RICE UNIVERSITY

Throughput Maximizing and Service Provisioning Strategies for Millimeter Wave Networks

by

David Ramirez

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Behnam Aazhang, Chair
J.S. Abercrombie Professor of Electrical and Computer Engineering, Chair

Edward Knightly
Professor of Electrical and Computer Engineering

Illya V. Hicks
Associate Professor of Computational and Applied Mathematics

Anthony Ephremides
Distinguished University Professor, University of Maryland

Houston, Texas
December, 2016
ABSTRACT

Throughput Maximizing and Service Provisioning Strategies for Millimeter Wave Networks

by

David Ramirez

Migrating wireless communications to millimeter wave (mmWave) frequencies can satiate the growing demand for higher throughput and delay sensitive service provision. Due to high path loss and poor penetration, mmWave communications are challenged by limited coverage and blockage. Dense access point (AP) deployment and beamforming can enable mmWave networks to increase coverage and combat blockage. Greater AP density gives users diversity in AP selection, and thus a user can probe multiple APs to opportunistically transmit over the best channel. Yet, the opportunity to transmit to a previously probed AP may be lost due to inherent network properties (e.g. blockage, deafness, or decentralized scheduling), and the overhead delay and cost of probing more APs may also be detrimental to the search itself.

We analyze the impact of AP diversity, beamforming, and overhead in opportunistic mmWave networks in terms of throughput and service provision capabilities. Decentralized opportunistic solutions are presented within a model that accounts for overhead delay and overhead bit cost. We assume no a priori knowledge of channel conditions, thus a deterministically optimal solution is unattainable in non-trivial scenarios. Stochastically optimal strategies are proposed and genie-aided solutions
are presented as tight upper bounds. Ultimately, we present a model which includes the inherent properties of mmWave networks and obtain opportunistically optimal strategies for throughput and service provisioning.

Conditions under which throughput maximization is a sub-martingale are presented, thus the problem is approached as a finite horizon stopping problem with unreliable recall. The optimal opportunistic strategy is a set of a priori computable thresholds. Bounds on average overhead, average delay, and average performance bound of the stopping strategy are presented. The throughput performance of optimal opportunistic strategies and the impact of imperfect network measurements are evaluated via simulations.

Service provision maximization is considered under strict delay or average delay constraints. Both constraints lead to unique non-linear knapsack problems, yet reformulations into linear optimization problems are obtained by expanding the variable space. Optimal probing orders, for specific network conditions, are presented and the intuition of the optimal orders is leveraged for general network conditions. Opportunistic strategies are presented and evaluated via simulations against computationally intense stochastic programming solutions and impractical genie-aided tight upper bounds.
Todo cabe en un jarrito, sabiéndolo acomodar.
Acknowledgments

Behnaam Aazhang has been instrumental in keeping me motivated, focused, and happy throughout my time at Rice University. Alongside my advisor, I must thank many professors and administrators for doing their part. Edward Knightly for teaching me more about wireless networks than I know. Illya Hicks for reminding me about the beauty and power of mathematics. Ashutosh Sabharwal for giving me fearless lessons. Ned Thomas for inspiring me to aim higher. Seiichi Matsuda for listening to me and supporting my every attempt to make Rice even more amazing. To Carrie, Caroline, Jennifer, Ann, Kathryn, Ben, Adria, and many other friendly faces that make my life brighter and easier. To Jan Hewitt for making me the presentation snob I have proudly become.

My time at Rice allowed me to make many amazing friends whom I wish to thank. Matt Nokleby for being a wonderful mentor who pushed me to keep a good pace at Rice and on the running trail. Gareth, Sam and Corina for being an amazing trio of senior students that helped me grow. Shuqiao for being the best academic younger sister I could ever ask for. Rakesh, Zhiting, Sudha, Joe, Boqiang and Shi for bringing so much to the group, go team! Pedro for inspiring the *futból* banter. Oscar for guiding me beyond the hedges. Evan for showing me what genuine friendliness looks like. David (i.e. Kao) for code sharing our name. Adriana for keeping me connected. Ryan and Rose for letting me be a part of so many wonderful events. Chris for being a lower bound to several of my weightlifting records. Mahdi for optimally time sharing our office. To everyone in RCEL and SHPE for allowing me to be a part of such amazing groups. Sarah for feeding my dragons. CJ for giving me a sense of purpose and community when I most needed it. Robert for showing me what dedication looks
like. Su for sharing what is most precious to us. Last and undoubtedly not least, to Jason for making Turk and J.D., by comparison to us, look like people in a courteous professional relationship.

De igual forma agradezco a tantas personas que han sido parte de mi vida desde antes de Rice. Puff por siempre hacerme sentir parte de algo mayor. Moyos por compartir su visión de altura. Charles for all the unfiltered commentary. Charol por ser mi Hefestion. Marcelle por ser mi Mortz. Decenas más de amigos que me han apoyado desde lejos y, cuando la ocasión permite, desde cerca. INSA Lyon DTC était incroyable, merci à Sam, Bruno, et tous les INSAliens. A mi famille lyonnaise por cada cascada, patinada, y aventura que hemos compartido. A Crong por ser mi crong. A mi Houston Mom, Umbe Cantú, por darme cariño y apoyo sin medida. A Josué por ser mi mejor cuñado. A Claudio por ser mi mejor amigo. A Carolina por ser mi mejor flepper. A Nour por ser mi mejor madruja. A mi abuelito por inspirarme. A mis abuelitas por darme tantas sonrisas. A mi Mamá y Papá por todo y nada menos.
## Contents

Abstract ................................................. ii
Acknowledgments ........................................... v
List of Illustrations ....................................... ix
List of Tables ............................................. xii

1 Introduction ............................................. 1
   1.1 State of the Art ....................................... 5
   1.2 Contributions ....................................... 11
   1.3 Disclaimer .......................................... 14

2 Network Model ........................................... 15

3 Throughput Maximization ................................ 21
   3.1 Throughput Maximization Model ....................... 22
      3.1.1 Throughput Maximization in mmWave Networks .... 26
      3.1.2 mmWave Effective Throughput as a Sub-Martingale ... 29
      3.1.3 mmWave Optimal Transmission Strategy ............. 31
      3.1.4 Properties of Proposed Strategy .................... 34
   3.2 Performance Bounds .................................. 37
      3.2.1 Expected Delay and Overhead ..................... 38
      3.2.2 Throughput Performance Bound ..................... 39
   3.3 Numerical Results ................................... 42
   3.4 Summary ............................................. 48
4 Service Provisioning

4.1 Service Model .................................................. 51

4.2 Service Provisioning with Strict Delay .................. 58

4.2.1 Optimal Schedule for Strict Delay ...................... 60

4.2.2 Opportunistic Strategies .................................. 63

4.2.3 Stochastic Optimization Strategies ...................... 67

4.2.4 Numerical Results .......................................... 70

4.2.5 Conclusion .................................................. 75

4.3 Service Provisioning with Average Delay ............... 77

4.3.1 Optimal Schedule for Average Delay Constraint and Large Transmission Times ......................... 80

4.3.2 Optimal Average Delay Service Scheduling with Probing Cost ............................................. 85

4.3.3 Practical Strategies for Providing Services with Average Delay .............................................. 88

4.3.4 Stochastic Programming ...................................... 89

4.3.5 Optimal Stopping Strategy for Average Delay Service Provisioning ........................................... 92

4.3.6 Discussion of Practical Strategies for Average Delay Service Provisioning ................................ 96

4.3.7 Numerical Results ........................................... 97

4.3.8 Summary .................................................. 107

5 Conclusion and Future Directions ........................... 108

Bibliography .................................................... 112
Illustrations

2.1 Model network with $N = 4$ APs and two users with graphical representation of beamwidth $\phi_n$ from AP $n$ a distance $d_n$ from a user intending to transmit. While the user searches for an appropriate AP, other users may cause blockage or start communicating with previously probed APs.

3.1 Avg. throughput as a function of $B$ for $N = 10$. The proposed strategy obtains $\approx 20\%$ increase compared to a single probing strategy.

3.2 Avg. throughput as a function of available APs $N$ with $B = 0.3$. For larger $N$ a higher throughput may be achieved at the cost more probed APs.

3.3 Ratio of the average delay against the maximum delay, i.e. $\frac{\sum_{i=1}^{N} \gamma_i}{\gamma_{\text{max}}}$.

3.4 CDF of transmission occurring against total number of probed APs for $N = \{5, 10\}$. A lower CDF implies a lower average delay and overhead.

3.5 CDF of transmission occurring against total number of probed APs for $N = \{15, 25\}$. A lower CDF implies a lower average delay and overhead.
4.1 Scheduling period example with collocated users, each with their service requests, and a single AP. After the polling phase, the AP selects one of the \( N! \) possible probing orders during the ordering phase. For exposition, the impact of two schedules are shown during the execution phase on the right. While the schedules \( S_{j,1} \) and \( S_{j,i} \) use the same transmission vector \( x_j = \{1, 0, 1, 1\} \), the distinct probing orders result in distinct average service delay. For sake of the example, assume that \( \tau_1(2) \gg 0 \) such that the second user does not transmit. Note that \( \tau_1(2) = \tau_i(3) \). While both schedules have the same sum probing and transmission time, the average completion time of \( S_{j,i} \) is lower due to the order in which transmissions occur.

4.2 Example of services provided with a strict delay for \( N = 4 \),
\[
t = \{\frac{T}{4}, \frac{T}{3}, \frac{T}{4}, T\}, \quad \gamma_n = T/6 \quad \forall \ n, \quad v = \{3, 2, 3, 1\}, \quad x^* = \{1, 0, 1, 0\}, \quad \text{and} \quad V(S^*) = 6.\]
Service 1 and 3 are transmitted, while user 2 only probes, and user 4 has no time to probe or transmit, thus \( z_i = 3 \).

4.3 Average \( \sum_{i=1}^N x_i v_i \) as function of \( N \) with \( v_i \in \{1, 2, 3\} \quad \forall \ i \).

4.4 Average delay of completed services as function of \( N \).

4.5 Average \( \sum_{i=1}^N x_i v_i \) as function of \( N \) with \( v_i \in \{1, 5, 10\} \quad \forall \ i \).

4.6 Average \( \sum_{i=1}^N x_i v_i \) as a function of \( \frac{T}{y} \).

4.7 Average payoff as a function of \( N \). As \( N \) grows the preferred strategy changes from greedy, to dynamic, to mean stopping.

4.8 Average throughput as a function of \( N \). If we consider the performance and computational cost, as \( N \) grows the preferred strategy changes from greedy, to stochastic programming, to Dantzig stopping strategy.
4.9 Average throughput per scheduling period, i.e. $\sum_{i=1}^{N} b_i x_i$, as a function of $N$. The Dantzig stopping strategy achieves the highest throughput among practical strategy, although not achieving the highest sum service value (see Fig. 4.7).

4.10 Average Payoff as a function of transmit power $P$. All strategies, except the Dantzig stopping strategy, exhibit a nearly linear increase in performance with transmit power.

4.11 Average Payoff as a function of probing length $\gamma$.

4.12 Average throughput per scheduling period as a function of probing length $\gamma$. 
4.1 Portion of the solution table for $d(v, n)$ with values $t = \{\frac{T}{2}, \frac{T}{3}, \frac{T}{5}, \frac{T}{10}\}$ and $v = \{2, 3, 2, 1\}$. Combinations of $(v, n)$ with no solution (e.g. $d(1, 1)$) are marked “No Soln.”

4.2 Update of Table 4.1 for $d(v, n) + \sum_{i=1}^{n} \gamma_i$ with $\gamma_n = \frac{T}{5}$ $\forall$ $n$ and infeasible schedules crossed out. Combinations of $(v, n)$ with no solution (e.g. $d(1, 1)$) are marked “No Soln.”

4.3 Scheduling Strategies

Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Portion of the solution table for $d(v, n)$</td>
</tr>
<tr>
<td>4.2</td>
<td>Update of Table 4.1 for $d(v, n) + \sum_{i=1}^{n} \gamma_i$</td>
</tr>
<tr>
<td>4.3</td>
<td>Scheduling Strategies</td>
</tr>
</tbody>
</table>

61, 62, 98
Chapter 1

Introduction
Networks have been developed with the goal of seamless service provision, but limited resources and the stochastic nature of channels make service provision over wireless networks a challenging task. Future wireless networks will be required to meet stricter delay constraints, achieve higher rates, and provide connectivity for much denser networks than contemporary wireless networks [1, 2, 3]. Next generation wireless networks are tasked with delivering services within milliseconds and gigabits per second rates to many more users and a growing plethora of internet enabled things. To achieve these goals network resources must be expanded and network resource allocation must be tailored to intrinsic properties of future networks [4, 5].

The resources available to a network can be greatly expanded by leveraging the abundant frequency in millimeter wave (mmWave) bands, for example the unlicensed 2 GHz band centered at 60 GHz frequency [6]. A larger frequency band can enable higher throughput, which in turn can reduce transmission times and thus reducing latency. With transmissions requiring less temporal use of resources a denser deployment can be supported by future networks. The transition into mmWave networks has been recognized as an enabler for future networks to achieve faster rates, lower latency, and higher density deployments [7, 8, 2, 9, 1].

Parallel to academic research, the IEEE 802.11ad standard for 60 GHz frequency transmissions was published in 2012 [10], and the first 802.11ad certified router became commercially available in 2016 [11]. In July 2016 the Federal Communications Commission (FCC) released close to 11 GHz of band in the upper microwave and
mmWave range, of which 7 GHz were deemed unlicensed [12]. The 802.11ad protocol, commercial developments, and decision of the FCC are evidence of the growing interest of transitioning wireless communications to higher frequencies.

A move into mmWave bands faces two obstacles: reduced coverage and blockage. From Frii’s Law, transmitting at a higher frequency incurs proportionally higher path loss, thus coverage is reduced relative to lower frequency transmissions. Blockage is the high signal attenuation due to the inability of mmWave to penetrate many common objects, e.g. 20–35 dB from penetrating a human body [13, 14, 15, 16]. Beamforming can extend coverage of a single device, and high density deployments of co-located access points may help avoid blockage [17].

Beamforming allows signal power to be focused in a specific direction by controlling the phase of multiple antennas. Generally speaking, increasing the number of antennas allows for tighter control over the directionality of the beam, increasing beamforming gains. A benefit of mmWave devices is that, for a fixed antenna gain, antenna form factor decreases proportionally with frequency. A $4 \times 4$ antenna array at 28 GHz and half wavelength spacing would occupy an area of $1.5\text{cm} \times 1.5\text{cm}$, which is roughly the same area of a single 2.4 GHz antenna. Thus, mmWave devices can leverage highly directional beamforming to overcome path loss [18, 8, 19].

A densely deployed mmWave network can provide the higher transmission rates required for future networks. Yet, to properly provide delay-constrained services an increase in rate may not be sufficient. Users will tolerate only so much delay before
being provided a service. For example, a two-second buffering delay leads users
to abandon high-quality video streams [20, 21] and causes a jump in users’ stress
levels [22]. Managing delay can be done by improving network resource allocation
[23]. Furthermore, resource allocation should be done with careful consideration of
network properties and service requirements.

A strict transmission time delay is necessary for certain future services, e.g. tactile
internet, yet the network service provisioning capabilities are greatly taxed when
attempting to provide strict delay services for many users. An average completion
time delay is a less demanding constraint which can lead to higher service provisioning
capabilities and adequate for services where some delay may be tolerated by the user,
e.g. a video service can have a smooth playback if the average delay is less than the
remaining buffered video.

The use of beamforming enables high transmission rates and adequate coverage
for mmWave networks, while simultaneously placing a non-negligible burden on any
scheduling attempt. For example, selecting a beam pattern may incur delays from tens
to hundreds of milliseconds [24]. Furthermore, an increase in transmission rate for a
fixed information packet size leads to a shorter transmission time and a proportionally
larger overhead delay.

In this thesis, we analyze the impact of non-negligible overhead on mmWave
networks with the goals of maximizing service provision or maximizing. Service
provisioning is considered under either a strict delay or average delay constraint.
Throughput maximization is achieved in the stochastic sense for a general model that encompasses both a fixed transmission time or a fixed transmission and probing time. Optimal solutions under genie aided assumptions are obtained as useful upper bounds, while practical strategies are formulated, analyzed, and evaluated. To place our work in context, the state of the art in the area is now discussed.

1.1 State of the Art

As mentioned before, two obstacles to mmWave communications are reduced coverage and blockage. A higher frequency transmission results in higher path loss and blockage occurring from higher penetration loss with many common objects [13, 14, 15, 16]. Higher path loss can be alleviated via beamforming and blockage is addressed via a dense access point (AP) deployment [17].

To enable beamforming, a user and an AP go through an initial access phase which establishes the physical link by obtaining channel knowledge and selecting an appropriate beamform. Beam selection procedures proposed and evaluated in [25, 26, 18, 27] produced relatively efficient access methods with a small, yet non-negligible, overhead. Work in [28] proposed and evaluated the use of lower frequency band communications to reduce overhead associated with beam steering overhead present in mobile mmWave networks. Analysis in [24] shows that the initial access overhead alone can range from tens to hundreds of milliseconds. The impact of beam selection overhead on the throughput of a mmWave network was studied in [29], and
more generally reconfiguration delay (e.g. initial access) was shown to reduce the stable throughput region [30].

Note that configuration delay is present in non-mmWave networks. For example, multiple antenna systems can incur a configuration delay of up to $745\mu s$ [31], phased-lock loops in oscillators may incur delays of $200\mu s$ [32], and certain networks (e.g. satellite networks) have an unavoidable non-negligible propagation delay (e.g. satellite communication systems).

Blockage may be avoided by a dense deployment of co-located APs, which also improves network coverage [33]. Throughput scaling of a multi-tier mmWave network can be achieved despite blockage via a sufficiently dense AP deployment [34]. Adding APs increases transmit diversity, thus increasing the likelihood of finding a non-blocked AP and in turn increasing throughput. Analysis on the coverage-rate tradeoff for a mmWave network with blockage was conducted in [35]. The system proposed in [36] for mmWave devices can quickly detect blockage and immediately select the best alternative direction for the transmission.

Throughput maximization for mmWave networks has focused on personal area networks. Work in [37] proposed a randomized allocation of so called exclusive regions, i.e. randomized scheduling of spatial areas for transmissions to occur. The pseudo-wired model, as proposed in [38], suggests that the highly directional transmissions required for mmWave networks greatly reduces the impact of interference. Work in [39] presents a computationally complex optimization that leverages density to relay
around blockage. A mmWave network utility maximization model is presented in [40] where tractability is gained by upper and lower bounding the impact of interference. Work in [41] considers users, APs, and back-haul links operating in mmWave and proposes various scheduling algorithms.

