Chapter 1

INTRODUCTION

The infusion of wireless technologies in the modern digital communication infrastructure and its impact on our day-to-day life has been phenomenal in the recent time. The inception of wireless data services with their extensive support for legacy applications, such as e-mail, web browsing, file transfer etc., complemented with more sophisticated applications like geo-positioning system (GPS), mobile games, mobile multimedia services, mobile commerce, social networking etc., has given new dimensions to work, leisure, information access and social interactions. In fact, once dreamed paradigm of ubiquitous computing (Lyytinen Kalle 2002) has just started to deliver part of its ultimate promise. It would not be wrong assertion that high speed, reliable and dynamically scalable wireless data communication services work at the core to enable true ubiquitous access for end users. There are numerous wireless access technologies, more commonly termed radio access technologies (RATs), already deployed and contending to provide data services to end users for their specific needs, within constraints of each technology. These include IEEE standards 802.11 a/b/e/g/n based Wireless Local Area Networks (WLANs) (IEEE-802.11 2007), IEEE 802.16e based Wireless Metropolitan Area Networks (WMANs) (802.21-2008 2009), and series of generations of Cellular Networks (2G, 2.5G, 3G, etc) (Lawton 2005).

Each of the above mentioned technologies has its own advantages as well as some limitations. For instance, WLAN is good for moderate data rates with restricted mobility at lower cost; whereas, WMAN provides higher data rates with low mobility support at higher cost (S. K. Leung 2008). Similarly; cellular networks provide intermediate data rates at higher mobility with high cost but are constrained by user density and ambient conditions (T. Huang 2009). Despite their momentous flexibility, wireless media lacks consistency due to various factors; such as, mobility events, channel fading, fluid user density and traffic load that cause service quality variations and generally result in less than optimum throughput to mobile sessions (Jamalipour, Wada and Yamazato 2005). Such perturbations result in unpredictable performance metrics (reduced throughput, extended variations in end-to-end latencies, higher drop rates), and occasionally, the network may become
completely incapable of meeting the requirements of ongoing sessions, particularly in case of real-time applications (Kameswari. C 2006). Generally, for the quality-aware sessions, the resources are reserved for the whole duration of the session during call admission control (CAC) process but mobility and varying signal conditions may not comply with such agreements due to physical constraints (Devarapalli 2005).

At the same time, processing capabilities of mobile communication devices have also hoisted from simple phones to very sophisticated pocket personal computers (PCs) and mobile routers (MRs) (Lucian Suciu 2005). We find today, mobile devices equipped with powerful processors, large memories and multiple wireless interfaces that provide redundant choices of connecting to any of available wireless networks (Chen Yiping 2007). The enhanced computing power of mobile devices has also brought much heavier applications on these devices, enabling them to be used as mobile servers and routers. These services are heavy in terms of both processing complexity as well as communication volume. Further, real-time applications like video streaming and video conferencing require consistent service guarantees; more commonly referred to as Quality of Service (QoS), in terms of data rates and bounded latencies. Keeping in view the critical nature of all these services; the guaranteed service provisioning in such environments is challenged by the mutable service of wireless data networks and has received significant interest of researchers in recent years (Devarapalli 2005) (X. Z. Chen 2008).

The availability of communication resources, like output buffers in intermediate routers and communication link’s time-slots, is crucial for providing service guarantees to QoS-aware sessions. In the wired subnets, these resources are used more efficiently due to the higher predictability and consistency of link. In mobility using Mobile Internet Protocol (MIP), the link availability changes significantly in time and space and causes greater degree of uncertainty about the next point-of-attachments (PoAs), that also leads to major change in the path of session/flow (Perkins 2002). It is quite unlikely that the same resource set availed at the previous PoA is also available at all subsequent PoAs. It is therefore, desirable that some level of over-provisioning of resources is maintained in the mobility supporting routers and associated paths to overcome expensive signaling cost of renegotiations with multiple PoAs (S. De Vuyst 2008). This approach results in non-scalable systems that may not accommodate even modest user loads despite being quite expensive. There are quite a few parameters, such as user density, network capacity, mobility profiles
and patterns, channel access mechanisms that influence the service consistency of mobile data networks and make related decision making more complex during mobility events. The QoS provisioning during mobility is therefore, a question that requires multi-facet analysis before a comprehensive solution may be possible (Sharma 2008).

