Proceedings of the 8th IEEE Mediterranean Conference
on Control and Automation (MED 2000)
Rio, Patras, GREECE 17-19 July, 2000

ACTIVE PERFORMANCE MONITORING FOR MULTIMEDIA ATM NETWORKS

ANTHONY BURRELL +, TITSA PAPANTONI ++

+ Oklahoma State University, Computer Science Department, 219 Math Science, Stillwater, Oklahoma 74078, tburrell@cs.okstate.edu
++ University of Alabama, Electrical and Computer Engineering Department, Box 870286,Tuscaloosa, Alabama 35487-0286, tpapanto@coe.eng.ua.edu

Abstract. We consider multimedia ATM networks with time-varying traffics and topologies. To deal effectively with the time-varying environment, the deployment of traffic and network performance monitoring techniques is necessary for the identification of traffic changes, network failures, and also for the facilitation of protocol adaptations and topological modifications. The objective of the paper is the design, analysis and evaluation of mobile intelligent agents that implement effective performance monitoring techniques, while capturing the dynamics inherent in the multimedia environments. Towards this objective, a core sequential algorithm which depicts the functionality and operations of the network performance monitoring techniques is adopted. Specific forms of the algorithm are used for the identification of networks failures. For a given network topology, the location of the minimum necessary set of agents for complete network and traffic "visibility" is be specified via identifiability methods.

Key Words. Mulitmedia ATM Networks, Algorithmic Monitoring, Failure Recognition, Identifiability.

1. INTRODUCTION

Multimedia ATM networks are of high interest in this information technology era, where efficient bandwidth utilization is a key issue due to its limited availability. These networks aim at the satisfaction of the Quality of Service (QOS) of multimedia heterogeneous traffics (e.g., image, audio, data, graphics, text), with simultaneous high utilization of network resources and bandwidths. The performance demands imposed on networks carrying multimedia connections are challenging and can be satisfied via the deployment of highly dynamic statistical multiplexing protocols which honor the traffic QOS, while utilizing the network resources efficiently. While the successful design of high-performance traffic multiplexing protocols requires accurate modeling of traffic statistical characteristics and QOS, variations in the latter characteristics, topological network changes (some due to mobility), and /or failures in the response of network components affect directly, and frequently dramatically, their performance characteristics. It is thus crucial that traffic and network performance monitoring techniques be deployed, first for the identification of traffic changes and network failures, and then for the subsequent adaptation of the protocol operations, the pertinent recovery of failures, and the appropriate reconfiguration of network topologies.The main theme of this paper is the active performance monitoring of multimedia ATM, via the deployment of intelligent agents. These agents are active computational threads which employ statistically sound monitoring algorithms and are placed at key network locations, to acurately and efficienlty track the dynamics of the system. These locations change as the

network topology does; thus, the agents are mobile, where mobility is a function of time-varying network topologies.

2. FUNDAMENTAL TRAFFIC AND NETWORK PERFORMANCE CHARACTERIZATIONS AND MODELINGS — A CORE MONITORING ALGORITHM

In multimedia network environments, various traffic classes are present (i.e., voice, image, high-speed data, etc.) and the statistical characteristics per class are time-varying. It is desired that traffics and network performance metrics be continuously monitored to identify, timely and accurately, changes in their statistical characteristics, for subsequent adaptation of network operations and protocols, for identification of network failures, and for possible topological reconfigurations as well. The development of statistically sound monitoring techniques requires that first the traffics and network performance metrics be effectively modeled, so that the models capture their statistical variations. We model them as sets of distinct and well-known stochastic processes (e.g., Poisson with known rates, geometric with known parameters, etc.). For example, for the k-th traffic class we consider the general scenario where the various possible arrival processes that characterize the traffic may be represented by the set $\{\mu_i^{(k)}\}_{1\le i\le m}$ where $\mu_i^{(k)}$ depicts a well-known process, such as a Poisson process with known fixed rate, and where all processes in the set are distinct (e.g., all Poisson with different and distinct rates); it is assumed that the set $\{\mu_i^{(k)}\}_{1\le i\le m}$ has been obtained via thorough traffic characterization/learning methods (see ref. [5]). The objective of monitoring the k-th traffic class is then to identify, timely and accurately, shifts from some process $\mu_j^{(k)}$ in the set to any other process in the same set. Below, we present a core monitoring algorithm that operates on models represented by sets of distinct stochastic processes (see ref. [1, 3]).

