Elastic service scaling optimization in cloud-based communication systems

Tang, Jianhua

2015

Tang, J. (2015). Elastic service scaling optimization in cloud-based communication systems. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/65478>

<https://doi.org/10.32657/10356/65478>

ELASTIC SERVICE SCALING OPTIMIZATION IN CLOUD-BASED COMMUNICATION SYSTEMS

TANG JIANHUA

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2015

Elastic Service Scaling Optimization in Cloud-based Communication Systems

by

TANG Jianhua

submitted to the Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
January, 2015

Abstract

Cloud computing has emerged as a widely adopted computing paradigm over the past several years. Due to its ability to scale service capabilities, enhanced hardware utilization and reduced capital and operation expenditure can be achieved. Therefore, many conventional communication systems (CCS) are migrating from hardware-defined infrastructures to software-defined cloud environment. In this dissertation, we study two cloud-based communication systems (CBCS): cloud-centric media network (CCMN) and cloud radio access network (C-RAN). A CCMN is a cloud-based platform for content delivery, which is evolved from content delivery network (CDN); A C-RAN is an evolution of cellular communication networks, which decouples the baseband processing functionalities from the cellular base stations (BSs) and migrates those baseband processing tasks to a cloud baseband unit (BBU) pool. With the ability to elastically scale service capacities in the cloud-based system component, many problems well-studied in the CCS have to be re-looked in the CBCS. For example, resource allocation schemes for CCS are typically oblivious to computation costs since these are fixed. In CBCS, however, the computation costs at the cloud computation resource pool can be dynamically scaled according to users' demands.

In this dissertation, we show how to approximately optimize the elastic service scaling in the cloud-based component of the CCMN and C-RAN, in tandem with other network parameters like dynamic traffic arrival rates and cross-layer quality-of-service (QoS) guarantees, respectively. The main contributions of this dissertation are as follows:

- We consider the problem of optimally redirecting user requests in a CCMN to multiple destination virtual machines (VMs), which elastically scale their service capacities in order to minimize a cost function that includes service response times, computing costs, and routing costs. We also allow the request arrival process to switch between normal and flash crowd modes to model user requests to a CCMN. We quantify the trade-offs in flash crowd detection delay and false alarm frequency, request allocation rates and service capacities at the

VMs.

- We investigate a cross-layer resource allocation model for C-RAN to minimize the overall system power consumption in the BBU pool, fiber links and the remote radio heads (RRHs). We characterize the cross-layer resource allocation problem as a mixed-integer nonlinear programming (MINLP), which jointly considers elastic service scaling, RRH selection, and joint beamforming. The MINLP is however a combinatorial optimization problem and NP-hard. We relax the original MINLP problem into an extended sum-utility maximization (ESUM) problem, and we propose two approaches to solve the ESUM problem. In addition, we also propose a low-complexity Shaping-and-Pruning (SP) algorithm to obtain a sparse solution for the active RRH set.
- We consider the problem of system cost minimization in C-RAN by allowing each user equipment to associate with multiple VMs in the BBU pool. Furthermore, each RRH can serve only a limited number of UEs. Under this model, we study the system cost minimization problem. We jointly consider the VM activation in the BBU pool and sparse beamforming in the coordinated RRH cluster, which has limited fronthaul capacity constraint, to minimize the system cost of C-RAN. We formulate this problem as a MINLP, and then propose two different methods to obtain the optimal number of active VMs, as well as the sparse beamforming vectors.

The algorithms we proposed in this dissertation have relatively lower complexities than most of the existing algorithms in the literature. Furthermore, extensive simulation studies demonstrate that our proposed algorithms are more cost-efficient than other algorithms.

Supervisor: Tay Wee Peng
Title: Assistant Professor

Acknowledgments

It has been a long, tough, memorable and stunning journey for the past 4 years' PhD study. I have been extremely fortunate to meet and interact with a lot of outstanding individuals, without whom, the journey would be much more arduous. I would like to express my sincere appreciations to them.

To my supervisor Prof. Wee Peng Tay: thanks a lot for giving me the freedom to explore my research interests. It has been a tremendous wealth to be under your tutelage. The guidance and support you provided throughout the way are absolutely priceless and heartfelt. Your knowledge is boundless, and your perspective on research problem is deep and unique. I learned and acquired plenty of significant characters and skills from you, including humble behaviour, scientific writing, technical presentation, and rigorous thinking.

To Prof. Yonggang Wen in the School of Computer Engineering, NTU: thank you for showing me the way to cloud computing. It has been a irreplaceable experience to have so many technical discussions during the meetings for my first research topic in NTU.

To Prof. Tony Q. S. Quek in Singapore University of Technology and Design (SUTD): thanks for cultivating my interests in wireless communication, especially in cloud radio access network. You always suggest and send me latest papers which are exactly in line with my expectation. I'm looking forward to collaborating with you in the near future.

To Prof. Ben Liang in University of Toronto: I really appreciate your hosting and help during my visiting to University of Toronto. It was an awesome stay in your research group. I really enjoyed the discussion with you and your group members, and obtained lots of insights from your weekly group meetings.

To Prof. Laiquan Han in Northeastern University at Qinhuangdao: thanks for enlightening me the way for research when I was an undergraduate. I would not be able to come up with the idea for the PhD programme without the positive effects from you.

Other than the research experience, friendship is another crucial components of my PhD study. Thanks for all the lovely friends I met in Singapore, who could detach my mind from the complicated research works, and make my life so colorful, including Guo Shuai, Zhou Tianyi, Yuan Jifeng, Song Jie, Candy, Tan Zijian, Derek, Amelia, Zhou Yifan, Wu Chao, Kuang Tingting, Vivi, Fang Shu, Mu Hao, Cao Bin, Wang Zuyuan, Gong Ling, Wang Lixiang, Yuan Yang, Huang Mingyue, Wen Xin, Yu Le, Yan Xihong, Ou Jiajue, Liu Hui, Yao Jun, Yu Ya, Qin Tian, Wang Chuncheng, Zhang Maojin, Zhang Jiye, Ding Yi, Zhang Yuchao and Zheng Huanhuan.

The time spent in INFINITUS has been a most unforgettable one, which I will always pick up fond memories of. Thanks for all the labmates and also close friends, including Luo Wuqiong, Leng Mei, Che Yueling, Cai Shengming, Sheng Yizhu, Li Xiang, Liu Zilong, Cheng Chi, Wen Kai, Cai Ruojun, Francois, Gabriel, Zhang Yi, Hu Wuhua, Liu Yunxiang, Wang Yuan, Luo Sheng, Zhao Zhi, Xiong Fei, Ooy Mei and Thida.

Thanks for the alumni, who I met (again) in Singapore, from Northeastern University. I am especially grateful to Feng Guangyin, Li Lu, He Peijun, Ding Lin, Li Yuting, Hou Lingyu, Fu Jingyuan, Zhou Jingjing, Zhu Yirong, Lin Yimin and Sam Zhang. I have been so proud of having some of you guys who could always support the activities I organised.

I would like to thank my friends who have made my life in Toronto so memorable: Wang Yong, Christy and Xu Yujie.

Last, but most importantly, I would like to express my heartfelt appreciations to my parents, my sister, and my entire family. I am extremely grateful to my beloved parents for their continuous support and standing by me throughout my endeavors. The freedom they gave me to pursue my dreams is highly critical in my life. To my aunties and uncles: thanks for taking care of me and provide me adorable support in lots of my key life stages.

I dedicate this dissertation to the memory of my dearest grandparents, who selflessly and tirelessly gave me the unparalleled love. I learned many essential life principles from them when I was still a child, such as kindness and optimism. I will

always remember them fondly in my heart.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

1	Introduction	17
1.1	Drawbacks of existing non-cloud based networks	18
1.1.1	Drawbacks of traditional core network	18
1.1.2	Drawbacks of traditional RAN	20
1.2	Cloud based communication systems	20
1.2.1	CCMN	22
1.2.2	C-RAN	23
1.3	Motivations and contributions	25
1.4	Notations	27
1.5	Thesis Outline	28
2	Background and Related Work	29
2.1	User request redirection	29
2.1.1	Cost-blind redirection with fixed service provisioning	30
2.1.2	Cost-aware redirection with fixed service provisioning	31
2.1.3	Cost-aware redirection with elastic service scaling	32
2.2	Resource allocation problems in C-RAN	34
2.2.1	RRH selection and user association	35
2.2.2	Weighted sum-rate maximization	36
3	Dynamic Request Redirection and Elastic Service Scaling in CCMN	39
3.1	Problem formulation	40
3.1.1	Approximation for detection delay and false alarm frequency	48

3.2	Alternating optimization	50
3.3	Optimal request allocation and service capacity	51
3.3.1	VM prices	53
3.3.2	Bounded service capacity	55
3.4	Simulation results	56
3.4.1	YouTube trace data	58
3.4.2	Simulated arrival requests	61
3.5	Summary	63
3.6	Proofs	64
3.6.1	Proof of Proposition 3.1	64
3.6.2	Proof of Lemma 3.1	65
3.6.3	Proof of Theorem 3.1	66
4	Cross-Layer Resource Allocation with Elastic Service Scaling in C-RAN	69
4.1	System model	71
4.1.1	Problem formulation	74
4.2	The QWSRM problem	75
4.3	Cross-layer power consumption minimization	80
4.3.1	Relaxing the ESUM problem to a QWSRM problem	84
4.3.2	A WMMSE approach to solve the ESUM problem	85
4.3.3	The RRH selection problem	88
4.4	Simulation results	91
4.4.1	Simulation setup	91
4.4.2	The effect of shaping	93
4.4.3	The importance of cross-layer design	95
4.5	Summary	97
4.6	Proofs	98
4.6.1	Proof of Theorem 4.1	98
4.6.2	Proof of Proposition 4.1	98

4.6.3	Proof of Proposition 4.2	99
5	Towards System Cost Minimization in C-RAN with Limited Fron-	
	thaul Capacity	101
5.1	System model	102
5.1.1	System description	102
5.1.2	Queueing system model	103
5.2	Problem formulation	106
5.2.1	An equivalent formulation for problem (OP0)	107
5.3	Approaches to solve problem (OP1)	109
5.3.1	Bisection search for price vector Γ	109
5.3.2	Reweighted l_1 -norm relaxation	110
5.3.3	Two approaches for optimizing the variable m	111
5.4	Numerical results	116
5.4.1	Simulation setup	116
5.4.2	The benchmarks	118
5.4.3	Performance	119
5.5	Summary	121
5.6	Proofs	122
5.6.1	Proof of Theorem 5.1	122
5.6.2	Proof of Proposition 5.1	123
6	Conclusion and Future Work	125
6.1	Conclusion	125
6.2	Future work	126
6.2.1	Embedding mobile cloud computing in C-RAN	126
6.2.2	Caching as a Service (CaaS) in C-RAN	127
6.2.3	Statistical QoS provisioning	128
	Bibliography	131
A	List of Acronyms	145

List of Figures

1-1	Two different file distribution schemes.	18
1-2	Cloud-centric media network (CCMN).	23
1-3	Cloud radio access network (C-RAN).	24
1-4	The overall network structure of cloud-based communication system.	25
2-1	The request flows to data centers.	30
3-1	Request redirection model.	42
3-2	The logic flow chart for Chapter 3.	52
3-3	VM prices sorted in non-decreasing order.	55
3-4	Histogram of inter-arrival times under different traffic modes. The red curve shows a fitted exponential distribution.	59
3-5	Rate of YouTube requests over a typical one day period, starting at 3:40am and ending at 3:40am next day. The dotted lines indicate two periods in which the average traffic arrival rate is significantly higher than other times.	60
3-6	Convergence of the alternating optimization procedure for computing the DRES strategy.	60
3-7	Performance simulated with YouTube data.	61
3-8	Performance comparison under different arrival rate ratios.	62
3-9	Performance comparison when π changes.	62
3-10	Allocation proportions for different strategies in the flash crowd mode.	63

4-1	Queueing model representation of a C-RAN processing and transmission path for a UE i	72
4-2	An intuitive explanation for the BnB algorithm.	80
4-3	The two-step approach to solve problem (P1).	82
4-4	Simulation setup with 7 RRHs.	92
4-5	Number of active RRHs using different RRH selection algorithms. . .	94
4-6	System power consumption ratio under different algorithms.	96
5-1	A general model of C-RAN.	102
5-2	Queueing network model representation of a C-RAN cloud processing and wireless transmission.	104
5-3	An iterative approach to solve problem (OP0).	107
5-4	Simulation setup in a heterogeneous C-RAN.	118
5-5	System cost under different algorithms.	120
5-6	Different delay requirements.	121

List of Tables

1.1	Comparison between CDN and CCMN.	22
1.2	Comparison between traditional RAN and C-RAN.	26
3.1	Summary of common notations in Chapter 3.	41
3.2	Amazon EC2 instances and prices	43
3.3	Simulation parameters in Chapter 3	58
4.1	Summary of common notations in Chapter 4 and 5.	71
4.2	Simulation parameters in Chapter 4	92
5.1	Simulation parameters in Chapter 5	118

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

The evolution of the Internet has recently been driven by fast data proliferation. Cisco predicts that global IP traffic will grow with a compound annual growth rate of 21% in next few years, especially for the busy-hour Internet traffic, which will increase at a compound annual growth rate of 28% from 2013 to 2018 [1]. This growth is mostly driven by video content. By year 2018, video traffic, including TV, Internet, video on demand and peer-to-peer, will constitute approximately 80 % to 90% of global consumer traffic. In terms of the access modes, mobile data traffic will increase 11-fold from 2013 to 2018, and traffic from wired devices will be exceeded by those from mobile and wireless devices by 2018. This is partially driven by the increasing popularity of smart hand-held devices, such as smartphones and tablets.

To maintain a high level of quality-of-service (QoS) for emerging video content and mobile data, on the supply side, the growth of the Internet infrastructures, including bandwidth, storage and server farms, have to meet the pace of growing traffic in the Internet. Based on the three-layer hierarchical internetworking architecture model, i.e., core layer, aggregation layer and access layer, in next section, we discuss the main drawbacks of existing non-cloud based technologies used in core network and radio access network (RAN) respectively. Note that the discussion for aggregation network is out of scope for this dissertation.

Figure 1-1: Two different file distribution schemes.

1.1 Drawbacks of existing non-cloud based networks

We introduce existing solutions for non-cloud based networks, including core network and RAN, and analyze their drawbacks respectively.

1.1.1 Drawbacks of traditional core network

In core network, there is an urgent need to design networks and protocols that specifically address many technical challenges introduced by the upsurge in Internet multimedia traffic [2, 3]. One potential solution is the deployment of content delivery networks (CDNs). In a CDN, contents are pushed towards the edge servers [4], which are closer to the end consumers. Specifically, the CDN is an overlay network, which caches contents at the network edge and pushes them to the consumers from an optimal location [5]. We show the typical structures of single server distribution and CDN in Figure 1-1.

There are many advantages by utilizing CDN, such as decreasing server load, reducing content retrieval latency, improving content availability and increasing the number of concurrent users. However, there are some inherent limitations of utilizing

CDN [6]:

- Due to the high upfront and operation costs, it is impractical for most organizations or content service providers to build up and maintain their own CDNs around the world to process the global requests.
- Although there are some third-party CDN vendors, pioneered by Akamai [7], the high setup fees and other hidden fees still keep small clients away. And the support availability is another issue once third-party vendors are utilized to maintain the CDN.

In the meantime, the traffic to CDN has two important features. One important feature in CDN traffic is the higher frequency of flash crowds compared to normal Internet traffic [8]. A flash crowd occurs when there is an unexpectedly high amount of user traffic during a short period of time [9]. For example, [3] records 2501 flash crowds appearing in the CoralCDN over a 4-year span, i.e., almost two flash crowds happen daily on average. Another feature of the traffic to CDN is the increasing amount of dynamic content (e.g. online shopping and social network traffic).

Therefore, even for CDN vendors themselves, they also face many problems, since existing CDN solutions are inadequate to deal with the exponential growth of video content traffic. In particular, first, with an increasing amount of dynamic and media contents in the CDN, web applications are becoming more intensive in computation capability. As a result, traditional server selection mechanisms (e.g. forwarding requests to the closest server) designed for static contents may no longer be optimal [10]. Second, static resource allocation mechanisms suffer from poor resource dimensioning as multimedia traffic is highly variable, as reflected by the high peak to valley ratio [11] of user traffic. A static allocation based on peak hour traffic may have utilization rates as low as 5% to 10% [10, 12], while an allocation based on average traffic may result in high latency during flash crowds.

1.1.2 Drawbacks of traditional RAN

The evolution of the RAN is lagging behind the speed of data proliferation since traditional RAN architecture has the following limitations:

- Conventional cellular communication system is a strict connection limited system, i.e., at any given time, each user only connects to one base station (BS) that gives it the best performance, and BSs work independently from each other.
- Furthermore, the conventional cellular communication system is an interference-limited system, so that the system capacity is hard to improve under conventional interference management methods.
- Moreover, BSs are built by service providers on their own proprietary platforms, and, hence, BSs processing power cannot be shared by other service providers [13].

To overcome the above limitations of traditional RAN, many techniques have been proposed to enhance the capacity and coverage of wireless or mobile communication, such as coordinated multipoint (CoMP) [14] and heterogeneous network (HetNet) [15]. However, to enhance capacity and coverage, the operator faces following challenges: Firstly, the explosive increase in network capacity demand (especially busy-hour demand) triggers an exponential increase in the number of BSs, which leads to a significantly higher power consumption. Secondly, costly capital and operating expenditure leads to falling average revenue per user. Moreover, with the dynamic nature of the mobile traffic, the utilization of the BSs is actually quite low after the busy-hour. Last but not least, traditional RAN structure lacks the ability and efficiency to deploy the sophisticated centralized processing functionalities, e.g., interference management, in both CoMP and HetNet.

1.2 Cloud based communication systems

Cloud computing has emerged as a widely adopted computing paradigm over the past several years. The increasing popularity of cloud computing is due to its attractive

properties as listed below [16]:

- Five essential characteristics, i.e., on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. For example, rapid elasticity is defined as “capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time” in [16].
- Three service models, i.e., Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). For instance, PaaS means that the consumer can rent fundamental computing resources, including processing, storage, networks and other resources, from cloud service providers to deploy and run arbitrary software.
- Four deployment models, i.e., private cloud, community cloud, public cloud and hybrid cloud. For example, Amazon EC2 and Windows Azure are famous public cloud platforms.

As a result, many conventional communication systems (CCS) are trying to migrate from hardware defined infrastructures to software defined cloud environment, since cloud computing offers following benefits [17]:

- Reducing capital expense. There’s no need for individuals or organizations to spent big money to build up their own data centers, including hardware, software or licensing fees. They can only pay when they consume computing resources, and only pay for how much they consume.
- Easy to maintain. Different from owning a data center, cloud service customers do not need to worry about maintaining the infrastructure. Or, they can conduct maintaince work with just a few clicks.
- Elasticity. Individuals or organizations can access as much or as little as they need, and scale up and down as required with only a few minutes’ notice.

Table 1.1: Comparison between CDN and CCMN.

	CDN	CCMN
Resource Provisioning	Physical machine static installation	Virtual machine dynamic assignment
Infrastructure Ownership	Owned by content service provider	Leased from cloud service provider
Costs	High upfront and operating costs	Low/No upfront cost, low operating cost

- Improving accessibility and agility. Cloud data centers are deployed worldwide, people can enjoy a lower latency and better experience.

In this dissertation, we study two cloud-based communication systems (CBCS) for core network and RAN respectively: cloud-centric media network (CCMN) and cloud radio access network (C-RAN).

1.2.1 CCMN

As elaborated in Section 1.1.1, although using CDN in the core network can reduce server load, decrease system latency and achieve distributed network traffic, the capabilities of CDNs to handle dynamic loads and flash crowd traffic are still lacking.

The emergence of cloud computing offers a natural way to extend the capabilities of CDNs to support the growth of content with dynamic loads. In cloud computing, an organization or individual rents remote server resources dynamically, and users can add or remove server capacity at any time to meet their requirements [18]. To combine the merits of CDN and cloud computing, CCMN was proposed [19, 20] as an architecture that migrates CDN into the cloud (see Figure 1-2). In a CCMN, virtual machines are carved out of an underlying hybrid cloud, forming a content distribution overlay. Thus, the content consumers can obtain contents from the cloud service provider, instead of the content service providers only. The amount of system resources in a CCMN can be dynamically scaled up and down, matching real-time application demands. As a result, it offers the feasibility of reducing the cost of content delivery. To be more specific, we compare CDN and CCMN in Table 1.1 [21].

Figure 1-2: Cloud-centric media network (CCMN).

1.2.2 C-RAN

As described in Section 1.1.2, remedying the existing RAN is just “curing the symptoms, not the disease”. Therefore, the RAN service providers need a “white paper design” to pull themselves out of the aforementioned dilemma.

It is believed by the service providers that Centralized base-band pool processing, Cooperative radio with distributed antenna equipped by remote radio head (RRH) and real-time Cloud infrastructures RAN (C-RAN) is a promising solution to the challenges that the service providers are facing [13]. Utilizing centralized signal processing in the cloud baseband unit (BBU) pool instead of the distributed BSs in the conventional cellular networks can significantly save the capital and operating expenditure. Moreover, cooperative radio techniques using distributed antenna equipped by a RRH improves the spectrum efficiency and the communication quality of the cell edge users. Furthermore, applying cloud computing as the infrastructure of the centralized processing pool can greatly reduce the power consumption and improve hardware utilization, by the nature virtues of cloud computing. Therefore, the combination of these technologies leads to a reduced cost, lower energy consumption, higher spectral efficiency, and easier to manage/upgrade system, the C-RAN.

A C-RAN system consists of three main parts (cf. Figure 1-3): firstly, in the

Figure 1-3: Cloud radio access network (C-RAN).

remote site, there are RRHs plus antennas to receive and transmit signals. Secondly, RRHs are connected with the BBU pool by the high bandwidth low-latency optical transport network. The third part is the cloud computing based BBU pool equipped with programmable general purpose processors and real-time virtualization technology, which can greatly improve the hardware utilization. To be more specific, we list the properties and power consumptions of the three main components as below:

- **BBU pool.** Having a cloud data center as the BBU pool not only allows centralized signal processing, but also facilitate elastic service scaling of the resources of the BBU pool. Specifically, the BBU pool can dynamically adjust each virtual machine (VM) computation capacity and the number of active VMs to optimize power consumption for the changing traffic and channel states.
- **Fiber links.** Each RRH is connected to the BBU pool via a high-bandwidth, low-latency fiber link. The power consumption in the fronthaul links has traditionally been ignored in the conventional cellular networks (CCNs) literature since it is relatively much lower than the power consumption in the BSs of the CCNs. However, in C-RAN, the power consumption in fronthaul links is comparable to the power consumption at the RRHs since the RRHs are architecturally much simpler compared to conventional BSs.
- **RRHs.** In C-RAN, the functionality of RRHs can be as simple as just a sig-

Figure 1-4: The overall network structure of cloud-based communication system.

nal transmission and reception point. RRHs can cooperate with each other to perform centralized joint beamforming to mitigate interference. Thus, the throughput of the wireless channels to the users can be significantly enhanced.

Decoupling the baseband signal processing from the RRHs is the most attractive feature of C-RAN, which means that RRHs only need to keep the basic transmission and reception functionalities, while computationally intensive tasks can be migrated to the BBU pool in a cloud data center. This centralized signal processing and scheduling feature in the BBU pool further makes a variety of prospective technologies feasible, including centralized encoding and decoding, centralized compression and decompression, and joint beamforming.

Table 1.2 compares traditional RAN and C-RAN [13, 22].

1.3 Motivations and contributions

With the assistance of cloud computing in both core network and radio access network, we depict an overall network structure of cloud-based communication system in Figure 1-4.

With the ability to elastically scale service capacities in the cloud-based system component, many problems well studied in the CCS have to be re-looked at in the

Table 1.2: Comparison between traditional RAN and C-RAN.

	Traditional RAN	C-RAN
Radio and Baseband Functionalities	Co-located in each BS site.	Split into RRHs and BBU pool. Each RRH keeps basic radio functionalities in the remote site, while baseband processing functionalities from many sites are gathered in a cloud BBU pool, which is 20-40 km away from RRHs.
Computation Resource Provisioning	Fixed once the BS site is installed. Resources are normally underutilized.	Elastic, can be dynamically adjusted.
Power Consumption	Each BS use high power consumption to support base station equipments and air conditions.	Each RRH just need very low power consumption to support basic signal transmission and reception.
Independence among BSs	Signal processing in a distributed manner, i.e., each BS works independently.	Signal processing in a centralized manner in BBU pool. Hence, it is easy to implement coordinated and joint processing, e.g., centralized encoding and decoding.
Costs	High upfront and operating costs, long construction cycle.	Low upfront and operating cost, short construction cycle.

CBCS. For example, resource allocation schemes for CCS are typically oblivious to computation costs since these are fixed. In CBCS, however, the computation costs at the cloud computation resource pool can be dynamically scaled according to users' demands.

In this dissertation, we show how to approximately optimize the elastic service scaling in the cloud-based component of the CCMN and C-RAN, in tandem with other network parameters like dynamic traffic arrival rates and cross-layer QoS guarantees, etc.

Our main contributions are as follows:

- Introducing cloud computing in both core network and radio access network.
- Investigating elastic service scaling feature in cloud-based communication systems and studying three elastic service scaling models:
 1. Both the number of VMs and each VM's computation capacity can be dynamically adjusted. This model is studied in Chapter 3.
 2. The number of VMs in the system is fixed, while each VM's computation capacity can be dynamically scaled. This model is used in Chapter 4.
 3. Each VM's computation capacity is fixed, while the number of active VMs in the system can be dynamically adjusted. This model is investigated in Chapter 5.
- In addition to studying the elastic service scaling feature, we also jointly consider some practical issues in cloud-based communication systems, such as flash crowd traffic detection in CCMN in Chapter 3, RRH selection in C-RAN in Chapter 4 and limited fronthaul capacity in heterogeneous C-RAN in Chapter 5.

1.4 Notations

Throughout this dissertation, $\mathbb{N}$, $\mathbb{C}$, $\mathbb{R}$ and $\mathbb{R}^+$ stand for the natural numbers, complex numbers, real numbers and non-negative real numbers respectively. For each function $f : \mathbb{R} \mapsto \mathbb{R}$, we use $\partial^- f(x)$ and $\partial^+ f(x)$ to denote the left and right derivatives of $f(x)$ with respect to (w.r.t.) x , respectively. For a convex function f , the subdifferential at x is the set $[\partial^- f(x), \partial^+ f(x)]$. We use the notation $x^+ = \max(0, x)$. We also use calligraphy letters to represent sets, and boldface lower case letters to denote vectors. The notation $\|\mathbf{x}\|_2$ denotes the Euclidean norm of $\mathbf{x}$, while $(\cdot)^T$ and $(\cdot)^H$ represent transpose and conjugate transpose, respectively. $\mathcal{CN}(0, \sigma^2)$ is the distribution of a circularly symmetric complex normal zero mean random variable with variance σ^2 . $\Re(\cdot)$ stands for the real part of a complex number.

The other notations are formally defined in the sequel where they first appear in each chapter. We summarize most commonly used abbreviations in Appendix A.

1.5 Thesis Outline

In Chapter 2, we summarize related works for this dissertation and include state-of-the-art solutions for some related problems. We present request redirection and elastic service scaling problem in CCMN and propose an efficient solution in Chapter 3. In Chapter 4, we study system power minimization problem with cross-layer QoS guarantee under an idealized C-RAN model. Then, the idealize model is generalized in Chapter 4, in which we consider the joint VM activation and sparse beamforming problem to minimize system cost. We conclude and discuss some future research directions in Chapter 6.

Chapter 2

Background and Related Work

In this dissertation, we study the elastic service scaling problems in cloud-based communication systems, including CCMN and C-RAN. We summarize background and state-of-the-art solutions for the related problems in this chapter.

2.1 User request redirection

In this dissertation, we first develop cost-aware optimal user redirection and service rate scaling mechanisms in a CCMN under switching arrival traffic modes. In a user request redirection system (cf. Figure 2-1), a dispatcher aims to redirect incoming service requests to multiple servers, while each server receiving requests scales its service rate in order to jointly minimize a cost function.

The topic of user request redirection or load balancing has been well studied in the literature (see [23, 24] for surveys on this topic).

Generally, based on the network layers and principle, request redirection (or, request routing) can be classified in DNS request routing, transport-layer request routing, and application-layer request routing [25].

Depending on the models and algorithms in the literature, we classify the mechanisms into two categories in terms of different parameters and metrics involved: cost-blind and cost-aware.

Figure 2-1: The request flows to data centers.

2.1.1 Cost-blind redirection with fixed service provisioning

For the cost-blind mechanisms, people take the network traffic as the only parameter to minimize the CPU and communication overhead [26–28]. The most common algorithms are as follows [29]:

- (i) Random balancing mechanism (RAND). In such a policy, the incoming requests are distributed to the servers in the network with a uniform probability [30,31].
- (ii) Round Robin algorithm (RR). This algorithm selects a different server for each incoming request in a cyclic mode, ignoring the state of the network and the servers, and allocates the same amount of requests to each server [32,33].
- (iii) Weighted Round Robin (WRR). As an upgraded version of RR, in the case of heterogeneous servers in the network, the traffic will be allocated to the servers, based on the server capacities, with specified ratios [34].
- (iv) Least-loaded algorithm (LL). A well-known dynamic strategy for load balancing adopted in several commercial solutions. The main idea is assigning the incoming requests to the server with least load in that moment. However, an obvious shortcoming for this scheme is the least loaded server is saturated frequently

until an updated state information is propagated [35].