In the backdrop of finding comprehensive solution for QoS during mobility, the mobile devices equipped with multiple wireless interfaces; experiencing a heterogeneous wireless network (HWN) environment, provide a window for enhancing QoS by opportunistic use of any or more than one networks simultaneously (X. Z. Chen 2008). The multimode mobile device (MMD); the term found commonly in literature for devices equipped with multiple wireless interfaces, not only can improve service guarantees for their application and services, but can also provide sufficient leverage to other single interfaced mobile devices by choosing network with lower traffic load. Therefore, simultaneous use of multiple interfaces of MMDs has been keenly studied as bandwidth aggregation (BAG) techniques in recent past (X. Z. Chen 2008). Though initial studies indicate considerable potential of BAG techniques (sometimes also referred to as capacity aggregation; abbreviated as CAG of multiple wireless channels), the usefulness of any such configuration is dependent on many parameters; such as the degree of asymmetric characteristics of wireless channels, the congestion state of each path of these wireless channels, the architectural placement of such BAG/CAG techniques in the protocol stack, number of hops on each path, ratio of load distribution on each path, and overhead of managing BAG/CAG (Kameswari. C 2006). In worst cases, the cumulative impact of all these factors can be catastrophic for end-to-end (E2E) performance metric (retransmissions due to out-of-sequence reception, higher packet drop rate, complex congestion control mechanisms). Even in best case scenarios, the performance of BAG/CAG techniques may not coincide with the theoretical sum of all available capacities due to its operational and maintenance overheads. Mobility of such MMDs adds another major constraint. It is therefore, important to quantitatively analyze the concerned parameters to make BAG/CAG techniques more useful and workable. The accuracy of such analysis can greatly facilitate process of developing suitable scheduling strategies and distribution of traffic load on available interfaces. The purpose of these scheduling strategies may be focused on subsiding counter-productive impact of heterogeneity of multiple paths and to achieve one or more desirable performance metrics; either for the individual flows or for the networks (Chebrolu K. 2005). This thesis investi-
gates the possibility of maximizing QoS of individual flows during mobility, through quantification of impact of multi-path heterogeneity.

1.1. Background

One of the possible consequences of multi-path traversal of a single flow is the out-of-sequence (OOS) reception of packets at the receiver or at intermediate proxy, deployed to harmonize these flows (X. Z. Chen 2008). The OOS reception could seriously increase complexity of congestion and flow control mechanisms of transport protocols such as transmission control protocol (TCP) and Stream Control Transport Protocol (SCTP) (Janardhan R. Iyengar 2006). The other QoS enabled protocols such as Real-time Transport Protocol (RTP) and Real-time Stream Transport Protocol (RSTP) that require stringent time and rate constraints may also face higher buffer occupancy and packet drop due to OOS reception of packets at the receiver (Piratla N M. 2008). The problem gets sterner during mobility as episodes of various mobility events, such as handovers cause repeated disconnections at layer 2 (S. K. Leung 2008). In a multiple care-of-address (CoA) registration scenario using Mobile IPv4 (MIPv4), multiple tunnels are available between MMD and home agent (HA) to simultaneously use these tunnels and transparently route backlogged packets over any of available tunnels (R. Wakikawa 2009).

The major challenge in such scenarios is to transport multiple flows with distinct QoS requirement, on these multiple tunnels in a manner that minimizes the probability of OOS reception for each flow (Devarapalli 2005). Figure 1-1 depicts one such scenario where there are multiple flows, such that flow $i$ has a data rate $r_i$, are queued in multiple buffers before being transmitted over the multiple tunnels through a multi-server scheduler\(^1\), working at MMD. The scheduled packets then traverse through any of the $M$, MIPv4 tunnels, as per deployed scheduling strategy. The major challenges in devising suitable scheduling strategy depends on identification and estimation of each path characteristics accurately, to support scheduling process in a way that E2E delay variations and OOS reception are minimized at the tunnel ends. A comprehensive model for analysis of such scenarios is therefore important before adopting any scheduling strategy. This thesis work de-

\(^1\) A multi-server scheduler is a terminology used in queuing system where multiple service points are available. In packet switched communication networks, the multiple network interfaces act as the service points.
develops foundations for QoS aware multi-server scheduling paradigm for mobile sessions. Under the assumption of sufficient availability of communication capacity of multiple networks, it minimizes OOS reception through proactively adaptive scheduling strategies that accurately estimate the characteristics of multiple tunnels and emulates multiple MIPv4 tunnels as a single virtual channel with known probability of OOS arrival. It develops both stochastic and deterministic analytical models to signify E2E path dynamics of such tunnels to minimize impact of E2E path asymmetries of these tunnels.

