Approaches to Joint Base Station Selection and Adaptive Slicing in Virtualized Wireless Networks

Kory A. Teague

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Electrical Engineering

Allen B. MacKenzie, Chair

Luiz DaSilva

R. Michael Buehrer

Mohammad J. Abdel-Rahman

August 1, 2018 (TBD)

Blacksburg, Virginia

Keywords: TBD

Approaches to Joint Base Station Selection and Adaptive Slicing in Virtualized Wireless Networks

Kory A. Teague

(ABSTRACT)

This work received support in part from the National Science Foundation via work involved with the Wireless @ Virginia Tech research group.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mobile carriers have seen explosive growth in both the volume of users and the demands of those users within the networks they operate. New and evolving data-driven applications, such as audio/video streaming, social networking, and the Internet of Things have also placed increasing demand upon the networks. In 2016, the amount of IP data handled by mobile networks exceeded 86 Exabytes; it is projected to reach almost 200 Exabytes in 2018, and 580 Exabytes in 2021 [1]. Due to this exponential growth, incremental approaches to improve the network will fail to satisfy demand. As this growth continues to progress into the near future, new architectures like 5G and its associated technologies will be needed to keep pace with the demand.

However, deployment of these technologies and networks can be a costly, prohibitive venture. To meet these demands requires a similar increase in capital (CapEx) and operational expenditures (OpEx). As volumes and costs rise and margins shrink, approaches to reduce these expenditures become increasingly necessary. Resource infrastructure sharing has been a common practice for MNOs going back to 2G and 3G networks. First, mobile network operators (MNOs) needed to offer coverage for their users in regions where they had no infrastructure, leading to the creation of roaming agreements, eliminating the need for deploying new infrastructure in that region and reducing CapEx. Second, by sharing passive elements of the infrastructure, such as physical sites, tower masts, power, and air-conditioning, the CapEx of deploying new backhaul and radio access networks (RANs), such as cellular base stations (BSs), has decreased [2].

CapEx reductions from passive resource sharing drove an interest in resource sharing of the active elements of the network. For example, MNOs might share RANs, core networks, BSs, antenna systems, or backhaul, which leads to reductions in both CapEx and OpEx. The ability to share these active resources removed the necessity for network operators to own and maintain a physical network while providing actual MNO-like mobile services. These mobile virtual network operators (MVNOs) function similarly to MNOs, but operate a virtualized wireless network (VWN) comprised of virtual resources instead of physical resources without the associated CapEx. It has been shown that virtualization in this manner can increase overall demand satisfaction of a set of VWNs while decreasing overall cost (i.e., OpEx) by decreasing the idle capacity of the networks [3].

In order to take advantage of increasing virtualizable resources and competition for those resources, a specific problem must be solved: how to select the set of virtual resources to form a VWN that meets its demands with necessary or maximum demand satisfaction at minimum cost. The solution to this problem is further complicated in the context of multiple MVNOs, each with one or more VWNs with unique demands, within a large pool of available virtual resources that can be adaptively allocated as demands shift.

This thesis addresses the topic of resource selection and adaptive slicing within cellular VWNs through the lens of a stochastic optimization problem and investigates the two approaches to efficiently reach a solution.

1.1 Trends in Wireless Networking

IP traffic is increasing across all types of network archetypes, and is trending to become more mobile focused. According to the Cisco Visual Networking Index, global IP traffic will increase nearly threefold over the 2016-2021 time period, reaching 3.3 Zetabytes (ZB) annually in 2021 from 1.2 ZB annually in 2016. Traffic across the fixed internet backbone is projected to match this threefold pace, growing from 790 Exabytes (EB) to 2.2 ZB. However, mobile data traffic is projected to have twice the growth of fixed internet over the same period, increasing almost sevenfold from 86 EB in 2016 to 580 EB in 2021. By the end of 2021, 17% of global IP traffic will be mobile data or internet traffic generated by handsets, notebook cards, and mobile broadband gateways, up from 7.5% in 2016. More broadly, traffic from wireless and mobile devices combined will reach 63% of total IP traffic by 2021, up from 49% in 2016. By 2021, smartphone IP traffic (33% of global IP traffic) will alone outnumber PC

IP traffic (25%). Overall, these projections show that population data use and the number of devices attached to the networks is increasing; in 2021, annual global IP traffic will reach 35 GB per capita associated with 3.5 networked devices per capita, up from 13 GB per capita and 2.3 networked devices per capita in 2016. [1]

Consumer IP traffic comprises a majority share of global IP traffic, and this share will continue to grow. From 2016 to 2021, business IP traffic will grow at an average of 21% compound annual growth rate (CAGR), below the growth rate of global IP traffic (24% CAGR). Similarly, business mobile data traffic (41% CAGR) will lag behind that of global mobile data traffic (46% CAGR). [1]

In both the consumer and business markets, this demand includes enormous growth in video applications, specifically that of video streaming. By 2021, global IP video traffic will reach 82% of all consumer internet traffic, up from 73% in 2016. This is represented by a threefold increase of all global IP video traffic over the period and a fourfold increase on just the internet backbone. By 2021, internet live video streaming will account for 13% of this video traffic, growing 15-fold over the period. Similarly, internet video surveillance will occupy 3.4% of internet video traffic, increasing sevenfold, and internet video to tv will increase by a factor of three and a half, growing to comprise 26% of internet video traffic. While these are all values for increases on the internet backbone, mobile data should see similar, but smaller, growth in these areas. Virtual and augmented reality uses will see the largest increase, growing at a 82% CAGR, and expected to reach a 20-fold increase between 2016-2021. [1]

The technology underlying the mobile data network needs to continue to evolve with these changing trends and growth. The primary focus of the 5G cellular standard has been to meet these targets in an effective and robust manner. Of specific interest is that of aggregate data rate (e.g., area capacity, the available amount of data a network can facilitate over a unit area) and edge rate (e.g., 5% rate, the minimum data rate that can be reasonably provided to all but 5% of users) of the network. For 5G, the general consensus is that aggregate data rate and edge rate must be 1000x and 100x that of 4G, respectively [4]. To supply these rates, several strategies are being investigated, with three primary technologies being (1) the continuing of cellular densification and offloading, (2) increased bandwidth by expanding into new spectra like Wi-Fi and millimeter wave, and (3) increasing spectral efficiency through advances such as those in massive multiple-input multiple-output (MIMO).

The first strategy is extreme densification and offloading. By making network cells smaller, the number of active nodes increases for the same unit area. This is a common strategy across cellular generations, and a large impetus behind the use of smaller range RANs like microcells and femtocells [5]. Cell sizes have shrunk, dropping from the order of hundreds of square kilometers to now fractions of a square kilometer. The most important benefit of cell densification is that it increases spectral reuse, which reduces the amount of users competing for the same resources. Theoretically, since signal-to-interference ratio is maintained as the cell shrinks, such densification can be repeated indefinitely as deployments allow [4, 6].

The second strategy is to increase bandwidth through the use of previously unused spectra such as millimeter wave (mmWave) and Wi-Fi. Cellular networks have utilized microwave frequencies ranging from a few centimeters to about a meter in wavelength; this range has become thoroughly occupied and to generate new bandwidth would require expanding to new frequencies [?]. Up to now, mmWave has been unused and in some cases unlicensed due to very poor propagation properties and high equipment costs. However, equipment costs are falling rapidly due to technological maturation. Further, the propagation qualities are increasingly surmountable as cell sizes shrink [?].

cite (1)

cite (2)

The third strategy involves the use of massive MIMO to increase spectral efficiency. MIMO uses multiple transmit and receive antennas to exploit multipath signal propagation, multiplying the capacity of a given radio link. The technology has been used for over a decade as a component of Wi-Fi before being introduced into the 3G, 4G, and 4G LTE standards [4]. A new approach to be used in 5G is that of "massive MIMO", where the number of transmit antennas at the BS greatly outnumber the number of active users [7]. For example, a given BS might have hundreds of antennas while maintaining data links for tens of users. This provides several benefits, most importantly that of vastly improving spectral efficiency.

5G must supply these rates at much higher energy and cost efficiencies, ideally matching or exceeding the capacity increases to avoid increasing overall network energy use and OpEx. However, technologies that have been investigated to adequately increase the capacity of the network have several major hurdles to meet in order to be implemented at the desired energy and cost efficiencies. Massive MIMO requires the deployment of a vast number of antennas, which requires new BS architectures that have issues with scalability and cost. Millimeter-wave is more expensive than the more mature hardware of typical cellular bands. Decreasing cell size for cellular densification allows for smaller, cheaper BSs, but this decreased cost may not keep pace with the required number of increased deployments. Additional measures to decrease expenditures on top of that of the aforementioned technologies is worth investigating. [4]

3GPP (the Third-Generation Partnership Project) is currently working on finalizing the standard for 5G implementations. In December 2017, 3GPP froze the first half of release 15 of the 5G standard, establishing the specifications of non-standalone 5G which utilizes existing LTE networks. It is expected that 3GPP will freeze the second half of release 15, covering 5G New Radio (5G-NR) which establishes specifications for new standalone 5G deployments, in Summer of 2018.

