Statistical Analysis of ATM Call Detail Records

Creighton T.R. Hager

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical Engineering

Scott F. Midkiff, Chair
Luiz A. DaSilva
Nathaniel J. Davis, IV
Clark K. Gaylord

December 1999
Blacksburg, Virginia

Keywords: ATM, Cluster Analysis, Quality of Service, Service Level Agreements

Statistical Analysis of ATM Call Detail Records

Creighton T.R. Hager

(ABSTRACT)

Network management is a problem that faces designers and operators of any type of network. Conventional methods of capacity planning or configuration management are difficult to apply directly to networks that dynamically allocate resources, such as Asynchronous Transfer Mode (ATM) networks and emerging Internet Protocol (IP) networks employing Differentiated Services (DiffServ). This work shows a method to generically classify traffic in an ATM network such that capacity planning may be possible. These methods are generally applicable to other networks that support dynamically allocated resources.

In this research, Call Detail Records (CDRs) captured from a “live” ATM network were successfully classified into three traffic categories. The traffic categories correspond to three different video speeds (1152 kbps, 768 kbps, and 384 kbps) in the network. Further statistical analysis was used to characterize these traffic categories and found them to fit deterministic distributions. The statistical analysis methods were also applied to several different network planning and management functions. Three specific potential applications related to network management were examined: capacity planning, traffic modeling, and configuration management.

Acknowledgments

I would like to express my deep felt gratitude to my advisor, Dr. Midkiff, for his patience, advice, and continuous support. The other members of my committee, Dr. L.A. DaSilva, Dr. N.J. Davis, IV, and Mr. C. K. Gaylord, have also provided valuable suggestions and constructive feedback for this work.

I would like to thank my parents, Albert and Louisiana, and my brother, Kristopher, who encouraged and supported me through all my years in school. I also give special thanks to Yongqin Zhuang, who helped me immensely throughout my Masters program.

This research was conducted as a collaborative project between the Bradley Department of Electrical and Computer Engineering and Communication Network Services (CNS). I give many thanks to Eric Brown and Clark Gaylord for their knowledge and advice. The funding provided by IBM University Partnership Program grant awarded to Virginia Tech in September 1998 helped support my education for the past couple of years.

Thanks to Xianrui Roger Wang, my research partner. He performed the initial tasks that enabled my research to commence. I appreciate his productivity and cooperation throughout our work together.

My colleagues and friends, who have worked with me on many projects unrelated to this thesis, must also be noted. These include but are not limited to, Rajesh Banerjee, Takeshi Ikuma, Andras Kovacs, Katina Reece, Haiyuan Wang, and Yu Wang.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

1.1 Motivation

Multimedia, e.g., voice and video, application classes have joined data as important traffic types. Examples of these application classes are video teleconferencing, video-on-demand, and distributed simulation. Such mixed traffic types are supported by Asynchronous Transfer Mode (ATM) technology [1] and emerging Internet Protocol (IP) Differentiated Services (DiffServ) [2,3]. While these classes can operate to some extent using best effort delivery, conventional best effort service for this type of traffic can result in reduced performance for the application and, potentially, inefficient use of network capacity. It is the purpose of this research to demonstrate an approach of providing reliable capacity planning strategies for multimedia traffic on networks that dynamically allocate network resources to provide QoS, with a specific focus on ATM networks. Such an approach is needed to manage networks that extend the existing best effort service model to meet the needs of multimedia classes with real-time constraints.

ATM Call Detail Records (CDRs) serve as the information basis for this research. CDRs provide multiple types of information, including Circuit Utilization and Call Times [4,5]. Possible functions of analysis of CDRs include resource allocation, establishing and monitoring service level agreements, and traffic modeling. Ultimately, these functions can be applied to networks other than ATM and help provide reliable QoS as well as an understanding of the inherent traffic.

1.2 Approach and Results

The focus of this work is a system approach for capacity planning, including data collection from the ATM network based on collection of CDRs, classification of the data into distinct traffic categories, algorithms for analysis of the data including determining statistical distributions, and potential network management tasks that can benefit from the analytical results. The general data collection and analysis process implemented in this project is shown in Figure 1-1. Note that this work focused on data analysis. Data collection is described in detail in [4].

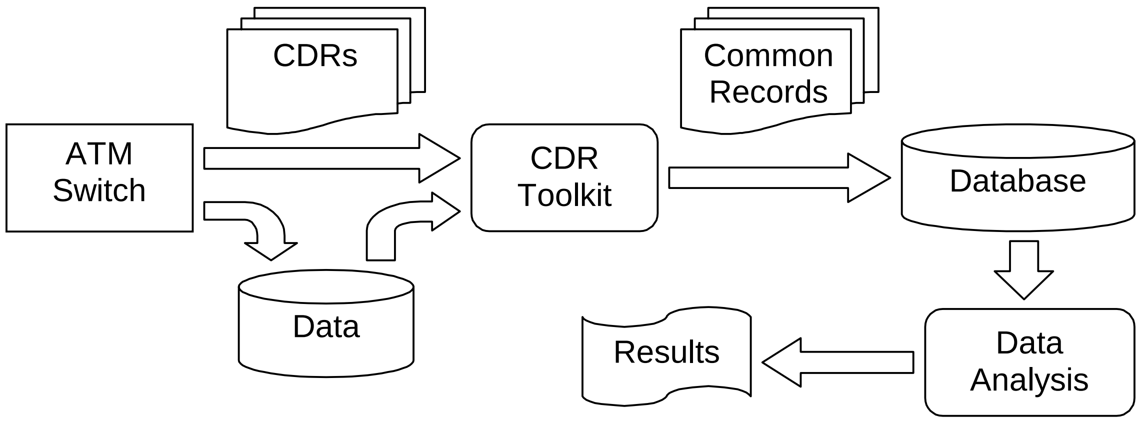

Figure 1-1. Overall data collection and analysis process.

The overall goal of this research is to investigate a method that resolves any arbitrary call in the network into a particular traffic category and determine relevant potential network applications from the analytical results.

Statistical analysis is performed on the data parameters of the CDRs. This procedure classifies network “calls” into separate traffic categories, as discussed in Chapter 3. The traffic categories are generally created from similar QoS and traffic parameters, summarized in Chapter 2. The data parameters are further analyzed to determine possible statistical distributions. Knowing the particular distribution of a traffic category can aid policy-based management, a potential network management application, described in Chapter 4. In addition, analysis of the network utilization by calls can yield vital information about provisioning in operational networks. Prospective strategies and features of network management needed for a dynamically allocated resource network are detailed in Chapter 5.

1.3 Organization

This thesis is divided into six chapters. Chapter 2 presents an overview of prior and current research related to networks, especially ATM, that dynamically allocate resources. Characterizations of ATM service categories and CDRs are reviewed. Cluster analysis techniques are also discussed as a related research topic. Chapter 3 discusses how the research data was used. The creation of the research data and the statistical analysis schemes and programs are presented. Research goals are also discussed. Chapter 4 reviews the results of the statistical analysis. The classification of network traffic is discussed. Analysis of the CDR data to examine statistical distributions is also examined. Chapter 5 presents potential network management applications based on the statistical analysis. Relevant statistical results are also presented and discussed. Chapter 6 summarizes the work, discusses contributions of this research, and presents potential future work. Detailed results from the data analysis are included in an appendix.

Chapter 2. Background

This chapter discusses relevant information related to ATM networks, such as service categories, traffic parameters, and call detail records. Concepts related to cluster analysis are also reviewed. An introduction to definitions and cluster analysis models are presented. The procedure for collecting CDRs from ATM switches is also discussed.

2.1 ATM Service Categories and Traffic Parameters

2.1.1 Service Categories

The ATM Forum [6] defines five service categories that relate to traffic characteristics and QoS requirements.

Constant Bit Rate (CBR): The CBR service category is used by connections that request a fixed amount of bandwidth that is continuously available during the lifetime of the connection. This service supports a constant or guaranteed rate to transport services, such as video, voice, or circuit emulation, that require rigorous control of timing and performance parameters.

Real-Time Variable Bit Rate (rt-VBR): The real-time VBR service category is intended for real-time applications, i.e., those requiring tightly constrained delay and delay variation, as would be appropriate for voice and video applications. Sources are expected to transmit at a rate that varies with time and can be described as “bursty.”

Non-Real-Time Variable Bit Rate (nrt-VBR): The non-real-time VBR service category is intended for non-real-time applications that have bursty traffic. For those cells that are transferred within the traffic contract, the application expects a low cell loss ratio.

Unspecified Bit Rate (UBR): The UBR service category is intended for non-real-time applications, i.e., those not requiring tightly constrained delay and delay variation. Examples of such applications are traditional data communications applications, such as file transfer and email. In addition, UBR service does not specify traffic related service guarantees and can be characterized as a “best effort” service.

Available Bit Rate (ABR): As defined by the ATM Forum, ABR is an ATM layer service category for which the limiting ATM layer transfer characteristics provided by the network may change subsequent to connection establishment. A flow control mechanism is specified that supports several types of feedback to control the source rate in response to changing ATM layer transfer characteristics. ABR provides for transport of traffic at the bit rate available at the time and on a dynamic basis.

2.1.2 Traffic Parameters

A traffic parameter describes an inherent characteristic of a traffic source [6]. It may be quantitative or qualitative. For every traffic class there are a number of parameters that characterize both the traffic and the QoS. The requested parameters have to be defined by the user when a connection setup is requested. Each traffic parameter is reviewed below.

Peak Cell Rate (PCR): The PCR traffic parameter specifies an upper bound on the rate, in cells/s, at which traffic can be submitted on an ATM connection.

Sustainable Cell Rate (SCR): The SCR is an upper bound on the average rate of the conforming cells of an ATM connection, in cells/s, over time scales that are long relative to those for which the PCR is defined.

Maximum Burst Size (MBS): MBS traffic parameter specifies the maximum number of cells that can be transmitted at the PCR, assuming the receiving buffers are empty at the beginning of the burst.

Minimum Cell Rate (MCR): An ABR service traffic descriptor, in cells/s, that is the rate at which the source is always allowed to send.

2.1.3 Association Between Service Categories and Traffic Parameters

In the previous sections, the ATM service categories and traffic parameters were described independently. Some traffic parameters are related to some service categories. Table 2-1 lists the relations among them.

Table 2-1. ATM Service Category and Attribute Associations

Attribute	**ATM Layer Service Category**				
	CBR	rt-VBR	nrt-VBR	UBR	ABR
Traffic Parameters:					
PCR, CDVT	Specified			Specified	Specified
SCR, MBS, CDVT	N/A	Specified		N/A	
MCR	N/A			N/A	Specified

The requested service category and associated traffic parameter values or a connection can be found in an ATM CDR. These parameters, and others, are used for the statistical analysis described in the next chapter.

2.2 Call Detail Records

A call detail record is created when a virtual circuit is established. The FORE ASX switch used in this research generates a raw CDR data file every 5 minutes. (Note that five minutes is the default interval value and this value can be adjusted [7].) The raw CDRs from FORE switches are temporarily stored in the memory at the switch and then automatically transferred to an external UNIX file server via FTP. The necessary data for analysis is extracted from the call detail records.

2.2.1 Duration Parameter

The call duration and its descriptive statistics are important factors for any type of statistical analysis involving traffic characterization. Finding this parameter from the call detail records requires some assumptions before integration into the data set to be analyzed. A call may be seen multiple times in the data set due to multiple samples of call records over the duration of a call. For example, choosing a 24-hour period sample for cluster analysis and having the records logged every 5 minutes will cause any call longer than 5 minutes to appear in multiple records.

The duration of the calls that span the sample period need to be considered. There are essentially four types of calls: (a) a call that was setup and terminated within the sample period, (b) a call that was setup before the beginning of the sample period and was terminated after the end of the sample period, (c) a call that was setup before the beginning of the sample period and was terminated during the sample period, (d) a call that was setup within the sample period and was terminated after the end of the sample period. These types of calls are illustrated in Figure 2-1.

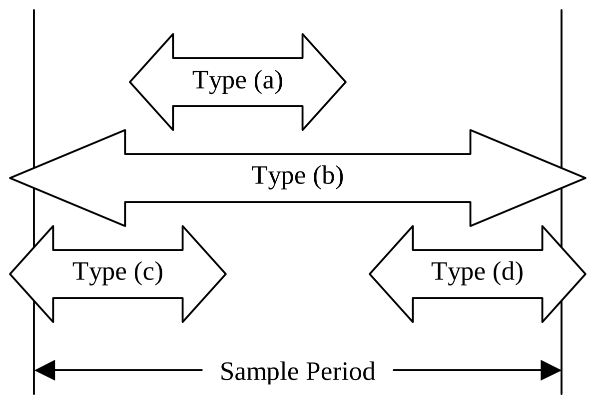

Figure 2-1. Classification of duration parameter.