2.1 Core Monitoring Algorithm

We first represent an initial restricted form of the algorithm and then a reinitialization extension. Both the initial restricted form and the reinitialization extension operate on observed data sequences $x_1, x_2, \ldots, x_n, \ldots$ that are generated by the acting processes (e.g., in the case of processes representing traffics, the x_i's may be numbers of arrivals within consecutive fixed-length time intervals). Below, we use the abbreviated form x_1^n, for the sequence $x_1, \ldots, x_n$.

Initial Restricted Algorithm

The restricted algorithm addresses the following problem. Let the process which initially generates the data sequence be the process μ_0. Let it be possible that a shift to any one of m-1 independent processes μ_i ; i=1, ..., m-1 may occur at any point in time, where if a $\mu_0 \to \mu_i$ shift occurs, then the process μ_i remains active thereafter. The objective is to detect the occurrence of a $\mu_0 \to \mu_i$ shift as accurately and as timely as possible, including the detection of the process μ_i which μ_0 changed to. Let us denote by f_i ; i=0, 1, ..., m-1 density or probability functions induced by the processes μ_i ; i=0, 1, ..., m-1. Then, for the present problem, we propose the following algorithm.

Algorithm

(a) Select a threshold $\delta_0 > 0$.

(b) Have m-1 parallel algorithms operating. The ith algorithm; i =1, ..., m-1 is monitoring a $\mu_0 \to \mu_i$ shift. $T_n^{0i}(x_1^n)$ denotes the operating value of the ith algorithm at time n, given the observation sequence. The operating value $T_n^{0i}(x_1^n)$ is updated as follows.

$$T_0^{0i} \equiv 0$$

$$T_n^{0i}(x_1^n) = \max\left(0,\ T_{n-1}^{0i}(x_1^{n-1}) + \log \frac{f_i(x_n|x_1^{n-1})}{f_0(x_n|x_1^{n-1})}\right)$$

TThe algorithmic system stops the first time n when either one of the m-1 parallel algorithms crosses the common threshold δ_0. If the ith algorithm is the one that first crosses the threshold, then it is declared that a $\mu_0 \to \mu_I$ shift has occurred.

The algorithm described above is clearly sequential and thus computationally simple and efficient. Its fundamental performance characteristics can be found in reference [3].

Reinitialization Extension

Let us consider a generalization of the above problem as follows. At any point in time, let the data be generated by one of m mutually independent and parametrically defined stochastic processes $\{\mu_i$; i = 0, 1, ..., m-1$\}$. At any point in time, the acting process may shift to either one of

the remaining processes, in an equally probable fashion. The objective is to detect such shifts as accurately and as timely as possible. The present problem is a reoccurrence generalization of the problem in the initial restricted algorithm. For its solution, we propose a reinitialization extension of the latter algorithm, described below.

Reinitializing Algorithm

With each process μ_i, we associate a positive threshold value δ_i. Let it be known that at time zero the process μ_0 is acting. Then, at time zero, the core algorithm is deployed, with operating common threshold δ_0. Let T_1 denote the time instant when the above algorithm stops, and let a $\mu_0 \rightarrow \mu_{i_1}$ shift be decided at T_1. Then, at T_1, the $\mu_0 \rightarrow \mu_{i_1}$ decision is accepted and the core algorithm is deployed again, with a common operating threshold δ_{i_1}, to monitor a shiftfrom the process μ_{i_1} to either one of the remaining processes. The common operating threshold δ_{i_1} is associated with the starting process μ_{i_1}. In general, let $\{T_l\}_{l \geq 0}$ denote the sequence of decision/reinitialization time instants induced by the algorithm, with $T_0 \triangleq 0$. Then, at T_l it is decided that the process μ_{i_l} starts acting, and the extended algorithm with a common operating threshold δ_{i_l} is immediately deployed, to monitor a change from μ_{i_l} to either one of the remaining processes. Within the time interval $[T_l, T_{l+1})$, it is decided that the process μ_{i_l} is continuously acting.