1.2 Virtualization, Virtualized Wireless Networks, and the Networks without Borders Paradigm

One approach towards minimizing CapEx and OpEx of networks has been the utilization of resource sharing. Resource sharing encompases the sharing of resources between multiple networks, and can take the form of *passive sharing* – referring to the sharing of physical sites, tower masts, cabling, power supplies, and other components that are not actively on part of the network architecture – and *active sharing* – referring to the sharing of the active network architecture itself, such as backhaul and RAN. The practice has been utilized since 2G and 3G networks as a tool for reducing CapEx in expanding the network [2]. Since then, resource sharing has become more common, where it is now available, standardized [8], and implemented in many major carrier networks. As reported by Costa-Perez et al. [9], a 2010 market survey [10] found that over 65% of European MNOs have deployed mobile infrastructure sharing in some form. It was further reported [9] that 20% of cells carry about 50% of total network traffic, with the remaining 80% of cells still causing OpEx without gain. Through active resource sharing, networks can reduce or avoid redundant deployments and wasted capacity, reducing overall CapEx and OpEx.

The increasing prominence of active resource sharing challenges the traditional model of ownership of the various network layers and elements. Once it became feasible for network operators to utilize resources owned and maintained by other operators, it became possible for these MNOs to operate networks primarily or only using these shared resources. A given shared resource can be decoupled from a specific physical resource, instead enabling it to be adaptively associated with any of a given pool of qualifying physical resources as network conditions allow, establishing the shared resource as a virtual resource. Further, virtual networks can adapt to changing network conditions, adding and removing virtual resources as capacity requirements change. For example, a MVNO with a network of virtual resources can add additional virtual resources during the peak hours when additional capacity for end user satisfaction is needed, and removing unneeded resources during times of low demand to reduce OpEx of the network.

Other research in virtualization has been shown to improve performance in wireless networks. Panchal and Yates [11] have shown on an LTE testbed that active inter-operator resource sharing improves performance of overloaded networks in terms of decreased drop probability and overloaded sectors. Virtualized sharing methods provided further, albeit marginal, performance improvements at an increased complexity, but would be more suitable on adaptive DSA (dynamic spectrum access) based systems. In this capacity, improved performance allows for smaller networks reducing CapEx and OpEx. Costa-Perez et al. [9] found in LTE testbeds that network virtualization substrate (NVS), a suggested virtualization technique, provides improved overall throughput compared to networks without resource sharing.

1.2.1 Virtualization and the Network Value Chain

This concept of virtualization partitions the classical wireless networking value chain, allowing for specialization of segments of the value chain into new entities such as resource providers and service providers [2]. Traditional MNOs control every segment of the typical mobile network value chain (Fig. 1.1 [12]), from spectrum to the end user. With the introduction of virtualization techniques, MVNOs can obtain access to bulk network services available from an MNO. This allows for MVNOs to specialize into specific formulations without the CapEx or responsibility to deploy and maintain the radio infrastructure while incurring no significant cost [12]. For example, an MVNO could focus on marketing, working solely within the distribution channel to the network's customers, or the MVNO could establish itself earlier in the value chain, focusing on operating the network from the core network.

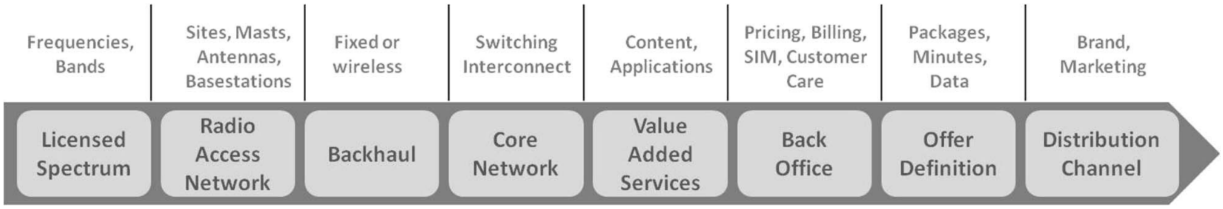

Figure 1.1: Typical MNO value chain [12]

Specialization of the networks and the entities involved in the network can improve the cost efficiency of the netowork. According to Beckman and Smith [2]: "Extensive vertical integration is a characteristic of an immature product. As the product increases in complexity, it is no longer possible to [provide] an end-to-end solution." In both examples, the MVNO adds value to the traditional value chain by specializing in segments (e.g., marketing or service creation) that are different from the segments (e.g., network maintenance) still handled by the owner and operator of the network resources.

By focusing on the strengths provided by virtualization, more value can be generated through specialization. Doyle et al. [12] investigates the value chain with this segmentation in mind and introduces the Networks without Borders (NwoB) approach as a new service-oriented network market with a proposed new value chain (Fig. 1.2 [12]). The network under the NwoB approach is entirely service-oriented, where the network responds to services and connectivity is tailored for the service. Services have a wider meaning than the voice, text, and data of a typical MNO. Services also include that of Netflix-like or real-time video streaming, Internet-of-Things (IoT) applications, or various types of over-the-top services. Each service would be provided by a service provider that compensates the virtual network

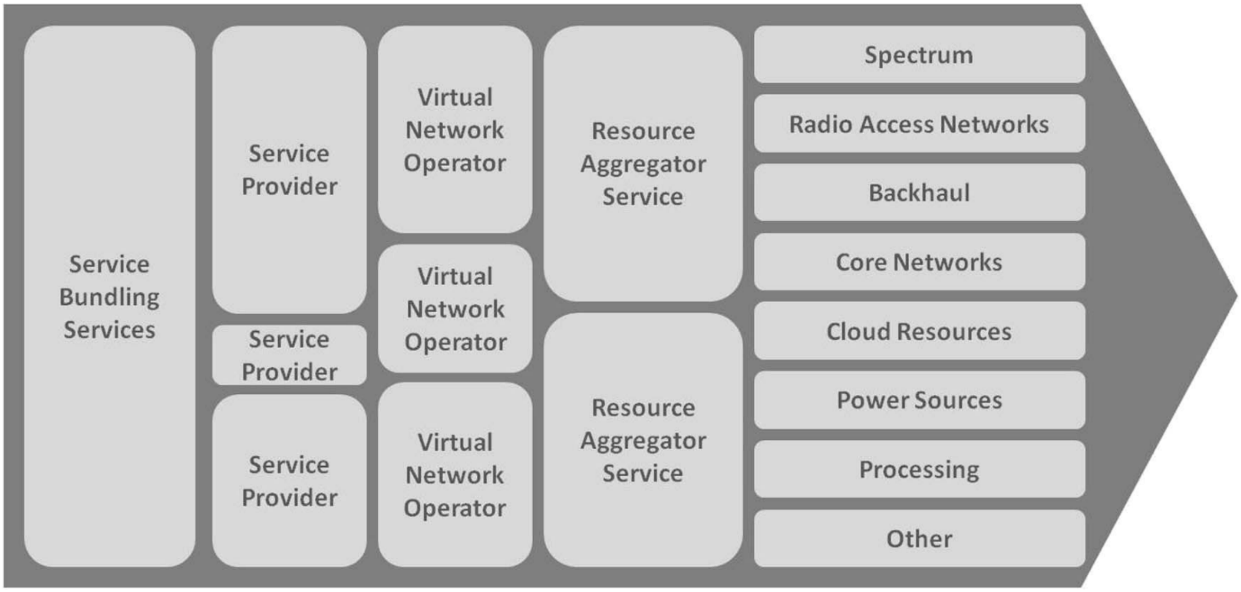

Figure 1.2: Proposed network value chain under the NwoB paradigm [12]

operator operating a virtual network constructed specifically for the purpose of that service; the virtual network is the service. Unlike an MVNO which manages resources provided to it by agreement, the virtual network operator manages slices of virtual resources from a pool of all resources as provided through resource aggregating services.

The benefits of this paradigm as proposed by Doyle et al. [12] are four-fold. First, it provides specialization and independence for each stage, allowing service providers to focus on generating value from services provided. Second, networks can be specialized for a service, reducing OpEx through extensive resource sharing. Third, as resources are virtualized and pooled together, any resource (e.g., typical RAN, Wi-Fi, mmWave, raw spectrum) could be added with the pool and utilized for a network as its properties fit the network's needs. Fourth, it lowers the barrier for entry and establishes services for new entities to fulfill.