Resource allocation criteria dictate the partition of network resources. Criterion selection can focus on the interests of network operators or of users. With the network operator’s interest in mind, one can aim to increase the information carrying capacity of a network by utilizing a rate maximization criterion [30, 42, 43, 44, 45, 46, 47, 48], in which resources are allocated based on which users can most increase their rate. Yet rate maximization may lead to few users enjoying minimal delay, while many other users despair in their wait. Alternatively, fairness criteria [49, 50] distribute resources to under-served users with the goal of providing similar service to all users. Partitioning resources fairly to service all equally can result in users suffering equal, yet still unacceptable, delay [51].

The completion time of a service accounts for the configuration delay and transmission time associated to said service plus the configuration delay and transmission time of every previous service in the schedule. Completion time has been shown to be the appropriate criterion to deliver service as expected by users [52]. Partitioning network resources by focusing on completion time ultimately leads to less delay for serviced users. In [53] a Lyapunov optimization algorithm sacrifices throughput to meet worst-case delay constraints. Generally, worst-case delay can be bounded by
the use of network calculus for deterministic networks [54, 55]. Network calculus has been proposed to provide quality of services guarantees over circuit switched networks [56]. A stochastic approach to network calculus was presented in [57], yet the so-called stochastic network calculus does not fully transfer all results from traditional network calculus to a stochastic regime as shown in [58]. The scheduling algorithm in [59] uses frame based scheduling to allocate resources to service requests with heterogeneous delay constraints. The utility maximization problem for static channels is analyzed in [60] and time-varying channels is analyzed in [61] where an auction mechanism is employed to guarantee fairness among users.

Previous work considering average completion time or average delay [62, 63, 64, 65, 66] leveraged Little’s Law, which relates the average delay, average arrival rate, and average queue size. Little’s Law allows one to focus on arrival rate and queue size by guaranteeing, via triangular relationship, a certain delay. Applying Little’s Law to systems with preference or systems where items may be dropped leads to an obtuse view of the system [67]. A system with preference may have unfavored items stay in the system for prolonged periods of times, thus limiting the time scale on which Little’s Law applies. Additionally, Little’s Law assumes that all items entering the system are serviced before departing, thus unamenable for a network in which a packet is dropped due to excessive delay.

A common assumption in previous work is that transmission time of services is freely available at the start of a scheduling period. With full knowledge of transmis-
sion times, scheduling problems are often modeled by the classical knapsack problem described as finding a subset of items to maximize the sum item value given a weight constraint on the items that may be selected [68]. The knapsack problem is known to be NP-hard, yet efficient near-optimal heuristics exist in the literature [69, 70]. One such heuristic is Dantzig’s Heuristic which selects the higher value-to-weight items until the knapsack is full.

Two non-linear variations of the knapsack problem of note are the collapsing knapsack and the expanding knapsack [71]. The collapsing knapsack problem assumes that the knapsack is constrained by a non-increasing function of the number of items so far included [72]. Work in [73] showed how to reduce a collapsing knapsack problem with $N$ items to a classical knapsack problem with $2N$ items. Opposite to a collapsing knapsack, an expanding knapsack is constrained by a non-decreasing function of the number of items.

For a fixed amount of information, an increase in transmission rate (e.g. rates obtained by leveraging the abundant mmWave band) leads to a decrease in the transmission time. A shorter transmission time leads to a proportionally larger overhead delay (e.g. due to beam pattern selection [24]), and thus our work assumes that transmission times are not readily available to allocate resources. To fully exploit mmWave networks one should carefully limit just how much overhead cost can be had before committing to establishing a single mmWave transmission.

Scheduling a mmWave network is daunting, considering the required network den-
sity [74] may lead to insurmountable network overhead, and that broadcast messages may go unheard due to high path loss [75]. Centralized scheduling is feasible for reasonably sized networks with broadcast capabilities, yet centralized coordination could fail to fully exploit a mmWave network.

Increasing AP density to avoid blockage heavily burdens mmWave scheduling attempts. To avoid the control overhead associated with scheduling, decentralized opportunistic strategies have been proposed. Opportunistic strategies wait for the right opportunity to occur before transmitting. Previous work has considered opportunistic strategies for lower GHz frequencies [76, 77]. Work by [78] proposed a sequential channel probing until the signal-to-noise-ratio exceeded a threshold, where the threshold was independent of the number of channels already probed. Similarly, [79, 80] propose opportunistic strategies for throughput maximization with fairness considerations. In [81, 82] delay constraints are applied to opportunistic approaches to satisfy unique service requests.

Previous work modeled the opportunistic communication problem as the well-studied optimal stopping problem without recall [83, 84]; where “stopping” is analogous to transmitting and “recall”, to being able to transmit on a previously probed channel. Previous work also often assumes an infinite number of probings are available provided that a user waits for channel conditions to change completely before probing again. However, 5G specifications focus on low latency communications [2, 3]. Probing \textit{ad infinitum} is untenable for any delay-conscious user. Thus, after observing
sufficient number of channels a delay-conscious user may decide to recall a previous channel, but to do so the inherent properties of a mmWave network must be considered.

1.2 Contributions

The goal of my work has been to provide insight to optimizing throughput and service provisioning in networks with a configuration delay and particularly mmWave networks. Thus, the proposed strategies do not assume that transmission time (i.e. channel state) information is readily available at any network element. Genie aided scenarios are considered only to obtain performance upper bounds and insights for practical strategies.

On the topic of service maximization, I analyze strict delay constraint and average delay constraints for wireless networks with non-negligible overhead. Service maximization refers to maximizing a payoff gained from completing packetized transmissions while meeting a temporal constraint. By assuming that transmission times are not readily available leads to the development of overhead conscious strategies which differ from those in the literature.

For service maximization with average delay constraint the use of Little’s Law is inappropriate since the system considers preference and some users may go unserviced. Service preference is derived from the notion that at a given time a user may greatly prefer a video stream over a social media update, and thus the network should be
incentivized to allocate resources accordingly via service values associated to each service requested made by a user. Without Little’s Law, instead I propose a wireless network optimization framework to maximize sum service value under an average completion time constraint. In essence, I guarantee that at each network instance the average delay is constrained, while Little’s Law guarantees that the average delay over many instances is constrained.

Constraining the average completion time results in a non-linear constraint that goes from non-decreasing when services are completed before the average delay constraint, to non-increasing as services are completed after the average delay constraint. Essentially, a completion time below the average constraint will bring down the average for the remaining services, thus allowing for a later final transmission. Alternatively, services completed after the average constraint pull the average up. Thus optimizing with an average delay constraint leads to an expanding-then-collapsing knapsack. Furthermore, while the sum completion time of multiple services is only impacted by the completion time of previous services, the average completion time of multiple services is impacted by, both, the individual completion time of previous services and the order in which previous services were completed.

I show that a reduction in the feasible set of solutions for the expanding-then-collapsing knapsack is possible without sacrificing optimality in the reduced set of solutions. A transformation from the non-linear knapsack problem to a classical knapsack problem that can be solved under a genie aided scenario is also presented.
As practical strategies for service maximization, stochastic programming and stopping strategies are presented.

Considering the intrinsic properties of mmWave networks, I consider a scenario where user may leverage a finite number of co-located APs to maximize the expected throughput. Having a finite number of APs to potentially probe fixes a hard limit on the delay before transmitting while considering mmWave AP diversity. A user may recall a previously probed AP as the intended destination for a transmission. Conscious of the probability of blockage and the probability of another user transmitting to a previously probed AP, I assume these APs may be available or unavailable to the user in the future with some known probability.

I analyze a generalized model for either fixed transmission time or a fixed time for probing and transmission. An existence proof of the optimal stopping strategies assuming only independence between the statistical distribution of transmission times is presented. An optimal stopping strategy is obtained via thresholding against a set of thresholds which are computable a priori to network realizations. Motivated by the importance of delay, bounds on average delay and throughput are presented. Finally, by focusing on two corner cases for recall capabilities, a performance lower bound due to inaccurate network measurements is presented.

I offer a new look at the opportunistic access strategies to consider AP diversity and unavailability of previously probed channels, both key properties of mmWave networks. I propose to unchain from a centralized scheduler which greatly taxes
network resources, and instead promote the use of optimal opportunistic access which can unleash the throughput needed for future mmWave networks.

1.3 Disclaimer

The results in this thesis and in large part the text of this thesis are adapted from:

D. Ramirez and B. Aazhang, “Service Centric Scheduling with Strict Deadlines”, *IEEE Global Communications Conference 2015*

D. Ramirez, L. Huang, Y. Wang, and B. Aazhang, “Optimal Opportunistic Transmissions Over Directional mmWave Channels”, *IEEE Personal Indoor and Mobile Communications 2016*

as well as the manuscripts currently under review:


D. Ramirez, L. Huang, Y. Wang, B. Aazhang, “On Opportunistic mmWave Networks with Blockage,” *submitted to IEEE Journal on Selected Areas in Communication Millimeter Wave Communications for Future Mobile Networks,*
Chapter 2

Network Model
Consider a dense wireless network operating in mmWave frequencies such that transmissions are directional and the probability of blockage is non-negligible. User devices and APs have multiple antennas and beamforming capabilities. While antennas may be plentiful on mmWave devices, user devices are limited to a single RF chain due to space and power constraints. Therefore, users are capable of transmitting to only one AP at a time while APs may establish multiple simultaneous transmissions (e.g., multi-user beamforming).

No centralized coordination exists between users; thus transmissions occur opportunistically from users to APs and vice versa. To establish a general model we use the term network element to refer to either a user or an AP. In the following subsections we adopt a more precise language to define which elements are transmitting and which are receiving.

For an arbitrary network element intending to transmit assume $N$ potential destinations, i.e., other network elements, are in communication range. A dense network deployment, as required for mmWave networks [7, 8, 2], allows the assumption $1 \leq N$. The value of $N$ can be known via a discovery algorithm, or for example if APs have fixed locations known to a location-aware user [3].

A probing phase, which allows channel conditions to be learned and a beam selection algorithm to be executed [25, 26, 18, 27], is required to establish a transmission between network elements. A beam selection algorithm searches for an adequate beam width to enable a highly directional transmission. Without loss of generality,
label all $N$ potential destinations in the order in which they are probed by the transmitting element, e.g. if a user is probing APs, then the AP that is probed first is AP 1. Define such a labeling as the ordered set $\mathbf{N} = \{1, ..., N\}$ and $n \in \mathbf{N}$.

Define the overhead cost $\delta_n$ as the total amount of information transmitted during the probing and beam searching phases when probing element $n$. Note that $\delta_n$ depends on the number of available beam widths and the selected beam searching method [29, 35]. The available beam widths in turn depend on the number of available antennas at an AP. Define the lower and upper bounds in overhead cost, respectively, as $\delta_{\min}$ and $\delta_{\max}$, such that $0 \leq \delta_{\min} \leq \delta_n \leq \delta_{\max} < \infty \ \forall \ n$.

Considering that during the beam selection phase time intervals elapse where no information is transmitted, e.g. short interframe space (SIFS), we must model the temporal duration of the probing and beam selection period to adequately capture the overhead burden. Define $\gamma_n$ as the total time duration of the probing and beam selection phases when probing AP $n$. Furthermore, define the lower and upper bounds of $\gamma_n$, respectively, as $\gamma_{\min}$ and $\gamma_{\max}$, such that $0 \leq \gamma_{\min} \leq \gamma_n \leq \gamma_{\max} < \infty \ \forall \ n$.

A user must probe at least one AP, but the user is not required to probe all $N$ APs. Users can be strongly motivated to probe more APs if the transmission rate to a probed AP is considerably low. After probing (i.e. after an appropriate beam pattern has been selected) a transmission may be established at the highest transmission rate possible with negligible error. We consider the information theoretic rate as an upper bound to a practical transmission rate. Define the transmission rate from to network
element $n$ as

$$r_n = \log \left( 1 + P_n \frac{h_n}{N} \right),$$

(2.1)

where $N$ is the noise power and $h_n$ the channel gain. $P_n$ is the received power defined as

$$P_n = \frac{P g_0 g_n}{d_n^\alpha},$$

(2.2)

where $P$ is the transmit power, $d_n$ is the distance from transmitter to the receiver, $\alpha$ is the path loss exponent, while $g_0$ and $g_n$ correspond to, respectively, transmitter and receiver beamforming gains. Under a flat-top model and assuming aligned beam directions between transmitting and receiving elements, define the beamforming gain as [85]

$$g_n = \eta_n \frac{2\pi - \phi_n}{\phi_n},$$

(2.3)

where $\eta_n$ is the beamforming efficiency and $\phi_n$ is the beam width of element $n$. Define the lower and upper bounds to the beam width, respectively, as $\phi_{\text{min}}$ and $\phi_{\text{max}}$ such that $\phi_{\text{min}} \leq \phi_n \leq \phi_{\text{max}} \forall n$. Note $g_0$ follows a similar definition as (2.3), with appropriate notational changes.

Beam width $\phi_n$ depends on the employed beam searching method. For example, a non-exhaustive beam search may terminate with a beam width $\phi_n > \phi_{\text{min}}$ at an overhead cost $\delta_n < \delta_{\text{max}}$. We assume that the rate $r_n$ is a monotonic non-decreasing function of $\delta_n$. A network example with graphical representation of $\phi_n$ is shown in Fig. 2.1. Note $\gamma_n \geq \frac{\delta_n}{r_{\text{base}}}$ where $r_{\text{base}}$ is the base rate at which overhead bits are transmitted, and the inequality follows from time periods without transmission
Figure 2.1: Model network with $N = 4$ APs and two users with graphical representation of beamwidth $\phi_n$ from AP $n$ a distance $d_n$ from a user intending to transmit. While the user searches for an appropriate AP, other users may cause blockage or start communicating with previously probed APs.
during beam searching, e.g. SIFS periods. For IEEE 802.11ad, the base rate is two orders of magnitude slower than the maximum supported rate [10]. Unless otherwise noted, we assume a priori known $\delta_n, \gamma_n \forall n$.

Our model is ultimately focused on when transmissions should occur, and not in the direction of which they occur. This slight on directionality is enabled by the highly directional transmissions required for mmWave networks. While our model will be amenable to both an uplink and downlink scenario, for ease of exposition we use the uplink scenario when considering throughput maximization and the downlink scenario when considering the service maximization problem. While the presented model can be used to address either uplink or downlink, the reader may find ease of understanding by assuming users are interested in obtaining a higher throughput while APs are focused on providing services in a timely manner.
Chapter 3

Throughput Maximization
3.1 Throughput Maximization Model

We now discuss aspects of our model that are particular for throughput maximization. Assume an arbitrary user in the network with the intention to transmit must select one AP out of $N$ APs in communication range. A user seeks to transmit to one AP, thus a user must probe at least one AP. If a probed AP offers “too low” a transmission rate, then a user may be motivated to probe multiple APs. Ultimately we mathematically define how low a transmission rate is to be deemed “too low” by the user.

After $n$ probings, distinguish APs labeled $k \in \{1, \ldots, n-1\}$ as previously probed APs. Define the act of transmitting to AP $k$ as recalling. When a user recalls AP $k$, there exists a probability that AP $k$ may be unavailable. This probability exists due to three factors evoked from intrinsic properties of mmWave networks. First, due to network density there exists a probability that a previously probed AP has begun communicating with another user in the network and becomes entirely unavailable. Second, after $n$ probings the rate $r_k$ may be unavailable due to interference present in mmWave networks, e.g. nodes unaware of ongoing transmissions could themselves transmit and cause interference [75]. Third, and most critical in mmWave networks, the link may be unavailable due to blockage. A mmWave link may face blockage due to user movement or objects unrelated reducing the rate $r_k$ to zero [15]. The combination of these three effects leads us to consider that the rate of previously probed APs may become unavailable.
A previously probed AP is unavailable with probability $B$ such that $0 \leq B \leq 1$ and available with probability $1 - B$. Note that $B$ is the probability of a channel being unavailable, and not only the probability of a channel being in blockage. When a user is unable to transmit to a previously probed AP, the user will instead transmit to the last probed AP. Consequentially, we define the highest transmission rate as a function of the probing stage $n$ as

$$\rho_n = \begin{cases} \max\{\rho_{n-1}, r_n\} & \text{with probability } 1 - B, \\ r_n & \text{with probability } B, \end{cases}$$

(3.1)

where the dependence of $\rho_n$ on $B$ is omitted to simplify notation, and for completeness define $\rho_1 = r_1$.

We assume that statistical knowledge of the channel gains $h_n \forall n$ is available to the user. The statistical distribution of the achievable rates depends on the statistical distribution of channel gains. Under our assumption that the statistical distribution of $h_n$ is known, the distribution of $r_n$ is known as well. Meaning that the cumulative distribution function $F_n$ and the derivative $dF_n$ are known, and $E[r_n]$ can be calculated. Ultimately we are interested in the statistical distribution describing the achievable rates.

Our model is general enough to consider two transmission modes for mmWave networks. The first with fixed transmission time and the second with fixed time for, both, transmission and probing. To model a fixed transmission time (i.e. once a decision to transmit is made the transmission lasts $T$ seconds), we could assume
that $\gamma_{\text{max}} = \gamma_{\text{min}} = 0$. A fixed time for both transmission and probing, by assuming $\gamma_{\text{min}} > 0$, is relevant for delay sensitive applications (e.g. transmission must end before $T$ seconds). We can also limit the total amount of time spent in probing (i.e. the overhead) by fixing $N$ to be a function of $\gamma_n$, e.g. for $T = 1$ and $\gamma_n = 0.01 \forall n$, fixing $N = 5$ results in a 5% maximum overhead.

Similar to [78], we focus on the total transmitted useful bits minus the overhead bits. Mathematically, define the achievable transmitted information after probing $n$ APs as

$$\Theta_n(\rho_n) = \rho_n t_n - \sum_{j=1}^{n} \delta_j. \quad (3.2)$$

where $t_n = (T - \sum_{i=1}^{n} \gamma_i)$ since we assume infinite backlogged users and not finite packet sizes as in the service model. Ideally, a user seeks to maximize their throughput. Realistically, a user only knows the transmission rates of $n$ APs, but does not know the future or what rate AP $n + 1$ may offer. Therefore, a user aims to maximize the average effective throughput by deciding when to stop probing APs, i.e.

$$\max_{n \in \mathbb{N}} E \left[ \frac{\Theta_n(\rho_n)}{T} \right] = \frac{1}{T} \max_{n \in \mathbb{N}} E \left[ \rho_n t_n - \sum_{j=1}^{n} \delta_j \right]. \quad (3.3)$$

An opportunistic strategy transmits after $n$ probes if the “right opportunity” is present, e.g. $r_n$ is sufficiently large. The “right opportunity” is a threshold which depends on what is available now and what is available in the future. A transmission occurs when an observation meets or exceeds the threshold. An optimal opportunistic strategy solves (3.3).