- (v) Fastest response time. Acting as the alternate of LL that the request is assigned to the server with the shortest expected response time. As a result, the algorithm is not stable enough [36].
- (vi) Two random choices algorithm (2RC). An interesting revision of RAND, which randomly chooses two servers and distributes the request to the one with less load. The performance is improved exponentially, compared to RAND [37].
- (vii) Next-neighbor load sharing. It's the advanced version of 2RC that combines the RAND and 2RC, i.e., firstly, selecting one server randomly, and then, allocate the request to either that server or its neighbor based on their respective loads (the least loaded server is chosen). It shows good performance especially when the load is heavy [38].

2.1.2 Cost-aware redirection with fixed service provisioning

In contrast to the cost-blind mechanisms, cost-aware mechanisms use some other factors as the additional metric, to reach a more reasonable load-balanced system. Other than the traffic itself, some typical metrics are as follows:

- (i) Resource-aware. As the network is becoming more complicated, the heterogeneity in the network (e.g. computing, bandwidth and memory availability) should not be ignored. The load balancing system should take a comprehensive consideration for all the resource availability information gathered. Lots of graph theory based models have been proposed to address the heterogeneous problem, for instance, octree partitioning [39], directed graph model [40], and multilevel graph model [41].
- (ii) Content-aware. In a CDN, replica placement and request redirection are two main technical problems to be considered equivalently. Most of the research works study them separately, i.e., one work only focus on one of the two aspects usually. However, some researchers argue that the two problems have

close inter-relation and they consider the two problem jointly as a newer approach for load balancing in CDNs [10, 42, 43]. In addition, the recent work [44] considers this joint problem in the cloud CDNs. A more complicated scenario which takes content eviction into consideration, together with the two aforementioned factors, is studied in [45]. Bjorkqvist *et al.* [46] address that, to retrieve a content in the CDNs, the system always needs to accommodate the trade-offs between vertical communication, peer communication, storage capacity and caching policies. They propose a optimization framework in [46] to minimize the total content retrieval cost in multi-layer CDNs. This problem is also investigated in the cloud CDNs [47].

- (iii) Energy-aware. Energy management in data centers has become an active area of research in recent years (e.g. [48] presents dynamic algorithms for turning servers on and off in data centers to save energy). [49] suggests routing requests to locations with the cheapest energy. [50] proposes the load balancing schemes in CDNs by maximizing energy reduction and minimizing the impact on client-perceived Service Level Agreements (SLAs). Green renewable energy is taken into consideration for load balancing in [51].

2.1.3 Cost-aware redirection with elastic service scaling

The earliest works on request redirection or load balancing focus on servers with known static service rates and capacities. Several mechanisms including random balancing and weighted round robin methods [30–34] were proposed. Subsequently, server loads [35, 37, 38] and response times [36] are taken into consideration when choosing the servers for redirection, leading to lower service response times. Due to the increasing complexity and heterogeneity of CDNs in terms of available computing, bandwidth and memory resources, several works have proposed request redirection methods based on octree partitioning [39], directed graph models [40], and multilevel graph models [41]. In addition, [49, 50] propose load balancing by routing requests to locations with the cheapest energy or by maximizing energy reduction and minimizing

the impact on on client-perceived SLAs.

All the above works address the issue of request redirection in a traditional CDN, where the servers have fixed service rates. In contrast, in a CCMN, the servers elastically scale their service rates to match the packet arrival rates, with the user paying a higher cost for a higher service rate. In [52], the request redirection problem is considered for a data center, in which a master server splits its request arrivals to all available computing servers, each with a different rate. The authors consider a linear computing cost for each computing server, and their goal is to optimize the service rates of all servers to minimize the system response time. [53] studies the interaction of service rate scaling with load balancing, which is defined in the strict sense that a performance metric (like mean response time) at all servers with positive redirection rates has the same value. Equilibrium redirection and service rates are derived under this load balancing constraint.

Minimizing content retrieval latency using caching and edge servers has been studied in [47], while [54] investigates the joint problem of building distribution paths and replica placement in cloud-based CDNs. Although finding optimal distribution paths and replica placements are topics of importance and related to the problem of user request redirection, these are not within the scope of this dissertation. Instead we assume that the redirection destinations have been fixed, and costs like the routing cost and computing cost function at each of these destinations are known. The joint problem of content placement and request redirection has been considered in [44] under various constraints. However, the proposed algorithm is an integer linear program that has a high complexity and is sensitive to variations in request traffic, i.e., a new optimization is required at each time step. It is still an open problem to find good approximations to the joint optimization problem that yield reasonable sub-optimal solutions.

2.2 Resource allocation problems in C-RAN

C-RAN aims to be a competitive and potential 5G framework, which has been attracting comprehensive research attention from both industry and academia since 2011 [55–58]. Many prototypes, test-beds and architecture designs have been done to show the feasibility and performance gain by adopting C-RAN [59–62]. The concept of RAN as a service (RANaaS) has also been developed based on the structure of C-RAN [63]. Moreover, the problem of dynamic frequency reuse in heterogeneous C-RAN is investigated in [64] and the computationally aware strategy is proposed to reduce the computational outage in C-RAN [65] recently. A survey for C-RAN was also conducted in [22].

C-RAN provides a centralized BBU pool, instead of the distributed base stations, to improve the resource utilization, such as the hardware and the energy, and enable the centralized processing, e.g., joint precoding. However, there are two main side effects, the first one is that the channel state information (CSI) matrix becomes huge, and the second one is the high amount of data transfer in the fronthaul, whose capacity is actually limited in practice [66]. On the one hand, for the large CSI overhead, [67] combines the instantaneous CSI and statistical CSI and [68] proposes a threshold-based channel matrix sparsification method to reduce the CSI overhead, while [69] considers the imperfect CSI in C-RAN. On the other hand, limited fronthaul capacity in C-RAN is studied in [70–79]. Specifically, [70–72] aims to minimize the number of active UE-RRH pairs (i.e., the fronthaul capacity is one of the terms in the objective function to be minimized) to mitigate the amount of data transfer in the fronthaul, while in [73–76], they set the fronthaul capacity as a constraint of certain optimization problems, e.g., sum-power minimization problem; [77–79] develop efficient signal compression/quantization algorithm to downsize the load in C-RAN fronthaul. In addition, to reduce the amount of data transfer in C-RAN, caching in the access points is also a promising approach [80, 81].

The BBU pool of C-RAN comprises many general purpose processors (GPP), which forms a cloud computing infrastructure using virtualization technology [66].

However, most of the previous works related to C-RAN BBU pool just make use of the centralized processing property offered by cloud computing to optimize the system. For example, a central encoder is developed in [79] to jointly encode the messages intended for the mobile stations, and a cloud decoder in [82] utilizes the joint statistics of the received correlated signals to decompress the received signal. These works do not consider the issue of elastic service scaling and resource allocation the BBU pool, which is one of the focus of this dissertation. In addition, unlike most of the works in the literature, which investigate methods to provide QoS guarantees for specific layers in the OSI stack (e.g., ensuring bandwidth or latency requirements are met only in the wireless transmission part), we consider methods to ensure cross-layer QoS guarantees in this dissertation.

2.2.1 RRH selection and user association

In the fronthaul of C-RAN, including the fronthaul links, the RRHs and the wireless channel, coordinated multipoint (CoMP) techniques are deployed to enhance the system throughput. In order to enhance energy efficiency, cell, BS or RRH selection for the fronthaul has been comprehensively studied over the past several years [83–89]. For example, the authors in [84] and [86] jointly consider the base station selection problem and linear precoding. Fronthaul link or RRH selection for C-RAN have been studied by [70, 71, 82, 90–92]. For instance, [82] considers joint BS selection and distributed compression in C-RAN to improve energy efficiency, while [91] considers RRH selection jointly with fronthaul beamforming to minimize the system power consumption. Reference [67] aims to reduce the overhead of obtaining instantaneous CSI in C-RAN by combining the instantaneous CSI and statistical CSI.

The joint RRH selection and resource allocation problem is NP-hard [91], therefore to solve it exactly is computationally intractable when the number of RRHs is large. We summarize some commonly used approaches here.

- In the first approach, the problem is formulated as a mixed-integer nonlinear programming (MINLP), and then solved by Branch and Bound (BnB) or Branch

and Cut (BnC) methods [89]. Both the BnB and BnC methods yield the optimal solution, but have high time complexity.

- Another approach is the “*sorting-and-removing*” heuristic method [85, 91], in which the RRHs are ranked according to some priority criteria in each iteration, and the RRH with the lowest priority is removed. The process continues until the problem becomes infeasible. This method can produce a near-optimal solution, but still has high computational complexity.
- The “*sparsity-inducing*” method is inspired by compressive sensing. Reweighted l_1 -norm relaxation and sparsity-inducing norms are used to obtain a sparse subset of RRHs [70, 82, 91]. This method is efficient in computational complexity but cannot guarantee optimality.
- Finally, constructing a Markov Chain Monte Carlo (MCMC) is a potential way to solve the RRH selection problem as well [88].

In Chapter 4, we propose a “*Shaping-and-Pruning*” method, which is a trade-off between the sorting-and-removing and sparsity-inducing methods, in order to obtain a near-optimal performance with lower computational complexity.

2.2.2 Weighted sum-rate maximization

Weighted sum-rate maximization (WSRM) is a very important problem in multiuser wireless networking system, especially in multiuser MISO/MIMO system. It has attracted many research attentions in recent years. However, the global optimal solution for this problem is normally intractable, since the problem is nonconvex (or even NP-hard [93]), and the problem was treated as an open problem [94]. We summarise some of the recently proposed methods to find the global/local optimal solutions as follows:

- The BnB algorithm is proposed in [95, 96] based on the fact that the objective function of the WSRM problem is monotonic w.r.t. each user’s achievable rate.

This algorithm can obtain the global optimal solution with finite iterations, but still with high time complexity.

- Based on the well-known relation between mean square error (MSE) and user's achievable rate, an iterative weighted minimum mean square error (WMMSE) algorithm is proposed in [97] to solve the WSRM problem. Although this algorithm involves alternating optimization and can only guarantee the local optimal solution, its much lower time complexity than BnB method makes the algorithm widely adopted, for instance, in the max-min fairness problem [98, 99]. A similar idea is also proposed in [100, 101].
- By observing that the user's achievable rate can be written as the difference of two convex functions (d.c.), the d.c. programming algorithm is investigated in [102, 103] to solve the WSRM problem. This algorithm converges to local optimal solution.
- By constructing a sequence of polyblocks, monotonic optimization methods [104] are used to solve the WSRM problem in [105]. This method provides global optimal solution.
- Exploiting the nonnegativity of the channel gain, transmit powers and user's achievable rates, the authors in [106, 107] make use of the nonnegative matrix theory, e.g. the Perron-Frobenius theory, to find out the global optimal solution for the WSRM problem.

Some of these methods can be extended to solve the weighted sum-utility maximization (WSUM) problem, such as the WMMSE and BnB methods. We utilize and generalize the WMMSE and BnB methods to solve problems in this dissertation.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Dynamic Request Redirection and Elastic Service Scaling in CCMN

Evolving from CDN to CCMN, the introduction of cloud computing poses additional technical challenges in its operations. Specifically, a lot of design and operational challenges that have been resolved in CDNs need to be revisited in CCMN. One particular example of importance is the user-redirection mechanism, which should take into consideration the new resource allocation paradigm under cloud computing.

In this chapter, we address the problem of dynamically scaling the request allocation rates and service capacities of VMs, where the request allocation rate of a VM is the number of transaction requests allocated by a dispatcher to the VM per unit time, and the service capacity of a VM is the number of transactions per unit time completed by the VM, typically measured in transactions per second (TPS) [108]. In order to mitigate the under-utilization and high latency problems alluded to earlier, we adopt the quickest detection framework of [109] to detect the onset of flash crowds, and derive optimal allocation and service capacity policies that take into account detection delays as well as false alarms.

This chapter differs from the most related works [52] and [53] in the following ways:

- We consider a different request redirection model that is somewhat simpler than

the one in [52] but more suited to a CDN. Specifically, we assume that there is a dispatcher that splits arrival requests to multiple servers optimally. Our goal is to minimize a weighted cost function consisting of the system response time, server computing costs and routing costs. In our model, a routing cost to each server is specified to more accurately reflect the operation of a CDN. Furthermore, we aim to optimize not only the service rates but also the redirection rates to each server. In contrast to [52], which fix the redirection rates in advance, we obtain the somewhat surprising conclusion that not all available servers should be utilized.

- Compared with [53], our work also have many novelties. Firstly, we derive optimal redirection and service rates to minimize a cost function similar to [53] but without the strict load balancing constraint. This models a CCMN better as a CCMN is designed for elastic rate allocation without the need to maintain the same equilibrium performance at each server. Secondly, we do not restrict to monomial power cost functions at the servers. Instead, we consider a general class of cost functions that is non-decreasing and convex in the service rate. This allows us to include various costs like computation, power, storage, and routing costs. Lastly, we allow the arrival traffic to switch between different modes to better model usage patterns in a CCMN, whereas [53] assumes that the arrival traffic statistics are fixed.

3.1 Problem formulation

In this section, we describe our model and system setup, define some notations, and state our assumptions. We summarize some commonly used notations for this chapter in Table 3.1.

We consider the problem of redirecting an arrival request process at a dispatcher or load balancer in a CCMN (cf. Figure 3-1), which tends to experience sporadic bursts of user traffic [9]. Since Poisson processes have been widely used to model request arrivals under both normal and flash crowd traffic conditions [110, 111], we

Table 3.1: Summary of common notations in Chapter 3.

Symbol	Definition
α_J	traffic arrival rate to the dispatcher in mode J , $J \in \{N, F\}$
$\lambda_{J,i}$	request allocation rate to VM S_i in traffic mode J
$\mu_{J,i}$	service capacity of VM S_i in traffic mode J
π	average fraction of time that the arrival traffic is in normal mode N
n	maximum number of VMs available to the dispatcher
$s_i(\cdot)$	computing cost of VM S_i , as a function of its service capacity
c_i	unit routing cost to VM S_i
$\varphi_i(\cdot)$	$= s_i(\cdot) + c_i$, service cost of VM S_i
β	weight for the total expected service cost incurred
$\mathcal{T}$	set of stopping times based on the number of request arrivals in each time interval
T_N	stopping time for detecting the switch of arrival traffic from mode N to mode F
T_F	stopping time for detecting the switch of arrival traffic from mode F to mode N
$\delta(T)$	worst case detection delay of stopping time T
$\theta(T)$	mean time between false alarms of stopping time T
c_F	average additional cost incurred per request arrival due to detection delays by T_N or false alarms by T_F
$v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)$	average additional cost per unit time due to false alarms by T_N or detection delays by T_F
l_J	mean time duration for each continuous time period that the arrival traffic is in mode J
t_f	maximum amount of time within which false alarms are resolved
p_i	VM price of VM S_i , see (3.16)

assume that the arrival process is a Poisson process that switches between a normal mode (denoted as N) with arrival rate α_N and a flash crowd mode (denoted as F) with arrival rate $\alpha_F > \alpha_N$ at unknown times [12, 110].

The dispatcher splits the arrival requests into independent sub-processes, which are redirected to a *maximum* of $n > 0$ destination VMs $S_1, \dots, S_n$. Suppose that the dispatcher has determined that the arrival process is in mode J , where $J \in \{N, F\}$. Let $\lambda_{J,i}$ be the *allocation rate* or the rate of the redirected Poisson arrival process at VM S_i in traffic mode J . A VM is said to be *active* in mode J if the allocation rate $\lambda_{J,i} > 0$ (i.e., the dispatcher may choose to use less than n VMs). Let $\mu_{J,i}$ be the *service capacity* provided by S_i , i.e., the number of arrival transactions that can be

Figure 3-1: Request redirection model.

served by the VM per unit time. Similar to assumptions commonly adopted in the CDN and multimedia cloud literature [52, 112], we assume that the service time is an exponential distribution with rate $\mu_{J,i}$. Each VM in Figure 3-1 is therefore modeled as an $M/M/1$ queue. To ensure stability, we require that $\lambda_{J,i} < \mu_{J,i}$ for all i and J . Let $\boldsymbol{\lambda}_J = (\lambda_{J,1}, \dots, \lambda_{J,n})$ and $\boldsymbol{\mu}_J = (\mu_{J,1}, \dots, \mu_{J,n})$ be vectors of the allocation rates and service capacities respectively under traffic mode J . A pair $(\boldsymbol{\lambda}_J, \boldsymbol{\mu}_J)$ is called a *policy*.

A redirection to VM S_i incurs a *service cost* $\varphi_i(\mu_{J,i}) = s_i(\mu_{J,i}) + c_i$ for each request, where c_i is the unit routing cost, and $s_i(\cdot)$ is the computing cost. For simplicity, we assume that the total routing cost $\lambda_{J,i}c_i$ scales linearly with the allocation rate $\lambda_{J,i}$. The computing cost $s_i(\cdot)$ at each VM is non-decreasing in the service capacity. This models a cloud infrastructure where one or more VMs are initiated for each application, and a larger cost is incurred if more computing resources are requested by each VM. The computing cost includes the cost of the power used by S_i , and the memory or storage costs required. The power cost is typically assumed to be monomial in the service capacity [53], but we do not restrict to such cost functions in this chapter. Instead, we make the following assumptions regarding the computing cost.

Assumption 3.1. For each VM S_i , $i = 1, \dots, n$, the computing cost $s_i(\cdot)$ has the

Table 3.2: Amazon EC2 instances and prices

Name	No. of EC2 Compute Units	Price
Small	1	0.06 USD/hour
Medium	2	0.12 USD/hour
Large	4	0.24 USD/hour
Extra Large	8	0.48 USD/hour

following properties:

- (1) $s_i(\mu) \geq 0$ for all $\mu \geq 0$;
- (2) $s_i(\mu)$ is a convex and non-decreasing function of μ .

Assumption 3.1 models many computing costs of interest in cloud architectures. To the best of our knowledge, all existing commercial cloud platforms like Amazon EC2, Rackspace, Google App Engine and Microsoft Windows Azure, charge flat hourly prices for a fixed amount of resources or each type of instance. An example is the Amazon EC2 Standard On-demand Instance pricing model [113]. Under this plan, four different types of VM instances are available (see Table 3.2). These are called the Small, Medium, Large, and Extra Large instances, with 1, 2, 4, and 8 EC2 Compute Units respectively. Each EC2 Compute Unit has a fixed service capacity of $\bar{\mu}$ transactions per hour, and the price of each unit is $p_{\text{EC2}} = 0.06/\bar{\mu}$ USD¹ per transaction. The cost of an instance S_i can then be modeled by

$$s_i(\mu) = \begin{cases} e_i p_{\text{EC2}}, & \text{if } \mu \leq e_i \bar{\mu}, \\ \infty, & \text{if } \mu > e_i \bar{\mu}, \end{cases} \quad (3.1)$$

where $e_i = 1, 2, 4$, or 8 depending on the type of instance that S_i belongs to. This cost can be regarded as the “computing cost” if a dispatcher is redirecting application requests to Amazon EC2 Standard On-demand instances, and it clearly satisfies Assumption 3.1.

The Amazon EC2 pricing model and other similar commercial cloud computing pricing models are designed with retail and enterprise consumers in mind, where lower level costs like power consumption and memory storage costs are transparent

¹USD is abbreviation for United States dollar.

to a user, and included in a single price p_{EC2} . For the cloud service providers, to optimize a data center, the computing cost can be more appropriately modeled as $s_i(\mu) = k_i\mu^{a_i}$, where $k_i > 0$ and $a_i > 1$ are positive constants. This cost corresponds to the power used by the VM when its service capacity is μ [114], and is adopted by [53]. In addition to the cost of power consumption, we can also include the cost of processor and storage memory by letting $s_i(\mu) = k_i\mu^{a_i} + r_i\mu$, where $r_i > 0$, and we have assumed that memory or storage cost scales linearly with service capacity [44]. (Such general cost functions are not considered in [53].) Furthermore, in practical systems, the maximum service capacity of any VM is often limited; a fact that is not captured in smooth cost functions. Our model allows us to define cost functions similar to (3.1), where an infinite cost is incurred when the service capacity exceeds the maximum VM computation capacity.

Assumption 3.1 implies that there is a trade off between the average mean response time [115] of the VMs given by

$$\sum_{i=1}^n \frac{\lambda_{J,i}}{\alpha_J} \frac{1}{\mu_{J,i} - \lambda_{J,i}},$$

and the total expected service cost incurred, given by

$$\sum_{i=1}^n \frac{\lambda_{J,i}}{\alpha_J} \varphi_i(\mu_{J,i}).$$

Let $\beta > 0$ be a fixed weight and for $J \in \{N, F\}$, $\boldsymbol{\lambda}_J = (\lambda_{J,1}, \dots, \lambda_{J,n})$, $\boldsymbol{\mu}_J = (\mu_{J,1}, \dots, \mu_{J,n})$, and $\alpha_J > 0$, let

$$f(\boldsymbol{\lambda}_J, \boldsymbol{\mu}_J, \alpha_J) = \sum_{i=1}^n \frac{\lambda_{J,i}}{\alpha_J} \frac{1}{\mu_{J,i} - \lambda_{J,i}} + \beta \sum_{i=1}^n \frac{\lambda_{J,i}}{\alpha_J} \varphi_i(\mu_{J,i}). \quad (3.2)$$

We are interested to find policies that minimize a weighted average of $f(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \alpha_N)$ and $f(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \alpha_F)$, while taking into consideration the switching traffic modes. To do this, we divide time into equal unit intervals, and suppose that $Z_1, \dots, Z_{\nu-1}$ are the number of arrivals in each interval when the arrival process is in mode N , and

$Z_\nu, Z_{\nu+1}, \dots$ are the number of arrivals in each interval when the arrival process has switched to mode F . The random variables Z_i , for $i = 1, \dots, \nu - 1$ are independent and identically distributed (i.i.d.) Poisson random variables with rate α_N , while Z_i , for $i \geq \nu$ are i.i.d. Poisson random variables with rate α_F . The index ν is called a change point, and can take values in $[1, \infty]$, with $\nu = \infty$ corresponding to the case where the arrival process stays in mode N throughout. As ν is unknown, it is inferred from the observations $Z_1, Z_2, \dots$. Let $\mathcal{T}$ be the set of stopping times associated with $Z_1, Z_2, \dots$ [109]. Each element of $\mathcal{T}$ is a detection policy that can be used by the dispatcher to determine if the arrival process has switched from mode N to mode F . Any detection policy $T \in \mathcal{T}$ has a trade-off in the worst case detection delay [109, 116]

$$\delta(T) = \sup_{t \geq 1} \text{ess sup } \mathbb{E}_t[(T - t + 1)^+ \mid Z_1, \dots, Z_{t-1}],$$

where $\mathbb{E}_t$ is the expectation operator when $\nu = t$, with its mean time between false alarms defined as

$$\theta(T) = \mathbb{E}_\infty[T].$$

One can similarly define the worst case detection delay and mean time between false alarms for the case where the arrival traffic switches from mode F to mode N .

Our goal is to find allocation rates λ_J and service capacities μ_J , for $J \in \{N, F\}$, that are optimal in the sense that they minimize a cost function that takes into account the trade-offs in performance at each VM, the service cost incurred at each VM, and the costs incurred due to detection delays and false alarms. Suppose that the arrival traffic switches from mode N to mode F . When there is a delay in detecting the change in the arrival traffic mode, there will be a transient increase in the buffer length at each VM as the underlying arrival rate has increased to α_F . This incurs additional memory or storage cost and an increase in the service response time. We assume that the average additional cost per unit time is given by $c_F(\alpha_F - \alpha_N)$, where c_F is a positive constant that can be interpreted as the average additional cost incurred per request arrival. On the other hand, if a false alarm occurs, the system adopts the

policy $(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)$ even though the policy $(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N)$ is optimal. This results in higher computing costs at the VMs. We assume that false alarms are resolved within a fixed bounded time period t_f , with the average additional cost per unit time given by a penalty function $v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)$ incurred at the active VMs. A similar consideration can be made when the arrival traffic switches from mode F to mode N . For simplicity, we assume that there is an ergodic stochastic process governing the traffic modes. For example, the sequence of traffic modes may be modeled as a continuous time Markov process, with transitions between modes N and F . We do not require full knowledge of this underlying process, except that the average fraction of time that the arrival process is in normal mode N is given by $\pi \in (0, 1)$, which can be estimated from historical data. In Section 3.4, we present simulation results to show how the performance of our proposed algorithm is impacted by the value of π .

Let T_N and T_F be stopping times for detecting the switch of arrival traffic from mode N to mode F , and vice versa, respectively. Consider a long interval consisting of l request arrivals, and let K_J be the set of requests at which the arrival process is in mode J , for $J \in \{N, F\}$. Let D_N and D_F be the sets of change points at which the arrival traffic switches from mode N to mode F , and vice versa, respectively. For each $J \in \{N, F\}$ and $i \in D_J$, let d_i be the length of the detection delay incurred by the stopping time T_J . We approximate the detection delays d_i to be independent and identically distributed for each J . Finally, let F_J be the set of arrivals at which T_J incurs false alarms. Assuming that VMs' response times reach equilibrium and are independent of the length of the traffic mode, the average equilibrium cost can be approximated as

$$\begin{aligned}
& \frac{|K_N|}{l} f(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \alpha_N) + \frac{|K_F|}{l} f(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \alpha_F) \\
& + \frac{1}{l} \sum_{i \in D_N} d_i c_F(\alpha_F - \alpha_N) + \frac{1}{l} \sum_{i \in D_F} d_i v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F) \\
& + \frac{|F_N|}{l} t_f v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F) + \frac{|F_F|}{l} t_f c_F(\alpha_F - \alpha_N).
\end{aligned} \tag{3.3}$$

For $J \in \{N, F\}$, let l_J be the mean time duration for each continuous time period that

the arrival traffic is in mode J . Letting $l \rightarrow \infty$, we see that (3.3) can be approximately upper bounded by (3.4) below. Our optimization problem is then formulated as

$$\begin{aligned}
& \min_{\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, T_N, T_F} \quad \pi f(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \alpha_N) + (1 - \pi) f(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \alpha_F) \\
& \quad + c_F(\alpha_F - \alpha_N) \frac{(1 - \pi)}{\alpha_F} \left(\frac{\delta(T_N)}{l_F} + \frac{t_f}{\theta(T_F)} \right) \\
& \quad + v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F) \frac{\pi}{\alpha_N} \left(\frac{\delta(T_F)}{l_N} + \frac{t_f}{\theta(T_N)} \right), \\
& \text{s.t.} \quad \sum_{i=1}^n \lambda_{J,i} = \alpha_J, \text{ for } J \in \{N, F\}, \\
& \quad 0 \leq \lambda_{J,i} < \mu_{J,i}, \text{ for } 1 \leq i \leq n, J \in \{N, F\}, \\
& \quad T_N, T_F \in \mathcal{T},
\end{aligned} \tag{3.4}$$

where “s.t.” stands for “subject to”.

In this chapter, we suppose that the penalty function $v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)$ is separable across VMs and satisfies the following assumptions.

Assumption 3.2. Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$.

(i) The penalty function $v(\boldsymbol{\lambda}, \boldsymbol{\mu})$ has the form

$$v(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{i=1}^n \lambda_i v_i(\mu_i) - \xi,$$

where ξ is a positive constant, and for each $i = 1, \dots, n$, $v_i(\cdot)$ is a non-negative, convex, and non-decreasing function.

(ii) We have $\min v(\boldsymbol{\lambda}, \boldsymbol{\mu}) \geq 0$, where the minimization is over all $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ such that $0 \leq \lambda_i < \mu_i$ for all $i = 1, \dots, n$, and $\sum_{i=1}^n \lambda_i = \alpha_F$.

Assumption 3.2 covers many, but not all, practical cases of interest. For example, in designing a CCMN system, we can interpret the constant ξ in Assumption 3.2 as the average cost (storage, service response time, computing cost etc.) incurred per unit time when adopting the policy $(\boldsymbol{\lambda}_N^*, \boldsymbol{\mu}_N^*) = \arg \min f(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \alpha_N)$ during normal traffic mode, while the function $v_i(\mu)$ is the average cost incurred when VM

S_i provides a service capacity of μ . Assumption 3.2(ii) ensures that the penalty function is always positive over all feasible allocation rates and service capacities, which is the case for a practical system. We note that Assumption 3.2 does not cover the most general case where the penalty function is of the form $v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \boldsymbol{\lambda}_N, \boldsymbol{\mu}_N)$, which however introduces technical difficulties into the solution of (3.4) and its interpretation. In Section 3.4, we provide simulation results that suggest that Assumption 3.2 does not significantly impact the performance of the policies derived from (3.4), compared to the “perfect” strategy that knows the exact points in time when the traffic switches its mode.