3. MONITORED METRICS – SOME GENERAL CONCEPTS

The metrics to be monitored in the ATM environment are determined by the objectives. The global objectives are traffic and network management.

3.1. Traffic Management — Monitoring Metrics

Traffic management refers to the development of effective signaling and transmission algorithms and protocols that, in conjunction with dynamic capacity allocation techniques, satisfy the various Quality of Service (QOS) characteristics of the time-varying multimedia heterogeneous traffics. The dynamic capacity allocation techniques are assisted by Traffic Monitoring Algorithms (TMA), such as those in Section II, which track effectively changes in the statistical characteristics of the traffics (such as rates). The various characteristics of both the external and the intranetwork traffics are modeled as sets of distinct stochastic processes and the monitored metrics are the corresponding traffic arrivals (see ref. [2, 4]).

3.2. Network Management — Functions and Monitored Metrics

Network management is a crucial issue, which has not been comprehensively addressed. It should be clear, however, that the key components of effective network management are performance monitoring, identification of network failures, and topological network reconfigurations assisted by traffic monitoring techniques. We proceed with the concretion of these concepts.

Performance Monitoring

Performance monitoring is the indispensable component in network management. The term is meaningful only if, at first, the important network performance metrics are identified, and then, statistically reliable algorithms are deployed for the continuous monitoring of these metrics. The key network performance metrics are: delays, traffic rejection rates, wasted network resource rates, and satisfaction of the various other (e.g., jittering, bit error rates, etc.) traffic QOS. Under normal network operational conditions (low error channels, fast ATM switches, etc.), the "other" traffic QOS are observed if comprehensively designed dynamic signaling and transmission multiplexing protocols are deployed, assisted by a TMA. In the latter case, the remaining key network performance metrics are delays (D), rejection rates per traffic message (MRR), and wasted capacity rates (WCR). Given normal network operational conditions, given comprehensively designed protocols for signaling and transmission, the D, MRR, and WCR metrics take predictable (via analysis and numerical evaluations) values for given statistical descriptions of the various network traffics (see [2, 8]). Thus, the reasons for the monitoring of the D, MRR, and WCR metrics are either to identify changes in traffic characterizations for normal operational network conditions, or, in conjunction with traffic monitoring, to identify failures in network components and functions.

Given the statistical descriptions of the various network traffics, given comprehensively designed

and fully analyzed and evaluated protocols for signaling and transmission, complete statistical characterizations of the D, MRR, and WCR network performance metrics are feasible (see [2, 8]). Thus, the monitoring of these metrics can be implemented by sequential algorithms as those in Section II. Consider, for example, one of the three metrics (D, MRR, or WCR) and complete statistical descriptions of the network traffics (provided by a priori traffic analyses in conjunction with decisions induced by the traffic monitoring protocol). Given the statistical descriptions of the network traffics, the distribution of the metric under normal operational conditions of the network is known (obtained via a priori performance analysis of the deployed signaling and transmission protocols). In addition, a set of "abnormal" distributions of the metric may be characterized then, each representing a distinct "abnormal" network state.

A performance monitoring system for the metrics may then include sequential algorithms as those in Section 2 to detect shifts from the distribution of the metric under normal operational conditions of the network to any of the "abnormal" distributions (each being associated with a specific "abnormal" network state).

Identification of Failures

Consider the simultaneous deployment of specific and comprehensively designed signaling and transmission protocols; together with their full evaluation and subsequent performance figures and tables, of traffic monitoring protocols as those in Section II, and of the-similar to the latter-monitoring protocols of the D, MRR, and the WCR network performance metrics. The decisions performed by the traffic monitoring system dictate the distributions of the D, MRR, and WCR metrics both under normal and "abnormal" operational network conditions. These distributions mat then be used by the — as that in Section II — sequential algorithmic system to detect shifts from normal to "abnormal" or failing states of network components. For example, if the traffic monitoring and the performance monitoring operations are performed at the origin versus the destination ends of a unidirectional network channel, failures at the channel/fiber level may be so detected.

