU, Pennsylvania State University, State College, PA; Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

ZHENMING LIU, Department of Computer Science, College of William and Mary, Williamsburg, VA, USA

ZHI LIU, Global Information and Telecommunication Institute, Waseda University, Tokyo, Japan

SATYAJAYANT MISRA, Department of Computer Science, New Mexico State University, Las Cruces, NM, USA

ANDREAS F. MOLISCH, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

ASHISH PATRO, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

STEFAN POLEDNA, TTTech Computertechnik AG, Wien, Austria

CONG SHI, School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA; Square, Inc., San Francisco, CA, USA

WILFRIED STEINER, TTTech Computertechnik AG, Wien, Austria

DEEPAK S. TURAGA, IBM T. J. Watson Research Center, Yorktown, New York, NY, USA

MIHAELA VAN DER SCHAAR, Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA, USA

DALE WILLIS, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

FELIX MING FAI WONG, Yelp Inc., San Francisco, CA, USA

ELLEN W. ZEGURA, School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

BO ZHANG, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

JUNSHAN ZHANG, School of ECEE, Arizona State University, Tempe, AZ, USA

TAO ZHANG, Corporate Strategic Innovation Group, Cisco Systems, Inc., San Jose, CA, USA

YI ZHENG, Corporate Strategic Innovation Group, Cisco Systems, Inc., San Jose, CA, USA

RAYMOND ZHENG, Corporate Strategic Innovation Group, Cisco Systems, Inc., San Jose, CA, USA

CONTENTS

Introduction

BHARATH BALASUBRAMANIAN,[1] MUNG CHIANG,[2] and FLAVIO BONOMI[3]

[1] *ATT Labs Research, Bedminster, NJ, USA*
[2] *EDGE Labs, Princeton University, Princeton, NJ, USA*
[3] *Nebbiolo Technologies, Inc., Milpitas, CA, USA*

The past 15 years have seen the rise of the cloud, along with rapid increase in Internet backbone traffic and more sophisticated cellular core networks. There are three different types of clouds: (i) data centers, (ii) backbone IP networks, and (iii) cellular core networks, responsible for computation, storage, communication, and network management. Now the functions of these three types of clouds are descending to be among or near the end users, as the "fog." Empowered by the latest chips, radios, and sensors, the edge devices today are capable of performing complex functions including computation, storage, sensing, and network management. In this book, we explore the evolving notion of the *fog architecture* that incorporates networking, computing, and storage.

Architecture is about the division of labor in modularization: who does what, at what timescale, and how to glue them back together. The division of labor between layers, between control plane and data plane, and between cloud and fog [1] in turn supports various application domains. We take the following as a working definition of the fog architecture: it is an architecture for the cloud-to-things (C2T) continuum that uses one or a collaborative multitude of end-user clients or near-user edge devices to carry out a substantial amount of storage, communication, and control, configuration, measurement, and management. Engineering artifacts that may use the fog architecture include 5G, home/personal networking, embedded AI, and the Internet of things (IoT) [2].

In Figure I.1, we highlight that fog can refer to an architecture for computing, storage, control, or communication network, and that as a network architecture it may support a variety of applications. We contrast between the fog architecture and the current practice of the cloud along the following three dimensions:

1. Carry out a substantial amount of storage at or near the end user (rather than stored primarily in large-scale data centers).
2. Carry out a substantial amount of communication at or near the end user (rather than all routed through the backbone network).

Fog for 5G and IoT, First Edition. Edited by Mung Chiang, Bharath Balasubramanian, and Flavio Bonomi.

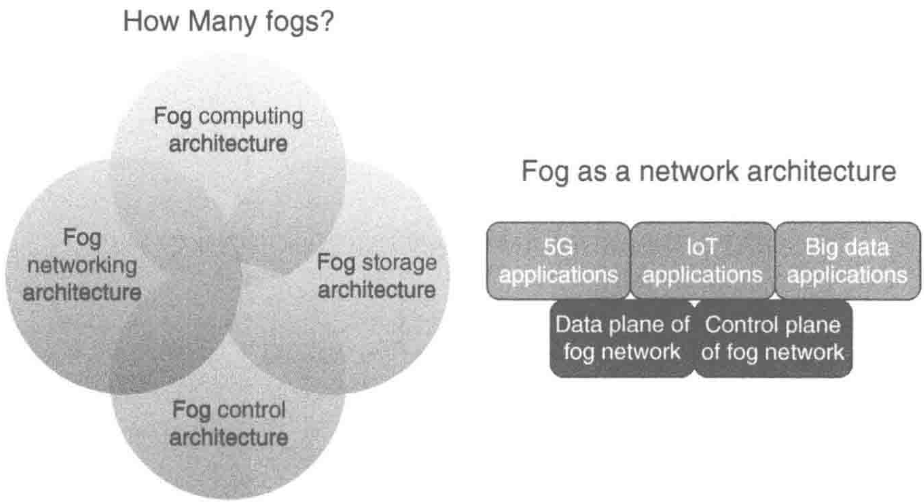


Figure I.1 Fog architectures and applications. Supported by such architectures.

3. Carry out a substantial amount of computing and management, including network measurement, control, and configuration, at or near the end user (rather than controlled primarily by gateways such as those in the LTE core).