3.1.1 Approximation for detection delay and false alarm frequency

In this section, we briefly review the theory of quickest detection in a non-Bayesian setting, and derive approximations to the change detection delay and mean time between false alarms, from which an approximation to the optimization problem (3.4) is then obtained.

The problem of quickest detection is to optimally detect a change in the underlying distribution from which a sequence of observations $Z_1, Z_2, \dots$ is drawn from, subject to certain false alarm constraints. The observations are drawn i.i.d. from distributions Q_0 and Q_1 before and after an unknown change point ν respectively. Since at each time t , we only have access to the previously observed random variables Z_k , for $k \leq t$, the change detection policy is a stopping time $T \in \mathcal{T}$.

A commonly used stopping time is Page’s CUSUM test [109, 117] given by

$$T_\sigma = \inf\{k \geq 0 : \max_{1 \leq j \leq k} \sum_{l=j}^k L(Z_l) \geq \log \sigma\}, \quad (3.5)$$

where $L(z) = \log \frac{dQ_1}{dQ_0}(z)$ is the log-likelihood ratio of Q_0 w.r.t. Q_1 . It is well known that Page’s test is an optimal change detection policy in the sense that for any σ , the test T_σ minimizes the detection delay $\delta(T)$ among all stopping times T satisfying

$\theta(T) \geq \theta(T_\sigma)$. The following result follows from the optimality of Page's CUSUM test, and Wald's approximations [109], and its proof is relegated to Section 3.6.1.

Proposition 3.1. Suppose A and B are positive constants, and $L(z)$ has no atoms under Q_0 . There exists a threshold $\sigma > 0$ such that an optimal solution to the following optimization problem

$$\begin{aligned} \min_T \quad & A\delta(T) + \frac{B}{\theta(T)} \\ \text{s.t.} \quad & T \in \mathcal{T}, \end{aligned} \tag{3.6}$$

has the form (3.5). Furthermore, if $\sigma \geq 1$, then the optimal detection delay and the mean time between false alarms can be approximated as

$$\delta(T_\sigma) \cong \frac{1}{M_1} \left(\frac{1}{\sigma} - 1 + \log \sigma \right), \tag{3.7}$$

$$\theta(T_\sigma) \cong \frac{1}{M_0} (\sigma - \log \sigma - 1), \tag{3.8}$$

and the optimal threshold σ can be approximated as the solution to

$$A(\sigma - \log \sigma - 1)^2 = BM_0M_1\sigma, \tag{3.9}$$

where for $i = 0, 1$, $M_i = |\mathbb{E}_{Q_i}[L(Z)]|$, and $\mathbb{E}_{Q_i}$ is the expectation operator under the probability distribution Q_i .

We now return to the optimization problem in (3.4). For any fixed rates $\{\boldsymbol{\lambda}_J, \boldsymbol{\mu}_J \mid j \in \{N, F\}\}$, finding the optimal stopping time T_N is equivalent to solving the optimization problem (3.6), with $A = (1-\pi)c_F(\alpha_F - \alpha_N)/(\alpha_F l_F)$ and $B = \pi t_f v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)/\alpha_N$, while to find the optimal T_F , we set $A = \pi v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)/(\alpha_N l_N)$ and $B = (1-\pi)c_F t_f(\alpha_F - \alpha_N)/\alpha_F$. To simplify the mathematics, we use the approximations in Proposition 3.1 to arrive at the following approximation for the objective function in (3.4),

$$\begin{aligned} & \mathbf{Q}(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \sigma_N, \sigma_F) \\ &= \pi f(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \alpha_N) + (1 - \pi) f(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \alpha_F) \end{aligned}$$

$$\begin{aligned}
& + c_F(\alpha_F - \alpha_N) \frac{(1 - \pi)}{\alpha_F} \left(\frac{1}{l_F M_F} \left(\frac{1}{\sigma_N} - 1 + \log \sigma_N \right) + \frac{t_f M_F}{\sigma_F - \log \sigma_F - 1} \right) \\
& + v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F) \frac{\pi}{\alpha_N} \left(\frac{1}{l_N M_N} \left(\frac{1}{\sigma_F} - 1 + \log \sigma_F \right) + \frac{t_f M_N}{\sigma_N - \log \sigma_N - 1} \right), \quad (3.10)
\end{aligned}$$

where for $J \in \{N, F\}$, σ_J is the threshold corresponding to T_J and $M_J = |\alpha_J \log(\alpha_N/\alpha_F) + (\alpha_F - \alpha_N)|$ (The M_J here is based on the assumption that the arrival processes are Poisson process in both normal and flash crowd traffic mode).

With this approximation, in the rest of this chapter, we focus on obtaining the solution to the following optimization problem:

$$\begin{aligned}
\min \quad & \mathbf{Q}(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \sigma_N, \sigma_F), \\
\text{s.t.} \quad & \sum_{i=1}^n \lambda_{J,i} = \alpha_J, \text{ for } J \in \{N, F\}, \\
& 0 \leq \lambda_{J,i} < \mu_{J,i}, \text{ for } 1 \leq i \leq n, J \in \{N, F\}, \\
& \sigma_J \geq 1, \text{ for } J \in \{N, F\}.
\end{aligned} \quad (3.11)$$

We note that to solve (3.11), the parameters $\pi, \alpha_N, \alpha_F, l_N$, and l_F need to be first estimated from historical data (using for example, the Maximum Likelihood Estimator (MLE) approach [118]), while the remaining parameters c_F and t_f , and penalty function $v(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)$ need to be chosen appropriately depending on the application. We show an example in Section 3.4.

3.2 Alternating optimization

In this section, we apply the alternating optimization method to simplify the optimization problem (3.11). We partition the variables in (3.11) into three subsets: $(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N)$, $(\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F)$ and the thresholds (σ_N, σ_F) . We start with a random initial guess $(\boldsymbol{\lambda}_N(0), \boldsymbol{\mu}_N(0), \boldsymbol{\lambda}_F(0), \boldsymbol{\mu}_F(0), \sigma_N(0), \sigma_F(0))$ for the optimization variables respectively.

At each iteration t , we perform the following series of optimizations,

$$(\boldsymbol{\lambda}_N(t), \boldsymbol{\mu}_N(t)) = \arg \min_{\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N} \mathbf{Q}(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \boldsymbol{\lambda}_F(t-1), \boldsymbol{\mu}_F(t-1), \sigma_N(t-1), \sigma_F(t-1)) \quad (3.12)$$

$$(\boldsymbol{\lambda}_F(t), \boldsymbol{\mu}_F(t)) = \arg \min_{\boldsymbol{\lambda}_F, \boldsymbol{\mu}_F} \mathbf{Q}(\boldsymbol{\lambda}_N(t), \boldsymbol{\mu}_N(t), \boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \sigma_N(t-1), \sigma_F(t-1)) \quad (3.13)$$

$$(\sigma_N(t), \sigma_F(t)) = \arg \min_{\sigma_N, \sigma_F} \mathbf{Q}(\boldsymbol{\lambda}_N(t), \boldsymbol{\mu}_N(t), \boldsymbol{\lambda}_F(t), \boldsymbol{\mu}_F(t), \sigma_N, \sigma_F) \quad (3.14)$$

where the minimizations in (3.12) and (3.13) are over all $\boldsymbol{\lambda}_J$ and $\boldsymbol{\mu}_J$ such that

$$\sum_{i=1}^n \lambda_{J,i} = \alpha_J,$$

and

$$0 \leq \lambda_{J,i} < \mu_{J,i}, \text{ for } 1 \leq i \leq n,$$

for $J = N$ and $J = F$ respectively, and the minimization in (3.14) is over all $\sigma \geq 1$. The minimization in (3.14) can be computed by solving (3.9) with the appropriate values of A and B substituted in. In the next section, we derive the solutions to the minimization problems (3.12) and (3.13).

Since $\mathbf{Q}(\boldsymbol{\lambda}_N, \boldsymbol{\mu}_N, \boldsymbol{\lambda}_F, \boldsymbol{\mu}_F, \sigma_N, \sigma_F) \geq 0$ and the objective function value is non-increasing at each iteration of the alternating optimization procedure, the iterates converge to a local minimum of $\mathbf{Q}(\cdot)$ as $t \rightarrow \infty$. To increase the chance of finding the global minimum, the procedure can be applied to several random initial guesses.

We show the logic flow chart for this chapter in Figure 3-2.

3.3 Optimal request allocation and service capacity

In this section, we derive optimal allocation rates and service capacities for the problems (3.12) and (3.13), which correspond to policies the dispatcher adopts after it has

Figure 3-2: The logic flow chart for Chapter 3.

determined that the arrival process is in mode N and F respectively.

To simplify notations, we drop the iteration index t in the alternating optimization procedure in this section. The problems (3.12) and (3.13) are both equivalent to the following optimization problem,

$$\begin{aligned}
 \min_{\mu_i, \lambda_i} \quad & \sum_{i=1}^n \frac{\lambda_i}{\alpha} \frac{1}{\mu_i - \lambda_i} + \beta \sum_{i=1}^n \frac{\lambda_i}{\alpha} w_i(\mu_i), \\
 \text{s.t.} \quad & \sum_{i=1}^n \lambda_i = \alpha, \\
 & 0 \leq \lambda_i < \mu_i, \text{ for } 1 \leq i \leq n,
 \end{aligned} \tag{3.15}$$

where we let $\alpha = \alpha_J$, $\lambda_i = \lambda_{J,i}$ and $\mu_i = \mu_{J,i}$, with $J = N$ and F for (3.12) and (3.13) respectively. In addition, we let $w_i(\mu) = \varphi_i(\mu)$ and

$$w_i(\mu) = \varphi_i(\mu) + \frac{\pi \alpha_F v_i(\mu)}{(1 - \pi) \beta \alpha_N} \left(\frac{1/\sigma_F - 1 + \log \sigma_F}{l_N M_N} + \frac{t_f M_N}{\sigma_N - \log \sigma_N - 1} \right),$$

for (3.12) and (3.13) respectively. We derive the general form of the optimal solution for (3.15), which is applicable for both (3.12) and (3.13).

3.3.1 VM prices

The form of the optimal solution to (3.15) is closely related to the price of each VM, which for $i = 1, \dots, n$, we define as

$$p_i = \min_{\mu > 0} \left\{ \frac{1}{\mu} + \beta w_i(\mu) \right\}. \quad (3.16)$$

Suppose that the dispatcher has to pay one dollar per unit service time, and β dollars per unit cost (routing and computation), then the price of VM S_i is the expected total price that the dispatcher pays to VM S_i . Interpreted in this way, it is natural that the dispatcher should choose VMs with the lowest prices such that the system is stable. Therefore, it follows that there is a threshold price, below which all VMs with a price cheaper than this threshold will be sent redirection requests. These are the active VMs. Furthermore, the VMs will provide service capacities that depend on this threshold price. We show that this intuitive argument holds in Theorems 3.1.

For each $i = 1, \dots, n$, and $p > 0$, let

$$g_i(p) = \sup \left\{ \mu : w_i(\mu) + \mu \partial^- w_i(\mu) \leq \frac{p}{\beta} \right\}. \quad (3.17)$$

Since w_i is convex, we have $\partial^- w_i(x)$ is non-decreasing, and $g_i(\cdot)$ is a non-decreasing function. If the computing cost $s_i(\mu)$ is continuously differentiable over all $\mu \in [0, \infty)$, $g_i(p)$ can be found as the implicit solution to

$$w_i(\mu) + \mu w'_i(\mu) = \frac{p}{\beta},$$

where $w'_i(\mu)$ is the first derivative w.r.t. μ . We make use of $g_i(\cdot)$ in Theorem 3.1 below to characterize the optimal number of active VMs, service capacities and allocation rates. We first prove a result regarding $g_i(p_i)$.

Lemma 3.1. For $i = 1, \dots, n$, let $\tilde{\mu}_i > 0$ be such that $p_i = 1/\tilde{\mu}_i + \beta w_i(\tilde{\mu}_i)$. Then, $\tilde{\mu}_i = g_i(p_i)$.

Proof. See Section 3.6.2. □

Theorem 3.1. Suppose that Assumptions 3.1 and 3.2 hold, and that $p_1 \leq \dots \leq p_n < p_{n+1} = \infty$. Then, the optimal service capacities for the optimization problem (3.15) are

$$\mu_i^* = \begin{cases} g_i(p), & \text{for } 1 \leq i \leq m^* \\ 0, & \text{for } m^* < i \leq n, \end{cases} \quad (3.18)$$

and the optimal allocation rates are

$$\lambda_i^* = \begin{cases} \mu_i^* - \sqrt{\frac{\mu_i^*}{p - \beta w_i(\mu_i^*)}}, & \text{for } 1 \leq i \leq m^* \\ 0, & \text{for } m^* < i \leq n, \end{cases} \quad (3.19)$$

where

$$m^* = \arg \max_{1 \leq k \leq n} \left\{ p_k \mid \sum_{i=1}^k \left(g_i(p_k) - \sqrt{\frac{g_i(p_k)}{p_k - \beta w_i(g_i(p_k))}} \right) < \alpha \right\}, \quad (3.20)$$

and $p \in (p_{m^*}, p_{m^*+1}]$ is such that $\sum_{i=1}^n \lambda_i^* = \alpha$.

Proof. See Section 3.6.3. □

An intuitive description for the relationship between server prices and the number of active VMs is depicted in Figure 3-3.

From Theorem 3.1, the allocation rates follow a water-filling solution, with the Lagrange multiplier p serving as a threshold price, and only VMs with prices below this threshold are sent redirection requests. In steps (3.12) and (3.13) of the alternating optimization procedure, the optimal policy can be found by using the following procedure:

1. The dispatcher sorts $\{p_i : i = 1, \dots, n\}$ in a non-decreasing order.
2. The dispatcher chooses a set of VMs using (3.20), the optimal allocation rates $\{\lambda_i^*\}_{i=1}^n$ using (3.19), and computes the price threshold p .
3. The dispatcher sends the price threshold p to the chosen active VMs.
4. Each active VM S_i provides service capacity $g_i(p)$.

Figure 3-3: VM prices sorted in non-decreasing order.

The complexity of the first step is $\mathcal{O}(n \log n)$.² To find the price threshold p and the set of active VMs in the second step, a binary search on the sorted array obtained in the first step produces an interval $(p_{m^*}, p_{m^*+1}]$ containing p , and the optimal number of VMs m^* . This has complexity $\mathcal{O}(n \log n)$. Assuming that the prices $\{p_i\}_{i=1}^n$ increases at most exponentially fast, a binary search in the interval $(p_{m^*}, p_{m^*+1}]$ takes $\mathcal{O}(n)$ iterations. Therefore, the computation complexity at the dispatcher is $\mathcal{O}(n \log n)$ for each iteration of the alternating optimization procedure.

3.3.2 Bounded service capacity

In this section, we consider the special case where each VM S_i , for $i = 1, \dots, n$, has bounded service capacity $\bar{\mu}_i$, and $\varphi_i(\mu) = v_i(\mu) = \bar{c}_i$. An example of such a system is given by the Amazon EC2 Standard On-demand Instance plan, as described in Section 3.1. We have the following corollary for solving (3.12), which follows from Theorem 3.1. A similar result holds for the problem (3.13).

Corollary 3.1. Suppose that Assumptions 3.1 and 3.2 hold, and the maximum ser-

²We say that $q(n) = \mathcal{O}(g(n))$ if $\limsup_{n \rightarrow \infty} |q(n)/g(n)| < \infty$.

vice capacity at VM S_i is $\bar{\mu}_i$ with constant service cost $\varphi(\mu) = \bar{c}_i$, for $i = 1, \dots, n$. Suppose further that $p_i = 1/\bar{\mu}_i + \beta\bar{c}_i$, for $i = 1, \dots, n$, are such that $p_1 \leq \dots \leq p_n < p_{n+1} = \infty$. Then the optimal number of active VMs for the optimization problem (3.12) is

$$m^* = \arg \max_{1 \leq k \leq n} 1 \left\{ p_k \left| \sum_{i=1}^k \left(\bar{\mu}_i - \sqrt{\frac{\bar{\mu}_i}{p_k - \beta\bar{c}_i}} \right) < \alpha \right. \right\}. \quad (3.21)$$

Furthermore, the optimal service capacity for $1 \leq i \leq m^*$ is $\mu_i^* = \bar{\mu}_i$, and the optimal allocation rates are

$$\lambda_i^* = \begin{cases} \bar{\mu}_i - \sqrt{\frac{\bar{\mu}_i}{p - \beta\bar{c}_i}} & , \text{ for } 1 \leq i \leq m^* \\ 0 & , \text{ for } m^* < i \leq n, \end{cases} \quad (3.22)$$

where $p \in (p_{m^*}, p_{m^*+1}]$ is such that $\sum_{i=1}^n \lambda_i^* = \alpha$.

3.4 Simulation results

In this section, we present simulation results to verify the performance of our proposed request allocation and service capacity scaling policy. We test the performance of our algorithm and various benchmark strategies on real YouTube request data collected by [119], and then on simulated request data in order to verify the impact of traffic arrival characteristics on the algorithms' performance. For ease of reference, we call the solution that we derive for the problem (3.11) via the alternating optimization method described in Section 3.2 the Dynamic Request Redirection and Elastic Service Scaling (DRES) strategy. We use the sum of service delay and service cost weighted by β as our performance criteria (cf. (3.2)).

We compare the performance of the DRES strategy with that of the following benchmark strategies:

- *N strategy.* This strategy assumes that the arrival traffic is always in the normal traffic mode N and the policy it adopts is obtained by minimizing $f(\boldsymbol{\lambda}, \boldsymbol{\mu}, \alpha_N)$.

Let the optimal allocation rates be $\boldsymbol{\lambda}^N$ and the service capacities be $\boldsymbol{\mu}^N$. For $J \in \{N, F\}$, the dispatcher redirects $\alpha_J \lambda_i^N / \alpha_N$ amount of traffic to VM S_i , which uses a service capacity of μ_i^N regardless of the arrival rate at that VM. When the arrival traffic is in mode F , the service capacity μ_i^N may be smaller than the actual traffic arrival rate at S_i , leading to a rapidly increasing service response time.

- *F strategy.* This strategy assumes that the arrival traffic is always in flash crowd mode F by adopting the allocation rates and service capacities obtained when minimizing $f(\boldsymbol{\lambda}, \boldsymbol{\mu}, \alpha_F)$. Suppose that the optimal allocation rates are $\boldsymbol{\lambda}^F$ and the service capacities are $\boldsymbol{\mu}^F$. In this strategy, we can guarantee that $\mu_i^F > \lambda_i^F > \alpha_J \lambda_i^F / \alpha_F$ for all $J \in \{N, F\}$. However, when the arrival traffic is in mode N , the policy used is not optimal, leading to higher service costs.
- *Perfect strategy.* We assume that we can perfectly detect the change points when the arrival traffic switches between mode N and mode F . We adopt the N or F policies when the arrival traffic is in mode N or F respectively. This strategy is unrealizable, and serves as a lower bound in our performance comparisons.

In our simulations, we adopt a cubic computing cost model, with $s_i(\mu) = k_c \mu^3$, where $k_c > 0$ for every $i = 1, \dots, n$. This cost function has been widely used to model the relationship between energy consumption and service capacities [53, 114]. We let the unit routing cost to VM S_i be $c_i = i c_0$, where $c_0 > 0$ is a constant.

To compute the DRES strategy, we let the false alarm penalty function in Assumption 3.2 be $v_i(\cdot) = \gamma \beta s_i(\cdot)$, for each $i = 1, \dots, n$, and where $\gamma > 0$ is a constant. We choose $\xi = \sum_{i=1}^n k_c \lambda_{N,i} s_i(\mu_{N,i})$, where $(\boldsymbol{\lambda}^N, \boldsymbol{\mu}^N)$ is the policy adopted by the N strategy. We choose γ to be sufficiently large to ensure that Assumption 3.2 holds. We also let the average additional cost per request during a flash crowd detection delay be $c_F = \bar{c}_F(f(\boldsymbol{\lambda}^F, \boldsymbol{\mu}^F, \alpha_F) - f(\boldsymbol{\lambda}^N, \boldsymbol{\mu}^N, \alpha_N))$, where $(\boldsymbol{\lambda}^F, \boldsymbol{\mu}^F)$ is the policy in the F strategy. The parameters used are given in Table 3.3, while the remaining parameters π , α_N , α_F , l_N and l_F are estimated from the YouTube trace data.

Table 3.3: Simulation parameters in Chapter 3

Parameter	Value	Parameter	Estimated
$\bar{c}_F$	100	π	0.497
γ	5	α_N	14.33
k_c	10	α_F	44.89
c_0	2	l_N	269.69 min.
t_f	5 min.	l_F	295.83 min.

3.4.1 YouTube trace data

YouTube request trace data was collected by [119] from a campus network over a period of 13 days. A profile of the number of requests per min is shown in Figure 3-5 for a typical day in the dataset. It is clear from Figure 3-5 that there are two periods (as indicated by the dotted lines) in which the average traffic arrival rate is significantly higher than that in other periods like the interval $[0,400]$. In order to define a flash crowd, we first take a running average of the number of requests over 30-minute windows to smooth out the data. Then, we define a flash crowd to be a period of at least 30 minutes, in which the running average traffic arrival rate is not less than 35 requests per minute [110].

The first 5 days of the YouTube dataset are used for parameter estimation. We find that on most of the days, flash crowds occur twice a day, and each flash crowd has a mean period of $l_F = 295.83$ minutes, while the normal traffic mode has a mean period of $l_N = 269.69$ minutes. We plot the histogram of the inter-arrival times for both normal traffic mode and flash crowd mode for the first 5 days of Youtube trace data in Figure 3-4. We see that the distribution of the inter-arrival time of the requests approximates the exponential distribution reasonably well, with a better fit for the flash crowd mode than the normal traffic mode. Therefore, it is reasonable to use Poisson arrival processes to model the arrival traffic in both traffic modes. It can then be shown that the MLE for the traffic arrival rate in each traffic mode is given by the average number of requests per minute over the traffic mode period [118]. We estimate that $\alpha_N = 14.33$ per minute and $\alpha_F = 44.89$ per minute. Finally, the fraction of time occupied by the normal traffic mode is given by $\pi = 0.497$.

We first show the objective function value (3.10) in each iteration of the alternating

(a) Normal traffic mode.

(b) Flash crowd mode.

Figure 3-4: Histogram of inter-arrival times under different traffic modes. The red curve shows a fitted exponential distribution.

Figure 3-5: Rate of YouTube requests over a typical one day period, starting at 3:40am and ending at 3:40am next day. The dotted lines indicate two periods in which the average traffic arrival rate is significantly higher than other times.

Figure 3-6: Convergence of the alternating optimization procedure for computing the DRES strategy.

optimization procedure used to compute the optimal policies for the DRES strategy. It can be seen from Figure 3-6 that the procedure converges in less than 10 iterations.

We next evaluate the performance of the different strategies using arrival request data from the remaining 8 days not used in the parameter estimation. It can be seen

Figure 3-7: Performance simulated with YouTube data.

from Figure 3-7 that DRES outperforms the N and F strategies over a wide range of β values. Our simulations suggest that since current request allocation and service capacity strategies implemented in CDNs are similar to one of the N or F strategies, they are poorly equipped to handle multiple traffic modes, with either an upsurge in service delay during flash crowd traffic mode or an under-utilization of resources in normal traffic mode.

3.4.2 Simulated arrival requests

In this subsection, we simulate various arrival traffic characteristics, which cannot be tested using the YouTube data set, in order to verify the performance of our proposed algorithm. In each simulation run, we consider an interval of 1440 minutes, and randomly generate two change points that are $1440(1 - \pi)$ apart. The arrival traffic in between the two change points correspond to flash crowd traffic, and is generated using a Poisson process with rate α_F . The arrival traffic in the rest of the interval is in normal mode, and is generated using a Poisson process with rate α_N .

We let $\beta = 25$, and use the same parameters $\pi = 0.497$ and $\alpha_N = 14.33$ estimated from the YouTube data set in our simulations, while α_F is varied to simulate different arrival rate ratios α_F/α_N . In Figure 3-8, we show the performance when we estimate

Figure 3-8: Performance comparison under different arrival rate ratios.

Figure 3-9: Performance comparison when π changes.

the value of π to be $\hat{\pi} = 0.2, 0.497$ or 0.8 in the DRES strategy. It can be seen that the DRES strategy outperforms the N and F strategies for all arrival rate ratios, even when $\hat{\pi} \neq \pi$.

Next, we fix the arrival rates α_N and α_F to be those estimated from the YouTube data set and let $\beta = 5$, but vary the proportion π of normal mode traffic. It can be seen from Figure 3-9 that DRES again outperforms the N and F strategies. In Figure

Figure 3-10: Allocation proportions for different strategies in the flash crowd mode.

3-10, we compare the normalized allocation rates for the different strategies. We see that DRES spreads the arrival requests more evenly amongst the VMs during flash crowd traffic arrivals. Although the VMs with a higher index have higher routing costs, DRES still redirects more traffic to these VMs than the F strategy because it tries to mitigate the additional penalty incurred when a false alarm occurs, which is ignored by the F strategy. On the other hand, if DRES determines that the arrival is in N mode, it adopts the same rates as the N strategy. In this case, it redirects requests to far fewer VMs than the F strategy, which uses the same rates regardless of the traffic arrival statistics. This allows the DRES strategy to be more energy efficient by turning off non-active VMs during N mode traffic arrivals.

3.5 Summary

In this chapter, we address the problem of dynamically scaling the request allocation rates and service capacities of VMs in CCMN, whose incoming traffic switches between normal and flash crowd modes. Our work can be summarized as follows:

- We develop a systematic framework for dynamic request allocation and service capacity scaling in a CCMN based on quickest detection of changes in the arrival

traffic mode, which switches between normal and flash crowd modes depending on the traffic arrival rate. We consider an elastic cost model in which the service capacity of a VM can be elastically scaled to model a cloud computing environment.

- We derive an optimal traffic mode change detection strategy, the optimal request allocation policy, and the optimal service capacity scaling policy under each traffic mode by introducing the concept of a VM price for each VM. The price of a VM depends only on the service cost function of the VM itself, and determines the desirability of the VM to be allocated requests. We also show that there exists a threshold price, which depends on the traffic arrival rate and the VM service costs. We interpret this threshold price as the maximum price the dispatcher is willing to pay for any VM. We show that the optimal VM selection policy consists of choosing only those VMs with a VM price less than the threshold price.
- We provide simulations that suggest that our proposed dynamic allocation and service capacity scaling mechanism outperforms other existing allocation methods when the arrival request traffic has both normal and flash crowd modes. We test all strategies on a real world YouTube request data set, and we also use extensive simulations to verify their performances when various arrival traffic characteristics are varied.

3.6 Proofs

3.6.1 Proof of Proposition 3.1

The optimal stopping time exists because one can construct a trivial stopping time to make the objective function finite. Suppose that the optimal stopping time is T' , with $\theta(T') = \gamma$. Since $L(z)$ has no atoms under Q_0 , we can find a threshold σ so that Page's CUSUM test (3.5) satisfies $\theta(T_\sigma) = \gamma$, and minimizes $\delta(T)$ over all

stopping times with the same false alarm constraint. We then have $\delta(T_\sigma) \leq \delta(T')$, which implies that an optimal solution to (3.6) is given by (3.5).

The approximations (3.7)-(3.8) follow from Theorem 6.5 of [109] and Wald's approximations for the expected stopping time value under Q_0 and Q_1 , which can be found in Chapter 4 of [109]. Finally, substituting these approximations in (3.6) and taking derivatives w.r.t. σ , we obtain (3.9). The proof is now complete.

3.6.2 Proof of Lemma 3.1

Since $\tilde{\mu}_i$ is the minimizer of the right hand side (R.H.S.) in (3.16), there exists a subgradient d such that

$$\frac{1}{\tilde{\mu}_i} = \beta \tilde{\mu}_i d. \quad (3.23)$$

Let $\mu = g_i(p_i)$. From (3.17), we have

$$\beta(w_i(\mu) + \mu \partial^- w_i(\mu)) \leq \frac{1}{\tilde{\mu}_i} + \beta w_i(\tilde{\mu}_i),$$

which together with (3.23) yields

$$w_i(\mu) - w_i(\tilde{\mu}_i) \leq \tilde{\mu}_i d - \mu \partial^- w_i(\mu). \quad (3.24)$$

Suppose that $\mu > \tilde{\mu}_i$. Since $w_i(\cdot)$ is non-decreasing, we have from (3.24) that $\tilde{\mu}_i d \geq \mu \partial^- w_i(\mu)$. This is a contradiction since $\partial^- w_i(\mu) \geq d$ as $w_i(\cdot)$ is convex. Therefore, we must have $\mu \leq \tilde{\mu}_i$. Similarly, from (3.17) and the convexity of $w_i(\cdot)$, we have

$$\beta(w_i(\mu) + \mu \partial^+ w_i(\mu)) \geq \frac{1}{\tilde{\mu}_i} + \beta w_i(\tilde{\mu}_i),$$

and the same argument as above implies that $\mu \geq \tilde{\mu}_i$. The lemma is now proved.