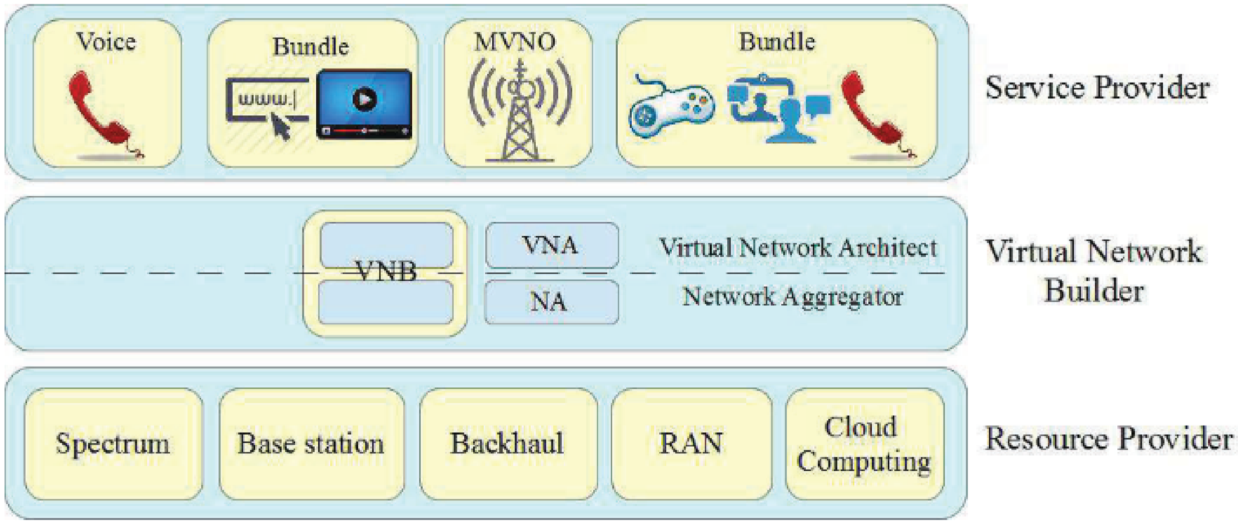

Figure 1.3: VWN architecture as used in this work [3,13]

1.2.2 Virtualization Architecture in this Work

Recognizing the critical nature of virtualization and resource allocation, this thesis develops and analyzes two methods for constructing virtualized wireless networks built on a virtualization architecture [3,13] inspired by the NwoB paradigm presented by Doyle et al. [12]. Fig. 1.3 [3,13] illustrates the three primary roles in this architecture: (1) the Resource Providers (RPs), (2) the Virtual Network Builders (VNBs), and (3) the Service Providers (SPs).

RPs deploy and maintain the physical resources that are to be virtualized and offered for use within the virtualization framework, and are the various entities that occupy the right-most column of segments (i.e., resources) in the NwoB value chain (Fig. 1.2 [12]). These resources can be in the form of any network-capable resource. For example, the resources could be BSs as provided by a traditional MNO, a company-or individual-owned WLAN, femtocell access points, available licensed or unlicensed spectrum, or cloud computing. They are any entity that offers a virtualizable resource, such as a traditional MNO, company, or individual. Further, they maintain the resources, but also determine how the resource would be sliced and shared. This large variety causes the underlying infrastructure of this virtualization framework to be highly heterogeneous.

The VNB acts as resource aggregator, VWN constructor, and as intermediary between SPs and RPs. Therefore, the VNB acts as a combination virtual network operator and resource aggregator in the NwoB value chain (Fig. 1.2 [12]). The VNB aggregates the resources maintained by individual RPs to establish the pool of available virtual resources. The VNB also coordinates with SPs to understand the demands of their services, and constructs VWNs tuned specifically to these demands. By understanding the needs of the services provided by the SPs, the VNB will evaluate which virtual and virtualizable resources available from the RPs are needed to construct the optimal[1] network for the SPs' needs, coordinate with the necessary RPs to obtain access to these resources for a given wholesale (OpEx) cost, and construct the network for the SPs to operate. Multiple VNBs can coexist, each with their potentially overlapping set of RPs from which to aggregate resources.

The SP operates very similarly to the service providers in the NwoB approach. Primarily, an

[1] In this network context, "optimal" is loosely defined to mean a network that provides the maximum demand satisfaction for the service provided by the SP at the minimum cost (OpEx) to be paid to the RP. These two requirements – maximum demand satisfaction and minimum cost – are frequently contradictory and need to be balanced by the VNB.

SP determines a service that they wish to provide, understands and enumerates the demands that are to be appropriately satisfied for that service, and provides the service over the VWN to their end users. SPs can provide a wide range of services over the network. The service could be a traditional MNO or be providing MNO-like services, such as voice calling and texting. Services could cover specific applications, such as IoT, teleconferencing, augmented or virtual reality, or emergency services. Other examples include traditional over-the-top services, such as Netflix-like or real-time (live) video streaming, social media (Facebook, Twitter, etc.), messaging (Skype, Groupme, etc.), or news/content feeds. Further, an SP could also bundle several services, either through a single VWN built for the bundle, or by bundling services provided by several SPs.

Between these three entity roles, various interactions become possible. The most common interactions are illustrated in Fig. 1.4 [13]. The interactions between the various entity roles are: (A) among SPs; (B) between the SPs and the VNBs; (C) among VNBs; (D) between the VNBs and the RPs; and (E) among RPs. It should be apparent that across each of these interactions is the imposition of costs as exchange for the transfer of services, networks, and resources.

Interaction (A) describes associations among various SPs. This would typically occur in situations where a SP desires to bundle the services of several SPs, or when a SP wishes to utilize a specialized network operation from another SP. Generally, this interaction would be performed manually over timescales of weeks or months.

Interaction (B) describes associations directly between SPs and VNBs. This would be one of

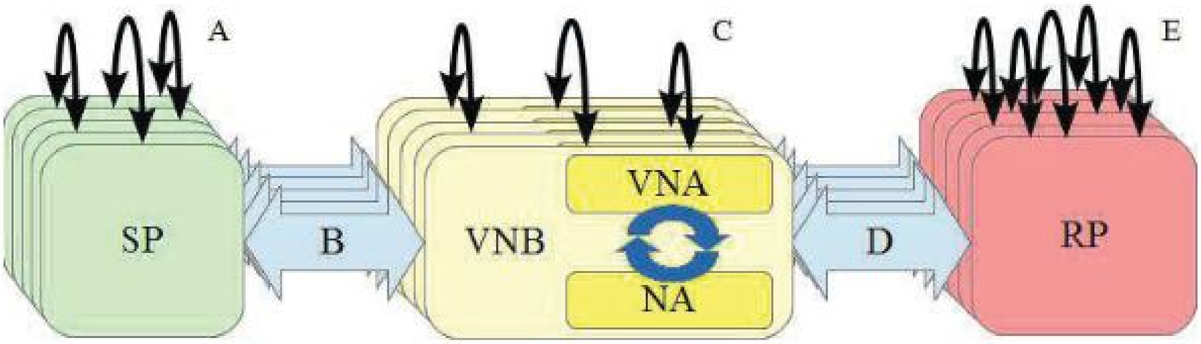

Figure 1.4: Interactions between roles in the VWN architecture [13]

the most common interactions within this framework. This interaction is bidirectional. In the first direction, SPs would provide the VNB they are coordinating with the specific demands and needs for the service they are providing. In the opposite direction, the VNB utilizes these conveyed needs and demands to construct a VWN and provide it for use to the service provider. Ideally, this interaction is highly or entirely automated, with the interactions varying from minutes or hours to weeks or months based on the level of automation and the specifics of the interaction. It will require optimization techniques and/or machine learning to achieve satisfactory results in this interaction.

Interaction (C) describes associations among various VNBs. Generally, such interactions may occur when a VNB does not have access to the appropriate virtual resources to satisfy interaction (B) interactions. Obvious examples include not having the necessary resources to provide adequate coverage over geographical areas or capacity in high-density environments. These interactions would generally be performed manually over timescales of weeks or months.

Interaction (D) describes associations between VNBs and RPs. Similar to interaction (B), this would be the other of the most common interactions within this framework. It is also very important, as it establishes the mapping between the virtual and physical resources and builds the substrate that the framework is built upon. VNBs interact with the RPs by making requests for new resources and releasing unneeded resources. RPs interact with the VNBs by issuing updates, such as any changes to the resources in the VNBs' available pool of resources. Updates such as these are potentially highly disruptive to the VNBs as the updates can impact a large number of VWNs managed by the VNBs. With further similarity to interaction (B), this interaction is highly dependent on automation; based on the level of automation, this interaction may occur over timescales of minutes or hours to weeks or months.

Interaction (E) describes associations among various RPs. In this interaction, various RPs establish connections with each other to facilitate proper mapping of physical resources to virtual through the use of quality of service (QoS) parameters that define the abstracted resources. For example, a small-scale RP containing only an individual-owned femtocell could connect with a larger RP via this interaction so that the resource within the small-scale RP is visible for association with a VNB over interaction (D) as handled by the larger RP. These interactions could take seconds to weeks depending on the complexities of the RPs, their resources, and the amount of human involvement.

Other work has been completed using this architecture. Abdel-Rahman et al. [3] constructed several resource allocation models, including one-stage programs, two-stage programs, and a one-stage stochastic program, to investigate the efficacy of this virtualization architecture upon a preexisting set of resources. The implementation focused on interaction (B) from the perspective of the VNB, and showed that virtualization decreased the cost and idle capacity of the networks and increased demand satisfaction of the networks.

Cardoso et al. [13] expanded on this work by introducing a two-stage stochastic program to optimize interaction (B). The two-stage stochastic resource allocation similarly reduces cost and idle capacity of the VWN compared to the network without sharing. However, no direct comparisons are made with the non-stochastic programs tested by Abdel-Rahman et al. [3].