If all $N$ APs are always available (i.e. $B = 0$), then a user could exhaustively
probe all APs before selecting the AP with the highest transmission rate at a cost of \( \sum_{i=1}^{N} \delta_i \) and \( \sum_{i=1}^{N} \gamma_i \) in overhead. An exhaustive strategy gains full knowledge of the network (i.e. learns \( r_n \forall n \)) at a potentially prohibitively high overhead. If \( 0 < B \leq 1 \), an exhaustive probing strategy cannot guarantee the highest transmission rate since a previously probed AP may be unavailable.

As an example, consider a cautious user probing one AP and finding \( t_1 r_1 < t_2 E[r_2] - \delta_2 \). Meaning, the user expects that transmitting to the second AP is better even at the cost of \( \delta_2 \). Assume the user probes AP 2 and finds \( t_2 r_1 > t_2 r_2 \). Should the user have based the decision to transmit only on the expected rate to the next AP? No, the user should consider the expected rate of the next AP and that the following APs may offer a sufficiently high rate to compensate the cost and risk of probing one more AP.

In the following section, we show that an optimal solution exists for (3.3) in the form of a threshold strategy. The thresholds depend on fixed network parameters and can be computed a priori, removing the computational burden from users. Thus, mmWave users need not be burdened by computations and can leverage the necessary AP density in future mmWave networks via a simple threshold based transmission strategy.
3.1.1 Throughput Maximization in mmWave Networks

We begin by describing stopping problems and how they relate to opportunistic transmissions. Then, we show how conditions necessary for the existence of optimal solutions in (3.3), are present in our model. Finally, we present our optimal strategy as a set of optimal thresholds which can be computed a priori if network statistics are available.

For a simple understanding of a stopping problem, consider a hypothetical game where a six-sided die can be rolled at most a finite number of times. In this game you receive a payoff equivalent to the face value of the die whenever you decide to stop rolling or you run out of rolls. If many rolls are available, a good strategy would be to stop only if you see the largest face value. As the number of available rolls decrease, a better strategy might be to stop if you see any of the largest two face values. If very few rolls are available, then a good strategy might be to stop only if you see a face value equal or higher to the expected face value. A stopping strategy mathematically defines the minimum value (i.e. threshold) that should be observed such that you decide to stop the game.

A stopping problem considers a sequential observation of random variable (RV) realizations where a decision to stop the sequence is based on a function of the realizations that have been observed. In our model, the realizations of RVs (i.e. the die rolls) are the transmission rates and occurrence of blockage, the observation (i.e. learning the face value) translates to probing, the cost of a roll is the overhead, and
the decision to stop is the decision to transmit.

The stopping problem references cited in [78] use i.i.d. assumptions. An approach to stopping problems based on martingales does not require the RVs to be identically distributed [83]. We opt for the martingale approach to allow our model to be applicable even when the underlying distributions of \( r_n \) are not identically distributed, thus applicable to a broader set of use cases for mmWave networks. We now establish the necessary mathematical formality to define a specific kind of martingales, called sub-martingales.

Define a sequence of RVs \( X_n \forall n \in \mathbb{N} \). We distinguish RVs by labeling with a super-index, i.e. \( X_n^{(i)} \forall i \in \{1, ..., I\} I \in \mathbb{Z}^+ \), but for defining martingales the super-index labeling is unnecessary and is thus omitted. Each RV \( X_n \) has an associated \( \sigma \)-algebra \( \mathcal{F}_n \). The \( \sigma \)-algebra \( \mathcal{F}_n \forall n \) is a subset of the universal set \( \Omega \), i.e. \( \cup_{n \in \mathbb{N}} \mathcal{F}_n \in \Omega \). With a probability measure \( P_n \) define the probability space \( \{ \Omega, \mathcal{F}_n, P_n \} \) which determines the distribution \( F_n \).

Ultimately, our interest is in the sequence of pairs \( \{X_n, \mathcal{F}_n\} n \in \mathbb{N} \) to represent the stochastic nature of our problem. By defining \( X_n \forall n \) as a sequence of RVs we mean that \( \mathcal{F}_k \subset \mathcal{F}_n \forall k < n \in \mathbb{N} \). Since our construction has the sequence of RVs follow the ordered set \( \mathbb{N} \) and \( \mathcal{F}_n \) as a subset of an algebraic structure, particularly the set \( \Omega \), the \( \sigma \)-algebra \( \mathcal{F}_n \) is by definition a filtration. A filtration is defined as an indexed set \( \mathcal{S}_i \), with \( i \in I \), of objects from an algebraic structure \( \mathcal{S} \) where the index the set \( I \) is a totally ordered set, e.g. the positive integers \( \mathbb{Z}^+ \), subject to the
condition that \( S_i \in S_j \forall i \leq j \). Finally, a RV \( X_n \) is said to be adapted to the filtration \( \mathcal{F}_n \) if \( X_n \) is \( \mathcal{F}_n \)-measurable for each \( n \in \mathbb{N} \). In other words, at \( n \) we cannot see into the future yet we perfectly know the present and the past (i.e. \( E[X_n|\mathcal{F}_n] = X_n \) and \( E[X_k|\mathcal{F}_n] = X_k \forall k < n \)).

At sequence step \( n \) an instance of the RV \( X_n \) is observed, i.e. \( E[X_n|\mathcal{F}_n] = X_n \). As the sequence progresses (i.e. probing continues) additional instances of the sequence pairs are obtained, e.g. \( X_{n+1} \) is observed with filtration \( \mathcal{F}_{n+1} \) for which \( \mathcal{F}_n \subset \mathcal{F}_{n+1} \). Essentially the filtration \( \mathcal{F}_n \) corresponds to the knowledge obtained from \( n \) probings.

One with a creative flair may think of \( \mathcal{F}_n \) as filtering out all randomness from \( X_n \).

We intend to use \( X_n \) to model the combined stochastic nature underlying the achievable rate and the occurrence of blockage. Similarly, \( \mathcal{F}_n \) is intended to represent the knowledge gained after \( n \) probings. We should allow probing, and our sequence, to continue if we expect to gain from continuing. A sequence where we expect to gain from continuing said sequence is a sub-martingale. Formally, define a sub-martingale as a sequence of pairs \( \{X_n, \mathcal{F}_n\}_n \in \mathbb{N} \) such that

- \( \mathcal{F}_k \subset \mathcal{F}_n \subset \mathcal{F}_N \forall k < n \in \{1, ..., N\} \),
- \( X_n \) is an \( \mathcal{F}_n \)-measurable RV with finite first moment \( \forall n \),
- \( \text{and } X_k \leq E[X_n|\mathcal{F}_k] \forall k < n \).

Thus, probing should only continue if our probing sequence is a sub-martingale. Returning to our die example, if rolling the dice is a sub-martingale then you should
continue rolling since the next roll is expected to increase your gain, e.g. if your die lands on the smallest valued face. For completeness, a martingale represents a dice game where, on average, after any roll you do not expect to gain or lose from additional rolls, i.e. $X_k = E[X_n | \mathcal{F}_k] \forall k < n$. A super-martingale is a game where you are more likely to lose than to win by continuing to roll, i.e. $X_k \geq E[X_n | \mathcal{F}_k] \forall k < n$.

Our optimal strategy defines the thresholds at which to stop probing, i.e. when a realization turns the sequence into a super-martingale. An optimal strategy exists if our objective function is a sub-martingale [83]. Thus, we move to show under which conditions (3.2) is a sub-martingale.

### 3.1.2 mmWave Effective Throughput as a Sub-Martingale

We first decompose (3.2) to an affine combination of sub-martingales. Then, we show that a weighted combination of sub-martingales is also a sub-martingale. Finally, conditions under which (3.2) is a sub-martingale are defined.

Throughput has a probability $1 - B$ of depending on previously probed AP $k$ and a probability $B$ of only depending on AP $n$. Each of these two scenarios can be considered as a RV. Thus, we expand $\Theta_n(\rho_n)$ into the weighted sum of two RVs $(1 - B)X_n^{(1)} + BX_n^{(2)}$, where $X_n^{(1)} = \max(\rho_{n-1}, r_n)t_n$ and $X_n^{(2)} = r_nt_n - \delta'_n$ with $\delta'_n = \sum_{i=1}^{n} \delta_i$. Note that we aim for the sum of RVs to represent throughput, hence the overhead cost is only counted once, i.e. in $X_n^{(2)}$ and not $X_n^{(1)}$. If (3.2) is a sub-martingale, then as shown in [83] an optimal stopping strategy exists for (3.3). The
following theorem states that if \( X_n^{(1)} \) and \( X_n^{(2)} \) are sub-martingales, then (3.2) is a sub-martingale.

**Theorem 1.** Let the non-negative RVs \( X_n^{(i)} \) \( \forall i \in \{1, ..., I\} \) be adapted to the filtration \( \mathcal{F}_n \) with \( \alpha_i \geq 0 \forall i \). If \( X_n^{(i)} \) are sub-martingales, then \( X_n = \sum_{i=1}^I \alpha_i X_n^{(i)} \) is a sub-martingale.

**Proof.**
\[
E[X_{n+1}|\mathcal{F}_n] = E[\sum_{i=1}^I \alpha_i X_{n+1}^{(i)}|\mathcal{F}_n] = \sum_{i=1}^I \alpha_i E[X_{n+1}^{(i)}|\mathcal{F}_n] \\
\geq \sum_{i=1}^I \alpha_i X_n^{(i)} = X_n.
\] (3.4)

Where the inequality follows from the assumption that \( X_n^{(i)} \) are sub-martingales, i.e. \( E[X_{n+1}^{(i)}|\mathcal{F}_n] \geq X_n^{(i)} \).

Now, we show under which conditions \( X_n^{(1)} \) and \( X_n^{(2)} \) are sub-martingales so that, by Theorem 1, the throughput is also a sub-martingale. Focusing on the third condition of sub-martingales we consider

\[
E[X_{n+1}^{(1)} - X_n^{(1)}|\mathcal{F}_n] = E[\rho_{n+1} t_{n+1}|\mathcal{F}_n] - \rho_n t_n,
\] (3.5)

where \( E[X_n^{(1)}|\mathcal{F}_n] = X_n^{(1)} \) since we assume the process is adaptive. Similarly

\[
E[X_{n+1}^{(2)} - X_n^{(2)}|\mathcal{F}_n] = E[r_{n+1} t_{n+1} - \frac{\delta_{n+1}}{B}|\mathcal{F}_n] - r_n t_n,
\] (3.6)

where the adaptive assumption allows \( E[X_n^{(2)}|\mathcal{F}_n] = X_n^{(2)} \). For the third condition of the definition of sub-martingales to hold, we require that \( E[\rho_{n+1} t_{n+1}] \geq \rho_n t_n \) for \( X_n^{(1)} \),
and $E[r_n t_n + 1 - \frac{\delta_{n+1}}{B}] \geq r_n t_n$ for $X^{(2)}_n$ to be sub-martingales. For any $\delta_n > 0$ and $B < 1$, if $X^{(2)}_n$ is a sub-martingale then so is $X^{(1)}_n$. Note that at $B = 1$ the sequence depends entirely on $X^{(1)}_n$. Considering $\gamma_n$ and $\delta_n$ are bounded, a condition which is independent of statistics on antenna availability and probing time lengths can be obtained as

$$E[r_{n+1}] \geq \frac{r_n t_n + \delta_{\min}}{t_n - \gamma_{\min}}. \quad (3.7)$$

If we want to model a fixed transmission time independent of time spent probing then $\gamma_n = 0 \ \forall \ n$ and (3.7) reduces to $E[r_{n+1}] \geq r_n + \frac{\delta_{\min}}{B}$. Essentially, we require that the overhead cost is outweighed by the additional throughput of the next AP.

What if a sequence is no longer a sub-martingale? When a sequence is no longer a sub-martingale, then we expect no gain in throughput from probing the remaining APs. While the mathematics suggest that we only probe if $\rho_n$ is a sub-martingale, the engineering perspective dictates that the user must probe at least once. Therefore, after the first probe, the user can follow the mathematical suggestion to know when to continue probing and when to transmit. In the following section we show how to obtain optimal thresholds such that the user knows when to optimally stop the sequence.

### 3.1.3 mmWave Optimal Transmission Strategy

We now identify the smallest observed throughput values at which the sequence is still a sub-martingale. To solve (3.3) we leverage $N$ being finite and find the optimal
strategy via backwards induction. Essentially, we recursively solve sub-problems of (3.3) where the optimization variable is constrained to a smaller subset of \( N \), and the solution of one sub-problem depends on the optimal solution of the sub-problems with an element-wise smaller subset of \( N \).

After probing \( N \) APs the user has one option: transmit. Thus transmitting is optimal after \( N \) probings. Next, consider a user that has probed \( N - 1 \) APs for which the expected benefit if the user probes AP \( N \) and continues optimally is

\[
W_{N-1} = (1 - B) \left[ \int_{\rho_{N-1}}^{\infty} t_N \rho dF_N + B t_N \int_{0}^{\infty} \rho dF_N - \delta_N \right], \tag{3.8}
\]

where \( \rho \) is the highest available rate among unprobed APs, and the derivative of the cumulative distribution function \( dF_N \) is with respect to \( \rho \). Note the overhead of probing up to stage \( n \) is canceled out by the overhead of probing up to stage \( n + 1 \). A user having probed \( N - 1 \) APs must decide to transmit and obtain throughput \( \frac{t_{N-1} \rho_{N-1}}{F_{N-1}} \), or probe AP \( N \) and expect at most a throughput of \( \frac{W_{N-1}}{F} \).

Mathematically, define the decision of a user following our proposed optimal strategy as

\[
R_n(\rho_n) = \max \left( t_n \rho_n, W_n \right) \tag{3.9}
\]

where if \( R_n(\rho_n) = t_n \rho_n \) the user transmits, and if \( R_n(\rho_n) = W_n > t_n \rho_n \) the user探. For completeness define \( W_N = 0 \). Note \( W_n \) is the maximum expected benefit of probing AP \( n + 1 \) and continuing optimally, i.e. \( W_k = \max_{n \in \{k < n \leq N\}} E[\Theta_n(\rho_n)] \).

Essentially \( W_n \) is a constrained sub-problem of (3.3).
For $n < N$ define the maximum expected benefit of probing AP $n + 1$ and continuing optimally as

$$W_n = (1 - B) \left[ R_{n+1}(\rho_n) F_{n+1}(\rho_n) + \int_{\rho_n}^{\infty} R_{n+1}(\rho) dF_{n+1} \right] + B \int_0^{\infty} R_{n+1}(\rho) \, dF_{n+1} - \delta_{n+1}.$$  

(3.10)

Although $W_0 = \max_{n \in \mathbb{N}} E \left[ \frac{\Theta_n(\rho_n)}{T} \right]$, i.e. (3.3), the term $W_0$ is an abuse of notation since $\rho_0$ is undefined, i.e. without a single probing there is no available transmission rate. A user following our proposed optimal strategy decides to transmit after probing $n$ APs if $R_n(\rho_n) = t_n \rho_n \geq W_n$.

The probing sequence is no longer a sub-martingale whenever $t_n \rho_n \geq W_n$ is observed. When a sequence is no longer a sub-martingale, the value of interest (i.e. throughput) is expected to decrease if the sequence continues. Thus, a user seeking to solve (3.3) should base their decision to transmit via comparing the currently available throughput against the maximum expected benefit of continuing the sequence.

A transmission strategy is a set of thresholds $\epsilon_n$ such that a transmission occurs when $\epsilon_n \leq t_n \rho_n$. Consider risk-averse and risk-seeking users with thresholds $\epsilon_n^a < W_n$ and $\epsilon_n^s > W_n$ respectively. These users coincide with the optimal strategy when $t_n \rho_n \geq W_n > \epsilon_n^a$ or $t_n \rho_n \geq \epsilon_n^s > W_n$, and when either condition is met the performance between those users and the optimal strategy is identical. Yet, when $W_n > t_n \rho_n \geq \epsilon_n^a$ or $\epsilon_n^s > t_n \rho_n \geq W_n$ the risk-averse and risk-seeking users differ from the optimal strategy, which from the definition of $W_n$ implies that they will obtain on average less throughput.
Define the smallest value of $\rho_n$ that satisfies the condition to transmit as $\epsilon_n^*$, or mathematically

$$
\epsilon_n^* = \min\{\rho : t_n\rho = W_n\}.
$$

(3.11)

Our proposed optimal strategy uses thresholds $\epsilon_n^* \forall n$.

For fixed $\delta_n$ and $\gamma_n$, e.g. $\delta_n = \delta \forall n$, thresholds $\epsilon_n^*$ can be computed a priori by solving (3.11). When probing overhead is not known a priori the values of $\delta_n$ and $t_n$ in (3.10) should be replaced with $E[\delta_n]$ and $E[t_n]$. Calculations of $\epsilon_n^*$ may be done by the user or offloaded to a remote processor.

3.1.4 Properties of Proposed Strategy

We show that thresholds are non-increasing with $n$, which in turn allows us to express (3.11) in a simpler form. Our proposed optimal strategy depends on computing $\epsilon_n^*$ which depends on how accurately $B$ is measured. Analyzing two corner cases for $B$ leads to a lower bound on performance for our strategy when one inaccurately captures how intrinsic properties of mmWave networks impact recall capabilities.

Via induction we show that throughput threshold is non-increasing with $n$. As a starting point we first show $t_{N-2}\epsilon_{N-2}^* \geq t_{N-1}\epsilon_{N-1}^*$. To this end, first consider $W_{N-2} - W_{N-1}$ and the fact $R_{N-1}(\rho) \geq t_{N-1}\rho \forall \rho$ with the assumption of the sequence is a sub-martingale it follows that $W_{N-2} \geq W_{N-1} \forall \rho$. From (3.11) recall that $t_n\epsilon_n^* = W_n$, thus $t_{N-2}\epsilon_{N-2}^* \geq t_{N-1}\epsilon_{N-1}^*$ with equality only when $\gamma_n = 0 \forall n$ and $B = 0$. 
Assume relations hold for $k > n$ with $k \leq N$, and consider

\[
W_n - W_{n+1} = \]

\[
(1 - B) \left( R_n(\rho) F_n(\rho) - R_{n+1}(\rho) F_{n+1}(\rho) + \int_{\rho}^{\infty} R_n(\varrho) dF_n - \int_{\rho}^{\infty} R_{n+1}(\varrho) dF_{n+1} \right) \]

\[
+ B \left( \int_{0}^{\infty} R_n(\varrho) dF_n - \int_{0}^{\infty} R_{n+1}(\varrho) dF_n + 1 \right) - \delta_n + \delta_{n+1}, \tag{3.12}
\]

Thus from induction assumptions $R_n(\rho) \geq R_{n+1}(\rho)$ $\forall \rho$ combined with assumption that the process is a sub-martingale leads to $W_n \geq W_{n+1}$ $\forall \rho$. As before, from (3.11), we obtain $t_n \epsilon_n^* \geq t_{n+1} \epsilon_{n+1}^*$ which implies that the minimum amount of information that the user must be able to transmit so that a decision to stop probing occurs is non-increasing with $n$. Thus confirming that as more options are available in the future, we should be more demanding in the present.