3.6.3 Proof of Theorem 3.1

The Lagrangian for the convex optimization problem (3.15) is

$$L = \sum_{i=1}^n \frac{\lambda_i}{\alpha} \frac{1}{\mu_i - \lambda_i} + \beta \sum_{i=1}^n \frac{\lambda_i}{\alpha} w_i(\mu_i) - p \sum_{i=1}^n \left(\frac{\lambda_i}{\alpha} - 1 \right) - \sum_{i=1}^n b_i \lambda_i - \sum_{i=1}^n h_i (\mu_i - \lambda_i),$$

where p , b_i , h_i are Lagrange multipliers, with $b_i \geq 0$, $h_i \geq 0$ for $i = 1, \dots, n$. For each $i = 1, \dots, n$, we obtain from $\partial L / \partial \lambda_i = 0$,

$$\frac{\mu_i^*}{(\mu_i^* - \lambda_i^*)^2} = p + \alpha(b_i - h_i) - \beta w_i(\mu_i^*), \quad (3.25)$$

and the Karush Kuhn Tucker (KKT) conditions yield

$$\lambda_i^* = \left(\mu_i^* - \sqrt{\frac{\mu_i^*}{p + \alpha b_i - \beta w_i(\mu_i^*)}} \right)^+, \quad (3.26)$$

where we have set $h_i = 0$.

Since $w_i(\mu_i)$ is a convex function, the KKT conditions for subdifferentiable functions [120] give

$$\begin{cases} \lambda_i^* \left(\frac{1}{(\mu_i^* - \lambda_i^*)^2} - \beta \partial^+ w_i(\mu_i^*) \right) \leq 0 \\ \lambda_i^* \left(\frac{1}{(\mu_i^* - \lambda_i^*)^2} - \beta \partial^- w_i(\mu_i^*) \right) \geq 0. \end{cases} \quad (3.27)$$

If $\lambda_i^* > 0$, we have $b_i = 0$, and from (3.26), we obtain

$$\frac{1}{(\mu_i^* - \lambda_i^*)^2} = \frac{p - \beta w_i(\mu_i^*)}{\mu_i^*}. \quad (3.28)$$

Substituting (3.28) into (3.27), we obtain

$$\beta(w_i(\mu_i^*) + \mu_i^* \partial^- w_i(\mu_i^*)) \leq p \leq \beta(w_i(\mu_i^*) + \mu_i^* \partial^+ w_i(\mu_i^*)), \quad (3.29)$$

which implies that $\mu_i^* = g_i(p)$. Observe from (3.15) that if $\lambda_i^* = 0$ for some i , then the optimal μ_i^* can be chosen to be any non-negative value without changing the objective

function value. This implies that there are an infinite number of optimal solutions.³ We can still take $\mu_i^* = g_i(p)$, let $h_i = 0$ and choose b_i appropriately to satisfy the KKT conditions.

We now show that VM S_i is active only if $p > p_i$, and is inactive only if either $g_i(p) = 0$ or $p \leq p_i$. The necessary condition for S_i to be active comes from (3.26), which implies that if $\lambda_i^* > 0$, we have $b_i = 0$ and

$$p > \frac{1}{\mu_i^*} + \beta w_i(\mu_i^*) \geq p_i.$$

Now suppose that $\lambda_i^* = 0$. From (3.26) and (3.29), we have

$$p + b_i \leq \frac{1}{g_i(p)} + \beta w_i(g_i(p)) \quad (3.30)$$

$$\leq p - g_i(p) \left(\beta \partial^- w_i(g_i(p)) - \frac{1}{g_i(p)^2} \right). \quad (3.31)$$

where the last equality follows from (3.17). Since $b_i \geq 0$, we have either $g_i(p) = 0$ or $\beta \partial^- w_i(g_i(p)) - 1/g_i(p)^2 \leq 0$. The second condition is equivalent to $g_i(p) \leq \tilde{\mu}_i$, where $\tilde{\mu}_i$ is the unique minimizer of the R.H.S. in (3.16). From Lemma 3.1, we have $g_i(p) \leq g_i(p_i)$, and since $g_i(\cdot)$ is a non-decreasing function, we obtain $p \leq p_i$. This implies that in the optimal policy, the VMs are chosen in non-decreasing order of p_i . The number of active VMs needs to meet the constraint $\sum_{i=1}^n \lambda_i^* = \alpha$, hence (3.20) holds, and the theorem is proved.

³The only physically reasonable solution however corresponds to choosing $\mu_i^* = 0$ when $\lambda^* = 0$.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Cross-Layer Resource Allocation with Elastic Service Scaling in C-RAN

In Chapter 3, we discuss the elastic service scaling problem in the cloud-based core network. In Chapters 4 and 5, we investigate the problem in cloud radio access network. Specifically, in this chapter we use a simple C-RAN model, i.e., each user equipment (UE) only associates with one VM in the BBU pool and the number of UEs that each RRH can serve is unlimited. While in Chapter 5, C-RAN model is extended to that each UE is allowed to associate with multiple VMs and each RRH can serve only a limited number of UEs.

After solving the elastic service scaling problem in core network in Chapter 3, we shift our focus to the RAN side. Although C-RAN brings in many significant features, there are two main challenges. The first one is how to support the huge amount of data transfer in the fronthaul, whose capacity is actually limited in practice [66], by dynamically adjusting the beamforming vectors and the active RRH set or user association set. The second one is, as all the computation resources are migrated from the BSs into a centralized BBU pool, how to effectively manage and dispatch the computation resources in C-RAN. In this dissertation, we look at these problems in two scenarios:

- Firstly, we assume each UE can only associate with one VM in the BBU pool, and each active RRH can transmit data to all UEs. However, under this model, the active RRH set is a variable to be determined. This scenarios is studied in this chapter.
- Secondly, we assume that the active RRH set is fixed, but each active RRH can only serve a limited number of UEs, hence the user association set is a variable. In the BBU pool, we assume each UE can associate with multiple VMs, while the number of active VMs is a variable to be optimized. This scenarios is investigated in next chapter, i.e, Chapter 5.

Although C-RAN makes it possible to transfer CCNs from hardware defined infrastructures to a software defined environment, many design and operational challenges that have been resolved in CCNs need to be revisited in C-RAN. One particular example of importance is the resource allocation problem. Specifically, in CCNs, power control and beamforming strategies have been used to minimize the system power consumption such that users' predefined QoS requirements are fulfilled. Unfortunately, these strategies cannot plug directly into the C-RAN framework. In CCNs, the BSs' computation capacity is fixed. As a result, resource allocation methods in CCNs are oblivious to the computation capacities of the BSs although users' achievable QoS levels are actually dependent on them. Under C-RAN architecture, the computational functionalities in conventional BSs are migrated to the cloud based VMs in BBU pool, whose computation capacity can be scaled according to users' QoS requirements and various parameters from different layers of the open systems interconnection (OSI) stack, including the incoming traffic rate from the application layer and wireless channel state information from the physical layer. Therefore, developing a cross-layer resource allocation scheme is required in order to fully utilize the features of a C-RAN, and to optimize the overall system power consumption.

Table 4.1: Summary of common notations in Chapter 4 and 5.

Symbol	Definition
a_i	expected delay in the BBU pool for the data to user equipment (UE) i
b_i	expected delay during wireless transmission for the data to UE i
B_i	wireless transmission bandwidth for UE i
c_i	achievable rate for UE i during wireless transmission
d_i	expected system delay for data to UE i
e_i	mean square estimation error for the signal to UE i
E_j	maximum transmitting power of RRH j
$\mathbf{h}_{ij}^H$	the channel from RRH j to UE i
K	the number of antennas in each RRH
L	the number of all RRHs
N	the number of UEs
u_i	the data symbol to UE i
$\mathbf{w}_{ij}$	transmit beamformer for UE i from RRH j
x_i	mean square error (MSE) weight for UE i
y_i	receive beamformer for the signal to UE i
λ_i	mean arrival data rate to UE i
τ_i	system delay requirement for user i
δ_i	additive white Gaussian noise at UE i , and $\delta_i \sim \mathcal{CN}(0, \sigma_i^2)$
$\mathcal{L}$	the set of all RRHs
$\mathcal{N}$	the set of all UEs

4.1 System model

In this section, we present our C-RAN system model and problem formulation. We summarize some commonly used notations in Chapters 4 and 5 in Table 4.1.

Suppose that there are N single-antenna UEs and L available RRHs, each with K antennas, in a C-RAN cluster. We denote the sets of all UEs and all RRHs as $\mathcal{N} = \{1, \dots, N\}$ and $\mathcal{L} = \{1, \dots, L\}$, respectively. We denote the set of active RRHs (i.e., the set of RRHs that are servicing the UEs in $\mathcal{N}$, and their associated fiber links) as $\mathcal{A}$. We have $\mathcal{A} \subseteq \mathcal{L}$. The amount of voice and data traffic associated with each UE $i \in \mathcal{N}$ up to time t is given by $\Delta_i(t)$ (bits), and each UE i is served by one VM with computation capacity μ_i in the BBU pool. After processing by the VM, the data is forwarded to the UE via $|\mathcal{A}|$ active RRHs (we assume data sharing among the RRHs), where $|\mathcal{A}| \leq L$ is the cardinality of the set $\mathcal{A}$. Let the achievable wireless

Figure 4-1: Queueing model representation of a C-RAN processing and transmission path for a UE i .

transmission rate to UE i using the active RRHs be c_i .

Queueing model, with the channel capacity as the queue's service rate, is widely used to characterize wireless communication systems [121]. We introduce a double-layer queueing network to represent each UE's data processing and transmitting behavior in the C-RAN downlink (cf. Figure 4-1). Our model can be easily extended to the C-RAN uplink as well. Specifically, in the BBU pool, the data of UE i is processed (e.g., encoded) by a VM, which is abstracted as a processing queue, with mean service rate μ_i . Then, the processed data is transmitted to UE i via the RRHs, which are modeled using a transmitting queue with mean service rate c_i . Note that the links between the BBU pool and the RRHs are high-bandwidth, low-latency optical fiber links with negligible transmission delay. However, the power consumption P_f of each fiber link cannot be neglected, compared with the power consumption in the associated RRH.

For each UE $i \in \mathcal{N}$, let a_i represent the expected delay in the processing queue (i.e., the expected delay in the BBU pool) and b_i be the expected delay in the transmitting queue (i.e., the expected delay during wireless transmission). Our goal is to design a cross-layer algorithm such that for each UE i , the system expected delay $d_i = a_i + b_i$ satisfies the cross-layer QoS constraint:

$$d_i \leq \tau_i, \quad (4.1)$$

where τ_i is a predefined QoS requirement for UE i .

We assume that UE i 's packet arrival process to the processing queue is a Poisson process with mean rate λ_i , where $\Delta_i(t) = \lambda_i t$, and the service time of each data

packet in the processing queue follows an exponential distribution with mean $1/\mu_i$. Then, the arrival process to the transmitting queue is the same as the one to the processing queue [122, 123]. We assume that the service time of each data packet in the transmitting queue follows an exponential distribution with mean $1/c_i$. Therefore, the data processing and transmitting for each UE i in our C-RAN model can be treated as two M/M/1 queues [115] in tandem ¹. We have for $\mu_i, c_i > \lambda_i$,

$$d_i = \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i}.$$

In the wireless transmission, C-RAN leverages CoMP transmission to enhance the throughput [124]. There are two types of CoMP techniques in the downlink: coordinated scheduling/coordinated beamforming (CS/CB) and joint transmission (JT). In this work, we consider JT as the CoMP technique in C-RAN, i.e., each UE's data can be shared among all the coordinated RRHs. Let x_i denote the data symbol for the i th UE with $E[|x_i|^2] = 1$, and $\mathbf{w}_{ij} \in \mathbb{C}^K$ denote the transmit beamformer for UE i from RRH j . We assume the channel does not change over the duration of transmission, i.e., block fading channel model. The channel from RRH j to UE i is denoted as $\mathbf{h}_{ij}^H$, where $\mathbf{h}_{ij} \in \mathbb{C}^K$, for $i \in \mathcal{N}$ and $j \in \mathcal{A}$. Thus, the received signal at UE i is given by

$$\hat{x}_i = \sum_{j \in \mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} x_i + \sum_{k \neq i} \sum_{j \in \mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{kj} x_k + \delta_i,$$

where the first term on the right hand side is the useful signal for UE i , the second term is the interference to UE i , and $\delta_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive white Gaussian noise (AWGN) at UE i .

As such, the signal-to-interference-plus-noise ratio (SINR) at UE i , with the active RRH set $\mathcal{A}$, becomes

$$\text{SINR}_i(\mathcal{A}) = \frac{|\sum_{j \in \mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}|^2}{\sigma_i^2 + \sum_{k \neq i} |\sum_{j \in \mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}|^2}. \quad (4.2)$$

¹Note that, we do not consider packet size changes caused by baseband processing.

The achievable rate of UE i , c_i , should satisfy ²

$$c_i \leq B_i \log(1 + \text{SINR}_i(\mathcal{A})), \quad (4.3)$$

where B_i is the bandwidth for UE i . Each RRH j has maximum transmitting power constraint given by

$$\sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} = \sum_{i=1}^N \|\mathbf{w}_{ij}\|_2^2 \leq E_j, \text{ for } j \in \mathcal{L}. \quad (4.4)$$

4.1.1 Problem formulation

The BBU pool of C-RAN utilizes a cloud computing infrastructure with elastic service scaling. In particular, the BBU pool can dynamically adjust the VMs' computation capacities to handle dynamic user traffic and channel states. We model VM i 's power consumption $\varphi_i(\mu_i)$ as a function of its computation capacity μ_i . We make the following assumptions regarding the VM's power consumption function φ_i .

Assumption 4.1. For each VM i , $i \in \mathcal{N}$, the power consumption function $\varphi_i(\mu_i)$ has the following properties:

- 1) $\varphi_i(\mu_i) \geq 0$ for all $\mu_i \geq 0$,
- 2) $\varphi_i(\mu_i)$ is a convex and increasing function of μ_i .

We can see that Assumption 4.1 is consistent with Assumption 3.1 in Chapter 3. The power consumption of a VM i is often modeled as $\varphi_i(\mu_i) = k_i \mu_i^{a_i}$, where $k_i > 0$ and $a_i > 1$ are positive constants. This power consumption function satisfies Assumption 4.1, and has been widely adopted in the literature [53, 114, 125, 126].

Our aim is to minimize the system power consumption in C-RAN, which consists of three components: the power consumption in the BBU pool, the power consumption in the fiber links, and the power consumption at the RRHs. Specifically, (i) the power consumption for each VM in the BBU pool with computation capacity μ_i is

²Throughout Chapter 4 and 5, all the log functions are based 2.

$\varphi_i(\mu_i)$, $\forall i \in \mathcal{N}$; (ii) the power consumption for each active fiber link is P_f ; and (iii) the power consumption at RRH $j \in \mathcal{A}$ is $(1/\eta) \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij}$, where $\eta \in (0, 1)$ is the inefficiency coefficient of the amplifier in each RRH. Our optimization problem can then be formulated as follows:

$$\begin{aligned} \text{(P0)} \quad & \min_{\mu_i, c_i, \mathbf{w}_{ij}, \mathcal{A}} \quad \sum_{i=1}^N \varphi_i(\mu_i) + |\mathcal{A}|P_f + \frac{1}{\eta} \sum_{i=1}^N \sum_{j \in \mathcal{A}} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\ \text{s.t.} \quad & \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \quad \forall i \in \mathcal{N}, \end{aligned} \quad (4.5)$$

$$\lambda_i < \mu_i, \lambda_i < c_i, \quad \forall i \in \mathcal{N}, \quad (4.6)$$

$$c_i \leq B_i \log(1 + \text{SINR}_i(\mathcal{A})), \quad \forall i \in \mathcal{N}, \quad (4.7)$$

$$\sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{A}, \quad (4.8)$$

where $\text{SINR}_i(\mathcal{A})$ is given by (4.2).

Problem (P0) is difficult to solve for the following reasons: (i) it is a combinatorial optimization problem and NP-hard [70]; and (ii) the problem is nonconvex even if the active RRH set $\mathcal{A}$ is known a priori. However, by relaxing (P0) into a quasi weighted sum-rate maximization (QWSRM) problem, we obtain a BnB solution. In the following section, we first discuss the QWSRM problem and its BnB solution.

4.2 The QWSRM problem

In this section, we extend the classical weighted sum-rate maximization (WSRM) problem to a QWSRM problem, whose solution will help us in tackling (P0). Throughout this section, we assume that the active RRH set $\mathcal{A}$ is known.

Mathematically, the WSRM problem is typically formulated as

$$\begin{aligned} \min_{c_i, \mathbf{w}_{ij}} \quad & \sum_{i=1}^N -\varepsilon_i c_i \end{aligned} \quad (4.9)$$

$$\text{s.t.} \quad c_i \leq B_i \log(1 + \text{SINR}_i(\mathcal{A})), \quad \forall i \in \mathcal{N}, \quad (4.10)$$

$$\sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{A}, \quad (4.11)$$

where c_i is the throughput of UE i , ε_i is an arbitrary nonnegative weight, and $\text{SINR}_i(\mathcal{A})$ is given by (4.2).

Since the phase rotation of the complex vector $\mathbf{w}_{ij}$ has no impact on the WSRM problem, we can recast constraint (4.10) as

$$\|\mathbf{r}_i(\mathcal{A})\|_2 \leq \sqrt{1 + 1/(2^{c_i/B_i} - 1) \Re[R_{ii}(\mathcal{A})]}, \quad \forall i \in \mathcal{N}, \quad (4.12)$$

where vector $\mathbf{r}_i(\mathcal{A}) = [R_{i1}(\mathcal{A}), \dots, R_{iN}(\mathcal{A}), \sigma_i]^T$, $R_{ik}(\mathcal{A}) = \sum_{j \in \mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}$, and $\Re(\cdot)$ stands for the real part of a complex number [127, 128]. Note that constraint (4.12) is a second-order cone (SOC) constraint only if c_i is a constant.

Applying the Cauchy-Schwarz inequality to (4.7), we have

$$\begin{aligned} c_i &\leq B_i \log \left(1 + \frac{1}{\sigma_i^2} \sum_{j \in \mathcal{A}} \|\mathbf{h}_{ij}\|_2^2 \sum_{j \in \mathcal{A}} \|\mathbf{w}_{ij}\|_2^2 \right) \\ &\leq B_i \log \left(1 + \frac{1}{\sigma_i^2} \sum_{j \in \mathcal{A}} \|\mathbf{h}_{ij}\|_2^2 E_j \right) \triangleq \bar{c}_i. \end{aligned} \quad (4.13)$$

Let $\bar{\mathbf{c}} = [\bar{c}_1, \dots, \bar{c}_N]^T$. We define a generalization of the WSRM problem, which has the same constraints as the WSRM problem but with an extended objective function as follows:

$$\begin{aligned} \min_{\mathbf{c}, \mathbf{w}_{ij}} \quad & f(\mathbf{c}) \\ \text{s.t.} \quad & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{A}, \\ & \|\mathbf{r}_i(\mathcal{A})\|_2 \leq \sqrt{1 + 1/(2^{c_i/B_i} - 1) \Re[R_{ii}(\mathcal{A})]}, \quad \forall i \in \mathcal{N}, \end{aligned} \quad (4.14)$$

where $\mathbf{c} = [c_1, \dots, c_N]^T$, and the objective function $f(\mathbf{c})$, for $0 \leq \mathbf{c} \leq \bar{\mathbf{c}}$, has the following properties:

- 1) $f(\mathbf{c})$ is a function only of $\mathbf{c}$ and

2) $f(\mathbf{c}) < \infty$ is continuously differentiable, and

3) $f(\mathbf{c})$ is convex in the feasible region defined by (4.11) and (4.12).

To avoid trivial solutions for (4.14), in this chapter, we assume $\frac{\partial f(\mathbf{c})}{\partial c_i} \big|_{\mathbf{c}=0} < 0$, for $i = 1 \cdots N$. We call (4.14) the QWSRM problem, which models a variety of problems in wireless communications. For example, we can interpret $f(\mathbf{c})$ as the *reverse utility* function, corresponding to the concave utility function in network congestion control problems [129, 130].

Let $\tilde{\mathbf{c}} = [\tilde{c}_1, \dots, \tilde{c}_N]^T$ be the root to the system of equations $\partial f(\mathbf{c})/\partial c_i = 0$, for all $i \in \mathcal{N}$, where each $\tilde{c}_i$ is set to $\bar{c}_i$ if the solution does not exist. Let $\mathcal{F}$ represent the feasible region of the variable $\mathbf{c}$ in (4.14), and $\mathbf{c}^* = [c_1^*, \dots, c_N^*]^T$ and $\mathbf{w}_{ij}^*$, $\forall i \in \mathcal{N}, \forall j \in \mathcal{A}$, be the optimal solution of the QWSRM problem.

Theorem 4.1. The optimal achievable rate $\mathbf{c}^*$ of the QWSRM problem falls inside or on the surface of the N -dimensional rectangle $\mathcal{Q}_{\text{init}} = \{\mathbf{c} \mid c_i \in [0, \min\{\tilde{c}_i, \bar{c}_i\}], i \in \mathcal{N}\}$.

Proof. See Section 4.6.1. □

A weighted minimum mean square error (WMMSE) approach to solve the WSRM problem based on the relationship between SINR and MMSE is proposed in [97]. However, the WMMSE approach cannot always find the global optimal solution. Subsequently, a BnB method is proposed in [96], which shows that this method can produce the global optimal solution. The proposed BnB method uses the fact that the objective function in WSRM problem is monotonically non-increasing in the achievable rates $c_i \geq 0$, for all $i \in \mathcal{N}$. In what follows, we first give a brief overview of the BnB algorithm from [96], and then show how to extend it to solve the QWSRM problem in (4.14).

The BnB approach is widely used in nonconvex optimization problems, e.g., the integer programming problems. For each iteration step of the BnB algorithm, one needs to generate a sequence of asymptotically tight upper and lower bounds for the objective function, with both bounds converging to the global optimal value eventually. The basic idea in [96] of using the BnB algorithm to solve the WSRM

problem is to first expand the unknown feasible region of the WSRM problem to a known initial N -dimensional rectangle, and then sequentially shrink the rectangle until it is small enough, where at each iteration, the variables $\mathbf{c}$ are fixed, and a feasibility problem w.r.t. to the variables $\{\mathbf{w}_{ij} : i \in \mathcal{N}, j \in \mathcal{A}\}$ is solved. This avoids solving the nonconvex WSRM problem w.r.t. $\{c_i, \mathbf{w}_{ij}\}$ directly.

Inspired by the BnB algorithm in [96], we develop a similar BnB procedure in Algorithm 4.1 for the QWSRM problem (4.14). We use $\mathcal{Q}_{\text{init}}$ given in Theorem 4.1 as the initial N -dimensional rectangle. We shrink the N -dimensional rectangle by making use of the following upper bound

$$\gamma_{\text{ub}}(\mathcal{Q}) = \begin{cases} f(\mathbf{c}_{\min}), & \mathbf{c}_{\min} \in \mathcal{F} \\ +\infty, & \text{otherwise} \end{cases} \quad (4.15)$$

and lower bound ³

$$\gamma_{\text{lb}}(\mathcal{Q}) = \begin{cases} f(\mathbf{c}_{\max}), & \mathbf{c}_{\max} \in \mathcal{F} \\ +\infty, & \text{otherwise,} \end{cases} \quad (4.16)$$

for every N -dimensional rectangle $\mathcal{Q} \triangleq \{\mathbf{c} \mid c_{i,\min} \leq c_i \leq c_{i,\max}, \forall i \in \mathcal{N}\} \subseteq \mathcal{Q}_{\text{init}}$, where $c_{i,\min}$ and $c_{i,\max}$ denotes the end points of the i th edge of $\mathcal{Q}$, $\mathbf{c}_{\min} = [c_{1,\min}, \dots, c_{N,\min}]^T$ and $\mathbf{c}_{\max} = [c_{1,\max}, \dots, c_{N,\max}]^T$. Note that $\mathbf{c}_{\max}$ need not be in the feasible region $\mathcal{F}$. At each iteration, for a given $\mathcal{Q}$, the following feasibility problem is solved:

$$\begin{aligned} \text{find} \quad & \mathbf{w}_{ij}, \forall i \in \mathcal{N}, \forall j \in \mathcal{A} \\ \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{A})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_{i,\min}/B_i} - 1}} \Re[R_{ii}(\mathcal{A})], \forall i \in \mathcal{N}, \\ & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \forall j \in \mathcal{A}. \end{aligned} \quad (4.17)$$

Note that (4.17) is a second-order cone programming (SOCP) feasibility problem w.r.t. $\mathbf{w}_{ij}$, which can be solved by using interior-point methods on an equivalent

³Although [96] provides an improved lower bound with additional computational overhead, we use the basic lower bound in this chapter for simplicity.

SOCP with a trivial objective function [131].

Algorithm 4.1 BnB algorithm for QWSRM problem

```

1: Input:  $\mathcal{Q}_{\text{init}}$ ,  $\mathcal{A}$ , and  $f(\mathbf{c})$ .
2: Initialize: Obtain  $\tilde{c}_i$  by solving  $\frac{\partial f(\mathbf{c})}{\partial c_i} = 0$ , for  $i \in \mathcal{N}$ . Set  $k = 1$ ,  $\mathcal{B}_1 = \mathcal{Q}_{\text{init}}$ ,
    $u_1 = \gamma_{\text{ub}}(\mathcal{Q}_{\text{init}})$  and  $l_1 = \gamma_{\text{lb}}(\mathcal{Q}_{\text{init}})$ .
3: Check the feasibility of problem (4.17) with given  $\tilde{\mathbf{c}}$ .
4: if feasible then
5:    $\mathbf{c}_o = \tilde{\mathbf{c}}$ ;
6: else
7:   while  $u_k - l_k > \epsilon$  do
8:     Branching:
       • Set  $\mathcal{Q}_k = \mathcal{Q}$ , where  $\mathcal{Q}$  satisfies  $\gamma_{\text{lb}}(\mathcal{Q}) = l_k$ .
       • Split  $\mathcal{Q}$  into  $\mathcal{Q}_I$  and  $\mathcal{Q}_{II}$ , along one of its longest edges.
       • Update  $\mathcal{B}_{k+1} = (\mathcal{B}_k \setminus \{\mathcal{Q}_k\}) \cup (\mathcal{Q}_I, \mathcal{Q}_{II})$ .
9:     Bounding:
       • Update  $u_{k+1} = \min_{\mathcal{Q} \in \mathcal{B}_{k+1}} \{\gamma_{\text{ub}}(\mathcal{Q})\}$ .
       • Update  $l_{k+1} = \min_{\mathcal{Q} \in \mathcal{B}_{k+1}} \{\gamma_{\text{lb}}(\mathcal{Q})\}$ .
10:    Set  $k = k + 1$ ;
11:   end while
12:   Set  $\mathbf{c}_o = \mathbf{c}_{\text{min}}$ ;
13: end if
14: Output:  $\mathbf{c}_o$ .
```

The rationale of using the BnB algorithm to solve the QWSRM problem is the same as that for the WSRM problem, i.e., we sequentially shrink the given N -dimensional rectangle $\mathcal{Q}_{\text{init}}$, where the optimal solution falls in, until the the lower and upper bounds satisfy $u_k - l_k \leq \epsilon$, where $\epsilon > 0$ is a predefined accuracy level. We show an intuitive explanation for Algorithm 4.1 in Figure 4-2, where the rectangle contains the optimal point, is shrunk step by step until the predefined accuracy level satisfies.

The following result shows that Algorithm 4.1 converges to the optimal solution of the QWSRM. The proof is similar to Theorem 1 in [96] and the convergence analysis in [132], which we omit for brevity.

Figure 4-2: An intuitive explanation for the BnB algorithm.

Theorem 4.2. The output $\mathbf{c}_o$ generated by Algorithm 4.1, converges arbitrarily close to the optimal solution $\mathbf{c}^*$ of the QWSRM problem, within a finite number of iterations, i.e., for any $\epsilon > 0$, there exists $M > 0$ such that $u_M - f(\mathbf{c}^*) \leq \epsilon$.

Remark 4.1. The reason that the upper and lower bounds in (4.15) and (4.16) are suitable for the QWSRM problem is that $f(\mathbf{c})$ is monotonic in each interval $c_i \in [0, \tilde{c}_i]$, for all $i \in \mathcal{N}$. Thus, by fixing the variable $\mathbf{c}$ in the QWSRM problem, instead of solving the nonconvex problem (4.14) directly, we just need to solve a SOCP (4.17) in each iteration of the BnB algorithm. For Algorithm 4.1, the input N -dimensional rectangle $\mathcal{Q}_{\text{init}}$ provided by Theorem 4.1 can be further shrunk if a priori upper and lower bounds of $\mathbf{c}$ are known.

4.3 Cross-layer power consumption minimization

In this section, we reformulate (P0) into a MINLP, which we further decompose into a QWSRM problem and a RRH selection problem. We introduce an auxiliary binary variable $\beta_j \in \{0, 1\}$, $\forall j \in \mathcal{L}$, where $\beta_j = 1$ if and only if RRH j is active (the fiber link j is turned on). In the case of $\beta_j = 0$, i.e., RRH j is inactive, there is no signal

transmitted from RRH j to all the UEs. Hence, RRH j is active or not is equivalent to $\sum_{i=1}^N \|\mathbf{w}_{ij}\|_2^2 > 0$ or $= 0$, respectively. In addition, for $i \in \mathcal{N}$,

$$\sum_{j \in \mathcal{A}} \mathbf{w}_{ij}^H \mathbf{w}_{ij} = \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij}, \quad (4.18)$$

since $\sum_{j \in \mathcal{A}^c} \mathbf{w}_{ij}^H \mathbf{w}_{ij} = 0$, where $\mathcal{A}^c$ is the complementary set of $\mathcal{A}$.