Figure 1-1: A simplified model of Multipath E2E flow management during mobility

1.2. State-of-the-Art in Bandwidth Aggregation

The simultaneous use of multiple available wireless interfaces has gained considerable interest of research community in recent years. The majority of researches attempted to achieve BAG to tap maximum service for specific set of applications, such as video streaming, online streaming. The upper layers are considered as more convenient means of service maximization for individual flows due to Application Programming Interface (API) support of operating systems. In such approaches, individual applications open multiple sockets over multiple ports and transmit multiple streams of a single flow to achieve desired service rates (T. Huang 2009). It is obvious that in such arrangement, large volumes of data may be transported in short time, but due to varying E2E dynamics of multiple networks, service may not be guarantee, particularly in terms of latencies. Similarly, transport layer BAG approaches use single or multiple congestion windows to maintain unified E2E stream through multiple paths (Chebrolu K. 2005), (Sharma 2008) & (Janardhan R. Iyengar 2006). The above mention approaches achieve desirable service
guarantees at the cost of service fairness, as multiple congestion windows may reduce the share of other flows considerably. The multiple congestion windows add significant overhead of congestion and flow control operations and may also induce congestions in the host networks. It is important to note that despite heavy overheads and congestion vulnerabilities, service guarantees in terms of bounded latencies and in-order reception of flow may not be guaranteed. In (Kameswari. C 2006), a network layer solution has been proposed that sends video traffic to an MMD using a proxy server at the base station (BS). The BAG is accomplished through transmission of packets, arrived at the proxy server over any one of available links that promises the fastest delivery of traffic to the MMD. Since it is a single hop multipath traversal, the OOS arrival is nominal but in case, packets are already arriving OOS at the proxy server, the in-order delivery to MMD cannot be guaranteed. In (K. R. Evensen 2009), a network layer, proxy server based approach of BAG is presented that ranks a link according to its service quality and transmits packets on the link with best link quality. This approach has significant benefits provided path characteristics don’t change significantly over the time. In mobility based scenarios, the OOS arrival may be much higher due to frequent path changes (Ming Yang 2004). The uncertainty in acquiring the desired QoS parameters at the next PoA has also motivated for pre attachment resource reservation for the mobile sessions at the probable PoA (Xiao 2004). In this approach, resources are reserved for a mobile session while it is still in the previous PoA. This approach seriously hampers the scalability of system and also results in poor resource utilization. Table 1.1, 1.2 and 1.3 summarize major contributions in this area. The review of above mentioned state-of-the-art techniques of BAG/CAG highly motivates thorough investigation of QoS maximization potential anticipated through BAG/CAG techniques, especially in presence of OOS reception due to multi-path traversal of a flow during mobility. The same is thus, the core research question investigated in this thesis.

1.3. Hypothesis and Contributions of the Thesis

The hypothesis of the thesis is the provision of QoS\(^2\) while in mobility, through multi-path transport of flows under the constraints of minimum OOS reception of packets at the receiver end. In multi-path flows, QoS may be compromised by the E2E delay variations of constituent paths that cause OOS reception of packets, as discussed before. Therefore, we

\(^2\) Though QoS has diverse meanings in communication networks, but in the scope of this thesis, it has been defined in terms of desired data rates within a bounded delay constraint.
need to minimize OOS in order to achieve required QoS. The OOS arrival is mainly attributed to some parameters; including E2E delay and its variations. In this thesis work, we have formalized the fundamental E2E parameters that influence the multi-path traversal of QoS enabled flows during mobility in HWN environment and contribute significantly in OOS reception. The fundamental question of interest is whether or not CAG of heterogeneous paths can provide QoS guarantees during mobility under the constraint of minimum OOS reception? The subsequent questions that arise include what parameters can be most crucial in quantifying heterogeneity of the multiple paths and the impact of heterogeneity on QoS guarantees of traffic flows? Can we upper-bound the impact of these parameters to guarantee desired QoS levels? How can we make such upper-bounds to be tight enough so as they don’t hinder the scalability of the system without compromising desired QoS levels? What kind of scheduling strategies can best absorb the impact of path heterogeneity? How closer we can get to the optimization of QoS guarantees during mobility, after answering the above questions?