The durations for type (a) and type (b) calls can be determined from the call records. However, calls of type (d) have unknown duration. Calls longer than the sample period, type (b), can be assigned a duration of the sample period, if necessary. Calls that are setup within the sample period and are terminated after the end, type (d), have more complications. Type (c) and type (d) calls balance each other, because portions of both call types are not included in the sample period. Since duration can be found for type (c) calls, type (d) calls were dropped from the sample and the type (c) calls were kept.

The continuous succession of calls plays an important role in the analysis. In an ideal environment, the switches sometimes have calls that are active for days. However, realistically, switches will release or terminate the calls and immediately attempt to set them up again. If certain parameters of two type (a) calls are identical, i.e., the same starting times and port addresses, and the difference between the termination time of the first call and the setup time of the second call is relatively small, the two calls can be assumed to be one.

Failed calls are assumed to be unimportant. These calls are ignored in the current analysis because the amount of time for which the switch resources are used is negligible. Only successful calls are of interest since they occupy significant network resources.

2.2.2 Extracted Data Parameters

In the previous section, a procedure was described to calculate the call duration parameter. This procedure was implemented due to the regular interval, e.g., every five minutes, that CDR records are stored to a file system and the irregular durations of the network calls. This parameter is used in the statistical analysis procedures described later. This parameter is just one of several needed for traffic characterization. The other parameters needed for the statistical analysis are extracted directly from the CDRs [4]. This was accomplished by first converting the raw CDR from the switch into a readable and usable format. A toolkit written specifically for extracting the data from the CDR was then used to obtain the important parameters for the statistical analysis phase [4]. The extracted parameters include some of the traffic parameters (PCR, SCR, and MBS), the class of service, and the total number of cells received and transmitted during the call.

2.2.3 Utilization Parameter

Knowing the utilizations of a set of calls can help determine the traffic intensity or bandwidth usage of a given link at a given time. This parameter is determined by dividing the total number of cells transferred in a call by the duration, or holding time, of the call. Unfortunately, this calculation is a rough average, at best, of utilization. By using this method, the variance of the utilization within a particular call is not known. There can be intermittent periods of time within a call where no cells are transferred because of the bursty nature of the traffic. Recording the total number of cells transferred at given points throughout a call allows the variance of utilization within the call to be estimated. However, this matter is left for possible future research.

2.3 Cluster Analysis

Cluster analysis [8, 9] is a multivariate procedure for detecting natural groupings in data. Classification is based on placing objects into groups or clusters suggested by the data, not into groups that are defined *a priori*, such that objects in a given cluster tend to be similar to each other in some sense and objects in different clusters tend to be dissimilar. The objective is to minimize the variation between members of each cluster while maximizing the difference between groups, all within certain constraints, such as a fixed number of clusters.

This method of analysis was chosen to analyze the CDRs because it was speculated that each class of network application would have similar characteristics, where the data parameters are a measure of the characteristics. Moreover, each call in the network is associated with several data parameters. The data parameters can be treated as input variables for the cluster analysis procedure to determine different categories of traffic, e.g., video teleconferencing. Each traffic category is linked with *n* data parameters, providing a multidimensional categorization. Resulting clusters from the CDRs can determine how the ATM network is growing or what types of calls require higher priorities. This analytical method leads to potential network applications, such as traffic modeling or simulation that require this form of knowledge.

2.4 Summary

This chapter reviewed general terms and concepts associated with ATM networks that are needed to understand the remainder of this thesis. Background on CDRs, extraction and addition of parameters, was also presented. Further details are available from the ATM Forum in the UNI 4.0 specification [10] and ATM Forum Traffic Management Specification V4.0 [6]. For further information on ATM technology, refer to [1, 6, 10, 11].

In addition, cluster analysis was introduced in this chapter. Cluster analysis is a powerful method of analyzing data associated with multiple parameters. Thus, it was chosen to analyze the CDRs in this work. Some general cluster analysis techniques are examined in the next chapter.

Chapter 3. Cluster Analysis Approach

As explained in Chapter 2, the log of call detail records contains parameters from the ATM switches that are useful for statistical characterization of the corresponding calls. These parameters are analyzed using the SAS System to determine possible traffic categories. Separating the traffic into distinct categories can improve manageability of the network. The SAS System is an integrated system of products that enables one to access, manage, analyze, and present data [12]. This chapter discusses the data parameters involved in the analysis, the particular approach used for cluster analysis, and some initial results for the classification of traffic.

For this research, CDRs were collected from a FORE Systems ATM switch in the Virginia Tech campus backbone. The campus backbone contains switches mostly from the FORE Systems ASX series (FORE ASX-200BX and FORE ASX-1000) and has an architecture similar to that of NET.WORK.VIRGINIA. The Virginia Tech network carries a variety of traffic including video teleconferencing, local area network (LAN) emulation, and voice data. The traffic with guaranteed QoS received the most attention, as it requires more accurate methods of capacity planning. This non-best effort traffic underwent statistical analysis to determine specific application categories within the traffic. In this research, three work weeks (Monday to Friday) of data from March 15, 1999 to April 2, 1999 were analyzed to illustrate the approach and to provide some concrete, albeit limited, statistical results.

3.1 Data Parameters

Prior to the cluster analysis, CDR parameters are divided into two categories, independent and dependent variables, as shown in Table 3-1. Independent variables are those known at the time of the call setup and are used to identify the underlying traffic category. Dependent variables, on the other hand, are those subsequent measurements of variable resource utilization incurred by that call. The parameter division was made to create clusters from the independent variables and then to associate the dependent variables with the generated clusters. The clusters generated from independent variables would be more closely related to the services and not affected by values measured at the end of calls (see below).

Table 3-1. Independent and Dependent Variables for Analysis

Independent Variables	Dependent Variables
Peak Cell Rate	Call Duration
Sustainable Cell Rate	Circuit Utilization
Maximum Burst Size	Cells Received (optional)
Class of Service (optional)	Cells Transmitted (optional)

Some of the variables in Table 3-1 are labeled as “optional.” The Class of Service is labeled as such because of the possibility that all the calls in a data set are of the same Class of Service. The outcome of the cluster analysis would remain the same, in this case, regardless of inclusion of the variable in the analysis. The two dependent variables are optional because their descriptive statistics can be found from the results of the Call Duration and Circuit Utilization variables. Only the non-optional variables were used in this research to reduce redundancy.

3.2 Cluster Analysis

A cluster analysis procedure is used to classify traffic. This analysis assumes that types of traffic flows can be identified based on common parameters and traffic characteristics. The statistical analysis procedure examines the data parameters from the CDRs of ATM switches, processed as described in Chapter 2, and places the CDRs into groups or clusters. Each cluster, specifically the centroid of the cluster, represents a subset of data parameters with relatively similar properties, such as peak cell rate or maximum burst size. These clusters can then be correlated with existing traffic categories to aid in capacity planning. Some general clustering techniques are shown in Figure 3-1. These techniques are further described in the sections below.

Important issues to consider when applying clustering are scaling the variables before calculation, the particular method of cluster analysis to use, and the appropriate number of groups for a given data set.

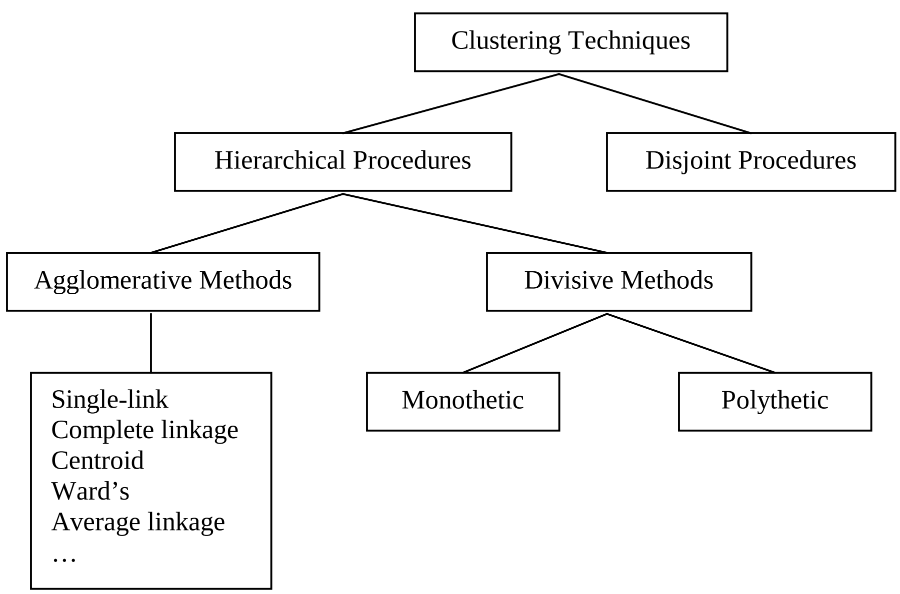

Figure 3-1. Chart of general clustering techniques.

The statistical analysis procedure is separated into four phases: (1) appropriately scaling and separating the independent variables, (2) applying a disjoint clustering procedure to the scaled independent variables, (3) applying a hierarchical clustering procedure to the results of the disjoint clustering procedure, and (4) using a canonical discriminant procedure on the output data to obtain useful plots. Each phase is discussed below.

3.2.1 Preliminary Analysis

The ATM calls can be inherently delineated as being either best effort or non-best effort. A best effort call designates a QoS class with no specified parameters and with no assurances that the traffic will be delivered across the network to the target device. ABR and UBR are both best effort service examples, with the collected data including only UBR as best effort calls. A non-best effort call indicates some form of guaranteed service for that particular call. CBR, rt-VBR, and nrt-VBR are examples of non-best effort calls, with the collected data containing only rt-VBR calls. Since new capacity planning methods are needed for this type of call, these non-best effort calls are the ones of particular interest for the statistical analysis.

The independent parameters of the non-best effort traffic define an *n*-dimensional space, where the *n* independent parameters represent the space. In particular, a 3-dimensional space (using three independent parameters) is considered; the Class of Service parameter is not used because all the non-best effort calls fall into the same Class of Service. There were 327 non-best effort calls out of a total of 25,713 recorded calls.

Cluster analysis was not performed for the best effort calls. This is due to the parameters in the CDRs for the best effort calls. The only independent parameter defined in the CDR for a best effort call is the Peak Cell Rate, which is not particularly useful for UBR traffic. The parameter is actually sometimes specified, but it is never used for capacity planning.

Since the independent parameters have different units and the variables with large variances tend to have more effect on the resulting clusters than those with small variances, each independent variable was first standardized to zero mean and unit variance before applying the clustering algorithm. The SAS STANDARD procedure [9] was used to properly normalize these data.

3.2.2 Disjoint Clustering

Following normalization, a disjoint clustering procedure is applied to the scaled independent variables. The hierarchical clustering procedure, explained below, is not practical for large data sets because with most methods the CPU time varies as the square or cube of the number of observations [9]. The disjoint clustering procedure requires time proportional to the number of observations and can, therefore, be used with much larger data sets [9]. Thus, the disjoint clustering procedure is used for a preliminary cluster analysis producing a large number of clusters and then the hierarchical clustering procedure is used to hierarchically cluster the preliminary clusters.

Disjoint clustering can be applied to coordinate data and is especially suitable for large data sets, from approximately 100 to 100,000 observations. This procedure combines an effective method for finding initial clusters with a standard iterative algorithm for minimizing the sum of squared distances from the cluster means. A set of points called “cluster seeds” is selected as a first guess of the means of the clusters. Each observation is assigned to the nearest seed to form temporary clusters, and the process is repeated until no further changes occur in the clusters. Disjoint clusters place each object in one and only one cluster.

The SAS FASTCLUS procedure [9] was used to analyze the data with the disjoint clustering algorithm. The number of observations for this analysis was relatively small (237 observations) so the maximum number of clusters allowed for this procedure was set to 10 clusters. For a larger data set, on the order of 100,000 observations, the maximum number of clusters should be increased to 100 for more accurate results. The convergence criterion for this algorithm was specified to equal zero. The iterations of the FASTCLUS procedure terminate when the maximum distance by which any seed has changed is less than or equal to the minimum distance between initial seeds multiplied by the convergence criterion [8]. The maximum number of iterations for recomputing the cluster seeds was set to 99. After all observations are assigned, the cluster seeds are replaced by the cluster means. This means that the seeds will generate clusters as accurately as possible to a final cluster or until 99 iterations have passed. Consequently, the algorithm converged to zero before reaching the maximum number of iterations and produced only six clusters. These clusters were used as input for the hierarchical clustering algorithm discussed in the next section.