Let the l_0 -norm of a vector $\beta = [\beta_1, \dots, \beta_L]^T$ be denoted by $\|\beta\|_0$, i.e., $\|\beta\|_0 = |\{j : \beta_j \neq 0\}|$, where β_j is the j th element of vector β . Let the vector $\mathbf{m} = [\sum_{i=1}^N \|w_{i1}\|_2^2, \sum_{i=1}^N \|w_{i2}\|_2^2, \dots, \sum_{i=1}^N \|w_{iL}\|_2^2]^T$. Hence, $|\mathcal{A}| = \sum_{j=1}^L \beta_j = \|\beta\|_0 = \|\mathbf{m}\|_0$. Combining (4.18), (4.12) and $|\mathcal{A}| = \|\mathbf{m}\|_0$, we can reformulate problem (P0) as

$$\begin{aligned} \text{(P1)} \quad & \min_{\mu_i, c_i, \beta_j, \mathbf{w}_{ij}} \sum_{i=1}^N \varphi_i(\mu_i) + \|\mathbf{m}\|_0 P_f + \frac{1}{\eta} \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\ \text{s.t.} \quad & \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \quad \forall i \in \mathcal{N}, \\ & \lambda_i < \mu_i, \lambda_i < c_i, \quad \forall i \in \mathcal{N}, \\ & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1} \Re \mathbf{e}[R_{ii}(\mathcal{L})]}, \quad \forall i \in \mathcal{N}, \end{aligned} \quad (4.19)$$

$$\sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq \beta_j E_j, \quad \forall j \in \mathcal{L}, \quad (4.20)$$

$$\beta_j \in \{0, 1\}, \quad \forall j \in \mathcal{L}, \quad (4.21)$$

where (4.19) is derived from (4.12) due to the fact that $\mathbf{w}_{ij} = 0$, for $j \in \mathcal{A}^c$.

Proposition 4.1. In problem (P1), constraint (4.5) is the active inequality constraint, i.e., the optimal $\{\mu_i, c_i\}$ for problem (P1) always satisfies the following equation:

$$\begin{aligned} & \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i} = \tau_i, \quad \forall i \in \mathcal{N}, \quad \text{or,} \\ & \mu_i = \lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2(c_i - \lambda_i) - \tau_i}, \quad \forall i \in \mathcal{N}. \end{aligned}$$

Proof. See Section 4.6.2. □

Figure 4-3: The two-step approach to solve problem (P1).

Thus, based on Proposition 4.1, we define

$$g_i(c_i) \triangleq \varphi_i(\mu_i) = \varphi_i \left(\lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2(c_i - \lambda_i) - \tau_i} \right), \quad (4.22)$$

where

$$c_i > \lambda_i + \frac{1}{\tau_i}, \forall i \in \mathcal{N}. \quad (4.23)$$

In what follows, we propose a two-step approach to solve problem (P1) (cf. Figure 4-3). Specifically,

1. we relax problem (P1) to problem (P2), and for problem (P2), we propose two different algorithms to solve it, i.e., the BnB algorithm in Section 4.3.1 and the WMMSE algorithm in Section 4.3.2 respectively.
2. Then, based on the optimal achievable rate obtained by solving problem (P2), problem (P1) becomes a RRH selection problem (cf. problem (Q2) in Section 4.3.3), we then propose an efficient Shaping-and-Pruning algorithm to obtain the sparse solution of the RRH selection problem.

For problem (P1), which is a MINLP, we can apply the following two relaxation techniques:

- A binary-to-continuous relaxation to the binary variable $\beta_j, \forall j \in \mathcal{L}$, i.e., β_j is relaxed to a continuous variable, taking values in the closed interval $[0, 1]$.
- Inspired by compressive sensing, l_1 -norm can be utilized as a convex relaxation of l_0 -norm since l_1 -norm is a convex envelop of l_0 -norm [70]. We can apply l_1 -norm relaxation to the objective function of problem (P1).

Therefore, applying the two techniques above to problem (P1) and combine Proposition 1, we have the following optimization problem:

$$\min_{c_i, \beta_j, \mathbf{w}_{ij}} \sum_{i=1}^N g_i(c_i) + (P_f + \frac{1}{\eta}) \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \quad (4.24)$$

$$\begin{aligned} \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1}} \Re[R_{ii}(\mathcal{L})], \quad \forall i \in \mathcal{N}, \\ & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq \beta_j E_j, \quad \forall j \in \mathcal{L}, \\ & c_i > \lambda_i + \frac{1}{\tau_i}, \quad \forall i \in \mathcal{N}, \\ & 0 \leq \beta_j \leq 1, \quad \forall j \in \mathcal{L}, \end{aligned} \quad (4.25)$$

which is a nonconvex optimization problem.

We can easily observe that the variable β_j is redundant in problem (4.24) and, hence, can be dropped. Thus, problem (4.24) can be further simplified as:

$$(P2) \quad \min_{c_i, \mathbf{w}_{ij}} \sum_{i=1}^N g_i(c_i) + (P_f + \frac{1}{\eta}) \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \quad (4.26)$$

$$\begin{aligned} \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1}} \Re[R_{ii}(\mathcal{L})], \quad \forall i \in \mathcal{N}, \\ & c_i > \lambda_i + \frac{1}{\tau_i}, \quad \forall i \in \mathcal{N}, \\ & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L}. \end{aligned} \quad (4.27)$$

We call problem (P2) as the *extended sum-utility maximization* (ESUM) problem.

4.3.1 Relaxing the ESUM problem to a QWSRM problem

Similar to (4.13), we apply the Cauchy-Schwarz inequality to (4.7), and combining with (4.18), we have

$$c_i \leq B_i \log \left(1 + \frac{1}{\sigma_i^2} \sum_{j=1}^L \|\mathbf{h}_{ij}\|_2^2 \sum_{j=1}^L \|\mathbf{w}_{ij}\|_2^2 \right), \forall i \in \mathcal{N}, \quad (4.28)$$

which further yields

$$\sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \geq \frac{(2^{c_i/B_i} - 1) \sigma_i^2}{\sum_{j=1}^L \|\mathbf{h}_{ij}\|_2^2}. \quad (4.29)$$

Hence, problem (P2) can be approximated as

$$\begin{aligned} \text{(Q1-1)} \quad & \min_{c_i, \mathbf{w}_{ij}} \quad \sum_{i=1}^N f_i(c_i) \\ \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1} \Re[R_{ii}(\mathcal{L})]}, \quad \forall i \in \mathcal{N}, \\ & c_i > \lambda_i + \frac{1}{\tau_i}, \quad \forall i \in \mathcal{N}, \\ & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L}, \end{aligned}$$

where

$$f_i(c_i) = g_i(c_i) + (P_f + \frac{1}{\eta}) \frac{(2^{c_i/B_i} - 1) \sigma_i^2}{\sum_{j=1}^L \|\mathbf{h}_{ij}\|_2^2}.$$

Proposition 4.2. Suppose that Assumption 4.1 holds. Then, problem (Q1-1) is a QWSRM problem, whose optimal solution $\mathbf{c}^*$ falls inside or on the surface of the N -dimensional rectangle $\hat{\mathcal{Q}}_{\text{init}} = \{\mathbf{c} = [c_1, \dots, c_N]^T \mid c_i \in (\lambda_i + 1/\tau_i, \min\{\tilde{c}_i, \bar{c}_i\}], \text{ for } i \in \mathcal{N}\}$, where $\tilde{c}_i$ is the root of the equation $\partial f_i(c_i)/\partial c_i = 0$, and $\bar{c}_i = B_i \log \left(1 + \frac{1}{\sigma_i^2} \sum_{j=1}^L \|\mathbf{h}_{ij}\|_2^2 E_j \right)$.

Proof. See Section 4.6.3. □

To obtain the optimal achievable rates $\mathbf{c}^* = [c_1^*, \dots, c_N^*]^T$ for (Q1-1), we utilize

Algorithm 4.1 with the following inputs:

- 1) $f(\mathbf{c}) = \sum_{i=1}^N f_i(c_i)$, where $\mathbf{c} = [c_1, \dots, c_N]^T$
- 2) $\mathcal{Q}_{\text{init}} = \hat{\mathcal{Q}}_{\text{init}}$
- 3) $\mathcal{A} = \mathcal{L}$

The optimal VM computation capacity for UE i as implied by (Q1-1) is then given by $\mu_i^* = \lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2(c_i^* - \lambda_i) - \tau_i}$. Note that the solution $\{(\mu_i^*, c_i^*) \mid i \in \mathcal{N}\}$ is in general sub-optimal for (P1) because of the relaxations we have done to obtain (Q1-1), but is guaranteed to be feasible for (P1).

4.3.2 A WMMSE approach to solve the ESUM problem

Although problem (P2) can be approximated as a QWSRM problem (Q1-1) and then be solved by the BnB algorithm, which obtains the global optimal solution for problem (Q1-1), the complexity of the BnB algorithm is still high. In this subsection, we aim to develop a lower complexity algorithm to obtain the local optimal solution for problem (P2) directly.

Based on the results in [97], which aims to solve the sum-utility maximization problem by the WMMSE algorithm, we extend the algorithm to solve our ESUM problem.

Let's define $\theta_i(\cdot) = g_i(-B_i \log(\cdot))$ and denote $\tilde{\theta}_i(\cdot)$ as the inverse mapping of the gradient map $\nabla \theta_i(e_i)$. It can be verified that $\theta_i(e_i)$ is a strictly concave function in the interval $(2^{-\hat{c}_i}, \infty)$, where $\hat{c}_i = (\lambda_i + 1/\tau_i)/B_i$.

Proposition 4.3. The optimal transmit beamformer vectors $\mathbf{w}_{ij}$ for the ESUM problem is identical with the optimal $\mathbf{w}_{ij}$ obtained from the following problem:

$$\begin{aligned}
 \text{(Q1-2)} \quad \min_{x_i, y_i, \mathbf{w}_{ij}} \quad & \sum_{i=1}^N x_i e_i + \sum_{i=1}^N \theta_i(\tilde{\theta}_i(x_i)) - \sum_{i=1}^N x_i \tilde{\theta}_i(x_i) \\
 & + (P_f + \frac{1}{\eta}) \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij}
 \end{aligned}$$

$$\begin{aligned} \text{s.t.} \quad & \|\mathbf{r}_i\|_2 < \sqrt{1 + 1/(2^{\hat{e}_i} - 1)} \Re[R_{ii}], \quad \forall i \in \mathcal{N}, \\ & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L}, \end{aligned}$$

where $x_i > 0$ and y_i are the MSE weight and receive beamformer for user i respectively, e_i is the MSE defined as

$$\begin{aligned} e_i &\triangleq \mathbb{E} [\|y_i^H \hat{u}_i - u_i\|_2] \\ &= \left| y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} - 1 \right|^2 + \sum_{l \neq i}^N \left| y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj} \right|^2 + \sigma_i^2 |y_i|^2, \\ &= \sum_{l=1}^N \left| y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj} \right|^2 - 2 \Re \left[y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} \right] + \sigma_i^2 |y_i|^2 + 1. \end{aligned} \quad (4.30)$$

Moreover, $\theta_i(\tilde{\theta}_i(x_i)) - x_i \tilde{\theta}_i(x_i)$ is strictly convex w.r.t. x_i .

Proof. The proof is similar with Appendix B of [97]. We omit it for brevity. $\square$

Based on Proposition 4.3, instead of solving problem (P2) directly, we can solve problem (Q1-2) to obtain the optimal transmit beamformer vectors $\mathbf{w}_{ij}$. Since problem (Q1-2) is convex w.r.t. each variable while keeping other variables fixed, problem (Q1-2) is much easier to solve than problem (P2). Specifically, problem (Q1-2) can be solved via the following alternating optimization procedure:

- For given $\mathbf{w}_{ij}$, $\forall i \in \mathcal{N}$ and $j \in \mathcal{L}$, the optimal receive beamformer of problem (Q1-2) can be calculated by the well-known MMSE receiver:

$$y_i = \frac{\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}}{\sum_{k \in \mathcal{N}} \left(\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj} \right) \left(\sum_{j \in \mathcal{L}} \mathbf{w}_{kj}^H \mathbf{h}_{ij} \right) + \sigma_i^2}. \quad (4.31)$$

- For fixed y_i and $\mathbf{w}_{ij}$, $\forall i \in \mathcal{N}$ and $j \in \mathcal{L}$, the optimal MSE weight x_i of problem (Q1-2) can be obtained by

$$x_i = \nabla \theta_i(e_i). \quad (4.32)$$

- Under fixed x_i and y_i , $\forall i \in \mathcal{N}$, the optimal transmit beamformer vector $\mathbf{w}_{ij}$ can be obtained by solving the following quadratically constrained quadratic program (QCQP), which can be easily reformulated as a SOCP:

$$\begin{aligned}
\min_{\mathbf{w}_{ij}} \quad & \sum_{i=1}^N x_i e_i + (P_f + \frac{1}{\eta}) \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
\text{s.t.} \quad & \|\mathbf{r}_i\|_2 < \sqrt{1 + 1/(2^{\hat{c}_i} - 1)} \Re[R_{ii}], \quad \forall i \in \mathcal{N}, \\
& \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L},
\end{aligned} \tag{4.33}$$

where e_i is given by (4.30).

Therefore, we can solve problem (P2) with the iteratively WMMSE method as elaborated in Algorithm 4.2, in which

$$O^{(p)} = \sum_{i=1}^N g_i(c_i^{(p)}) + (P_f + \frac{1}{\eta}) \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}^{(p)}\|_2^2,$$

and

$$c_i^{(p)} = B_i \log \left(1 + \frac{|\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}^{(p)}|^2}{\sigma_i^2 + \sum_{k \neq i} |\sum_{j \in \mathcal{L}} \mathbf{h}_{ik}^H \mathbf{w}_{kj}^{(p)}|^2} \right).$$

Algorithm 4.2 Iteratively WMMSE approach for the ESUM problem

- 1: Initialize: $\mathbf{w}_{ij}^{(0)}$ and $p = 1$.
 - 2: **while** $|O^{(p)} - O^{(p-1)}| > \xi$ **do**
 - 3: For given $\mathbf{w}_{ij}^{(p-1)}$, obtain receive beamformer $y_i^{(p)}$ by (4.31);
 - 4: Fix $\mathbf{w}_{ij}^{(p-1)}$ and $y_i^{(p)}$, obtain the MSE weight $x_i^{(p)}$ from (4.32) and (4.30);
 - 5: For given $x_i^{(p)}$, $y_i^{(p)}$ and $z_{ij}^{(p)}$, obtain the transmit beamformer $\mathbf{w}_{ij}^{(p)}$ by solving the convex optimization problem (4.33);
 - 6: Update $c_i^{(p)}$;
 - 7: Let $p = p + 1$.
 - 8: **end while**
 - 9: Output: $\mathbf{c}_o = [c_1^{(p)}, \dots, c_N^{(p)}]^T$.
-

4.3.3 The RRH selection problem

After obtaining the achievable rates $\mathbf{c}^* = [c_1^*, \dots, c_N^*]^T$ via solving problem (P2) by Algorithm 4.1 or Algorithm 4.2, we now turn to the RRH selection problem since the integer variable β_j that indicates whether RRH j is active or not in (P1) is relaxed to a continuous variable in problem (4.24), and then dropped in (P2). The main focus of this subsection is to recover the integer values of $\beta_j, \forall j \in \mathcal{L}$, based on the given optimal UE achievable rate vector $\mathbf{c}^*$ from Algorithm 4.1 or Algorithm 4.2.

Replacing μ_i and c_i in problem (P1) by the solutions μ_i^* and c_i^* obtained from Algorithm 4.1 respectively, and utilizing $\|\mathbf{m}\|_0 = \sum_{j=1}^L \beta_j$, we have the following RRH selection problem:

$$\min_{\beta_j, \mathbf{w}_{ij}} \sum_{j=1}^L \beta_j E_j \quad (4.34)$$

$$\text{s.t.} \quad \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq \beta_j E_j, \quad \forall j \in \mathcal{L},$$

$$\beta_j \in \{0, 1\}, \quad \forall j \in \mathcal{L},$$

$$\|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + 1/(2^{c_i^*/B_i} - 1) \Re[R_{ii}(\mathcal{L})]}, \quad \forall i \in \mathcal{N}, \quad (4.35)$$

which is a MINLP. To solve (4.34) efficiently, we apply the binary-to-continuous relaxation on the variable β_j , for $j \in \mathcal{L}$. Furthermore, we add an additional constraint $\sum_{j=1}^L \beta_j \geq 1$, which is redundant for problem (4.34), to improve the accuracy of binary-to-continuous relaxation. As a result, the relaxed RRH selection problem becomes

$$\begin{aligned} (\text{Q2}) \quad & \min_{\beta_j, \mathbf{w}_{ij}} \sum_{j=1}^L \beta_j E_j \\ & \text{s.t.} \quad \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq \beta_j E_j, \quad \forall j \in \mathcal{L}, \\ & \quad 0 \leq \beta_j \leq 1, \quad \forall j \in \mathcal{L}, \\ & \quad \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + 1/(2^{c_i^*/B_i} - 1) \Re[R_{ii}(\mathcal{L})]}, \quad \forall i \in \mathcal{N}, \end{aligned}$$

$$\sum_{j=1}^L \beta_j \geq 1, \quad (4.36)$$

which is a SOCP and can be solved easily by standard convex programming tool boxes [131]. Let the optimal solution of problem (Q2) be $\{\tilde{\beta}_j, \tilde{w}_{ij} \mid i \in \mathcal{N}, j \in \mathcal{L}\}$. From this solution, our goal is to infer a sparse integer solution for $\beta_1, \dots, \beta_L$.

We can interpret $\tilde{\beta}_j$ to be the priority of RRH j being chosen to be active, where a RRH with relatively lower priority value should be turned off. However, [89] suggests that some incentive algorithms can help improve the RRH selection results. We utilize the reweighted l_1 -norm relaxation as the incentive strategy and propose the following *Shaping-and-Pruning (SP)* algorithm, which has two main steps:

1. **Shaping.** We use the reweighted l_1 -norm relaxation [133] in (Q2) to “shape” the solutions into a sparse form. Specifically, we solve the reweighted problem

$$\begin{aligned} \min_{\beta_j, w_{ij}} \quad & \sum_{i=1}^N \sum_{j=1}^L \rho_j \beta_j E_j \\ \text{s.t.} \quad & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq \beta_j E_j, \quad \forall j \in \mathcal{L}, \\ & 0 \leq \beta_j \leq 1, \quad \forall j \in \mathcal{L}, \\ & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + 1/(2^{c_i^*/B_i} - 1) \Re[R_{ii}(\mathcal{L})]}, \quad \forall i \in \mathcal{N}, \\ & \sum_{j=1}^L \beta_j \geq 1, \end{aligned} \quad (4.37)$$

where $\rho_j = 1/(\tilde{\beta}_j + \xi)$, ξ is adaptively chosen by $\xi = \max\left\{\min(\tilde{\beta}_1, \dots, \tilde{\beta}_L), \phi\right\}$, and ϕ is a small positive value to ensure numerical stability [71]. We denote the optimal solution obtained from problem (4.37) as the *shaped priorities* $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$.

2. **Pruning.** Sort the shaped priorities $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$ in ascending order, so that $\hat{\beta}_{\pi_1} \leq \hat{\beta}_{\pi_2} \leq \dots \leq \hat{\beta}_{\pi_L}$, for some permutation $(\pi_1, \dots, \pi_L)$ of the set $\mathcal{L}$. We define the J th active RRH set to be $\mathcal{A}_J \triangleq \{\pi_{J+1}, \dots, \pi_L\}$. Then, we apply the

bisection search to find J^* , which is the largest index J such that $\beta_{\pi_1} = \dots = \beta_{\pi_J} = 0$ and $\beta_{\pi_j} > 0$, for all $j \geq J + 1$, form a feasible solution to (Q2). Finally, take the active set to be $\mathcal{A}^* = \mathcal{A}_{J^*}$.

After obtaining the active set $\mathcal{A}^*$, the corresponding beamforming weights $\mathbf{w}_{ij}^*$ can be found by solving the following SOCP:

$$\begin{aligned}
& \min_{\mathbf{w}_{ij}} \quad \sum_{i=1}^N \sum_{j \in \mathcal{A}^*} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
& \text{s.t.} \quad \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{A}^*, \\
& \quad \|\mathbf{r}_i(\mathcal{A}^*)\|_2 \leq \sqrt{1 + 1/(2^{c_i^*/B_i} - 1) \Re[R_{ii}(\mathcal{A}^*)]}, \quad \forall i \in \mathcal{N},
\end{aligned} \tag{4.38}$$

where $\mathbf{w}_{ij}^* = 0$, $\forall i \in \mathcal{N}$, and $\forall j \notin \mathcal{A}^*$.

The proposed SP algorithm makes a trade-off between the conventional sorting-and-removing and sparsity-inducing algorithms [70,91]. Specifically, the computational complexity is reduced by utilizing the bisection search, instead of the sequential and iterative search in sorting-and-removing algorithms. We summarize the Shaping-and-Pruning algorithm in Algorithm 4.3.

Remark 4.2. Note that, in the Shaping step in line 2 of Algorithm 4.3, we only need to solve problem (4.37) once, instead of iteratively updating the weights $\{\rho_j \mid j \in \mathcal{L}\}$, as is done in [70]. Suppose that the interior-point method is applied to solve the feasibility problem of (Q2), which is a SOCP, in each iteration. The time complexity to solve each SOCP is $\mathcal{O}((NLK)^{3.5})$, where K is the number of antennas in each RRH [134]. The complexity of solving (4.37) and (4.38) are also both $\mathcal{O}((NLK)^{3.5})$. Therefore, the complexity of our SP algorithm is $\mathcal{O}((NLK)^{3.5} \log L)$.

In summary, the proposed solution for problem (P1) is obtained as follows: the optimal VM computation capacities and achievable rates $\{(\mu_i^*, c_i^*) \mid i \in \mathcal{N}\}$ are obtained from Algorithm 4.1 or Algorithm 4.2, which are then used to determine the optimal active RRH set $\mathcal{A}^*$ and its corresponding beamforming weights w_{ij}^* from

Algorithm 4.3 Shaping-and-Pruning Algorithm

- 1: Initialization. Solve problem (Q2) to obtain $\{\tilde{\beta}_j \mid j \in \mathcal{L}\}$. Let $J_{\min} = 0$ and $J_{\max} = L$.
 - 2: Shaping. Solve problem (4.37), where $\rho_j, j = 1, \dots, L$, are defined by $\{\tilde{\beta}_j \mid j \in \mathcal{L}\}$, to obtain the shaped priorities $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$.
 - 3: Sorting. Sort the shaped priorities $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$ in ascending order, to obtain $\hat{\beta}_{\pi_1} \leq \hat{\beta}_{\pi_2} \leq \dots \leq \hat{\beta}_{\pi_L}$.
 - 4: **while** $J_{\max} - J_{\min} \geq 2$ **do**
 - 5: $J = \lfloor (J_{\max} + J_{\min})/2 \rfloor$;
 - 6: Check the feasibility of problem (Q2) if $\mathcal{A} = \mathcal{A}_J$;
 - 7: **if** feasible **then**
 - 8: $J_{\min} = J$;
 - 9: **else**
 - 10: $J_{\max} = J$;
 - 11: **end if**
 - 12: **end while**
 - 13: Output $J^* = \lfloor (J_{\max} + J_{\min})/2 \rfloor$, $\mathcal{A}^* = \mathcal{A}_{J^*} = \{\pi_{J^*+1}, \dots, \pi_L\}$ and its corresponding beamforming weights $\mathbf{w}_{ij}^*$ by solving (4.38).
-

Algorithm 4.3. For ease of reference, we call this procedure the Cross-Layer Shaping-and-Pruning (CLSP) algorithm.

4.4 Simulation results

In this section, we present simulation results to verify the performance of the proposed CLSP algorithm, and compare it with several existing algorithms in the literature. Note that the performance comparison between BnB algorithm and WMMSE algorithm has been done in [96], in which BnB algorithm shows better performance than WMMSE algorithm. Therefore, in this section, we only focus on utilizing BnB algorithm as the approach to solve the WSUM problem.

4.4.1 Simulation setup

We consider a C-RAN system of 7 RRHs, where RRH 1 to 6 are located on a circle centered at RRH 7, with radius 0.5 km. The RRHs 1 to 6 are placed at equal distances apart, as shown in Figure 4-4. UEs are randomly, uniformly and independently

Table 4.2: Simulation parameters in Chapter 4

Parameter	Value	Parameter	Value
L	7	K	2
σ^2	-83.98 dBm	η	0.2
P_f	5 W	E	1 W
ϑ	5 dB	s	10 dB

Figure 4-4: Simulation setup with 7 RRHs.

distributed within this disk. The wireless transmission bandwidth is 10 MHz. We adopt the path loss model used by the 3GPP specification for Evolved Universal Terrestrial Radio Access in [135], where the received power at a UE d km from a RRH is given by

$$p \text{ (dB)} = 128.1 + 37.6 \log_{10} d.$$

The transmit antenna gain at each RRH is ϑ . The lognormal shadowing parameter is s . In our simulations, we consider homogeneous RRHs and UEs with $E_1 = E_2 = \dots = E_L = E$, $\sigma_1 = \sigma_2 = \dots = \sigma_N = \sigma$, and $\tau_1 = \tau_2 = \dots = \tau_N = \tau$.

For the power consumption function $\varphi_i(\mu_i)$, we adopt the formula $\varphi_i(\mu_i) = k_i \mu_i^3$, where $k_i > 0$ is a constant. This power consumption formula was proposed by [114] and adopted by [53, 125]. We summarize our simulation parameters in Table 4.2 [70, 127].

4.4.2 The effect of shaping

In this subsection, we show the performance of the proposed SP algorithm for the RRH selection problem, compared with the following benchmark algorithms:

- *Exhaustive Search (ES) Algorithm.* This algorithm solves the RRH selection problem (4.34) using an exhaustive search over all possible RRH selections to obtain the optimal solution for problem (4.34). It has a high complexity of $\mathcal{O}((NLK)^{3.5}2^L)$, which makes the algorithm intractable when L becomes large. This is used as a benchmark to compare other algorithms against.
- *Bisection Search (BS) Algorithm.* This algorithm, which was proposed in [136], skips the shaping step in the SP algorithm, and uses $\tilde{\beta}_j$ in place of $\hat{\beta}_j$ for all $j \in \mathcal{L}$ in the pruning step of the SP algorithm. We use this algorithm as a benchmark to show the effect of the shaping step in the SP algorithm. The complexity of the BS algorithm is $\mathcal{O}((NLK)^{3.5} \log L)$.
- *Full Cooperation (FC) Algorithm.* This algorithm assumes all the RRHs are chosen to be active, i.e., $\mathcal{A}^* = \mathcal{L}$. The complexity of the FC algorithm is $\mathcal{O}((NLK)^{3.5})$.
- *Successive Selection (SS) Algorithm.* This algorithm was proposed in [127], and lets all RRHs to be active in the initial iteration, with a RRH having the least power consumption removed at each subsequent iteration. The iterations are performed until the problem (4.34) becomes infeasible. The complexity of the SS algorithm is $\mathcal{O}((NLK)^{3.5}L)$.
- *Greedy Selection (GS) Algorithm.* This algorithm was proposed in [91]. It considers all RRHs to be active in the initial iteration, and then removes the RRH that reduces the system power consumption by the largest amount at each iteration until the problem (4.34) becomes infeasible. Simulation results in [91] suggest that this algorithm produces a near-optimal solution, compared with the global solution obtained by solving a MINLP. The complexity of the GS algorithm is $\mathcal{O}((NLK)^{3.5}L^2)$.

(a) Different required achievable rates.

(b) Different number of UEs.

Figure 4-5: Number of active RRHs using different RRH selection algorithms.

To compare the performance of all the RRH selection algorithms, we suppose the optimal achievable rate c_i^* for each user i is identical in problem (4.34), i.e., $c_1^* = \dots = c_N^* = c$. In Figure 4-5(a), we show the mean number of active RRHs versus each UE's achievable rate c under different RRH selection algorithms when the number of UEs $N = 6$. We see that SP algorithm outperforms FC, SS and BS algorithms over all achievable rates. When $c \leq 15$ Mb/s, the SP algorithm has comparable or even better

sparsity performance than the GS algorithm. When $c \geq 15$ Mb/s, compared to the SP algorithm, the GS algorithm produces a solution with about 5% less active RRHs, but at the expense of $L^2/\log L = 17.5$ times computational complexity overhead.