Gomez et al. [14] utilized this architecture from an economics perspective. Using a matching markets framework, they investigated the interaction of association between SPs and VNBs, such as the methods for how SPs indicate their needs and how VNBs indicate their VWN capabilities, and the fees that SPs will pay to partner for a VNB. Gomez expanded on this work in her Ph.D. dissertation [15].

Still reading/parsing these papers. Double check.

The focus of this thesis is on optimization approaches largely in the context of interaction (B). This problem involves establishing how SPs convey the demands needed by the VNB to construct an optimal VWN for the service provided by the SP. Further, the construction of the optimal VWN is sought within a short amount of time so that interaction (B) can be completed over shorter timescales (e.g., minutes or hours) instead of longer (e.g., days, weeks, months). With an optimal VWN in mind, construction of the VWN is inherently an optimization problem, and the seeking of expedient solutions lays within the study of optimization.

1.3 Review of Optimization Methods

In this thesis, I approach the problem of the creation of optimal networks by the VNB that satisfy the specific demands of SPs using a pool of resources provided by a set of RPs. This is naturally a form of optimization problem, in which some objective function is either minimized or maximized. At it's most basic, optimization techniques (e.g., linear programming, integer programming) will find the set of input parameters that minimize or maximize a single decision variable – the value of the objective function – in context of a set of constraints. More complex optimization problems can solve for multiple decision variables by establishing weights based on their relative importance within the objective function.

1.3.1 Stochastic Programming

Standard linear and integer programming requires complete, certain knowledge of all parameters that affect the functions or model being optimized (i.e., the model's parameters and functions must be deterministic). Communications, especially wireless communications, can be highly non-deterministic as the communication channel introduces a large amount of uncertainty. Stochastic programming provides a powerful mathematical tool to handle optimization under such uncertainties.

Stochastic programming has been recently exploited to optimize resource allocation in various types of wireless communications operating under uncertainties. Abdel-Rahman et al. [3] exploit stochastic optimization within the framework of the virtualization architecture presented in Section 1.2.2 to minimize the cost of resource allocation by introducing probabalistic QoS guarantees. Cardoso et al. [13] expand on that work by introducing a second stage to balance maximizing demand satisfaction while minimizing cost. Abdel-Rahman and Krunz [16] utilize stochstic programming for resource allocation in DSA networks considering satisfaction of link demand as a stochastic constraint. Abdel-Rahman et al. [17] propose a stochastic optimization formulation to optimally orchestrate LTE-U networks that utilize Wi-Fi access points considering stochastic QoS guarantees. Abdel-Rahman et al. [18] utilize stochastic optimization for resource allocation in opportunistic LTE-A networks considering the probability of rate demand satisfaction for end users. Soltani et al. [19] utilize stochastic optimization as a tool for resource allocation tasks for OFDMA-based cognitive radios considering interference from primary user systems. Atawia et al. [20] utilize predictive resource allocation techniques, which include the use of stochastic optimization, for improving energy-efficient video streaming for mobile end-users, such as those riding buses and trains.

Introducing stochastic parameters and constraints allows the optimization model to consider probabilities within the optimization. In the case of resource allocation in networks, it may be possible to allocate enough resources to satisfy all end-user demand. Such an optimization may require too many resources to be economical considering the law of diminishing returns, with the solution being cost prohibitive. It is much cheaper to solve such that 95% or 99% of demand is satisfied, leaving the remainder as unfortunate momentary edge cases.

However, standard linear programming techniques cannot solve models with stochastic parameters. Stochastic programming therefore requires converting the stochastic program into its deterministic equivalent program (DEP) which replaces all stochastic variables with deterministic variables [21]. The process of forming a DEP from a stochastic program involves converting each stochastic variable into a set of all possible scenarios and scenario probabilities. These scenarios and scenario probabilities are present within the model as a new dimension and weight for the now-deterministic variable. To fully encapsulate the stochastic variable, the deterministic equivalent variable may be composed of an uncountably infinite set.

Is it, or is it just countable? Maybe the textbook [21] talks of this. Read and double check.

Resource allocation problems are typically some form of integer programming – in which all decision variables (unknowns) are integers – or mixed integer programming – in which some decision variables (unknowns) are integers. Both integer and mixed integer programs are considered generally NP-hard[2] and thereby intractable. As the programs increase in scope, they become enormously more computationally complex to solve; accounting for the scenarios of the previously stochastic variables further increases this complexity. Finding the optimal solution may require more time than is feasible; in the worst case, these problems run in exponential time complexity.

[2] Finding the minimum resource allocation that provides coverage over a geographic area falls within a category of problems referred to as *minimum set cover problems*. It is apparent that the problem considered in this thesis – specifically the stochastic program proposed in Section 2.2 – is some form of minimum set cover problem; specifically, it might be referred to as a capacitated set cover problem. Minimum set cover problems are provably NP-hard and typically rely on approximation solutions to solve in a feasible amount of time [22].

1.3.2 Metaheuristic Approaches

The use of heuristic or metaheuristic algorithms can provide close-to-optimal solutions in much better time. Examples include hill climbing, simulated annealing, ant colony optimization, and particle swarm optimization. Each of these approaches are iterative techniques.

Hill climbing starts with an arbitrary solution and makes incremental changes to variables, finding a new solution. If the new solution is more optimal than the previous, the new solution is iterated upon. This continues until no further improvements are made. Hill climbing will only find the local maximum close to the initial arbitrary solution, and is best in convex problems where the only local maximum is guaranteed to be the global maximum.

Simulated annealing is inspired by the process of annealing found in metallurgy, where metal is heated to the point where atoms can migrate, reducing defects in the crystalline structure. In simulated annealing, the model has some notion of internal energy of the system to be minimized, temperature, and discrete states with one being an initial state. On each iteration, the temperature drops, or cools, slightly, and the system selects a new state that neighbors the current one, and switches to the new state with a probability dependent on the temperature and the energies of the current and new state. For temperatures greater than zero, the system can transition to a state with higher energy (i.e., that is less optimal) to better investigate the search space. As temperature drops, the system's energy will tend to drop. When the temperature reaches zero, the system will only transition to states of lower energy (i.e., that are more optimal), reducing to the hill climbing algorithm.

Ant colony optimization is inspired by the behavior of ants. A colony of ants move around independently trying to find food, laying pheromones on the taken path. Upon crossing paths, ants have a probabilistic chance to follow the new path based on the strength of the pheromones of the new and old paths. Over time, pheromones evaporate, and paths less taken will weaken. Longer paths, since they take longer to traverse and will be reinforced less often, will also weaken. This has benefits over approaches like simulated annealing because it adapts in real time.

In particle swarm optimization, a number of candidate solutions, called particles, are created that move semi-chaotically. In each iteration, every particle will move according to its velocity. Each particle has it's best known position, and it's velocity updates in a way guided by their own best known position and the swarm's best known position. This allows a large portion of the search space to be investigated, investigate candidate solutions to explore for regions containing local maxima until it settles to exploit and find the best found local maxima.

In this thesis, I utilize a genetic algorithm as an approach for optimization. A genetic algorithm is a form of evolutionary algorithm, a set of algorithms which are inspired by biological evolution and natural selection. Each iteration is called a *generation* and is composed of a number of candidate solutions called *individuals*. Each individual is defined by a *chromosome* which details the specific candidate solution. During each generation, every individual is evaluated on its *fitness*, a function dependent on the individual's chromosome; the higher the individual's fitness, the more optimal the individual. Individuals called *parents* are then randomly selected to pass their chromosome onto the next generation in a process called *selection*; in selection, more fit individuals are more likely to be selected. With a certain probability, groups of parents will undergo *crossover* and exchange the data contained within their chromosomes to form new *children* that are a mixture of the parents; if mixing does not occur, the parents are cloned into the next generation as children. Then, individual bits within the children's chromosomes have a chance to flip, or *mutate*. The resulting children from crossover and mutation form the entire next generation.

Since chromosomes from fitter individuals are more likely to pass on to subsequent generations, generations gradually become fitter. Through crossover, fit chromosomes may combine to form fitter children that proliferate; less fit children are often also formed, but are generally not selected for later generations. Mutation introduces diversity into the generations, which expand the exploration of the search space. More details, including that of implementation and variants, will be expanded upon in Section 3.2.

Genetic algorithms have been used previously as approaches for simplifying constraints of large, complex stochastic optimization problems. Cui et al. [23] used a genetic algorithm where each chromosome defined a subproblem of the larger optimization problem, and the fitness was evaluated by solving the subproblem with linear programming optimization methods. Hybrid approaches, such as the one used by Cui et al. [23], and other effective metaheuristic algorithms (e.g., ant colony optimization, particle swarm optimization, neural networks and machine learning) are worth investigating in the context of the posed VWN architecture, but beyond the scope of this thesis.

