Abstract

Enhancing Communication Availability with Hybrid Networks

Mei Ya Chan
Master of Applied Science
Graduate Department of Electrical and Computer Engineering
University of Toronto
2015

Communication-Based Train Control (CBTC) systems have been adopted worldwide to maximize traffic capacity and enforce railway safety. In CBTC systems, wireless communication between trains and the control station is typically realized by homogeneous networks, such as 802.11 WLANs or GSM-R. As a safety-critical application, communication availability is of paramount importance in CBTC systems and is enhanced by component redundancy. We present a novel solution to enhance communication availability in CBTC systems by exploiting heterogeneous networks to provide alternative communication paths. We employ overlay networks over the heterogeneous networks. Experimental results show that overlay networks over LTE networks are an advisable alternative communication path. We design a joint horizontal and vertical handoff decision algorithm which utilizes historical measurements. The proposed algorithm demonstrates effectiveness at reducing number of handoffs, service interruption and low QoS, eliminating ping-pong effects, increasing handoff intervals, and maintaining robustness to partial network failure or degradation.
Acknowledgements

First of all, I would like to express my sincere thankfulness to my supervisor, Prof. Jörg Liebeherr, for his continuous support of my MASc studies. His great mentoring experience and expertise has provided me precious and insightful guidance in all the time of research and writing of this thesis.

Further, I also want to thank the project partner, Thales, for providing me the precious training opportunity, and all the past and present fellows at the Network Research Lab, Ashraf Al Daoud, Aditya Shankar, Alireza Shekaramiz, Boris Drazic, Chang Liu, Gagandeep Singh, HaiZhou Wang, Hossein Mortazavi, Hussein Al-Zubaidy, Jiyu Chen, Parsa Mirdehghan, and Yongbo Tang, for the stimulating discussions.

Last but not the least, I would like to thank my family: my parents, TsaiChin Chan, HsiangLan Lu, and to my sisters, TingTing Chan and YaYing Chan, and also my friends, KeYi Li, Sara Al-Kokhon, Zhen Dai, and QiuHua Tian, for supporting me spiritually throughout writing this thesis and my life in general.
## Contents

1 Introduction ................................. 1
   1.1 Thesis Contributions ...................... 5
   1.2 Thesis Organization ...................... 6

2 Related Work .............................. 7
   2.1 Horizontal Handoff Decision for 802.11 WLANs or Cellular Networks ............. 7
   2.2 Vertical Handoff Decision Between 802.11 WLANs and Cellular Networks .......... 9
      2.2.1 Criteria ......................................... 9
      2.2.2 Vertical Handoff Decision Algorithm ...................... 10
   2.3 Handoff Decision Incorporating Locations of PoAs and Location of MS .......... 16
   2.4 Horizontal Handoff Decision in CBTC Systems .................................. 17

3 ASON Over Hybrid Network ................. 18
   3.1 Communication Model and Network Environment .......................... 18
   3.2 HyperCast ........................................ 20
      3.2.1 Overlay Node .................................. 20
      3.2.2 Forwarding Engine ............................. 22
      3.2.3 CSA Processor .................................. 22
      3.2.4 Adapter, Interface, and Address Repository ...................... 22
   3.3 Switchover Using ASONs ................................ 22
      3.3.1 ASON Configuration ............................. 23
      3.3.2 ASON Operation .................................. 23
   3.4 Evaluation of ASON over LTE ............................. 28
      3.4.1 Switchover Delay Experiment ................................ 28
      3.4.2 Bandwidth Consumption Experiment .......................... 31

4 Design of Handoff Decision Algorithm ..... 34
   4.1 Data Collection Process .......................... 34
      4.1.1 Score Database .................................. 35
   4.2 Location-based Handoff Decision Algorithm with Adaptive Planning ............. 37
      4.2.1 Coverage Area and Coverage Profile .......................... 38
List of Tables

2.1 Normalization Methods .................................................. 12
2.2 Weighting Methods ......................................................... 13
3.1 Statistics of the switchover delay measurements of ASON over LTE .............. 31
5.1 Parameters for path loss .................................................... 46
5.2 Maximum segment size (m) under shadowing with $\sigma_s = 4$ dB .................. 49
5.3 Maximum segment size (m) under Rayleigh fading .................................... 50
5.4 Maximum segment size (m) under composite shadowing with $\sigma_s = 4$ dB and Rayleigh fading .................................................. 51
5.5 Minimum sample size per segment and minimum segment size for different $p$ and $\sigma_s$ under shadowing with $\sigma_s = 4$ dB ........................................ 54
5.6 Minimum sample size per segment and minimum segment size for different $p$ under Rayleigh fading .................................................. 54
5.7 Minimum sample size per segment and minimum segment size for different $p$ and $\sigma_s$ under composite shadowing with $\sigma_s = 4$ dB and Rayleigh fading ............... 55
6.1 Configuration parameters .................................................. 63
6.2 Comparison of the three algorithms using experimental data ......................... 72
6.3 Comparison of the three algorithms using the simulated data ......................... 75
List of Figures

1.1 Architecture of a CBTC system ............................................ 2
1.2 Architecture of the hybrid network ........................................ 3
2.1 RSS of two PoAs as functions of MS location .......................... 8
3.1 Communication paths in a CBTC system ................................. 19
3.2 Communication paths in a hybrid network ............................... 19
3.3 HyperCast Version 4 overlay socket ....................................... 21
3.4 The primary and secondary overlays for a CBTC system with \( n \) MSs and a central control station ........................................ 24
3.5 ASON operation modes ...................................................... 25
3.6 Setup of switchover delay experiment (ASON over LTE) .......... 28
3.7 Bell Canada’s 4G LTE Novatel Wireless U679 Turbo Stick ........ 29
3.8 Experiment testbed at the University of Toronto ..................... 29
3.9 Empirical distribution of switchover delay of ASON over LTE .... 31
3.10 Bandwidth consumed by an overlay socket to maintain the network topology in an ASON with two nodes for different heartbeat times ............. 33
4.1 Data collection process ..................................................... 36
4.2 The tabular structure of the score database ............................. 37
4.3 Location-based handoff decision algorithm with adaptive planning ........................................................................................................... 39
5.1 \( \epsilon \) and \( 1 - \alpha \) as functions of \( s \) under shadowing with \( \sigma_S = 4 \) dB ......................................................... 57
5.2 Using \( \epsilon \) and \( 1 - \alpha \) as functions of \( s \) to calculate the range of feasible values of \( s \) under shadowing with \( \sigma_S = 4 \) dB ......................................................... 58
5.3 Using \( \epsilon \) and \( 1 - \alpha \) as functions of \( s \) to calculate the minimum feasible value of \( \epsilon \) under shadowing with \( \sigma_S = 4 \) dB ......................................................... 58
5.4 Using \( \epsilon \) and \( 1 - \alpha \) as functions of \( s \) to calculate the maximum feasible value of \( 1 - \alpha \) under shadowing with \( \sigma_S = 4 \) dB ......................................................... 59
5.5 \( \epsilon \) and \( 1 - \alpha \) as functions of \( s \) under Rayleigh fading .................. 60
5.6 \( \epsilon \) and \( 1 - \alpha \) as functions of \( s \) under composite shadowing with \( \sigma_S = 4 \) dB and Rayleigh fading .................................................. 61
6.1 PoA deployment from an actual CBTC system ........................................... 65
6.2 Location of the MS as a function of time from the measurement trace ...... 65
6.3 Coverage areas of PoAs obtained from experimental measurement traces . 66
6.4 An example of simulated Received Signal Strength (RSS) and experimental RSS
   of PoA 24 .................................................................................................. 68
6.5 Coverage areas of PoAs obtained from simulated measurement traces ...... 69
6.6 Handoff locations and locations of the selected PoAs of the three algorithms using
   the experimental data .................................................................................. 70
6.7 Distribution of handoff intervals of the three algorithms using the experimental
data .............................................................................................................. 70
6.8 Distribution of service interruption durations of the three algorithms using the
   experimental data ......................................................................................... 71
6.9 Distribution of low QoS durations of the three algorithms using the experimental
data .............................................................................................................. 71
6.10 Handoff locations and locations of the selected PoAs of the three algorithms using
   the simulated data ......................................................................................... 73
6.11 Distribution of handoff intervals of the three algorithms using the simulated data ................................................................. 73
6.12 RSS of the selected PoA by the three algorithms using the simulated data . 74
6.13 Distribution of low QoS durations of the three algorithms using the simulated data ................................................................. 74
6.14 Number of handoffs as a function of \( k \) for partial network failure .......... 76
6.15 Average handoff interval as a function of \( k \) for partial network failure ...... 76
6.16 Maximum service interruption duration (top) and number of service interruption
   incidents (bottom) as a function of \( k \) for partial network failure ............... 77
6.17 Maximum low QoS duration (top) and number of low QoS incidents (bottom)
   as a function of \( k \) for partial network failure ............................................ 78
6.18 Number of handoffs as a function of \( k \) for partial network degradation ...... 79
6.19 Average handoff interval as a function of \( k \) for partial network degradation .. 80
6.20 Maximum low QoS duration (top) and number of low QoS incidents (bottom) as
   a function of \( k \) for partial network degradation ............................................ 81
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project</td>
</tr>
<tr>
<td>AHP</td>
<td>Analytic Hierarchy Process</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>ASON</td>
<td>Application-layer Service Overlay Network</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>BPNN</td>
<td>Back-Propagation Neural Network</td>
</tr>
<tr>
<td>CBTC</td>
<td>Communication-Based Train Control</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>CSA</td>
<td>Cross-Substrate Advertisement</td>
</tr>
<tr>
<td>DSRC-WAVE</td>
<td>Dedicated Short Range Communications and Wireless Access in Vehicular Environment</td>
</tr>
<tr>
<td>ENN</td>
<td>Elman Neural Network</td>
</tr>
<tr>
<td>FL</td>
<td>Fuzzy Logic</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>GNSS</td>
<td>Global Navigation Satellite System</td>
</tr>
<tr>
<td>GRA</td>
<td>Grey Relational Analysis</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile</td>
</tr>
<tr>
<td>ICDF</td>
<td>Inverse Cumulative Distribution Function</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>LTE</td>
<td>Long Term Evolution</td>
</tr>
<tr>
<td>MADM</td>
<td>Multiple Attribute Decision Making</td>
</tr>
<tr>
<td>MANET</td>
<td>Mobile Ad Hoc Network</td>
</tr>
<tr>
<td>MDP</td>
<td>Markov Decision Process</td>
</tr>
<tr>
<td>MEW</td>
<td>Multiplicative Exponent Weighting</td>
</tr>
<tr>
<td>MS</td>
<td>Mobile Station</td>
</tr>
<tr>
<td>NAT</td>
<td>Network Address Translation</td>
</tr>
<tr>
<td>OFDM</td>
<td>Orthogonal Frequency Division Multiplex</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>PoA</td>
<td>Point of Attachment</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RSS</td>
<td>Received Signal Strength</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>SAW</td>
<td>Simple Additive Weighting</td>
</tr>
<tr>
<td>SCTP</td>
<td>Stream Control Transmission Protocol</td>
</tr>
<tr>
<td>SINR</td>
<td>Signal-to-Interferences-plus-Noise Ratio</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
<tr>
<td>TOPSIS</td>
<td>Technique for Order Preference by Similarity to Ideal Solution</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunications System</td>
</tr>
<tr>
<td>WLAN</td>
<td>Wireless Local Area Network</td>
</tr>
</tbody>
</table>
Nomenclature

1 – \( \alpha \)  Statistical confidence level

\( S_{i,j}^{\phi} \)  Random variable denoting the score of a Point of Attachment (PoA) \( \phi \) at the \( i \)th spatial section and the \( j \)th temporal section

\( T_c \)  Threshold that determines whether a spatial section is covered by a PoA or coverage threshold

\( \lambda_n \)  Number of samples of each PoA per meter or sample density

\( \phi \)  Label of a PoA

\( \sigma_S \)  Standard deviation of shadowing in dB

\( \{ \phi^d \} \)  List of detected PoAs

\( \{ s_{\phi^d} \} \)  List of scores of respective PoAs in \( \{ \phi^d \} \)

\( k_u \)  Number of consecutive samples to identify an unpredictable event

\( p_c \)  Probability that a spatial section is covered by a PoA or coverage probability

\( p_{u_0} \)  Per-sample unpredictable event probability

\( s_{i,j}^{\phi} \)  Measurement of the score of a PoA \( \phi \) at the \( i \)th spatial section and the \( j \)th temporal section

\( s \)  Size of a spatial section or segment size in meters
Chapter 1

Introduction

The global growth of cities and the increase of population density have intensified the demand for railway transit systems, such as tram, light rail, subway, metro, and so forth. To maximize the efficiency of a transit system, trains are scheduled on a track in a pipelined manner while a safe distance between any two consecutive trains is preserved to ensure security. Safety distances are enforced by assigning each train a movement authority which is the permission to move through a designated section or block of track with supervised speed. Movement authorities are computed and updated based on the locations and speeds of all trains in the system at any time. Hence, it is pivotal that trains are capable to transfer these two pieces of information to a regional or central control station in exchange for a movement authority within a stringent time frame. Moreover, to guarantee safe train operations, especially under emergency situations, it is essential that train attendants are able to maintain communications with the ground personnel and the control station at all times to override automatic train control decisions. Collectively, these two indispensable functions have driven the incorporation of radio communication technology as an essential component in train control systems, referred to as Communication-Based Train Control (CBTC) systems.

The components of a CBTC system can be simplified into a central control station, a number of trains, referred to as Mobile Stations (MS), and a communication network, referred to as the infrastructure network. The infrastructure network connects the central control station and the MSs. The infrastructure network contains a wired network that connects the central control station to a number of access points. These access points have been deployed along the track, and they are referred to as the wayside access points. Each MS equips wireless radios to associate with a wayside access point in order to connect to the central control station. As an example, Figure 1.1 shows a central control station, an MS, and an infrastructure network in a CBTC system. Since train control messages are highly time-critical, communications in CBTC systems are subject to stringent latency and availability requirements to assure railway safety. In terms of latency, IEEE 1474 [1] establishes the performance and functional requirements of a CBTC system that the typical bidirectional communication delay between train and wayside should range from 0.5 to 2 seconds. In terms of availability, hardware redundancy is widely
Figure 1.1: Architecture of a CBTC system

adopted where each MS equips redundant wireless radios, one at the front and one at the rear of the train, and excessive access points are deployed.

If consecutive train control messages are lost or the delay of the messages is beyond the requirements, numerous trains in travel, or most likely all trains in the system, are forced to halt to ensure safe operation. For example, if the disconnection lasts for $l$ minutes, and supposing that the entire route is $d$ km long and comprises $n$ stations, trains are separated by a headway, which is the separation between two trains measured in time, of $m$ minutes with an average speed of $v$ km per minute, and each train stays at a station for $x$ minutes, a total of $\frac{nx+d}{vm}$ trains are forced to stop for a delay of $l$ minutes until the communication link recovers. Furthermore, if the delay is prolonged beyond the departure of prospective trains that are scheduled during the $l$ minutes, an additional $\lfloor \frac{l}{m} \rfloor$ trains are also postponed. Consequently, a total of $\frac{nx+d}{vm} + \lfloor \frac{l}{m} \rfloor$ trains are affected by an $l$-minute communication delay incident between the central control station and an MS. Taking the Yonge-University Line of Toronto rapid transit as an example, where $d$ is 30.2 km, $n$ is 32, $m$ is 2.5 minutes during peak hours, $v$ is roughly 0.67 km per minute, and $x$ is nominally 0.5 minutes, a communication delay incident involving a single MS influences a number of $\frac{nx+d}{vm} + \lfloor \frac{l}{m} \rfloor \geq 24.4$ MSs in total. To avoid such a decrease in the transit efficiency, a CBTC system seeks to maximize the availability of communication in the train-to-wayside communication system.

In this thesis, we explore methods that improve the overall communication availability of a CBTC system by considering alternative communication paths provided by heterogeneous networks which are networks that use different access technologies. We refer to an infrastructure network augmented by one or more heterogeneous network as a hybrid network where the
augmenting heterogeneous networks are referred to as the backup networks. Examples of backup networks include 3rd Generation Partnership Project (3GPP) cellular networks, such as 2G Global System for Mobile (GSM), 3G Universal Mobile Telecommunications System (UMTS) and 4G Long Term Evolution (LTE), that have been deployed and are operated by third-party providers and Mobile Ad Hoc Networks (MANETs) formed by groups of MSs. Figure 1.2 shows a hybrid network consisted of an infrastructure network, a cellular network, and a MANET. The cellular network connects an MS and the central control station over the Internet. The MANET is consisted of two MSs forming an ad hoc link. The MANET allows an MS that cannot directly connect to a wayside access point or a base station in a cellular network to obtain access to the central control station through other MSs which act as relays. To utilize the aforementioned heterogeneous networks, an MS needs to equip multiple network adapters, an 802.11-based adapter and a cellular adapter. Furthermore, the 802.11-based adapter needs to be able to support both the infrastructure mode and the ad-hoc mode. For generalization, we use the generic term Point of Attachment (PoA) to refer to a node that provides access to a network. A PoA can be an access point in an 802.11 Wireless Local Area Network (WLAN), a base station in a cellular network, or an MS in a MANET. The proposed hybrid network solution is a novel approach to enhance communication availability as communication between train and wayside in state-of-the-art CBTC systems is carried out by homogeneous networks.

We explore two questions in this thesis. The first question is, can a hybrid network be run as Application-layer Service Overlay Networks (ASON) to satisfy the Quality of Service (QoS) requirements of time-critical applications? An ASON is an overlay network running at the
application layer. An overlay network is a virtual network built on top of one or more existing substrate networks [2]. A substrate network is a group of nodes that adopt a compatible or common addressing and routing scheme. Substrate networks can refer to networks at different layers, such as the data link layer, the network layer, or the application layer. The nodes in a substrate network are referred to as physical nodes. The nodes in an overlay network are referred to as logical nodes. A link between two logical nodes, called a logical link, can be comprised of a number of physical links in one or more substrate networks. Logical nodes are able to exchange information periodically via physical links in the substrate networks to form and maintain a network topology based on a specific protocol in a self-organizing manner. We adopt ASON to organize the hybrid network since ASON provides a higher level of abstraction which simplifies the engineering of heterogeneous networks. For example, the problem of a node changing the attachment from one substrate network to another can be realized by switching between overlays, referred to as a switchover.

The second question is to determine which PoA to associate with to assure the received QoS meets the given constraints. The process of transferring the association and ongoing sessions to another PoA is called a handoff, or handover, and the selection of the PoA is called a handoff decision. Handoffs are commonly categorized as horizontal or vertical, where the two PoAs involved in a horizontal handoff belong to a homogeneous network, whereas the two PoAs involved in a vertical handoff belong to two heterogeneous networks. Horizontal handoffs and vertical handoffs are widely regarded as distinct problems. Horizontal handoff decisions typically consider a small set of criteria which are mainly related to signal strength [3, 4]. On the other hand, vertical handoff decisions consider multiple criteria from a wide range of metrics, for instance, signal strength, bandwidth, delay, mobility, user preference, and application-specific QoS requirements [5, 6, 7, 8, 9]. In our vision, we do not differentiate PoAs by their access technologies, that is, we do not make a distinction between horizontal handoff and vertical handoffs in designing the handoff decision algorithm. Rather, we treat PoAs by the QoS they provide and design a joint horizontal and vertical handoff decision algorithm. The handoff decisions generated by the algorithm govern the behaviours of the overlay networks as a vertical handoff corresponds to a change in the attachment to the substrate networks and requires a switchover to be performed.

Among various vertical handoff decision algorithms in the literature, a group of studies has attempted to incorporate locations of PoAs in handoff decision algorithm and achieved a reduction in handoff latency [10, 11]. In the studies on handoff decisions specifically tailored for CBTC systems [12, 13, 14], all of them consider a homogeneous network environment, and only one of them attempts to utilize location information about access points and MSs [12]. Therefore, research efforts on vertical handoff decision for CBTC system with incorporation of location of PoAs and MSs are still demanding.