Knowing $t_n \epsilon_n^* \geq t_{n+1} \epsilon_{n+1}^*$ allows us to simplify (3.11) by constraining the feasible values of $\rho$ such that $\rho \geq \epsilon_{n+1}^*$ without losing optimality. Constraining $\rho \geq \epsilon_{n+1}^*$ simplifies calculating $W_n$ in (3.11) by reducing the decision rule for the next stage as

\[
R_{n+1}(\rho) = t_{n+1} \rho \quad \forall \rho > \epsilon_{n+1}^*.
\]

We now explore two corner cases in our model, Case $B$ where recall is impossible and Case $NB$ where recall is perfect (e.g. no blockage). Case $B$ assumes $B = 1$ and models a network where previously probed APs are ignored or where blockage is highly likely to occur. Case $NB$ assumes $B = 0$ and that blockage is non-existent in a network, i.e. ideal yet unrealistic conditions for a mmWave network. Denote the optimal thresholds for these cases by $\epsilon_n^B$ and $\epsilon_n^{NB}$ respectively.
Our interest in these two cases is motivated by the importance of \( B \) in our model. The performance of a strategy using \( \epsilon^*_n \) depends on how accurately \( B \) represents probability of previous APs being available. Assume that due to measurement errors \( \epsilon'_n \) is computed with \( B' \neq B \). The performance of a strategy using \( \epsilon'_n \) relative to a strategy using \( \epsilon^*_n \) is lower bounded by the minimum between the performance of the strategies using \( \epsilon^B_n \) and \( \epsilon^{NB}_n \). For example, if \( B < 0.5 \) then at worst \( B' = 1 \) which is equivalent to Case \( B \) and if \( B > 0.5 \) then at worst \( B' = 0 \) which is equivalent to Case \( NB \).

For Case \( B \), i.e. \( B = 1 \), the optimal thresholds are

\[
\epsilon^B_n = \min \left\{ \rho : \rho = \frac{t_{n+1}}{t_n} \int_{0}^{\infty} R_{n+1}(\rho) dF_{n+1} - \frac{\delta_{n+1}}{t_n} \right\}. \tag{3.13}
\]

Since the right-hand side inside the minimization in (3.13) is independent of \( \rho \), we obtain the recursive function

\[
\epsilon^B_n = \frac{t_{n+1}}{t_n} \left( \epsilon^B_{n+1} F_{n+1}(\epsilon^B_{n+1}) + \overline{p_n} - \int_{0}^{\epsilon^B_{n+1}} \rho dF \right) - \frac{\delta_{n+1}}{t_n}, \tag{3.14}
\]

with \( \overline{p_n} = \int_{0}^{\infty} \rho dF_n \); the sub-index on \( \rho \) is suppressed since \( B = 1 \), and \( \epsilon^{B}_n = 0 \) for completeness. Note that for \( \gamma_n > 0 \) the ratio \( \frac{t_{n+1}}{t_n} < 1 \), thus as \( n \) grows the threshold decreases, i.e. \( \epsilon^B_n > \epsilon^B_{n+1} \).

For Case \( NB \), i.e. \( B = 0 \), the optimal thresholds are

\[
\epsilon^{NB}_n = \min \left\{ \rho : \rho = \frac{t_{n+1}}{t_n} \left( \rho F_{n+1}(\rho) + \int_{0}^{\infty} \rho dF_{n+1} \right) - \frac{\delta_{n+1}}{t_n} \right\}, \tag{3.15}
\]

and \( \epsilon^{NB}_n = 0 \). Obtaining the form in (3.15) requires \( t_n \epsilon^*_n \geq t_{n+1} \epsilon^*_n \) which we have already shown. When \( \gamma_n > 0 \), the thresholds decrease, but so does the achievable
throughput obtained via recall since the remaining time to transmit is reduced. For \( \gamma_n = 0 \), optimal thresholds are single valued, i.e. \( \epsilon_n^{NB} = \epsilon_{n+1}^{NB} \forall n < N \), which implies that recall only occurs after probing all \( N \) APs. Furthermore, for \( \gamma_n = 0 \) clearly \( \epsilon_n^{NB} \geq \epsilon_n^B \) which implies that when recall is perfectly available we should be more demanding of what the network should offer.

A corner case of our model with \( B = 1 \) (i.e. Case B), \( N = \infty \) and \( \gamma_n = 0 \) was considered in [78] which resulted in single valued optimal threshold, i.e. \( \epsilon_n^* = \epsilon_{n+1}^* \forall n \).

Note that for \( \gamma_n = 0 \forall n \), Case B is not solved by a single optimal threshold value due to \( N < \infty \). Intuitively, having an infinite number of opportunities can lead, rightfully so, to an optimally stubborn attitude. Even if we willingly decide to forgo the ability to recall, committing ourselves the possibility of probing ad infinitum until the ideal situation arises out of an intrinsically stochastic mmWave network may very well lead to large delays. Thus, we next analyze the expected delay of using opportunistic access strategies.

### 3.2 Performance Bounds

Probing more APs increases delay but may lead to higher throughput. Delay is paramount for many applications in mmWave networks, thus we characterize the delay incurred by the use of our optimal opportunistic strategy.
3.2.1 Expected Delay and Overhead

The probability that a transmission occurs after exactly one probing is equivalent to the probability of rate $r_1$ being above the threshold, i.e. $P(r_1 \geq \epsilon_1^*)$. To calculate the probability of a transmission occurring exactly after two probings we combine the probability of the first AP being selected after two probings, i.e. $P(\epsilon_1^* > r_1 \geq \epsilon_2^*)$, and the probability of transmitting to the second AP if the first AP was unacceptable, i.e. $P(r_1 < \epsilon_1^*)P(r_2 \geq \epsilon_2^*)$. Recall that $\epsilon_n^* \geq \epsilon_{n+1}^*$, hence we require only to consider that rates fall between consecutive thresholds such that transmission occurs.

Define $n^*$ as the value of $n$ at which a transmission occurs, i.e. $n^* = \min_{n \in \mathbb{N}} \{n : t_n \rho_n = \epsilon_n^*\}$. Define the probability that a transmission occurs after exactly $n$ probings as

$$P(n = n^*) = P(r_n \geq \epsilon_n^*) \prod_{i=1}^{n-1} P(r_i < \epsilon_n^*) + \sum_{j=1}^{n-1} P(\epsilon_{n-1}^* > r_j \geq \epsilon_n^*) \prod_{k=1}^{j-1} P(r_k < \epsilon_n^*).$$

(3.16)

Recall that $\epsilon_N^* = 0$ and $r_n \geq 0$ thus $P(r_n < \epsilon_N^*) = 0$. For strategies using single valued thresholds, e.g. $\epsilon_n = \epsilon_{n+1}$, the equation above reduces to a simple binomial probability with $n-1$ failures and one success.

Define the expected average delay as

$$\overline{D} = \sum_{n=1}^{N} P(n = n^*) \sum_{j=1}^{n} \gamma_j,$$

(3.17)

and similarly the expected average overhead cost as

$$\overline{C} = \sum_{n=1}^{N} P(n = n^*) \sum_{j=1}^{n} \delta_j.$$

(3.18)
Note that both $D$ and $C$ implicitly depend on the selected thresholds. With knowledge of the associated statistical distributions, one can find thresholds such that a fixed average delay $\bar{D}$ is met by an opportunistic strategy.

Opportunistic strategies are not exempt from the known throughput-delay trade-off present in wireless networks. To glean insights from how the throughput-delay tradeoff is present we present an average $c$-bound for our optimal opportunistic strategy.

3.2.2 Throughput Performance Bound

We bound the performance of our proposed stochastic strategy relative to an optimal deterministic strategy, i.e. a strategy achieving optimality at each network instance. The presented bound is a value $0 \leq c \leq 1$ such that the stochastic strategy is on average $c$ times the performance of the optimal strategy. For example, guessing the outcome of a fair coin flip results in $c = \frac{1}{2}$, since random guessing is right half the times compared to a deterministic optimal solution.

To obtain a deterministically optimal solution of (3.3), consider a genie-aided user with a priori knowledge of $r_n \forall n$. Of course, the genie-aided user could simply probe a single AP to obtain the highest throughput, i.e. $\text{argmax}_{i \in N}(r_it_1 - \delta_i)$. For the sake of obtaining a tighter upper bound, let us assume that the genie-aided user probes in the same order as a non genie-aided user. The genie-aided user solves (3.3) by transmitting to AP $m^* = \text{argmax}_{i \in N}(r_it_i - \sum_{j=1}^{i} \delta_j)$ and probing exactly $m^*$ APs.
On average the genie-aided user obtains an expected throughput of

\[ E\left[ \frac{\Theta_m^*(r_m^*)}{T} \right] = \frac{1}{T} \int_0^\infty \left( 1 - \prod_{i=1}^N P(r_i t_i - \sum_{j=1}^i \delta_j \leq y) \right) dy, \]

where we assume independence among achievable rates of distinct APs. Note that (3.19) upper bounds (3.3) due to the convexity of the max function and Jensen’s inequality.

When overhead is non-existent (or ignored in throughput calculations), i.e. \( \delta_n = 0 \) and \( \gamma_n = 0 \), then (3.19) grows with \( N \), essentially tending towards the maximum rate possible as \( N \) goes to infinity. Recall that \( t_n = T - \sum_{i=1}^N \gamma_n \), and thus a fixed \( T \) results in \( t_n < 0 \) as \( n \) grows. For non-negligible overhead, i.e. \( \delta_n > 0 \) or \( \gamma_n > 0 \), the \( \frac{y + \sum_{i=1}^i \delta_j}{t_i} \) tends to a negative value as \( N \) grows and (3.19) remains finite. Essentially, the overhead makes even a genie-aided user weary of too much probing.

Computing the expected throughput of a given opportunistic strategy requires accounting of all the combinations in which a transmission may occur after \( n \) probings. Note that the probability of a transmission occurring exactly after \( n \) probes, i.e. \( P(n = n^*) \), is simplified by only requiring to count the conditions of when a transmission occurs and not keep track of which AP is selected. Accounting for both “when” and “which” results in a lengthy combinatorial exercise.

For exposition and tractability, we account only for when a transmission occurs and generalize the accountability of which AP is selected. Thus, we present an upper
bound to $c$ as

$$c \leq \frac{\sum_{n=1}^{N} \left( (1 - B)E[\max_{i \leq n}(\Theta_n(r_i))] + BE[\Theta_n(r_n)] \right) P(n = n^*)}{E[\Theta_{m^*}(r_{m^*})]}$$

(3.20)

with equality at $N = 1$.

As before, consider Case $B$ with $B = 1$ and Case $NB$ with $B = 0$. For Case $B$ the average $c$-bound is

$$c^B \leq \frac{\sum_{n=1}^{N} E[\Theta_i(r_i)] P(n = n^*)}{E[\Theta_{m^*}(r_{m^*})]} \leq \frac{\max_{i \in \mathbb{N}} E[\Theta_i(r_i)]}{E[\max_{i \in \mathbb{N}} E[\Theta_i(r_i)]]}$$

(3.21)

where the second inequality follows from the law of total probability and Jensen’s inequality. Note $c^B \leq 1$ as expected.

For Case $NB$ the we obtain the following

$$c^{NB} \leq \frac{\sum_{n=1}^{N} E[\max_{i \leq n}(\Theta_n(r_i))] P(n = n^*)}{E[\Theta_{m^*}(r_{m^*})]}$$

$$\leq \frac{\max_{n \leq N, i \leq n} E[\Theta_n(r_i)]}{E[\max_{i \in \mathbb{N}} E[\Theta_i(r_i)]]} = \frac{\max_{i \in \mathbb{N}} E[\Theta_i(r_i)]}{E[\max_{i \in \mathbb{N}} E[\Theta_i(r_i)]]}$$

(3.22)

where the second inequality follows from Jensen’s inequality and the law of total probability and the equality from the fact that $\delta_n \geq 0$ and $\gamma_n \geq 0$. The bound on $c^{NB} \leq 1$ as expected.

As noted before, Cases $B$ and $NB$ provide a lower bound to an implementation of our algorithm with inaccurate measurement of $B$. Therefore, from (3.21) and (3.22) we obtain the best case scenario for throughput for the largest possible error in measuring $B$. The average delay, i.e. (3.17), and the throughput bound, i.e. (3.20), are both tied to threshold selection via $P(n = n^*)$. Intuitively, smaller thresholds
to start a transmission faster than larger thresholds, while larger thresholds have the possibility of reaping benefits from exploring more APs. In the following section we explore such tradeoffs via simulation.

3.3 Numerical Results

To benchmark our strategy we consider a genie-aided user, a user performing exhaustive search, a user transmitting after only one probe, a user assuming $B = 1$, i.e. Case $B$, and a user assuming $B = 0$, i.e. Case $NB$. A genie-aided user is infeasible in practice, but serves as a meaningful upper bound. For our numerical analysis, channels $h_n$ are drawn from a known Ricean distribution. Throughput is presented as bits per slot per Hz by fixing $T = 1$ and a unit bandwidth.

Unless otherwise stated numerical performance analysis parameters are $N = 10$, $P = 20$ dB, $N = 1$ dB, $\delta_n = \delta = 0.02 T E[r_n] \forall n$, $\gamma_n = \gamma = \frac{0.05 T}{N}$ and $B = 0.3$. Parameter selection for $\delta$ and $\gamma$ is in accordance with published works, e.g. [29] considers the time to probe ten beams $0.02 T$, [78] considers overhead time of 5%$T$ with base rate of $\frac{2 \text{Mb}}{\text{sec}}$ and median rate of $\frac{5.5 \text{Mb}}{\text{sec}}$ leads to a proportion of overhead to information bits $\frac{2 \times 0.05}{5.5 \times 0.95} \approx 0.02$, and [35] considers a 5% maximum temporal overhead., and $N = 10$ coincides with the need for multiple APs to provide mmWave network coverage [8, 7]

Fig. 3.1 shows the throughput as a function of $B$. Note that throughput scales with the bandwidth size, which is in the order of GHz for mmWave. When $B = 0$ and
Figure 3.1: Avg. throughput as a function of $B$ for $N = 10$. The proposed strategy obtains $\approx 20\%$ increase compared to a single probing strategy.
Figure 3.2: Avg. throughput as a function of available APs $N$ with $B = 0.3$. For larger $N$ a higher throughput may be achieved at the cost more probed APs.
$B = 1$ the optimal strategy coincides with the strategies making such assumptions. The assumption that APs are always unavailable offers a better degradation over all scenarios of $B$, relative to the opposite assumption. When $B = 0$, no practical strategy achieves the genie aided strategy due to probing more APs than the genie aided strategy.

Fig. 3.2 shows the impact of increasing the available APs $N$ on throughput with $B = 0.3$. Since our strategy, assume $B = 1$, and assume $B = 0$ tend toward the same thresholds as $N$ grows, it follows that all three tend toward the same throughput performance. Note that the average throughput gain from increasing $N$ diminishes as $N$ grows since a user is likely to only probe so many, e.g. $n < N$, APs before finding a suitable AP. Fig. 3.4 shows the CDF of the probability of a transmission occurring at $n$ for various values of $N$. A higher CDF implies less delay since a transmission is more likely to have occurred (i.e. the closer to the top left of the plot equals less delay). Note that as $N$ grows, the CDF of all three strategies tend towards a very similar performance.

The ratio of average delay to maximum delay against the number of available APs, i.e. $\frac{D}{\sum_{i=1}^{N} \gamma_i}$, is shown in Fig. 3.3. Note that the average overhead follows the same behavior as the average delay, cf. (3.17) and (3.18). In terms of average delay Case $B$ outperforms Case $NB$, which contrasts with the throughput performance, cf. Fig. 3.2. Note the decline in average delay of our proposed strategy is sharp for $N < 5$ which implies the motivating idea that mmWave networks can greatly benefit from
Figure 3.3: Ratio of the average delay against the maximum delay, i.e. $\frac{\overline{D}}{\sum_{i=1}^{N} \gamma_i}$. 
Figure 3.4: CDF of transmission occurring against total number of probed APs for \( N = \{5, 10\} \). A lower CDF implies a lower average delay and overhead.

Figure 3.5: CDF of transmission occurring against total number of probed APs for \( N = \{15, 25\} \). A lower CDF implies a lower average delay and overhead.
a higher AP density without having to extremely high AP density per user. All but the exhaustive strategy tend to incur a smaller average delay, which suggests a strong tendency to probe only a few APs independent of the total number of APs.

The cumulative distribution function (CDF) of a transmission occurring for Case B, Case NB, and the proposed strategy against the number of APs probed is shown for $N \in \{10, 25\}$ in Fig. 3.4. The CDF of Case B for all values of $N$ is identical over the overlapping range, e.g. the CDF of a transmission occurring for Case B after 9 probes is $0.6 \forall N$. In contrast, the CDF of Case NB and the proposed strategy maintain a similar shape across values of $N$, e.g. the CDF of a transmission occurring for Case NB after $\lceil \frac{N}{2} \rceil$ probes is $\approx 0.8 \forall N$. Such a behavior follows from Case B being independent of $N$, i.e. Case B is also the solution for $N = \infty$, while Case NB and our proposed strategy incorporate the value of $N$ directly into the threshold decisions.

3.4 Summary

We proposed an optimal opportunistic transmission strategy for wireless mmWave networks. Users should leverage AP diversity by probing multiple APs in search of a channel that provides a high transmission rate. Our results suggest that opportunistic strategies can benefit from a denser AP deployment, but the rate at which throughput increases is decreased for high density deployments. Our work can be combined with efficient beam selection and network discovery algorithms to enable a higher
throughput in future mmWave networks. When the number of potential probings is finite and relatively small, then our optimal transmission strategy outperforms existing strategies.
Chapter 4

Service Provisioning
4.1 Service Model

We now discuss aspects in our model that are particular for service provisioning. Development of future networks must consider the diverse requirements imposed by a variety of services. Therefore, we consider services with a strict delay constraint and an average delay constraint. Time sensitive services (e.g. stock options) have a strict delay requirement, which if the network is unable to meet then the service becomes obsolete. Services with a less strict requirement may be satisfied with being provided for on average in a timely manner.