3.2.3 Hierarchical Clustering

A hierarchical clustering procedure is applied to refine the initial disjoint clusters. There are two general procedures for hierarchical clustering, agglomerative and divisive [8]. The agglomerative procedure fuses *N* data points into larger groups, whereas the divisive procedure separates the *N* data points into finer groupings. The divisive procedure, in turn, contains two schemes for clustering, monothetic and polythetic. In a monothetic scheme, cluster membership is based on the presence or absence of a single characteristic. Polythetic schemes use more than one characteristic (variables).

Hierarchical clustering of observations can use many methods applied to coordinate data or distance data. The most common methods include average linkage, complete linkage, single linkage, and Ward's minimum variance method [9]. Of the hierarchical clustering methods mentioned, average linkage was chosen because it tends to join clusters with small variances and is relatively robust with respect to outliers. All methods described above are based on an agglomerate hierarchical procedure.

In the average linkage method, each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one is left. The various clustering methods differ in how the distance between two clusters is computed. Hierarchical clusters are organized so that one cluster may be entirely contained within another cluster, but no kind of overlap between clusters is allowed.

The SAS CLUSTER and TREE procedures [9] were used for the hierarchical clustering algorithm. The six clusters generated from the disjoint analysis were used as input for this particular algorithm. The CLUSTER procedure finds hierarchical clusters of observations in a SAS data set [9]. As indicated above, the average linkage method, also known as group average or the unweighted pair-group method using arithmetic averages, (UPGMA) [9], was utilized for its robustness towards small variances. In addition an index that can be used for choosing the number of clusters was selected by using the cubic clustering criterion (CCC) option [9]. The approximate expected R^2 value under the uniform null hypothesis and semipartial R^2 statistics were also chosen for this particular analysis. The semipartial R^2 statistic is the between-cluster sum of squares divided by the corrected sum of squares, and gives the decrease in the proportion of variance resulting from joining the clusters. R^2 is the proportion of variance accounted for by the clusters. In Table 3-2, the change in semipartial R^2 (SPRSQ) and R^2 (RSQ) are relatively small until the stage moving from three to two clusters (discounting the stage moving from two to one cluster). This provides some evidence for a three-cluster solution. The CCC would have offered more evidence for a three-cluster solution but the variables were relatively correlated with one another, rendering the CCC statistic invalid for this analysis.

Table 3-2. Average Linkage Cluster Analysis

Clusters	SPRSQ	RSQ	CCC
5	0.0001	1	87.92
4	0.0004	1	69.53
3	0.0238	0.976	23.74
2	0.0739	0.902	17.57
1	0.9019	0	0

The CLUSTER procedure creates an output data set that can be used by the TREE procedure to draw a tree diagram of the cluster hierarchy or to output the cluster membership at any desired level. The TREE procedure was employed to create a data set containing an additional variable (*cluster*) that indicates where each observation belongs in a specified number of clusters. The data set was then sorted by cluster for the next step of the analysis.

The statistics for the dependent variables can now be grouped with the independent variables in their appropriate clusters for further analysis. Table 3-3 provides an example of results from cluster analysis showing variables for each of the three identified clusters.

Table 3-3. Basic Descriptive Statistics for Data Variables

Cluster	N Obs	Name	Mean	Standard Deviation
1	201	Peak Cell Rate (cells/s)	3486.00	0.00
		Sustainable Cell Rate (cells/s)	3485.00	0.00
		Maximum Burst Size (cells)	16.00	0.00
		Call Duration (s)	6984.63	23989.90
		Circuit Utilization (cells/s)	3249.79	1.88
2	116	Peak Cell Rate (cells/s)	2361.02	44.85
		Sustainable Cell Rate (cells/s)	2357.74	43.78
		Maximum Burst Size (cells)	11.57	1.09
		Call Duration (s)	8922.83	8645.15
		Circuit Utilization (cells/s)	2232.57	130.72
3	10	Peak Cell Rate (cells/s)	1317.00	0.00
		Sustainable Cell Rate (cells/s)	1316.00	0.00
		Maximum Burst Size (cells)	8.00	0.00
		Call Duration (s)	2538.22	1826.67
		Circuit Utilization (cell/s)	1249.78	0.45

3.2.4 Canonical Discriminant Procedure

To obtain useful plots, a canonical discriminant procedure is applied to the clustered data. Canonical discriminant analysis is a dimension-reduction technique that derives canonical variables based on two or more groups of observations with measurements on several quantitative variables [8]. Canonical discriminant analysis derives a linear subspace of the variables that has the highest separation between the groups. The canonical discriminant procedure takes the output data from the hierarchical cluster analysis and computes the canonical variables for discriminating among the clusters. The first two canonical variables can be plotted to show cluster membership.

For each canonical correlation, the SAS procedure CANDISC [9] tests the hypothesis that it and all smaller canonical correlations are zero in the population. Both standardized and unstandardized canonical coefficients are output, as well as correlations between the canonical variables and the original variables. The results are displayed in Table 3-2 and the plot of the first two canonical variables is shown in Figure 3-2.

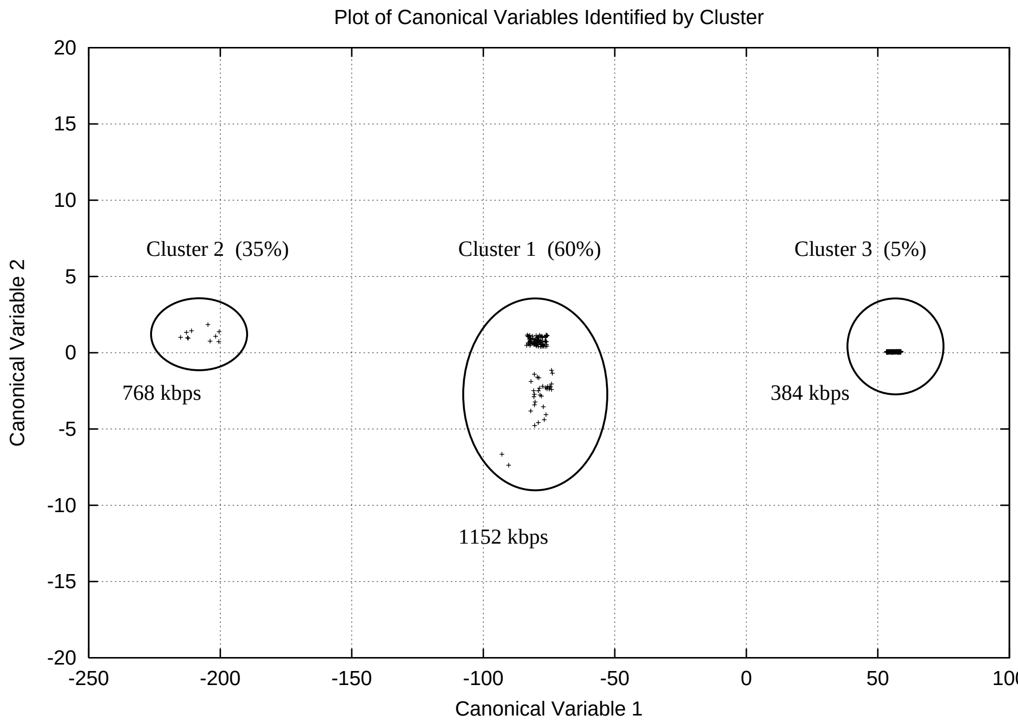

Figure 3-2. Clusters generated from non-best effort traffic.

3.3 Summary

This chapter provided an overview of the approach that was implemented for the statistical analysis of the ATM CDRs. The algorithmic procedures within the cluster analysis were discussed. CDRs were examined and statistically classified using the clustering algorithm. Per traffic category usage patterns may aid in planning for future resource needs given appropriate service objectives. This topic is further discussed in Chapters 4 and 5. The Virginia Tech campus backbone network provided a first test case for further investigation, extension, and validation of these techniques. Ultimately, the network must scale to meet expected future usage. This requires further data analysis on the traffic categories as well as the aggregate data. It is important to understand the statistical distributions of the dependent variables for the ATM calls. Specific strategies for resolving distributions are discussed in Chapter 4.

Chapter 4. Data Analysis

As discussed in Chapter 3, there are two types of ATM calls that were collected from the FORE switches. The best effort calls provide service on a first-come-first-serve basis without any guarantees for proper delivery. Cluster analysis was not applied to data for best effort calls since the network does not explicitly allocate resources for such calls. The non-best effort calls underwent this type of analysis because of their capability to offer some form of guaranteed service and the associated need to reserve network resources. After the initial clustering procedures the calls were further analyzed to determine if they fit a specific statistical distribution. This analysis could help to ascertain appropriate capacity planning methods for the matching calls. In addition, utilization was analyzed to aid in provisioning for networks.

4.1 Analysis of Best Effort Calls

Within the three work weeks of data used in this research, there exist 25,375 best effort calls out of a total of 25,702 recorded calls. Thus, the best effort calls account for nearly 99% of all calls. Once more this supports the separation of the two types of calls for individual statistical analysis. A cluster analysis based on all calls would, in essence, be a cluster analysis of the best effort calls, since the parameters of the non-best effort calls would be overshadowed by data for the best effort calls.

Table 4-1. Basic Descriptive Statistics for Best Effort Dependent Variables

Variable	Mean	Std. Deviation	Minimum	Maximum
Cells Transferred	420,873	28,973,444	0.0	3,939,890,607
Call Duration (s)	6,572	197,003	4.7	9,590,000
Circuit Utilization (cells/s)	32	366	0.0	15,008

In Table 4-1, descriptive statistics are listed for the dependent variables of the best effort calls. These statistics can be used to roughly describe the best effort traffic.

4.2 Analysis of Non-Best Effort Calls

As stated in Chapter 3, there are 327 non-best effort calls out of a total of 25,702 recorded calls. The cluster analysis procedure was used to determine traffic categories from the independent variables originating from the traffic parameters of the CDRs. This procedure was also discussed in Chapter 3. In addition to finding traffic categories, analyzing the data associated with the links of the ATM network can produce results that assist in provisioning for the network.

4.2.1 Traffic Categories

A traffic category is a portion of the network traffic that can be associated with a particular user service or network function such as voice or video. Each traffic category has certain parameters that make it unique from traffic in other categories. Understanding these parameters and how they affect the overall network traffic can provide better manageability of related services.

In this work, network traffic categories were obtained based on the results of the cluster analysis of the non-best effort CDRs. The cluster analysis yielded three clusters, which relate to three traffic categories. The centroid of each cluster establishes the mean parameters that are associated with the corresponding traffic category. Some descriptive statistics of the clusters are given in Table 3-3.

Tracing the traffic categories found from the cluster analysis back to the actual associated services in the network results in an accurate match. The three traffic categories correspond to varying data transfer rates of videoconferencing traffic. The data transfer rates are 1152 kbps, 768 kbps, and 384 kbps, which correspond to 3000 cells/s, 2000 cells/s, and 1000 cells/s, respectively, when converted into ATM traffic rates. From this knowledge, the standard circuit utilization and other parameters can be used to further analyze the individual traffic categories as well as the aggregate traffic.

4.2.2 Link Analysis

Another important aspect of capacity planning is to understand the traffic load carried by the network links. The entire network carries traffic composed of different traffic categories and best effort data, and the network consists of individual links that each carries a portion of the overall traffic. In other words, a link in the network can also be analyzed to ascertain inherent traffic categories that are carried and can be individually analyzed and managed.

The “call_out_port_vp_vc” parameter in the CDRs aided in identifying the links. The parameter has the FORE Systems CDR format of “switch fabric | slot | switch port / VPI / VCI” (e.g., 2A1/0/239) [5]. The virtual channel identifier (VCI) is a unique numerical tag, defined by a 16-bit field in the ATM cell header, that identifies a virtual channel (VC) over which a stream of cells is to travel for the duration of a connection between devices. The virtual path identifier (VPI) is an eight-bit field in the ATM cell header that indicates the virtual path (VP) over which the cell should be routed. Combined, these fields identify a connection on an ATM network.

In this data set, the links can be simply represented by the switch fabric, slot, and switch port of the Out Port VP VC parameter. One reason to support this claim is that the first VPI (xxx/0/xxxxx) is used in all the links. Four links were found using this parameter and the representation given above: 2A1, 2D3, 2D4, and 2E1. Each link was analyzed to find inherent clusters.

The non-best effort cluster analysis results from the aggregate data produced the same three traffic categories that are present throughout the network. Thus, when analyzing the links for the clusters, the maximum number of clusters in the hierarchical portion of the procedure was set to the maximum number of clusters possible in the disjoint analysis. The centroids of any clusters found can be matched to the centroids of the three traffic categories with a small percentage of difference. In this way, analysis to determine the number of clusters in the hierarchical procedure can be skipped, since only three possible clusters are present in this particular network.