With the goal to enhance communication availability, we outline the design objectives of the handoff decision algorithm: minimizing service interruption durations, defined as the time
duration when the selected PoA is out of service due to reasons such as severely weak signal or strong interference; minimizing low QoS durations, defined as the time duration when the selected PoA fails to provide the QoS requirements; eliminating ping-pong effects, that are the phenomena when a handoff decision algorithm oscillates the selection among a set of PoAs; minimizing the number of handoffs, and increasing handoff intervals, defined as the distance travelled by the MS between two consecutive handoffs. In addition, the handoff decision algorithm needs to be reliable in the sense that it is resilient to partial partial network failure or degradation.

We observe that MSs in CBTC systems have predictable motion paths (train lines). Under the assumption that all PoAs and all MSs have uniform hardware (radios), we conjecture that MSs of a train line that depart in a certain time range should experience similar QoS at the same location if they associate with the same PoA. The assumptions allow making use of historical measurements to predict real-time QoS and plan out a sequence of handoffs in advance that fulfils the aforementioned objectives.

1.1 Thesis Contributions

This thesis investigates using a hybrid network to enhance communication availability for time-critical applications, specifically a train control system. The thesis consists of two parts. In the first part, we adopt ASONs to facilitate the switchover process. Specifically, we organize and construct ASONs on top of the hybrid network. We propose several different operation schemes to achieve the switchover process. Next, we examine the feasibility of an ASON over an LTE network to support the delay and bandwidth requirements of the time-critical train control application, which has not been evaluated before. In the second part, we propose a location-based joint horizontal and vertical handoff decision algorithm with adaptive planning for CBTC systems. The algorithm exploits predictable motion paths and utilizes historical measurements. The algorithm requires a data collection process that surveys the QoS of all PoAs in the system at different times and locations. The historical data is used to determine the coverage area of each PoA. An overlapping coverage graph is derived from the coverage areas of PoAs where a PoA is represented by a node and PoAs with overlapping coverage areas are connected by a link. A sequence of handoffs is planned based on the graph in advance. An MS follows the sequence and executes a handoff when arriving at the next PoA. On the fly, an MS compares the historical and real-time QoS. If the historical and real-time QoS are inconsistent, which is referred to as an unpredictable event, the sequence of handoffs is recomputed. In this way, the algorithm is resilient to network failure or degradation. We evaluate the performance and resilience of the proposed algorithm with experimental and simulated data.
1.2 Thesis Organization

This thesis is structured as follows. Chapter 2 presents a survey of horizontal and vertical handoff decision algorithms for wireless networks. Chapter 3 illustrates the architecture of ASONs over a hybrid network and investigates the feasibility of using an ASON over a backup network (an LTE network) to support the time-critical train control application. Chapter 4 presents the design of the proposed location-based handoff decision algorithm with adaptive planning. Chapter 5 discusses the design issues and provides justifications. Chapter 6 presents the performance evaluation of the proposed scheme with comparisons to two existing schemes. Finally, Chapter 7 concludes the thesis and provides directions for future work.
Chapter 2

Related Work

In this chapter we review related studies on horizontal and vertical handover decision algorithms for 802.11 WLANs and cellular networks where we highlight a group of algorithms that utilizes the locations of access points and the location of the MS as well as horizontal handoff decision schemes tailored for CBTC systems. The distinction between handoff for 802.11 WLANs or cellular networks and handoff for CBTC systems is that the MSs considered in the context of 802.11 WLANs or cellular networks are personal mobile devices and they have random motion paths (unpredictable motion), whereas the MSs in the context of CBTC systems are trains and they have fixed motion paths (predictable motion). Horizontal handoff and vertical handoff share similar handoff procedure, which typically consists of three phases: handoff criterion/criteria measurement, decision, and execution.

2.1 Horizontal Handoff Decision for 802.11 WLANs or Cellular Networks

The most prevalent handoff criterion in horizontal handoff in 802.11 WLANs and in cellular networks [3, 4, 15] is RSS, which is a measure of the power of a received signal originating at a transmitter, usually measured in Decibel-milliwatts (dBm). As wireless channels are subject to rapid fluctuations by various types of fading, RSS is usually averaged over time. Consider Figure 2.1 which depicts the RSS of two PoAs as an MS moves from PoA 1 toward PoA 2. The MS was originally connected to PoA 1, and in order to receive a satisfactory QoS, the MS needs to perform a handoff from PoA 1 to PoA 2 somewhere in the middle. Various schemes to determine the handoff time/location are summarized in [3], and these schemes are also employed in practice:

- Relative RSS: The MS chooses the PoA with the strongest average RSS measurement. This approach is prone to introducing unnecessary handoffs, as RSS measurements suffer drastic fluctuations, and the PoA with the strongest RSS may change rapidly in overlapping regions.
• Relative RSS with threshold: The MS hands over to another PoA if the RSS of that PoA is stronger than the RSS of the current PoA, and the RSS of the current PoA is below a threshold. As the threshold value affects the point of handover, a proper choice of the threshold needs to be made while the effectiveness of the threshold requires prior knowledge of the crossing point of the RSS values of the competitive PoAs.

• Relative RSS with hysteresis: The MS chooses the PoA with an RSS higher than the RSS of the current PoA by a hysteresis margin to prevent a ping-pong effect. This approach may introduce early handovers where the RSS of the original PoA is still sufficiently high.

• Relative RSS with hysteresis and threshold: The MS hands over to the new PoA when the RSS of the old PoA is lower than the threshold, and the RSS of the new PoA is stronger than the RSS of the old PoA by a hysteresis margin to prevent a ping-pong effect and premature handover.

• A dwell timer can be incorporated into any of the approaches above to prevent unnecessary handover operations. The dwell timer activates a handover execution when it expires if a certain condition is satisfied and endures until its expiration. The dwell timer is started once the condition used in the approach is satisfied. For example, if the relative RSS approach is used, the condition is that the RSS of a new PoA becomes stronger than the RSS of the current PoA. If the condition remains true when the dwell timer expires, a handover to the new PoA will be executed. If the condition does not hold during the duration of the dwell timer, no handover will take place. The duration of the dwell timer can be fixed or adaptive.
2.2 Vertical Handoff Decision Between 802.11 WLANs and Cellular Networks

Unlike a horizontal handover decision which usually considers one or a small number of criteria, a vertical handoff decision typically takes multiple criteria into account. We first introduce different types of criteria and review the strategies to combine multiple criteria to produce a handoff decision.

2.2.1 Criteria

Many criteria are correlated and have different levels of importance based on application. Handoff decision criteria can be classified into four categories: network-related, terminal-related, service-related and user-related [6, 16].

2.2.1.1 Network-related

Network-related criteria are indicators of the network conditions.

- **Network coverage**: Different wireless technologies have different ranges of coverage. For example, UMTS has a wide area coverage, while a WLAN generally has a short range of coverage. Hence, if an MS is crossing through a WLAN coverage and the MS moves at a high speed, then it may not be necessary to hand off to the WLAN.

- **RSS**: RSS indicates the received signal strength, measured in dBm, dB, mW or W, and therefore a large RSS value is preferred. RSS is a widely adopted criterion as it has a direct impact to bandwidth and Bit Error Rate (BER) and can be readily acquired.

- **Bandwidth**: Bandwidth is the rate of data transfer provided by the network. Bandwidth is a dominant criterion for applications that requires a high data transfer rate.

- **Latency**: Latency can be a dominant criterion for application- or service-specific QoS requirements. For example, a network with high latency fails to support conversational services.

- **Signal-to-Noise Ratio (SNR)**: SNR measures the ratio of signal strength to background noise in dB. If the SNR is too low, the signal cannot be decoded correctly. A similar measure that includes interferences is the Signal-to-Interferences-plus-Noise Ratio (SINR).

- **BER**: By comparing the transmitted sequence and the received sequence, BER is calculated by taking the ratio of the number of erroneous bits in the received sequence to the total number of bits received. BER can be a dominant criterion for specific services requiring high data integrity.
• Load: Network load reflects the spare capacity of the network. Network load can be measured as a ratio of the current traffic to the maximum traffic that can be carried by the network. A network with a high load may block new calls and degrade existing calls.

• Security: Certain types of applications, such as financial applications, require secure networks to ensure confidentiality and integrity. Unencrypted wireless networks, or public wireless networks, may be untrustworthy and malicious and make sensitive data vulnerable to exposure.

2.2.1.2 Terminal-related

This group concerns the characteristics and current status of MSs.

• Mobility: The mobility of an MS, including location and velocity, can affect the handoff decision. For example, a fast moving MS may not benefit from numerous handovers between WLANs and cellular network due to the large amount of overheads and delays and may benefit more by staying in the cellular network.

• Power: As the power of MS is constrained, connecting to a network that consumes less MS power can be important when the battery is low.

2.2.1.3 Service-related

Different service types impose different combinations of requirements of latency, bandwidth, BER, and so forth. For example, voice or conversational services require low packet delay and jitter, streaming services demand low jitter and high bandwidth, while interactive services require low BER and moderately low delay.

2.2.1.4 User-related

User-related criteria reflect users’ preferences or satisfaction.

• User preferences: Users may have preferences to connect to certain types of wireless networks for different QoS.

• Monetary cost: Wireless network technologies are charged differently and therefore affecting users’ choices.

2.2.2 Vertical Handoff Decision Algorithm

In this section, we present a representative set of vertical handoff decision algorithms and classify them by strategy.
### 2.2.2.1 Cost Function Strategy

An early vertical handoff decision algorithm introducing a cost function strategy appeared in [17]. The criteria considered are the bandwidth $B_n$, power consumption $P_n$, and monetary cost $C_n$ provided or consumed by using network $n$. The cost function is formulated as

$$f_n = w_b N(1/B_n) + w_p N(P_n) + w_c N(C_n),$$

where $N(x)$ is a normalization function and $w_b$, $w_p$, and $w_c$ are the weights of the criteria. An example of a normalization function is the logarithm. A lower cost indicates a better network.

A network needs to be better than the current network for the duration of a stability period, $T_s$, before a handoff occurs to ensure the handoff is worthwhile. The stability period is the sum of the handover latency and the time to make up the loss (cost, bandwidth or other factors depending on the considered criteria) incurred by the handover procedure. Consider a simple case where bandwidth, denoted by $B_{\text{target}}$ and $B_{\text{current}}$, is the only criterion, then $T_{\text{makeup}}$ can be derived in terms of $T_{\text{handoff}}$ by

$$(B_{\text{target}} - B_{\text{current}})T_{\text{makeup}} = B_{\text{current}}T_{\text{handoff}}.$$ 

Let $B = \frac{B_{\text{target}}}{B_{\text{current}}}$ and assume $B > 1$, the stability period becomes

$$T_s = T_{\text{handoff}} + T_{\text{makeup}} = T_{\text{handoff}} + \frac{T_{\text{handoff}}}{B - 1}.$$ 

Zhu and McNair generalize the cost function for multi-service networks in [18, 19] to be

$$C_n = \sum_s E^n_s \sum_j W^n_{s,j} Q^n_{s,j},$$

where $C^n$ is the cost of network $n$; $E^n_s$ is an elimination factor, which is $\infty$ if network $n$ fails to satisfy constraints for service $s$, or 1 if the constraints are satisfied; and $Q^n_{s,j}$ is the normalized criterion value provided by network $n$ for criterion $j$ and service $s$. The authors demonstrate that a cost function strategy prevails over the RSS strategy by providing higher throughput without increasing the blocking rate.

### 2.2.2.2 Multiple Attribute Decision Strategy

Multiple Attribute Decision Making (MADM) is a class of methods that deal with problems involving multiple variables in operations research. Decision methods generally require three components: alternatives, attributes, and weights of the attributes. An alternative is decided by attributes and weights of the attributes. Attributes can be categorized into beneficial attributes and cost attributes. Beneficial attributes are attributes that favour large values. Cost attributes are attributes that prefer small values. Attributes are normalized to be comparable.
Because of their distinct characteristic, beneficial and cost attributes have different normalization functions. Common normalization methods are listed in Table 2.1 [20, 21]. The weight or importance of each attribute can be objectively or subjectively determined. Table 2.2 lists different weighting methods [20, 21, 22, 23]. The entropy weighting method was developed by Zeleny (1982) from the entropy concept proposed by Shannon and Weaver (1947). In the entropy weighting method, the weight assigned to each attribute corresponds to the uncertainty associated with that attribute. The standard deviation method assigns weight based on the standard deviation of each attribute, and it is used in [24]. Analytic Hierarchy Process (AHP) was developed by Saaty [25]. AHP provides a comprehensive framework to structure a decision problem. AHP has been adopted in the vertical handoff decision studies as a method to generate the weights for criteria. User defined methods collect user preferences of each attribute. Service type and service subscription methods were introduced in [23] where the authors define fixed weights based on the service type or the service subscription. MADM methods score and rank alternatives based on the weights and the attributes. In the context of vertical handoff decision, the alternatives are the candidate networks, and the attributes are the criteria. MADM methods that have been applied in vertical handoff decision are Simple Additive Weighting (SAW) [26], Multiplicative Exponent Weighting (MEW) [9], AHP [27], Grey Relational Analysis (GRA) [27], Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [23], ELECTRE [28] and VIKOR [29]. Comparisons of different MADM techniques have been conducted in [9] and [30].

Stevens-Navarro et al. conduct a comparison of GRA, MEW, SAW and TOPSIS using weights determined by AHP described above [9]. The chosen criteria are BER, delay, jitter and bandwidth. The first result was on the performance of bandwidth, where the four algorithms

<table>
<thead>
<tr>
<th>Normalization function for Beneficial Attributes</th>
<th>Normalization function for Cost Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear $x_{ij}^{\text{normal}} = \frac{x_{ij}}{\max_i x_{ij}}$</td>
<td>Linear $x_{ij}^{\text{normal}} = \frac{\min_i x_{ij}}{x_{ij}}$</td>
</tr>
<tr>
<td>Max-Min $x_{ij}^{\text{normal}} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}$</td>
<td>Max-Min $x_{ij}^{\text{normal}} = \frac{\max_i x_{ij} - x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}$</td>
</tr>
<tr>
<td>Square root $x_{ij}^{\text{normal}} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2}}$</td>
<td>same</td>
</tr>
<tr>
<td>Sum $x_{ij}^{\text{normal}} = \frac{x_{ij}}{\sum_i x_{ij}}$</td>
<td>Sum $x_{ij}^{\text{normal}} = \frac{1}{\sum_i (1/x_{ij})}$</td>
</tr>
</tbody>
</table>
Table 2.2: Weighting Methods

<table>
<thead>
<tr>
<th>Category</th>
<th>Weight Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>$w_j = \frac{d_j}{\sum_k d_k}$, $d_j = 1 - \frac{1}{\ln N} \sum_i y_{ij} \ln(y_{ij})$, $y_{ij} = (x_{ij})_{normal}$ with square root method</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$w_j = \frac{d_j}{\sum_k d_k}$, $d_j = \sqrt{\frac{\sum_i (x_{ij} - \bar{x}<em>j)^2}{N \bar{x}<em>j}}$, $\bar{x}<em>j = \frac{1}{N} \sum_i y</em>{ij}$, $y</em>{ij} = (x</em>{ij})_{normal}$ with square root method</td>
</tr>
<tr>
<td>AHP</td>
<td>Subjective [25]</td>
</tr>
<tr>
<td>User Defined</td>
<td>Subjective Obtained from user preference</td>
</tr>
<tr>
<td>Service Type</td>
<td>Subjective Self-defined weights based on the service type, e.g. voice, streaming, and web</td>
</tr>
<tr>
<td>Service Subscription</td>
<td>Subjective Obtained from network operator or self-defined based on the subscription, e.g. gold, silver, and bronze</td>
</tr>
</tbody>
</table>
performed similarly for voice and streaming traffic classes, but GRA outperformed the other three for interactive and background traffic classes. The second evaluation was on the performance of delay, where the four algorithms performed similarly for conversational and streaming traffic classes, but GRA outperformed the other three for interactive and background traffic classes by having lower delay. A few years later, Stevens-Navarro et al. expand the comparison to include the ELECTRE and VIKOR methods in [30]. The criteria considered are cost, available bandwidth, total bandwidth, packet delay and packet jitter. However, unlike [9], the results do not reveal that any of the methods outperforms the others. In fact, GRA provided the worst performance for packet delay and jitter for voice traffic class among all methods. The inconsistency of the results may be caused by a different assignment of weights.

2.2.2.3 Fuzzy Logic Strategy

Fuzzy Logic (FL) is a robust mathematical framework to deal with uncertainties with the capability to handle dynamically changing criteria such as RSS, BER, and so on. In [31], Hou and O'Brien propose a vertical handoff decision algorithm with FL for the case of two collocated heterogeneous wireless networks. They first develop two basic handoff schemes: an immediate handoff and a dwell handoff. An immediate handoff is preferred when there is a long interruption (duration without connectivity), while a dwell handoff is preferred if there is a short interruption. They found that the combination of two basic schemes, if triggered at the correct time, can result in better average transmission latency than the two schemes alone. Hence, they employ FL to make the proper vertical handoff decision scheme selection according to the network conditions. Comparisons of the average transfer delay of the two basic schemes and the FL scheme in all cases (different interruption mode, handover failure probability, message length and handoff execution delay) show that the FL scheme combines the merits of the two basic schemes and always inclines to the better of the two. Moreover, FL is generally incorporated with MADM methods (fuzzy MADM) and neural networks (neuro-fuzzy), which will be discussed in the following sections.

2.2.2.4 Neural Network Strategy

A neural network solution has been adopted in vertical handoff decision for (1) providing a simple representation for a physical implementation, and (2) producing accurate results for new inputs that the neural network has not seen during the training process [22]. In [22], Nasser et al. used a Back-Propagation Neural Network (BPNN) with an input layer, a hidden layer and an output layer. The hidden layer and the output layer employed a sigmoid activation function. The weight of each connection and the bias at each neuron is randomly generated at the beginning. The inputs to the BPNN are the (user-defined) weights of the four criteria. This is very different from other approaches as the network environment is assumed to be static so there is no need to collect real-time measurements for the criteria. The BPNN was trained with samples before use. Each sample defines the four weights and the network to be selected.
The four weights are fed into the BPNN and the BPNN outputs the network that it deems the most suitable. The network selected by the BPNN is compared with the network to be selected defined in the sample to determine if the network is correctly selected. The authors report the average successful network selection rate to be 87%.

2.2.2.5 Fuzzy MADM Strategy

The combination of FL and MADM is used in two approaches. In the first approach, when criteria consist of variables that are expressed in human language, called linguistic variables (for example, user preference), FL is used as a normalization method that converts the linguistic variables into numerical values which are inputs of the MADM calculation. The second approach adopts FL as a handover initiation mechanism which decides the necessity of entering the MADM phase.

The first approach arose as MADM methods do not work efficiently with imprecise data and user preferences. FL can take the role to convert linguistic terms into numeric values. Hence, formulating the vertical handoff decision problem into a fuzzy MADM gained some popularity [32, 24]. Zhang illustrates a thorough tutorial on formulating and solving a fuzzy MADM in [32]. The approach consists of two main steps: (1) defuzzify fuzzy data and (2) apply the MADM method. Liao et al. adopt this approach and further restrict the execution of a handover decision if the target network is better than the second by a margin to reduce unnecessary handovers in [24].

In the second approach, an FL-based handover initiation phase is inserted before the handover decision phase. Without the handover initiation phase, each MT routinely repeats the vertical handoff decision process which periodically updates the decision even if the current network is satisfactory, which may introduce unnecessary handovers. The handover initiation phase evaluates if there is a need for a handover. In the handover initiation phase, criteria, such as RSS, bandwidth, network coverage, and velocity, are inputs to the FL procedure. The inputs pass through fuzzification, inference and defuzzification. The output of the defuzzification is compared with a predefined threshold to determine whether a handover should be initiated. If so, the handover procedure advance to the decision phase. This approach is adopted by [33, 34, 35, 36, 37].

2.2.2.6 Neuro-fuzzy Strategy

The combination of fuzzy logic and neural network is termed neuro-fuzzy in the field of artificial intelligence. In this direction, FL takes the role of the decision maker with the aid of a neural network. Guo et al. proposed an adaptive multi-criteria vertical handoff (AMVHO) [38] where a modified Elman Neural Network (ENN) is responsible to predict the number of users of the target network, which is a pivotal input to the FL based decision procedure. An ENN is a recurrent network that can maintain state information by inserting “context units” to store memory, which features a good dynamic advantages in prediction. The AMVHO method is compared to
the traditional RSS policy where the AMVHO method outperforms the traditional RSS policy in terms of lower BER, higher SNR, lower number of dropped packets and retransmissions.