1.4 Thesis Objective

The objective of this thesis is to develop two approaches of joint resource allocation to construct a set of a VWNs and adaptively slice the selected resources to the individual VWNs. A model will be presented as the context for these approaches, expanding upon the VWN architecture proposed in Section 1.2.2. The validity of this model will be restricted to the scope of cellular networks using generic base stations as its resources with perfect connections to demand points within range. The two proposed approaches will be performed within the VNB, and evaluated in four cases that differ in the resources provided by the RPs and service demands to be satisfied by the SPs. Accordingly, the efficacy of these approaches will be measured primarily by the optimality of the solutions, such as cost and network service demand satisfaction, and the run time, providing the VNB with a sufficient solution in a reasonable amount of time.

1.5 Thesis Outline

This thesis is organized as follows. Chapter 1 establishes the motivation for investigating resource selection for virtual network construction. Chapter 1 also presents the associated background information and literature review regarding virtualization, wireless networks, and optimization. Chapter 2 defines the model used for the resource allocation methods explored in this thesis. Further, Chapter 2 also details the two-stage stochastic optimization problem which optimally performs resource selection and slicing as a basis of approaches presented within this work. Chapter 3 establishes the two approaches investigated to provide solutions to the optimization problem posed in Chapter 2: a sampled Deterministic Equivalent Program which solves the problem as a whole and a genetic algorithm that simplifies the problem by providing an estimated optimal resource selection. Chapter 4 tests these two approaches by presenting four data sets that mimic real world cellular networks and evaluates the results. Chapter 5 contains the conclusions and proposed future work in this area.

Chapter 2

Virtual Network Builder Model

This chapter establishes the mathematical foundation for the work completed in this thesis. First, a geographic model is presented, defining an area of interest, the pool of resources maintained by the RPs for use by the VNB, a characterization for service demand communicating the needs for the SPs' VWNs, and the SPs' end users to be satisfied. Second, a two-stage stochastic program utilizing this model is proposed to solve the posed problem of resource selection and adaptive slicing for use in VWN construction within the VNB.

After finishing chapter, reread lede

2.1 Network Area Definitions

Consider a geographic area of width X meters and length Y meters that contains a VNB and a set $\mathcal{S} \overset{\text{def}}{=} \{1, 2, \ldots, S\}$ of virtualized resources the VNB has aggregated for use in the construction of VWNs. The pool of resources, $\mathcal{S}$, is mapped to physical resources owned

See Notes

and maintained by RPs and are made available for use through contracts between the RPs and the VNB. The contract-negotiated cost for the VNB to lease resource $s \in \mathcal{S}$ is denoted by c_s. The costs for the resources used within a constructed VWN are passed to the SPs as part of the overall cost of the network. The rate capacity of resource $s \in \mathcal{S}$ is denoted by r_s and its coverage radius is denoted by b_s.

Does this belong here?

Let $\mathcal{N} \overset{\text{def}}{=} \{1, 2, \ldots, N\}$ be the set of SPs seeking a VWN to host their services with coverage within the geographical area. An SP $n \in \mathcal{N}$ associates with the VNB to create their desired VWN. Through this association, SP n must coordinate with the VNB to indicate the demands of the intended service the VWN would need to satisfy. SP n must know and communicate to the VNB the estimated geographic distribution of the service's traffic demand density as a function, $\rho\left(x, y\right), x \in [0, X], y \in [0, Y],$ in terms of $\frac{\text{bits}}{\text{km}^2}$. This could be in the form of a continuous function or as discrete pixels, and could be generalized as a heatmap indicating locations of necessary coverage and the desired capacity within specific regions of the area. Examples of possible maps could be for services such as localized video streaming (specific, localized coverage with high regional capacities) or MNO-like voice lines (broad coverage with comparatively low capacity). Further, the SP would also provide the desired or needed percent demand satisfaction rate for the service. Some services have high priority, such as those related to emergency services, and must have nearly if not perfect 100% demand satisfaction. Others, such as the aforementioned generic voice lines or video streaming, can withstand some demand to remain unsatisfied as a trade off for decreased network leasing or operational costs.

The current model doesn't need this, but adjusting α (sDEP) or the map (GA) effectively controls for this. Can the model be updated formally for this?

Let $\mathcal{M}_n \overset{\text{def}}{=} \{1,\, 2,\, \ldots,\, M_n\}$ be the set of demand points SP $n \in \mathcal{N}$ is attempting to satisfy with its service. Each demand point $m \in \mathcal{M}_n$ is seeking to connect to the VWN operated by SP $n \in \mathcal{N}$ with total point traffic demand denoted by d_m. Let $u_{ms} \in [0,\, 1]$ represent the normalized capacity (with respect to r_s) of resource $s \in \mathcal{S}$ at point $m \in \mathcal{M}_n, n \in \mathcal{N}$, (i.e., the normalized maximum rate that a user can receive at point m from resource s). Specifically,

$$u_{ms} \overset{\text{def}}{=} \begin{cases} 0, & \text{if demand point } m \text{ is located more than } b_s \text{ meters away from resource } s, \\ 1, & \text{if demand point } m \text{ is located within a small distance of resource } s \\ (0,\, 1), & \text{otherwise.} \end{cases} \tag{2.1}$$

It is apparent that u_{ms} will vary according to the path-loss characteristics of the environment and other various factors. In some instances, it can be beneficial to simplify this definition such that

$$u_{ms} \overset{\text{def}}{=} \begin{cases} 1, & \text{if demand point } m \text{ is located within } b_s \text{ meters of resouce } s, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

It wouldn't be that hard to alter u_{ms} to utilize a pathloss model of some sort. It might complicate the GA, as the cells would need to be cap point capacity according to it, but it should be fairly simple with a known model.

This simplification allows the available pool of resources, or any subset thereof, to be easily visualized as a Voronoi tessellation [?]. In a Voronoi tessellation, a two-dimensional plane is tessellated into a set of convex polygons, each of which is defined by a single point contained within. All points comprising the area enclosed by a polygon is closer to that polygon's defining point than any other polygons'. Using the simplified definition for u_{ms}, the Voronoi tessellation of a set of resource locations functions as a coverage map for those resources, assuming that all polygons are within the respective ranges for their associated resources; each polygon represents a region where $u_{ms} = 1$. Assuming this binary condition is a simplification that aids in the implementation of the approach used in Section 3.2, and is generally assumed in the rest of this thesis.

Each demand point has a location stochastically determined by the distribution of traffic demand density that SP $n \in \mathcal{N}$ communicates to the VNB. Throughout this thesis, stochastic variables will be differentiated from deterministic variables with a tilde ($\sim$) placed above the symbol (e.g., $\tilde{u}_{ms}$).

At times, such as for demonstration and visualization or computation, wholly deterministic variables are needed in place of stochastic variables. In general, this is accomplished for the specific instance by finding a realization of the distribution associated with the stochastic variable. For demand points with stochastic locations, they are generated according to a two-dimensional non-stationary Poisson point process (PPP) using the traffic demand density function or distribution, $\rho\left(x_i,\, y_i\right)$, as the PPP's spatial intensity measure. In this process, I use an acceptance-rejection method [?]. A stationary PPP is generated according to the maximum value within the region of the traffic demand density function. That is, a number of points generated within the region is selected from a Poisson random variable with mean $\rho_{\max} * X * Y$. Each point is then independently and uniformly distributed (i.e., each point has a location $(x,\, y)$ with $x \sim \mathcal{U}\left(0,\, X\right)$ and $y \sim \mathcal{U}\left(0,\, Y\right)$) over the region. Then, each point undergoes an acceptance-rejection procedure to inhomogenize the PPP. Each point is kept

cite
See note

with a probability of the ratio of the value of the demand density function at that point's location to the maximum value of the demand density function. That is, for each point in the PPP a uniformly-distributed random number, P, is generated over $[0,1]$ and the point is either *accepted* and kept or *rejected* and discarded according to

$$\begin{cases} \text{if } P \leq \frac{\rho(x_i, y_i)}{\rho_{\max}}, & \text{the point at } (x_i,\, y_i) \text{ is kept} \\ \\ \text{otherwise,} & \text{the point at } (x_i,\, y_i) \text{ is discarded} \end{cases} \tag{2.3}$$

where x_i and y_i are the x- and y-coordinates of the i^{th} point of the stationary PPP. So that the distribution of demand realized by the non-stationary PPP corresponds to the traffic demand density function, the overall demand

$$D = \int_0^X \int_0^Y \rho\,(x,\, y)\, dydx \tag{2.4}$$

of the demand density distribution is evenly distributed such that

$$d_m = \frac{D}{M_n}, \forall m \in \mathcal{M}_n,\, n \in \mathcal{N} \tag{2.5}$$

Generally, PPPs and non-stationary PPPs generate a number of points correlating to the intensity value and function, respectively. A specific number, M_n, of points can also be generated as necessary to populate a realization for $\mathcal{M}_n$, $n \in \mathcal{N}$. Instead of generating a random number of points according to a Poisson random variable, points are generated one at a time and individually either kept or discarded as defined in Eq. 2.3. Once M_n points have been generated and kept, a non-stationary PPP of $\mathcal{M}_n$ has been generated.