Next, we let $c = 20$ Mb/s, and show the mean number of active RRHs versus the number of UEs N under different RRH selection algorithms in Figure 4-5(b). We see from Figure 4-5(b) that the SP algorithm has similar sparsity performance as the GS algorithm, and outperforms the FC, SS and BS algorithms. From both Figures 4-5(a) and 4-5(b), we can see that although we incur an overhead to perform the shaping step in the SP algorithm, its solution sparsity is improved by 5% - 10 %.

4.4.3 The importance of cross-layer design

In this subsection, we present simulation results to verify the performance gain using a cross-layer design in which both the BBU pool power consumption and the RRH power consumption are jointly optimized. Most of the previous work in C-RAN optimizes only the power consumption for the wireless transmission layer and BBU pool independently. We call this class of algorithms the decoupled-layer (DL) algorithms. We assume that, for the DL algorithms, the delay in the BBU processing queue a_i and the delay in the RRH transmitting queue b_i satisfy $a_i \leq \tau_i/2$ and $b_i \leq \tau_i/2$ respectively. We formulate optimization programs, similar to problem (P0), for finding the solutions for UEs' achievable rates, VM computation capacities, beamformer vectors and active RRH set separately. The RRH selection problem can then be solved using either the SS or GS algorithms. We call these the DLSS and DLGS method respectively.

In this chapter, we have provided a general framework that allows us to perform cross-layer (CL) optimization of the overall system power consumption. Our two-step approach (cf. Figure 4-3) allows us to first solve a QWSRM and then a RRH selection problem. In particular, for the RRH selection problem, we can again adopt the SP, BS, FC, SS, and GS algorithms. We call these the CLSP, CLBS, CLFC, CLSS, and CLGS methods respectively. From our discussion in the previous subsection, we see that the GS algorithm in general gives the most sparse active RRH set. Therefore in Figure

(a) Different arrival rates.

(b) Different number of UEs.

Figure 4-6: System power consumption ratio under different algorithms.

4-6, we use the performance of the CLGS method as the normalizing benchmark, and for each method, we plot the ratio of the average system power consumption versus that of the CLGS method.

In Figure 4-6, we let users' mean arrival rates to be identical, i.e., $\lambda_1 = \dots = \lambda_N = \lambda$. We show the relationship between the UEs' mean arrival rate and the system power consumption for $N = 6$ in Figure 4-6(a). We observe that, firstly,

CL algorithms outperform DL algorithms. That is, CLSS and CLGS perform better than DLSS and DLGS respectively. Secondly, CLSP outperforms CLSS and CLBS, and CLSP outperforms DLGS under the high traffic rate regime, i.e., $\lambda \geq 15$ Mb/s. Finally, as the incoming traffic rate increases, the performance gap between CLFC and CLSP becomes smaller since CLSP needs more active RRHs to support the higher rate demand. The performance of system power consumption versus the number of UEs is depicted in Figure 4-6(b) when $\lambda = 20$ Mb/s. We see again that CL algorithms are better than DL algorithms. CLSP also outperforms CLBS, which again shows the importance of the shaping step in the SP algorithm.

4.5 Summary

In this chapter, we consider the problem of optimizing the allocated VM computation capacities in the BBU pool, the set of active RRHs, and the beamforming strategies at the active RRHs in order to minimize the overall system power consumption for C-RAN. Specifically, this chapter can be summarized as follows:

- We formulate the cross-layer resource allocation problem as a MINLP by minimizing the system power consumption, which consists of three parts: the power consumption in the BBU pool w.r.t. the VM computation capacity, the power consumption in the fiber links w.r.t. the number of active links (or, active RRHs) and the transmission power on the RRHs w.r.t. the transmit beamformer.
- We relax the MINLP into a ESUM problem, and propose two different approaches to solve it: a BnB algorithm and a WMMSE algorithm. Based on the achievable rates found by solving the ESUM problem, we propose an efficient Shaping-and-Pruning algorithm to perform RRH selection. Our proposed algorithm achieves a trade-off between computational complexity and solution optimality.
- We provide simulation results that suggest that our proposed approach outperforms the recently proposed greedy selection algorithm of [91] and successive

selection algorithm of [127] in terms of overall system power consumption, since these methods only optimize the RRH selection and RRH beamforming strategies. This shows that cross-layer optimization can result in higher energy efficiencies for a C-RAN.

4.6 Proofs

4.6.1 Proof of Theorem 4.1

Suppose that there exists some $i \in \mathcal{N}$ such that $\tilde{c}_i < \bar{c}_i$, and $c_i^* \in (\tilde{c}_i, \bar{c}_i]$. Since $f(\mathbf{c})$ is convex and finite, there exists $\hat{c}_i \in (\tilde{c}_i, c_i^*)$ such that $f(\hat{\mathbf{c}}) < f(\mathbf{c}^*)$, where $\hat{\mathbf{c}}$ is $\mathbf{c}^*$ with the i -th element c_i^* replaced by $\hat{c}_i$. In addition, we have

$$\begin{aligned} \|\mathbf{r}_i^*(\mathcal{A})\|_2 &\leq \sqrt{1 + 1/(2^{c_i^*/B_i} - 1)} \Re[R_{ii}^*(\mathcal{A})] \\ &< \sqrt{1 + 1/(2^{\hat{c}_i/B_i} - 1)} \Re[R_{ii}^*(\mathcal{A})], \end{aligned}$$

which implies that $\hat{\mathbf{c}}$ is a feasible rate vector for the QWSRM (4.14). But $f(\hat{\mathbf{c}}) < f(\mathbf{c}^*)$, which contradicts the assumption that $\mathbf{c}^*$ is optimal for the QWSRM. The theorem is now proved.

4.6.2 Proof of Proposition 4.1

On the one hand, if we fix the variables $\{\mu_i, c_i, \beta_j\}$ in problem (P1), then problem (P1) is reduced to

$$\begin{aligned} \min_{\mathbf{w}_{ij}} \quad & \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\ \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1}} \Re[R_{ii}(\mathcal{L})], \forall i \in \mathcal{N}, \\ & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq \beta_j E_j, \forall j \in \mathcal{L}, \end{aligned} \tag{4.39}$$

which is a SOCP. Then, we can observe that, if we slightly increase the value of constant c_i , the feasible region of problem (4.39) is shrunk accordingly. That means the optimal value of problem (4.39) is nondecreasing w.r.t. c_i .

On the other hand, from Assumption 4.1, $\varphi_i(\mu_i)$ is increasing w.r.t. μ_i . Therefore, the optimal $\{\mu_i, c_i\}$ of problem (P1) must achieve equality in the system delay constraint (4.5) since the left hand side of (4.5) is monotonically decreasing w.r.t. μ_i and c_i respectively. The proposition is now proved.

4.6.3 Proof of Proposition 4.2

On the one hand, from Assumption 4.1, $\varphi_i(\cdot)$ is convex and increasing; on the other hand, $\lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2(c_i - \lambda_i) - \tau_i}$ is convex w.r.t. $c_i > \lambda_i + 1/\tau_i$. Then it can be proved that, for each $i \in \mathcal{N}$, if $c_i > \lambda_i + 1/\tau_i$, $f_i(c_i)$ is convex, based on the composition rules which can preserve convexity [137].

Therefore, $\sum_{i=1}^N f_i(c_i)$ is a convex function over $\hat{\mathcal{Q}}_{\text{init}}$, and it also satisfies the three properties of the objective function $f(\mathbf{c})$ in a QWSRM problem (4.14) in Section 4.2. Therefore, problem (Q1-1) is a QWSRM problem. In addition, $\bar{c}_i$ is an upper bound of c_i derived from (4.28), since for any $i \in \mathcal{N}, j \in \mathcal{L}$, we have $\|\mathbf{w}_{ij}\|_2^2 \leq E_j$. The proposition is now proved.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Towards System Cost Minimization in C-RAN with Limited Fronthaul Capacity

After solving the problem in Chapter 4, which makes use of an idealized model for C-RAN, we consider a more general model for C-RAN. We list the differences between the idealized C-RAN model and the general C-RAN model as follows:

1. We consider the general model that each UE can associate with multiple VMs in the BBU pool, and we need to choose the optimal number of active VMs in the BBU pool. In contrast, in the idealized model, we just assume each UE can only associate with one VM, and the number of active VMs is hence equivalent to the number of UEs..
2. In this general C-RAN model, we consider limited fronthaul capacity, which is assumed to be unlimited in the idealized C-RAN model. As a consequence, each UE can only access to a subset of RRHs in the coordinated RRH cluster. However, we assume each UE can access every RRH in the active cluster in the idealized C-RAN model.

In this chapter, we consider the cost interaction between cloud processing and wireless transmission in C-RAN. In particular, we take the joint consideration of

Figure 5-1: A general model of C-RAN.

VM activation in the BBU pool and sparse beamforming in the coordinated RRHs cluster, which have limited fronthaul capacity constraint for each RRH, to minimize the system cost of C-RAN.

5.1 System model

In this section, we present a general C-RAN model, which is a more practical version than the one in Chapter 4.

5.1.1 System description

Suppose that there are N single-antenna UEs and L coordinated RRHs, each with K antennas, in C-RAN cluster. We denote the set of all UEs and all RRHs as $\mathcal{N} = \{1, \dots, N\}$ and $\mathcal{L} = \{1, \dots, L\}$ respectively. There are M homogenous VMs in the BBU pool, each has the computation capacity μ and cost φ when it is active. We denote the number of active VMs as m , hence $m \in \mathbb{N}$ and $m \leq M$.

In the downlink of C-RAN (cf. Figure 5-1), all UEs' incoming traffic is received by a dispatcher firstly. Let's assume the mean arrival data rate in the dispatcher (for UE i) is λ_i , $\forall i \in \mathcal{N}$, and denote $\alpha = \sum_{i \in \mathcal{N}} \lambda_i$. Then each transport block [65] (or even a code block within each transport block) in the data flow to UE i can be routed to one of m active VMs for processing (e.g., turbo coding) with probability $1/m$ by the dispatcher. And hence, the mean incoming traffic rate routed to each active VM

is α/m .

In the wireless transmission part, we consider joint transmission as the CoMP technique in C-RAN, i.e., each UE's data can be shared among all the coordinated/associated RRHs, while some of the RRHs have limited fronthaul link capacity (Note that the fronthaul links between the BBU pool and the RRHs are heterogeneous, they can be fiber links, copper cables or wireless channels.). After processed by the VMs, each UE's data is forwarded to the UE via (up to) L RRHs (since the data is shared among the limited fronthaul RRHs). Let the achievable wireless transmission rate to UE i in C-RAN be c_i .

5.1.2 Queueing system model

Each active VM in the BBU pool can be captured as a queue. Specifically, for each queue, the mean arrival rate is α/m and the mean service rate is μ . We assume the tasks within each queue are served in a first in first out (FIFO) manner and the buffer length is infinite. Furthermore, queueing model, with the channel capacity as the queue's service rate, is also widely used to characterize wireless communication systems [121].

Therefore, we introduce a double-layer queueing network to represent each UE's data processing and transmitting behavior in the C-RAN downlink (cf. Figure 5-2). Our model can be easily extended to the C-RAN uplink as well. Specifically, in the BBU pool, the transport blocks to each UE are processed (e.g., encoded) by m parallel active VMs, each of them is abstracted as a cloud processing queue, with mean service rate μ . Then, the processed data is transmitted to UE i via RRHs and wireless channels, which are characterized by a wireless transmission queue with mean service rate c_i .

We denote the mean expected delay for data to UE i in the BBU pool as a_i (i.e., the expected delay during cloud processing). Let b_i be the expected delay of the data to UE i in the wireless transmission queue (i.e., the expected delay during wireless transmission). We assume that UE i 's packet arrival process to the dispatcher is a Poisson process with mean rate λ_i . Hence, the arrival process to each VM also

Figure 5-2: Queueing network model representation of a C-RAN cloud processing and wireless transmission.

forms Poisson process with mean arrival rate α/m . Suppose that the service time of each data packet in cloud processing queue follows an exponential distribution with mean $1/\mu$, for $\mu > \alpha/m$. Then, for each UE's data, the arrival rate to the wireless transmission queue is the same as the one to the dispatcher [122, 123]¹. We assume that the service time of each data packet in the wireless transmission queue follows an exponential distribution with mean $1/c_i$. Therefore, the data processing and transmitting in our C-RAN model can be treated as two layers of M/M/1 queues in tandem. In addition, from queueing theory [115], we have

$$a_i = \frac{m}{m\mu - \alpha}, \quad (5.1)$$

$$b_i = \frac{1}{c_i - \lambda_i} \quad (5.2)$$

where $\mu > \alpha/m, c_i > \lambda_i, \forall i \in \mathcal{N}$. Hence, the expected system delay for the data to UE i is

$$d_i = a_i + b_i.$$

We assume the channel does not change over the duration of transmission, i.e., block fading channel model. Let u_i denote the data symbol for the i th UE with $E[|u_i|^2] = 1$, and $\mathbf{w}_{ij} \in \mathbb{C}^K$ denote the transmit beamformer for UE i from RRH j .

¹Note that, we do not consider the impact, introduced by cloud processing queue, on the arrival data rate to the wireless transmission queue. Specifically, we assume that, after being processed by the cloud processing queue, the size of transport blocks and the inter-arrival times to the gateway still remain the same as the one to the dispatcher.

The block fading channel from RRH j to UE i is denoted as $\mathbf{h}_{ij}^H$, where $\mathbf{h}_{ij} \in \mathbb{C}^K$, for $i \in \mathcal{N}$ and $j \in \mathcal{L}$. Thus, the received signal at UE i is

$$\hat{u}_i = \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} u_i + \sum_{l \neq i} \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj} u_l + \delta_i,$$

where the first term is the useful signal for UE i , the second term is the interference to UE i and $\delta_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive white Gaussian noise (AWGN) at UE i .

Thus, the signal-to-interference-plus-noise ratio (SINR) at UE i is

$$\text{SINR}_i = \frac{|\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}|^2}{\sigma_i^2 + \sum_{k \neq i} |\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}|^2}. \quad (5.3)$$

The downlink achievable rate to UE i , c_i , should satisfy

$$c_i \leq B_i \log(1 + \text{SINR}_i),$$

where B_i is the wireless transmission bandwidth for UE i . And, in the meanwhile, each RRH j has its maximum transmitting power E_j , i.e.,

$$\sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \text{ for } j \in \mathcal{L}.$$

On top of the basic system model above, we have the following two additional constraints to capture the features of general C-RAN model:

1. **System delay constraint:** To couple the cloud processing and wireless transmission in C-RAN, we propose the the cross-layer system delay constraint:

$$d_i \leq \tau_i, \forall i \in \mathcal{N},$$

where τ_i is a predefined QoS requirement for UE i .

2. **Fronthaul capacity constraint:** We denote S_j as the fronthaul link capacity w.r.t. the maximum number of UEs that can be connected with this fronthaul

link. In other words, if S_j is limited, there's only a subset of UEs which can be associated with RRH j . We can cast this fronthaul capacity constraint as

$$\sum_{i \in \mathcal{N}} \|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\|_0 \leq S_j, \forall j \in \mathcal{L},$$

where $\|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\|_0 = 1$ if and only if RRH j is associated with UE i .

5.2 Problem formulation

Our aim is to minimize the system cost in C-RAN, which consists of two components: the cost in BBU pool, and the power consumption at the RRHs. Specifically, (i) the cost for all the active VMs is $m\varphi$; (ii) the cost in RRHs, which is a linear function w.r.t. the power consumption at RRHs. In particular, the cost in RRHs is $\eta \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij}$, where $\eta > 0$ is the weight between the cost in BBU pool and RRHs. For example, for $\eta > 1$, we can interpret $1/\eta$ as the inefficiency coefficient of the amplifier in each RRH.

Therefore, our optimization problem can be formulated as

$$\begin{aligned} \text{(OP0)} \quad & \min_{m, c_i, \mathbf{w}_{ij}} \quad m\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\ \text{s.t.} \quad & \frac{m}{m\mu - \alpha} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \quad \forall i \in \mathcal{N}, \end{aligned} \quad (5.4)$$

$$\alpha < m\mu, \lambda_i < c_i, \quad \forall i \in \mathcal{N}, \quad (5.5)$$

$$0 < m \leq M, \quad m \in \mathbb{N}, \quad (5.6)$$

$$c_i \leq B_i \log(1 + \text{SINR}_i), \quad \forall i \in \mathcal{N}, \quad (5.7)$$

$$\sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L}, \quad (5.8)$$

$$\sum_{i=1}^N \|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\|_0 \leq S_j, \quad \forall j \in \mathcal{L}, \quad (5.9)$$

where SINR_i is given by (5.3).

Figure 5-3: An iterative approach to solve problem (OP0).

We assume the feasible region of problem (OP0) is nonempty. Let the optimal solution for problem (OP0) be $\{m^*, c_i^*, \mathbf{w}_{ij}^*\}$. We define the $l_{2,0}$ -norm of vector $\mathbf{w}_{ij}$ as $\|\mathbf{w}_{ij}\|_{2,0} \triangleq \|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\|_0$.

Problem (OP0) is a MINLP. It is difficult to solve for the following reasons: (i) it has two integer constraints, i.e., (5.6) and (5.9); and (ii) the problem is nonconvex even if we assume $m \in \mathbb{R}^+$ and the fronthaul capacity constraint (5.9) are removed.

In this chapter, we propose an iterative approach (cf. Figure 5-3) to solve problem (OP0), which includes has two integer constraints. In particular, in Subsection 5.2.1, we propose a reformulation for problem (OP0), which aims to handle integer constraint (5.9) and makes problem (OP0) much more tractable. Then, in Subsection 5.3.3, we propose two different approaches to deal with integer constraint (5.6).

5.2.1 An equivalent formulation for problem (OP0)

In problem (OP0), one of the solution challenges is the $l_{0,2}$ -norm constraint (5.9). Two commonly used approaches that deal with $l_{0,2}$ -norm the are: smoothing function approximation [72, 138] and reweighted $l_{1,2}$ -norm approximation [76, 133]. However, if we just relax the left hand side of constraint (5.9) with smoothing function or $l_{1,2}$ -norm approximations, and solve the relaxed problem directly, then we have to guarantee that the optimal solution derived from the relaxed problem is also feasible for the original problem (OP0). That involves extra overheads for feasibility checking. In this subsection, we propose an equivalent formulation for problem (OP0), which

avoids the overhead caused by feasibility checking.

Let's consider the following problem, which considers the trade-off between the system cost and fronthaul capacity:

$$\begin{aligned}
(\text{OP1}) \quad & \min_{m, c_i, \mathbf{w}_{ij}} \quad m\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} + \sum_{i=1}^N \sum_{j=1}^L \gamma_j \|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\|_0 \\
& \text{s.t.} \quad \frac{m}{m\mu - \alpha} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \quad \forall i \in \mathcal{N}, \\
& \quad \alpha < m\mu, \lambda_i < c_i, \quad \forall i \in \mathcal{N}, \\
& \quad 0 < m \leq M, \quad m \in \mathbb{N}, \\
& \quad c_i \leq B_i \log(1 + \text{SINR}_i), \quad \forall i \in \mathcal{N}, \\
& \quad \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L},
\end{aligned}$$

where γ_j is the price for RRH j . Let $\Gamma \triangleq [\gamma_1, \dots, \gamma_L]^T$.

We denote $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ as the optimal solution for problem (OP1) for given price vector Γ . Define $\beta_j(\Gamma) = \sum_{i=1}^N \|\mathbf{w}_{ij}(\Gamma)\|_{2,0}$. The following theorem builds up the relationship between problem (OP1) and (OP2), we relegate its proof to Section 5.6.1.

Theorem 5.1. Problem (OP0) and (OP1) are equivalent, in sense of the optimal system cost, if $\beta_j(\Gamma) = S_j, \forall j \in \mathcal{L}$.

Based on Theorem 5.1, solving problem (OP0) now becomes finding the price vector Γ , such that the optimal solution for problem (OP1) stratifies

$$\beta_j(\Gamma) = S_j, \forall j \in \mathcal{L}. \quad (5.10)$$

In next section, we introduce properties of price vector Γ and propose a algorithm to find out the optimal Γ such that equation (5.10) holds.

5.3 Approaches to solve problem (OP1)

Instead of solving problem (OP0) directly, in what follows, we propose a step-by-step relaxation and reformulation approach to simplify the problem. Specifically,

1. In subsection 5.3.1, we introduce some properties of price vector Γ in problem (OP1), and we propose a price adjusting algorithm to find out the proper value of Γ that satisfies equation (5.10) in Theorem 5.1.
2. For fixed price vector Γ , we apply reweighted l_1 -norm relaxation on problem (OP1), then problem (OP1) is simplified into problem (OP2) in subsection 5.3.2.
3. We propose two different approaches to solve problem (OP2) in subsection 5.3.3.

5.3.1 Bisection search for price vector Γ

Inspired by [73], the following properties help shedding lights on how to iteratively adjust the price vector in problem (OP1) such that equation (5.10) holds.

Proposition 5.1. Price γ_j , $j \in \mathcal{L}$, has following properties, if we fix γ_k as a constant $\bar{\gamma}_k$, $\forall k \in \mathcal{L}/j$, and denote $\bar{\Gamma}_j = [\bar{\gamma}_1, \dots, \bar{\gamma}_{j-1}, \gamma_j, \bar{\gamma}_{j+1}, \dots, \bar{\gamma}_L]^T$.

1. $\beta_j(\bar{\Gamma}_j)$ is a decreasing function w.r.t. γ_j .
2. There is a threshold price for RRH j , $\theta_j = \varphi M + \sum_{j \in \mathcal{L}} E_j + \eta \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k S_k$, such that, for $\gamma_j \geq \theta_j$, $\beta_j(\bar{\Gamma}_j) \leq S_j$.

Proof. The proof is similar with the one in [73]. We provide it for completeness in Section 5.6.2. □

Recall that the feasible region of problem (OP0) is nonempty, therefore, we can always satisfy equation (5.10) by iteratively searching $\gamma_j \in [0, \theta_j]$.

We elaborate the algorithm to solve problem (OP1) by iteratively adjusting the price vector in problem (OP1) in Algorithm 5.1.

Algorithm 5.1 Price adjusting algorithm for problem (OP1)

- 1: Initialize: Let $\Gamma^{(0)} = \mathbf{0}_{L \times 1}$
 - 2: Iteration l : Solve problem (OP1) with given $\Gamma^{(l-1)}$, obtaining $\beta_j(\Gamma^{(l-1)})$, for $j \in \mathcal{L}$.
 - 3: **if** $\beta_j(\Gamma^{(l-1)}) \leq S_j, \forall j \in \mathcal{L}$, **then**
 - 4: problem (OP1) achieves optimal solution, stop iteration;
 - 5: **else**
 - 6: for those $\tilde{j} \in \{j : \beta_j(\Gamma^{(l-1)}) > S_j, \forall j \in \mathcal{L}\}$, fix $\bar{\gamma}_k = \gamma_k^{(l-1)}, \forall k \in \mathcal{L}/\tilde{j}$, and find $\gamma_{\tilde{j}}$ in $[\gamma_{\tilde{j}}^{(l-1)}, \theta_{\tilde{j}}^{(l-1)}]$ with bisection search such that equation $\beta_{\tilde{j}}(\bar{\Gamma}_{\tilde{j}}) = S_{\tilde{j}}$ holds. Then, let $\gamma_{\tilde{j}}^{(l)} = \gamma_{\tilde{j}}$.
 - 7: **end if**
 - 8: Let $l = l + 1$, go to step 2.
-

In Algorithm 5.1, the main iteration in Step 2 involves a algorithm to solve problem (OP1). Although we avoid the feasibility problem by reformulating problem (OP0) in to (OP1), the $l_{0,2}$ -norm still remains unsolved in the objective function. In next subsection, we introduce reweighted $l_{1,2}$ -norm approximation for problem (OP1).

5.3.2 Reweighted l_1 -norm relaxation

In compressive sensing [139], reweighted l_1 -norm is regarded as an effective way to deal with the l_0 -norm in the objective function, since l_1 -norm is the convex relaxation for l_0 -norm [133] .

Specifically, we can relax the terms involving l_0 -norm in the objective function of problem (OP2) as:

$$\|\mathbf{w}_{ij}^H \mathbf{w}_{ij}\|_0 \approx \rho_{ij}^{(p)} \mathbf{w}_{ij}^H \mathbf{w}_{ij}, \quad (5.11)$$

where $\rho_{ij}^{(p)} = 1 / \left(\left\| \mathbf{w}_{ij}^{(p-1)} \right\|_2^2 + \phi \right)$ is the weight in iteration p , $\mathbf{w}_{ij}^{(p-1)}$ is a constant vector obtained from previous iteration and ϕ is a small positive constant to guarantee the numerical stability. The intuition behind the weight $\rho_{ij}^{(p)}$ is that the beamformer which has smaller norm in iteration $p-1$ is allocated a larger weight $\rho_{ij}^{(p)}$ in iteration p , and hence, the norm of the beamformer is further reduced after solving the problem in iteration p .

With the l_1 -norm relaxation, problem (OP1) can be approximated as the following problem:

$$\begin{aligned}
(\text{OP2}) \quad & \min_{m, c_i, \mathbf{w}_{ij}} \quad m\varphi + \sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
& \text{s.t.} \quad \frac{m}{m\mu - \alpha} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \quad \forall i \in \mathcal{N}, \\
& \quad \alpha < m\mu, \lambda_i < c_i, \quad \forall i \in \mathcal{N}, \\
& \quad 0 < m \leq M, \quad m \in \mathbb{N}, \\
& \quad c_i \leq B_i \log(1 + \text{SINR}_i), \quad \forall i \in \mathcal{N}, \\
& \quad \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L},
\end{aligned}$$

where $z_{ij}^{(p)} = \eta + \gamma_j \rho_{ij}^{(p)}$.

Problem (OP2) is still a MINLP. In next subsection, we discuss two different approaches to solve it.

5.3.3 Two approaches for optimizing the variable m

From constraints (5.4) and (5.5), we can derive that,

$$\begin{aligned}
m & \geq \alpha \left(\frac{\tau_i}{n_i} + \frac{1}{n_i(n_i(c_i - \lambda_i) - \mu)} \right) \\
& \geq \alpha \left(\frac{\tau_i}{n_i} + \frac{1}{n_i(n_i(\check{c}_i - \lambda_i) - \mu)} \right) \triangleq m_i, \quad \forall i \in \mathcal{N},
\end{aligned} \tag{5.12}$$

or,

$$c_i \geq \lambda_i + \frac{\mu_i}{n_i} + \frac{\alpha}{n_i(n_i m - \alpha \tau_i)} \triangleq g_i(m), \quad \forall i \in \mathcal{N}, \tag{5.13}$$

where $n_i = \tau_i \mu - 1 > 0$, and $\check{c}_i$ is an upper bound of c_i , which can be derived from (4.28), i.e.,

$$\check{c}_i = B_i \log \left(1 + \frac{1}{\sigma_i^2} \sum_{j=1}^L \|\mathbf{h}_{ij}\|_2^2 E_j \right), \quad \forall i \in \mathcal{N}.$$

Based on (5.12) and (5.13), in what follows, we discuss two different approaches to solve problem (OP2).

A searching integer (SI) m approach

Once m is fixed as an integer $\bar{m}$, problem (OP2) is reduced into the following weighted sum-power minimization (WSPM) problem:

$$\begin{aligned}
(\text{OP2-1}) \quad & \min_{\mathbf{w}_{ij}} \quad \sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
\text{s.t.} \quad & \bar{c}_i \leq B_i \log(1 + \text{SINR}_i), \quad \forall i \in \mathcal{N}, \\
& \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L},
\end{aligned} \tag{5.14}$$

where $\bar{c}_i = g_i(\bar{m})$ is a constant.

Similar with Chapter 4, since phase rotation of $\mathbf{w}_{ij}$ cannot affect problem (OP2-1), we can recast constraint (5.14) as the following second-order cone (SOC) [127, 128]:

$$\|\mathbf{r}_i\|_2 \leq \sqrt{1 + 1/(2^{\frac{\bar{c}_i}{B_i}} - 1)} \Re[R_{ii}], \quad \forall i \in \mathcal{N}, \tag{5.15}$$

where $R_{ik} = \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}$, $\mathbf{r}_i = [R_{i1}, \dots, R_{iN}, \sigma_i]^T$ and $\Re(\cdot)$ stands for the real part of a complex number. Thus, problem (OP2-1) can be reformulated as a second-order cone programming (SOCP), which can be easily solved by interior point method with some standard optimization tool boxes, e.g., CVX [131].

Therefore, a straightforward approach to obtain optimal solution for problem (OP2) is, firstly, searching integer (SI) m in the following interval

$$\left[\max_{i \in \mathcal{N}} \lceil m_i \rceil, M \right], \tag{5.16}$$

and solving the SOCP (OP2-1) with given $\bar{m}$. Then, comparing the system costs corresponding to different $\bar{m}$ to find out the optimal m^* . Therefore, SI approach obtains the global optimal solution for problem (OP2).

Denote $\mathbf{w}_{ij}(\bar{\mathbf{c}})$ as the optimal solution of the WSPM problem for given $\bar{\mathbf{c}} \triangleq [\bar{c}_1, \dots, \bar{c}_N]^T$. Moreover, we denote $\bar{\mathbf{c}}' \triangleq [\bar{c}'_1, \dots, \bar{c}'_N]^T$. Suppose that, for $i \in \mathcal{N}$, $\bar{c}'_i > \bar{c}_i$, and $\bar{c}'_k = \bar{c}_k$, for all $k \in \mathcal{N}/i$.