2.2.2.7 Markov Decision Process Strategy

A Markov Decision Process (MDP) model consists of five elements: decision epochs, states, actions, transition probabilities, and rewards. An MS is assumed to make a decision to either stay in the current network or handover to another network at every decision epoch based on its current state. The state includes information such as the network identifier that the MS is currently connected to, the available bandwidth, the average latency and the jitter information of all available networks. After an action is chosen, a immediate reward that reflects the QoS provided by the chosen network during the epoch is assigned. The transition probability to the next state depends on the current state and the action. An MDP finds the optimal policy that determines the sequence of actions based on the current state such that the expected total reward is maximized. Stevens-Navarro et al. formulate the vertical handoff decision problem into an MDP and compare it with SAW, TOPSIS, GRA, bandwidth-centric and never-handoff strategies. They reported that the MDP approach always has the maximum expected total reward and fewest handovers.

2.3 Handoff Decision Incorporating Locations of PoAs and Location of MS

Different approaches that incorporates locations of PoAs and location of the MS into handoff decision have shown improved handoff performances [10, 11].

Shin et al. [10] present a handoff decision algorithm to expedite the latency of 802.11 handoffs using neighbour graphs. An 802.11 handoff procedure consists of a discovery or probing, an authentication and a reassociation stage. The probe stage is the dominating factor in handoff latency, accounting for more than 90% of the overall latency [39]. During the discovery stage, an 802.11 device repeatedly sends probe requests and waits for probe responses from PoAs in the proximity for all channels (a total of 11 channels at 2.4 GHz in North America). The discovery stage can be significantly accelerated by omitting channels that are not used by any of the surrounding PoAs. The authors curtail the number of channels to be scanned by collecting information of the (relative) locations and the operating channels of the PoAs in advance. The authors collect association patterns, where an association pattern is a sequence of PoAs (and their operating channels) associated along the movement of an MS, over all MSs. The relative locations of PoAs are obtained from the association patterns and are represented in a neighbour graph. Nodes in the neighbour graph are PoAs in the area of interest. An edge is inserted between two nodes if the two PoAs are successive in an association pattern. A neighbour graph is stored at the MSs. An MS can obtain the set of PoAs in the neighbourhood (and their operating channels) and skip the idle channels. The authors show that their algorithm is able
to reduce the probing latency by 80.7% compared to the full-scanning method over 250 handoffs in an indoor environment.

Tseng et al. [11] illustrate a location-based fast handoff scheme for 802.11 WLANs with the assistance of a location server. The location server possess the information of location, authentication, and reassociation of all PoAs in the system. The handoff latency is shortened by omitting the discovery stage in a handoff procedure. An MS requests the location server for a list of PoAs (and the respective authentication and reassociation information to complete a handoff procedure) resided in the direction toward which the MS will roam imminently. The MS predicts future movements by using the most recent position and displacement vector.

### 2.4 Horizontal Handoff Decision in CBTC Systems

In this section, we discuss handoff decision algorithms that are specific to CBTC systems where the radio communication network consists of homogeneous 802.11 WLANs [12, 13, 14].

Jiang et al. [12] propose a location-based handoff algorithm that aims at reducing handoff latency and ping-pong effects by exploiting the fixed motion paths and predictable handoff patterns of MSs. The authors first identify that general handoff algorithms for WLANs are not suitable for a CBTC system, since (1) ping-pong effects are more likely due to the fast changing wireless channel between an MS and a wayside PoA, especially in a tunnel environment, and (2) as MSs travel at faster speeds (compared to human) and hand off more frequently, delay of the discovery stage in a handoff procedure becomes a more severe problem that undermines the effective communication time. The authors recognize fixed motion paths and predictable handoff patterns in a CBTC system and propose predetermination of handoff locations. MSs monitor the states of the surrounding PoAs by periodically sending probe requests while maintaining ongoing connections, and only execute a handoff when the MSs have arrived at the predetermined handoff locations. Compared to an RSS-based scheme, the authors demonstrate their method can effectively eliminate ping-pong effects in a simulated tunnel environment under the case that all PoAs in the system are functioning correctly.

Zhu et al. [13] propose employing 802.11p, Dedicated Short Range Communications and Wireless Access in Vehicular Environment (DSRC-WAVE), and Stream Control Transmission Protocol (SCTP) in substitution of 802.11a/b/g and User Datagram Protocol (UDP). Also, they apply the MDP framework for the handoff decision with the objective to optimize the composite performance of packet delay and throughput. In [14], the authors incorporate control theory and formulate an MDP problem with the objective to minimize the (linear quadratic) error between the actual trajectory (location and velocity) of an MS and the trajectory guided by the train controller.
Chapter 3

ASON Over Hybrid Network

In this chapter we first provide an overview of the communication model between an MS and the central control station. Then we illustrate how a hybrid network, consisting of an infrastructure network and one or more backup networks, can be utilized to enhance communication availability. We adopt ASONs to facilitate the switchover between the heterogeneous networks. We first present an introduction of an ASON software system. Then we demonstrate how we organize and configure the ASONs and propose several operation schemes. Finally, we evaluate the feasibility of using an ASON over an LTE network to support the delay and bandwidth requirements of the time-critical train control application.

3.1 Communication Model and Network Environment

We consider a CBTC system with an 802.11-based infrastructure network. The communication paths between an MS and the central control station in a CBTC system are illustrated in Figure 3.1. To increase communication reliability, train control messages between an MS and the central control station are duplicated by equipping an MS with two 802.11 wireless radios, one at the front and one at the rear of the MS. Each radio associates with one of the 802.11-based wayside PoAs. The wayside PoAs are selected using an RSS-based technique. The wayside PoAs are connected to the central control station via a wired infrastructure network. The infrastructure network is assumed to be a private Internet Protocol (IP) network. The radios have directional antennae to cover longer ranges at the cost of reduced beamwidth, which is a preferable trade-off in a near line-of-sight propagation environment. To account for partial network failure due to breakdowns of a subset of wayside PoAs, wayside PoAs are deployed in a redundant manner.

Our proposed solution seeks to enhance communication availability by utilizing an alternative communication path provided by heterogeneous networks. Figure 3.2 illustrates a scenario of multiple equipment breakdowns and two different fail-safe communication paths over the heterogeneous networks. For the MS on the right, the 802.11 wireless radio on the right cannot attach to an operating wayside PoA. If the MS has a cellular wireless radio, it can transfer
Figure 3.1: Communication paths in a CBTC system

Figure 3.2: Communication paths in a hybrid network
the messages via the cellular wireless radio and the messages will traverse the cellular network and arrive at the central control station via Internet. In this case, the central control station is required to have a static public IP address to be reachable over the Internet. In Figure 3.2, for the MS on the left, the 802.11 wireless radio on the right is not able to directly connect to any wayside PoA. If this 802.11 wireless radio is in range of the 802.11 wireless radio of another MS, the two radios can form an ad hoc link to build a MANET. If the 802.11 wireless radio of the latter MS is able to connect to a functioning wayside PoA, it can forward the messages for the original MS. In this case, the 802.11 wireless radios are required to operate in infrastructure mode and ad hoc mode simultaneously, or another set of 802.11 wireless radios in ad hoc mode is required to operate in parallel.

In our solution, we assume (1) the central control station has an attachment to the infrastructure network and to each backup network, and (2) each MS is equipped with two radio sets where an radio set includes a 802.11 wireless radio for the infrastructure network and a wireless radio for each backup network. A radio set is installed at the front and the other at the rear for redundancy. Each radio set performs switchover between the infrastructure network and the backup network(s) independently. The decision algorithm will be presented in Chapter 4. We adopt ASONs to facilitate the switchover process. The ASONs are constructed using an open source software system called HyperCast.

### 3.2 HyperCast

HyperCast [40] encapsulates a logical node and its corresponding physical node(s) into an overlay socket which acts as the endpoint of communication in an overlay network. An overlay socket provides an Application Programming Interface (API) that supports functions of constructing a new overlay network, joining or leaving an existing overlay network, and sending and receiving data messages for the application, which are called the application messages. The main components in an overlay socket of HyperCast Version 4 include an overlay node, a forwarding engine, a Cross-Substrate Advertisement (CSA) processor and a multi-interface adapter as illustrated in Figure 3.3.

#### 3.2.1 Overlay Node

The overlay node is responsible for discovering nodes in the overlay network and maintaining neighbourhood relationships and the overlay topology.

The discovery is achieved by one of the three mechanisms: broadcast, buddy list, or rendezvous server. In the broadcast method, a new overlay node announces itself to existing members of the overlay network by broadcasting. This method is only viable when broadcast or multicast functions are available in the substrate. In the buddy list method, each overlay node maintains a list of overlay nodes that are likely to be in the overlay network. The list is referred to as a buddy list, and the overlay nodes in the buddy list are referred to as ren-
Figure 3.3: HyperCast Version 4 overlay socket
dezvous buddies. The rendezvous buddies provide information about joining the network to a new overlay node. In the rendezvous server method, each overlay node contacts a server which has a well-known address. The server contains information of active members in the overlay.

The supported overlay topologies include Delaunay triangulation, spanning tree, and a distributed hash table. Overlay nodes achieve node discovery and topology maintenance by exchanging periodic control messages, called protocol messages. The time between the transmissions of two successive protocol messages is called the heartbeat time. The heartbeat time controls the responsiveness and convergence rate of the overlay network after a change of the overlay topology.

3.2.2 Forwarding Engine

The forwarding engine handles the forwarding decision of an application message. If the recipient of the application message is the local overlay socket, the forwarding engine passes the application message up to the application program. If the application message needs to be further forwarded, the forwarding engine requests the addresses of the next-hop(s) from the overlay node and passes the application message down to the adapter.

3.2.3 CSA Processor


3.2.4 Adapter, Interface, and Address Repository

The adapter provides the send and receive functions to the other components in the overlay socket and interacts with the substrate networks. The adapters contain a number of interfaces and an address repository. An interface is designated to one substrate network. Each interface implements the send and receive functions specific to the protocol (e.g. UDP, Transmission Control Protocol (TCP), Ethernet, and Bluetooth) and delivery mode (e.g. unicast and multicast) of the substrate network. Incoming messages from different interfaces are collected into two buffers, one for protocol messages and one for application messages. Protocol messages are directed to the CSA processor and the overlay node. Application messages are directed to the forwarding engine. The address repository stores and maintains bindings of a logical address and substrate addresses and performs address resolution.

3.3 Switchover Using ASONs

We illustrate the usage of ASONs to facilitate the switchover process between the infrastructure network and a backup network using a cellular network. With the ability to support multiple
substrate networks, we can construct ASONs with two different schemes, single-substrate overlay and multi-substrate overlay. In the single-substrate overlay scheme an overlay network is constructed over a substrate network, whereas in the multi-substrate overlay scheme a single overlay network is constructed over multiple substrate networks. In the single-substrate overlay scheme, the switchover mechanism is realized by changing the overlay network to/from which messages are sent and received. On the other hand, in the multi-substrate overlay scheme, the switchover mechanism is realized by changing the interface within an overlay socket. However, as dynamic interface selection is not yet available in HyperCast, we adopt the single-substrate overlay scheme and leave the multi-substrate overlay scheme and its evaluation for future endeavour.

3.3.1 ASON Configuration

A CBTC system has a number of $n$ MSs, where each MS has two radios sets and each radio set has a radio for the infrastructure network and a radio for the cellular network, and 1 central control station, where the central control station has an attachment to the infrastructure network and an attachment to the Internet. An overlay socket is created on each radio for each radio set for each MS and also on each attachment of the central control station. Two overlay networks are created. One overlay network contains the overlay sockets that are created on top of the radios of the MSs for the infrastructure network and the attachment of the central control station to the infrastructure network. This overlay network is referred to as the primary overlay. The other overlay network contains the overlay sockets that are created on top of the radios of the MSs for the backup and the attachment of the central control station to the Internet. This overlay network is referred to as the secondary overlay.

We refer to the ASONs over the infrastructure network as the primary overlays and the ASONs over the backup network as the secondary overlays. Figure 3.4 illustrates the primary and secondary overlays created for the considered CBTC system.

3.3.2 ASON Operation

Each train control message between an MS and the central control station is duplicated into two replicas. Each replica is sent/received to/from one of the two radio sets. Within a radio set, a replica can be sent/received via one of the overlays, or can be further duplicated and sent/received via both overlays. We establish three modes for a train control message between a radio set and the central control station to be sent/received, namely complete-redundancy, join-on-switchover, and request-to-switchover. Before an overlay socket can send/receive messages, it needs to join the overlay, that is to exchange protocol messages to form and maintain an overlay network topology. In the complete-redundancy mode a train control message is sent/received in both overlays, whereas in the latter two modes a train control message is sent/received in one of the overlays at any time.
Figure 3.4: The primary and secondary overlays for a CBTC system with \( n \) MSs and a central control station
Figure 3.5: ASON operation modes
In the complete-redundancy mode, the overlay sockets of a radio set and the overlay sockets of the central control station join the primary and secondary overlay. The radio set sends/receives train control messages (application messages) to/from the central control station via the primary and secondary overlay simultaneously. There is no switchover process required in this mode. The flows of protocol messages and application messages are illustrated in Figure 3.5a.

In the join-on-switchover mode, a radio set and the central control station join and send/receive train control messages via the primary overlay. In addition, as the central control station is unaware of the handoff decision of the radio set, the central control station also needs to join and send train control messages to the secondary overlay. At this point, the radio set does not receive any train control message on the secondary overlay as it has not yet joined the secondary overlay. When triggered by a vertical handoff decision, the radio set leaves the current overlay, that is it sends a message to announce its departure to the neighbours and stops sending further protocol messages, the current overlay, and joins the secondary overlay and starts sending/receiving train control messages. Figure 3.5b demonstrates the process of a switchover from the primary overlay to the secondary overlay.

In the request-to-switch mode, a radio set and the central control station join the primary overlay and send/receive train control messages. In addition, the radio set and the central control station also join the secondary overlay. When triggered by a vertical handoff decision, the radio set sends a message to request the central control station to send/receive train control messages via the selected overlay. The message is referred to as the switchover request message. Upon receiving the switchover request message, the central control station stops sending train control messages to the current overlay and starts sending train control messages to the selected overlay. Figure 3.5c depicts a switchover process from the primary overlay to the secondary overlay in two steps. First, the radio set sends the switchover request message, and, second, the radio set and the central control station change the overlay to which they send/receive protocol messages and application messages.

The adoption of ASONs facilitates the switchover process into simple manipulation of the primary overlay and the secondary overlay. Next, we discuss the performance implications of the three modes in terms of delay and bandwidth consumption. Depending on the QoS and the cost of each substrate network, one can determine which mode to be adopted.

### 3.3.2.1 End-to-end Packet Delay, End-to-end Overlay Message Delay, Overlay Convergence Delay, and Switchover Delay

The end-to-end packet delay refers to the time elapsed from the instant a source sends a message to a destination to the instant when the destination receives the message. The end-to-end packet delay is decomposed into propagation delay, transmission delay, node processing delay, and queueing delay.

We define the end-to-end overlay message delay to be the time elapsed from the instant
a source overlay socket sends an overlay (protocol or application) message to a destination overlay socket to the instant the destination receives the overlay message. The end-to-end overlay message delay is composed of the end-to-end packet delay and the additional overlay socket processing delays of the source and the destination overlay sockets.

We also define the overlay convergence delay to be the time elapsed from the instant an overlay socket begins to join an ASON to the instant logical links are formed and the overlay has stabilized after exchanging protocol messages.

The switchover delay is defined to be the time elapsed from the instant a switchover process begins to the instant an overlay socket can receive application messages in the new overlay. In the complete-redundancy mode, as there is no switchover process and the new overlay has been maintained and has stabilized, the switchover delay is one end-to-end overlay message delay incurred by the transmission of an application message. In the join-on-switchover mode, since an overlay socket needs to join the new overlay, the switchover delay includes an overlay convergence delay in addition to an end-to-end overlay message delay caused by the transmission of an application message. In the request-to-switchover mode, the switchover process involves the transmission of a request message, which takes an end-to-end overlay message delay, and then an application message is transmitted through the new overlay. Therefore, the switchover delay consists of two end-to-end overlay message delays. Among the three modes, the switchover delay of the join-on-switchover mode yields the longest delay. We will evaluate the switchover delay of the join-on-switchover mode with the backup network being an LTE network in Section 3.4.1.

3.3.2.2 Bandwidth Consumption

The bandwidth consumed by an overlay socket can be separated into two parts, the bandwidth consumed by exchanging protocol messages and that by exchanging application messages. Denote the rate at which protocol messages sent (uplink) and received (downlink) per time unit by an overlay socket measured in bits per second (bps) by \( B_P \) and the rate of application messages sent and received per time unit measured in bps by \( B_A \). The request message used in the request-to-switch mode on average occupies a data rate of \( B_R \). \( B_P \) depends on heartbeat time of the ASON. \( B_A \) is independent to the configuration parameters of the ASON and is only subject to the train control application. In contemporary train control protocols, an MS typically generates a data traffic of 70 kbps with 50 kbps for the downlink and 20 kbps for the uplink [41].

Nominally, in the complete-redundancy mode, each overlay socket consumes \( B_P + B_A \) bps; in the join-on-switchover mode, the selected overlay socket consumes \( B_P + B_A \) bps while the other overlay socket does not consume any bandwidth; in the request-to-switch mode, the selected overlay socket consumes \( B_P + B_A + B_R \) bps and the other overlay socket consumes \( B_P \) bps. We will evaluate \( B_P \) in Section 3.4.2 to find out if overlay networks can be established on the considered wireless networks.
Chapter 3. ASON Over Hybrid Network

3.4 Evaluation of ASON over LTE

To examine if ASON over a cellular network, specifically an LTE network, is a feasible solution to enhance communication availability, we conduct experiments to address the following two questions:

- whether an ASON over an LTE network is capable of satisfying the time constraint, and
- whether an LTE network can support the aggregate bandwidth consumed by exchanging protocol and application messages.

3.4.1 Switchover Delay Experiment

The objective of this experiment is to assess whether the switchover delay in the join-on-switchover mode, that is an overlay convergence delay plus an end-to-end overlay message delay, with the substrate network being an LTE network is able to meet the delay constraint set by [1].

3.4.1.1 Experiment Setup

The setup of the experiment consists of two parts, the physical network setup and the ASON setup, which is illustrated in Figure 3.6. We start with the physical network setup. The experiment uses two computing devices, a MacBook Pro and a Mac Pro. The MacBook Pro is equipped with a 2.7 GHz Intel Core i7, and a 8 GB DDR3 RAM; the Mac Pro is equipped with two 2.8 GHz Quad-Core Intel Xeon, and a 4 GB DDR2 RAM. Both of them run the operating system OS X 10.6.8. The MacBook Pro is connected to a commercial cellular adapter, 4G LTE Novatel Wireless U679 Turbo Stick from Bell Canada, via the USB port. The cellular adapter is shown in Figure 3.7. The Mac Pro is connected to the campus network of the University of Toronto via an Ethernet cable. The devices and the network connections are illustrated in Figure 3.8. The experiment is completed indoors in a lab environment.

The Mac Pro is used to emulate the central control station and the MacBook Pro is used to emulate one of the radios for the backup network of an MS. The Ethernet network to which
Figure 3.7: Bell Canada’s 4G LTE Novatel Wireless U679 Turbo Stick

Figure 3.8: Experiment testbed at the University of Toronto
the Mac Pro is connected represents the wired infrastructure network. The Mac Pro is assigned a static public IP address. The MacBook Pro is connected to a commercial LTE network via the aforementioned cellular adapter. The cellular adapter has a dynamically assigned private IP address.

We construct an ASON with two overlay sockets using HyperCast, one at the Mac Pro and the other at the MacBook Pro. Both overlay sockets run on top of Internet, specifically using the unicast UDP for the transport layer protocol and IP for the network layer protocol. We use a DT protocol that provides a buddy list rendezvous method. The address of the overlay socket of the Mac Pro, specified by the public IP address and a port number, is entered as the buddy at the overlay socket of the MacBook Pro. As the overlay socket of the MacBook Pro joins the ASON, it will spontaneously send a protocol message to contact the buddy (the overlay socket of the Mac Pro and the message) will establish a Network Address Translation (NAT) entry that enables the overlay socket of the Mac Pro to contact it.