The percentages of non-best effort calls in each traffic link are: 23.24% in 2A1, 48.01% in 2D3, 25.38% in 2D4, and 3.36% in 2E1. These percentages give an indication of which links are possibly over or under provisioned. In addition, Table 4-2 lists the percentages of calls in each traffic category from the corresponding links. The table shows that roughly two-thirds of the non-best effort calls within each link fall into the 1152 kbps traffic category, roughly one-third fall into the 768 kbps traffic category, and only a small percentage of the calls fall into the 384 kbps traffic category.

Table 4-2. Percentage of Non-Best Effort Calls from Links into Traffic Categories

Traffic Category	2A1	2D3	2D4	2E1
1152 kbps	61.84%	61.78%	60.24%	63.64%
768 kbps	32.89%	35.67%	37.35%	36.36%
384 kbps	5.26%	2.55%	2.41%	0%

4.3 Call Distributions

Knowing the traffic categories within the network and the parameters associated with them, and identifying the links are not enough for proper network management. The links, traffic categories, as well as the aggregate data need to be analyzed to determine statistical distributions, if they exist. Understanding these distributions can assist in predicting or managing a class of traffic or even the entire network.

4.3.1 Characterizing Data

Characterizing the calls in the network involves finding the distributions of certain inherent data parameters. Interarrival times, holding times, and circuit utilization are three useful data parameters where distributions may be able to be found. Sorting the CDRs by Call Start Time and calculating the differences between adjacent Call Start Times can easily yield the interarrival times. The holding times are simply the Call Durations of the CDRs.

Many statistical tests are used to help in determining statistical distributions of data. Two of the most common tests are discussed in the next two subsections of this chapter: the normality test [12, 13] and the exponential test [13]. These tests are not necessarily sufficient to declare a particular type of distribution, but they are a good first choice of tests.

4.3.2 Normality Test

The SAS UNIVARIATE procedure [12] was used to test for normality. With a random sample from a Gaussian population, the probability of obtaining a sample that does not deviate from a Gaussian distribution as much as this sample does is the p-value from the normality test. By looking at the distribution of a small sample of data, it is hard to tell whether or not the values came from a Gaussian distribution. Running a formal test does not make it easier. The tests simply have little power to discriminate between Gaussian and non-Gaussian populations with small sample sizes. The test does not really have much power to detect deviations from Gaussian distribution unless there are around 100 values. Thus, the interpretation of the results of a normality test depends on the p-value calculated by the test and on the sample size.

The normality test shows that the interarrival times, holding times, and circuit utilizations of the traffic categories, links, and aggregate data do not have Gaussian distributions. The results of these analyses can be found in Appendix A.1.

4.3.3 Exponential Test

Of the tests for exponential distributions, the Chi-Square test formalizes the idea of frequency comparisons [13]. It is an approximate test but is widely used because of its universal applicability and relative simplicity. The test is comprised of the following steps.

Step 1: Divide the entire range into k adjacent intervals $[a_0, a_1)$, $[a_1, a_2)$, ...$[a_{k-1}, a_k)$.
Let N_j = number of data values in the j^{th} interval $[a_{j-1}, a_j)$. This is the *observed* frequency for this interval.

Step 2: From the theoretical (hypothesized) distribution one can calculate the probability that $a_{j-1} <= X < a_j$. Let $p_j = P\ (a_{j-1} <= X < a_j)$ be this probability.
Then np_j = the *expected* number of observations in the j^{th} interval.

Step 3: Calculate the Chi-Square statistic.

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

If the fit is good the Chi-Square value will be small (e.g., $\chi^2 \le 5$ with less than 10 degrees of freedom).

<u>Case 1</u> - All parameters are known. If none of the parameters of the hypothesized distribution were estimated from the observed data, then Chi-Square has an χ^2_{k-1} approximate distribution with *(k-1)* degrees of freedom.

<u>Case 2</u> - If m of the parameters of the hypothesized distribution were estimated from the hypothesized distribution then Chi-Square has an approximate χ^2_{k-m-1} distribution with *(k-m-1)* degrees of freedom.

Step 4: Postulate the Null hypothesis
H_0: Sample data are from the hypothesized distribution

Step 5: Select the significance level - α
Reject the null hypothesis H_0 if

$$\chi^2 > \chi^2_{d,f,1-\alpha}$$

The significance level says that there is a 100 α% probability of rejecting the null hypothesis when it is true, i.e., the hypothesized distribution is correct.

A program, *ChiSquared*, was written in C++ to formulate the procedure discussed above for Pearson's Chi-Square goodness of fit test. All of the data parameters in this research that were tested using the program did not fit to an exponential distribution. The test results can be found in Appendix A.1.

4.3.4 Analysis of Distributions

Various methods of network management or capacity planning can be applied to traffic with known distributions. For example, network traffic that fits a Poisson distribution can apply Poisson process queuing systems and use Erlang's formulas to predict blocking probability or expected delay [14]. Unfortunately, the collected data do not appear to fit common distributions and are deemed to be largely deterministic. This makes sense since the non-best effort calls are made according to an academic schedule for video conferencing or distance learning to and from Virginia Tech. The interarrival times increase overnight as there is not any non-best effort activity. However, in some networks, such as large MANs or WANs, it would be difficult and time consuming to trace each call back to its matching original service.

Devising a capacity planning strategy for growth in deterministic traffic can prove to be more difficult than with traffic that fit into a distribution type. The next section discusses one method for managing non-best effort traffic.

4.4 Utilization Analysis

The Circuit Utilization parameter from the ATM CDRs can be analyzed to support provisioning in operational networks. The total utilization at a given point in time represents the traffic intensity of non-best effort calls. Using the traffic intensities, one can create a level or threshold for managing the non-best effort calls based on utilization. In other words, a threshold of utilization can act similarly to a service level with similar characteristics and parameters. This topic is discussed in detail below.

4.4.1 Traffic Intensity

The traffic intensities must be expressed as a function of time before further analysis. A C++ program, *Traffic*, was written for this specific task. The *Traffic* program receives input from the collected CDRs. It reads the Call Start Time, Call End Time, and Circuit Utilization parameters into a linked list. The Call Start Time and Circuit Utilization parameters are placed into one item on the list and the Call End Time parameter and the same Circuit Utilization parameter are placed into another item on the list. This is repeated for every call until these parameters from all the non-best effort calls are stored in the linked list. The items in the list now contain a Time parameter (stored as a *long* data type) and a Utilization parameter (typecast as a *long* data type). In addition, the Utilization parameter within the same list item as the Call End Time is made negative. This is done to calculate the total utilization or traffic intensity at a given time. The negative value is added to the total utilization value at the Call End Time, decreasing the resources used because the call has just ended. The list is then sorted by the Time parameter. Finally, the list is modified to calculate the overall Utilization value. The first Time and Utilization parameters are placed in the list, then the next Time parameter is already the next item, since the items were sorted by Time, and the matching Utilization value is added to the previous Utilization parameters and the result is kept. This is repeated until all the items have been modified. The resulting list now contains coordinates to plot traffic intensity as a function of time. Figure 4-1 shows a process diagram, with examples, for finding the traffic intensity.

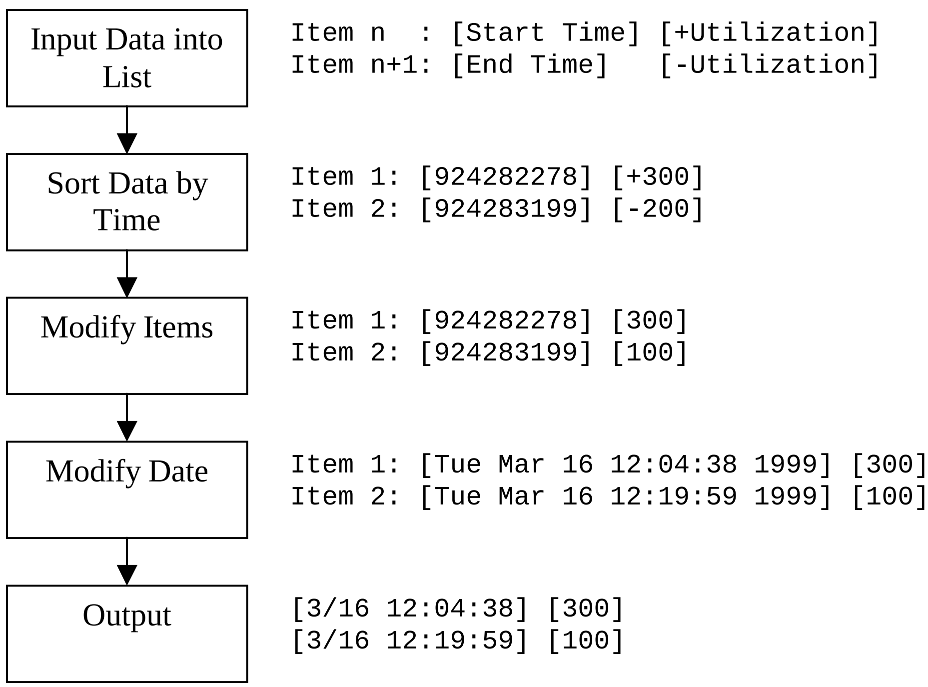

Figure 4-1. Process to determine traffic intensity.

The output of the *Traffic* program is a two-column list that can be plotted with the timestamp converted from epoch format to a more readable human format. However, the plot would be meaningless to the human eye since the Time parameters are represented as the amount of time in seconds that has passed since the UNIX time epoch of 1 January 1970. These large numbers need to be converted to a more readable format. A simple C++ program, *TConvert*, was written to complete this job. The *ctime* procedure is used to convert the Time parameter into a more readable format. The results of the *TConvert* program can finally be plotted.

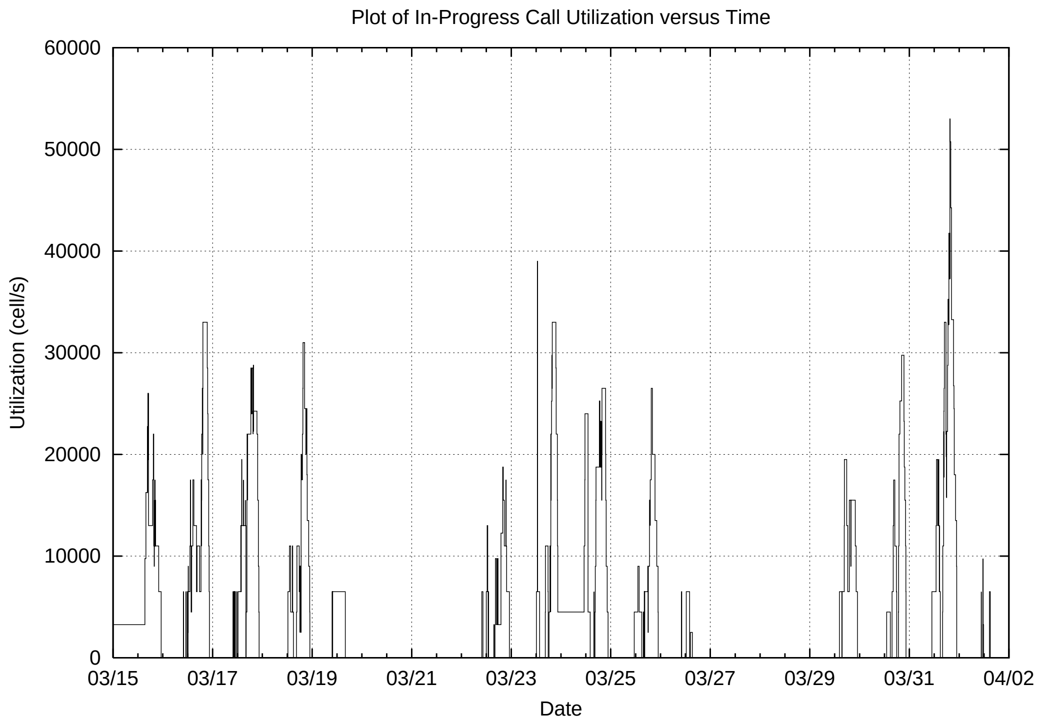

Figure 4-2. Traffic intensity plot of aggregate traffic.

The step plot displayed in Figure 4-2 represents aggregate traffic intensity as a function of time and was plotted using GNUPLOT [15]. Traffic intensity plots for individual traffic categories can be found in Appendix A.2.

4.4.2 Threshold Prediction

Once the traffic intensities are found as a function of time, they can be analyzed to determine a threshold or level of utilization for future traffic. In other words, analysis of a traffic sample period can produce a utilization rate bound such that future traffic intensities will likely fall under this bound with a minimal difference. The algorithm described below attempts to predict a limit for future traffic intensities based on a period for analysis. The prediction algorithm used for this analysis was modified from the one presented in [16].