Proposition 5.2. For the WSPM problem, the optimal objective function value $\sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}(\bar{\mathbf{c}})^H \mathbf{w}_{ij}(\bar{\mathbf{c}}) \leq \sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}(\bar{\mathbf{c}}')^H \mathbf{w}_{ij}(\bar{\mathbf{c}}')$.

Proof. Since the WSPM problem can be reformulated as a SOCP by recasting constraint (5.14) into (5.15). Then, we can easily observe from constraint (5.15) that the feasible region of this SOCP is shrunk when we use $\bar{\mathbf{c}}'$ instead of $\bar{\mathbf{c}}$ as the input of the WSPM problem. Hence, the optimal objective value $\sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}(\bar{\mathbf{c}}')^H \mathbf{w}_{ij}(\bar{\mathbf{c}}')$, obtained with a shrunk feasible region, is no less than $\sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}(\bar{\mathbf{c}})^H \mathbf{w}_{ij}(\bar{\mathbf{c}})$. The proposition is now proved. $\square$

Proposition 5.2 shows that the optimal objective function value of the WSPM problem is non-decreasing w.r.t. $\bar{c}_i$, $i \in \mathcal{N}$. Based on this proposition, we have the following corollary.

Corollary 5.1. Constraint (5.4) are active inequality constraints, i.e., the optimal $\{m^*, c_i^*\}$ for problem (OP0) always satisfies the following equation

$$\frac{m^*}{m^* \mu - \alpha} + \frac{1}{c_i^* - \lambda_i} = \tau_i, \quad \forall i \in \mathcal{N}. \quad (5.17)$$

Proof. We can observe that, firstly, the objective function of (OP0) is monotonically increasing w.r.t. m ; secondly, based on Proposition 5.2, if we fix m , the objective function of problem (OP0) is non-decreasing w.r.t. c_i , for all $i \in \mathcal{N}$. Therefore, the optimal solution of problem (OP0) $\{m^*, c_i^*\}$ must achieve equality in the system delay constraint (5.4) since the left hand side of (5.4) is monotonically decreasing w.r.t. m and c_i respectively. The corollary is now proved. $\square$

Note that Corollary 5.1 is applicable for problem (OP1) and (OP2) as well.

A joint optimization (JO) approach

However, if the number of all available VMs M is very large, the aforementioned SI algorithm may not be applicable. For the large M case, we can relax m from natural numbers to non-negative real numbers, i.e., $m \in \mathbb{R}^+$.

For the received signal at UE i , $\hat{u}_i$, we denote $y_i \in \mathbb{C}$ as the receive beamformer. Then, the mean square error (MSE) e_i is defined as

$$\begin{aligned} e_i &\triangleq \mathbb{E} [\|y_i^H \hat{u}_i - u_i\|_2^2] \\ &= \left| y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} - 1 \right|^2 + \sum_{l \neq i}^N \left| y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj} \right|^2 + \sigma_i^2 |y_i|^2, \\ &= \sum_{l=1}^N \left| y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj} \right|^2 - 2\Re \left[y_i^H \sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} \right] + \sigma_i^2 |y_i|^2 + 1. \end{aligned} \quad (5.18)$$

Hence, for given transmit beamformer $\mathbf{w}_{ij}$, the minimum mean square error (MMSE) is

$$e_i^{\text{mmse}} = 1 - \sum_{j \in \mathcal{L}} \mathbf{w}_{ij}^H \mathbf{h}_{ij} y_i^{\text{mmse}}, \quad (5.19)$$

where y_i^{mmse} is the well-known MMSE receive beamformer given by:

$$y_i^{\text{mmse}} = \frac{\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}}{\sum_{k \in \mathcal{N}} \left(\sum_{j \in \mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj} \right) \left(\sum_{j \in \mathcal{L}} \mathbf{w}_{kj}^H \mathbf{h}_{ij} \right) + \sigma_i^2}. \quad (5.20)$$

Lemma 5.1. Each UE i 's achievable rate $B_i \log(1 + \text{SINR}_i)$ satisfies the following equation [98]:

$$B_i \log(1 + \text{SINR}_i) = \max_{x_i, y_i} (\log x_i - x_i e_i + 1) B_i, \quad (5.21)$$

where $x_i \in \mathbb{R}^+$ is the MSE weight.

With Lemma 5.1, we have the following Proposition.

Proposition 5.3. For $m \in \mathbb{R}^+$, problem (OP2) can be represented as:

$$\begin{aligned}
 \text{(OP2-2)} \quad & \min_{x_i, y_i, m, \mathbf{w}_{ij}} \quad m\varphi + \sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
 \text{s.t.} \quad & g_i(m) \leq \max_{x_i, y_i} (\log x_i - x_i e_i + 1) B_i, \forall i \in \mathcal{N}, \quad (5.22)
 \end{aligned}$$

$$\begin{aligned}
 & \max_{i \in \mathcal{N}} [m_i] \leq m \leq M, \quad m \in \mathbb{R}^+, \quad (5.23) \\
 & \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L},
 \end{aligned}$$

where $g_i(m)$ and e_i are given by (5.13) and (5.18) respectively.

Let's partition the variables in problem (OP2-2) into three groups, i.e., x_i , y_i and $\{m, \mathbf{w}_{ij}\}$. The reasons that we introduce two new groups of variables x_i and y_i in problem (OP2-2) are:

- Firstly, if we fix x_i and y_i , then problem (OP2-2) can be easily recast as a convex optimization problem w.r.t. $\{m, \mathbf{w}_{ij}\}$, since $g_i(m)$ is a convex function.
- In addition, for the right hand side of (5.22), by checking the first order optimality condition, optimal receive beamformer y_i can be obtained by equation (5.20).
- The optimal MSE weight x_i in the right hand side of (5.22) is given by

$$x_i = (e_i^{\text{mmse}})^{-1}, \quad (5.24)$$

for fixed $\{m, \mathbf{w}_{ij}\}$ and y_i .

Since problem (OP2-2) is convex w.r.t. each variable group while keeping other variable groups fixed. Therefore problem (OP2-2) is much easier to solve than solve problem (OP2) directly. Specifically, in problem (OP2-2), the optimal y_i and x_i can be obtained with closed-form solutions if we fix the other variable groups as constants respectively, and $\{m, \mathbf{w}_{ij}\}$ can be obtained by solving the following convex

optimization problem, for fixed x_i and y_i :

$$\begin{aligned}
(\text{OP2-2.1}) \quad & \min_{m, \mathbf{w}_{ij}} \quad m\varphi + \sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
\text{s.t.} \quad & g_i(m) + B_i x_i e_i \leq B_i (\log x_i + 1), \forall i \in \mathcal{N}, \\
& \max_{i \in \mathcal{N}} \lceil m_i \rceil \leq m \leq M, \quad m \in \mathbb{R}^+, \\
& \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L}.
\end{aligned} \tag{5.25}$$

That means problem (OP2-2) can be efficiently solved by alternating optimization method.

Thus, we can conclude that the algorithm to solve problem (OP1) includes two tiers of loops, i.e., a outer loop to update the weights $z_{ij}^{(p)}$ and a inner loop to solve problem (OP2-2.1) by the iteratively WMMSE method. To simplify the complexity, we can combine the two tiers of loops together with a single loop [76], as elaborated in Algorithm 5.2, in which

$$O^{(p)} = m^{(p)}\varphi + \sum_{i=1}^N \sum_{j=1}^L z_{ij}^{(p)} \left\| \mathbf{w}_{ij}^{(p)} \right\|_2^2.$$

It is proved in [97] that Algorithm 5.2 can produce a local optimal solution for problem (OP2-2).

5.4 Numerical results

In this section, we conduct simulation to verify the performance of our proposed algorithm, and compare it with latest proposed benchmark algorithms.

5.4.1 Simulation setup

We consider a heterogeneous C-RAN system of 7 RRHs, where RRH 1 to 6 are located on a circle centered at RRH 7, a macro RRH, with radius 0.5 km. RRHs 1 to 6 are

Algorithm 5.2 Joint reweighted l_1 -norm relaxation and iteratively WMMSE approach for problem (OP2)

- 1: Initialize: $\mathbf{w}_{ij}^{(0)}$ and $p = 1$.
 - 2: **while** $|O^{(p)} - O^{(p-1)}| > \xi$ **do**
 - 3: Given $\mathbf{w}_{ij}^{(p-1)}$, obtain receive beamformer $y_i^{(p)}$ by (5.20);
 - 4: Fix $\mathbf{w}_{ij}^{(p-1)}$ and $y_i^{(p)}$, obtain the MSE weight $x_i^{(p)}$ from (5.19) and (5.24);
 - 5: Given $x_i^{(p)}$, $y_i^{(p)}$ and $z_{ij}^{(p)}$, obtain the number of active VMs $m^{(p)}$ and transmit beamformer $\mathbf{w}_{ij}^{(p)}$ by solving the convex optimization problem (OP2-2.1);
 - 6: Update $z_{ij}^{(p)}$;
 - 7: Let $p = p + 1$.
 - 8: **end while**
 - 9: Output: optimal number of active VMs is $\lceil m^{(p)} \rceil$ and optimal beamformer vectors can be obtained by solving problem (OP2-1), in which $\bar{c}_i = g_i(\lceil m^{(p)} \rceil)$.
-

placed at equal distances apart, as shown in Figure 5-4. UEs are randomly, uniformly and independently distributed within this disk. The wireless transmission bandwidth is 10 MHz for each UE. We adopt the path loss model used by the 3GPP specification for Evolved Universal Terrestrial Radio Access in [135], where the received power at a UE d km from a RRH is given by

$$p \text{ (dB)} = 128.1 + 37.6 \log_{10} d.$$

The transmit antenna gain at each RRH is ϑ . The lognormal shadowing parameter is s . The difference between the normal RRHs (i.e., RRH 1 to 6) and the macro RRH, i.e., RRH 7, are:

1. the maximum transmitting power: $E_1 = E_2 = \dots = E_6 = E_n < E_7$,
2. the fronthaul capacity constraint: $S_1 = S_2 = \dots = S_6 = S_n < S_7$.

In our simulations, we consider homogeneous UEs with $\sigma_1 = \sigma_2 = \dots = \sigma_N = \sigma$, $\lambda_1 = \dots = \lambda_N = \lambda$, and $\tau_1 = \tau_2 = \dots = \tau_N = \tau$. We summarize our simulation parameters in Table 5.1.

Table 5.1: Simulation parameters in Chapter 5

Parameter	Value	Parameter	Value
φ	10	K	2
σ^2	-83.98 dBm	η	5
E_n	1 W	E_7	10 W
S_n	2	S_7	6
ϑ	5 dB	s	10 dB

Figure 5-4: Simulation setup in a heterogeneous C-RAN.

5.4.2 The benchmarks

In Section 5.3, we find out the number of active VMs and the sparse beamforming vectors by solving problem (OP1). The sparsity of the beamforming vector can be achieved by our proposed price adjusting (PA) algorithm, and the number of active VMs can be obtained by either SI or JO approaches. For ease of reference, we call this two different approaches as PASI and PAJO respectively.

It is noted that once the sparsity of the beamforming vector is obtained, the association relationship between the RRHs and the UEs is also determined. Specifically, $\|\mathbf{w}_{ij}^*\|_2 > 0$ implies that if and only if UE i is associated with RRH j . However, recently, [76] proposed a static clustering (SC) scheme, which helps obtaining the sparse beamforming vectors in two steps:

1. Utilizing the heuristic Algorithm 3 in [76] to obtain the UE set $\mathcal{N}_j$ that associates with RRH j , such that $\|\mathbf{w}_{ij}^*\|_2 > 0$ for $i \in \mathcal{N}_j$ and $\|\mathbf{w}_{ij}^*\|_2 = 0$ for $i \in \mathcal{N}_j^c$, where

$\mathcal{N}_j^c$ is the complementary set of $\mathcal{N}_j$.

2. Based on the given user association set $\mathcal{N}_j$, solving the following problem

$$\begin{aligned}
(\text{OP0-S}) \quad & \min_{m, c_i, \mathbf{w}_{ij}} \quad m\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
& \text{s.t.} \quad \frac{m}{m\mu - \alpha} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \quad \forall i \in \mathcal{N}, \\
& \quad \alpha < m\mu, \lambda_i < c_i, \quad \forall i \in \mathcal{N}, \\
& \quad 0 < m \leq M, \quad m \in \mathbb{N}, \\
& \quad c_i \leq B_i \log(1 + \text{SINR}_i), \quad \forall i \in \mathcal{N}, \\
& \quad \sum_{i=1}^N \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \quad \forall j \in \mathcal{L}, \\
& \quad \|\mathbf{w}_{ij}\|_2 = 0, \quad \forall i \in \mathcal{N}_j^c, j \in \mathcal{L}.
\end{aligned}$$

Thus, problem (OP0-S) is very similar with problem (OP2). Combing our proposed methods i.e., SI and JO, to solve problem (OP2), we have two benchmark algorithms, i.e., static clustering searching integer (SCSI) and static clustering joint optimization (SCJO) respectively.

5.4.3 Performance

In Figure 5-5, we present the system cost with different traffic rate and number of UEs under different algorithms. We show the relationship between UEs' mean arrival rate and system cost for $N = 8$ in Figure 5-5(a). We observe that, firstly, PA algorithms outperform the SC algorithms. That is, PAJO and PASI perform better than SCJO and SCSI respectively. Secondly, JO algorithms have very close performance with SI algorithms, i.e., the performance of PAJO and SCJO almost overlaps with PASI and SCSI respectively. Finally, as the incoming traffic rate increases, the performance gap between PA algorithms and SC algorithms becomes larger since the user association sets obtained from SC algorithms are oblivious from the UE's traffic rate. The performance of system cost versus the number of UEs is depicted in Figure 5-5(b) when

(a) Different arrival rates.

(b) Different number of UEs.

Figure 5-5: System cost under different algorithms.

$\lambda = 30$ Mb/s. We see again that PA algorithms are better than SC algorithms. In addition, for $N \geq 10$, SC algorithms are unable to guarantee the feasibility, which again because that the user association sets obtained from SC algorithms are oblivious from UE's incoming traffic rates and QoS requirements.

In Figure 5-6, we depict how different delay requirements affect system cost. It suggests that, firstly, SC algorithms cannot support stringent delay requirements,

Figure 5-6: Different delay requirements.

since the user association sets obtained from SC algorithms are oblivious from UEs' QoS requirements and incoming traffic rates. Secondly, the performance gap between PA algorithms and SC algorithms becomes closer with looser delay requirements. Finally, since SI algorithm achieves the global optimal solution for problem (OP2), while JO algorithm only obtains the local optimal solution, it can be verified in Figure 5-6 that SI algorithms have better performance than JO algorithms.

5.5 Summary

In this chapter, we jointly consider the VM activation and sparse beamforming problem in order to minimize the overall cost for a C-RAN with limited fronthaul capacity. In particular,

- we formulate the joint problem as a MINLP by minimizing the overall cost, which consists of two parts: the cost in the BBU pool w.r.t. the number of active VMs, and the cost (power consumption) on the RRHs w.r.t. the transmit beamformers.
- To avoid the feasibility problem caused by relaxing the l_0 -norm constraint directly, we reformulate the original MINLP into an equivalent problem, which

introduces a price vector. By adjusting the value of the price vector, the problem can be simplified into a subproblem. We propose two different approaches to solve this subproblem, i.e., searching integer approach and joint optimization approach.

- We provide simulation results that suggest that our proposed approach can provide better feasibility guarantee and obtain lower system cost than the recently proposed static clustering algorithm.

5.6 Proofs

5.6.1 Proof of Theorem 5.1

Based on $\{m^*, \mathbf{w}_{ij}^*\}$ and $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ are the optimal solutions for problem (OP0) and (OP1) respectively, we have

$$\begin{aligned}
& m(\Gamma)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\Gamma)\|_2^2 + \sum_{j=1}^L \gamma_j \beta_j(\Gamma) \\
& \leq m^* \varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}^*\|_2^2 + \sum_{i=1}^N \sum_{j=1}^L \gamma_j \|\mathbf{w}_{ij}^*\|_{2,0} \\
& \leq m^* \varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}^*\|_2^2 + \sum_{j=1}^L \gamma_j S_j \\
& \leq m(\Gamma)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\Gamma)\|_2^2 + \sum_{j=1}^L \gamma_j S_j,
\end{aligned}$$

where the first inequality is based on that $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ is the optimal solution for problem (OP1), the second inequality is based on constraint (5.9) in problem (OP0), the third inequality is based on that $\{c_i^*, \mathbf{w}_{ij}^*\}$ is the optimal solution for problem (OP0) and $\beta_j(\Gamma) = S_j$ (which implies $\{m(\Gamma), \mathbf{w}_{ij}(\Gamma)\}$ is also feasible for problem (OP0)). Then, let's substitute the equation $\beta_j(\Gamma) = S_j$ into the right hand side of

the third inequality above, we can have

$$m(\Gamma)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\Gamma)\|_2^2 = m^*\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}^*\|_2^2. \quad (5.26)$$

Therefore, the theorem is now proved.

5.6.2 Proof of Proposition 5.1

Let $\bar{\Gamma}'_j \triangleq [\bar{\gamma}'_1, \dots, \bar{\gamma}'_j, \dots, \bar{\gamma}'_L]^T$ be a different price vector with $\bar{\Gamma}_j$, such that $\gamma'_j > \gamma_j$ and $\bar{\gamma}'_k = \bar{\gamma}_k$, for $k \in \mathcal{L}/j$. We have

$$\begin{aligned} & m(\bar{\Gamma}_j)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\bar{\Gamma}_j)\|_2^2 + \gamma_j \beta_j(\bar{\Gamma}_j) + \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k \beta_k(\bar{\Gamma}_k) \\ & \leq m(\bar{\Gamma}'_j)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\bar{\Gamma}'_j)\|_2^2 + \gamma'_j \beta_j(\bar{\Gamma}'_j) + \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k \beta_k(\bar{\Gamma}'_k), \end{aligned}$$

and

$$\begin{aligned} & m(\bar{\Gamma}'_j)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\bar{\Gamma}'_j)\|_2^2 + \gamma'_j \beta_j(\bar{\Gamma}'_j) + \sum_{k \in \mathcal{L}/j} \bar{\gamma}'_k \beta_k(\bar{\Gamma}'_k) \\ & \leq m(\bar{\Gamma}_j)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\bar{\Gamma}_j)\|_2^2 + \gamma'_j \beta_j(\bar{\Gamma}_j) + \sum_{k \in \mathcal{L}/j} \bar{\gamma}'_k \beta_k(\bar{\Gamma}_k). \end{aligned}$$

Adding up both sides of the two inequalities above and simplifying it, we have

$$(\bar{\gamma}'_j - \bar{\gamma}_j) \beta_j(\bar{\Gamma}'_j) \leq (\bar{\gamma}'_j - \bar{\gamma}_j) \beta_j(\bar{\Gamma}_j).$$

Hence, the first statement is now proved.

We denote $\{\hat{m}, \hat{\mathbf{w}}_{ij}\}$ as a feasible solution for problem (OP0), whose feasible region

is nonempty. Then, we have

$$\begin{aligned}
& m(\bar{\Gamma}_j)\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\mathbf{w}_{ij}(\bar{\Gamma}_j)\|_2^2 + \gamma_j \beta_j(\bar{\Gamma}_j) + \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k \beta_k(\bar{\Gamma}_j) \\
& \leq \hat{m}\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\hat{\mathbf{w}}_{ij}\|_2^2 + \gamma_j \sum_{i=1}^N \|\hat{\mathbf{w}}_{ij}\|_{2,0} + \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k \sum_{i=1}^N \|\hat{\mathbf{w}}_{ik}\|_{2,0},
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
& \beta_j(\bar{\Gamma}_j) - \sum_{i=1}^N \|\hat{\mathbf{w}}_{ij}\|_{2,0} \\
& < (\hat{m}\varphi + \eta \sum_{i=1}^N \sum_{j=1}^L \|\hat{\mathbf{w}}_{ij}\|_2^2 + \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k \sum_{i=1}^N \|\hat{\mathbf{w}}_{ik}\|_{2,0}) / \gamma_j \\
& \leq (M\varphi + \eta \sum_{j=1}^L E_j + \sum_{k \in \mathcal{L}/j} \bar{\gamma}_k S_k) / \gamma_j.
\end{aligned}$$

Therefore, if $\gamma_j \geq \theta_j$, then $\beta_j(\bar{\Gamma}_j) - \sum_{i=1}^N \|\hat{\mathbf{w}}_{ij}\|_{2,0} < 1$. Since $\beta_j(\bar{\Gamma}_j)$ and $\sum_{i=1}^N \|\hat{\mathbf{w}}_{ij}\|_{2,0}$ are both integers, then we have $\beta_j(\bar{\Gamma}_j) = \sum_{i=1}^N \|\hat{\mathbf{w}}_{ij}\|_{2,0} \leq S_j$.

This completes the proof.

Chapter 6

Conclusion and Future Work

In this chapter, we conclude our work firstly, and then propose some potential future research directions.

6.1 Conclusion

In this dissertation, we studied the elastic service scaling problem in CCMN and C-RAN, which are promising solutions for core network and RAN respectively. We summarize the research results as follows:

Since the emergence of cloud computing technologies allows the design of elastic cost-aware user redirection mechanisms that scale flexibly with the arrival traffic, in Chapter 3, we derived optimal redirection policies under a cloud-centric framework, by jointly detecting the arrival traffic mode and adapting the allocation and service capacities accordingly. We show that the optimal redirection policy involves choosing those VMs with the lowest prices, up to a threshold price. We also show the approach to compute the optimal allocation and service capacities of the active VMs. The simulation results show that the proposed mechanism performs better than other benchmark strategies.

In Chapter 4, we investigated the problem of minimizing the overall system power consumption (including the power consumption in the BBU pool, the fiber links and the RRHs) in a C-RAN, such that the cross-layer QoS and per-RRH power

constraints are satisfied. We formulated a MINLP and then relax it to an ESUM problem to solve the cross-layer resource allocation problem, which gives the optimal achievable rate for each user. Based on the optimal achievable rate, we propose an efficient SP algorithm, with lower computational complexity than several state-of-the-art RRH selection methods, to recover a sparse solution for the RRH selection problem. Simulation results suggest that our proposed SP algorithm outperforms various other methods, and the proposed cross-layer algorithm is more energy efficient than existing decoupled-layer methods.

In Chapter 5, we considered the joint VM activation and sparse beamforming problem in heterogeneous C-RAN, which has limited fronthaul capacity. We aim to minimize the system cost of C-RAN, including VM cost (w.r.t. the number of active VMs) in the BBU pool and RRH cost (w.r.t. the beamformer vectors). To tackle the limited fronthaul capacity constraint, we propose a price adjusting algorithm; To find out the optimal number of VMs, we proposed two different algorithms: searching integer and joint optimization. Simulation results suggest that our proposed algorithms have more robust performance and lower system cost than the newly proposed static clustering algorithms.

6.2 Future work

We summarize some potential future works and the challenges therein in this section.

6.2.1 Embedding mobile cloud computing in C-RAN

Mobile cloud computing (MCC) emerged as the combination of mobile computing and cloud computing. Specifically, it is defined as [140] “Mobile cloud computing at its simplest, refers to an infrastructure where both the data storage and the data processing happen outside of the mobile device. Mobile cloud applications move the computing power and data storage away from mobile phones and into the cloud, bringing applications and mobile computing to not just smartphone users but a much broader range of mobile subscribers.” By utilizing MCC, we can improve data storage

capacity, enhance processing power and extend battery lifetime.

C-RAN has the cloud data center (i.e., the BBU pool), which is normally much nearer to the UEs than the public cloud service providers. For the C-RAN service providers, it is possible to utilize the cloud data center to offer the MCC as a service (MCCaaS) as well. The benefit for service providers are multi-fold: Firstly, C-RAN service providers can rent their redundant computation resources to the third party applications to make profit. Secondly, C-RAN service providers can also utilize their computation resources in the BBU pool to offer MCC service to their own subscribers to improve the subscribers's quality-of-experience (QoE) [141,142], e.g., extending the battery lifetime.

To achieve the aforementioned visions, there are some research works need to be done for each scenario:

- To maximize the C-RAN service providers profit, how to optimally segment the computation resources to the third party application providers? A related work for capacity segmentation in public cloud can be found in [143], however, for C-RAN and MCC, the features of wireless transmission have to be taken into consideration.
- How to design the offloading strategies to improve the subscribers' QoE, while also minimizing the overhead caused, e.g., system power consumption for C-RAN? Most of the research works on MCC are just in a selfish manner, i.e., they just consider saving UEs' power consumption, but the power consumption (overhead) in the RAN is usually ignored [126,144].

6.2.2 Caching as a Service (CaaS) in C-RAN

In CDN, there are two basic problems: request redirection and replica placement. We discussed the request redirection problem in CCMN in Chapter 3 and the replica placement problem in CDN and cloud CDN has been studied in [44–46, 54] as well. However, no matter in conventional CDN or cloud CDN, the replicas or contents are still in the core network. As data and content consumers in the Internet is proliferating

very fast, the current cloud CDN or CCMN structure is becoming arduous for the huge amount of (mobile) traffic. Specifically, since CDN is deployed in the core layer of network, the huge amount of traffic transmitted from core layer is causing increasing pressure to the aggregation layer and (radio) access layer, which normally have much less bandwidth than the core layer.

To reduce the amount of traffic in the aggregation layer and the backhaul of radio access layer, and also provide high QoS guarantee to mobile users, “Femoto-Caching” [145–148] or “Caching in the RAN” [80, 81, 149, 150] is attracting increasing research attentions. Owing to the nature of C-RAN, it is easy to extend one more functionality of C-RAN: Caching as a Service (CaaS). That means we can just cache the content in the BBU pool, which consists of many general purpose servers. Then the traffic load on the backhaul links can be reduced accordingly.

There are some potential issues to be considered in CaaS:

- CaaS aims to push the content from the edge of core network to the access network, the contents are, hence, distributed in a hierarchical manner, i.e., a portion of the contents should be placed in the BBU pool while some of the contents are still in the core network only. How to design an optimal hierarchical caching strategy to minimize the content retrieval cost while also guarantee the QoS of mobile communication?
- One of the main features of CaaS is that the traffic load in the RAN backhaul is reduced. However, the huge amount of traffic still needs to go through the C-RAN fronthaul, whose capacity is limited. Then a cross-layer strategy should be proposed to minimize the maximum UE’s system delay (i.e., delay in the fronthaul + delay in the backhaul).

6.2.3 Statistical QoS provisioning

As the increasing popularity of real time applications, e.g., VoIP, streaming video and online gaming, the operators are facing more stringent challenges, i.e., providing bounded delay or guaranteed service bandwidth for these applications. However, it is

almost impossible to guarantee the hard delay bound for every packet in the network, especially in the wireless network, because of the dynamical nature of network traffic and wireless channel conditions. Therefore, instead of providing the hard delay bound or the average delay bound (like we did in Chapter 4 and 5), people consider an alternative way of guaranteeing the delay bound violation probability, a method of statistical QoS provisioning.

To set up the statistical QoS provisioning model, we can make use of the preliminary knowledge in the following part.

Effective bandwidth and effective capacity

The concept of *effective bandwidth* was proposed, based on large deviation theory, and extensively studied in the early 1990s, in order to obtain statistical QoS guarantees over the wireline asynchronous transfer mode (ATM) networks [151–153]. Let the amount of arrival data (in bits) to the channel i , over the time interval $[0, t)$, be a random process $\Delta_i(t)$. Assuming that the asymptotic log-moment generating function of $\Delta_i(t)$, expressed as $\Lambda_i(z_i) = \lim_{t \rightarrow \infty} (1/t) \log \mathbb{E}[e^{z_i \Delta_i(t)}]$, exists for all $z_i \geq 0$. Then the effective bandwidth function of the random arrival process $\Delta_i(t)$ is defined as,

$$\alpha_i(z_i) = \frac{\Lambda_i(z_i)}{z_i} = \lim_{t \rightarrow \infty} \frac{1}{z_i t} \log \mathbb{E}[e^{z_i \Delta_i(t)}], \quad (6.1)$$

where $\mathbb{E}$ denotes the expectation, $z_i \geq 0$ is defined as the QoS exponent for the channel i .

For the channel with random input $\Delta_i(t)$ and fixed service rate $\mu_i \geq 0$, let $\tilde{D}_i$ be the steady state delay of the arrival packets in channel i and D_{max} be the delay bound. Then the probability of $\tilde{D}_i$ exceeding D_{max} satisfies

$$\Pr\{\tilde{D}_i \geq D_{max}\} \approx e^{-\mu_i \alpha_i^{-1}(\mu_i) D_{max}}, \quad (6.2)$$

where $\alpha_i^{-1}(\mu_i) = z_i$ is the inverse function of the effective bandwidth $\alpha_i(z_i) = \mu_i$. This shows the probability that the delay exceeds a certain threshold D_{max} decays

exponentially fast with the increasing of the threshold.