3.3. Distributed Intelligent Agents and Algorithmic Performance Characteristics

The traffic and network performance monitoring operations are performed by intelligent agents that are generally located at key nodes of the backbone network and execute distributed functions. Each agent monitors a subset of network elements (nodes, channels, etc.). The performance monitoring characteristics of the monitoring algorithms , as executed by the distributed agents, affect the overall system performance and the induced tradeoffs.

4. ACTIVE MONITORING OF TRAFFIC REJECTION RATES FOR FAILURE RECOGNITION

The performance of the network, as perceived by the users, is measured by the "success" of message transmission attempts, where "success" means satisfaction of all the imposed QOS as well as good delay characteristics, and where "failure" is synonymous to rejection. Thus, from the point of view of the network users, the identification of a transmission attempt as successful or non-successful is based on the overall origin-to-destination performance, where origins and destinations are local base stations or nodes. We note that local base stations are where traffics are both generated and disseminated.

4.1. A Single Channel Approach

Consider a single network channel and a single traffic class. Let all multiplexing hybrid protocols be designed and well studied (as for example in [2, 4]), and let all their dynamics and performance metrics be well known. Let then p denote the message rejection rate attained by the system for the traffic class at hand, under "normal" channel conditions . We note that the multiplexing protocols deployed by the system are designed to satisfy the QOS of the traffic (see ref. [2, 4]); thus, the rejection rate p lies within the tolerance region, as dictated by the traffic class. Let q (where $q > p$) be the value of the traffic rejection rate designated as "alarming" regarding a shift of the channel conditions from "normal" to "abnormal". Then, it is desirable that a possible $p \rightarrow q$ shift in the traffic rejection rate be continuously monitored.

To deploy a high performance $p \rightarrow q$ shift detection algorithm, models of the stochastic processes that generate message rejections are needed. A worst case model which also reflects reality under some network conditions is the Bernoulli model. The Bernoulli model is a worst case model because it leads to worst performance for a whole class of algorithms, including the core algorithm in Section II.

Drawing from the above discussion, let the messages generated by the traffic class be indexed by the natural numbers {i; $1 \le i$ }. Let then x_i be equal to 1 ; if the i^{th} message is rejected in its transmission (failing some of its QOS) through the channel, and let x_i be equal to 0; otherwise. Let the sequence$\{x_i\}$ be a Bernoulli process with Pr ($x_i = 1$) = p; under "normal" channel conditions, and with Pr ($x_i = 1$) = q; under "abnormal" channel conditions. We deploy the core algorithm in Section II, to monitor a possible shift from the Bernoulli parameter p to the Bernoulli parameter q. Some straight forward algebraic manipulations lead to the following simplified form of the algorithm, where V (n) denotes the operational value of the algorithm at time n.

Algorithm

The algorithm deploys a threshold $\delta > 0$. It decides $p \rightarrow q$ shift has occurred the first time n such that $V(n) \ge \delta$ .The operational values of the algorithm are updated as follows:

$V(0) \equiv 0$

$V(n+1) = \max (0 , V(n) + [x_{n+1} - \zeta (p , q)])$

where,

$$p < \zeta (p , q) \equiv \left[\log \frac{(1-p)q}{(1-q)p} \right]^{-1} \log \frac{q}{p} < q$$

The threshold δ is selected as a tradeoff between the probability of correct decision and the probability of false alarm. As the Kullback information number, $I(p,q) \equiv p \log (p/q) + (1-p) \log ((1-p) / (1-q))$, increases, high probability of correct decisions and low probability of false alarms can be simultaneously attained.

4.2. Minimal Set of Monitoring Agents for Given Network Topology

Data measurements taken solely at the local base stations suffice to identify and monitor overall network traffic rates and network performance metrics. The questions to be addressed at this point are: Can end-to-end (local base station-to-local base station) data alone be used to effectively locate faulty network lines *anywhere in the network?* If yes, how; and if not, what kind of additional information is necessary and from where in the network should it be collected? The answers to these questions will provide the cardinality of the minimal set of intelligent agents and their localities. To explore the above questions, a mathematical maximum likelihood statistical approach may be applied to a fixed network topology model*: a network with fixed nodes and connectivity, with point-to-point loads, and with fixed routing probabilistic structure*. This approach allows sufficient conditions for the identifiability (estimation) of a fixed set of per link failure probabilities to be developed. These conditions depend only on the routing probabilities for the point-to-point traffic, and indicate that, while link failure probabilities for wireless networks cannot in general be completely determined from end-to-end observations, ambiguities can be resolved by the addition of observations on a minimal number of selected links from tandems (central processing nodes). The maximum likelihood approach also indicates the appropriate extension of the core algorithm in Section 2, when applied to detect shifts in per link call (communication attempt) failure probabilities for a large-scale wireless network. The extension uses the basic simplified assumption that the network routing structure is known and remains unchanged. This assumption may reflect average load network conditions (no large load fluctuations present), and may be closer to realistic when limited network portions are considered.