Prediction Algorithm: Let $U(t)$ be the utilization at time t. The utilization after one time step in the future, i.e., at time t+1, is estimated from $U(t)$, as given by

$$\hat{U}(t+1) = a(t)U(t)$$

where $a(t)$ is an estimated weight factor at time t.

The error of the prediction at time t is

$$e(t) = U(t) - \hat{U}(t)$$

The prediction scheme uses the error to modify the weighting factor whenever the error is available at each time step. A modified normalized least mean square error (NLMS) prediction algorithm is used to estimate the weighting factor. Given an initial value of $a(0) = 0$, the weighting factors are updated by

$$a(t) = a(t-1) + \frac{e(t)}{\mu^{(m)}(N)}$$

and

$$\mu^{(m)}(N) = \frac{1}{m} \sum_{j=N-m+1}^{N} \mu(j)$$

where $\mu^{(m)}(N)$ is the mean utilization using the Moving Average [16] method for the prediction, with m being the number of time steps used in the moving average. In this case, $m = 3$ so the last three utilization values are averaged to help estimate the weighting factor. Smaller values of m would be less effective in smoothing the prediction from sudden changes in utilization, but larger values of m would require more processing time.

For this research, only three work weeks of data were available for analysis. To test the prediction algorithm, data from the first two weeks were analyzed and the results were checked using data from the third week. The data were first manipulated into even time steps with a given utilization value at each time step. A time-weighted average within each time step was used to calculate the utilization values. The time step values of 5 minutes, 10 minutes, and 60 minutes were used. The Prediction Algorithm with the Moving Average Estimator was then executed at each time step. Thus, the algorithm predicted the utilization values of the next time step. The predictions were used as upper bounds for utilization. The results for the 5-minute time step are shown below in Table 4-3. The remaining results can be found in Appendix A.3.

Table 4-3. Aggregate Data Analysis Results Using 5-Minute Time Step

Analysis Time	**Bound (cells/s)**	**Percent Time Exceeded (Activity)**	$\sum \frac{\lvert x-\hat{x}\rvert}{N}$	**Over Utilization Rate (cells/s)**	**Excess Capacity Rate (cells/s)**	**Percent Time Exceeded (Total)**
Median	10000	50%	7038	–	–	–
1st day	31519	5.73%	19240	241.0606	27650	1.78%
1st week	40049	0.66%	27404	63.8676	36127	0.20%
1st 2 weeks	50749	0.33%	37977	7.0158	46812	0.10%

The column for Analysis Time in Table 4-3 represents the amount of time slots the prediction algorithm analyzed. For example, the entry of "1st day" denotes that the non-best effort calls within the first day were analyzed using the prediction algorithm. The Bounds predicted from the Analysis Times of "1st day, 1st week, and 1st 2 weeks" are shown in Figure 4-3. The Median entry was used as a "control" for the results produced by the prediction algorithm analysis.

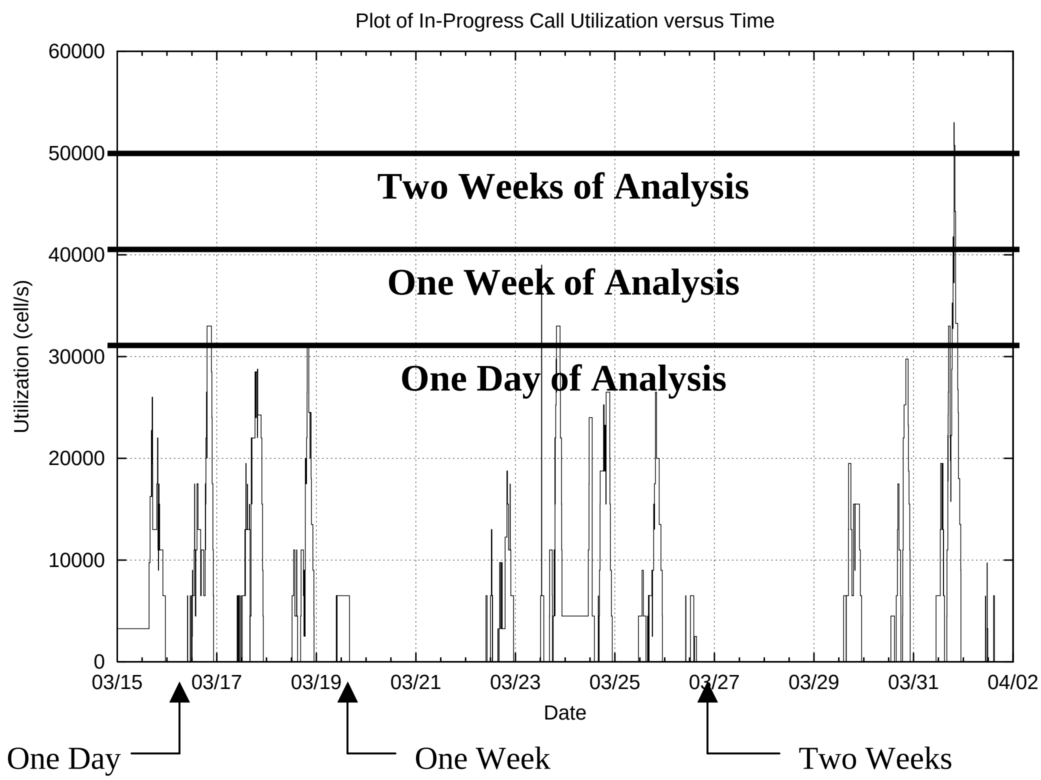

Figure 4-3. Traffic intensity plot showing predicted bounds.

The Bounds, or predicted maximum utilization values, generated from these Analysis Times were tested on the remaining time slots to find the other parameters in Table 4-3. In Figure 4-4, a sample Bound is shown with a traffic intensity plot. The values of the Bounds were based on the variance of the utilization rates within the analyzed time slots. A large variance in a set of time slots would lead to a large Bound. These Bounds can be used for creating Service Level Agreements.

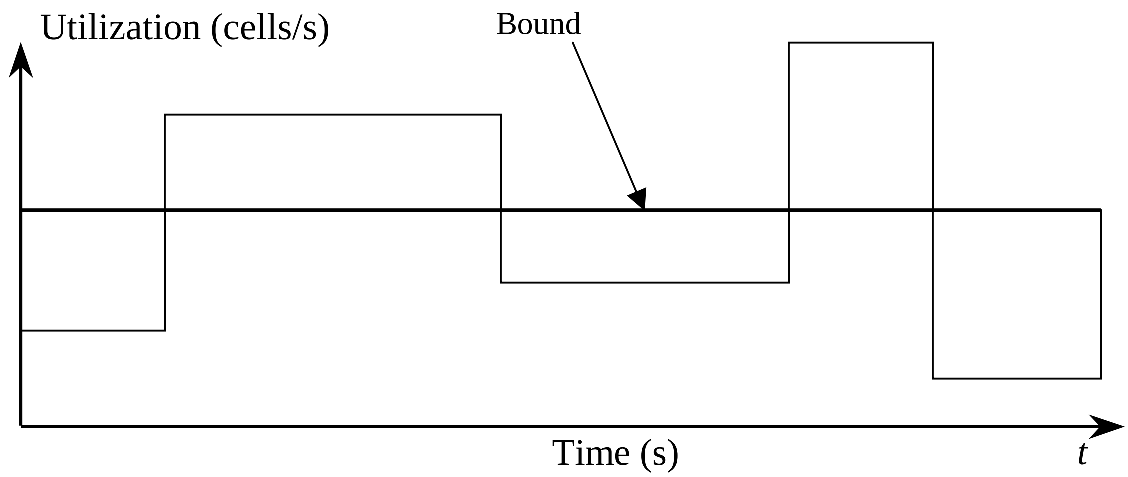

Figure 4-4. Traffic intensity plot with bound.

The Percent Time Exceeded (Activity and Total) designates the percent of time slots that contain utilizations that exceed the Bounds. The Percent Time Exceeded (Total) parameter is found by dividing the number of time slots with utilization values exceeding the Bound by the total number of time slots. The Percent Time Exceeded (Activity) parameter is calculated similarly, except instead of using the total number of time slots, the number of time slots with utilization values greater than zero is used. A traffic intensity plot with time slots corresponding to utilization values that exceed a Bound is shown in Figure 4-5. The number of the "over Bound" slots divided by the total time slots in *t*, represented in the figure, would result in the Percent Time Exceeded value.

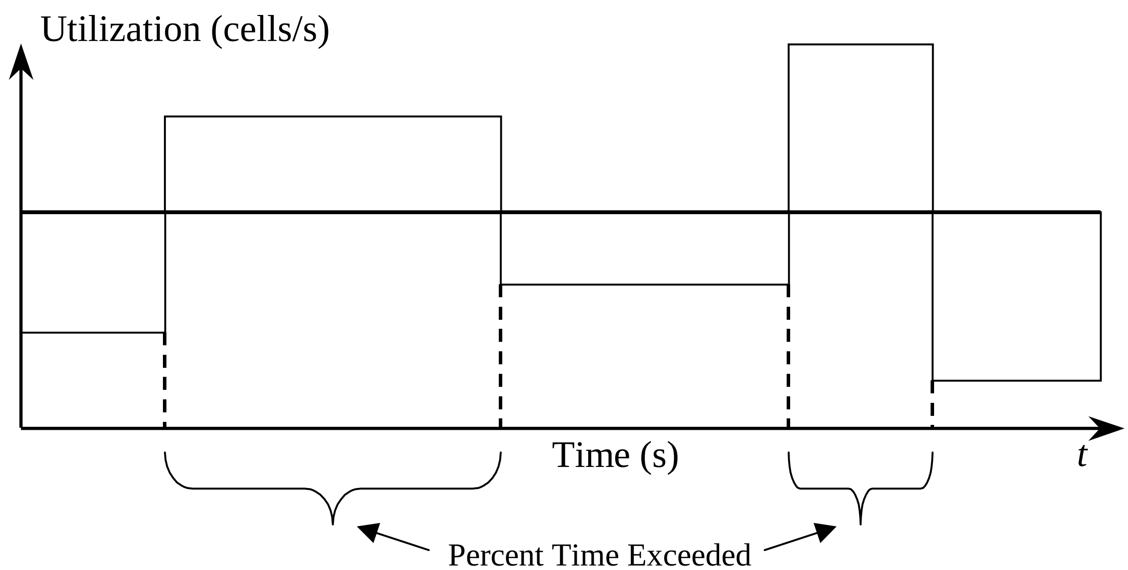

Figure 4-5. Calculation of percent time exceeded.

The formula of

$$L_1 = \sum \frac{|x - \hat{x}|}{N}$$

represents an error calculation based on the Bound values and number of time slots.

The Over Utilization Rate, given in cells per second, is the rate at which cells will not make the designated service levels. To find this value, the curve of traffic intensity versus time surpassing the bound is first integrated, and the result is divided by the total number of time slots (Figure 4-6). This data parameter can be used to determine sudden increases in network traffic.

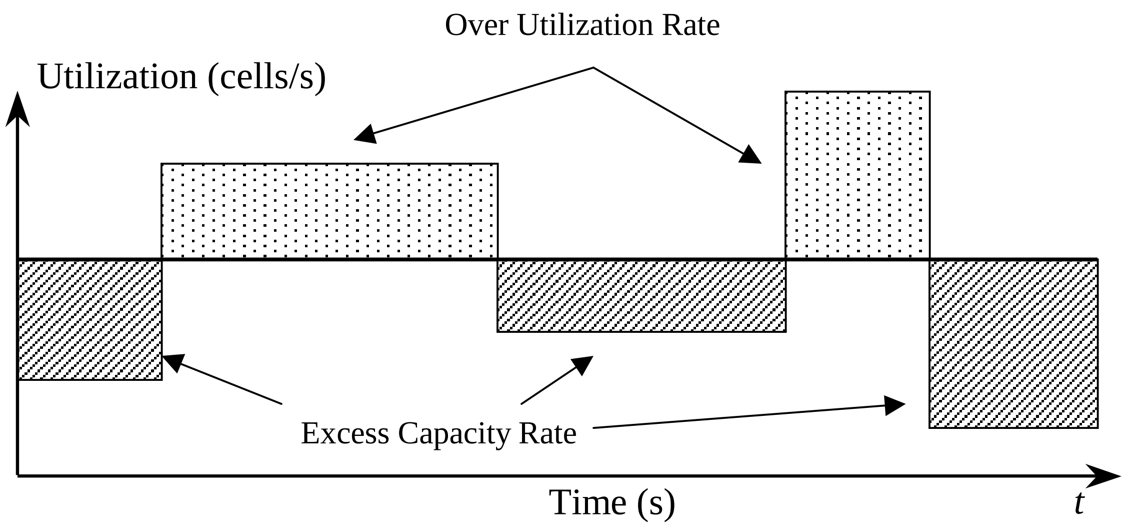

Figure 4-6. Calculation of over utilization rate and excess capacity rate.