While our model is amenable for both downlink and uplink, we refer to the AP as the transmitting element and the users as the receiving elements. Thus, assume an AP is tasked with the singular goal of providing services to users via wireless transmissions of information packets. Services are said to be “provided” if all transmitted packets meet their temporal constraint. We consider two temporal constraints: a strict delay and an average completion time delay. The network receives no benefit if a subset of transmitted packets does not meet the requirement (i.e. an “all or nothing” approach).

Without loss of generality, assume each node sends a single service request; a single user with multiple service requests can be treated as the case where multiple users have individual requests. At the benevolence of the reader, we use $N$ as both the total number of users and service requests.

For the service model we assume users have packetized requests of finite bit size. A service request $n$ consists of a packet size $b_n$ bits. Define the transmission time for
service $n$ as

$$t_n = \frac{b_n}{r_n}.$$

(4.1)

Note that $t_n$ is known only after probing. We assume that $E[t_n] + \gamma_n \leq \frac{b_n}{r_{base}} \ \forall n \in N$, else probing would not be a desirable action to undertake since user $n$ could just transmit at $r_{base}$ and expect a shorter completion time.

To provide services, the AP seeks uses a temporal partition of resources in the form of a schedule. A schedule defines when and which services are transmitted. A scheduling period is a three phase cycle: polling, ordering, and execution. We now describe each phase alongside the necessary mathematical formulation.

The polling phase is a fixed period of time when a user may transmit a packet of $b_{poll}$ bits with a service request to the AP. Transmissions during the polling phase are uncoordinated, but temporally short enough such that probability of collision is negligible [86] and probing is undesirable (i.e. $\frac{b_{poll}}{r_{base}} \leq \gamma_{min}$).

The ordering phase is a fixed period of time in which the AP selects and broadcast the order in which users will probe and decide to transmit during the execution phase. Define the probing order $\vartheta_i$ as an ordered set of size $N$, with $i \in \{1, ..., N!\}$. Furthermore, define $\vartheta_1$ as the probing order that strictly follows the arbitrary labeling of the $N$ nodes, i.e. $\vartheta_1 = \{\vartheta_1(n) : \vartheta_1(n) = n \ \forall n \in N\}$. Note that for any probing order label $i \in \{1, ..., N!\}$ the probing order $\vartheta_i$ is a permutation of $\vartheta_1$.

Finally, the execution phase occurs when nodes follow the order established in the ordering phase to sequentially probe and decide to transmit. After probing, a
decision to *not* transmit may be taken if, for example, the channel is in outage. The decision to transmit, or not, is announced during the execution phase at the end of the probing period to allow the sequence to continue without interference. Alternatively, if a decision to transmit is made, then the AP announces the transmission time via broadcast at the end of the probing period.

Define the transmission vector $x_j = \{x_j(n) : x_j(n) \in \{0, 1\} \forall n \in \mathbb{N}\}$ where $x_j(n) = 1$ if service $n$ is transmitted and $x_j(n) = 0$ if service $n$ is not transmitted. Note that $j \in \{1, ..., 2^N\}$ since there are $2^N$ possible combinations of transmission vectors. It will be useful to know the position of the last transmitting element, thus define $z_i = \max\{n : x_j(\vartheta_i(n)) = 1 \forall n \in \mathbb{N}\}$; i.e. $z_i$ is the position of the last service in the probing order $\vartheta_i$. Note that the dependence of $z_i$ on an activation vector $x_j$ is notationally suppressed for simplicity. Thus, the polling phase defines $N$, the ordering phase defines $\vartheta_k$, and the execution phase defines $x_j$.

Formally, define a schedule as the pair $S_{j,i} = \{x_j, \vartheta_i\}$. For a given value of $N$ there exist $N!2^N$ distinct schedules each with, potentially, a distinct delay experienced by the users. A complete measure of delay would consider the duration of the polling and ordering phase. The polling and ordering phase have a fixed temporal duration, while the length of the execution phase may vary depending on the temporal constraint being met and network conditions. For a strict delay constraint the duration of the execution phase is upper bounded, while the average delay constraint only has a known average duration. Given that the polling and ordering phases have a fixed
duration, and thus we focus on the delay in the execution phase for a given schedule \( S_{j,i} \).

Individually, each service experiences delay up until the time their packet transmission is completed, i.e. completion time. For a schedule \( S_{j,i} \) the completion time of the \( n \)-th service in order \( \vartheta_i \) (i.e. \( \vartheta_i(n) \)) is mathematically defined as

\[
tc(S_{j,i}, n) = \sum_{l=1}^{n} (t\vartheta_i(l)x_j(\vartheta_i(l)) + \gamma\vartheta_i(l)),
\]

(4.2)

and depends on which nodes probed and what services were transmitted before \( \vartheta_i(n) \).

Note that a user with multiple service requests (which are modeled as users with single requests) may not require equal lengths for each probing duration. For example, if services \( n \) and \( n+1 \) correspond to the same user then \( \gamma_{\vartheta_i(n)} > \gamma_{\vartheta_i(n+1)} \). For services requiring a strict delay, we will consider the completion time as the temporal metric of interest. Define \( T \) as the strict delay constraint. The strict delay constraint can be used to limit the number of service requests received during the polling phase as

\[
N < \left\lceil \frac{T}{\gamma_{\min}} \right\rceil \text{ for } \gamma_{\min} > 0.
\]

A service \( n \) with a strict deadline is defined as successfully transmitted by schedule \( S_{j,i} \) if \( x_j(\vartheta_i(n)) = 1 \) and \( tc(S_{j,i}) \leq T \). Define a feasible schedule \( S_{j,i} \) as a schedule for which all transmitted services are successfully transmitted. A common deadline among pairs is applicable for pairs with equal or similar services (e.g. industrial sensors monitoring the same process). Since there is no penalty for early transmission completion, the strictest deadline among a group of dissimilar services can be considered as the common deadline. The network seeks to obtain the highest sum of service
payoffs possible.

For services interested in the average service experience delay of all transmitted services, we consider the average completion time as the temporal metric of interest. For a given schedule $S_{j,i}$ define the average completion time as

$$
\bar{t}^c (S_{j,i}) = \frac{\sum_{z_i} m=1 \left[ \sum_{l=1}^{m} \left( t_{\vartheta_i(l)} x_j(\vartheta_i(l)) + \gamma_{\vartheta_i(l)} \right) x_j(\vartheta_i(m)) \right]}{\sum_{z_i} m=1 x_j(\vartheta_i(m))},
$$

or in compact form, by substituting (4.2), as

$$
\bar{t}^c (S_{j,i}) = \frac{\sum_{z_i} m=1 t^c(S_{j,i}, m)x_j(\vartheta_i(m))}{\sum_{z_i} m=1 x_j(\vartheta_i(m))},
$$

where the numerator is the summation of all the completion times of transmitted services, and the denominator is the total number of transmitted services.

The average service delay that the network aims to meet is defined as the average completion time constraint $\bar{T}$. A service $n$ is defined as successfully transmitted by schedule $S_{j,i}$ if $x_j(\vartheta_i(n)) = 1$ and $\bar{t}^c (S_{j,i}) \leq \bar{T}$. Define a feasible schedule $S_{j,i}$ as a schedule for which all transmitted services are successfully transmitted. Fig. 4.1 shows a graphical example of the three phase scheduling period and the difference between average completion time and sum transmission time.

The execution phase ends when no more services can be successfully transmitted. The AP broadcasts a predefined message to signal the end of the execution phase and the start of a new polling phase. The polling phase and ordering phase have a fixed duration, but the exact duration of the execution phase for average delay constrained services depends on how many transmissions occur and when they are completed.
Figure 4.1: Scheduling period example with collocated users, each with their service requests, and a single AP. After the polling phase, the AP selects one of the $N!$ possible probing orders during the ordering phase. For exposition, the impact of two schedules are shown during the execution phase on the right. While the schedules $S_{j,1}$ and $S_{j,i}$ use the same transmission vector $x_j = \{1, 0, 1, 1\}$, the distinct probing orders result in distinct average service delay. For sake of the example, assume that $t_{\vartheta_1(2)} \gg 0$ such that the second user does not transmit. Note that $\vartheta_1(2) = \vartheta_i(3)$. While both schedules have the same sum probing and transmission time, the average completion time of $S_{j,i}$ is lower due to the order in which transmissions occur.
Nonetheless, the average delay of the entire schedule for transmitted services is on average no greater than the sum of the duration of the polling phase, the ordering phase, and the average completion time constraint.

If a service is not transmitted in a scheduling period, the user may send a request in the polling phase of the next scheduling period. Alternatively, a user may decide not to make a request in the next polling phase (e.g. a packet is dropped) if the packet has gone unserviced for an unacceptable amount of time. We assume users have a queued number of service requests, else they would not make a service request. Users may drop a request from their local queue if a predetermined number of attempts has been exceeded.

Considering all possible transmission vectors (i.e. $x_j \forall j \in \{1, \ldots, 2^N\}$) and probing orders (i.e. $\theta_i \forall i \in \{1, \ldots, N!\}$), the size of the set of feasible schedules is upper bounded by $N!2^N$. In an ideal world, a feasible schedule $S_{j,i}$ would successfully transmit all $N$ services. A more realistic expectation is to admit that not all services may be transmitted in every schedule. Thus, we require a means of deciding which services are, and which are not, transmitted.

The decision to transmit one service over another, we argue, should be done by considering the preference a user may have of one service over another. A quantification of said preference allows the network to differentiate the service requests. To allow a differentiation of services, define a service value $\nu_n \forall n \in N$ associated to each service. From a network operator’s perspective, the value of $\nu_n$ could also be a
function of service priority, a function of user’s starvation (e.g. inversely proportional to how long since a user has been serviced), related to a higher level service metric (e.g. inversely proportional to remaining buffer of a video), or simply a mathematical abstraction of user preference.

The service provision capabilities of a network are quantified by the sum value of transmitted services. Note that when $v_n = 1 \ \forall \ i \in N$, the service provision capabilities are equivalent to the total number of transmitted packets. A feasible schedule $S_{j,i}$ achieves a sum service value of

$$V(S_{j,i}) = \sum_{i=1}^{N} x_i w_i. \quad (4.5)$$

An optimal schedule $S^*$ achieves a service value no less than the service value of any feasible schedule $S_{j,i}$, i.e. $V(S^*) \geq V(S_{j,i})$. Note that optimality depends on feasibility, which in turn is defined depending on the temporal constraint being considered.

### 4.2 Service Provisioning with Strict Delay

For an AP providing services with a strict deadline, $S^*$ can be obtained as the solution to

$$\max_{S_{j,i}} \ V(S_{j,i})$$

s.t. $t^c(S_{j,i}, n) \leq T. \quad (4.6)$

The problem in (4.6) represents a single channel, and the case for multiple channels can be treated as separate instances of the formulation above. Fig. 4.2 visually presents an example and optimal solution for services provided with a strict delay.
Figure 4.2: Example of services provided with a strict delay for $N = 4$, $t = \{\frac{T}{4}, \frac{T}{3}, \frac{T}{2}, T\}$, $\gamma_n = T/6 \forall n$, $v = \{3, 2, 3, 1\}$, $x^* = \{1, 0, 1, 0\}$, and $V(S^*) = 6$.

Service 1 and 3 are transmitted, while user 2 only probes, and user 4 has no time to probe or transmit, thus $z_i = 3$.

Like many resource allocation problems, (4.6) is similar to the famous knapsack problem. The knapsack problem is described as follows: select the set of items with highest sum value to place in a knapsack without exceeding a weight limit. Unlike the typical knapsack problem we must pay a cost, i.e. $\gamma_n$, to discover the weight, i.e. $t_n$, of each item. Additionally, if an item $n$ is not added to the knapsack (a pair does not transmit) and the weight of another item $n+1$ is discovered then item $n$ cannot be added to the knapsack (since the pair is assumed to move to a different channel or direction). In the trivial case when $\gamma_n = 0 \forall n$ the problem (4.6) reduces to the knapsack problem.

The knapsack problem is known to be an NP-hard problem. Therefore, the intractability of obtaining $S^*$ goes beyond the impractical necessity of full network information. Nonetheless, solving (4.6) serves as an upper bound for practical strate-
4.2.1 Optimal Schedule for Strict Delay

We now present a method of solving (4.6) inspired by the known dynamic programming solution for the knapsack problem [87]. First define the set \( v = \{1, ..., \sum_{i=1}^{N} v_i\} \).

For any \( v \in v \), \( n \in N \), and probing order \( \vartheta_1 \) consider the problem

\[
\begin{align*}
\min_{x} & \quad \sum_{i=1}^{n} x_i t_i \\
\text{s.t.} & \quad \sum_{i=1}^{n} x_i v_i = v,
\end{align*}
\]

(4.7)

where (4.7) is the dual of (4.6) for \( n = N \) if \( \gamma_n = 0 \ \forall n \). The variables \( n \) and \( v \in v \) are constraints in (4.7). Define \( d(v, n) \) as the solution to (4.7) for given values of \( v \) and \( n \). Note that \( d(v, n) \) is the minimum sum transmission time such that a payoff of \( v \) is achieved by the first \( n \) pairs of a probing order \( \vartheta_1 \).

Obviously the schedule achieving \( d(\sum_{i=1}^{N} v_i, N) \) is not the same as the schedule achieving \( S^* \), except for the trivial case when all services can be transmitted. Since \( v \) contains all possible sum service payoff values then obviously \( V(S^*) \in v \). Thus, it follows that iteratively finding \( d(v, n) \) for all \( n \in N \) and \( v \in v \) leads to the optimal solution to (4.6).

The method to find the optimal solution to (4.6) is as follows: Construct a table of \( d(v, n) \) for all values of \( v \in v \) and \( n \in N \). Construction can be done recursively thanks to the optimal substructure property present in the knapsack problem [87]. The solution for \( d(v, n) \) ignores the probing cost and deadline constraint in (4.6).
Table 4.1: Portion of the solution table for \(d(v, n)\) with values \(t = \{T_2, T_3, T_5, T_{10}\}\) and \(v = \{2, 3, 2, 1\}\). Combinations of \((v, n)\) with no solution (e.g. \(d(1, 1)\)) are marked “No Soln.”.

As an example a portion of the table constructed with \(d(v, n)\) is presented in Table 4.1. Therefore, to consider the probing cost update the table of solutions \(d(v, n)\) by discarding all entries in the table, due to being infeasible, for which

\[
d(v, n) + \sum_{i=1}^{n} \gamma_i > T,
\]

then select among the remaining entries the one with highest value for \(v\). The example in Table 4.1 is updated in Table 4.2 which shows the discarded infeasible schedules crossed out. Note discarding an infeasible schedule does not discard a payoff that is achievable by a feasible schedule.

Note that there may exist values of \(v \in \mathbf{v}\) for which no solution of \(d(v, n)\) exists for any \(n\) (e.g. \(v = \{2, 2\}\) and \(v = 3\)). Identifying which values of \(v\) cannot be achieved
Table 4.2: Update of Table 4.1 for $d(v, n) + \sum_{i=1}^{n} \gamma_i$ with $\gamma_n = T^5/6$ for all $n$ and infeasible schedules crossed out. Combinations of $(v, n)$ with no solution (e.g. $d(1, 1)$) are marked “No Soln.”.

by a combination of values becomes a combinatorial problem in itself, thus we allow $v$ to iterate over all values. More than one entry with highest sum payoff may exist (e.g. if $N = 2, t = \{T_2, T_2\}, \gamma_n = T/4$ for all $n$, and $v = \{2, 2\}$), therefore arbitrarily select the one with smallest $d(v, n) + \sum_{i=1}^{n} \gamma_i$ among them which, as we will now prove, is $S^*$.

**Theorem 2.** The optimal solution to (4.6) for given $N, T,$ and $\vartheta_1$ can be obtained by iteratively solving $d(v, n)$ for $v \in \{0, ..., \sum_{i=1}^{N} v_i\}$ and $n \in \{0, ..., N\}$, then selecting the feasible schedule achieving the highest value of $v$ among all the solutions.

**Proof.** Define $v^* \in \mathbf{v}$ as the value selected by the algorithm described in the proof. Consider any $v' > v^*$ obtained by some schedule $S'$. If $\sum_{i=1}^{N} x'_i t_i + \sum_{j=1}^{z_1} \gamma_j > T$ then the schedule is infeasible, and the existence of $v'$ does not contradict the optimality
of \( v^\ast \). Therefore if \( v' \) exists then \( \sum_{i=1}^{N} x'_i t_i + \sum_{j=1}^{z_1} \gamma_j \leq T \), from which \( \sum_{i=1}^{N} x'_i t_i < T \) follows if \( \exists \gamma_j > 0 \). Obviously \( v' \leq \sum_{i=1}^{N} v_i \), therefore the value of \( v' \) was considered when recursively constructing the table for \( d(v, n) \), but discarded so that \( v' > v \). If \( d(v', n) \) was discarded it implies that \( d(v', n) + \sum_{i=1}^{n} \gamma_i > T \), thus infeasible. Therefore, if \( v' \) exists it is obtained only by an infeasible schedule. Since \( v^\ast \) is the highest payoff achievable by a feasible schedule, by definition \( v^\ast = V(S^\ast) \).

Construction of the table based on \( d(v, n) \forall v \in v, n \in N \) requires \( O(N \sum_{i=1}^{N} v_i) \) time, which makes the algorithm run in pseudo polynomial time. For strict deadlines, as those envisioned in future wireless services [1], reducing the computation time before making a decision is paramount. Considering the computation time as well as the time required to acquire the necessary network information for computing and computing a solution to (4.6) forces us to label finding \( S^\ast \) as impractical. Fortunately, producing a solution for (4.6) provides insights, such as the relation to (4.7) and an upper bound. In the pursuit of practicality, in the following section we approach our problem from the opposite direction in which pairs have local network information.

### 4.2.2 Opportunistic Strategies

We now present a strict delay service provisioning scheme using local network information in the form of an opportunistic strategy. Information regarding the cumulative distribution function of channels to the \( N \) users is available to the AP. The local network information scenario requires information that can be gained via implemented
protocols such as IEEE 802.11. We also present a greedy scheme which upper bounds the performance of random access protocols that do not consider scheduling.