The thesis answers most of the questions of the research hypothesis. The focal point in this work is to analyze the probable causes of OOS reception, and the impact of such reception. We have adapted an incremental approach in which asymmetric characteristics of collocated subnets are analyzed through both deterministic and stochastic models to estimate intra-path and inter-path E2E delay variations and then combined it into a model that minimizes OOS reception. These models provide closed form upper bounds for the probable OOS reception of packets of the same flow. The deterministic model provides useful insight for achieving upper bounds for the two key metrics i.e. Intra-path and inter-path E2E delay variation and corresponding extent of OOS reception. The deterministic model, though is useful for basic system design, provides only best or worst case statistics for the above mentioned E2E metrics. In order to derive complete spectrum of the distribution of these metrics, stochastic models are developed to achieve tighter bounds and improve scalability of installed capacity in the system.

To our best knowledge, both deterministic as well as stochastic models are not available in literature for QoS-enabled multi-server scheduling over heterogeneous networks. Similarly, the proactive description of OOS reception of traffic over the multiple asymmetric paths is also a novel contribution of this thesis. The bounds achieved through the modeling process are used to devise novel scheduling algorithms to distribute traffic on multiple
paths. The validation of these models and accompanying scheduling algorithms is performed through simulations in ns-2 simulator (The Network Simulator - ns-2 n.d.).

Table 1-1: BAG/CAG proposals for the network layer of protocol stack

<table>
<thead>
<tr>
<th>BAG Proposal</th>
<th>Aggregation Technique</th>
<th>Impact</th>
</tr>
</thead>
</table>
| (Kameswari, Bhaskaran and R., 5 2005)                                                                                                                                                                       | Capacity Estimation over the E2E path and Weighted Traffic Distribution                                                          | **Contribution:** Basic E2E multi-path solution for TCP  
**Drawback:** Capacity estimation is expensive in terms of bandwidth |
| (Evensen, et al. 2009)                                                                                                                                                                                      | Intermediate Proxy deployment based traffic integration and distribution                                                        | **Contribution:** General solution for multi-class traffic an better resource utilization  
**Drawback:** Location of proxy not evaluated as the multi-hop parallel path adds complex E2E delay distribution |
| (Kameswari and R, Bandwidth Aggregation for Real-time Applications in Heterogeneous Wireless Networks 2006)                                                                                                  | EDPF based scheduling integrated at the Proxy server                                                                            | **Contribution:** Lesser OOS arrival and improved throughput.  
**Drawback:** Restricted to a single hop wireless section solution |
| (Jui-Tang, Yuan-Ying and Chien-Chao 2008)                                                                                                                                                                   | Redundant resource reservation on available paths                                                                             | **Contribution:** QoS is ensured through reservations at multiple probable PoA  
**Drawback:** Redundant reservation results in poor resource utilization |
| (Phatak, Goff and Plusquellic 2003)                                                                                                                                                                          | FIFO scheduling on multiple IP-in-IP encapsulated Tunnels                                                                      | **Contribution:** Higher bandwidth for TCP applications  
**Drawback:** Path asymmetries not handled causing poor E2E delay performance for real-time applications |
| (Jiang and Yao 2003)                                                                                                                                                                                       | FIFO based Scheduling                                                                                                           | **Contribution:** E2E delay variation is accounted for service guarantees  
**Drawback:** The distribution of E2E delay variation not studied |

The results of analytical models and simulations indicate significantly reliable QoS provisioning during mobility through multi-server scheduling while demonstrating low order OOS reception at the tunnel end. It has been found through the proposed deterministic model that the OOS reception at the tunnel ends is primarily based on the variations in server latencies that eventually lead to intra-path as well as inter-paths E2E delay variations. It has been further investigated that delay variations are also directly proportional to the number of servers (nodes) in the paths. The variations in packet length also contributed
significantly in overall E2E delay variations through variable transmission delay. The variations in propagation delays of heterogeneous wireless channels also contributed in OOS reception of packets. The simulations of proposed scheduling scheme has revealed that about 80% of traffic experience delay variations of fewer than 10% of mean, and the accompanying OOS receptions are even lower than 8% of all received packets. It was further investigated that QoS guarantees were accomplished with lesser buffer overhead at the tunnel ends for holding OOS reception that helps in maintaining in-order forwarding of packets to the application or the next subnet. The QoS guarantees were maintained up to a very high reliability of 90%. The results of proposed models were validated through simulations and were found in close binding with analytical model results within a narrow range of 10% deviations.