### 3.4.1.2 Experiment Scenario

To find the earliest time an application message can be received after an overlay socket joins the overlay, we set the overlay socket of the Mac Pro to be a sender and the overlay socket of the MacBook Pro to be a receiver. The sender creates the ASON and attempts to send application messages to the receiver. An application message is sent every 10 ms to imitate continuous transmission of application messages. The receiver joins the overlay, records the time of the action, waits for the application messages, and records the time when the first application message is received. The switchover delay is the difference between the two timestamps. We collect measurements of the switchover delay while varying the heartbeat time which controls the convergence speed of the overlay. The experiment is repeated for 250 times.

### 3.4.1.3 Experiment Results

Figure 3.9 depicts the empirical distribution of the switchover delay for different heartbeat times and Table 3.1 summarizes the median, 95\textsuperscript{th} percentile, and the maximum of the measurements.

From Figure 3.9 and Table 3.1, we first observe that the median of the switchover delay increases linearly when the heartbeat time increases from 25 ms to 1000 ms. This is expected since the overlay convergence delay is expected to increase with the heartbeat time.

Next, we observe the 95\textsuperscript{th} percentiles. We discover that for a heartbeat time from 250 to 1000 ms the 95\textsuperscript{th} percentile of the measurements is within two heartbeat times, whereas for a heartbeat time from 25 ms to 150 ms, the 95\textsuperscript{th} percentile does not exhibit a significant correlation to the heartbeat time. This may be an indication that, when the heartbeat time is less than 150 ms, the end-to-end packet latency of LTE dominates the impact of the heartbeat time and prevents the switchover from being further decreased. We suggest operating an ASON at a heartbeat time about 150 ms, for two reasons: first, to satisfy the delay constraint
Figure 3.9: Empirical distribution of switchover delay of ASON over LTE

Table 3.1: Statistics of the switchover delay measurements of ASON over LTE

<table>
<thead>
<tr>
<th>Heartbeat Time</th>
<th>Median</th>
<th>95th Percentile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 ms</td>
<td>150 ms</td>
<td>300 ms</td>
<td>13520 ms</td>
</tr>
<tr>
<td>150 ms</td>
<td>235 ms</td>
<td>324 ms</td>
<td>3118 ms</td>
</tr>
<tr>
<td>250 ms</td>
<td>320 ms</td>
<td>496 ms</td>
<td>1797 ms</td>
</tr>
<tr>
<td>500 ms</td>
<td>590 ms</td>
<td>969 ms</td>
<td>2760 ms</td>
</tr>
<tr>
<td>1000 ms</td>
<td>1109 ms</td>
<td>1831 ms</td>
<td>5819 ms</td>
</tr>
</tbody>
</table>

specified in [1], and, second, to balance the trade-off between overlay convergence delay and the bandwidth consumption.

For the remaining 5% of cases (the measurements between the 95th percentile and the maximum), the packets have been dropped in the LTE network and the overlay converges in more than two heartbeat times. For example, one measurement of the switchover delay for the heartbeat time equal to 25 ms even took 13520 ms. Note that the maximum of the measurements does not display a correlation to the heartbeat time.

### 3.4.2 Bandwidth Consumption Experiment

The objective of this experiment is to evaluate whether the aggregate bandwidth consumption of protocol message and application message, $B_P + B_A$, can be supported by an LTE network. Specifically, we seek to assess the bandwidth of protocol message, $B_P$, consumed by an overlay
socket in an ASON with two nodes. We evaluate $B_P$ for different values of heartbeat time. Then we compare $B_P + B_A$ to the available bandwidth of an LTE network.

### 3.4.2.1 Experiment Setup

This experiment involves one computing device. We use the same MacBook Pro as in the previous experiment. The MacBook Pro does not need to be attached to any physical network. This experiment is conducted indoors at the University of Toronto.

We construct an ASON with two overlay sockets running on top of unicast UDP/IP using the loopback address with two different ports on the MacBook Pro. One of the overlay sockets is configured as the rendezvous buddy and its address is added to the buddy list of the other overlay socket. The two overlay sockets have the same heartbeat time.

### 3.4.2.2 Experiment Scenario

The two overlay sockets join the ASON simultaneously, stay for 20 seconds, and leave the overlay. No application message is sent during their stay. After leaving the overlay, each overlay socket records the total protocol messages sent and received, measured in bits and number of packets, and the total time elapsed, measured in milliseconds, and computes the bandwidth in bps. We repeat the measurement and vary the heartbeat time from 25 to 1000 ms.

### 3.4.2.3 Experiment Results

Since the overlay network only contains two overlay sockets, the messages sent (uplink) by one overlay socket are the messages received (downlink) by the other and vice versa. Therefore, the sum of the uplink and the downlink traffic is the same for the two overlay socket. The total bandwidth consumed by the uplink and the downlink traffic as a function of heartbeat time is illustrated in Figure 3.10. The bandwidth consumption decreases as the heartbeat time increases. We consider if an LTE network is able to support the bandwidth consumed by topology maintenance as well as the train control application. Since LTE adopts Orthogonal Frequency Division Multiplex (OFDM), LTE connections operate on orthogonal sub-carriers and the available bandwidth is dedicated for each LTE connection. According to the provider, the minimum available bandwidth of an LTE connection is 12 Mbps [42]. From the previous experiment 3.4.1, we recommend operating an ASON at a heartbeat time of 150 ms, which generates a traffic of 22.72 kbps at the data link layer. Along with the traffic generated by the train control application, which is 70 kbps [41], the aggregate bandwidth of protocol and application messages, $B_P + B_A$, is less than 100 kbps and should be sufficiently supported by an LTE network.
Figure 3.10: Bandwidth consumed by an overlay socket to maintain the network topology in an ASON with two nodes for different heartbeat times.
Chapter 4

Design of a Location-based Joint Horizontal and Vertical Handoff Decision Algorithm with Adaptive Planning

In this chapter, we present our proposed handoff decision algorithm. The handoff decision algorithm selects a PoA from the hybrid network. The objectives of the algorithm are reducing service interruption, low QoS, and number of handoffs, increasing handoff interval, eliminating ping-pong effects, and maintaining resilience to network failure or degradation.

We evaluate each PoA by a score which is a composite measure that combines multiple criteria where the application has the freedom to select the criteria as well as define the normalization function of each criterion and the method to combine the criteria.

We observe predictable motion paths in CBTC systems. We assume that the score of a PoA measured by different MSs arriving at the same location should be similar. This means that we can take advantages of historical measurements to predict future measurements. If this assumption is valid, we can enhance handoff performance by collecting measurements and planning out handoffs in advance.

We present our proposal in two steps. We first demonstrate the data collection process and then the handoff decision algorithm.

4.1 Data Collection Process

We illustrate a method to collect measurement data for all PoAs of a train line without introducing significant changes to normal train operation. We make the following three necessary assumptions: (1) all MSs and PoAs are assumed to function properly, that is no hardware failure or hardware degradation, and no physical obstruction is present, (2) each PoA is represented
by a score at any time, where a score is computed based on the instantaneous measurements of network-related and terminal-related criteria and the preset values of service-related and user-related criteria.

The data collection process for a train line involves multiple MSs. The process is illustrated as a flow diagram in Figure 4.1. Each MS starts at the same origin and travels toward the same end-point. While an MS travels, each radio installed on the MS periodically performs a scanning process where a radio searches for nearby PoAs and measures the scores of the PoAs on all channels. After a scanning process is completed, each radio generates a scan result that includes the following fields: a timestamp \((t)\), which is expressed as an absolute time in seconds, a location-stamp \((l)\), which specifies the position of the MS on the track with respect to the origin in meters, a list of detected PoAs \(\{\phi^d\}\) and their respective scores \(\{s^d\}\). For simplicity, the radios belonging to a radio set are assumed to be synchronized and perform scanning at the same time. The scan results produced by radios of a radio set are accumulated into a score measurement trace and are sorted chronologically. After an MS arrives at the end-point of the train line, it uploads the score measurement trace for the front and the rear radio sets to the central control station. The central control station appends the measurements of the score of each PoA to the score database.

4.1.1 Score Database

The score database consists of two tables, a table for the front radio set and another for the rear radio set. Each table has two dimensions where one dimension specifies space and the other specifies time. The range of the space corresponds to the length of the train line and the space is partitioned into equal-sized spatial sections. For instance, the \(j^{th}\) spatial section refers to the range \([r+(j-1)s, r+js]\) m of the train line where \(r\) is the reference point (may be different from the origin) and \(s\) is the size of a spatial section. The spatial section size should be selected to differentiate the location-induced randomnesses. We will investigate appropriate spatial section sizes in Section 5.1. The time dimension corresponds to the transit service period, that is the departure time of the first train and the arrival time of the last train, and is partitioned into uneven temporal sections based on peak hours (presence of interferences) and off-peak hours (absence of interferences). Suppose that the first train departs at 5 AM, the last train arrives at 1 AM, and the peak hours are 6 to 10 AM and 3 to 7 PM; the temporal sections are \(\{T^h_i\} = \{[5,6), [6,10), [10, 15), [15, 19), [19,24) \cup [0,1]\} in hours. The temporal sections can also be extended to account for potential differences between weekdays, weekends and holidays. The intersection of a spatial section and a temporal section is referred to as a cell as indicated in Figure 4.2. Conceptually, the score of a PoA \(\phi\) in a cell is viewed as a random variable \(S_{i,j}^\phi\), while the scores of \(\phi\) from different cells are viewed as different random variables. Each cell contains a number of vectors where a vector stores the measurements of the score of a particular PoA. The measurements in the vector regarding a PoA \(\phi\) are realizations of \(S_{i,j}^\phi\) and yield the empirical distribution of \(S_{i,j}^\phi\). After the central control station finishes appending data to the
Chapter 4. Design of Handoff Decision Algorithm

36

Yes

Start

Assign a terminal station of a train line as the origin and the other as the end-point

MS departs from the origin

MS continuously collects scan results until reaches the end-point

MS uploads scan results to central control station

Central control station appends scan results to score database

Number of measurements in each vector in each cell > minimum sample size?

No

Yes

End

Figure 4.1: Data collection process
The score database is maintained at the central control station to avoid data synchronization and concurrent access. Upon receipt of the score measurement traces uploaded by an MS, the central control station appends the measurements to the corresponding vector (identified by the PoA), cell (identified by the timestamp and the location-stamp), and table (identified by front or rear).

4.2 Location-based Handoff Decision Algorithm with Adaptive Planning

The handoff decision algorithm starts by downloading a copy of the score database from the central control station. The score database is used to compute the coverage area of each PoA which is the area in which the score an PoA exceeds a certain threshold, which will be elaborated in Section 4.2.1. The coverage areas of all PoAs are stored into a table called the coverage profile. The coverage profile is further used to compute a sequence of minimum number of handoffs, called the handoff map. Computation of the handoff map is described in Section 4.2.2. A radio set follows the handoff map and performs a handoff to the next PoA when the MS enters the coverage area of the PoA. A radio set, while associating with a PoA, takes periodic measurements of nearby PoAs. These real-time measurements are used to identify unpredictable events. The handoff decision algorithm updates the handoff map adaptively.
to *unpredictable events*, which are discrepancies between the collected measurements and the real-time measurements and indicate potential network failure or degradation. We will discuss unpredictable events in detail in Section 4.2.3. If the unpredictable events involve a subset of PoAs that are included in the current handoff map, the algorithm recomputes a new handoff map that excludes those PoAs; if the unpredictable events involve all PoAs in the system, or if no handoff map is available, the algorithm falls back to an algorithm uses the relative RSS with a threshold value. Unpredictable events are reported to the central control station for further investigation. The central control station may disseminate the unpredictable events to prospective MSs to reserve ample time for reaction. When the MS completes the trip, it may upload the score measurement traces to the central control station to enrich the score database. Figure 4.3 illustrates the operation of an MS traveling on a train line and the handoff decision algorithm.

4.2.1 Coverage Area and Coverage Profile

The coverage area of a PoA, as briefly defined previously, is the area in which the score of the PoA exceeds a predefined threshold value, called the *coverage threshold* and denoted by $T_c$, at a probability greater than or equal to a predefined *coverage probability* $p_c$. The coverage threshold and the coverage probability are determined by the application where the coverage threshold corresponds to the application-specific QoS requirements and the coverage probability should be a high probability value close to 1. Since we partition the space into discrete spatial sections, the coverage area of a PoA $\phi$ is represented by a set of spatial sections. A spatial section $j$ is said to be $p_c$-covered by a PoA $\phi$ at the $i^{th}$ temporal section if

$$P(S_{i,j}^{\phi} > T_c) = 1 - F_{S_{i,j}^{\phi}}(T_c) \geq p_c, \quad (4.1)$$

where $P(\cdot)$ denotes a probability and $F_{S_{i,j}^{\phi}}(\cdot)$ denotes the Cumulative Distribution Function (CDF) of $S_{i,j}^{\phi}$. If $S_{i,j}^{\phi}$ has a strictly increasing and continuous CDF, Equation 4.1 can be expressed using the Inverse Cumulative Distribution Function (ICDF) or the quantile function as

$$F_{S_{i,j}^{\phi}}^{-1}(F_{S_{i,j}^{\phi}}(T_c)) = T_c < F_{S_{i,j}^{\phi}}^{-1}(1 - p_c). \quad (4.2)$$

Finally, the coverage area of $\phi$ comprises all spatial sections that are $p_c$-covered by $\phi$. Note that, the $p_c$-covered spatial sections may not be contiguous, that is, the coverage area of $\phi$ may consist of multiple partitions, such as the coverage area of an LTE base station.

We use the measurements in the score database to estimate the coverage areas of PoAs. The measurements of $S_{i,j}^{\phi}$ form the empirical distribution function of $S_{i,j}^{\phi}$. By asymptotic theory [43], or large sample theory, the empirical distribution function drawn from a number of $n$ measurements converges almost surely to the true distribution function as $n$ approaches infinity. Thus, when the number of measurements of $S_{i,j}^{\phi}$ is sufficiently large, first, the left-hand-side of Equation 4.1 becomes the ratio of the number of measurements that are greater than
Chapter 4. Design of Handoff Decision Algorithm

Figure 4.3: Location-based handoff decision algorithm with adaptive planning
\( T_c \) to the total number of measurements in the \((i,j)\)th cell, and, second, the right-hand-side of Equation 4.2 becomes the \((1 - \phi c)\)th sample quantile. For example, if the \(j\)th spatial section is \(\phi c\)-covered by \(\phi\) during temporal section \(i\) with \(\phi c = 0.95\), then there are at least 95\% of the measurements in the \((i,j)\)th cell greater than \(T_c\), or, equivalently, the 0.05\th sample quantile of the measurements in the \((i,j)\)th cell is greater than \(T_c\). We compare the \((1 - \phi c)\)th sample quantile to \(T_c\) for each spatial section for a PoA to determine its coverage area and repeat the procedure for all PoAs. The results are stored into the coverage profile in the format of a key-value pair where the key is the label of the PoA and the value is the set of \(\phi c\)-covered spatial sections.

Finally, the coverage areas of all PoAs in the system at different temporal sections (since we suppose that a PoA may have a smaller coverage area when interfering devices are present and vice versa) are stored into a lookup table called the coverage profile.

### 4.2.2 Handoff Map

A handoff map specifies the sequence of PoAs that an MS hands off to on its route given the location and the destination of the MS. The PoAs in a handoff map are selected such that their coverage areas overlap to form a continuous coverage from the location of the MS to the destination. A handoff map for a route given the location of the MS is computed as follows. First, we construct a undirected graph with PoAs as vertices, (specifically, each partition of the coverage area of a PoA is a vertex). Vertices are connected by an edge when (the partitions of) the coverage areas of the corresponding PoAs overlap. We refer to this graph as the overlapping relationship graph. A path between two vertices in the overlapping relationship graph means that an MS can travel from one PoA to another PoA while maintaining satisfactory QoS. If there exist many paths between two specified vertices, the shortest path is preferred. Since any path between two specified vertices guarantees an MS to receive satisfactory QoS (at the coverage probability \(\phi c\)), a shortest path prevails other non-shortest paths as the number of handoffs is minimized. Then, since each handoff introduces a period of disconnection, referred to as the disconnection time, minimizing the number of handoffs also minimizes the total disconnection time. If there exist multiple shortest paths, they are deemed equivalent where one is randomly chosen and becomes the handoff map.

Using a handoff map strategy has the following advantages. First, the handoff latency is reduced as the PoA that will be associated with is foreknown and therefore the discovery stage in the handoff procedure can be omitted. Second, if the real-time scores of the PoAs in the map correspond to their scores stored in the score database, a handoff map ensures an MS to be always \(\phi c\)-covered and thus decreases the low QoS duration. Third, as a handoff map is a shortest path between the two specified vertices, a handoff map minimizes the number of handoffs, which in turn increases the handoff intervals. Fourth, a handoff map can effectively eliminate ping-pong effects as a PoA appears at most once on a shortest path.
4.2.3 Unpredictable Event

We define an unpredictable event to be an occurrence of real-time score measurements of a PoA being smaller (worse) than the score measurements recorded in the score database with a statistical significance. An unpredictable event can be caused by a hardware failure, a hardware degradation, or an appearance of a physical obstacle between a PoA and an MS. If an unpredictable event involves a PoA in the handoff map, the handoff map can no longer ensure an MS to receive satisfactory QoS at a minimum probability of $p_c$. Therefore, it is crucial to monitor for unpredictable events in real-time.

Denote by $s_{i,j}^\phi$ a score measurement of a PoA $\phi$ taken by an MS at temporal section $i$ and spatial section $j$. If $\phi$ has a hardware failure, then $\phi$ can no longer be detected by an MS and $s_{i,j}^\phi$ is represented by $-\infty$.

Identification of hardware failure has a different procedure from identification of hardware degradation, physical obstruction, or antenna misalignment. To identify hardware failure, we first search the set of PoAs that cover the cell $(i,j)$ from the coverage profile and denote the set by $\{\phi_{i,j}\}$. We compare $\{\phi_{i,j}\}$ to the list of detected PoAs $\{\phi^d\}$ in a scan result. Then $\{\phi_{i,j}\} \setminus \{\phi^d\}$ is the set of failed PoAs.

To identify hardware degradation or physical obstruction, for each PoA $\phi$ in the set $\{\phi^d\}$ we check if the following condition is satisfied:

$$s_{i,j,1}^\phi \leq F^{-1}_{S_{i,j}^\phi}(p_{u_0}), \quad s_{i,j,2}^\phi \leq F^{-1}_{S_{i,j}^\phi}(p_{u_0}), \quad \cdots, \quad s_{i,j,k_u}^\phi \leq F^{-1}_{S_{i,j}^\phi}(p_{u_0}),$$  \hspace{1cm} (4.3)

where $s_{i,j,1}^\phi, s_{i,j,2}^\phi, \ldots, s_{i,j,k_u}^\phi$ denote a sequence of $k_u$ consecutive and independent score measurements, $F^{-1}_{S_{i,j}^\phi}(\cdot)$ is the quantile function of $S_{i,j}^\phi$ and is approximated by the empirical measurements from the score database, and $p_{u_0}$ is a small probability value close to 0. $k_u$ is referred to as the number of consecutive samples for unpredictable event identification, and $p_{u_0}$ is referred to as the per-sample unpredictable event probability.

We analyze the sensitivity, which measures the probability of correct identification given an unpredictable event has occurred, and the specificity, which measures the probability of correct rejection given that an unpredictable event has not occurred, of the identification scheme given by Equation 4.3. A test with high sensitivity is useful for confirming the absence of an unpredictable event; a test with high specificity is useful for confirming the presence of an unpredictable event. Denote by $I$ the event that an unpredictable event is identified and $U$ to be the event that an unpredictable event occurs, then $P(I|U)$ is the sensitivity and $P(\overline{I}|\overline{U})$ is the specificity.

$P(I|U)$ is the probability of observing $k_u$ consecutive real-time measurements that are all below the $p_{u_0}$th quantile of $S_{i,j}^\phi$ given that an unpredictable event has occurred. Therefore,

$$P(I|U) = P(s_{i,j}^\phi \leq F^{-1}_{S_{i,j}^\phi}(p_{u_0}))^{k_u}$$
by independence of the measurements. \(P(I|U)\) increases as \(p_{u_0}\) increases, as \(k_u\) decreases, or as the distribution of \(s_{i,j}^\phi\) shifts to the left.