Users in the local network information scenario still know their position in the probing order. Meaning, user \( i \in \mathbb{N} \) does not probe until after having listened to \( i - 1 \) other probes (and the potentially accompanying transmissions). This assumption allows users to avoid collisions, thus solutions presented in this section upper bound a system with collisions.

Due to not knowing the transmission rate to each user, the AP cannot solve (4.6), but the dynamic programming solution suggests a minimization of transmission time. Note that the minimization of the total transmission time requires full network information [88]. Instead we look at minimizing the time to first transmission, particularly minimizing the average time to first transmission, which is achievable with local network information. Given the stochastic nature of wireless networks we derive the optimal transmission decision in terms of minimum average time to first transmission.

Users can be considered as unselfish and efficient since they will transmit only if the transmission time is sufficiently small. By presenting our problem as a stopping problem we derive mathematically what is meant by sufficiently small in terms of the distribution function of \( t_i \) and \( \gamma_i \). Broadly speaking, a stopping problem considers a series of random events \( (t_i) \) for which there is a cost \( (\gamma_i) \) to transition between events and “stopping” (transmitting) at an event provides some payoff \( (\upsilon_i) \). An optimal
stopping decision is one which when used to evaluate if the series should stop at an event meeting a specific criteria then the average sum payoff is maximized.

Intuitively, for a very small $\gamma_i$ it is best for user $i$ to transmit only if $t_i = t_{\text{min}}$. Alternatively, for a large $\gamma_i$ then the first pair should always transmit. For simplicity, in this section we assume $\gamma = \gamma_i \forall i$. At the end of this section an example with a uniform distribution is provided to further elucidate intuition.

Without loss of generality, define the first transmitting pair as the $n$-th pair, where $1 \leq n \leq N$. Simply stated, if the $n$-th pair decides to transmit then there have been $n$ probings and no transmissions. The objective function in this section is defined as

$$\min_n E[n\gamma + t_n].$$  \hfill (4.9)

Note that even the first pair must probe, thus $n \geq 1$. While $n = 1$ minimizes the temporal cost of probing, having the first pair transmit may be suboptimal (e.g. if $t_1 = t_{\text{max}} = T - \gamma$). From optimal stopping theory \cite{84} we seek a strategy to stop at the optimal $n^*$ which minimizes (4.9). Define

$$n^* = \min\{n \geq 1 : t_n \leq \epsilon^*\},$$  \hfill (4.10)

where $\epsilon^*$ is the optimal threshold value which is used to decide if service $n$ should be transmitted or not. Thus, service $i$ is transmitted only if it takes time no greater than $\epsilon^*$. The value of $\epsilon^*$, assuming $\gamma > 0$, can be derived from (4.9) as follows
\[ \epsilon^* = E[\min\{t_n, \epsilon^*\}] + \gamma, \]

then

\[ \epsilon^* = \int_{-\infty}^{\epsilon^*} t_n \, dF(t_n) + \int_{\epsilon^*}^{\infty} \epsilon^* \, dF(t_n) + \gamma, \]

and by rearranging terms

\[ \gamma = \int_{-\infty}^{\epsilon^*} (\epsilon^* - t_n) \, dF(t_n), \tag{4.11} \]

where \( dF(t_n) \) corresponds to the derivative of the cumulative distribution function of the transmission time. Note that \( n \) is the first transmitted service, but all services after \( n \) need only update (4.11) with their variable values. The optimal stopping threshold has a dependency on \( \gamma \) since for large probing costs learning a new channel becomes prohibitively expensive. We refer to this solution as the local network information solution for strict delay services.

We define the greedy solution as the scheme using local network information and a threshold \( \epsilon^G = t_{\max} \), meaning that a service is transmitted after probing independent of channel conditions. Note that the greedy solution in our model, due to being collisionless, upper bounds random access medium access control schemes such as IEEE 802.11.

Consider the case when the transmission times follow a uniform distribution, that is \( t_i \sim U(t_{\min}, t_{\max}) \forall i \), then

\[ \epsilon^* = t_{\min} + \sqrt{2\gamma(t_{\max} - t_{\min})}. \tag{4.12} \]
Note 4.12 defines how “small” a time $t_i$ should be for service $i$ to be transmitted.

The service centric notion of distinguishable services is ignored in this scenario. Instead this section presents a relevant lower bound, in the form of 4.11, to our opportunistic scheme. The practicality of the local network information solution stems from not requiring a centralized scheduler or gossiping of network information. Unfortunately, the local network information solution ignores payoffs and therefore is suboptimal in terms of maximizing service payoffs.

### 4.2.3 Stochastic Optimization Strategies

We now present our opportunistic scheme for strict delay service provisioning with statistical network information scenario. The network information in this section assumes the AP has local network information plus service payoffs of all requested services and $t_n \forall n \in \mathcal{N}$ is i.i.d.. Given the stochastic nature of wireless networks and that statistical information is available, we seek to optimize the average sum service payoff, that is $E[V(S_{j,i})]$. We present a decision rule to opportunistically maximize the average sum service value and prove its optimality. Leveraging the problem formulation, a suboptimal algorithm with potentially less complexity is described.

Since statistical network information is available it is possible to replace $t_i$ in (4.6) with $E[t_i] \forall i$. Therefore, for a given $T$, $N$, and $\vartheta$ we can construct the optimization problem
\[
\begin{align*}
\max_x & \quad \sum_{i=1}^{N} x_i v_i \\
\text{s.t.} & \quad \sum_{i=1}^{N} x_i E[t_i] + \gamma_i \leq T,
\end{align*}
\tag{4.13}
\]

and define the optimal solution as \( S^E \). Note that \( S^E \) can be derived in the same way as \( S^* \) with the appropriate substitution. The solution of (4.13) results in a schedule that can be broadcast and strictly followed, thus representing a non-opportunistic scheduler only utilizing statistical network information. While obtaining \( S^E \) can offload computations away from individual users to a single AP, it fails to leverage the information gained from probing.

For our opportunistic scheme, decisions to transmit, or not, are done after probing. Each pair must decide what is the most efficient use of the available resource (time) given the attainable service payoff. An efficient use of resources must consider what can be gained from taking either the decision to transmit or not. Without loss of generality consider the decision made by the first pair after probing the channel. We propose that the first pair should transmit if

\[
v_1 + E[V(S_{j,i}|x_1 = 1)] \geq E[V(S_{j,i}|x_1 = 0)]. \tag{4.14}
\]

where \( V(S_{j,i}|x_1 = 1) \) is the maximum sum service value of a schedule given that the first service is transmitted, and similarly \( V(S_{j,i}|x_1 = 0) \) is the maximum sum service value of a schedule given that the first service is not transmitted.

The left-hand side of (4.14) represents what the first pair has gained plus what
is expected that other pairs will gain in the remaining time. The right-hand side of (4.14) represents what is expected that other pairs would gain if the first pair does not transmit. While equality between both sides implies that on average both options would result in equal payoff, we arbitrarily favor the slightly more deterministic left hand side. As a trivial example, for a very large $v_1$ and very small $t_1$ then the first pair will most likely transmit, but for a small $v_1$ and large $t_1$ the first pair will most likely not transmit.

The decision 4.14 holds for every transmitting pair by adjusting the service payoff, remaining probing order, and remaining time on both sides. Define the schedule $\psi$ as the schedule obtained if all pairs follow their version of decision 4.14. We refer to the schedule $\psi$ derived from following (4.14) for each service as the opportunistic solution.

Note that $\psi$ is derived from 4.14 and achieves the maximum average sum service payoff, i.e. $\max E[V(S_{j,i})]$, which is by definition different from the maximum sum payoff, i.e. $\max V(S_{j,i})$. As before, the problem at present is analogous to the knapsack problem. In this case, we have an estimate on the average weight (time) of the following items (pairs) to consider. In (4.14), note that $E[V(S_{j,i}|x_1 = 1)]$ is a smaller “knapsack” than $E[V(S_{j,i}|x_1 = 0)]$, thus there are fewer possible combinations to consider. A smaller search space, generally, leads to a shorter computation time, thus a more attractive situation to be in when strict deadlines are imposed. Simply stated, optimally filling a small knapsack is easier than filling a large knapsack.
For (4.14) the reduction in search space relates to the combinations enabled by $t_1$. For example, if $t_1 = 2(t_{\text{min}} + \gamma)$, $N = 3$, and $T = t_1 + \gamma$ then the combinations of the following two transmission times $t_2 = t_3 = t_{\text{min}}$ and $t_{\text{min}} < t_2 \leq 2(t_{\text{min}} + \gamma)$ must be considered in the right-hand side of (4.14) and not on the left-hand side. Reducing the complexity of the general knapsack problem is beyond the scope of this work (since it involves proving $P = NP$). Instead, our formulation offers a reduction in the size of the knapsack considered, thus gaining tractability in computing a solution.

When one considers strict deadlines as paramount then a reduced computation time becomes essential. Our proposed opportunistic scheme is upper bounded by the optimal solution of the full network information scenario and lower bounded by the local network information. The bounds are respectively a product of the greater and lesser available network information, but serve to benchmark the performance of our proposed opportunistic scheme.

4.2.4 Numerical Results

Through simulations we evaluate the performance of our opportunistic scheme. The optimal solution derived with full network information (Full NI) serves as an upper bound. The local network information (Local NI) solution is meaningful as a lower bound and, by contrast, details the benefits of service centric networks. The greedy solution is presented as a general benchmark and to represent an upper bound to random access medium access control protocols that do not consider scheduling. The
fixed scheme, derived under the statistical network information scenario, highlights the benefits of opportunistic operation. Unless otherwise noted, numerical evaluations consider $N = 20$, $T = 600$, $\gamma = 30$, $v_i \sim U\{1, 2, 3\}$ $\forall i$ and $t_i \sim U\{1, ..., 100\}$ $\forall i$.

Fig. 4.3 shows the average sum service payoffs as a function of the number of pairs. While the network is over resourced ($N \leq 6$) all schemes, except Local NI, select all pairs to transmit because there are enough resources. Recall Local NI will not select pairs with a high transmission time. As the network transitions to become under
resourced ($6 < N < 12$) the Full NI and Opportunistic scheme begin to diverge, due to our scheme relying on network statistics. When the network becomes under resourced ($N \geq 12$) the schemes stabilize since, on average, only so many nodes can be serviced. Note that Local NI achieves a higher performance than the Fixed scheme when payoffs between services are similar. Fig. 4.5 shows a scenario that differs from Fig. 4.3 by using $v_i \sim U\{1, 5, 10\}$ $\forall i$. When there is a greater difference in service payoffs the local network information schemes (Greedy and Local NI) begin to underperform even the Fixed scheme since they are non-service centric. Generally speaking, in the under resourced scenario the service centric schemes consider if the benefit (payoff) of resource consumption (transmission time) is sufficient to transmit or not. Therefore, when the distinction between service payoffs becomes more pronounced then the value of a service centric networks grows. Additionally, a higher service payoff also increases the value of additional network information in terms of performance since the performance difference in Fig. 4.5 is greater than in Fig. 4.3.

Fig. 4.4 shows the average delay for completed services as a function of the number of pairs. In terms of average delay the optimal solution is outperformed by local network information schemes (Greedy and Local NI). The optimal and opportunistic schemes have a higher delay per serviced pair when compared to the local network information schemes due to servicing more pairs. The local network information scheme outperforms in 4.4 since, from construction, the scheme is minimizing the average time to transmit.
Figure 4.4: Average delay of completed services as function of $N$. 
Figure 4.5: Average $\sum_{i=1}^{N} x_i v_i$ as function of $N$ with $v_i \in \{1, 5, 10\}$ $\forall i$. 
Fig. 4.6 shows the average payoff as a function of increasing probing cost to deadline ratio \((\gamma/T)\), which permits exploring the limits of deadline strictness within our model. Recall that the decision threshold for Local NI depends on \(\gamma\), hence the sharp spike for low values of the ratio \(\frac{\gamma}{T}\). An increase in \(\gamma/T\) quickly leads to the under resourced scenario since learning the channel for a few pairs quickly takes up the entire time before the deadline is reached. Given that the time to learn a channel is fixed, note that the fixed scheme has more abrupt changes (e.g. \(\frac{\gamma}{T} = 1/6\), \(\frac{\gamma}{T} = 1/4\), and around \(\frac{\gamma}{T} = 4/10\)) in considering stricter deadlines. The abruptness from the fixed scheme further promotes the use of opportunistic schemes to address tighter or varying deadlines.

The opportunistic and fixed schemes require more overhead than the local network information scenario. The overhead asymmetry depends on the implementation and would translate to the local network information scenario in having more time available to transmit before its deadline. Similarly, computation time depends heavily on implementation issues and decisions, and impact the relative performance of the schemes.

4.2.5 Conclusion

We presented a formulation for service centric networks with strict deadlines and probing cost. We approach our proposed problem under distinct network information availability scenarios. Our problem is analogous to the knapsack problem if there
Figure 4.6: Average $\sum_{i=1}^{N} x_i v_i$ as a function of $\frac{\gamma}{T}$. 
were a cost associated to learning the weight of each item. A method to achieve the optimal solution is proposed for the full network information scenario. Under the local network information scenario an optimal stopping solution is presented along with a greedy solution which upper bounds random access schemes. Our opportunistic scheme uses statistical information which in practice is more feasible to obtain than full network information.

The presented schemes serve to analyze the tradeoffs in performance relative to the available network information. Considering the temporal cost of probing a channel is relevant for future networks with much stricter delay requirements. If services are prioritized to gain sufficient distinction, then the benefits of a service centric network outperforms alternatives. Opportunistic schemes depend on gaining local information which can become prohibitive under stricter deadlines. Future work may explore the impact of imperfect local information on service centric networks.

4.3 Service Provisioning with Average Delay

For an AP providing services with an average delay, the optimal schedule $S^*$ can be obtained as the solution to

$$
\max_{S_{j,i}} V(S_{j,i})
$$

s.t. $\bar{T}(S_{j,i}) \leq \bar{T}$. \hfill (4.15)

Deciding if a user transmits (or not) depends on their transmission time, service value, and which transmissions have previously occurred. Unlike the strict delay scenario
the order in which transmissions occur gains importance due to impacting the average delay $\bar{t}(S_{j,i})$. Due to the increased feasible space of solutions (i.e. $N^2N!$), deciding both a probing order and transmission vector is a much harder problem than deciding either one when the other is fixed (e.g. fixing a probing order and searching for an ideal transmission vector).

Solving (4.15) requires a priori knowledge of $t_n \forall n \in N$, which is infeasible in practice. Yet, solving (4.15) gives a theoretical upper bound against which we benchmark practical strategies. We impose the requirement that users with a priori knowledge of $t_i$, and consequently $h_i$, must probe before transmitting during the execution phase such that the schedule obtained with a priori knowledge is a tighter upper bound than a schedule with a priori knowledge which ignores probing.

At first glance, (4.15) is reminiscent of the famous classical knapsack problem [71]. The knapsack problem considers a set of items, each with a value and weight, along with a weight limited knapsack. The objective in the knapsack problem is to maximize the sum value of items inside the knapsack without exceeding the weight limit. Among the many variations of the classical knapsack problem [71], the non-linear collapsing knapsack problem is closest to (4.15). The collapsing knapsack problem assumes that the weight constraint on the knapsack is a function of the items in the knapsack. For a wireless themed example of such a knapsack, consider the frequency partition problem among multiple users when a guard-band is required. A single user could occupy the entire spectrum for a transmission, but two users could not each use exactly half the
spectrum for transmission since part of the spectrum must be used for a guard-band.

The collapsing knapsack is known to be an NP-hard problem [71].

In the present scenario, the items are services, value is the service value, and weight is the transmission time. Constraining the average completion time results in a non-linear constraint which depends on items previously introduced in the knapsack (i.e. previously transmitted services). Unlike the collapsing knapsack problem, our non-linear constraint can go from “expanding” the knapsack to “collapsing” the knapsack, depending on when individual services are completed.

To illustrate the constraint expansion in our problem, consider two services with transmission times $\gamma_i + t_i < T$ and $\overline{T} = \gamma_j + t_j$. If service $j$ is transmitted first, the constraint is met (i.e. the knapsack is filled) and no other service can be transmitted. Alternatively, if service $i$ is transmitted first, then service $j$ can be transmitted second as long as $\gamma_j + t_j \leq 2(\overline{T} - \gamma_i - t_i)$. Thus, if service $i$ is transmitted first, the knapsack is apparently expanding from a constraint of $\overline{T}$ to $2\overline{T}$, but at the cost of an additional $\gamma_i + t_i$ to account for the completion time of $j$. Similarly, the constraint collapses when the completion time is greater or equal to $\overline{T}$, since the knapsack constraint expands by an additional $\overline{T}$ at a cost greater than $\overline{T}$. The fact that our non-linear constraint is not monotonic (i.e. can go from expanding to collapsing), impedes us from applying the reduction proposed in [73] while simultaneously motivating our approach at solving the “expanding then collapsing” knapsack present in our formulation.

Using average completion time as a constraint exacerbates our problem by con-
necting the transmission decision to the probing order. Furthermore, if the channel is probed and no transmission occurs, then resources have been wasted. Thus, it is unclear when a user should be placed in the probing order. In the following section we simplify our problem to gain intuition in solving (4.15).

4.3.1 Optimal Schedule for Average Delay Constraint and Large Transmission Times

For this section, and only this section, assume $\gamma_n \approx 0 \forall n$. Our assumption on $\gamma_n$ is applicable to systems where the transmission time is large (i.e. slow transmission rates or large packet lengths) relative to the probing time. Assuming $\gamma_n \approx 0$ simplifies our problem and leads to intuition on when a greedy order may help in obtaining a solution for (4.15).

First, we show that for a given transmission vector the average completion time is minimized by a greedy ordering. Then, we prove that the sum service value for a greedy order is no less than any other schedule under any ordering. Therefore, we can reduce the set of feasible schedules from $N!2^N$ to only $2^N$ without sacrificing optimality. Finally, we introduce a set of binary variables to linearize (4.15).

Define a probing order $\vartheta_g$ as a greedy order such that the transmission time of elements is non-decreasing for elements further along in the probing order, i.e.

$\vartheta_g = \{ \vartheta_g(i) : t_{\vartheta_g(i)} \leq t_{\vartheta_g(i+1)} \ \forall \ i \in N \}$, and define $t_{\vartheta_g(N+1)} = \infty$ for mathematical correctness. More than one greedy order may exist if any two services have equal
transmission time. The following theorem shows that any schedule \(S_{j,k}\) without a greedy order may reduce the average completion time \(T(S_{j,k})\) by adopting a greedy order.