The Excess Capacity Rate is the rate at which cells undershoot the designated service levels. To determine the Excess Capacity Rate, the number of cells between the level and the traffic intensity curve beneath it is first summed, and the result is then divided by the total number of time slots to determine the cell rate of excess capacity (Figure 4-6).

4.5 Summary

This chapter presented results from analyzing the best effort and non-best effort calls. Cluster analysis was not applied to data for best effort calls because such calls do not provide guaranteed services, do not require resource allocations and do not have relevant data parameters. Analysis of best effort calls was also presented in this chapter. The non-best effort calls and their inherent traffic categories were analyzed for possible statistical distributions. However, the data clearly failed the tests to determine distributions. In addition, traffic intensity thresholds were performed for these data. Some possible tight bounds for the traffic intensities were found to work particularly well. These bounds and other results can be applied to the creation and monitoring of Service Level Agreements. This topic is discussed in detail in Chapter 5.

Chapter 5. Applications for Network Management

The results of the data analysis presented in Chapter 4 point to interesting potential applications that are directly related to network management. The potential network management applications derived from this research include network capacity planning, configuration management, and traffic modeling. Relevant components from each application will be discussed in this chapter to link the analytical results of this work to possible practical applications in network management.

5.1 Capacity Planning

Capacity planning for networks involves monitoring of workloads, utilizations, and response times and predicting traffic growth so as to assess the future capacity needs and associated recommendations for equipment or service purchases or tuning [17]. The difficulties in quantifying user needs and demand for resources is one of the main challenges involved in improving network capacity planning [18]. It is time-consuming and labor-intensive to plan the changes required to keep a communication network effective as a strategic asset while maintaining acceptable expenses.

In the environment today that emphasizes quality of service, the ultimate goal of network capacity planning for a service provider or network operator is to meet service level agreements (SLA) using optimal resource capacity at reasonable cost. The ultimate goals of network capacity planning for a service consumer is to purchase a communications services with an associated SLA that meets organizational needs at minimum cost and to ensure that the contracted SLA is being met by the service provider. We focus here on capacity planning from the perspective of the network operator. Network optimization can be viewed as the process of balancing design factors to attain the best network configuration to minimize cost within the constraints including availability and performance. The general process of network capacity planning is explained in the following sections.

5.1.1 Determining and Quantifying Current Workload

The first step in capacity planning is to measure the current workload and decompose it into components that represent discrete applications or services. This first phase in planning network capacity should be repeated periodically, e.g., semiannually, quarterly, or as needed, and followed by a comparison of actual versus projected resource consumption for the period. Measurement tools and procedures need to be identified and, if necessary, developed and documented so that the characteristics of the current workload can be understood.

The regular collection of data is necessary to determine the present workload of the network. As explained in Chapter 2, the tool described in [4] was used to collect CDRs from an ATM switch every five minutes throughout three work weeks. After this initial step, the relevant data was extracted from the CDRs using another tool [4].

In Chapter 3, cluster analysis was used as a technique to understand the traffic in the ATM network. This procedure separated the workload into distinct traffic categories. The traffic categories resulting from the cluster analysis can be correlated with applications or services. The cluster analysis procedure can be used or modified as seen fit to determine and quantify the current network traffic for this phase of the capacity planning process.

Some descriptive statistics were calculated to help quantify the best effort traffic. The statistics include mean, standard deviation, minimum value and maximum value. These data can be found in Table 4-1.

The non-best effort traffic was successfully separated into three traffic categories. Each category was traced to a distinct multimedia service, and these services corresponded to three different videoconferencing traffic data transfer rates as specified in Chapter 4. Thus, the network workload can be organized into some best effort traffic and three non-best effort services for the first phase of capacity planning.

For non-best effort traffic, the resulting clusters may not completely divide the traffic into distinct categories; some overlap is possible in the hierarchical analysis. To help address this possible problem, a segment of the network, or the entire network, can be analyzed to trace the calls back to original services and then compared with cluster analysis results of this particular segment. If overlap is negligible then the cluster analysis procedure may be used for this phase.

5.1.2 Projecting Future Workload

This phase requires the planning analyst to communicate with the end-user community, with application developers, and with strategic business planners to obtain input regarding future application workload. The following are some relevant techniques for performance prediction [18].

Analytical Modeling: These models are actually programs using analytic queuing equations for predicting performance of a design. Heuristics are employed to reduce the number of possible network designs to a manageable level. Most frequently, segments of networks are evaluated separately. The model integration happens in the final design and optimization step. Usually, the integration has to be accomplished manually.

Traffic Engineering: Voice-related load estimates and Erlang B equations for sizing communication facilities are techniques usually used. These techniques can be automated by using pre-calculated spreadsheets and tables.

Simulation: The estimated traffic is coupled with a discrete event simulation model representing the topology and network element attributes. These attributes include speed, estimated queuing and processing delays, distance propagation, turnaround times, segmentation of message units, and overhead due to control characters. Simulation may be very accurate, but it requires considerable preparation and long run times. Conversely, when there is too much simplification, the results may not be accurate enough for proper capacity planning.

As presented in Chapter 4, the analysis of network utilization produced results that aid the analyst in comprehending and predicting traffic. These results can be used in an analytical model, along with proper network designs, to project future network traffic. For example, the Bounds and parameters derived (i.e., Percent Time Exceeded or Over Utilization Rate) can be used to predict new additions to the current network. If similar traffic categories were to be used in the new network, then these utilization parameters would assist in projecting the future workload.

A traffic category can also have similar characteristics with traffic load for Plain Old Telephone Service (POTS) or other audio applications. In this case, the traffic engineering technique can be applied to the variables linked with the traffic category. This “audio traffic category” can be analyzed similarly to how telephone companies analyze traffic in their network [14].

The results of this work can be applied to the techniques for performance prediction. Each technique can utilize the results or combinations of techniques can be used to project future workload. For example, an analytical model can be created and tested in conjunction with implementing Erlang equations for audio applications. The utilization parameters can then be used to support simulations.

In terms of analytical modeling, the results of the utilization analysis in Chapter 4 show some rudimentary performance predictions of the traffic. The threshold prediction algorithm improves resource prediction of the aggregate traffic if more analysis time is allowed (shown in Table 5-1). The size of the time step also affects the prediction. A larger time step would be more inaccurate (60 minutes), but a smaller one would require more processing time for the prediction. Future traffic is not expected to exceed the predicted threshold value of utilization.

Table 5-1. Accuracy of Threshold Prediction Algorithm

Analysis Time	**Percent Time Exceeded [5-minute time step]**	**Percent Time Exceeded [10-minute time step]**	**Percent Time Exceeded [60-minute time step]**
1st day	1.78%	1.72%	9.83%
1st week	0.20%	0.20%	0.49%
1st 2 weeks	0.10%	0.20%	0%

5.2 Configuration Management

Configuration management ensures the reliability of all of the components necessary to support the delivered system throughout the project life cycle [19, 20]. It is a set of middle- and long-range activities for controlling physical, electrical, and logical inventories; maintaining vendor files and trouble tickets; supporting provisioning and order processing; defining and supervising service level agreements; managing changes;

and distributing software [18]. The analysis and results of this research can be employed to create and monitor service level agreements.

5.2.1 Service Level Agreements

Network administrators are expected to evaluate service levels over a long period of time. Service level management involves the application of a standard methodology to ensure that commitments for customer service levels are consistently met. The effect is increased planning and decreased "crisis" management. An SLA is, most simply, a formal written contract. Simple or elaborate, service level agreements contain the following elements [21].

- Identification of the contracting parties
- A description of the work to be processed, including type, volume, mix, and time of arrival
- The service levels to be provided, including response time, turnaround time, deadlines, accuracy, and availability
- A performance-reporting procedure with frequency and types of reports to be provided to users and data-center management specified
- Penalties for noncompliance
- Provisions for modifying the agreement
- An expiration date

Of the elements listed above, the service levels will be reviewed in detail and the results of Chapter 4 will be applied to this topic.

5.2.2 Service Levels

The areas of service that matter most to users are responsiveness, accuracy, and availability. All time measures should be reported and discussed in customer-perceived terms. This means that response time should not be the internal response time provided by software tools, but the time for the response to return to the terminal. The customer cares little whether the cause of poor response is due to mismanagement of network performance, improperly allocated main storage, or hardware problems. It is critical to define when a job or transaction is complete.

Responsiveness varies by need and purpose. On-line work is measured by response time, usually in seconds [18]. The key to constructing an achievable response-time objective is to build in a margin of safety. A few high values of response time could skew the average above the acceptable level. Instead, an average response range and the percentage of time that the average will be met should be specified.

Turnaround time refers to the length of time it takes to complete a job after the job is input to the system. As with response time, service objectives should be expressed in averages to be met a certain percentage of the time.

Availability refers to the proportion of time the network services can be used. In most instances, this measure is meaningful only for on-line work but is a critical issue if the customer is being charged for use of equipment on a regular basis.

Accuracy, also expressed as a percentage of time, refers to the number of jobs run without errors. Errors include mounting the wrong tape, printing on incorrect forms, retransmission reruns, and so on. Information transmission accuracy is becoming an essential part of overall accuracy.

One important element in an SLA is the workload description [21], or the expected volume of activity associated with each traffic category. The traffic categories should each be characterized in terms of resource consumption as well as having an overall allocation for the aggregate data. In this work, the three traffic categories have allocation bounds that are close to the aggregate allocation bound; however, if other traffic categories are introduced into the network, then the parameters might change. Changes in parameters such as average utilization per traffic category or other statistics gathered from the centroids of the corresponding clusters could lead to large differences in expected utilizations of each traffic category.

The results of the prediction algorithm described in Chapter 4 can be utilized to set certain levels of availability. Since performance reports should be scheduled regularly, the workload and service can be compared to determine if missed service levels were due to excess work or inadequate performance on the part of the network. The predicted utilization levels represent an estimated bound for over utilization. The expected over utilization values for the predicted workloads are relatively small, as shown in Table 5-2, implying high levels of availability and accuracy for this network. The over utilization rates are significantly lower than the utilization bounds, which also support the above conjecture. Using more analysis time improves the utilization bounds, which decreases the over utilization rate. However, as stated above, larger time steps provide a more inaccurate predicted bound and leads to larger over utilization rates.

Table 5-2. Comparison of Over Utilization Rates

Analysis Time	**Over Utilization Rate (cells/s) [5-minute time step]**	**Over Utilization Rate (cells/s) [10-minute time step]**	**Over Utilization Rate (cells/s) [60-minute time step]**
1st day	241.0606	230.9076	2174.987
1st week	63.8676	64.0552	94.3467
1st 2 weeks	7.0158	64.0552	0

5.3 Traffic Modeling

Traffic modeling is intended to enable statistical characterization (modeling) for further analysis and for use in simulation models. The focus is to study the characteristics of user traffic in communication or computer networks, such as the Internet or telephone switching systems.

5.3.1 Introduction

Information, whether voice, video, or traditional data, is packed into fixed-size cells or variable-size packets, and is forwarded through telecommunication links, satellites, switches, and other components of networks. Every component must be designed to optimize the performance of the network and to guarantee a specific quality of service to the user. In addition, the network components must avoid congestion in other components and links resulting from variation in the rate of the flow of information. The knowledge of the characteristics of the traffic flow is important in network design. One can simulate, analyze, and predict the behavior of the networks under different conditions and take appropriate measures in case of any problem. Simulation efficiency and the ability to reflect realistic behavior are becoming extremely important. Traffic patterns and services are continuously changing in the current networks.

There are several different models proposed for every type of traffic. The following list includes some of the most commonly used models for different types of traffic [22].

- Number of busy circuits (Calls): Poisson model
- PCM voice telephony source: On-Off source, Interrupted Poisson (IPP)
- Aggregate of voice sources: Markov-Modulated Poisson Process (MMPP)
- Variable-rate Video: MMPP, ARMA
- Bursty Data Traffic: Self-similar processes like FBM (Fractional Brownian Motion) and Pareto-Modulated process

5.3.2 Aggregate and Categorical Distributions

The output of performance models depends on the accuracy of the modeling assumptions, including those concerning the traffic arrival process. A realistic characterization of packet traffic is an essential prerequisite to developing accurate traffic models. Accurate packet traffic models are also essential to evaluate the relative merits of packet switching with alternatives such as circuit switching. Furthermore, without an understanding of packet traffic the planning and dimensioning of packet networks cannot be properly achieved.