\(P(I|U)\) is the probability of not observing \(k_u\) consecutive real-time measurements that are all below the \(p_{u_0}\)th quantile of \(S_{i,j}^\phi\) given that an unpredictable event has not occurred. Therefore,

\[
P(I|U) = 1 - P(s_{i,j}^\phi \leq F_{S_{i,j}^\phi}^{-1}(p_{u_0}))^{k_u} = 1 - p_{u_0}^{k_u},
\]

where the first equality follows from independence of the measurements and the second equality follows from \(F_{S_{i,j}^\phi}(\cdot) = F_{S_{i,j}^\phi}(\cdot)\). \(P(I|U)\) increases as \(p_{u_0}\) decreases or as \(k_u\) increases.

From the above, we see that there is a trade-off between sensitivity and specificity. A test with high sensitivity but low specificity is likely to raise many false alarms, that is many unpredictable events are identified, but only a small proportion of the identifications truly indicates the occurrence of an unpredictable event. On the other hand, a test with low sensitivity but high specificity may miss identifying many unpredictable event occurrences, yet if an unpredictable event is identified, it is almost certain to occur. To achieve moderate sensitivity and specificity, the values of \(p_{u_0}\) and \(k_u\) need to be carefully selected. A small \(p_{u_0}\) requires more measurements to be taken in advance, which lengthens the data collection process, whereas a large \(k_u\) delays the time for an unpredictable event to be identified, which is undesirable for time-critical systems. For example, to fulfil the requirement of a two-second long train-to-wayside message delay, we reserve one second for unpredictable event identification and one second for handoff latency and train-to-wayside message delay via the new PoA. A measurement is obtained every 0.25 seconds or at 4 Hz, which yields \(k_u = 4 \text{ Hz} \times 1 \text{ s} = 4\) measurements. Suppose the score is solely based on the RSS, then under log-normal shadowing \(S_{i,j}^\phi\) has a normal distribution with mean \(\mu\) and standard deviation \(\sigma\) when expressed in dB scale. Further, suppose an unpredictable event causing a degradation of \(\delta = 20 \text{ dB}\) has occurred. Then \(s_{i,j}^\phi\) is now drawn from a normal distribution with mean \(\mu - \delta\) and standard deviation \(\sigma\). It follows that the sensitivity of the identification scheme is

\[
P(I|U) = \left(\frac{1}{2}(1 + \text{erf}(\text{erf}^{-1}(2p_{u_0} - 1) + \frac{\delta}{\sigma\sqrt{2}}))\right)^{k_u}.
\]

Finally, selecting a value of 0.01 for \(p_{u_0}\) yields \(P(I|U) = 0.9888\) and \(P(I|U) = 1 - 10^{-8}\).

\[\text{RSS degradation resulted from a line-of-sight path to a non-line-of-sight path typically causes a drop by 20 to 30 dB} \, [44].\]
Chapter 5

Design Issues and Justifications

In Chapter 4 we assumed that the randomness in the score of a PoA at a location within a temporal section can be represented by a random variable. This was exploited for the development of our algorithm. Since the distribution of the random variable is unknown, measurements are used to obtain the empirical distribution. To obtain the empirical distribution, we discretize the space into spatial sections, referred to as segments, and in such a way that the score of a PoA within a segment can be represented by a single random variable.

The size of segments must be decided with care. A small segment size provides a better granularity and yields a highly precise and accurate coverage estimate. However, under a fixed number of available measurements, a small segment size results in a small amount of samples in each segment, which may lead to an empirical distribution that diverges from the actual distribution of the random variable. Hence, selecting a segment size becomes an important engineering question that determines the trade-off between the accuracy and the fidelity of the coverage estimates, where the former restricts the maximum segment size and the latter limits the minimum segment size.

5.1 Maximum Segment Size

The coverage area of a PoA is computed from the distribution of the score received by an MS at, ideally, any location. However, this is impractical. Therefore, we divide the space into discrete segments. However, if a segment is too large such that the $(1-p_c)$th quantile of the score measured at one end of the segment is above the threshold while the score measured at the other end of the segment is below the threshold, then the segment cannot be accurately classified. We investigate how large a segment can be such that the difference between the score measured at different points within a segment is bounded by a specified amount.

Consider a random variable $S_d$ representing the score of a PoA measured at a distance $d$ relative to the PoA. Suppose that, for a small value $\Delta \geq 0$, the random variable $S_{d+\Delta}$ is similar to $S_d$. We quantify the difference between $S_{d_1}$ and $S_{d_2}$ where $d_2 > d_1$ in terms of the difference between their respective $p^{th}$ quantiles, i.e., $|q_{p,d_1} - q_{p,d_2}|$. Let $d_1$ and $d_2$ represent any two points
inside a segment. We require the difference $|q_{p,d_1} - q_{p,d_2}|$ to be bounded by a given value $\epsilon > 0$, that is

$$|q_{p,d_1} - q_{p,d_2}| \leq \epsilon.$$  \hfill (5.1)

This requirement results in a maximum distance between any two points inside the same segment, and, hence, provides an upper bound on the segment size. Since $q_{p,d}$ is unlikely to be a linear function of $d$, $|q_{p,d_1} - q_{p,d_2}|$ cannot be simplified as a function of the difference of $d_1$ and $d_2$. This means that the value of $|d_1 - d_2|$, resulting from Equation 5.1 when the equality holds, depends not only the difference between $d_1$ and $d_2$, but also on the values of $d_1$ and $d_2$. Specifically, $|d_1 - d_2|$ may become very small when $d_1$ and $d_2$ correspond to two points that are very close to the PoA as the variation in $S_d$ increases at locations closer to the PoA. Therefore, it is important that the equality in Equation 5.1 should be satisfied at a significant location.

We decide that, as a handoff decision should only be generated when the score is around the coverage threshold $T_c$, the location where the equality in Equation 5.1 holds is when $d_1$ (or $d_2$) equals a distance such that $q_{p,d_1}$ (or $q_{p,d_2}$) equals $T_c$.

We attribute the randomness in $S_d$ to the underlying wireless propagation, where, specifically, $S_d$ is related to the RSS. Further, we assume the coverage threshold $T_c$ and $\epsilon$ can both be mapped to values in terms of RSS. Then, we seek to identify restrictions on the maximum segment size using Equation 5.1 for given values of $T_c$ and $\epsilon$ and assumptions on the distribution of $S_d$ with wireless propagation models.

### 5.1.1 Wireless Propagation Models

We start with investigating the fundamentals of radio propagation over a fading channel and the components causing the variations in the received power. Without loss of generality, we use the generic terms transmitter and receiver to refer to PoAs and MSs, respectively. Consider a simple case of a static transmitter (PoA) and a dynamic receiver (MS). The transmitted signal arriving at the receiver is subject to large-scale fading and small-scale fading where large-scale fading includes two components: attenuation due to path loss and shadowing due to the signal being blocked by obstacles between the transmitter and the receiver. Small-scale fading, also called multipath fading, is due to multiple copies of the desired signal arriving at the receiver causing constructive and destructive interferences.

#### 5.1.1.1 Path Loss

Suppose that both shadow fading and multipath fading are absent. Then, the far field relationship between the transmitted power and the received power can be described by Friis transmission equation [45] as

$$\frac{P_r}{P_t} = G_t G_r \left( \frac{\lambda}{4\pi d} \right)^\gamma,$$  \hfill (5.2)
where $P_t$ is the transmitted power in milliwatt (mW), $P_r$ is the received power in mW, $G_t$ is the transmitter’s antenna gain, $G_r$ is the receiver’s antenna gain, $\lambda$ is the wavelength in meters, $d$ is the Euclidean distance between the transmitter and the receiver in meters, and $\gamma$ is the path loss exponent that describes the slope of the path loss characteristics in dB as a function of distance, which is typically determined by curve fitting to measured data [46]. In a measurement trace available to us, $\gamma = 2$ shows a decent curve fitting (see Chapter 6). In dB scale, Equation 5.2 becomes

$$P_{r_{dB}} = P_{t_{dB}} + G_{t_{dB}} + G_{r_{dB}} + 10\gamma \log_{10} \frac{\lambda}{4\pi} - 10\gamma \log_{10} d.$$  

(5.3)

Setting $k = P_{t_{dB}} + G_{t_{dB}} + G_{r_{dB}} + 10\gamma \log_{10} \frac{\lambda}{4\pi}$, Equation 5.3 reduces to

$$P_{r_{dB}} = k - 10\gamma \log_{10} d.$$  

(5.4)

For numerical computations, we set the parameters $P_{t_{dB}}$, $G_{t_{dB}}$, $G_{r_{dB}}$ and $\lambda$ based on the rules and regulations published by Federal Communications Commission (FCC) [47]. To operate in three of the industrial, scientific and medical (ISM) radio bands: 902 to 928 MHz, 2.4 to 2.4835 GHz and 5.725 to 5.875 GHz, where the second band is adopted by 802.11 b/g/n and the third band is adopted by 802.11 a/h/j/n/ac, the devices must comply to the following two restrictions. First, the maximum transmitter output power before being amplified by the antenna is 30 dBm (1 W). Therefore, we decide to set $P_{t_{dB}} = 20$ dBm (100 mW). Second, the Effective Isotropic Radiated Power (EIRP) is 36 dBm. EIRP is calculated by EIRP = $P_{t_{dB}} + G_t - L_c$ where $L_c$ is the cable loss. We assume there is no loss introduced by the cable and EIRP = $P_{t_{dB}} + G_t$. We consider the transmitter has a low gain omni-directional antenna of $G_{t_{dB}} = 2$ dB [48]. Then, the EIRP is $20 + 2 = 22$ dBm, which is within FCC’s restrictions. The receiver is equipped with a directional antenna with a gain of $G_{r_{dB}} = 8$ dB for the main lobe and -2 dB for the back lobe [49]. Finally, with the assumption that the WLANs operate at 2.4 GHz, we obtain $\lambda = \frac{c}{f} = \frac{3\times10^8}{2.4\times10^9} = 0.125$ m, where $c$ is the speed of light. With these choices, $k$ in Equation 5.4 yields -10.05 dBm for the main lobe and -20.05 dBm for the back lobe. We will use the value of $k$ for the back lobe in our numerical computations since the smaller value of $k$ imposes the more stringent constraint. The parameters are summarized in Table 5.1.

### 5.1.1.2 Shadowing

In the presence of blocking objects between the transmitter and the receiver, the received signal experiences an additional attenuation on top of path loss due to the reflection and the absorption of the transmitted signal by the blocking objects. Generally, the physical properties such as the location, size, and dielectric permittivity of the blocking objects are unknown. Therefore, the shadowing attenuation requires statistical modelling. The most common model is log-normal shadowing. Combining path loss and log-normal shadowing, the far field received power in dB
Table 5.1: Parameters for path loss

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>2</td>
</tr>
<tr>
<td>$P_{t\text{dB}}$</td>
<td>20 dBm</td>
</tr>
<tr>
<td>$G_{t\text{dB}}$</td>
<td>2 dB</td>
</tr>
<tr>
<td>$G_{r\text{dB}}$ (main lobe)</td>
<td>8 dB</td>
</tr>
<tr>
<td>$G_{r\text{dB}}$ (back lobe)</td>
<td>-2 dB</td>
</tr>
<tr>
<td>$f$</td>
<td>2.4 GHz</td>
</tr>
<tr>
<td>$k$ (main lobe)</td>
<td>-10.05 dBm</td>
</tr>
<tr>
<td>$k$ (back lobe)</td>
<td>-20.05 dBm</td>
</tr>
</tbody>
</table>

scale, $P_{r\text{dB}}$, is modified from Equation 5.4 to

$$P_{r\text{dB}} = k - 10\gamma \log_{10} d + S_{\text{dB}},$$

(5.5)

where $S_{\text{dB}}$ is the shadowing component that is normally distributed with zero mean and a standard deviation of $\sigma_S$, which is called the shadow standard deviation and is typically in the range between 4 and 12 dB [46, 50]. Furthermore, $\sigma_S$ has been observed to be nearly independent of the path length $d$ [46]. Explicitly, the Probability Density Function (PDF) of the received power at a distance $d$ from the transmitter in dB scale is

$$f_{P_{r\text{dB}}}(x) = \frac{1}{\sqrt{2\pi\sigma_S}} e^{-\frac{(x-(k-10\gamma \log_{10} d))^2}{2\sigma^2_S}}, \quad x \in \mathbb{R}.$$  

(5.6)

Specifically for tunnel environments, Zhang and Hwang [51] showed the slow variations of the received signal follow the log-normal distribution with an averaged standard deviation of 4.7 dB for a 900 MHz channel from measurements conducted in tunnels. Moreover, Briso-Rodríguez [52] presented measurements of the shadowing effect caused by two trains passing inside a tunnel.

### 5.1.1.3 Multipath Fading

Multipath fading is a phenomenon caused by reflection and refraction of the transmitted signal as it travels through different mediums and terrestrial objects. The copies of the transmitted signal arrive at the receiver at different times (due to different path lengths), resulting in constructive or destructive interferences. Different statistical models, for example, Rayleigh, Rician, Nakagami, and Weibull fading, have been proposed to characterize the effect of multipath fading under distinct propagation environments. Rayleigh fading is acknowledged as a reasonable model propagation environments consisting a large number of scattering objects [53],
such as a dense urban environment, whereas, the propagation inside a tunnel is better described by Rician fading [51]. We assume that the received signal of an MS at open areas, e.g. above ground, follows Rayleigh fading and the received signal of an MS at closed areas, e.g. underground or tunnels, follows Rician fading. We will use Rayleigh fading in later analytical derivations for its tractability. The amplitude of the received signal over a Rayleigh fading channel is distributed as

\[ f_{V_r}(v) = \frac{2v}{P_r} e^{-\frac{v^2}{P_r}}, \quad v \geq 0, \]

where \( P_r \) is the average received power based on path loss (and shadowing). The impact on the power of the received signal is obtained by transforming the variable \( p = v^2 \) which results in

\[ f_{P_r}(p) = \frac{1}{P_r} e^{-\frac{p}{P_r}}, \quad p \geq 0, \]

which becomes an exponential distribution with parameter \( \lambda = \frac{1}{P_r} \). The expression in dB scale is obtained by applying another variable transformation, \( x = 10 \log_{10} p \), to convert \( P_r \) into \( P_r \text{dB} \) as

\[ f_{P_r \text{dB}}(x) = \int_{-\infty}^{\infty} \frac{1}{x^2} e^{-\frac{10}{x} \frac{P_r \text{dB}}{10}} e^{-\frac{10}{x} \frac{P_r \text{dB}}{10}} \cdot 20 \left( \frac{\ln 10}{\sqrt{2\pi} \sigma_S} \right)^2 \frac{e^{-\frac{10}{x} \frac{P_r \text{dB}}{10} \ln 10}}{2\sigma_S^2} \text{d}w, \]

(5.7)

If the shadowing component is absent, then \( P_r \text{dB} \) only comes from path loss and is \( k - 20 \log_{10} d \). Otherwise, as shown in [46], the amplitude or the power is expressed as a conditional density conditioned on shadowing as

\[ f_{V_r}(v) = \int_0^\infty f_{V_r|V}(v|w)f_{V}(w)\text{d}w = \int_0^\infty \frac{\pi v}{2w^2} e^{-\frac{\pi v^2}{4w^2}} \frac{20}{\ln 10 \sqrt{2\pi} \sigma_S w} e^{-\frac{(10 \log_{10} w - (k - 10 \gamma \log_{10} d))^2}{2\sigma_S^2}} \text{d}w, \]

for amplitude (which is also called the Suzuki distribution), and as

\[ f_{P_r}(p) = \int_0^\infty f_{P_r|P}(p|w)f_{P}(w)\text{d}w = \int_0^\infty \frac{1}{w} e^{-\frac{p}{w}} \frac{10}{\ln 10 \sqrt{2\pi} \sigma_S w} e^{-\frac{(10 \log_{10} w - (k - 10 \gamma \log_{10} d))^2}{2\sigma_S^2}} \text{d}w, \]

for power. Converting into dB scale, the received power under composite shadowing and Rayleigh fading yields

\[ f_{P_r \text{dB}}(x) = \int_{-\infty}^{\infty} \ln 10 \frac{10}{w} e^{-\frac{10}{w} \frac{P_r \text{dB}}{10}} \cdot 1 \left( \frac{\ln 10}{\sqrt{2\pi} \sigma_S} \right)^2 \frac{e^{-\frac{10}{w} \frac{P_r \text{dB}}{10} \ln 10}}{2\sigma_S^2} \text{d}w. \]

(5.8)

5.1.2 Numerical Results

Under the assumption of identical fading conditions within a segment, for example, equal \( \sigma_S \) value for shadowing, we use Equation 5.1 and derive restrictions on the maximum segment size under the cases of shadowing, Rayleigh fading, and their composite. Should the fading conditions vary within a segment in terms of distribution parameters or even the shape of the distribution, the expressions of \( q_{p,d_1} \) and \( q_{p,d_2} \) are required to change accordingly.
Our road-map to derive the maximum segment size can be summarized as follows:

1. Find an expression of $q_{p,d}$ by computing its CDF (integrating Equation 5.6 for shadowing, Equation 5.7 for Rayleigh fading with $\mathcal{P}_{\text{dB}} = k - 10\gamma \log_{10} d$, and Equation 5.8 for composite shadowing and Rayleigh fading, with respect to $x$) at $d$, then find its inverse or the quantile function and evaluate it at $p$.

2. Assume $d_2 > d_1$. Set $d_1$ to the value such that $q_{p,d_1}$ is equal to the coverage threshold, $T_c$, compute $d_2$ such that $|q_{p,l_1} - q_{p,l_2}| = \epsilon$, and record $|d_1 - d_2|$ as $s_{\text{max},1}$. Then, set $d_2$ to the value such that $q_{p,d_2}$ is equal to the coverage threshold, $T_c$, compute $d_1$ such that $|q_{p,l_1} - q_{p,l_2}| = \epsilon$, and record $|d_1 - d_2|$ as $s_{\text{max},2}$. The maximum segment size, denoted by $s_{\text{max}}$, is the more stringent result, i.e., $s_{\text{max}} = \min\{s_{\text{max},1}, s_{\text{max},2}\}$.

In our numerical computations, we set $\epsilon$ to reference values of 1 and 2 dB. The selected values for $p$ are 0.05, 0.1, and 0.15 as the coverage probability $p_c$ should be a probability close to 1, and, therefore, the $1 - p_c$ quantile should be close to 0. The coverage threshold $T_c$ after mapping to RSS is set to $-70 \pm 1, 2 \text{ dBm}$. We suppose $T_c$ corresponds to a QoS that can supply real-time train control and voice communication traffic, where the earlier requires 50 kbps [54] and the latter consumes 647–916 kbps [55]. The total data rate is thus about 1 Mbps. The conversion from data rate to RSS is exemplified using 802.11b where a 4 dB surplus is required to support a data rate of 1 Mbps (7 dB for 2 Mbps, 11 dB for 5.5 Mbps and 16 dB for 11 Mbps) on top of the larger of the two: the radio sensitivity level, which is the minimum power for a signal to be detected, or the background noise floor, where the radio sensitivity level is usually the dominating one [56]. The radio sensitivity level is obtained from a major manufacturer as $-84 \text{ dBm}$ [57]. Consequently, a minimum RSS of $-80 \text{ dBm}$ is required to support 1 Mbps ($-77 \text{ dBm}$ for 2 Mbps, $-73 \text{ dBm}$ for 5.5 Mbps, and $-68 \text{ dBm}$ for 11 Mbps). Thus, setting $T_c$ to around $-70 \text{ dBm}$ suffices to afford the aggregate train control and voice traffic.

5.1.2.1 Shadowing

Using the normal distribution, the quantile function of Equation 5.6 can be readily obtained as

$$q_{p,d} = F_{\mathcal{P}_{\text{dB}}}^{-1}(p) = k - 10\gamma \log_{10} d + \sigma S \sqrt{2} \text{erf}^{-1}(2p - 1),$$

(5.9)

where erf($\cdot$) is the error function with $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, and erf$^{-1}(\cdot)$ is the inverse error function.