Table 1-2: BAG/CAG proposals at the network layer of protocol stack

<table>
<thead>
<tr>
<th>BAG Proposal</th>
<th>Aggregation Technique</th>
<th>Impact</th>
</tr>
</thead>
</table>
| (Puneet, et al. 2009)                 | Collaborative sharing of community resources| Contribution: Scalable solution for end-users through higher communication resource availability  
Drawback: Group membership is essential which is difficult to keep up-to-date during mobility |
| (T. Huang 2009)                       | API and socket level supported link aggregation | Contribution: Easy to use aggregation  
Drawback: Control at application that may be unfair for other applications |
| (Fernandez, et al. 2009)              | QoS Negotiation through Multiple Interfaces | Contribution: QoS provision is more scalable  
Drawback: Negotiation cost is generally high |
| (Yao, Guo and Bhuyan 2008)            | FIFO Scheduling over homogeneous channels    | Contribution: Maximum resource utilization  
Drawback: Generally good for only one hop or over a well defined tunnel |
### Table 1-3: BAG/CAG proposals for Transport layer of protocol stack

<table>
<thead>
<tr>
<th>BAG Proposal</th>
<th>Aggregation Technique</th>
<th>Impact</th>
</tr>
</thead>
</table>
| (Kalyanaraman, Ramakrishnan and V. 2008) | Weighted Congestion window approach | **Contribution:** Higher throughput through multi-stream transmission  
**Drawback:** oscillation in congestion window based on historic data |
| (Ken and Hiroshi 2006) | Congestion Control tuned to fast retransmission problem | **Contribution:** Reduces fast retransmission issues  
**Drawback:** Mobility issues not discussed |
| (Magalhaes and Kravets 2001) | Congestion control based on differentiation of link and network loss | **Contribution:** Improved stability of consistent throughput  
**Drawback:** Complexity of Congestion control and capacity estimation is high |
| (Al, Saadawi and Lee 2004) | Modified Congestion Control of SCTP | **Contribution:** Higher Throughput with multi-streaming  
**Drawback:** The congestion control mechanism is tuned for specific applications. |
| (Iyengar, Amer and Stewart 2006) | Multi-streamed, multiple congestion window based link aggregation | **Contribution:** Improved throughput  
**Drawback:** The behavior of multiple congestion windows adds unpredictable retransmission scenarios |

### 1.4. Outline of Thesis

The flow of the thesis is given in the following text. Chapter 2 describes the system model along with underlying assumptions of the multi-tunnel mobility using multiple CoA registration at HA. The symbols and notation used in the stochastic and deterministic models are also described here. In chapter 3, a novel deterministic analytical model for multipath traversal of a flow is presented. The basic motive is to quantify E2E delay variation and bounded OOS reception. The deterministic model upper-bounds E2E delay variation on the basis of server latencies, transmission delays, propagation delays and packet arrival time at the scheduling server. The guaranteed rate service model is used to describe E2E path delay. The major source of delay variation is cause by the size of the packets and numbers of hop on each path. Under proposed model intra-path delay variation are much lesser than the inter-path delay variations. In the second model, the strict upper bounds of the E2E delay and its variations were relaxed to lower levels at significantly realistic thresholds to improve scalability of the service. In this model, it is identified that the delay variations are rooted in variable packet sizes and average queue lengths on each server in
the E2E path. It is proven that the proposed model has some salient properties that make it suitable for estimation of E2E delay variations and associated OOS reception. We also provide proof of theorems that enable characterization and quantification of E2E delay related metrics.

In chapter 4, a novel bounded variance network calculus (BVNC) based stochastic model for E2E delay variations, OOS receptions and buffer overflow bounds is presented. The stochastic study of these properties helps in optimizing the buffer management used to handle OOS reception. The stochastic model provides two important metrics, i.e., the probability of violation of a delay threshold at any server, and probability of violation of a given E2E delay threshold. These two metrics help in designing the appropriate strategy of buffer management to cope with OOS reception problem. The proof of theorems for violation of thresholds for E2E delay variations, OOS reception, and buffer occupancy are also given.

In chapter 5, we present multi-server scheduling algorithms that are based on the outcome of modeling exercise given in chapter 3 & 4. The proposed scheduling algorithms use E2E metrics to rank the available path, and schedule multi-class traffic, accordingly. It also provides simulation results of the cross-layer algorithms developed to minimize E2E delay variation and OOS reception. The results are compared with some of the existing schemes for comparison.

Chapter 6 summarizes the overall contribution of the thesis and findings of the research. Some future dimensions of research in this domain are also discussed. The importance of BAG/CAG is discussed with the support of some useful results as discussed in chapter 6.