The analysis of statistical distributions associated with calls discussed in Chapter 4 produced results that indicate that the network calls are largely deterministic, or at least cannot be characterized by common distributions. The occurrence of most of the calls can be predicted since they follow a schedule, as demonstrated in Figure 4-2. The large gaps or interarrival times between resource consumption represent nights and weekends of no non-best effort activity. This is a useful result, as it points out that a trace or deterministic model is reasonable for this network while, for example, a Poisson model is not.

It was also shown in Chapter 3 that the independent parameters of the non-best effort traffic define an n-dimensional space. The cluster analysis procedure essentially compresses the traffic into a one-dimensional space (the clusters themselves). This technique can be applied as an initial step to traffic modeling. Since the cluster analysis procedure can separate the calls into traffic categories linked to a particular service, the categories can be individually modeled. This could simplify the modeling procedures because services can be modeled separately instead of as a whole. However, a comparison of how well the categories and the aggregate data are modeled can only be verified through simulation and evaluation of actual results.

The three non-best effort traffic categories correspond to the video data transfer rates of 1152 kbps, 768 kbps, and 384 kbps. Each of these services can be modeled independently or together for simulation purposes (Figure 5-1). In this case, the 1152 kbps service had a much larger impact on the aggregate data than the 768 kbps or 384 kbps video data transfer rates. Problems occurring with the services associated with the lower rates would be less noticeable on the aggregate data model. However, an increase in usage of a particular service can still affect the entire network. Through modeling and simulating, it can be determined whether such a change would critically reduce performance or availability.

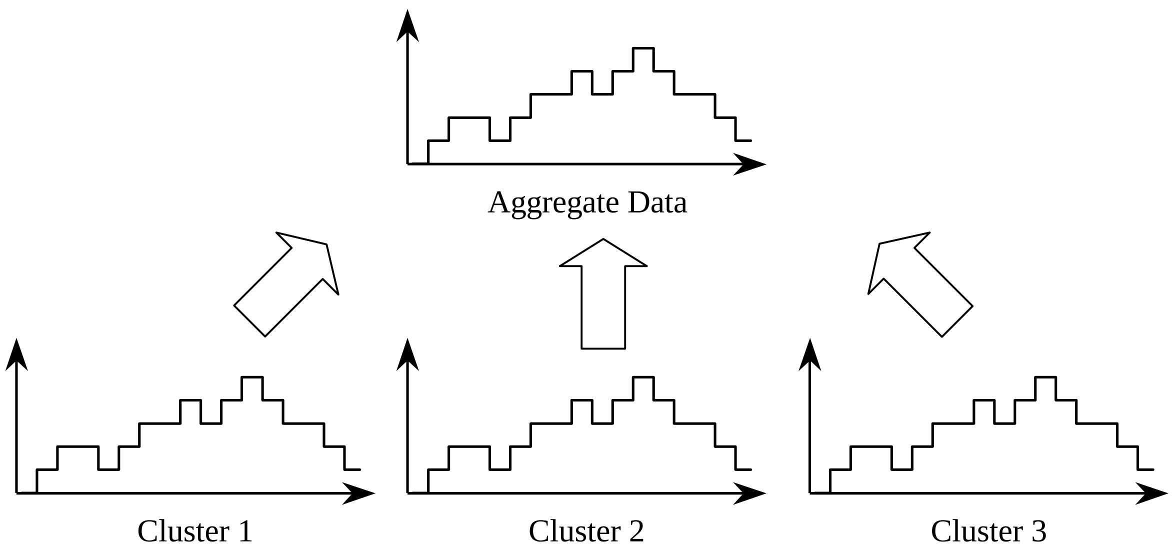

Figure 5-1. Aggregate and categorical distributions for traffic modeling.

5.4 Summary

Three interesting potential applications directly related to network management were discussed in this chapter. Methods for network capacity planning, configuration management, and traffic modeling can be developed based on results of the previous chapters. The analytical results of this research were linked to the practical pertinent components or processes of each application.

In terms of capacity planning, the workload is a necessary component that needs to be analyzed and projected for future development. The cluster analysis technique has proven to help to structure and quantify these workloads. In addition, cluster analysis can potentially be used to simplify traffic modeling procedures. Utilization and data analysis from Chapter 4 provide useful results to create, monitor, and update service level agreements for configuration management.

Chapter 6. Conclusions

Network management of multimedia classes constitutes an important area of research and has the potential to improve quality of service for related network services used today. Networks that support dynamically allocated resources may carry many types of traffic including data, time-division multiplexed calls, and circuit-switched voice as well as newer types of traffic, such as for distributed simulation or videoconferencing applications. Classical methods of capacity planning do not work well for these types of networks. Traffic flows are created on demand and there is no way to know the resources that the flow will allocate ahead of time. The approach for resolving multimedia traffic is the focus of this work. This chapter summarizes the thesis, discusses objectives, and outlines future areas of research.

6.1 Objectives

The first objective of this research was to use a cluster analysis procedure to investigate potential strategies for the statistical characterization of CDRs in an ATM network. This investigation included determining the data parameters and statistical cluster analysis procedure, as presented in Chapter 3. The second objective of this research was to analyze the results of the statistical procedure and determine distributions of the data, as discussed in Chapter 4. The third and final objective of this research was to associate selected portions of practical network management applications to the results. This association is presented in Chapter 5.

6.2 Approach

The research approach was to determine traffic categories within the ATM calls that are linked back to actual services, as described in Chapter 3. Once the traffic categories were created, the data analysis approach of Chapter 4 was used to investigate the traffic categories. This investigation included determination of the appropriate distributions and evaluation of traffic utilizations. The assumptions of the traffic parameters, proposed approach, and evaluation methodology are presented in the thesis.

6.3 Results

The cluster analysis of the ATM CDRs, as presented in Chapters 3, was based on three variables, Peak Cell Rate, Sustainable Cell Rate, and Maximum Burst Size. For the data collected for this research, three traffic categories resulted from the cluster analysis procedure that define three different videoconferencing speeds in the Virginia Tech campus network, 1152 kbps, 768 kbps, and 384 kbps. Unfortunately, the distribution analysis of the traffic categories and aggregate traffic, discussed in Chapter 4, provide a more or less negative result for the collected data. The traffic was characterized as being largely deterministic, or at least not fitting a common distribution, and would need to be modeled with a deterministic or trace model. However, the analysis of utilization produced results that could prove useful for some network management applications.

These applications include capacity planning, configuration management, and traffic modeling, and the relevant components of each network application were portrayed in Chapter 5.

Determining and quantifying the workload and projecting future traffic are components of capacity planning that are related to this work. The service levels in configuration management can be correlated to the results of the utilization analysis in Chapter 4. Finally, traffic modeling of the network can be matched with the clustering and distribution analyses for statistical characterization and for use in simulation models.

6.4 Future Research

Data analysis for network management of multimedia classes of service represents a relatively new area of research and is likely to be an important research area for some time to come. This work focused on an approach that can be used for potential network management and did not directly investigate many related issues. Possible future areas of research include: (1) real-time calculation of statistical parameters while extracting data, (2) investigation of how the results can support a network, (3) research into how this approach could be applied to other architectures, and (4) development of more precise models of traffic categories and distributions.

In the future, CDR formats can be standardized such that CDRs may be used by a large class of devices that support guaranteed services, as suggested in [4]. Usage patterns within traffic categories may aid in planning for future resource needs given appropriate service objectives. NET.WORK.VIRGINIA can provide a large-scale first test case for further investigation, extension, and validation of these techniques.

References

[1] D. McDysan and D. Spohn, *ATM Theory and Application*s, McGraw-Hill, Hightstown, N.J., 1998.

[2] I. Mahadevan and K.M. Sivalingam, "Quality of Service Architectures for Wireless Networks: IntServ and DiffServ Models," *Fourth International Symposium on Parallel Architectures, Algorithms, and Networks*, I-SPAN, 1999, pp. 420-425.

[3] X. Xiao and L.M. Ni, "Internet QoS: A Big Picture," *IEEE Network*, Vol. 13, No. 2, Mar.-Apr. 1999, pp. 8-18.

[4] X.R. Wang, "Analysis of ATM Call Detail Records and Recommendations for Standards," M.S. Thesis, Virginia Polytechnic Institute and State University, June 1999.

[5] FORE Systems, Inc., *Call Data Record (CDR) Detail*, http://www.fore.com/products/fv/so/Accountant/callrecorddetail.htm.

[6] ATM Forum, *Traffic Management Specification*, *Version 4.0*, April 1996, ftp://ftp.atmforum.com/pub/.

[7] FORE Systems, Inc. *AMI Configuration Commands Reference Manual*, March 1999.

[8] B.S. Everitt and G. Der, *A Handbook of Statistical Analyses using SAS*, Chapman & Hall, New York, N.Y., 1996, pp. 115-127.

[9] *SAS Institute Inc.*, *SAS/STAT User's Guide*, *Version 6*, *Fourth Edition*, *Volume 1*, SAS Institute Inc., Cary, N.C., 1989.

[10] ATM Forum, *UNI Signaling 4.0*, October 1995, ftp://ftp.atmforum.com/pub/.

[11] ATM Forum, *UNI 3.1 Signaling 3.1*, September 1994, ftp://ftp.atmforum.com/pub/.

[12] SAS Institute Inc., *SAS Language and Procedures: Usage 2*, *Version 6*, *First Edition*, SAS Institute Inc., Cary, N.C., 1991.

[13] J.S. Milton and J.C. Arnold, *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*, McGraw-Hill, New York, N.Y., 1995.

[14] H. Newton, *Newton's Telecom Dictionary*, Miller Freeman, New York, N.Y., 1999.

[15] T. Williams, C, Kelley, *et al.*, *GNUPLOT – An Interactive Plotting Program*, http://www.cs.dartmouth.edu/gnuplot/gnuplot.html.

[16] S. Moriguchi, M. Munctomo, and Y. Sato, "A Moving Average Method for Predicting Process Resource Usage Based on a State Transition Model," *Conference of the North American Fuzzy Information Processing Society*, NAFIPS, 1998, pp. 82-85.

[17] S.K. Jha and B.R. Howarth, "Capacity Planning of LAN using Network Management," *19th Conference on Local Computer Networks*, 1994, pp. 425-430.

[18] K. Terlpan, *Communication Networks Management*, Prentice Hall, Englewood Cliffs, N.J., 1992.

[19] G. Gopal, C. Kim, and A. Weinrib, "Dynamic Network Configuration Management," *IEEE International Conference on Communications*, ICC, 1990, pp. 95-301.

[20] K.H. Lee, "Design And Implementation Of A Configuration Management System," *IEEE Technical Program Conference Record, Global Telecommunications Conference*, 1993, pp. 1563-1567.

[21] C.N. Witzel, "Service-Level Agreements: A Management Tool for Technical Staff," *Journal of Capacity Management*, Vol. 1, No. 4, June 1983, pp. 344-362.

[22] A. Adas, "Traffic Models in Broadband Networks," *IEEE Communications Magazine*, Vol. 35, No. 7, July 1997, pp. 82-89.

[23] J. Crowder, *et al*., *NET.WORK.VIRGINIA*, http://www.networkvirginia.net.

[24] B.S. Jang, B.G. Kim, and G. Pecelli "A Prediction Algorithm for Feedback Control Models with Long Delays," *2nd IEEE International Workshop on Broadband Switching Systems Proceedings*, IEEE BSS, 1997, pp. 131-136.

Appendix A. Statistical Results

This appendix presents results from the statistical distribution tests and utilization analysis. In addition, plots of traffic intensity are included.

A.1 Distribution Analysis Results

This section presents the output of the distribution analysis performed on the holding times, interarrival times, and circuit utilizations. The following tables represent the results of the normality and exponential tests. The tables created from the normality tests include some descriptive statistics of the analyzed data as well as the p-values. The tables from the exponential test include the separation of expected and observed results and the Chi-Squared values.

A.1.1 Holding Times

Table A-1. Results of Normality Test for Holding Times

N	327	Sum Wgts	327
Mean	7536.211	Sum	2464341
Std Dev	19523.62	Variance	3.81E+08
Skewness	15.17057	Kurtosis	254.9387
USS	1.43E+11	CSS	1.24E+11
CV	259.0642	Std Mean	1079.659
T:Mean=0	6.980178	Pr>|T|	0.0001
Num ^= 0	327	Num > 0	327
M(Sign)	163.5	Pr>=|M|	0.0001
Sgn Rank	26814	Pr>=|S|	0.0001
W:Normal	0.235012	Pr<W	0.0001

Table A-2. Results of Exponential Test for Holding Times

Cell#	Cell Range	Observed	Expected
1	250.7	33	10.6991
2	617.8	33	15.0382
3	1981.4	33	49.8633
4	4487.5	33	71.1211
5	5920.9	33	31.2256
6	7233.1	33	23.8191
7	10163	33	40.3395
8	11602.5	33	14.7611
9	11823.9	33	2.03041
10	11823.9	30	68.1026

The value of Chi-Squared with 8 degrees of freedom is 615.278.