Why would we be interested in the fog view now? There are four main reasons summarized as CEAL. Many examples in recent publications, across mobile and landline, and from physical layer beamforming to application layer edge analytics have started leveraging these advantages [3–8]:

1. *Cognition: Awareness of Client-Centric Objectives.* Following the end-to-end principle, some of the applications can be best enabled by knowing the requirements on the clients. This is especially true when privacy and reliability cannot be trusted in the cloud or when security is enhanced by shortening the extent over which communication is carried out.
2. *Efficiency: Pooling of Local Resources.* There are typically hundreds of gigabytes sitting idle on tablets, laptops, and set-top boxes in a household every evening, across a table in a conference room, or among the passengers of a public transit system. Similarly, idle processing power, sensing ability, and wireless connectivity on the edge may be pooled within a fog network.
3. *Agility: Rapid Innovation and Affordable Scaling.* It is usually much faster and cheaper to experiment with client and edge devices. Rather than waiting for vendors of large boxes inside the network to adopt an innovation, in the fog world a small team may take advantages of smartphone API and SDK, the proliferation of mobile apps, and offer a networking service through its own API.

4. *Latency: Real-Time Processing and Cyber–Physical System Control.* Edge data analytics, as well as the actions it enables through control loops, often have stringent time requirement and can only be carried out on the edge or the "things", here and now. This is particularly essential for Tactile Internet: the vision of millisecond reaction time on networks that enable virtual–reality-type interfaces between humans and devices.

We further elaborate on the previous potential advantages of fog. Client and edge devices have increasing strength and capabilities. For instance, the original iPhone had a single core 412 MHz ARM processor with 128MB RAM and 8GB storage space. The iPhone 5S on the other hand carries a dual-core 1.3 GHz Apple A7 processor with 1GB RAM, 64 GB storage space, and enhanced GPU capabilities. Intel's mobile chip Atom and Nvidia's Tegra too promise near similar specifications. The increase in strength and capabilities implies complex functionality such as CPU/GPU intensive gaming, powerful location/context tracking sensors, and enhanced storage. Further, as suggested in [9], these interconnected edge devices will play a crucial role in orchestrating the IoT. Edge devices including mobile phones and wearable devices use a rich variety of sensors including gyroscopes, accelerometers, and odometers to monitor the environment around them. This enables the crucial notion of exploiting context both personal in terms of location and physical/psychological characteristics and context in the communal sense of how devices are interacting with other devices around them.

As the need for cloud-based services increases, the amount of data traffic generated in the core networks is increasing at an alarming rate. Cisco predicts that cloud traffic will increase almost four to five times over the next 5 years [10]. Further, they predict that cloud IP traffic will account for nearly two-thirds of all data center traffic by 2017. Can the fog alleviate some of this by satisfying application needs locally? For example, can part of cloud storage be moved closer to the user with edge/client devices acting as micro-data centers? Can videos be cached efficiently at the edge devices to reduce accesses to the cloud? Or more broadly, can edge devices perform an active role in orchestrating both data plane-based cloud services and control plane-based core network services?

Accesses to the cloud often span geographically distant entities with round-trip times of nearly 150–200 ms. Access latency is a crucial factor in the end-user experience with studies showing that a 20% decrease in RTTs results in a 15% decrease in page load time [11]. A significant way to decrease the RTT for content access is to place as much of the content physically close to the end user as possible. While decreasing latency is beneficial to all services, it may be a necessity for many services in the future. For example, services involving augmented reality applications may not tolerate latencies of more than 10–20 ms [12]. Hence, any computation/processing for these kind of services need to be performed locally. Fog services may play a significant part in addressing this challenge.

The fog R&D will leverage past experience in sensor networks, peer-to-peer systems, and mobile ad hoc networks while incorporating the latest advances in devices, systems, and data science to reshape the "balance of power" in the ecosystem between powerful data centers and the edge devices. Toward that end, this book serves as the first introduction to the evolving fog architecture, compiling work traversing many different areas that fit into this paradigm.

In this book, we will encounter many use cases and applications that in many ways are not necessarily new and revolutionary and have been conceived in the context of distributed computing, networking, and storage systems. Computing resources have been always distributed in homes, in factories, along roads and highways, in cities, and in their shopping centers. The field of pervasive or ubiquitous computing has been active for a long time. Networking has always deployed switches, routers, and middleboxes at the edge. Caching media and data at the edge has been fundamental to the evolution of Web services and video delivery.

As is typical of any emergent area of R & D, many of the themes in the fog architecture are not completely new and instead are evolved versions of accumulated transformations in the past decade or two:

- Compared with peer-to-peer (P2P) networks in the mid-2000s, fog is not just about content sharing (or data plane as a whole) but also network measurement, control and configuration, and service definition.
- Compared with mobile ad hoc network (MANET) research a decade ago, we have much more powerful and diverse off-the-shelf edge devices and applications now, together with the structure/hierarchy that comes with cellular/broadband networks.
- Compared with generic edge networking in the past, fog networking provides a new layer of meaning to the end-to-end principle: not only do edge devices optimize among themselves, but also they collectively measure and control the rest of the network.

Along with two other network architecture themes, ICN and SDN, each with a longer history, the fog is revisiting the foundation of how to think about and engineer networks, that is, how to optimize network functions: who does what and how to glue them back together:

- *Information-Centric Networks.* Redefine functions (to operate on digital objects rather than just bytes)
- *Software-Defined Networks.* Virtualize functions (through a centralized control plane)
- *Fog Networks.* Relocate functions (closer to the end users along the C2T continuum)

While fog networks do not have to have any virtualization or to be information centric, one could also imagine an information-centric, software-defined fog network (since these three branches are not orthogonal).