To obtain $s_{\text{max},1}$ we first set $q_{p,d_1} = T_c$ and obtain

$$d_1 = 10 \frac{k - T_c + \sqrt{\sigma S} \text{erf}^{-1}(2p - 1)}{10\gamma}.$$ 

(5.10)
Table 5.2: Maximum segment size (m) under shadowing with $\sigma_S = 4$ dB

<table>
<thead>
<tr>
<th>$T_c$ (dBm)</th>
<th>$\epsilon = 1$ dB</th>
<th>$\epsilon = 2$ dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 0.05$</td>
<td>$p = 0.1$</td>
</tr>
<tr>
<td>-68</td>
<td>12.7</td>
<td>15.1</td>
</tr>
<tr>
<td>-69</td>
<td>14.3</td>
<td>16.9</td>
</tr>
<tr>
<td>-70</td>
<td>16.0</td>
<td>19.0</td>
</tr>
<tr>
<td>-71</td>
<td>18.0</td>
<td>21.3</td>
</tr>
<tr>
<td>-72</td>
<td>20.2</td>
<td>23.9</td>
</tr>
</tbody>
</table>

Next, we substitute Equation 5.9 into $|q_{p,d_1} - q_{p,d_2}| = \epsilon$, which yields

$$|q_{p,d_1} - q_{p,d_2}| = 10\gamma \log_{10} \left| \frac{d_2}{d_1} \right| = \epsilon. \quad (5.11)$$

With the assumption $d_2 > d_1$, Equation 5.11 imposes a restriction on the relationship between $d_1$ and $d_2$ as

$$\frac{d_2}{d_1} = 10^{\frac{\epsilon}{10\gamma}}. \quad (5.12)$$

Then we combine Equation 5.10 and 5.12 to obtain

$$s_{max,1} = |d_1 - d_2| = 10^{\frac{k - T_c + \sqrt{2\sigma_S \text{erf}}^{-1}(2p-1)}{10\gamma}} (10^{\frac{\epsilon}{10\gamma}} - 1). \quad (5.13)$$

To obtain $s_{max,1}$ we set $q_{p,d_2} = T_c$ and obtain

$$d_2 = 10^{\frac{k - T_c + \sqrt{2\sigma_S \text{erf}}^{-1}(2p-1)}{10\gamma}}. \quad (5.14)$$

Then, we combine 5.12 and Equation 5.14 to get

$$s_{max,2} = |d_1 - d_2| = 10^{\frac{k - T_c + \sqrt{2\sigma_S \text{erf}}^{-1}(2p-1)}{10\gamma}} (1 - 10^{-\frac{\epsilon}{10\gamma}}). \quad (5.15)$$

Comparing the right hand sides of Equation 5.10 and 5.14, we conclude $s_{max} = s_{max,2}$ as $s_{max,2}$ is always smaller and becomes the limiting condition on the maximum segment size.

Finally, we compute numerical results for different values of $\epsilon$, $p$, $T_c$, and $\sigma_S = 4$ dB as listed in Table 5.2 where the maximum segment size increases in $\epsilon$, increases in $p$, and decreases in $T_c$. The table should be interpreted as follows: The segment size given in the table for $\epsilon = 1$ dB, $T_c = -70$ dBm, and $p = 0.05$, the segment size with a coverage probability of 0.95 and a coverage threshold of -70 dBm, should be less than 16.0 m. The maximum segment size guarantees the 0.05th quantiles of the RSS at any two points within the segment where the score crosses $T_c$ is always less than 1 dB.
Table 5.3: Maximum segment size (m) under Rayleigh fading

<table>
<thead>
<tr>
<th>$T_c$ (dBm)</th>
<th>$\epsilon = 1$ dB $p = 0.05$ $p = 0.1$ $p = 0.15$</th>
<th>$\epsilon = 2$ dB $p = 0.05$ $p = 0.1$ $p = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-68</td>
<td>6.2 8.8 11.0</td>
<td>11.6 16.7 20.7</td>
</tr>
<tr>
<td>-69</td>
<td>6.9 9.9 12.3</td>
<td>13.1 18.7 23.2</td>
</tr>
<tr>
<td>-70</td>
<td>7.7 11.1 13.8</td>
<td>14.7 21.0 26.1</td>
</tr>
<tr>
<td>-71</td>
<td>8.7 12.5 15.5</td>
<td>16.4 23.6 29.3</td>
</tr>
<tr>
<td>-72</td>
<td>9.8 14.0 17.4</td>
<td>18.4 26.4 32.8</td>
</tr>
</tbody>
</table>

5.1.2.2 Rayleigh Fading

After integrating Equation 5.7, the CDF of $P_{\text{dB}}$ is given by

$$F_{P_{\text{dB}}}(x) = 1 - e^{-10 \frac{x-(k-10\gamma \log_{10} d)}{10\gamma}}, \quad x \in \mathbb{R},$$

and the quantile function is

$$q_{p,d} = F_{P_{\text{dB}}}^{-1}(p) = 10 \log_{10}(\ln \frac{1}{1-p}) + k - 10\gamma \log_{10} d.$$ (5.16)

Following the same approach, we obtain $s_{\text{max},1}$ and $s_{\text{max},2}$ where

$$s_{\text{max},1} = 10^{k-T_c+10\log_{10}(\ln \frac{1}{1-p})} (10^{\frac{k}{10\gamma}} - 1),$$ (5.17)

and

$$s_{\text{max},2} = 10^{k-T_c+10\log_{10}(\ln \frac{1}{1-p})} (1 - 10^{\frac{k}{10\gamma}}).$$ (5.18)

As the right hand sides of Equation 5.17 is always greater than the right hand side of Equation 5.18, we again conclude $s_{\text{max}} = s_{\text{max},2}$. Numerical results for selected values of $\epsilon$, $p$, and $T_c$ can be found in Table 5.3.

5.1.2.3 Composite Shadowing and Rayleigh Fading

The CDF of the received power in dB is found by integrating Equation 5.8 with respect to $x$, which gives

$$F_{P_{\text{dB}}}(x) = \int_{-\infty}^{\infty} (1 - e^{-10 \frac{x-w}{10\gamma}}) \times \frac{1}{\sqrt{2\pi\sigma_S}} e^{-\frac{(w-(k-10\gamma \log_{10} d))^2}{2\sigma_S^2}} dw.$$ (5.19)

While an analytical expression is not available, we compute the CDF and its inverse numerically for different values of $d$ until an $d_1$, such that $q_{p,d_1} \approx T_c$ for a specific $p$ and $T_c$ is found. Then, with the assumption $d_2 > d_1$, we find $d_2$ such that $|q_{p,d_2} - q_{p,d_1}| \approx \epsilon$ and obtain $s_{\text{max},1}$. Similarly,
Table 5.4: Maximum segment size (m) under composite shadowing with $\sigma_S = 4$ dB and Rayleigh fading

<table>
<thead>
<tr>
<th>$T_c$ (dBm)</th>
<th>$\epsilon = 1$ dB</th>
<th>$\epsilon = 2$ dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 0.05$</td>
<td>$p = 0.1$</td>
</tr>
<tr>
<td>-68</td>
<td>5.0</td>
<td>7.3</td>
</tr>
<tr>
<td>-69</td>
<td>5.7</td>
<td>8.3</td>
</tr>
<tr>
<td>-70</td>
<td>6.3</td>
<td>9.2</td>
</tr>
<tr>
<td>-71</td>
<td>7.1</td>
<td>10.3</td>
</tr>
<tr>
<td>-72</td>
<td>8.0</td>
<td>11.7</td>
</tr>
</tbody>
</table>

$s_{max}^{1,2}$ is found in the same way where the expression $q_{p,d_1} \approx T_c$ in the first step is changed to $q_{p,d_2} \approx T_c$. Finally, the maximum segment size is $s_{max} = \min\{s_{max}^{1}, s_{max}^{2}\}$. The results of $s_{max}$ for $\sigma_S = 4$ dB and specified values of $\epsilon, p$ and $T_c$ are enumerated in Table 5.4.

In this section, we explored the maximum segment size under different wireless propagation models. Specifically, for the given parameter values $p = 0.05$, $T_c = -70$ dBm, and $\epsilon = 1$ dB, the maximum segment size is 16.0 m, 7.7 m, and 6.3 m for shadowing, Rayleigh fading, and composite shadowing and Rayleigh fading, respectively. Although these values seem small, they do not hinder the practicability of our algorithm as the precise point positioning service provided by Global Navigation Satellite System (GNSS) nowadays is capable to support centimeter-level precision. Where GNSS may not be available, such as in a tunnel, an MS can obtain its location with speedometer or odometer and calibrate the reading once GNSS service resumes.

### 5.2 Minimum Segment Size

A second constraint of the segment size results from the number of measurements or samples of the score of a PoA inside a cell in the score database. Recall in Chapter 4 that a cell is the intersection of a temporal section and a spatial section. Since we have fixed the size of temporal sections, the number of measurements per cell is only affected by the segment size. If the number of samples, or sample size, of a segment is too small, the population quantiles may be poorly estimated by the samples. Particularly, we are interested in two different population quantiles, the $(1 - p_c)\text{th}$ quantile as used in estimating coverage areas and the $p_{uo\text{th}}$ quantile as used in identifying unpredictable events. We denote by $n$ the required amount of samples of the score of a PoA per segment such that the population parameters can be estimated with a high confidence level. Then $n$ can restrict the lower bound of segment size as follows. Suppose that the number of samples of the score of a PoA within a temporal section per meter is $\lambda_n$, referred to as the sample density, where $\lambda_n$ is uniform for all PoAs, all temporal sections and
at all locations. Then the segment size in meters, \( s \), must be large enough such that
\[
\lambda_n s \geq n. \tag{5.20}
\]
Equation 5.20 yields the minimum segment size, denoted by \( s_{\text{min}} \), when the equality holds, i.e.,
\[
s_{\text{min}} = \frac{n}{\lambda_n}. \tag{5.21}
\]
To determine \( n \), we adopt the confidence interval concept to estimate the population quantile, which is explained in the following section.

### 5.2.1 Estimating Population Quantiles with Confidence Interval

Denote the random variable of interest by \( X \), the CDF of \( X \) by \( F_X(x) \), and the \( p^{\text{th}} \) \((0 \leq p \leq 1)\) quantile of \( X \) by \( q_p \). If \( X \) is continuous, then \( q_p = F_X^{-1}(p) \); if \( X \) is discrete, then \( q_p = \min\{q : F_X(q) \geq p\} \). The point estimator of \( q_p \) is the \( p^{\text{th}} \) sample quantile, denoted by \( Q_n^p \), where \( Q_n^p = F_n^{-1}(p) \) and \( F_n(x) \) is the empirical distribution function of \( X \) from a sample size of \( n \). \( F_n(x) \) is obtained by
\[
F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1(X_j \leq x),
\]
where \( X_j \) is the \( j^{\text{th}} \) sample of \( X \).

In [58], Bahadur proved that, by the Central Limit Theorem (CLT), the \( p^{\text{th}} \) sample quantile, \( Q_n^p \), is asymptotically normally distributed with mean \( q_p \) and variance \( \sigma_p^2 \) where
\[
\sigma_p^2 = \frac{p(1-p)}{f_X'(q_p)}, \tag{5.22}
\]
under the condition that \( F_X'(x) \) is bounded near \( q_p \) and \( F_X'(q_p) = f_X(q_p) > 0 \).

We adopt confidence interval [59] to estimate the \( p^{\text{th}} \) population quantile. A confidence interval is an interval estimate calculated from a sample of an experiment and is different if the experiment is to be repeated. This means that, suppose that an experiment is repeated, the confidence intervals derived from some experiments contain the true value of the \( p^{\text{th}} \) population quantile, while the confidence intervals of the other experiments do not. The frequency that the confidence intervals resulted from the repeated experiments contain the true value follows a parameter called the confidence level, \( 1 - \alpha \), usually expressed as a percentage value, \((1-\alpha)100\%\), where \( \alpha \) is referred to as the significance level. The typical values of \( 1 - \alpha \) in the field of statistics are 0.95 and 0.99. Since \( Q_n^p \) is normally distributed, the \((1 - \alpha)100\%\) confidence interval for \( q_p \) is
\[
[Q_n^p - \frac{z_{\alpha/2} \sigma_p}{\sqrt{n}}, Q_n^p + \frac{z_{\alpha/2} \sigma_p}{\sqrt{n}}],
\]
where \( z_{\alpha/2} \) is given by \( P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha \) and \( Z \sim \mathcal{N}(0,1) \). Expressing the confidence
interval by $Q_p^n \pm E$ where $E$ is called the margin of error and $E = \frac{z_{\alpha} \sigma_p}{\sqrt{n}}$, the sample size, $n$, can be obtained as

$$n \geq \left( \frac{z_{\alpha} \sigma_p}{E} \right)^2,$$

after some target value of $E$ is specified.

### 5.2.2 Numerical Results

Although the distribution of the random variable representing the score of a PoA measured by an MS at a segment is typically unknown, we again argue that the randomness results from the underlying wireless propagation system. We will illustrate our derivation of the required sample size per segment and the minimum segment size given values of $\lambda_n$ under the cases of log-normal shadowing, Rayleigh fading and composite shadowing and Rayleigh fading. Our road-map is as below:

1. Find an expression of $q_{p,d}$ by computing its CDF (integrating Equation 5.6 for shadowing, Equation 5.7 for Rayleigh fading with $P_{\text{dB}} = k - 20 \log_{10} d$, and Equation 5.8 for composite shadowing and Rayleigh fading, with respect to $x$) at $d$ then find its inverse or the quantile function and evaluate it at $p$.

2. Utilize Equation 5.22 to obtain $\sigma_p$ by substituting $q_{p,d}$ into $f_X(\cdot)$.

3. Utilize Equation 5.23 to obtain $n$ by selecting a value of $\alpha$ and $E$.

4. Utilize Equation 5.21 to obtain $s_{\text{min}}$ from $n$ and a given value of $\lambda_n$.

For numerical computations, we set $p$ to $1 - p_c$ and $p_{u_0}$ where $p_c$ is set to 0.85, 0.9, and 0.95 as before, and $p_{u_0}$ is a small probability set to 0.01 in our numerical examples. We set $1 - \alpha$ to 0.95 as is commonly done in the field of statistics. The margin of error, $E$, is set to a reference value of 1 dB. The sample density, $\lambda_n$, is set to 1000 samples per meter\(^1\).

#### 5.2.2.1 Shadowing

The PDF and the quantile function of the received power in dB scale under shadowing have been derived in Equation 5.6 and 5.9. Substituting the latter into the former yields

$$f_{P_{\text{dB}}} (p) = \frac{1}{\sqrt{2\pi} \sigma_{S_{\text{dB}}}} e^{-\left(\text{erf}^{-1}(2p-1))^2\right)} , \quad p \in [0, 1].$$

Following Equations 5.22 and 5.23, the expression of the required sample size is readily available. Numerical results of $n$ for $\sigma_S = 4, 6, 8, 10, 12$ dB and selected values of $p$, $1 - \alpha$ and $E$ are listed

\[\text{Suppose that an MS travels at a constant speed of 36 km/hr and each adapter generates a scan result at a constant rate of 5 Hz. Each MS contributes a sample density of 0.5 samples per meter after a trip. To achieve a sample density of 1000 samples per meter, a number of 2000 MS for each temporal section needs to be dispatched during the data collection phase.}\]
Table 5.5: Minimum sample size per segment and minimum segment size for different \( p \) and \( \sigma_S \) under shadowing with \( \sigma_S = 4 \) dB

(a) Minimum sample size per segment

<table>
<thead>
<tr>
<th>( \sigma_S ) (dB)</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>857</td>
<td>275</td>
<td>180</td>
<td>145</td>
</tr>
<tr>
<td>6</td>
<td>1928</td>
<td>618</td>
<td>405</td>
<td>325</td>
</tr>
<tr>
<td>8</td>
<td>3427</td>
<td>1098</td>
<td>719</td>
<td>577</td>
</tr>
<tr>
<td>10</td>
<td>5354</td>
<td>1716</td>
<td>1123</td>
<td>901</td>
</tr>
<tr>
<td>12</td>
<td>7710</td>
<td>2471</td>
<td>1617</td>
<td>1298</td>
</tr>
</tbody>
</table>

(b) Minimum segment size (m)

<table>
<thead>
<tr>
<th>( \sigma_S ) (dB)</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.857</td>
<td>0.275</td>
<td>0.180</td>
<td>0.145</td>
</tr>
<tr>
<td>6</td>
<td>1.928</td>
<td>0.618</td>
<td>0.405</td>
<td>0.325</td>
</tr>
<tr>
<td>8</td>
<td>3.427</td>
<td>1.098</td>
<td>0.719</td>
<td>0.577</td>
</tr>
<tr>
<td>10</td>
<td>5.354</td>
<td>1.716</td>
<td>1.123</td>
<td>0.901</td>
</tr>
<tr>
<td>12</td>
<td>7.710</td>
<td>2.471</td>
<td>1.617</td>
<td>1.298</td>
</tr>
</tbody>
</table>

Table 5.6: Minimum sample size per segment and minimum segment size for different \( p \) under Rayleigh fading

(a) Minimum sample size per segment

<table>
<thead>
<tr>
<th>( p )</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>7246</td>
<td>1450</td>
<td>726</td>
<td>485</td>
<td></td>
</tr>
</tbody>
</table>

(b) Minimum segment size (m)

<table>
<thead>
<tr>
<th>( p )</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.246</td>
<td>1.450</td>
<td>0.726</td>
<td>0.485</td>
<td></td>
</tr>
</tbody>
</table>

in Table 5.5a. An entry in the table is the minimum number of samples required to estimate the \( p^{th} \) quantile of \( S_d \) with a 95% confidence interval of length 2 dB (\( \pm E \)) where \( S_d \) is subject to log-normal shadowing with a shadow standard deviation of \( \sigma_S \) dB. For example, for \( \sigma_S = 4 \) dB and \( p = 0.01 \), the minimum number of samples required for each segment is 857. Moreover, the minimum sample size per segment decreases as \( p \) increases for the considered values.

5.2.2.2 Rayleigh Fading

The PDF and the quantile function of the received power in dB scale under Rayleigh fading have been obtained previously and are expressed in Equation 5.7 and 5.16, respectively. After substitution, the PDF is

\[
f_{P_{rdB}}(p) = -\frac{\ln 10}{10}(1-p)\ln(1-p), \quad 0 \leq p \leq 1. \tag{5.25}
\]

Following Equations 5.22 and 5.23, the sample size per segment can be expressed analytically where numerical results of \( n \) are provided in Table 5.6a. Similar to the case of shadowing, the sample size per segment also decreases as \( p \) increases for the considered values.
Table 5.7: Minimum sample size per segment and minimum segment size for different $p$ and $\sigma_S$ under composite shadowing with $\sigma_S = 4$ dB and Rayleigh fading

<table>
<thead>
<tr>
<th>$\sigma_S$ (dB)</th>
<th>$p$</th>
<th>$p$</th>
<th>$p$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.05</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>7287</td>
<td>1528</td>
<td>815</td>
<td>578</td>
</tr>
<tr>
<td>6</td>
<td>7665</td>
<td>1759</td>
<td>985</td>
<td>726</td>
</tr>
<tr>
<td>8</td>
<td>8385</td>
<td>2119</td>
<td>1258</td>
<td>961</td>
</tr>
<tr>
<td>10</td>
<td>9598</td>
<td>2662</td>
<td>1641</td>
<td>1271</td>
</tr>
<tr>
<td>12</td>
<td>11586</td>
<td>3348</td>
<td>2100</td>
<td>1663</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\sigma_S$ (dB)</th>
<th>$\sigma_S$ (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.05</td>
</tr>
<tr>
<td>4.0</td>
<td>7.287</td>
</tr>
<tr>
<td>6.0</td>
<td>7.665</td>
</tr>
<tr>
<td>8.0</td>
<td>8.385</td>
</tr>
<tr>
<td>10.0</td>
<td>9.598</td>
</tr>
<tr>
<td>12.0</td>
<td>11.586</td>
</tr>
</tbody>
</table>

5.2.2.3 Composite Shadowing and Rayleigh Fading

The PDF of the received power in dB scale under composite shadowing and Rayleigh fading is illustrated in Equation 5.8, where the quantile function is computed numerically. For $p = 0.01$, 0.05, 0.1, and 0.15, the computed values of $f_{P_{\text{dB}}}(p)$ are 0.00227, 0.0108, 0.0205, and 0.0292. Then utilizing Equation 5.22 and 5.23 we obtain the sample sizes for different values of $\sigma_S$, $p$, $1 - \alpha$, and $E$, which are listed in Table 5.7a. Comparing Table 5.5a, 5.6a, and 5.7a, we observe that the required sample size under composite shadowing and Rayleigh fading is always the largest among the three cases. When the standard deviation of shadowing is relatively small, say 4 dB, the sample size under composite shadowing and Rayleigh fading is close to that under Rayleigh fading, whereas when the standard deviation of shadowing is large, say 12 dB, the sample size under composite shadowing and Rayleigh fading is closer to that under shadowing. Furthermore, the sample size per segment also decreases in $p$ as in the previous two cases.