This is allowed because, by definition, each point in a PPP is independent and identically distributed; each point is generated independently and identically according to a uniform distribution. Generating more points is only indicative of a higher intensity PPP. Specifically, the number of points generated in the initial, stationary PPP is linearly dependent on a Poisson random variable with mean $\rho_{\max} * X * Y$; doubling the number of generated points correlates with a doubling of $\rho_{\max}$. Scaling ρ according to any desired number of points does not change the overall structural characteristics of the underlying distribution described. The only change is that with more demand points generated, each point provides less demand according to Eq. 2.5.

From the perspective of the VNB, each SP is only distinguishable by its overall demand characteristics. These demand characteristics are defined by its demand density distribution which defines $\tilde{u}_{ms}$. The VNB must construct a VWN for each SP, but for optimal VWNs to be created, the VNB must consider the demands of all SPs simultaneously and in context of each other. For the VNB, all SP demand points are indistinguishable. Therefore, the VNB considers single set of demand points $\mathcal{M} \overset{\text{def}}{=} \bigcup_{i=1}^{N} \mathcal{M}_i$ with demands d_m, $m \in \mathcal{M}$ and normalized capacities $\tilde{u}_{ms}$, $m \in \mathcal{M}$, $s \in \mathcal{S}$.

I assume that a resource $s \in \mathcal{S}$ can be allocated between multiple demand points, and $\delta_{ms} \in [0, r_s]$, $m \in \mathcal{M}$, $s \in \mathcal{S}$, represents the rate of resource s that is allocated to point m.

2.1.1 Example Demand Distribution Model

Review chapter thus far and pick up from here!

It has been shown that a log-normal distribution can approximate traffic demand in real-

Move onto talking about a

world cellular networks [24, 25].

To generate this spatial distribution over the area of consideration, an initial Gaussian field, $\rho^G = \rho^G(x,\, y)\,,\, x \in [0,\, X]\,,\, y \in [0,\, Y]$, is generated by

$$\rho^G(x,\, y) = \frac{1}{L} \sum_{l=1}^{L} \cos(i_l x + \phi_l)\, \cos(j_l y + \psi_l) \tag{2.6}$$

where $\mathcal{L} \overset{\text{def}}{=} \{1,\, 2,\, \ldots,\, L\}$ is a set of the products of two cosines with angular frequencies $i_l,\, j_l \sim \mathcal{U}(0,\, \omega_{\max})\,,\, l \in \mathcal{L}$ and phases $\phi_l,\, \psi_l \sim \mathcal{U}(0,\, 2\pi)\,,\, l \in \mathcal{L}$. As L increases, ρ^G approaches a Gaussian random field with a spatial autocorrelation dependent on $\omega_{\max}$ according to the central limit theorem.

Consider expanding on these definitions autocorrelation functions maybe add an image of the resulting ρ^G

The approximate Gaussian distribution ρ^G is then normalized to a standard normal distribution. The final log-normal distribution, $\rho = \rho(x,\, y)\,, x \in [0,\, X], y \in [0,\, Y]$, is determined by assigning location and scale parameters

$$\rho(x,\, y) = \exp\left(\frac{\sigma}{\sqrt{\mathrm{Var}\left(\rho^G\right)}}\, \rho^G(x,\, y) + \mu\right) \tag{2.7}$$

where $\mathrm{Var}\left(\rho^G\right)$ is the variance of ρ^G.

Consider an image showing the standardized and log-normal fields, and accompanying pdfs and histograms maybe separate ρ^S and ρ

$\rho(x,\, y)$ can be sampled over the space into individual pixels as per Lee with each pixel's value indicating the number of homogeneous demand points within the pixel [26]. In contrast, I allow $\rho(x,\, y)$ to provide a continuous, spatially-correlated log-normal distribution depicting the demand density over the region for the SP.

Expand on the differences between the model proposed by Lee and as implemented here

Change "I"?

2.2 Stochastic Optimization

Replace “we”s to “I”s or find alternate wording/tense/voice. Ensure equations are spaced appropriately for full page column and that there are no unnecessary vertical or horizontal spacing. Expounding on the various components of the stochastic optimization problem might be worthwhile. Might be worthwhile to also mention that the stochastic nature of this specific formulation is limited to handling stochastic demand point locations. As with 2.1, modify wording and phrasing accordingly to accommodate the possibility for multiple RPs and SPs in the model. When referring to “Problem”s (e.g., Problem 1 from section 2.2), refer to the equations that make up that problem; see first reference of Problem 1 in Chapter 3 for reference.

We formulate the presented problem as a two-stage stochastic optimization problem. We introduce $z_s, s \in \mathcal{S}$ as a binary decision variable defined as

$$z_s \overset{\mathsf{def}}{=} \begin{cases} 1, & \text{if BS } s \text{ is selected for the created VWN,} \\ 0, & \text{otherwise.} \end{cases}$$

To balance the interest of maximizing demand satisfaction against minimizing cost, we introduce the positive real number α as a weighting coefficient between the two stages.

Problem 1 (Two-Stage Stochastic Optimization Problem)

$$\underset{\{z_s,\, s\in\mathcal{S}\}}{\text{minimize}} \left\{ \sum_{s\in\mathcal{S}} c_s\; z_s + \alpha \mathbb{E}\left[h\left(z,\, u\right)\right] \right\} \tag{2.8}$$

subject to:

$$z_s \in \{0,1\}, \forall s \in \mathcal{S} \tag{2.9}$$

where $h(z,u)$ is the optimal value of the second-stage problem, which is given by:

$$\underset{\{\delta_{ms}, m\in\mathcal{M}, s\in\mathcal{S}\}}{\text{minimize}} \left\{ - \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}} \delta_{ms}\; \tilde{u}_{ms} \right\} \tag{2.10}$$

subject to:

$$z_s = \mathbb{1}_{\left\{\sum_{m\in\mathcal{M}} \delta_{ms} > 0\right\}}, \forall s \in \mathcal{S} \tag{2.11}$$

$$\sum_{s\in\mathcal{S}} \delta_{ms}\; \tilde{u}_{ms} \leq d_m, \forall m \in \mathcal{M} \tag{2.12}$$

$$\sum_{m\in\mathcal{M}} \delta_{ms} \leq r_s, \forall s \in \mathcal{S}. \tag{2.13}$$

The first stage objective function (2.8) minimizes the total cost of the selected network with respect to that network's ability to satisfy the demand contained within the region. The second stage objective function (2.10) maximizes demand satisfaction by maximizing the total demand allocated to the resources comprising the network, as specified by δ_{ms} as the decision variable of the second stage.

Constraints (2.9), (2.11), and (2.13) implement the defined ranges and values of the decision variables z_s and δ_{ms}, with (2.11) ensuring that demand is allocated only to selected resources.

For constraint (2.11), $\mathbb{1}_{\{*\}}$ is defined by

$$\mathbb{1}_{\{*\}} \overset{\mathsf{def}}{=} \begin{cases} 1, & \text{if condition } \{*\} \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$$

Constraint (2.12) ensures a demand point $m \in \mathcal{M}$ is not allocated more resources than it demands.

Chapter 3

Approximation Approaches

Go through chapter notes, and break them down to where and what.

Replace with introduction paragraph to this chapter. This chapter lays out the two major approaches I am using in this thesis: the sampled DEP and genetic algorithm, and the foundation those approaches are built on. These approaches are built on the stochastic optimization problem as laid out in 2.2, and meant to provide a solution (i.e., the DEP) or an estimate (i.e., the sampled DEP and genetic algorithm) as the original problem is not directly solvable.

In this chapter, I work on defining the approximation approaches used in my work. Lead in to discussing the need to approximate the stochastic optimization problem from section 2.2 to adequately solve my work, then introduce the two approaches I used to approximate the optimization problem: the DEP/its sampling/generalized post-selection slicing and the *genetic algorithm as a selection method.*

3.1 Deterministic Equivalent Program

Replace all "we"s with "I"s or with alternate phrasing/tense/voice. When referring to "Problem"s (e.g., Problem 1 from section 2.2), refer to the equations that make up that problem; see first reference of Problem 1 in Chapter 3 for reference.

Introduce the idea of a DEP as an approach for solving the original stochastic problem. Present the solved problem here in the form of the true deterministic equivalent program - as in, it is actually an equivalent to the original stochastic problem - with all the necessary expansions and additional variables. Focus on how this formulation no longer includes any stochastic variables and is purely deterministic. Mention that the trade off is that the deterministic variables are part of a infinitely large set of potential scenarios.

In order to solve the two-stage stochastic optimization formulation (Problem 1, eqs. (2.8)–(2.13)), we need to convert it to a deterministic equivalent program (DEP) that does not contain any stochastic variables (only deterministic variables) [21].

Let Ω be defined as the sample space, i.e., the set of all scenarios. Let $\hat{\Omega} \overset{\mathsf{def}}{=} \{1, 2, \ldots, O\}$ be a discrete set containing sampled scenarios. The probability a given scenario $\omega \in \hat{\Omega}$ occurs is denoted by $p^{(\omega)}$, $\omega \in \hat{\Omega}$, where $\sum_{\omega \in \hat{\Omega}} p^{(\omega)} = 1$. Variables that are dependent on the scenario are shown with a superscript (ω) with the specific scenario it is dependent on indicated by ω.