A.1.2 Interarrival Times

Table A-3. Results of Normality Test for Interarrival Times

N	166	Sum Wgts	166
Mean	10838.46	Sum	1799185
Std Dev	39659.31	Variance	1.57E+09
Skewness	6.363906	Kurtosis	43.62655
USS	2.79E+11	CSS	2.60E+11
CV	365.9126	Std Mean	3078.159
T:Mean=0	3.521086	Pr>|T|	0.0006
Num ^= 0	166	Num > 0	166
M(Sign)	83	Pr>=|M|	0.0001
Sgn Rank	6930.5	Pr>=|S|	0.0001
W:Normal	0.288037	Pr<W	0.0001

Table A-4. Results of Exponential Test for Interarrival Times

Cell#	Cell Range	Observed	Expected
1	117	17	1.78231
2	355	17	3.56672
3	675	17	4.6738
4	953	17	3.94985
5	1468	17	7.05479
6	2001	17	6.95681
7	3259	17	15.1245
8	6476	17	31.5606
9	45021	17	88.7237
10	45021	13	2.60692

The value of Chi-Squared with 8 degrees of freedom is 391.034.

A.1.3 Circuit Utilizations

Table A-5. Results of Normality Test for Circuit Utilizations

N	327	Sum Wgts	327
Mean	2827.779	Sum	924683.9
Std Dev	564.1641	Variance	318281.2
Skewness	-0.92907	Kurtosis	-0.14405
USS	2.72E+09	CSS	1.04E+08
CV	19.95078	Std Mean	31.19835
T:Mean=0	90.63876	Pr>|T|	0.0001
Num ^= 0	327	Num > 0	327
M(Sign)	163.5	Pr>=|M|	0.0001
Sgn Rank	26814	Pr>=|S|	0.0001
W:Normal	0.660616	Pr<W	0.0001

Table A-6. Results of Exponential Test for Circuit Utilizations

Cell#	Cell Range	Observed	Expected
1	2249.93	33	179.429
2	2249.97	33	0.00220172
3	2249.98	33	0.000664788
4	3245.95	33	43.8081
5	3249.78	33	0.140449
6	3249.96	33	0.0063817
7	3249.99	33	0.00113495
8	3250	33	0.000585958
9	3250.03	33	0.000984381
10	3250.03	30	103.61

The value of Chi-Squared with 8 degrees of freedom is 6.23513e+006.

A.2 Traffic Intensity Plots

This section presents the traffic intensity plots of in-progress call utilization as a function of time. The three traffic intensity plots shown below represent the three traffic categories described in Chapter 4.

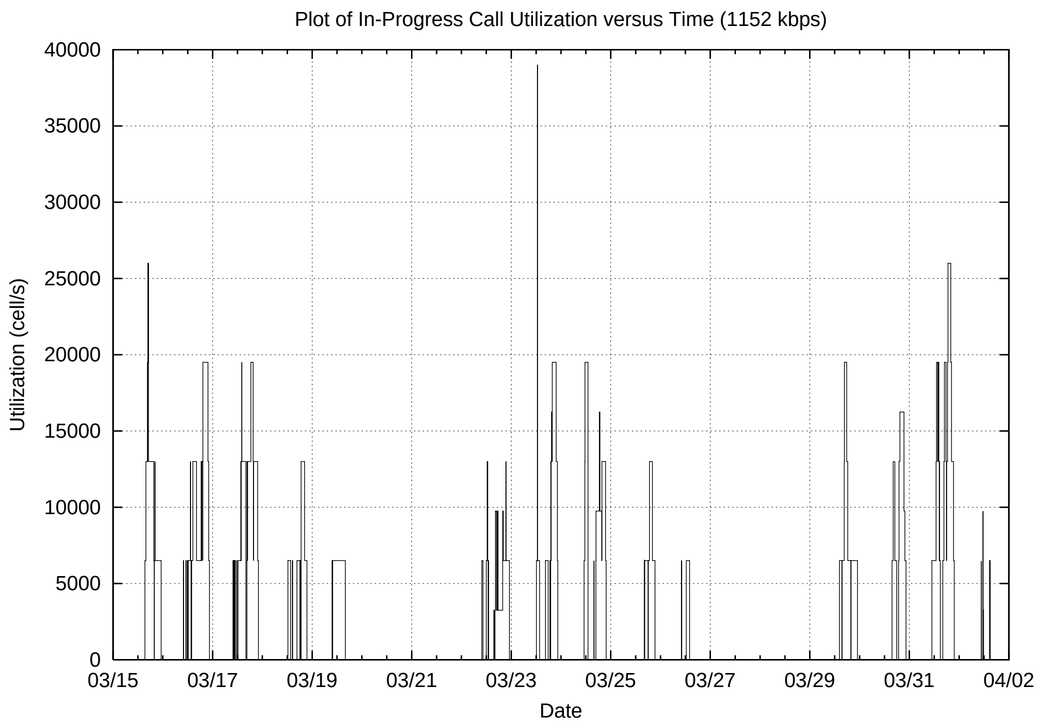

Figure A-1. Traffic intensity plot of 1152 kbps traffic category.

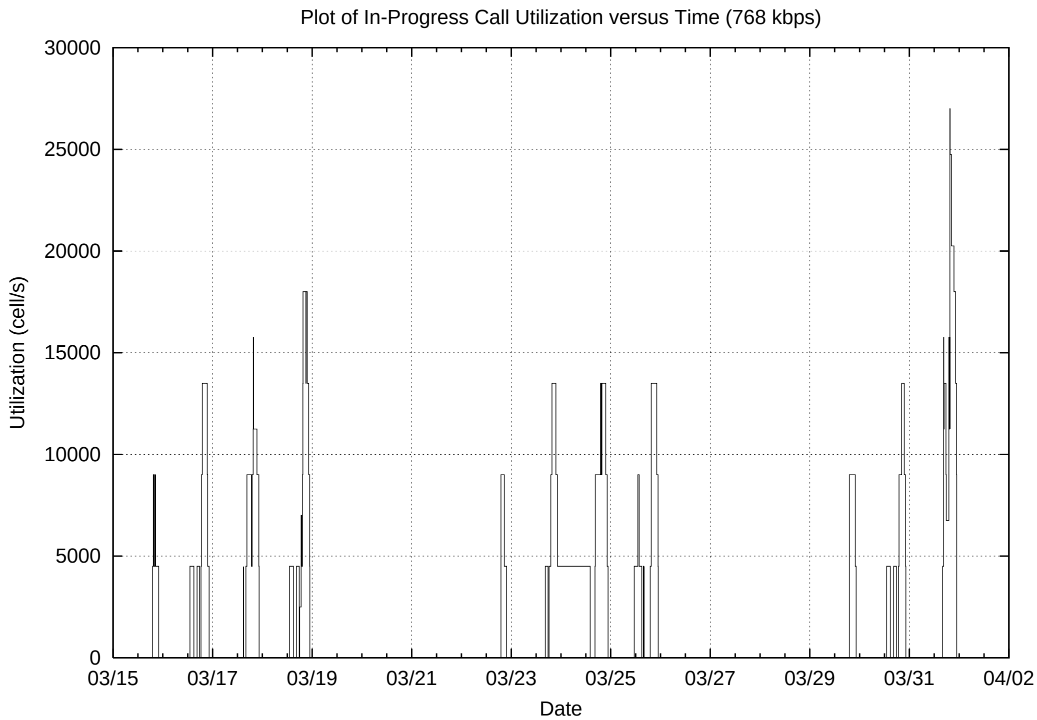

Figure A-2. Traffic intensity plot of 768 kbps traffic category.

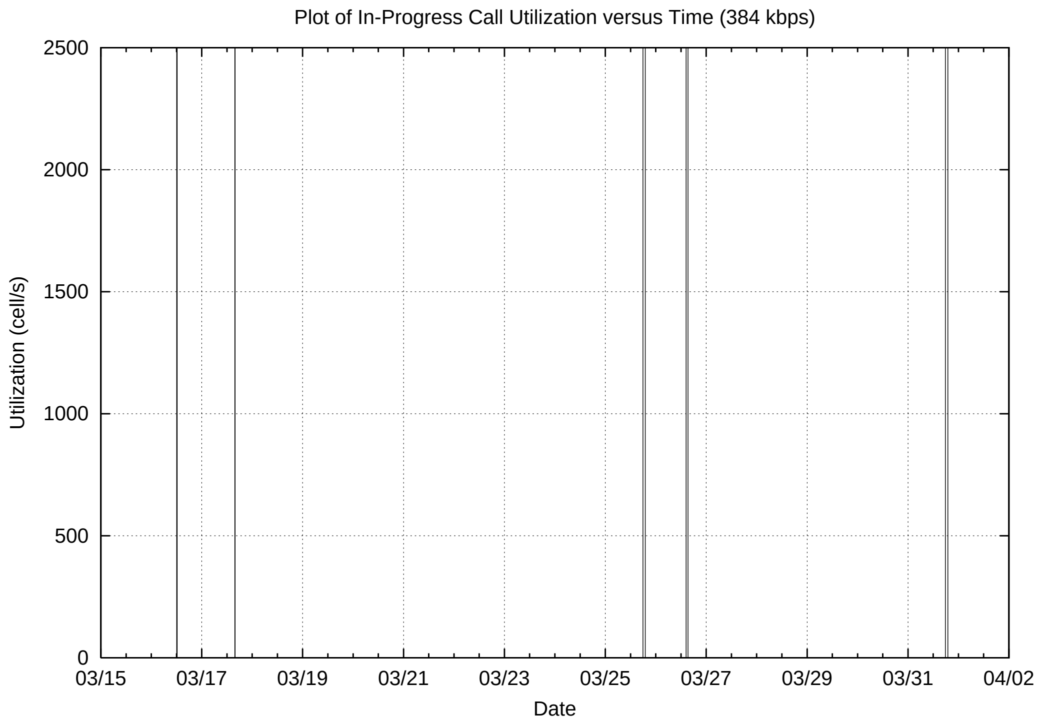

Figure A-3. Traffic intensity plot of 384 kbps traffic category.

A.3 Utilization Analysis Results

This section presents the output of the aggregate data analysis performed on the circuit utilizations. The following tables represent the results of the prediction algorithm and subsequent calculations following the predicted utilization bounds. Each table corresponds to a different time step used in the prediction calculations. The procedure for these calculations is discussed in Chapter 4.

Table A-7. Aggregate Data Analysis Results Using 10-Minute Time Step

Analysis Time	**Bound (cells/s)**	**Percent Time Exceeded (Activity)**	$\sum \frac{\lvert x-\hat{x}\rvert}{N}$	**Over Utilization Rate (cells/s)**	**Excess Capacity Rate (cells/s)**	**Percent Time Exceeded (Total)**
Median	9199	50%	7055	–	–	–
1 day	31556	5.39%	19580	230.9076	27691	1.72%
1 week	39764	0.64%	27444	64.0552	35849	0.20%
2 weeks	39764	0.64%	27444	64.0552	35849	0.20%

Table A-8. Aggregate Data Analysis Results Using 60-Minute Time Step

Analysis Time	Bound (cells/s)	Percent Time Exceeded (Activity)	$\sum \frac{\lvert x-\hat{x}\rvert}{N}$	Over Utilization Rate (cells/s)	Excess Capacity Rate (cells/s)	Percent Time Exceeded (Total)
Median	6599	50%	6803	–	–	–
1 day	15179	26.67%	8779	2174.987	12069	9.83%
1 week	35171	1.33%	24477	94.3467	31344	0.49%
2 weeks	53830	0%	42823	0	50014	0%

Vita

Creighton Tsuan-Ren Hager

Creighton Hager was an assistant in the Fiber and Electro-Optics Research Center and an Engineering Tutor for the College of Engineering while obtaining his Bachelor of Science in Electrical Engineering at Virginia Tech. While pursuing his Master of Science degree in Electrical Engineering, also at Virginia Tech, he worked as a Graduate Research Assistant analyzing ATM call detail records. He has held summer jobs as an Engineering Assistant, testing Common Operating Environment software, and as a Program Analyst developing macros for mathematical models for the U.S. Navy Submarine Program. His professional memberships include IEEE and Eta Kappa Nu.