**Theorem 3.** If \(\gamma_n = 0 \quad \forall \ n \in N\), for any set of transmission times \(t_i \quad \forall \ i \in N\) and a given transmission vector \(x_j\) the average completion time is minimized by a greedy order, i.e. \(\bar{T}(S_{j,k}) - \bar{T}(S_{j,g}) \geq 0 \quad \forall \ k \in \{1, ..., N!\}\).

**Proof.** We first prove the statement above for when the greedy order is unique, and then extend to the general non-unique case. Define the first differing service between \(S_{j,k}\) and \(S_{j,g}\) as \(\min_{l \in N}\{l : \vartheta_k(l) \neq \vartheta_g(l), x_j(\vartheta_g(l)) = 1\}\), meaning that \(\vartheta_k\) first differs from \(\vartheta_g\) at the \(l\)-th element, and the service in the \(l\)-th position of \(\vartheta_g\) is transmitted. Recall that \(x_j\) is independent of the probing order, and thus the service in the \(l\)-th position of \(\vartheta_g\) is transmitted in the \(l'\)-th order (where \(l' > l\)) in \(\vartheta_k\).

If the first differing service \(l\) is replaced in \(S_{j,k}\) by another transmitted service, i.e. \(x_j(\vartheta_k(l)) = 1\), it follows that

\[
t^c(S_{j,k}, l) > t^c(S_{j,g}, l), \quad (4.16)
\]

and

\[
t^c(S_{j,k}, l') > t^c(S_{j,g}, l'), \quad (4.17)
\]

since the transmission times in \(\vartheta_g\) are ordered such that \(t_{\vartheta_g(l)} < t_{\vartheta_g(l')}\) for \(l < l'\).

When the first differing service \(l\) is replaced in \(S_{j,k}\) by a non-transmitted service, i.e.
\( x_j(\vartheta_k(l)) = 0 \), it follows that
\[
\tau^c(S_{j,k}, l') \geq \tau^c(S_{j,g}, l),
\]
(4.18)
with equality when \( x_j(\vartheta_g(\lambda)) = 0 \ \forall \ \lambda \in \{l + 1, \ldots, l' - 1\} \). When \( x_j(\vartheta_k(l)) = 0 \) the relation between \( \tau^c(S_{j,k}, l) \) and \( \tau^c(S_{j,g}, l') \) is irrelevant since \( x_j(\vartheta_k(l)) = x_j(\vartheta_g(l')) = 0 \) and do not impact the average completion time (4.4). For a fixed transmission vector \( \mathbf{x}_j \) the denominator in the average completion time (4.4) is independent of the probing order. Thus, if the first differing service \( l \) is replaced by another transmitted service or by a non-transmitted service it follows that
\[
\tau(S_{j,k}) - \tau(S_{j,g}) \geq 0.
\]
(4.19)
The same argument above holds for any other differing elements. To complete the proof note that when the greedy order is not unique, the relationships in (4.16) and (4.17) are replaced with \( \geq \) and the same conclusion as above can be obtained.

While minimizing the average completion time is beneficial, ultimately we seek to maximize the payoff of completed services. As a corollary from Theorem 1, it can be shown that for any feasible schedule \( S_{j,k} \) with sum service value \( V(S_{j,k}) \), the schedule \( S_{j,g} \) can obtain the same sum service value \( V(S_{j,g}) = V(S_{j,k}) \). For a fixed greedy order \( \vartheta_g \) define \( S_{*,g} = \{x_*, \vartheta_g\} \) as the schedule which achieves the optimal solution to
\[
\max_{\mathbf{x}_j} \ V(S_{j,g})
\]
(4.20)
\[
s.t. \quad \bar{\tau}^c(S_{j,g}) \leq \bar{\tau}.
\]
Note that (4.20) searches only through the set of feasible transmission vectors $x_j$, while (4.15) searches through the set of feasible schedules $S_{j,k}$. The following theorem shows that when $\gamma_n = 0 \ \forall \ n \in N$, to obtain $V(S^*)$ it suffices to solve (4.20).

**Theorem 4.** If $\gamma_n = 0 \ \forall \ n \in N$, then $V(S^*_{s,g}) \geq V(S_{j,k})$ for any feasible schedule $S_{j,k}$ with $j \in \{1, \ldots, 2^N\}$ and $k \in \{1, \ldots, N!\}$.

**Proof.** Without loss of generality, assume that $S_{j,k} = S^*$, since any other feasible schedule for (4.15) would not achieve a higher sum service value. To arrive at a contradiction assume that $V(S^*_{s,g}) < V(S_{j,k})$ where $S^*_{s,g}$ is the solution to (4.20).

First, consider the case when $S_{j,g}$ is not a feasible schedule (i.e. not a feasible solution of (4.20)). By definition, a schedule is feasible if the average completion time constraint is met. From Theorem 1, we know that the average completion time is minimized by adopting $\vartheta_g$. Therefore, if $S_{j,g}$ is infeasible, then $S_{j,k}$ is also infeasible for (4.15). An infeasible schedule, by definition, cannot be optimal.

Next, consider the case when $S_{j,g}$ is a feasible schedule. Given that $S_{j,g}$ is a feasible schedule and has a fixed greedy order it follows that $V(S^*_{s,g}) \geq V(S_{j,k})$, else $S^*_{s,g}$ would not achieve the optimal solution of (4.20). Thus we arrive at a contradiction, which implies that $V(S^*_{s,g}) \geq V(S_{j,k})$ as desired. \qed

Theorem 4 shows that when $\gamma_n = 0 \ \forall \ n \in N$, solving (4.20) is sufficient to obtain $S^*$, i.e. the solution to (4.15). Note that the form in which (4.20) is presented is intractable due to the underlying non-linearity present in the calculation of the average delay.
To avoid the non-linearity in (4.15), we expand the parameter space by introducing the binary matrix $y_j$ with $y_j(n,i) \in y_j \forall n, i \in N$. For any $i \leq n \leq N$ the element $y_j(n,i) = 1$ only if both $x_j(i) = 1$ and $x_j(n) = 1$, and $y_j(n,i) = 0$ otherwise. Essentially, if service $n$ is transmitted, then the value of $y_j(n,i)$ indicates if the service $i$ was previously transmitted. Having a linear optimization problem allows the use of well known and studied solving techniques [89, 71].

Using $y_j$ we can reformulate (4.20) as

$$\max_{x_j} V(S_{j,k})$$

$$s.t. \sum_{n=1}^{N} \sum_{i=1}^{n} t_{\partial_{y}(i)} y_j(n,i) - T x_j(n) \leq 0$$

$$x_j(i) + x_j(n) - y_j(n,i) \leq 1$$

$$y_j(n,i) - x_j(i) \leq 0$$

$$y_j(n,i) - x_j(n) \leq 0$$

where $i \in N$ and $i \leq n$. The first constraint is the average completion time constraint. The second, third, and fourth set of constraints in (4.21) guarantee that if node $i$ transmits before $n$, then $t_{\partial_{y}(i)}$ is counted towards the completion time of $n$ and the completion time of $n$ is non-zero only if $n$ transmits. Introducing $y_{i,n}$ creates $N(N+1)/2$ additional constraints, but takes away the non-linear property of (4.15). Note that (4.21) has at most $2^N$ feasible solutions while (4.15) has at most $N!2^N$ feasible solutions, yet from Theorem 4 we retain optimality.

With our problem well formulated as a linear integer problem, a solution can be computed via typical optimization methods [89]. Numerical results presented in this
thesis are derived via the branch and bound method. Note that computing $V(S^*)$ requires a priori knowledge of $t_n \forall n \in N$ (i.e. full network information) which is available only via the impractical genie aided approach. Thus, the solution $V(S^*)$ obtained from solving (4.21) is achievable only in theory, but is an upper bound for us to benchmark possible implementable strategies.

4.3.2 Optimal Average Delay Service Scheduling with Probing Cost

Now we extend our analysis on average delay service provisioning to non-negligible probing cost scenarios (i.e. $\exists n : 0 < \gamma_n \approx t_n$). Such scenarios are motivated by future services demanding higher levels of interactivity and shorter network response times. We obtain an optimal schedule $S^*$ by re-ordering the solution of a reformulation of (4.21) inspired from the intuition gained in the previous Section.

When $\gamma_n > 0$ the benefit of a greedy order $\vartheta_g$ in obtaining $S^*$ is lost whenever a user probes without transmitting. For example, consider $N = 2$, $t = \{1, 1.5\}$, $v = \{1, 2\}$, $T = 1.6$, and $\gamma_n = 0.1 \forall n$. The optimal payoff $V(S^*) = 2$ is obtained only with the non-greedy ordering $\vartheta_k = \{2, 1\}$ and $x_j = \{0, 1\}$. Therefore, the result from the previous section does not hold for $\gamma > 0$.

From the previous example, intuition suggests that non-transmitting elements, independent of their transmission time, should be pushed to the end of the order. Following this intuition, for a schedule $S_{j,k}$ define a transmission first schedule $S_{j,f}$ with order $\vartheta_f$ defined as $\vartheta_f = \{\vartheta_f(i) : t_{\vartheta_f(i)} \leq t_{\vartheta_f(i+1)} \forall x_j(\vartheta_f(i)) = 1, x_j(\vartheta_f(i+1)) = $
\{1 \cup \{ \vartheta_f(i) : \vartheta_k(i) \neq x_j(i) = 0 \}. While the \vartheta_f can be described as a greedy ordering over only transmitted services, we use the term “transmission first” to avoid confusion with the previously defined greedy order \vartheta_g which is independent of \( x_j \). Essentially, non-transmitted services utilize temporal resources (due to probing) without benefit for the network. The following Lemma showcases the benefits of the transmission first schedule.

**Lemma 1.** For any schedule \( S_{j,k} \) such that \( \exists i \in N : x_j(i) < x_j(i+1) \) a transmission first schedule \( S_{j,f} \) achieves \( V(S_{j,f}) = V(S_{j,k}) \) with \( T(S_{j,f}) < T(S_{j,k}) \).

The proof for Lemma 1 is omitted due to similarity with the proof of Theorem 3.

Of course, obtaining \( \vartheta_f \) is only possible for a given transmission vector \( x_j \). Since the decision to transmit is done only after probing, we cannot practically obtain \( \vartheta_f \) for each schedule. Following the intuition that a transmission first schedule avoids wasting resources, we reformulate (4.15) such that for any probing order only transmitted services incur the probing penalty, i.e.

\[
\max_{x_j} V(S_{j,k})
\]

\[\text{s.t.} \quad \sum_{n=1}^{N} \sum_{i=1}^{n} (t_{\vartheta_g(i)} + \gamma_i) y_j(i,n) - T x_j(n) \leq 0\]

\[x_j(i) + x_j(n) - y_j(i,n) \leq 1 \quad (4.22)\]

\[y_j(i,n) - x_j(i) \leq 0\]

\[y_j(i,n) - x_j(n) \leq 0,\]

where \( i \in N, i \leq n, \) and \( y_j \) as previously defined. Note that since (4.22) considers a
probing penalty only if the service is transmitted, it follows that a schedule \( S_{j,k} \) that optimally solves (4.22) may incur an average completion time such that it is infeasible for (4.15).

We emphasize that (4.22) does not precisely represent the problem of providing services with average delay when users probe before deciding to transmit, since probing penalty is avoided when a user does not transmit. Yet the optimal solution to (4.22), i.e. the vector \( x^* \), is the same as the vector which solves (4.15). The value of our reformulation is derived from the fact that we can reorder the solution of (4.22) to obtain a feasible schedule for (4.15). Now, we show how to obtain \( S^* \) if \( \exists \gamma_n > 0 \) by solving the reformulation of (4.15) with a constrained set of feasible schedules.

Define \( S_{*,f} \) as the solution of (4.22) for \( \theta_k = \theta_g \) after being reordered as a transmission first schedule. Meaning, to obtain \( S_{*,f} \) we solve (4.22) for a greedy order \( \theta_g \) and then reorder the resulting schedule as a transmission first schedule. The following theorem shows that \( V(S^*) \) can be obtained by the reordered solution of (4.22).

**Theorem 5.** If \( \gamma_n > 0 \) then \( V(S_{*,f}) \geq V(S_{j,k}) \) for any feasible schedule \( S_{j,k} \) with \( j \in \{1, ..., 2^N\} \) and \( k \in \{1, ..., N!\} \).

**Proof.** Without loss of generality assume \( S_{j,k} = S^* \) since any other feasible schedule would achieve a lower sum service value. Note that from Lemma 1 we may reorder any feasible schedule into a transmission first schedule which is also feasible. Thus, as in Theorem 4, the schedule \( S_{j,f} \) is feasible and \( V(S_{*,f}) \geq V(S_{j,f}) = V(S_{j,k}) \). Alternatively \( S_{j,f} \) is infeasible which, from Lemma 1, implies that \( S_{j,k} \) is also infeasible.
Theorem 5 shows that the optimal sum service value $V(S^*)$ can be obtained by adopting a transmission first schedule from the solution to (4.22) for the greedy order $\vartheta_g$. The greedy ordering reduces the average completion time while the transmission first property discards non-transmitting elements from the schedule. Note that to solve (4.22) we still need to know $t_i \forall i \in \mathcal{N}$ a priori. As before, we are able to reduce the feasible solutions from $N!2^N$ to only $2^N$ without sacrificing optimality. The optimal solution is impractical and, simultaneously, useful as a theoretical upper bound for practical strategies described in the following section.

4.3.3 Practical Strategies for Providing Services with Average Delay

We now detail strategies to provide services with average delay constraint that do not require a priori knowledge of $t_i \forall i \in \mathcal{N}$. Two strategies based on stochastic programming and a stopping theory strategy are described. The strategies based on stochastic programming are appropriate for networks where computation capabilities are available during the scheduling period. The stopping theory strategy is far less onerous in terms of computation for users, since the majority of the computations can be done a priori to the scheduling period. All presented strategies are practical since they can be executed without a priori knowledge of $t_i \forall i$, meaning that they fit our assumption that probing is required before knowing $t_i$. The performance of the proposed strategies is evaluated in the following section.
4.3.4 Stochastic Programming

For the strategies in this subsection, we assume that all network elements (or a single centralized scheduler) know the deterministic aspects of the network (i.e. \(N\) and \(v_i \forall i \in N\)), and have statistical information regarding transmission times. The information required by this assumption can be obtained during the polling and ordering phases.

Given a packet of \(b_n\) bits, a user can compute an estimate of the transmission time as

\[
E[t_n] = \frac{b_n}{E[r_n]} \quad \forall \ n \in N.
\]  

(4.23)

The value of \(E[t_n]\) can be estimated if the statistical distribution of the channel is known, or calculated as a running average. If the network topology is mostly static (i.e. low mobility), then a running average over many samples could be used. If the topology is highly dynamic (i.e. high mobility), then a shorter running average might be preferred to calculate \(E[t_n]\). Following the intuition obtained from the previous Section, define the stochastic greedy order as

\[
\vartheta_{E[g]} = \{ \vartheta_{E[g]}(n) : E[t_{\vartheta_{E[g]}(n)}] \leq E[t_{\vartheta_{E[g]}(n+1)}] \ \forall \ n \in N \}.
\]

The stochastic greedy order is intuitive and throughput optimal for medium access control with opportunistic transmissions [90].

By substituting \(t_n\) for \(E[t_n]\) we can establish a stochastic optimization version of
(4.22) as

\[
\max_{x_j} V(S_{j,k}) \\
\text{s.t. } \sum_{n=1}^{N} \sum_{i=1}^{n} \left( E[t_{\vartheta_g(i)}] + \gamma_i \right) y_j(i, n) - T x_j(n) \leq 0 \\
x_j(i) + x_j(n) - y_j(i, n) \leq 1 \\
y_j(i, n) - x_j(i) \leq 0 \\
y_j(i, n) - x_j(n) \leq 0.
\] (4.24)

Define \( x_E \) as the transmission vector which solves (4.24) for \( \vartheta_E \), and the schedule \( S_E = \{ x_E, \vartheta_{E[g]} \} \). Note the AP can solve (4.24) and broadcast \( S_E \) at the end of the ordering phase. Define as the static stochastic strategy a network which strictly follows \( S_E \). While the static stochastic strategy limits the computational requirements to a single network element, e.g. the AP, it does not leverage information obtained after each probing.

A network may benefit from appropriately using the available network information. For example, if during the execution of \( S_E \), with \( x_n \), service \( n \) finds that \( t_n \gg E[t_n] \), then perhaps \( n \) should not transmit. User \( i \in N \) can leverage knowledge
of previous transmissions by solving

\[
\max_{x_j(\theta_k(i)) \ldots x_j(\theta_k(N))} V(S_{j,k})
\]

\[
\text{s.t. } \sum_{n=1}^{N} \sum_{l=1}^{n} (t'(\theta_k, i, l) + \gamma_l) y_j(i, n) - T x_j(n) \leq 0
\]

\[
x_j(i) + x_j(n) - y_j(i, n) \leq 1
\]

\[
y_j(i, n) - x_j(i) \leq 0
\]

\[
y_j(i, n) - x_j(n) \leq 0
\]

where

\[
t'(\theta_k, i, l) = \begin{cases} 
  t_{\theta_g(l)} x_j(\theta_g(l)), & \text{if } l \leq i \\
  E[t_{\theta_g(l)}], & \text{if } l > i.
\end{cases}
\]

Note that the optimization in (4.25) is only over the variables \( \{x_j(\theta_k(i)), \ldots, x_j(\theta_k(N))\} \) since all transmission variables before \( i \in \mathbf{N} \) have been decided by the time that user \( i \) probes. Define a network in which users solve (4.25) to decide to transmit (or not) during the execution phase as the dynamic stochastic strategy. Define \( S_D = \{x_D, \theta_E[g]\} \) as the dynamic stochastic schedule.

The static stochastic strategy decides a schedule (i.e. decides when and which users transmit) at the start of the execution phase with no knowledge of channel gains. The dynamic stochastic strategy decides the remainder of the schedule after each probing. While learning a channel gain offers more information that can be leveraged for better performance, the dynamic strategy carries a high computational cost. Therefore we next present a stopping theory approach which sequentially probes before deciding to transmit without incurring a large computational cost.
4.3.5 Optimal Stopping Strategy for Average Delay Service Provisioning

We now approach scheduling as a stopping problem [83] to obtain a simple threshold-based strategy. Stopping problems consider a stochastic sequence, where a decision to stop the sequence is made to optimize a function of sequence realizations so far observed. For us, the decision to stop is analogous to the decision to transmit, the stochastic sequence is a function of the transmission times, and realizations occur in the order defined by the probing order.