Problem 2 (Deterministic Equivalent Program of Problem 1)

$$\underset{\left\{\substack{z_s,\,\delta_{ms}^{(\omega)},\\ s\in\mathcal{S},\,m\in\mathcal{M},\\ \omega\in\hat{\Omega}}\right\}}{\text{minimize}} \left\{\sum_{s\in\mathcal{S}} c_s\, z_s - \alpha \sum_{\omega\in\Omega} p^{(\omega)} \left(\sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}} \delta_{ms}^{(\omega)}\, u_{ms}^{(\omega)}\right)\right\} \tag{3.1}$$

subject to:

$$\sum_{s\in\mathcal{S}} \delta_{ms}^{(\omega)}\, u_{ms}^{(\omega)} \leq d_m,\ \forall m\in\mathcal{M},\ \forall\omega\in\hat{\Omega} \tag{3.2}$$

$$\sum_{m\in\mathcal{M}} \delta_{ms}^{(\omega)} \leq r_s\, z_s,\ \forall s\in\mathcal{S},\ \forall\omega\in\hat{\Omega} \tag{3.3}$$

$$z_s \in \{0,\,1\},\ \forall s\in\mathcal{S}. \tag{3.4}$$

The objective function (3.1) combines both objective functions (2.8) and (2.10) of the initial formulation into a deterministic form. Constraints (3.2) and (3.3) ensure demand is not overallocated and is only allocated to selected resources and within capacity for all scenarios.

Problem 2 provides an equivalent deterministic form of Problem 1 for the (finite) sampled state space, $\hat{\Omega}$, containing O scenarios. With sufficiently large O, $\hat{\Omega}$ approaches a tight approximation of the original sample space. Within each scenario $\omega \in \hat{\Omega}$, the SSLT demand field ρ is sampled to provide a set of M discrete demand points. Each sampling of ρ is generated by creating a non-stationary 2D PPP with M points as described in Section 2.1.1.

3.1.1 Sampling Approaches

As the infinitely large set of scenarios renders the problem unable to be solved, it needs to be sampled into a finite set to be solved. Present the structure and nomenclature used to imply a sampled set of scenarios, and describe the structure of how the scenarios are sampled into a truncated set. Might be worth mentioning that there are other methods that might be better for sampling beyond the completely random sampling approach I am using. Worth consideration?

Sample Average Approximation

At what point is the sampling enough? As the set of scenarios considered within the sampled DEP increases, it more closely compares to the original DEP and the stochastic optimization problem, but it also becomes increasingly difficult to solve as the number of scenarios considered increases. So, it is beneficial to understand that a certain known number of scenarios provides a reasonably tight - what does reasonable mean? - solution to the original DEP to avoid being unnecessarily computationally expensive to solve. Finding this minimum necessary number of scenarios can be done via a sample average approximation (SAA) analysis, which should not be too complicated to do.

3.1.2 Adaptive Slicing

Now that we have a (close) approximation to the DEP and the original stochastic optimization problem, we have a method for deriving the minimum cost BS selection and adaptive slicing for the desired VWN. However, this selection is overly time consuming to constantly run, and the BSs selected for the VWN(s) by the VNB are fairly constant, so all that is needed is to dynamically (read: adaptively) slice the selected BSs to the various SPs. To do this, we simplify the sampled DEP such that it has only one scenario - ostensibly, the current scenario in time - and the BSs selected set to be a constant rather than a decision variable. The resulting problem is a single stage linear program that is much simpler to solve. This is used to adaptively slice resources to the demand.

After the solution to the sampled DEP of Section 3.1.1 has been found, the VNB has determined the joint BS selection that forms the VWN and a proposed resource slicing of considered possible scenarios, $\hat{\Omega}$, that allocates the resources to the SP's demand points. Since O is not infinite, any given scenario present in the formed VWN is unlikely to be an element of $\hat{\Omega}$. Further, as demand points move between BSs or enter or exit the VWN, a new scenario $\omega \notin \hat{\Omega}$ is formed. The VWN must adapt its resource slicing to these new demand points to maintain maximal demand satisfaction. With the VWN built, the joint BS selection, z_s, becomes a constant of the network, simplifying Problem 2 to a single-stage optimization problem.

Problem 3 (Deterministic Adaptive Slicing)

$$\underset{\{\delta_{ms},\, s\in\mathcal{S},\, m\in\mathcal{M}\}}{\text{maximize}} \left\{ \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}} \delta_{ms}\, u_{ms} \right\} \tag{3.5}$$

subject to:

$$\sum_{s\in\mathcal{S}} \delta_{ms}\, u_{ms} \leq d_m,\ \forall m \in \mathcal{M} \tag{3.6}$$

$$\sum_{m\in\mathcal{M}} \delta_{ms} \leq r_s\, z_s,\ \forall s \in \mathcal{S}. \tag{3.7}$$

It is worth noting that Problem 3 is more tractable than Problem 2 as it only contains the single continuous decision variable for resource slicing, simplifying the objective function (3.5) and constraint (3.7) from a mixed integer linear program to a linear programming problem.

3.2 Genetic Algorithm

Now that the first approach - DEP and its sampling - has been tackled, and the necessary tool to evaluate it has been derived from it - the simplified adaptive slicing program - move on to the genetic algorithm approach for approximating the BS selection process. Discuss the core algorithm of a genetic algorithm, then the various approaches that I used in its process (e.g., binary chromosomes, elitism, uniqueness, uniform crossover, bitwise mutation).

The Problem 2 formulation becomes intractable as O, S, or M increases. Most importantly, the accuracy of the sampled DEP is directly dependent on the size of $\hat{\Omega}$, O, directly causing a trade off between the accuracy of the sampled DEP and its computability in a reasonable amount of time. In this subsection, we reformulate the problem of joint BS selection for the VWN as a genetic algorithm, circumventing the need to discretize demand or to establish $\hat{\Omega}$, thereby simplifying the original problem into a more scalable form.

A genetic algorithm is an iterative metaheuristic in which an approximate solution to a given optimization problem is arrived at via a series of progressive generations. Each generation contains a number of candidate solutions, called individuals, each of which is defined by a chromosome. During a given generation, a fitness heuristic is assessed for each individual based on its chromosome. Then individuals are selected at random, with more fit individuals being selected with higher probability. Pairs of selected individuals will crossover with probability p_{xov}, a process similar to genetic recombination in biology. The resulting chromosomes then have probability p_{mut} to mutate, altering the chromosome slightly. Once enough new individual chromosomes have been selected and possibly undergone crossover and mutation, this set of new individuals, called children, forms the next generation to repeat the process.

For the genetic algorithm, ρ is not sampled for discrete demand points. Instead, we assume that all demand over the region is allocated to the closest resource. The subset of $\mathcal{S}$, $\mathcal{S}'$, that is selected for a given possible VWN forms a Voronoi tessellation from the point locations of the selected resources. The total demand allocated to a selected resource $s \in \mathcal{S}' \subseteq \mathcal{S}$ is $\iint_{V_s} \rho\,(x,\, y)\; dx\, dy$, where V_s is the region bounded by the cell of resource s in the Voronoi tessellation. If the total demand allocated to s exceeds r_s, s is considered to be *overcapacity.*

If V_s is not wholly contained within the coverage area of resource s, s is considered to be *overcoverage.*

Let $\mathcal{G} \overset{\text{def}}{=} \{1, 2, \ldots, G\}$ be the set of generations used in the genetic algorithm and $\mathcal{I}_g \overset{\text{def}}{=} \{1, 2, \ldots, I\}, g \in \mathcal{G}$ be the set of individuals within generation g. Each individual $i \in \mathcal{I}_{g \in \mathcal{G}}$ has a binary chromosome $z^{\{ig\}}$ of length S. $z_s^{\{ig\}}, s \in \mathcal{S}$, denoting each individual bit of the chromosome, is defined as follows:

$$z_s^{\{ig\}} = \begin{cases} 1, & \text{if BS } s \text{ is selected for the VWN forindividual } i \text{ in generation } g, \\ 0, & \text{otherwise} \end{cases}$$

The fitness heuristic of each individual chromosome, $z^{\{ig\}}$, is assessed as the reciprocal of the chromosome's cost, which is defined as

$$\text{fitness}\left(z^{\{ig\}}\right) = \frac{1}{\text{cost}\left(z^{\{ig\}}\right)} \tag{3.8}$$

$$\text{cost}\left(z^{\{ig\}}\right) = \sum_{s \in \mathcal{S}} \Bigg(c_s \; z_s^{\{ig\}} + c_{\text{cov}} \; \mathbb{1}_{\{V_s \not\subseteq R_s\}} + \left(c_{\text{cap}}^g - 1\right) \; \max\left(0, \iint_{R_s} \rho\,(x,\, y)\; dx\, dy - r_s\right)\Bigg) \tag{3.9}$$

where R_s is the coverage area region of resource $s \in \mathcal{S}$.

The cost function (3.9) indicates cost increases not only based on the cost of the resources selected, but also with imperfection costs c_{cov} and c_{cap}, the costs of a selected resource being overcoverage or overcapacity, respectively. The overcapacity cost grows with each successive generation. For early generations, this allows for imperfect solutions to temporarily exist to seed later generations and improve diversity to increase the probability of finding a better final approximate solution.