5.3 Adaptive Segmentation

We seek to investigate the dependency between parameters $s$, $p$, $\epsilon$, $T_c$, $\lambda_n$, $1 - \alpha$, and $E$ to determine (1) adequate values of $s$, (2) the minimum value of $\epsilon$, or (3) the maximum value of $1 - \alpha$ under scenarios of different constraints as the size of the score database increases, where $p$, $T_c$, and $E$ are treated as fixed system parameters and $\lambda_n$ is one of the limiting conditions. Firstly, we evaluate the range of feasible values of $s$ given the sample density of the current score database, denoted by $\lambda^*$, and some target values of $\epsilon$ and $1 - \alpha$. Secondly, we explore the minimum value of $\epsilon$ under $\lambda^*$ when an objective value of $1 - \alpha$ is desired. Thirdly, we compute the maximum confidence level of the confidence interval under $\lambda^*$ for a target value of $\epsilon$. 
5.3.1 Numerical Results

To calculate the range of feasible segment sizes, we collect the derivations of maximum and minimum segment sizes from the previous sections for fixed system parameters, objective values $\epsilon$ and $1 - \alpha$, and $\lambda^*$. To determine the minimum value of $\epsilon$, we invert the derivation of the maximum segment size to obtain an expression of $\epsilon$ as a function of $s$, where $s$ is substituted using the minimum segment size that can be attained for given parameters. To compute the maximum confidence level $1 - \alpha$, we invert the derivation of the minimum segment size to obtain an expression of $1 - \alpha$ as a function of $s$, where $s$ is substituted using the maximum segment size that can be achieved for given parameters. For a clear presentation, we first derive $\epsilon$ and $1 - \alpha$ as functions of $s$, where analytical expressions are available, and provide an illustration showing $\epsilon$ and $1 - \alpha$ which are jointly parametrized by $s$. The three studies of the sensitivity among parameters are evaluated in the case of shadowing.

For numerical computations, the system parameters are set as follows: for the maximum segment size, we set $p = 1 - p_c = 0.05$ and $T_c = -70$ dBm; for the minimum segment size, we set $p = \min(1 - p_c, p_u) = p_u = 0.01$ and $E = 1$ dB. The objective value of $\epsilon$ in the first and third studies is 2 dB and the objective value of $1 - \alpha$ in the first and third usage is 0.95.

5.3.1.1 Shadowing

Solving for $\epsilon$ in Equation 5.15 yields

$$\epsilon = -10\gamma \log_{10} \left( 1 - 10^{-\frac{k - T_c + \sqrt{2} \sigma_S \text{erf}^{-1}(2p - 1)}{\log_{10} s_{\text{max}}}} \right).$$

Secondly, expressing $1 - \alpha$ as a function of $s$ from Equation 5.21 with substitutions of Equations 5.22, 5.23, and 5.24 gives

$$1 - \alpha = \text{erf} \left( \sqrt{\frac{\lambda_n s_{\text{min}}}{2p(1 - p)}} \frac{E}{\sqrt{2\pi \sigma_S}} e^{-\left(\text{erf}^{-1}(2p - 1))^2\right)} \right).$$

Figure 5.1 illustrates Equation 5.26 (blue) and Equation 5.27 (green and red) with $\sigma_S = 4$ dB. We plot Equation 5.27 for two values of $\lambda_n$, where the green curve indicates the result when $\lambda_n = 53.6$ and the red curve indicates the result when $\lambda_n = 28.3$. These two values result from equating $s_{\text{max}}$ in Equation 5.26 and $s_{\text{min}}$ in Equation 5.27 when $1 - \alpha$ is set to 0.95, and $\epsilon$ is set to 1 dB (green) and 2 dB (red), respectively. In other words, $\lambda_n = 53.6$ is the minimum value such that $\epsilon = 1$ dB and $1 - \alpha = 0.95$ are achieved simultaneously, and $\lambda_n = 28.3$ is the minimum value for $\epsilon = 2$ dB and $1 - \alpha = 0.95$.

Calculating segment sizes: In Figure 5.2 we illustrate how to obtain the range of feasible segment sizes. We plot two curves in Figure 5.2, where the blue curve indicates $\epsilon$ as a function of $s$ and is the identical blue curve in Figure 5.1, and the green curve indicates the confidence
level $1 - \alpha$ for a given value of $\lambda^*$. We impose two constraints, $\epsilon \leq 2$ and $1 - \alpha \geq 0.95$. The constraint $\epsilon \leq 2$ restricts the maximum segment size, $s^\text{max}$, to 30.3 m, as shown by the bottom dashed line in Figure 5.2, and the constraint $1 - \alpha \geq 0.95$ restricts the minimum segment size, $s^\text{min}$, to $s^*$, as shown by the top dashed line. If $s^* > 30.3$ m, there is no feasible solution of $s$ that satisfies the constraints simultaneously, i.e., the score database does not contain sufficient samples; if $s^* \leq 30.3$, any segment size within the range $[s^*, 30.3]$ is a feasible value. For example, in Figure 5.2 we set $\lambda^* = 39.0$, which results in $s^* = 22.0$ m.

**Calculating minimum $\epsilon$:** In Figure 5.3 we illustrate how to obtain the smallest feasible value of $\epsilon$ for a target value of $1 - \alpha$ and a given value of $\lambda^*$. The two curves for $\epsilon$ and $1 - \alpha$ in Figure 5.3 are identical to those in Figure 5.2. Given the constraint $1 - \alpha \geq 0.95$ and $\lambda^*$, the minimum segment size is $s^*$ which is indicated by the dashed line in Figure 5.3. The constraint $s \geq s^*$ in turn imposes the restriction $\epsilon \geq \epsilon^*$ by Equation 5.26, which is illustrated by the dotted lines in Figure 5.3. With $\lambda^* = 39.0$, $s^*$ results in 22.0 m and, therefore, $\epsilon^*$ results in 1.40 dB.

**Calculating maximum $1 - \alpha$:** In Figure 5.4 we demonstrate how to obtain the largest feasible value of $1 - \alpha$ for a target value of $\epsilon$ and a given value of $\lambda^*$. The two curves for $\epsilon$ and $1 - \alpha$ in Figure 5.4 are identical to those in Figure 5.2. We impose a constraint $\epsilon = 2$, which restricts $s$ to be less than 30.3 m as indicated by the dashed line in Figure 5.4. The constraint $s \leq 30.3$ m then restricts the maximum confidence level to $(1 - \alpha)^*$ by Equation 5.27, which is indicated by the dotted lines. With $\lambda^* = 39.0$ and $\epsilon = 2$, $(1 - \alpha)^*$ results in 0.98.
Figure 5.2: Using $\epsilon$ and $1 - \alpha$ as functions of $s$ to calculate the range of feasible values of $s$ under shadowing with $\sigma_S = 4$ dB.

Figure 5.3: Using $\epsilon$ and $1 - \alpha$ as functions of $s$ to calculate the minimum feasible value of $\epsilon$ under shadowing with $\sigma_S = 4$ dB.
5.3.1.2 Rayleigh Fading

Expressing the relationship given in Equation 5.18 in terms of $\epsilon$ results in

$$
\epsilon = -10\gamma \log_{10}(1 - 10^{-\frac{k - T_c + 10 \log_{10} (\ln 10)}{10\gamma} s_{\text{max}}}.
$$

(5.28)

On the other hand, expressing $1 - \alpha$ as a function of $s$ involves Equation 5.21 with substitutions of Equation 5.23, 5.22, and 5.25, which results in

$$
1 - \alpha = \text{erf}\left(\sqrt{\frac{\lambda_n s_{\text{min}}}{2p(1-p)}} E \times \left(-\frac{\ln 10}{10} (1-p) \ln (1-p)\right)\right).
$$

(5.29)

Figure 5.5 illustrates Equation 5.28 in blue and Equation 5.29 in green and red. The green curve corresponds to a sample density $\lambda_n = 941.0$ such that $s_{\text{max}}$ in Equation 5.28 is equal to $s_{\text{min}}$ in Equation 5.29 where $\epsilon$ is set to 1 dB and $1 - \alpha$ is set to 0.95. On the other hand, the red curve corresponds to a sample density $\lambda_n = 492.9$ for $\epsilon = 2$ dB and $1 - \alpha = 0.95$. Compared to Figure 5.1, Rayleigh fading always yields higher $\epsilon$ value and lower $1 - \alpha$ level than shadowing with $\sigma_S = 4$ dB for all $s$. This result has three implications: First, compared to Figure 5.2, for target values of $\epsilon$ and $1 - \alpha$, the value of $s_{\text{max}}$ given by Rayleigh fading is always smaller than the value of $s_{\text{max}}$ given by shadowing with $\sigma_S = 4$ dB, and the value of $s_{\text{min}}$ given by Rayleigh fading is always larger than the value of $s_{\text{min}}$ given by shadowing with $\sigma_S = 4$ dB, i.e., the range of feasible segment size is smaller. Second, compared to Figure 5.3, for an objective value of $1 - \alpha$ and a given value of $\lambda_e$, the minimum feasible value of $\epsilon$ given

Figure 5.4: Using $\epsilon$ and $1 - \alpha$ as functions of $s$ to calculate the maximum feasible value of $1 - \alpha$ under shadowing with $\sigma_S = 4$ dB.
by Rayleigh fading is always greater than the minimum feasible value of $\epsilon$ given by shadowing with $\sigma_S = 4$ dB. Third, compared to Figure 5.4, for an objective value of $\epsilon$ and a given value of $\lambda_*$, the maximum feasible value of $1 - \alpha$ given by Rayleigh fading is always less than the maximum feasible value of $1 - \alpha$ given by shadowing with $\sigma_S = 4$ dB.

5.3.1.3 Composite Shadowing and Rayleigh Fading

Since the analytical expression of Equation 5.1 for composite shadowing and Rayleigh fading cannot be obtained (specifically, Equation 5.19 cannot be solved analytically for the inverse function), we compute $\epsilon$ numerically using the following steps. First, for any value of $d$, we compute the cumulative probability of the received power under composite shadowing ($\sigma_S = 4$ dB) and Rayleigh fading at fixed values of $x$ using Equation 5.19. Second, we approximate $q_{1-p_c,d}$ by $q_{1-p_c,d} \approx x_{1-p_c}$, where $x_{1-p_c}$ is the value of $x$ whose cumulative probability is the closest to $1 - p_c$. Third, we repeat the previous two steps for different values of $d$ until a value is found that satisfies $q_{1-p_c,d} \approx T_c$. This value is denoted by $d_1$. Next, we repeat the first two steps to find $q_{1-p_c,d_2}$ where $d_2 = d_1 \pm s$. Finally, $\epsilon$ can be parametrized by $s$ as

$\epsilon = |q_{1-p_c,d_1} - q_{1-p_c,d_2}| = \max(|q_{1-p_c,d_1} - q_{1-p_c,d_1} + s|, |q_{1-p_c,d_1} - q_{1-p_c,d_1} - s|)$.

To solve for $1 - \alpha$ in terms of $s$, we require the value of the probability density function evaluated at the $p^{th}$ quantile which again has no analytical expression. Thus, we compute numerical values for $1 - \alpha$ with the following procedure. First, we obtain numerical value of $q_{p_{u0},d_1}$ by $q_{p_{u0},d_1} \approx x_{p_{u0}}$. Second, we obtain $f_X(q_{p_{u0}})$ by substituting $q_{p_{u0},d_1}$ for $x$ in Equation 5.8. Finally, inverting Equation 5.20 with substitutions of Equations 5.23 and 5.22, and the value of $f_X(q_{p_{u0}})$ computed in the previous step yields numerical results of $1 - \alpha$ for given values of $s$.

Figure 5.6 illustrates $\epsilon$ (blue) and $1 - \alpha$ (green and red). The green curve corresponds
Figure 5.6: $\epsilon$ and $1 - \alpha$ as functions of $s$ under composite shadowing with $\sigma_S = 4$ dB and Rayleigh fading to a value of $\lambda_n$ ($\lambda_n = 1156.7$) such that the minimum segment size constrained by $1 - \alpha = 0.95$ coincides with the maximum segment size constrained by $\epsilon = 1$ dB, and the red curve corresponds to a value of $\lambda_n$ ($\lambda_n = 607.2$) such that the minimum segment size constrained by $1 - \alpha = 0.95$ coincides with the maximum segment size constrained by $\epsilon = 2$ dB.

Compared to the case of shadowing and Rayleigh fading, composite shadowing and Rayleigh fading yields the highest $\epsilon$ value for any given $s$ value and the lowest $1 - \alpha$ level under same $\lambda_n$ and $s$ values. Hence, composite shadowing and Rayleigh fading yields the smallest range of feasible segment sizes, the largest minimum feasible value of $\epsilon$, and the smallest maximum feasible value of $1 - \alpha$. 


Chapter 6

Evaluation of Handoff Decision Algorithm

In this chapter we evaluate our handoff decision algorithm in two parts. In the first part, we evaluate the performance of the proposed algorithm (Experiment 1), where we compare our algorithm to two other handoff decision algorithms, one RSS-based and one location-based. We conduct the evaluation using two experimental measurement traces, one for the front and one for the rear, recorded by a real-life light rail train. The measurement traces contain the RSS of the 8 strongest PoAs in range of the train. Note that these measurement traces are used to construct the score database as well as the RSS measured by the MS. We also synthesize measurement traces with simulated RSS values where the traces for the RSS measured by the MS and the traces used to construct the score database are generated separately. We describe the experimental measurement traces followed by the simulated measurement traces in Section 6.2 and 6.3, respectively. In the second part, we evaluate the robustness of our algorithm under partial network failure (Experiment 2) and under partial network degradation (Experiment 3).

In both parts, the parameters of our algorithm are assigned in Table 6.1. The segment size $s$ is assigned based on the maximum segment size for $T = T_c = -70$ dBm, $\epsilon = 1$ dB, and $p = 1 - p_c = 0.05$, which is 16.0 m as given in Table 5.2. The sample density $\lambda_n$ is assigned based on the minimum sample size per segment for $p = p_{u0}$, which is 857, and, since $s = 10$ m, $\lambda_n$ must be greater than $857/10 = 85.7$ m$^{-1}$.

The performance of each algorithm is evaluated using the following metrics:

- **Ping-pong Effect**: Ping-pong effects refer to the phenomenon that an algorithm produces oscillating selections between PoAs. As MSs only travel in one direction, ping-pong effects indicate handoffs to earlier PoAs, which is generally undesirable.

- **Number of Handoffs**: This is the total number of handoffs executed over the entire travel. Since an MS is disconnected during a handoff execution, an algorithm with a small number of handoffs is preferred.
Table 6.1: Configuration parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage Probability ($p_c$)</td>
<td>0.95</td>
</tr>
<tr>
<td>Coverage Threshold ($T_c$)</td>
<td>-70 dBm</td>
</tr>
<tr>
<td>Per-Sample Unpredictable Event Probability ($p_{u0}$)</td>
<td>0.01</td>
</tr>
<tr>
<td>Number of Consecutive Samples to Identify an Unpredictable Event ($k_u$)</td>
<td>3 samples</td>
</tr>
<tr>
<td>Segment Size ($s$)</td>
<td>10 m</td>
</tr>
<tr>
<td>Sample Density ($\lambda_n$)</td>
<td>100 samples/m</td>
</tr>
</tbody>
</table>

- **Handoff Interval**: The handoff interval is the distance between two consecutive handoff executions. The distribution of handoff intervals reflects the ability of an algorithm to prevent frequent handoffs.

- **Service Interruption Duration**: A service interruption is experienced by an MS when the associated PoA is undetectable. As the most important objective to preserve continuous communication, any long service interruption duration is intolerable in evaluating the performance of an algorithm.

- **Low QoS Duration**: The time duration when an MS experiences a QoS lower than the target level of -70 dBm from the associated PoA. This metric measures the effectiveness of an algorithm to maintain satisfactory QoS at all times.

All evaluations are conducted using Matlab on a MacBook Pro equipped with a 2.7 GHz Intel Core i7 and a 8 GB DDR3 RAM and running OS X 10.9.5.

## 6.1 Handoff Decision Algorithms Used for Comparison

We compare the proposed algorithm, which we will refer to as the adaptive planning (AP) algorithm, to an RSS-based algorithm and a location-based algorithm.

### 6.1.1 Relative Threshold Algorithm

The relative threshold (RT) algorithm starts by selecting the PoA with the highest RSS and preserves the association with the PoA until the RSS of the PoA drops below a predefined threshold. Meanwhile, a moving average of 5 samples is employed to counter RSS fluctuations. The threshold level is set to the same target QoS as in the proposed algorithm, which is -70 dBm.
6.1.2 Location-based Algorithm

The location-based (LB) algorithm requires foreknowledge of the PoA locations as well as the real-time MS location. The front radio connects to the nearest PoA ahead, whereas the rear radio associates with the nearest PoA behind. In the case of more than one PoAs deployed at the same location, the algorithm selects the one with the highest RSS and remains associating with it until another PoA becomes the nearest. Note that this algorithm does not utilize the RSS measured by the MS.

6.2 Experimental Measurement Trace

We obtain two experimental measurement traces, one recorded by the front radio and the other recorded by the rear radio, of the same MS that moves on a 2.4 km-long train line. Along the train line, a set of PoAs is deployed and their placement is illustrated in Figure 6.1a. Each PoA is marked by a star and identified by a number as labelled. The distribution of the distance between PoAs is shown in Figure 6.1b. About 80% of the PoAs are located at a distance of 75 to 195 m apart.

Each measurement trace consists of a series of scan results. Each scan result has a timestamp, a location-stamp, a list of (up to 8) detected PoAs, and a list of corresponding RSS measurements. The scan results are generated every 0.2 to 0.25 seconds and are arranged in chronological order. We will use the RSS as the only handoff decision criterion. If a handoff is decided, we assume the handoff execution (authentication and re-association) is completed within 10 ms [39], well before the next scan result is generated. Information of each handoff, including time, location, and the PoA to hand off to, is recorded into a trace called the handoff information trace. The handoff performance is evaluated using the measurement trace and the handoff information trace.

From the timestamps and the location-stamps we recover the movement of the MS and the location of the MS as a function of time is shown in Figure 6.2. The location is expressed as the distance traveled on the track with respect to a selected origin. The plateaus in the figure indicate time periods with no motion of the MS. Note that the trace includes a long time period where the train is not in motion.

From the location-stamps, the lists of PoAs, and the lists of RSS measurements, we can plot the RSS measurement against the location of the MS for each PoA, which can be found in Appendix A. We emphasize again that these experimental measurement traces are used as the RSS measured by the MS as well as the source of the score database at the same time. The coverage areas of the PoAs computed using the experimental measurement traces are illustrated in Figure 6.3 where the parameters used in estimating the coverage area, $s$, $p_c$, and $T_c$, are listed in Table 6.1. The lengths of the coverage areas vary by a noticeable amount. For example, in Figure 6.3a a PoA (PoA 15) has a coverage area that is about 230 m long while another PoA (PoA 27) has a coverage area that is about 770 m long.
Chapter 6. Evaluation of Handoff Decision Algorithm

Figure 6.1: PoA deployment from an actual CBTC system

(a) PoA placement

(b) Distribution of distance between PoAs

Figure 6.2: Location of the MS as a function of time from the measurement trace
Figure 6.3: Coverage areas of PoAs obtained from experimental measurement traces.
6.3 Simulated Measurement Trace

We also synthesize measurement traces with simulated RSS values and generate separate traces for the RSS measured by the MS and the score database.