A Sample of the Maximum Likelihood Approach

Here, we give a sample of the maximum likelihood approach. Consider a given portion of the whole network, involving a number of local base stations that are connected with each other through a number of links and tandems. We use the term *call* from now on for communication request.

Denote:

(kl): ordered end-to end communication, i.e., call originating at end k and addressing end l.

i; $l \le i \le M$: link index, where M is the overall number of links in the network portion considered.

r_{kl}: the relative load associated with pair (kl) over the network portion considered. Alternatively, probability that a random communication attempt made somewhere in the network portion is a (kl) attempt. Then, $\Sigma_{kl} r_{kl} = 1$, where the summation is over all communicating network end pairs.

$q_{i,\,(kl)}$: probability that a (kl) call uses the link indexed by i. This is a routing probability.

v_i: failure rate of the link indexed by i. This is the probability that a call going through the link indexed by i fails.

p_i: the probability that a call attempt made somewhere in the network portion considered fails due to the link indexed by i.

p_o: the probability that one random call generated somewhere in the network does not fail.

$f_{(kl)}(x)$: the probability that a random call attempt is generated on the network, it is a (kl) attempt, and an outcome x is observed. When this outcome is a failure or success, and these two concepts are disjointly defined, then

$$x = \begin{cases} 1, & \text{if attempt fails} \\ 0, & \text{if attempt succeeds.} \end{cases}$$

We will assume that both the relative loads and the routing probabilities remain unchanged and they are well-known.

Under these assumptions, each end-to-end pair (kl) in the network portion considered generates a constraint fraction r_{kl} representing the relative call load generated from (kl) the way an outside observer sees it. The outside observer sees, in addition, the outcome (communication success or failure) of every communication attempt in the network, while the individual users see only the outcomes of their own attempts.

The outside observer evaluates the overall performance of the network portion considered through the appropriate processing of the observed dispositions of end-to-end call attempts; the observations may be made either by the assumed outside observer, or they may be reported by the individual users.

Let us now make the following important assumption: A call failure is caused by just one link and the call continues being routed (or flowing) after a failure. This assumption excludes the possibility of a call being actually stopped at the link where the failure occurs.

Under our assumption, the contribution of link i to communication failures in the network is represented by the number p_i, i.e., p_i represents the probability that link i fails a random call on the network. Obviously, we have then:

$$\sum_{i=0}^{M} p_i = 1$$

We will show now that the probabilities p_i can also represent the point-to-point performance of the network, i.e., performance as seen by the network users.

Using binary classification of the successful and unsuccessful call attempts, we obviously have:

$$f_{(kl)}(1) = \sum_{i=1}^{M} p_i \bullet P_r\{(kl) \text{ attempt / random attempt, fails at link } i\}$$

$$= \sum_{i=1}^{M} p_i \frac{r_{kl} \bullet q_{i,(kl)}}{\sum_{m\rho} r_{m\rho} q_{i,(m\rho)}}$$

where in $\Sigma_{m\rho} r_{m\rho} q_{i,(m\rho)}$, the summation is overall the communicating (mr) pairs in the network portion considered.

Generalizing we easily see that:

$$f_{(kl)}(x) = r_{kl} \left[\sum_{i=1}^{M} p_i \frac{q_{i,(kl)}}{\sum_{m\rho} r_{m\rho} q_{i,(m\rho)}} \right]^x \bullet$$

$$\bullet \left[1 - \sum_{i=1}^{M} p_i \frac{q_{i,(kl)}}{\sum_{m\rho} r_{m\rho} q_{i,(m\rho)}} \right]^{1-x}$$

From the above expression, one observes that the influence of link i to the point-to-point performance between nodes k and l is represented by the expression:

$$v_i q_{i,(kl)} = p_i \frac{q_{i,(kl)}}{\sum_{m\rho} r_{m\rho} q_{i,(m\rho)}}$$

Summing over all the links gives the total point-to-point failure rate for pair kl. Also, it is clear that there is a direct relationship between the p_i and the corresponding link failure rate v_i.