For $1 < N < \infty$, more than one decision to stop may be made and at most finite number $N$ of probings are made; thus we have a multiple-stopping problem with finite horizon. A multiple-stopping problem can be treated as multiple single-stopping problems with random time (e.g. transmission time) between each stop without sacrificing optimality [91].

Recall the definition for a sequence of pairs $\{X_n, F_n\}$ $\forall n \in \mathbb{N}$, where $X_n = g(h_1, ..., h_n)$ is a function of the random variables $h_m \forall m \leq n$ observed so far (i.e. $X_n$ is also a random variable) and $F_n$ is the filtration at $n$ of the $\sigma$-algebra $\mathcal{F}$. Meaning that $F_n$ is indexed by the order set $n \in \{1, ..., N\}$ such that $F_n \subseteq F_{n+1} \forall n$ and $E[X_n | F_n] = X_n$. Note that $X_n \forall n \in \mathbb{N}$ need not be i.i.d. to obtain an optimal solution of a stopping problem [92].

When trying to find the maximum of a sequence, we do not stop the sequence while the sequence is a sub-martingale. Recall, a sequence of pairs is defined as a sub-martingale if
\[ F_k \subset F_n \subset F_N \ \forall \ k < n \in \{1, \ldots, N\}, \]

\[ X_n \text{ is an } F_n\text{-measurable RV with finite first moment } \forall n, \]

\[ \text{and } X_k \leq E[X_n|F_k] \ \forall \ k < n. \]

Alternatively, when trying to maximize a function, as long as the sequence remains a sub-martingale we should continue probing. Essentially, probing continues as long as continuing probing is expected to be a more profitable action, where “profitable” is defined by the function we seek to maximize.

Considering the knapsack-esque quality of our optimization problem, we note that Dantzig’s heuristic is a well studied and widely accepted heuristic to solve the classical knapsack problem [89]. Dantzig’s heuristic sequentially adds the items with the highest payoff-to-weight ratio until the constraint is met. Since the transmission time (i.e. weight) is not known a priori, we cannot directly apply Dantzig’s heuristic to our problem. Nonetheless, inspired by Dantzig’s heuristic for the traditional knapsack problem, we define the stopping problem

\[ \max_{n \in \mathbb{N}} E[\beta_n r_{\vartheta_k(n)}], \] (4.27)

where \( \beta_n = \frac{\upsilon_{\vartheta_k(n)}}{b_{\vartheta_k(n)}} \) is the service value-to-bits ratio of the \( n \)-th user. While (4.27) incurs no penalty by continuing the sequence, the finite horizon deters us from waiting for perfect channel conditions. Unlike the infinite horizon for a stopping problem, as explored in [78, 79], the finite horizon problem has a threshold value which depends on the number of remaining elements in the sequence. Recall that knowledge of service
values and number of bits (i.e. \(v_n\) and \(b_n\)) is obtained during the polling phase.

If \(\beta_n r_{\vartheta_k(n)}\) is large enough then, intuitively, we should stop and transmit. Alternatively, if \(\beta_n r_{\vartheta_k(n)}\) is too small, intuitively, we should continue to the next user at the risk of running out of users to probe. An optimal solution to (4.27) is obtained by deriving a set of thresholds which precisely and mathematically define what is meant by \(\beta_n r_{\vartheta_k(n)}\) being “too small” or “large enough”. Leveraging the finite horizon property, we obtain the set of optimal thresholds via backwards induction.

Obviously, the \(N\)-th user to probe should always transmit regardless of \(\beta_N r_{\vartheta_k(N)}\), as long as the average completion time constraint is satisfied. Although trivial, for completeness define the maximum expected service value-to-bits ratio if the \(N\)-th user does not transmit as \(\epsilon_N = 0\). We say that the threshold for the \(N\)-th user is \(\epsilon_N = 0\), i.e. the \(N\)-th user to probe transmits if \(\beta_N r_{\vartheta_k(N)} \geq 0 = \epsilon_N\). Note that for a network using the stopping strategy, every user knows that the last user to probe will transmit if schedule feasibility is maintained.

Now, the \(N-1\)-th user to probe will know \(\beta_{N-1} r_{\vartheta_k(N-1)}\) and should transmit if the expected service value-to-bits ratio of the \(N\)-th user to probe is less than \(\beta_{N-1} r_{\vartheta_k(N-1)}\). The \(N-1\)-th user must compare what they have, i.e. \(\beta_{N-1} r_{\vartheta_k(N-1)}\), against what is expected to be gained if the sequence continues. Formally, define \(\epsilon_N\) as the expected maximum service value-to-bits ratio of the \(N\)-th user if the \(N-1\)-th user does not transmit, which mathematically is

\[
\epsilon_{N-1} = \max \{ E[\beta_N r_{\vartheta_k(N)}|\mathcal{F}_{N-1}], \epsilon_N \}.
\]
which can be developed to

\[ \epsilon_{N-1} = \beta_N E[r_{\vartheta_k(N)}] = \beta_N \int_{0}^{\infty} r dF_N(r), \quad (4.28) \]

where \( r \) is an integration variable and \( F_N \) the cumulative distribution function of the transmission rate from user \( N \). The \( N - 1 \)-th user to probe should transmit if \( \beta_{N-1}r_{\vartheta_k(N-1)} \geq \epsilon_{N-1} \) and the completion time constraint is satisfied. Note that the value of \( \epsilon_{N-1} \) is independent of \( F_{N-1} \), and thus \( \epsilon_{N-1} \) can be computed a priori.

The \( N - 2 \)-th user to probe knows that the \( N - 1 \)-th user will transmit only if \( \beta_{N-1}r_{\vartheta_k(N-1)} \geq \epsilon_{N-1} \) and that the \( N \)-th user will transmit regardless of what probing reveals for the \( N \)-th user. Therefore, the expected maximum service value-to-bits ratio if the \( N - 2 \)-th user does not transmit is

\[ \epsilon_{N-2} = \max\{E[\beta_{N-1}r_{\vartheta_k(N)}|F_{N-1}], \epsilon_{N-1}\}, \]

which can be developed to

\[ \epsilon_{N-2} = \epsilon_{N-1}F_{N-1}(\epsilon_{N-1}) + \beta_{N-1} \int_{\epsilon_{N-1}}^{\infty} r dF_{N-1}(r), \quad (4.29) \]

where, again, \( r \) is used as a variable of integration. Thus a transmission should occur whenever user \( N - 2 \) observes a value of \( r_{\vartheta_k}(N - 2) \) such that \( \beta_{N-2}r_{\vartheta_k(N-2)} \geq \epsilon_{N-2} \), since on average the user is expected to obtain a higher pay off by transmitting than by continuing the probing phase.

From (4.29) we can generalize a formula for \( \epsilon_n \), where \( n \) is the total number of users that have already probed, as

\[ \epsilon_n = \epsilon_{n+1}F_{n+1}(\epsilon_{n+1}) + \beta_{n+1} \int_{\epsilon_{n+1}}^{\infty} r dF_{n+1}(r). \quad (4.30) \]
Recall that $\epsilon_N = 0$ and that $\epsilon_n \forall 1 \leq n \leq N$ can be calculated recursively and a priori from a scheduling period.

Define the schedule $S^Z = \{x_z, \vartheta_z\}$, where the transmission vector $x_z$ is such that transmission at the $n$-th probing occurs only if $\beta_n r_{\vartheta_z(n)} \geq \epsilon^Z_n$ and feasibility of the schedule is maintained. The Dantzig stopping strategy uses schedule $S^Z$.

### 4.3.6 Discussion of Practical Strategies for Average Delay Service Provisioning

The strategies presented in this section require only information that can be practically obtained (i.e. non-genie aided). The strategies leveraging stochastic programming can become highly demanding in terms of computation. Particularly, the dynamic stochastic strategy requires each user to be able to solve a stochastic program. The complexity of the stochastic programs considered will grow with the number of users, and therefore the dynamic strategies may be infeasible for many applications.

Alternatively, the stopping strategy require users to compare their measurements against the a priori computed thresholds. Stopping strategies demand from users minimal computations, since the threshold computation can be done elsewhere (e.g. the AP, or even outside the network). Unlike the stochastic programming strategies, the decision obtained from the stopping strategies is done without regard to what other users may decide.

The following section shows numerical results to evaluate the performance of the
presented strategies against the optimal genie aided solution. Intuitively, strategies utilizing more of the available information should obtain a higher performance. Ultimately, selecting a solution depends on the application, capabilities, and required performance of a network.

4.3.7 Numerical Results

We evaluate our proposed strategies via simulation and benchmark against the optimal solution and two baseline strategies. To serve as baseline strategies, we consider a greedy strategy and an exhaustive strategy. The greedy strategy uses the same probing order as the Dantzig stopping strategy and transmits regardless of what the channel conditions are (i.e. the greedy strategy would have stopping thresholds \( \epsilon^G = \epsilon^G_n = 0 \ \forall \ n \in \mathcal{N} \)). The exhaustive strategy is allowed to probe all channels before making any decision to transmit or not. The greedy strategy serves as a simple alternative to the practical strategies and the exhaustive strategy showcases the burden of learning the entire network. Strategies, implementation requirements, and qualities are presented in Table tab:dwn. The Dantzig stopping strategy and greedy strategy use a probing order of decreasing expected value-to-time ratio.

For numerical results, channels \( h_i \) are drawn from a zero-mean unit variance normal distribution. Unless otherwise noted, a transmit power of \( P = 15dB \) to the unit noise power \( Z \) is available to each node, service values (i.e. \( w_i \)) and packet lengths (i.e. \( b_i \)) are drawn from uniform distribution between (1, 5), \( N = 10 \), the average com-
pletion time deadline is set to $T = 2$ time units, and probing time duration $\gamma = 0.1$ time unit. Our decision for the initial values of $\gamma$ and $T$ follows from noting that typical probing durations are in the order of a tenth of the total transmission time [29, 78, 86]. Channel bandwidth $W$ is set to an arbitrary unit size to allow packet sizes to scale to the available channel bandwidth (i.e. larger channel bandwidth would carry a greater amount of information, ergo greater number of bits per packet).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Requires</th>
<th>Preferable if</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal</td>
<td>Genie</td>
<td>Genie available</td>
</tr>
<tr>
<td>Dynamic Stoch.</td>
<td>Solve (4.25)</td>
<td>Highest payoff at high cost</td>
</tr>
<tr>
<td>Static Stoch.</td>
<td>Solve (4.24) once</td>
<td>Require fixed schedule</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>Probe all users first</td>
<td>Not preferable</td>
</tr>
<tr>
<td>Dantzig Stopping</td>
<td>Precompute $\epsilon_n$</td>
<td>High payoff at low cost</td>
</tr>
<tr>
<td>Greedy Stopping</td>
<td>Any probing order</td>
<td>No computations</td>
</tr>
</tbody>
</table>

Fig. 4.7 shows the average sum service payoff as a function of $N$. For higher values of $N$ there is a higher probability of having several good channels (i.e. small $t_i$) for high value services (i.e. larger $w_i$) thus more higher value transmissions may occur. The exhaustive strategy peaks at $N \approx 5$ since the cost of learning all channel conditions (i.e. $N\gamma$) overtakes the benefit of finding the optimal solution. The
Figure 4.7: Average payoff as a function of $N$. As $N$ grows the preferred strategy changes from greedy, to dynamic, to mean stopping.
greedy strategy and Dantzig stopping strategy utilize the same probing order, but the Dantzig stopping strategy benefits from avoiding adverse channel conditions, as defined by the thresholds $\epsilon_n$. Both stochastic programming strategies slightly outperform the Dantzig stopping strategy at the cost of increased computational complexity. Note that at large $N$, solving even a single instance of (4.25) may be extremely cumbersome for many wireless communication devices.

The average throughput per scheduling period as a function of $N$ is shown in Fig. 4.8. Among the practical strategies the Dantzig stopping strategy, interestingly, achieves a higher throughput per scheduling period. Recall that the stochastic programming-based strategies are focused on maximizing the service value, shown in Fig. 4.7, while the Dantzig stopping strategy balances service value and throughput by the ratio $\beta_n \forall n$.

The average time to completion delay as a function of $N$ is presented in Fig. 4.9. Note that a higher value in Fig. 4.9 implies a higher utilization of the frequency spectrum (i.e. the schedule produced a smaller gap to $\overline{T}$). As with Fig. 4.7, the exhaustive strategy is overburdened by the cost of learning the network with performance dropping after $N \approx 4$. The greedy strategy achieves a higher average completion time delay than the other practical strategies since all users that probe are guaranteed to transmit. Alternatively the other practical strategies (i.e. not greedy strategy) incur the cost of some users probing without transmitting, which consumes temporal resources by raising the completion deadline of the next transmitting users.
Figure 4.8: Average throughput as a function of $N$. If we consider the performance and computational cost, as $N$ grows the preferred strategy changes from greedy, to stochastic programming, to Dantzig stopping strategy.
Figure 4.9: Average throughput per scheduling period, i.e. $\sum_{i=1}^{N} b_i x_i$, as a function of $N$. The Dantzig stopping strategy achieves the highest throughput among practical strategy, although not achieving the highest sum service value (see Fig. 4.7).
Figure 4.10: Average Payoff as a function of transmit power $P$. All strategies, except the Dantzig stopping strategy, exhibit a nearly linear increase in performance with transmit power.
The impact of varying the transmit power $P$ on sum service value is shown in Fig. 4.10. A lower transmit power invariably leads, on average, to lower achievable rate, thus a higher transmission time per service. When the transmission time is large, the Dantzig stopping strategy is likely to probe many times before transmitting, and thus achieving a lower sum payoff than the greedy strategy. At $P \approx 8dB$ the Dantzig strategy overcomes the greedy strategy, which underpins a cutoff for the rate available to a network and the decision to use either Dantzig stopping strategy or the greedy strategy.

To account for variability in the configuration delay that distinct networks can encounter, we analyze the impact of the probing period duration $\gamma$ in Fig. 4.11. Recall that $N = 10$. Hence at the smallest ratio shown the exhaustive strategy, due to the overhead delay, achieves a performance comparable to the greedy strategy at a larger ratio (i.e. $\approx 0.05 = \frac{2}{20}$). At a value of $\gamma = 0$ the exhaustive strategy would match the performance of the Optimal strategy since there would be no overhead cost to learn the entire network.

Fig. 4.12 shows the throughput achieved by the various strategies as a function of the probing period duration $\gamma$. Note that only for a certain range (i.e. $\gamma \leq 0.65$) does the Dantzig stopping strategy achieve a throughput higher than the Dynamic stochastic strategy (cf. 4.8). Recall that the threshold $\epsilon_n$ obtained by the Dantzig stopping strategy is independent of the probing period duration $\gamma$. Thus as $\gamma$ grows, the cost of not transmitting grows as well and the strategy achieves a lower through-
Figure 4.11: Average Payoff as a function of probing length $\gamma$. 
Figure 4.12: Average throughput per scheduling period as a function of probing length $\gamma$.

Note that the curvature of the plots in Fig. 4.11 is distinct from Fig. 4.12, more so than Fig. 4.7 and Fig. 4.8, showing that the sum service value is not simply scaling the achieved throughput. Thus the notion of service value could become disjoint of the throughput performance of a network.

Overall, we found that the dynamic stochastic strategy performs the best in terms of service value. Considering that forcing each node to solve an optimization problem
may very well be infeasible, we propose that the Dantzig stopping strategy be used due to the similar performance in service value and higher throughput. Alternatively, when optimization is trivial (i.e. small $N$) or the received signal-to-noise ratio is small (i.e. small achievable rates), a greedy solution should be pursued and all computations should be avoided.

4.3.8 Summary

A key performance criterion for wireless networks is the capability of providing services. To this end, we presented a framework which quantifies the service providing capabilities by assigning value to services with an average completion time. Service values are an abstraction of the priority that could be set by the user or network operator. The underlying optimization problem is made tractable by linearization which expands the variable space, with our results counteracting the increase in space by showing that an optimal schedule uses an activation first ordering.

Although the stochastic programing strategy achieves the highest service provision and does not require a priori information, said strategy incurs a heavy computational burden on network elements. Based on our evaluation, the Dantzig inspired stopping strategy would achieve similar performance at a much smaller computational cost.
Chapter 5

Conclusion and Future Directions
In this thesis, we proposed decentralized opportunistic strategies to maximize throughput and service provisioning capabilities of mmWave networks. Our approach assumes no a priori channel state information, ergo no a priori knowledge of the achievable rate or transmission time over any of the potential transmission channels. Such an assumption allows us to directly consider the overhead cost into the optimization of the network.

Motivated by directional transmissions and high AP density present in mmWave networks, we showed that the throughput maximization problem can be modeled as a stopping problem with finite horizon and unreliable recall. The “finite horizon” is due to the finite APs and the “unreliable recall” is due to the possibility of a previously probed AP being unavailable at present. The solution to a stopping problem is a set of thresholds that can be computed a priori. Our solution hinges on knowing with what probability a previously probed AP is available, or not, at present. In lieu of experimental evaluation to characterize said probability, we characterize the worst case error on the probability and thus lower bound the performance of our strategy. Furthermore, analytical expressions for the expected delay and overhead as functions of the threshold selection are presented.

The throughput-delay tradeoff for wireless network is well known and studied under network conditions and assumptions [93, 94, 95, 96]. The analytical expressions here presented can help guide fine tuning thresholds so that various points on the throughput-delay curve may be obtained. Our approach to opportunistic transmis-
sions as stopping problem with finite horizon and unreliable recall can be implemented for objective functions other than throughput maximization, e.g. energy minimization and underlay networks. For a given set of coding and modulation rates, similar to [78], a quantization of the rate function can lead to a quantization of the optimal thresholds.

In terms of service provisioning, two distinct delay constraints were considered: strict delay and average delay. For each constraint a genie aided optimal solution is provided to benchmark the opportunistic and stochastic optimization strategies. Providing services within an average delay is analogous to a non-linear knapsack with an expanding-then-contracting constraint. A reformulation from non-linear to linear, at the cost of a variable space expansion, was presented to obtain a genie aided optimal solution. Practical strategies for service provisioning were presented in the form of optimal stopping and stochastic optimization, both of which adhere to our assumption of no a priori network information. The service provisioning strategies for average delay constraints are guaranteed to meet the average delay at each scheduling instance.

Our approaches for service provisioning allows the wireless component of an interconnected network to become amenable with network calculus [54, 55] or stochastic network calculus [57], if we view the wireless component of the network as a single element with the appropriate constraint. Quantifying the value of services remains beyond the scope of engineering, yet may very well guide how a network should handle
information as services.
Bibliography