In the measurement traces for the RSS measured by the MS, we generate a series of scan results the timestamps of which are equally separated by 0.2 seconds. The timestamps are converted to locations using the movement record in Figure 6.2. The list of PoAs in each scan result includes all PoAs in the system. An RSS measurement is generated for each PoA using the path loss with shadowing model\(^1\) given in Equation 5.5, where the distance between the MS and a PoA is computed from the location-stamp and the location of the PoA. The path loss exponent is set to 2 and the shadow standard deviation is set to 4 dB. All PoAs are assumed to have identical radio hardware with parameters listed in Table 5.1. The receiver sensitivity of each PoA is assumed to be -84 dBm [57]. A PoA is undetectable if the RSS falls below the sensitivity of the MS. As an example, Figure 6.4 illustrates the simulated (blue) and the experimental (red) RSS of PoA 24, which is situated at the 1418 meter mark, measured by the front (left) and the rear (right) radios of the MS. The horizontal axis is the location on the track, expressed as the distance relative to the origin. The vertical axis is the RSS in dBm. The drops to -128 dBm indicate the incidents that the PoA is missing from a scan result. For the simulated data, the PoA is missing when the RSS is below the receiver sensitivity of the MS. For the experimental data, the PoA may sometimes disappear from the scan results even when the RSS is high. We suspect the cause to be the finite length of the list of PoAs (up to 8) such that a PoA with strong RSS may still be excluded from the list. In both the simulated and the experimental data, the RSS drops once the MS passes the PoA, as the antenna gain changes from the gain for the front lobe to the gain for the back lobe. The rear radio experiences the opposite effect of a sudden gain once it passes the PoA.

In the measurement traces for the score database, instead of generating scan results with equally spaced timestamps, we generate scan results with equally spaced location-stamps that are separated by \(\frac{1}{\lambda_n}\) m where \(\lambda_n\) is listed in Table 6.1. In this way, \(\lambda_n\) is uniform at all locations and is the same for all PoAs. Then, the RSS of each PoA in each scan result is simulated using the aforementioned path loss with shadowing model. The coverage area of each PoA using the simulated measurement traces is depicted in Figure 6.5 where the parameters \(s, p_c,\) and \(T_c\) are provided in Table 6.1. Most of the PoAs have an approximately 600 m long coverage area.

6.4 Experiment 1: Default Experiment

This experiment shows the handoff performance when there are no unpredictable events in the network. This experiment is conducted for the experimental traces and the simulated traces.

\(^1\)The path loss with shadowing model appears to be a better fit to our experimental data than the path loss with Rayleigh fading model and the path loss with shadowing and Rayleigh fading model.
6.4.1 Results of Experimental Data

We first evaluate the proposed handoff algorithm with the experimental data. In Figure 6.6, we show the handoff locations and the locations of the associated PoAs. First of all, we observe that RT exhibits ping-pong effects where it hands off to earlier PoAs. These events are highlighted by blue arrows. LB and AP do not exhibit ping-pong effects. At the front radio, AP has 7 handoff events, whereas RT has 12 such events and LB has 20. At the rear radio, AP has 8 handoff events, RT has 13 and LB has 19. The statistics of the handoff interval is provided in Figure 6.7 using histograms where AP shows the best performance in spacing out handoffs into longer and more even intervals.

Next, we evaluate the performances regarding service interruption and low QoS. Figure 6.8 illustrates the distribution of the durations of service interruption incidents where RT has a total of 6 short service interruption that last for at most 0.25 (front) and 0.4 (rear) seconds, whereas LB experiences a total of 6 (front) and 8 (rear) service interruption and most of them are excessively long. AP has no service interruption due to the score database\(^2\). Finally, Figure 6.9 depicts the distribution of low QoS durations, where all algorithms generate short durations with low QoS, where RT yields the largest amount of incidents compared to AP and LB.

Table 6.2 summarizes the results of the performance. AP demonstrates performances better than RT and LB in terms of ping-pong effects, number of handoffs, average handoff interval, service interruption, and low QoS duration.

6.4.2 Results of Simulated Data

We enrich the evaluation with the simulated data. Figure 6.10 depicts the location of the MS and the location of the selected PoA of each handoff event produced by RT, LB, and AP.

\(^2\)Recall that the measurement traces are used as the RSS measured by the MS as well as the score database for the experimental data.
Figure 6.5: Coverage areas of PoAs obtained from simulated measurement traces
Figure 6.6: Handoff locations and locations of the selected PoAs of the three algorithms using the experimental data

Figure 6.7: Distribution of handoff intervals of the three algorithms using the experimental data
Figure 6.8: Distribution of service interruption durations of the three algorithms using the experimental data

Figure 6.9: Distribution of low QoS durations of the three algorithms using the experimental data
Table 6.2: Comparison of the three algorithms using experimental data

<table>
<thead>
<tr>
<th>Performance Metrics</th>
<th>Front</th>
<th>Rear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Handoffs</td>
<td>RT 12, LB 20, AP 7</td>
<td>RT 13, LB 19, AP 8</td>
</tr>
<tr>
<td>Average Handoff Interval (m)</td>
<td>196.5</td>
<td>118.0</td>
</tr>
<tr>
<td></td>
<td>320.9</td>
<td>177.7</td>
</tr>
<tr>
<td></td>
<td>280.8</td>
<td>121.9</td>
</tr>
<tr>
<td>Number of Service Interruption Incidents</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Service Interruption Duration (s)</td>
<td>0.25</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>4.40</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of Low QoS Incidents</td>
<td>79</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Maximum Low QoS Duration (s)</td>
<td>0.5</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

The numbers in the legend indicate the number of handoffs produced by RT, LB, and AP, respectively. From Figure 6.10, we recognize that there are no ping-pong effects for any of the three algorithms, i.e. the MS never performs a handoff back to a previously selected PoA. In terms of the number of handoffs, LB executes the highest amount compared to RT and AP. Moreover, RT performs more handoffs for the front than the rear by a noticeable amount, while AP produces similar numbers of handoffs for the front and the rear. In Figure 6.11, we illustrate the distribution of handoff intervals using histograms. We make two observations. First, LB has shorter handoff intervals compared to RT and AP. Second, LB and AP have similar results for the front and the rear, whereas RT has shorter intervals that are shorter for the front than the rear where the handoff intervals for the front are shorter.

Next, we discuss performances of service interruption and low QoS. We first show the measurements of the PoA selected by each algorithm as a function of the location of the MS in Figure 6.12. First of all, we notice that there is no service interruption incident (signified by a value of -128) for any of the algorithms. Moreover, the RT algorithm exhibits an intriguing phenomenon for the rear where handoff decisions seem to be delayed such that the algorithm always misses the peak strengths of the PoAs as highlighted by the blue ellipses. In Figure 6.13, we present the distribution of the low QoS durations. We recognize that LB does not have any low QoS events, RT and AP have low QoS durations of at most 0.8 and 0.4 seconds, respectively. AP also outperforms RT in terms of the number of low QoS incidents.

In Table 6.3 we summarize the results of this experiment. We conclude that (1) AP performs the least amount of handoffs, while RT generates 3 more handoffs for the front and has the same amount of handoffs as AP for the rear, whereas LB executes about 4 times as many handoffs compared to AP, (2) AP and RT have longer handoff intervals than LB where AP has similar performance for the front and rear and RT has better results for the rear than the front, (3) all three algorithms avoid service interruption, and (4) LB satisfies the QoS requirements at all times, AP and RT have low QoS durations which lasts less than 1 second where AP has fewer and shorter low QoS durations than RT.
Figure 6.10: Handoff locations and locations of the selected PoAs of the three algorithms using the simulated data

Figure 6.11: Distribution of handoff intervals of the three algorithms using the simulated data
Figure 6.12: RSS of the selected PoA by the three algorithms using the simulated data

Figure 6.13: Distribution of low QoS durations of the three algorithms using the simulated data
Table 6.3: Comparison of the three algorithms using the simulated data

<table>
<thead>
<tr>
<th>Performance Metrics</th>
<th>Front</th>
<th>Rear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Handoffs</td>
<td>RT 7</td>
<td>LB 20</td>
</tr>
<tr>
<td></td>
<td>RT 4</td>
<td>LB 19</td>
</tr>
<tr>
<td>Average Handoff Interval (m)</td>
<td>290.3</td>
<td>118.0</td>
</tr>
<tr>
<td></td>
<td>577.5</td>
<td>121.9</td>
</tr>
<tr>
<td>Number of Service Interruption Incidents</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Service Interruption Duration (s)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of Low QoS Incidents</td>
<td>116</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>220</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Low QoS Duration (s)</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0</td>
</tr>
</tbody>
</table>

6.5 Experiment 2: Partial Network Failure

The objective of this experiment is to gain insights of the resilience of each algorithm under partial network failure which may be due to hardware failures or power outage. This experiment is only evaluated with the simulated data. We simulate partial network failure scenarios by removing different combinations of PoAs from the network according to a parameter $k$ such that there is a failed PoA between every $k$ functioning PoAs. Moreover, a value of $k$ corresponds to $k+1$ combinations of failed PoAs. Explicitly, the combinations of failed PoAs are $\{l, l + (k + 1), l + 2(k + 1), \cdots, l + \lfloor \frac{K-l}{k+1} \rfloor (k + 1)\}$ for $l = 1, 2, \cdots, k + 1$ where $K$ is the total number of PoAs in the system. For example, $k = 1$ indicates that there is a failed PoA for every 1 functioning PoAs and corresponds to two combinations, $\{1, 3, 5, \cdots\}$ and $\{2, 4, 6, \cdots\}$. Since there is a total of 32 PoAs in the network, we evaluate $k$ from 1 to 15, and all combinations for each value of $k$ are evaluated. Note that, for each value of $k$, different combinations may yield identical results and numerous data points in the figures presented in this section overlap.

Figure 6.14 illustrates the number of handoffs as a function of failure interval $k$. Each star marker represents the result of a specific PoA combination. The solid lines indicate the maximum results over the different combinations for each value of $k$, and the dashed lines indicate the minimum results. From Figure 6.14, we observe that (1) LB does not show any correlation between the number of handoffs and $k$, RT and AP show a weak correlation when $k$ is less than 2, (2) LB always has the same number of handoffs, (3) RT and AP produce different numbers of handoffs when the PoA combination changes even if $k$ is the same, (4) RT and AP generate similar number of handoffs, and (5) AP shows similar results for the front and the rear, whereas RT produces more handoff events for the front than the rear. In Figure 6.15 we illustrate the average handoff interval against $k$. We recognize that RT has different average handoff intervals for the front (about 300 m) and the rear (about 600 m), RT has constant average handoff intervals for the front and the rear (both about 100 m), and AP has similar average handoff intervals for the front and the rear (both about 400 to 500 m).
Chapter 6. Evaluation of Handoff Decision Algorithm

Figure 6.14: Number of handoffs as a function of $k$ for partial network failure

Figure 6.15: Average handoff interval as a function of $k$ for partial network failure
Figure 6.16 depicts the maximum service interruption duration and the number of service interruption incidents. We observe that RT and AP have no service interruption (as they did in the default experiment), since they are sensitive to RSS measured by the MS, whereas LB has long service interruptions that are more than 2 seconds as it selects the PoAs that have been removed. The extraordinary instance of a maximum service interruption duration lasting for almost 800 seconds is due to the removal of the PoA that is nearest to the MS during the long motionless period. Furthermore, LB shows a noticeable correlation between the number of service interruption and $k$. Figure 6.17 shows the maximum low QoS duration and the number of low QoS incidents. We see that LB has no low QoS incident, RT and AP has short low QoS incidents (less than 0.8 and 0.4 seconds, respectively), and RT has more incidents than AP.

In conclusion, under partial network failure, (1) RT and AP have fewer handoffs and longer average handoff interval than LB, AP has similar number of handoffs and average handoff interval for the front and rear, whereas RT produces different results and the performance for the rear is better, (2) RT and AP has no service interruption while LB has several service instability issues.
Figure 6.17: Maximum low QoS duration (top) and number of low QoS incidents (bottom) as a function of $k$ for partial network failure
Chapter 6. Evaluation of Handoff Decision Algorithm

6.6 Experiment 3: Partial Network Degradation

In this experiment we investigate the robustness of our algorithm under partial network degradation due to a degraded wireless channel, which may be a consequence of degraded hardware or appearance of physical obstacles. This experiment is evaluated with the simulated data. We simulate partial network degradation by decreasing the RSS’s of different combinations of PoAs by 20 dB. The combinations of PoAs are determined by the same parameter \( k \) as used in the previous experiment. Note that, for each value of \( k \), different combinations may yield identical results and numerous data points in the figures presented in this section overlap.

Figure 6.18 shows the number of handoffs as a function of \( k \) where AP and RT has fewer handoff events than LB. Compared to the previous experiment (Figure 6.14), all three algorithms show similar results where LB has identical number of handoffs, AP has similar results for the front and the rear, and RT has better results for the rear than the front. In Figure 6.19 we depict the average handoff intervals. The results display similar trends as in the previous experiment (Figure 6.15) where, first, RT and AP have longer average handoff interval than LB, second, AP has similar average handoff interval and RT has shorter average handoff interval for the front than the rear, and, third, RT and AP demonstrate a weak correlation between average handoff interval and \( k \) when \( k \leq 2 \) and no correlation when \( k > 2 \). Figure 6.20 illustrates the maximum low QoS durations and the number of low QoS incidents. In terms of the maximum low QoS duration, LB may have low QoS durations that are about 7 seconds for the front and 5 seconds for the rear, whereas RT and AP always experience short low QoS incidents that are within 0.8 and 0.4 seconds, respectively. In terms of the number of low QoS incidents, AP has the least amount of events, followed by RT than LB. Moreover, LB shows a correlation between

![Figure 6.18](image_url)

Figure 6.18: Number of handoffs as a function of \( k \) for partial network degradation
the number of low QoS incidents and $k$ where the number of low QoS incidents decreases as $k$ increases. On the other hand, RT and AP do not depict any noticeable correlation. Compared to the previous experiment (Figure 6.17), RT and AP show similar results while LB has different results, where in the previous experiment LB has no low QoS incidents.

To summarize, RT and AP are both robust to partial network degradation where (1) their performances on the number of handoffs and average handoff interval do not have any significant changes compared to the default experiment and the previous experiment, and (2) they both produce results of short low QoS durations that are less than 1 second and AP has fewer low QoS incidents than RT.
Figure 6.20: Maximum low QoS duration (top) and number of low QoS incidents (bottom) as a function of $k$ for partial network degradation.
Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we presented a novel solution to enhance communication availability in CBTC systems with a hybrid network, defined as an infrastructure network augmented by one or more backup network. The objective of the backup network is to provide alternative communication paths when the infrastructure network fails to bridge a connection between the central control station and an MS or when the connection quality cannot satisfy the required QoS for train control, i.e., stringent time constraint. The specific types of wireless technologies that have been considered as prudent choices for the backup network are cellular networks and a MANET formed by trains.

We have proposed to adopt ASON as the network structure to overcome the interoperability and dynamic network topology and routing problem. To validate the practicability of this framework, we assessed the performance implications of ASON over LTE in terms of bandwidth overhead and latency where we found the bandwidth demanded by an ASON, at the size of interest, is affordable by the considered wireless technologies, and the latency of ASON over LTE is capable to meet the time constraint of the safety-critical application for more than 95% of the time with appropriate parameter selection.

Since MSs in a CBTC system have a fixed motion path and almost deterministic and identical mobility patterns, we explained that planning handoffs based on foreknown QoS information can enhance handoff performances in terms of ping-pong effect, number of handoffs, handoff interval, service interruption and low QoS. By selecting appropriate segment sizes and collecting a sufficient amount of samples, the proposed handoff decision algorithm not only showed superior results in eliminating ping-pong effect, minimizing number of handoffs, increasing handoff interval, and reducing service interruption and low QoS to an RSS-based and a location-based scheme, but also exhibits robustness to partial network failure and degradation.
7.2 Future Work

This thesis can be complemented by future endeavour in the following directions.

1. **ASON over MANET:** This thesis has suggested the possibility of using train-formed MANETs as a backup network. The MANETs constitute communication paths for MSs that cannot connect to the infrastructure network. Since cellular networks are unlikely to cover the entire trajectory, especially the underground and tunnel sections, MANET is the remaining choice. The challenge in this approach includes locating the relaying node, i.e., the MSs that can form a direct physical link to a wayside or cellular PoA. To conduct real-life experiments in this direction will require MSs to equip radios that operate in infrastructure and ad-hoc mode simultaneously.

2. **Multi-Substrate Overlay:** We have mentioned the option to use a multi-substrate overlay scheme. The benefit of this scheme is the ability to facilitate the formation of overlay networks in a heterogeneous environment of multiple substrate networks. Unlike a single-substrate overlay, which designates an overlay network for a substrate network, a multi-substrate overlay allows integrated control of all the substrate networks within a single overlay network. More importantly, this scheme allows more complex relaying or message forwarding for MSs that are inaccessible to the infrastructure network. As an extreme example, supposing that the entire wireless infrastructure network is shut down and the only alternative for MSs to communicate with the central control station is via Internet, then the MSs that do not have cellular reception (those underground or inside tunnels) may request other MSs that have cellular reception to forward their messages via, and the overall communication path becomes: MS $\rightarrow$ MANET $\rightarrow$ cellular network $\rightarrow$ Internet $\rightarrow$ central control station. In contrast, options of communication path in the single-overlay scheme are either MS $\rightarrow$ MANET $\rightarrow$ infrastructure network $\rightarrow$ central control station, or MS $\rightarrow$ cellular networks $\rightarrow$ Internet $\rightarrow$ central control station.

3. **Propagation Model:** In the derivation of sample size per segment we only consider simple propagation models, whereas a more comprehensive model that considers correlated shadowing, Rician fading, interferences, or even takes the physical environments such as the geometries of tunnels and stations into account can enhance the accuracy of derivation and the overall handoff decision.
Appendix A

Experimental Measurement Traces

In this appendix we present the experimental measurement traces measured by the front and the rear adapters. Each trace contains a series of scan results. A scan result includes a timestamp, a location-stamp that indicates the location information of the MS on the track, a list of up to 8 PoAs and a list of the corresponding RSS measurements. We plot the series of location-stamps in the horizontal axis and the RSS of each PoA in the vertical axis. When a scan result does not include a PoA, the RSS of the PoA is indicated by -128 dBm. Although there are a total of 32 PoAs in the system, only 28 PoAs (PoA 5 to PoA 32) are found in the measurement trace recorded by the front adapter, and 30 PoAs (PoA 1, PoA 3, PoA 5 to PoA 32) are found in the measurement trace recorded by the rear adapter.

A.1 Front

![Graphs showing measurement traces for PoA 5 and PoA 6](image-url)

(1) PoA 5

(2) PoA 6
Appendix A. Experimental Measurement Traces

<table>
<thead>
<tr>
<th>Location of the MS (m)</th>
<th>RSS (dBm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-120</td>
</tr>
<tr>
<td>50</td>
<td>-100</td>
</tr>
<tr>
<td>100</td>
<td>-80</td>
</tr>
<tr>
<td>150</td>
<td>-60</td>
</tr>
<tr>
<td>200</td>
<td>-40</td>
</tr>
<tr>
<td>250</td>
<td>-20</td>
</tr>
<tr>
<td>300</td>
<td>0</td>
</tr>
<tr>
<td>350</td>
<td>20</td>
</tr>
<tr>
<td>400</td>
<td>40</td>
</tr>
<tr>
<td>450</td>
<td>60</td>
</tr>
<tr>
<td>500</td>
<td>80</td>
</tr>
<tr>
<td>550</td>
<td>100</td>
</tr>
<tr>
<td>600</td>
<td>120</td>
</tr>
</tbody>
</table>

(3) PoA 7

(4) PoA 8

(5) PoA 9

(6) PoA 10

(7) PoA 11

(8) PoA 12

(9) PoA 13

(10) PoA 14
Appendix A. Experimental Measurement Traces

(19) PoA 23

(20) PoA 24

(21) PoA 25

(22) PoA 26

(23) PoA 27

(24) PoA 28

(25) PoA 29

(26) PoA 30
Appendix A. Experimental Measurement Traces

(27) PoA 31

(28) PoA 32
A.2 Rear

(1) PoA 1

(2) PoA 3

(3) PoA 5

(4) PoA 6

(5) PoA 7

(6) PoA 8
Appendix A. Experimental Measurement Traces

(7) PoA 9
(8) PoA 10
(9) PoA 11
(10) PoA 12
(11) PoA 13
(12) PoA 14
(13) PoA 15
(14) PoA 16
Appendix A. Experimental Measurement Traces

(15) PoA 17

(16) PoA 18

(17) PoA 19

(18) PoA 20

(19) PoA 21

(20) PoA 22

(21) PoA 23

(22) PoA 24
Bibliography