The follow-up theory in the wireless channel model is *effective capacity* [121], which is the dual theory of effective bandwidth, under fixed arrival traffic rate with random service process. The authors in [121] derived the similar formulation for the delay-bound violation probability as (6.2) in the wireless channel based on effective capacity. In [154], effective bandwidth and effective capacity theory are jointly considered for the channel with both random arrival process and random service process. The effective bandwidth and effective capacity theories has been widely adopted and applied [85, 155, 156].

Potential issues

Based on the preliminary knowledge above, we can utilize the concept of effective bandwidth and effective capacity to construct the optimization problem. However, the term of delay-bound violation probability (6.2) is usually hard to tackle with. One possible way is to approximate the term in a posynomial form and then utilize the geometric programming (GP) [157, 158] method to solve the problem. However, solving a GP is still time consuming. Therefore, proposing a simple but efficient algorithm to solve the statistical QoS provisioning becomes one of the potential issues.

As we utilize effective bandwidth and effective capacity as the basis of statistical QoS provisioning, if we aim to provide cross-layer statistical QoS guarantee, then the cross-layer statistical QoS constraint should be based on the convolution of two single layer QoS constraint terms. Therefore, it is very hard to decouple the main problem into subproblems (like we did in Chapter 4). This is another potential challenge.

Other than the perspective of effective bandwidth and effective capacity, there is another practical issue which also requires statistical QoS provisioning. Specifically, all works in this dissertation is based on the assumption of unlimited buffer space for each queue. However, if the buffer space is limited, then there would be a packet loss probability in each queue. How to maintain a low packet loss probability in the CBCS is also a potential research direction in the future.

Bibliography

- [1] Cisco Systems, “Cisco visual networking index: Forecast and methodology, 2013-2018,” Cisco Systems, Technical Report, 2014.
- [2] H. Yin, X. Liu, F. Qiu, N. Xia, C. Lin, H. Zhang, V. Sekar, and G. Min, “Inside the bird’s nest: measurements of large-scale live VoD from the 2008 olympics,” in *Proc. ACM SIGCOMM IMC*, Chicago, IL, 2009, pp. 442–455.
- [3] P. Wendell and M. J. Freedman, “Going viral: flash crowds in an open CDN,” in *Proc. ACM SIGCOMM IMC*, Berlin, Germany, Nov. 2011, pp. 549–558.
- [4] A.-J. Su, D. R. Choffnes, A. Kuzmanovic, and F. E. Bustamante, “Drafting behind Akamai: Inferring network conditions based on CDN redirections,” *IEEE/ACM Trans. Netw.*, vol. 17, no. 6, pp. 1752–1765, Dec. 2009.
- [5] C. Kaufmann, “Peering with content distribution networks,” Akamai Technologies, Technical Report, 2009.
- [6] CDN Reviews, “Advantages and disadvantages of using a content delivery network,” 2012. [Online]. Available: <http://www.cdnreviews.com/cdn-advantages-disadvantages/>
- [7] Akamai Technologies, 2013. [Online]. Available: <http://www.akamai.com/>
- [8] Y. Chen, C. Chen, and D. Guo, “A measurement study of online internet video views through a large-scale search engine,” in *Proc. CyberC*, Zhangjiajie, China, Oct. 2009, pp. 180–183.
- [9] S. Ranjan and E. Knightly, “High-performance resource allocation and request redirection algorithms for web clusters,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 9, pp. 1186–1200, Sep. 2008.
- [10] F. L. Presti, C. Petrioli, and C. Vicari, “Distributed dynamic replica placement and request redirection in content delivery networks,” in *Proc. IEEE/ACM MASCOTS*, Istanbul, Turkey, Oct. 2007, pp. 366–373.
- [11] M. Kasbekar, “On efficient delivery of web content,” Akamai Technologies, Technical Report, 2010.

- [12] R. Buyya, M. Pathan, and A. Vakali, *Content Delivery Networks*. Berlin, Germany: Springer, 2008.
- [13] China Mobile Research Institute, “C-RAN: The road towards green RAN,” China Mobile Research Institute, White Paper, V2.5, Oct. 2011.
- [14] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. Mayer, L. Thiele, and V. Jungnickel, “Coordinated multipoint: Concepts, performance, and field trial results,” *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.
- [15] A. Khandekar, N. Bhushan, T. Ji, and V. Vanghi, “Lte-advanced: Heterogeneous networks,” in *European Wireless Conference (EW)*, Lucca, Italy, Apr. 2010, pp. 978–982.
- [16] P. Mell and T. Grance, “The NIST definition of cloud computing,” NIST, Gaithersburg, MD, Tech. Rep., Sep. 2011.
- [17] Amazon Web Services, “What is cloud computing?” 2014. [Online]. Available: <http://aws.amazon.com/what-is-cloud-computing/>
- [18] U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, “A cost-aware elasticity provisioning system for the cloud,” in *Proc. IEEE ICDCS*, Minneapolis, MN, Jun. 2011, pp. 559–570.
- [19] Y. Jin, Y. Wen, G. Shi, G. Wang, and A. Vasilakos, “CoDaaS: An experimental cloud-centric content delivery platform for user-generated contents,” in *Proc. IEEE ICNC*, Maui, HI, Jan. 2012, pp. 934–938.
- [20] Y. Wen, G. Shi, and G. Wang, “Designing an inter-cloud messaging protocol for content distribution as a service (codaas) over future internet,” in *Proc. ACM CFI*, Seoul, Korea, Jun. 2011, pp. 91–93.
- [21] Y. Wen, “Cloud-centric media network (CCMN): Empowering rich media experience with social interactive features,” oct. 2011. [Online]. Available: http://pdcc.ntu.edu.sg/cloud11/slides/talk5_pm.pdf
- [22] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, “Cloud RAN for mobile networks - a technology overview,” *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, pp. 1–24, 2014.
- [23] K. D. Devine, E. G. Boman, R. T. Heaphy, B. A. Hendrickson, J. D. Teresco, J. Faik, J. E. Flaherty, and L. G. Gervasio, “New challenges in dynamic load balancing,” *Applied Numerical Mathematics*, vol. 52, pp. 133–152, Feb. 2005.
- [24] K. Gilly, C. Juiz, and R. Puigjaner, “An up-to-date survey in web load balancing,” *World Wide Web*, vol. 14, no. 2, pp. 105–131, 2011.

- [25] A. Barbir, B. Cain, R. Nair, and O. Spatscheck, "Known content network (CN) request-routing mechanisms," United States, Tech. Rep., 2003, RFC 3568.
- [26] A. Barak and A. Shiloh, "A distributed load-balancing policy for a multicomputer," *Software Practice Experience*, vol. 15, no. 9, pp. 901–913, Sep. 1985.
- [27] A. Barak, A. Shiloh, and R. Wheeler, "Flood prevention in the mosix load-balancing scheme," *IEEE-TCOS Newsletter*, vol. 3, no. 1, pp. 24–27, Winter 1989.
- [28] A. Goscinski and M. Bearman, "Resource management in large distributed systems," *SIGOPS Oper. Syst. Rev.*, vol. 24, no. 4, pp. 7–25, Sep. 1990.
- [29] S. Manfredi, F. Oliviero, and S. P. Romano, "A distributed control law for load balancing in content delivery networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 55–68, Feb. 2013.
- [30] N. R. Mahapatra and S. Dutt, "Random seeking: a general, efficient, and informed randomized scheme for dynamic load balancing," in *Proc. IEEE IPDS*, Honolulu, HI, Apr. 1996, pp. 881–885.
- [31] A. Anagnostopoulos, A. Kirsch, and E. Upfal, "Stability and efficiency of a random local load balancing protocol," in *Proc. IEEE FOCS*, Cambridge, MA, Oct. 2003, pp. 472–481.
- [32] M. Shreedhar and G. Varghese, "Efficient fair queuing using deficit round-robin," *IEEE/ACM Trans. Netw.*, vol. 4, no. 3, pp. 375–385, Jun. 1996.
- [33] M. Jaseemuddin, A. Nanthakumaran, and A. Leon-Garcia, "TE-Friendly content delivery request routing in a CDN," in *Proc. IEEE ICC*, Istanbul, Turkey, Jun. 2006, pp. 323–330.
- [34] D.-C. Li and F. M. Chang, "An in-out combined dynamic weighted round-robin method for network load balancing," *The Computer Journal*, vol. 50, no. 5, pp. 555–566, Sep. 2007.
- [35] M. Dahlin, "Interpreting stale load information," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 10, pp. 1033–1047, Oct. 2000.
- [36] R. L. Carter and M. E. Crovella, "Server selection using dynamic path characterization in wide-area networks," in *Proc. IEEE INFOCOM*, vol. 3, Kobe, Japan, Apr. 1997, pp. 1014–1021.
- [37] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 12, no. 10, pp. 1094–1104, Oct. 2001.
- [38] C.-M. Chen, Y. Ling, M. Pang, W. Chen, S. Cai, Y. Suwa, and O. Altintas, "Scalable request routing with next-neighbor load sharing in multi-server environments," in *Proc. IEEE AINA*, Tamsui, Taiwan, Mar. 2005, pp. 441–446.

- [39] T. Minyard, Y. Kallinderis, and K. Schulz, "Parallel load balancing for dynamic execution environments," in *Comput. Methods Appl. Mech. Engrg*, 1996, pp. 96–0295.
- [40] J. D. Teresco, M. W. Beall, J. E. Flaherty, and M. S. Shephard, "A hierarchical partition model for adaptive finite element computation," *Comput. Methods Appl. Mech. Engrg*, vol. 184, pp. 269–285, 1998.
- [41] C. Walshaw and M. Cross, "Multilevel mesh partitioning for heterogeneous communication networks," *Future Generation Comput. Syst*, vol. 17, pp. 601–623, 2001.
- [42] F. L. Presti, N. Bartolini, and C. Petrioli, "Dynamic replica placement and user request redirection in content delivery networks," in *Proc. IEEE ICC*, vol. 3, May 2005, pp. 1495–1501.
- [43] C. Vicari, C. Petrioli, and F. L. Presti, "Dynamic replica placement and traffic redirection in content delivery networks," *SIGMETRICS Performance Evaluation Review*, vol. 35, pp. 66–68, Dec. 2007.
- [44] X. Qiu, H. Li, C. Wu, Z. Li, and F. C. M. Lau, "Cost-minimizing dynamic migration of content distribution services into hybrid clouds," in *Proc. IEEE INFOCOM*, Orlando, FL, Mar. 2012, pp. 2571–2575.
- [45] M. M. Amble, P. Parag, S. Shakkottai, and L. Ying, "Content-aware caching and traffic management in content distribution networks," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 2858–2866.
- [46] M. Bjorkqvist, L. Y. Chen, and X. Zhang, "Minimizing retrieval cost of multi-layer content distribution systems," in *Proc. IEEE ICC*, Kyoto, Japan, Jun. 2011, pp. 1–6.
- [47] M. Bjorkqvist, L. Y. Chen, M. Vukolic, and X. Zhang, "Minimizing retrieval latency for content cloud," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 1080–1088.
- [48] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 1098–1106.
- [49] A. Qureshi, R. Weber, H. Balakrishnan, J. Gutttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *Proc. ACM SIGCOMM*, Barcelona, Spain, Aug. 2009, pp. 123–134.
- [50] V. Mathew, R. K. Sitaraman, and P. Shenoy, "Energy-aware load balancing in content delivery networks," in *Proc. IEEE INFOCOM*, Orlando, FL, Mar. 2012, pp. 954–962.

- [51] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, “Greening geographical load balancing,” in *Proc. ACM SIGMETRICS*, San Jose, CA, Jun. 2011, pp. 233–244.
- [52] X. Nan, Y. He, and L. Guan, “Optimal resource allocation for multimedia cloud based on queuing model,” in *Proc. IEEE MMSP Workshop*, Hangzhou, China, Oct. 2011, pp. 1–6.
- [53] L. Chen, N. Li, and S. H. Low, “On the interaction between load balancing and speed scaling,” in *Proc. ITA Workshop*, San Diego, CA, Feb. 2011.
- [54] F. Chen, K. Guo, J. Lin, and T. La Porta, “Intra-cloud lightning: Building CDNs in the cloud,” in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2012, pp. 433–441.
- [55] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, “Toward green and soft: a 5G perspective,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.
- [56] W. H. Chin, Z. Fan, and R. Haines, “Emerging technologies and research challenges for 5G wireless networks,” *IEEE Wireless Commun. Mag.*, vol. 21, no. 2, pp. 106–112, Apr. 2014.
- [57] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, “Cloud technologies for flexible 5G radio access networks,” *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.
- [58] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
- [59] B. Sourjya, C. S. Preeth, J. M. Kashyap, K. Gautam, M. Anand, P. Paul, S. Vikram, and W. Thomas, “CloudIQ: A framework for processing base stations in a data center,” in *Proc. ACM MobiCom*, Istanbul, Turkey, Aug. 2012, pp. 125–136.
- [60] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, “FluidNet: A flexible cloud-based radio access network for small cells,” in *Proc. ACM MobiCom*, Miami, FL, Sep. 2013, pp. 99–110.
- [61] Y. Mao, L. Yong, J. Depeng, S. Li, M. Shaowu, and Z. Lieguang, “OpenRAN: A software-defined RAN architecture via virtualization,” in *Proc. ACM SIGCOMM*, Hong Kong, China, Aug. 2013, pp. 549–550.
- [62] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, “The case for re-configurable backhaul in cloud-RAN based small cell networks,” in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1124–1132.

- [63] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. D. Girolamo, and G. Giuliani, "RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network," in *Proc. Future Network and Mobile Summit*, Lisbon, Portugal, Jul. 2013, pp. 1–8.
- [64] K. Wang, M. Zhao, and W. Zhou, "Traffic-aware graph-based dynamic frequency reuse for heterogeneous Cloud-RAN," in *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 2343–2348.
- [65] M. C. Valenti, S. Talarico, and P. Rost, "The role of computational outage in dense cloud-based centralized radio access networks," in *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 1489–1495.
- [66] J. Bartelt, G. Fettweis, D. Wubben, M. Boldi, and B. Melis, "Heterogeneous backhaul for cloud-based mobile networks," in *Proc. IEEE VTC*, Las Vegas, NV, Sep. 2013, pp. 1–5.
- [67] Y. Shi, J. Zhang, and K. B. Letaief, "CSI overhead reduction with stochastic beamforming for cloud radio access networks," in *Proc. IEEE ICC*, Sydney, Australia, Jun. 2014, pp. 5165–5170.
- [68] C. Fan, Y. J. A. Zhang, and X. Yuan, "Scalable coordinated uplink processing in cloud radio access networks," in *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 4101–4106.
- [69] Y. Cai, F. Yu, and S. Bu, "Cloud radio access networks (C-RAN) in mobile cloud computing systems," in *Proc. IEEE INFOCOM Workshops*, Toronto, ON, Apr. 2014, pp. 369–374.
- [70] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.
- [71] B. Dai and W. Yu, "Sparse beamforming for limited-backhaul network mimo system via reweighted power minimization," in *Proc. IEEE GLOBECOM*, Atlanta, GA, Dec. 2013, pp. 1962–1967.
- [72] F. Zhuang and V. K. N. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, Feb. 2014.
- [73] V. N. Ha, L. B. Le, and N.-D. Dao, "Energy-efficient coordinated transmission for Cloud-RANs: Algorithm design and trade-off," in *Proc. Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, Mar. 2014, pp. 1–6.

- [74] —, “Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints,” in *Proc. IEEE WCNC*, Istanbul, Turkey, Apr. 2014, pp. 1862–1867.
- [75] V. N. Ha and L. B. Le, “Joint coordinated beamforming and admission control for fronthaul constrained cloud-rans,” in *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 4397–4402.
- [76] B. Dai and W. Yu, “Sparse beamforming and user-centric clustering for downlink cloud radio access network,” *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.
- [77] L. Liu, S. Bi, and R. Zhang, “Joint power control and fronthaul rate allocation for throughput maximization in OFDMA-based cloud radio access network,” *IEEE Trans. Wireless Commun.*, submitted.
- [78] Y. Zhou and W. Yu, “Optimized backhaul compression for uplink cloud radio access network,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.
- [79] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, “Joint precoding and multi-variate backhaul compression for the downlink of cloud radio access networks,” *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov 2013.
- [80] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, “Joint data assignment and beamforming for backhaul limited caching networks,” in *Proc. IEEE PIMRC*, Washington, DC, Sep. 2014, pp. 1370–1374.
- [81] D. Liu and C. Yang, “Will caching at base station improve energy efficiency of downlink transmission?” in *Proc. IEEE GlobalSIP*, Atlanta, GA, Dec. 2014, pp. 328–332.
- [82] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, “Robust and efficient distributed compression for cloud radio access networks,” *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.
- [83] D. Feng, C. Jiang, G. Lim, L. J. J. Cimini, G. Feng, and G. Y. Li, “A survey of energy-efficient wireless communications,” *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, pp. 167–178, 2013.
- [84] Y. Zeng, E. Gunawan, Y. L. Guan, and J. Liu, “Joint base station selection and linear precoding for cellular networks with multi-cell processing,” in *Proc. IEEE TENCON*, Fukuoka, Japan, Nov. 2010, pp. 1976–1981.
- [85] Q. Du and X. Zhang, “QoS-aware base-station selections for distributed MIMO links in broadband wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1123–1138, Jun. 2011.

- [86] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3214–3228, Jun. 2013.
- [87] D. Amzallag, R. Bar-Yehuda, D. Raz, and G. Scalosub, "Cell selection in 4G cellular networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1443–1455, Jul. 2013.
- [88] C.-P. Chien, K.-M. Yang, and H.-Y. Hsieh, "Selection of transmission points for delay minimization in LTE-A heterogeneous networks with low-power RRHs," in *Proc. IEEE WCNC*, Shanghai, China, Apr. 2013, pp. 783–788.
- [89] Y. Cheng, M. Pesavento, and A. Philipp, "Joint network optimization and downlink beamforming for CoMP transmissions using mixed integer conic programming," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, Aug. 2013.
- [90] A. Liu and V. Lau, "Joint power and antenna selection optimization for energy-efficient large distributed MIMO networks," in *Proc. IEEE ICCS*, Singapore, Nov. 2012, pp. 230–234.
- [91] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.
- [92] S. Luo, R. Zhang, and T. J. Lim, "Coordinated downlink and uplink user association and beamforming for energy minimization in cloud radio access network," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 459–463.
- [93] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.
- [94] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1426–1438, Aug. 2006.
- [95] P. Weeraddana, M. Codreanu, M. Latva-aho, and A. Ephremides, "Weighted sum-rate maximization for a set of interfering links via branch and bound," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3977–3996, Aug. 2011.
- [96] S. Joshi, P. Weeraddana, M. Codreanu, and M. Latva-aho, "Weighted sum-rate maximization for MISO downlink cellular networks via branch and bound," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 2090–2095, Apr. 2012.
- [97] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a mimo interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

- [98] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear transceiver design for a MIMO interfering broadcast channel achieving max-min fairness," in *Proc. IEEE ASILOMAR*, Nov. 2011, pp. 1309–1313.
- [99] W.-C. Liao, M. Hong, H. Farmanbar, X. Li, Z.-Q. Luo, and H. Zhang, "Min flow rate maximization for software defined radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1282–1294, Jun. 2014.
- [100] S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
- [101] T. Bogale and L. Vandendorpe, "Weighted sum rate optimization for downlink multiuser MIMO coordinated base station systems: Centralized and distributed algorithms," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1876–1889, Apr. 2012.
- [102] H. Kha, H. Tuan, and H. Nguyen, "Fast global optimal power allocation in wireless networks by local d.c. programming," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 510–515, Feb. 2012.
- [103] E. Che and H. Tuan, "Sum-rate based coordinated beamforming in multicell multi-antenna wireless networks," *IEEE Commun. Lett.*, vol. 18, no. 6, pp. 1019–1022, Jun. 2014.
- [104] Y. J. A. Zhang, L. Qian, and J. Huang, "Monotonic optimization in communication and networking systems," *Foundations and Trends in Networking*, vol. 7, no. 1, pp. 1–75, 2012.
- [105] L. P. Qian, Y. Zhang, and J. Huang, "MAPEL: Achieving global optimality for a non-convex wireless power control problem," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1553–1563, Mar. 2009.
- [106] L. Zheng and C. W. Tan, "Maximizing sum rates in cognitive radio networks: Convex relaxation and global optimization algorithms," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 667–680, Mar. 2014.
- [107] C. W. Tan, "Wireless network optimization by perron-frobenius theory," in *Proc. Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, Mar. 2014, pp. 1–6.
- [108] J. Gray, B. Good, D. Gawlick, P. Homan, and H. Sammer, "One thousand transactions per second," in *Proc. IEEE COMPCON*, Feb. 1985, pp. 96–101.
- [109] H. Poor and O. Hadjiladis, *Quickest Detection*. Cambridge, U.K.: Cambridge University Press, 2008.

- [110] I. Ari, B. Hong, E. L. Miller, S. A. Brandt, and D. D. E. Long, “Managing flash crowds on the Internet,” in *Proc. IEEE/ACM MASCOTS*, Orlando, FL, Oct. 2003, pp. 246–249.
- [111] Y. Chen, B. Zhang, and C. Chen, “Modeling and performance analysis of P2P live streaming systems under flash crowds,” in *Proc. IEEE ICC*, Kyoto, Jun. 2011, pp. 1–5.
- [112] K. Hosanagar, R. Krishnan, M. D. Smith, and J. Chuang, “Optimal pricing of content delivery network (CDN) services,” in *Proc. IEEE HICSS*, Hawaii, HI, Jan. 2004, pp. 1–10.
- [113] Amazon Elastic Compute Cloud, 2013. [Online]. Available: <http://aws.amazon.com/ec2/>
- [114] S. Kaxiras and M. Martonosi, *Computer Architecture Techniques for Power-Efficiency*, 1st ed. Morgan and Claypool Publishers, 2008.
- [115] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. New Jersey, U.S.: Prentice Hall, 1992.
- [116] G. Lorden, “Procedures for reacting to a change in distribution,” *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.
- [117] G. V. Moustakides, “Optimal stopping times for detecting changes in distributions,” *The Annals of Statistics*, vol. 14, no. 4, pp. pp. 1379–1387, 1986.
- [118] H. V. Poor, *An introduction to signal detection and estimation (2nd ed.)*. New York, NY, USA: Springer-Verlag New York, Inc., 1994.
- [119] UMassTraceRepository, 2013. [Online]. Available: <http://traces.cs.umass.edu/index.php/Network/Network>
- [120] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*, ser. Athena Scientific optimization and computation series. Athena Scientific, 2003.
- [121] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
- [122] P. J. Burke, “The output of a queuing system,” *Operations Research*, vol. 4, no. 6, pp. 699–704, Dec. 1956.
- [123] E. Reich, “Waiting times when queues are in tandem,” *The Annals of Mathematical Statistics*, vol. 28, no. 3, pp. 768–773, Sep. 1957.
- [124] D. W. H. Cai, T. Q. S. Quek, C. W. Tan, and S. H. Low, “Max-min SINR coordinated multipoint downlink transmission — duality and algorithms,” *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5384–5395, Oct. 2012.

- [125] J. Tang, W. P. Tay, and Y. Wen, "Dynamic request redirection and elastic service scaling in cloud-centric media networks," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1434–1445, Aug. 2014.
- [126] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.
- [127] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud radio access networks," in *Proc. IEEE GLOBECOM*, Atlanta, GA, 2013, pp. 4635–4640.
- [128] A. Wiesel, Y. C. Eldar, and S. S. (Shitz), "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.
- [129] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.
- [130] X. Zheng, F. Chen, Y. Xia, and Y. Fang, "A class of cross-layer optimization algorithms for performance and complexity trade-offs in wireless networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 20, no. 10, pp. 1393–1407, Oct. 2009.
- [131] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0," <http://cvxr.com/cvx>, Aug. 2012.
- [132] S. Boyd and J. Mattingley, "Branch-and-bound methods," Mar. 2007. [Online]. Available: http://see.stanford.edu/materials/lsoctee364b/17-bb_notes.pdf
- [133] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.
- [134] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, Mar. 2011.
- [135] 3GPP, "LTE; Evolved universal terrestrial radio access (E-UTRA); Radio frequency (RF) requirements for LTE Pico Node B (release 9)," 3rd Generation Partnership Project (3GPP), TS 36.931, May 2011, v9.0.0.
- [136] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation in cloud radio access network," in *Proc. IEEE GlobalSIP*, Atlanta, GA, Dec. 2014, pp. 313–317.
- [137] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

- [138] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed l^0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, Jan. 2009.
- [139] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [140] Mobile Cloud Computing World Forum, 2012. [Online]. Available: <http://www.mobilecloudcomputingforum.com/>
- [141] T. Ebrahimi, "Quality of multimedia experience: past, present and future," in *Proc. ACM MM*, Beijing, China, Oct. 2009, pp. 3–4.
- [142] U. Reiter, "Quality of experience: a buzzword or the key to successful multimedia delivery across networks?" in *Proc. ACM LCNC*, Quito, Ecuador, Oct. 2011, pp. 20–24.
- [143] W. Wang, B. Li, and B. Liang, "Towards optimal capacity segmentation with hybrid cloud pricing," in *Proc. IEEE ICDCS*, Macau, China, Jun. 2012, pp. 425–434.
- [144] W. Zhang, Y. Wen, and D. O. Wu, "Collaborative task execution in mobile cloud computing under a stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 81–93, Jan. 2015.
- [145] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "Femto-caching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Shanghai, China, Mar. 2012, pp. 1107–1115.
- [146] H. AhleHagh and S. Dey, "Video caching in radio access network: Impact on delay and capacity," in *Proc. IEEE WCNC*, Paris, France, Apr. 2012, pp. 2276–2281.
- [147] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femto-caching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
- [148] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
- [149] Y. Xu, Y. Li, Z. Wang, T. Lin, G. Zhang, and S. Ci, "Coordinated caching model for minimizing energy consumption in radio access network," in *Proc. IEEE ICC*, Sydney, Australia, Jun. 2014, pp. 2406–2411.
- [150] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

- [151] C.-S. Chang, “Stability, queue length, and delay of deterministic and stochastic queueing networks,” *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.
- [152] C.-S. Chang and J. Thomas, “Effective bandwidth in high-speed digital networks,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.
- [153] G. Choudhury, D. Lucantoni, and W. Whitt, “Squeezing the most out of ATM,” *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 203–217, Feb. 1996.
- [154] J. Tang and X. Zhang, “Cross-layer modeling for quality of service guarantees over wireless links,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4504–4512, Dec. 2007.
- [155] —, “Quality-of-service driven power and rate adaptation over wireless links,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.
- [156] —, “Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, Jun. 2008.
- [157] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi, “A tutorial on geometric programming,” *Optimization and Engineering*, vol. 8, no. 1, pp. 67–127, 2007.
- [158] M. Chiang, “Nonconvex optimization of communication systems,” *Advances in Mechanics and Mathematics, Special Volumn on Strang’s 70th Birthday*, 2007.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

List of Acronyms

AWGN	additive white Gaussian noise
BBU	baseband unit
BnB	branch and bound
BnC	branch and cut
BS	base station
CBCS	cloud-based communication system
CCMN	cloud-centric media network
CCS	conventional communication system
CCN	conventional cellular network
CDN	content delivery network
CL	cross-layer
CoMP	coordinated multipoint
C-RAN	cloud radio access network
CSI	channel state information
DL	decoupled-layer
ESUM	extended sum-utility maximization
GP	geometric programming
HetNet	heterogeneous network

MCC	mobile cloud computing
MCMC	Markov chain Monte Carlo
MINLP	mixed-integer nonlinear programming
MLE	maximum likelihood estimator
MMSE	minimum mean square error
MSE	mean square error
OSI	open systems interconnection
QCQP	quadratically constrained quadratic program
QoE	quality-of-experience
QoS	quality-of-service
QWSRM	quasi weighted sum-rate maximization
RAN	radio access network
RRH	remote radio head
SINR	signal-to-interference-plus-noise ratio
SLA	service level agreement
SOCP	second-order cone programming
UE	user equipment
VM	virtual machine
WMMSE	weighted minimum mean square error
WSRM	weighted sum-rate maximization

Appendix B

List of Publications

Journals

1. **J. Tang**, W. P. Tay, and Y. Wen, “Dynamic request redirection and elastic service scaling in cloud-centric media networks,” *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1434-1445, Aug. 2014.
2. **J. Tang**, W. P. Tay, and T. Q. S. Quek, “Cross-layer resource allocation with elastic service scaling in cloud radio access network,” *IEEE Trans. Wireless Commun.*, to appear.
3. **J. Tang**, W. P. Tay, T. Q. S. Quek and B. Liang, “System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity,” *IEEE Trans. Wireless Commun.*, submitted.

Conferences

1. **J. Tang**, W. P. Tay, and T. Q. S. Quek, “Cross-layer resource allocation in cloud radio access network,” in *Proc. IEEE GlobalSIP*, Atlanta, GA, Dec. 2014, pp. 313-317, **Invited Paper**.
2. **J. Tang**, W. P. Tay, T. Q. S. Quek and B. Liang, “Towards the Interaction Between Cloud Processing and Wireless Transmission in Cloud RAN,” in *Proc. IEEE ASILOMAR*, Pacific Grove, CA, Nov. 2015, **Invited Paper**.