Elitism is used, where the n most fit individuals of a given generation are automatically selected without crossover or mutation to be the first children of the next generation. Selection occurs via the roulette wheel selection method. Every individual i of a given generation g has a probability of being selected given by

$$\frac{\text{fitness}\left(z^{\{ig\}}\right)}{\sum_{i\in\mathcal{I}}\text{fitness}\left(z^{\{ig\}}\right)}$$

When crossover is performed on selected individuals, it is via the uniform crossover method with a mixing ratio of 0.5. That is, if two selected parent individuals crossover, each equivalent bit in the parents will swap with a probability of 50%. Mutation occurs on a bit-by-bit level, with each bit mutating (i.e., flipping) with probability $\frac{1}{S}$. The uniqueness property is then enforced on the resulting children to ensure diversity; if a child chromosome is identical to another child chromosome in the next generation, the child is discarded and a new child generated, ensuring that each individual of any given generation is unique within that generation.

The genetic algorithm iterates for a number of generations G. If the genetic algorithm settles on a single individual for a number of continuous generations, G_{halt}, it will halt and present that individual's chromosome as the final approximate solution for z_s. Otherwise, the chromosome of the fittest individual of generation G determines z_s.

The genetic algorithm only determines an approximate solution to the BS selection forming the VWN, informing the VNB of which BSs to obtain from the RPs. With this selection, z_s, the SP's demand points can be dynamically allocated resource slices as described by Problem 3 in Section 3.1.2.

Chapter 4

Testing and Simulations

Start this chapter ASAP. Start running Case I data now.

In this chapter I will be introducing four different cases to test the provided approximation approaches. The first will be the test case used in my conference paper (one SP, with homogeneous resources). The second will be an expansion of the test case used in my conference paper, but with heterogeneous resources. The third will extend to service multiple similar cellular SPs. The fourth will extend to a case with multiple SPs with various, specialized demands.

4.1 VWN Construction for a Single SP

Lead into the first two cases, which test the approaches while using a single SP.

4.1.1 Case I: Homogeneous Urban Cellular Network

Start this!

Basically as presented in my conference paper. One SP, homogeneous resources within the RPs. Might need to use a new data set, though, with a larger data set.

4.1.2 Case II: Impact of Heterogeneous Resources

Same as Case I, but with heterogeneous resources within the RPs. Need to understand how this changes the approaches.

4.2 VWN Construction for Multiple SPs

Lead into the second two cases (should I have more?), which test using multiple SPs to satisfy from the same set of resources.

4.2.1 Case III: Two Similar Urban Cellular Networks

First consider a case with two SPs with similar demands. Overlapping cellular networks. Could see how the approaches behave while two SPs partially overlap.

Homogeneous Resources

If it appears that the difference between Case I and Case II (sections 4.1.1 and 4.1.2) is worth further consideration, then analyze here with homogeneous resources. Otherwise, a single comparison should be sufficient.

Heterogeneous Resources

As for the previous subsubsection (4.2.1), but consider with heterogeneous resources.

4.2.2 Case IV: SPs with Specialized Demands

This is the major case that is the extension of my work. Case I (4.1.1) analyzed what happens with a single SP, Case II (4.1.2) expanded that to heterogeneous resources, and Case III (4.2.1) added an additional similar SP, but Case IV considers when there are several SPs and with their own considerations and unique demands. Need to consider what these SPs look like. One would be a cellular network like in Case I (moderate to high number of users, moderate demand). Another could be a streaming service (few users, high individual demand). Another an emergency service (very low number of users and demand, but requiring virtually 100% demand satisfaction - see note below). What other SPs should I consider?

Note*: I need to consider how to accurately label demand satisfaction within the approaches. In effect, this would be controlled by α for the (sampled) DEP and controlled by β or some such for the genetic algorithm. I should investigate this at some point of the thesis, probably* *within their appropriate sections in chapter 3 (DEP: 3.1 and GA: 3.2).*

Homogeneous Resources

As for Case III (4.2.1), if a considerable difference was detected between Cases I and II (4.1.1 and 4.1.2), consider analyzing the case with homogeneous resources and

Heterogeneous Resources

also with heterogeneous resources.

Chapter 5

Conclusions

Consider conclusions of my work. I don't think this chapter would be long, but condense my findings into some coherent thoughts, and redirect to what they are. Also expound on some of the further work that my research could be expanded to (e.g., further use cases investigating my approaches, use of (meta)heuristics other than a genetic algorithm to approximate the optimization problem, improve the basic capacity function used in my optimization model).

Old work Update and Remove!

5.1 Considerations for Future Work

Filled as ideas come to mind during writing

– Other traffic demand models (log-normal mixtures, α-stable)

– Rate normalization that varies with distance (i.e., non-binary u_{ms})

–

Bibliography

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2016-2021," [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf, June 2017, white paper at Cisco.com.

[2] C. Beckman and G. Smith, "Shared networks: making wireless communication affordable," *IEEE Wireless Communications Magazine*, vol. 12, no. 2, pp. 78–85, April 2005.

[3] M. J. Abdel-Rahman, K. Cardoso, A. B. MacKenzie, and L. A. DaSilva, "Dimensioning virtualized wireless access networks from a common pool of resources," in *Proceedings of the IEEE CCNC Conference*, January 2016, pp. 1049–1054.

[4] 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 Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[5] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, September 2008.

[6] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, “Modeling and analysis of k-tier downlink heterogeneous cellular networks,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, April 2012.

[7] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[8] 3GPP TS 23.251, “Network sharing; architecture and functional description,” v. 14.1.0, September 2017.

[9] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, “Radio access network virtualization for future mobile carrier networks,” *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, July 2013.

[10] “Mobile network sharing report 2010-2015 - developments, analysis & forecasts,” Visiongain, Tech. Rep., 2010.

[11] J. S. Panchal, R. D. Yates, and M. M. Buddhikot, “Mobile network resource sharing options: Performance comparisons,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4470–4482, September 2013.

[12] L. Doyle, J. Kibida, T. K. Forde, and L. DaSilva, “Spectrum without bounds, networks without borders,” *Proceedings of the IEEE*, vol. 102, no. 3, pp. 351–365, March 2014.

[13] K. V. Cardoso, M. J. Abdel-Rahman, A. B. MacKenzie, and L. A. DaSilva, "Virtualization and programmability in mobile wireless networks: Architecture and resource management," in *Proceedings of the Workshop on Mobile Edge Communications (MECOMM'17)*, 2017, pp. 1–6.

[14] M. Gomez, M. B. Weiss, G. McHenry, and L. Doyle, "Matching markets for spectrum sharing," in *Telecommunications Policy Research Conference*, September 2017. [Online]. Available: http://d-scholarship.pitt.edu/33631/

[15] M. Gomez, "Secondary spectrum markets: from "naked" spectrum to virtualized commodities," Ph.D. dissertation, September 2017. [Online]. Available: http://d-scholarship.pitt.edu/33130/

[16] M. J. Abdel-Rahman and M. Krunz, "Stochastic guard-band-aware channel assignment with bonding and aggregation for DSA networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3888–3898, July 2015.

[17] M. J. Abdel-Rahman, M. AbdelRaheem, A. B. MacKenzie, K. Cardoso, and M. Krunz, "On the orchestration of robust virtual LTE-U networks from hybrid half/full-duplex Wi-Fi APs," in *Proceedings of the IEEE WCNC Conference*, April 2016.

[18] M. J. Abdel-Rahman, M. AbdelRaheem, and A. B. MacKenzie, "Stochastic resource allocation in opportunistic LTE-A networks with heterogeneous self-interference cancellation capabilities," in *Proceedings of the IEEE DySPAN Conference*, September/October 2015, pp. 200–208.

[19] N. Y. Soltani, S. J. Kim, and G. B. Giannakis, “Chance-constrained optimization of OFDMA cognitive radio uplinks,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1098–1107, March 2013.

[20] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, “Joint chance-constrained predictive resource allocation for energy-efficient video streaming,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1389–1404, May 2016.

[21] P. Kall and S. W. Wallace, *Stochastic Programming.* John Wiley and Sons, 1994.

[22] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 3rd ed. Springer Publishing Company, Incorporated, 2000.

[23] T. Cui, R. Bai, A. J. Parkes, F. He, R. Qu, and J. Li, “A hybrid genetic algorithm for a two-stage stochastic portfolio optimization with uncertain asset prices,” in *2015 IEEE Congress on Evolutionary Computation (CEC)*, May 2015, pp. 2518–2525.

[24] U. Gotzner and R. Rathgeber, “Spatial traffic distribution in cellular networks,” in *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, May 1998, pp. 1994–1998, vol. 3.

[25] M. Michalopoulou, J. Riihijrvi, and P. Mhnen, “Towards characterizing primary usage in cellular networks: A traffic-based study,” in *Proceedings of the IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, May 2011, pp. 652–655.

[26] D. Lee, S. Zhou, and Z. Niu, "Spatial modeling of scalable spatially-correlated log-normal distributed traffic inhomogeneity and energy-efficient network planning," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 1285–1290.