Let us now consider that the outside observer has collected a fixed number of observations for the network portion considered, and he is given the pair index (kl) for each of them and the outcome of each attempt (success or failure). Then, an ML estimation algorithm to estimate the p_i's assuming the r_{kl}'s, $q_{i,(kl)}$'s known, requires the maximum likelihood function:

$$f(\bar{p}) = \sum_{kl} \sum_{j=1}^{N_{kl}} \left[x_{j,kl} \log \left(\sum_{i=1}^{M} p_i \frac{q_{i,(kl)}}{\sum_{m\rho} r_{m\rho} q_{i,(m\rho)}} \right) + \right.$$

$$\left. + (1 - x_{j,kl}) \log \left(1 - \sum_{i=1}^{M} p_i \frac{q_{i,(kl)}}{\sum_{m\rho} r_{m\rho} q_{i,(m\rho)}} \right) \right] + \sum_{kl} \sum_{j=1}^{N_{kl}} \log r_{kl}$$

The ML optimal $\overline{p}$ value is found when the gradient of $f(\overline{p})$ is set equal to zero, where the identifiability of $\overline{p}$ is determined via the second gradient matrix of $f(\overline{p})$. In particular, $\overline{p}$ is ML identifiable if and only if the latter matrix is strictly negative definite. Alternatively, the maximum set of identifiable links is determined by the maximum set of linearly independent routing vectors $\left[q_{i,((kl)_1)},\ldots,q_{i,((kl)_m)},\ldots\right]$. The latter set is generally *not unique* and determines the links that are "visible" via the end-to-end measurements, where the remaining links should be monitored via direct (end-to-end per link) observations.

5. CONCLUSIONS

We presented an approach towards the design of mobile intelligent agents for network performance monitoring of multimedia ATM networks. Our approach is based on a core algorithm that monitors key network performance metrics to subsequently perform diagnoses on the network service conditions and to dictate appropriate actions. We also presented an mathematical approach for the selection of an "optimal" minimal set of agents and their localities.

6. REFERENCES

[1] P.K. Bansal and P. Papantoni-Kazakos, "An Algorithm for Detecting a Change in a Stochastic Process," *IEEE Transactions on Information Theory*, Vol. IT-32, pp. 227-235, March 1986.

[2] A. Burrell, D. Makrakis, and P. Papantoni-Kazakos, "Traffic Monitoring for Capacity Allocation of Multi-Media Traffic in ATM Broadband Networks," *Telecommunications Systems Journal*, 1998, Vol 9, pp 173-206.

[3] A. Burrell and P. Papantoni-Kazakos, "Extended Sequential Algorithms for Detecting Changes in Acting Stochastic Processes," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 28, No. 5, September 1998, pp. 703-710.

[4] A.T. Burrell and P. Papantoni-Kazakos, "An Integrated Approach to Signaling, Transmission, and Traffic Monitoring for Dynamic Capacity Allocation in Mobile ATM Networks," *Hawaii International Conference on Systems Sciences*, Wailea, Hawaii, Jan. 7-10, 1997.

[5] A. T. Burrell and P. Papantoni-Kazakos, "On-Line Learning and Dynamic Capacity Allocation in the Traffic Traffic Management of Integrated Services Networks", *European Transactions on Telecommunications*, Special Issue on Architectures, Protocols, and Quality of Service for the Internet of the Future, Vol 10, No.5, March/April 1999, pp. 202-214.

[6] D.P. Chandler, A.P. Hulburt, and M.J. McTiffin, "An ATM-CDMA Air Interface for Mobile Personal Communications," COST 231 TD (94) 91.

[7] I. Chlamtac and A. Farago, "Making Transmission Schedules Immune to Topology Changes in Multi-Hop Packet Radio Networks," *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp. 23-29, February 1994.

[8] D. Raychaudhuri, "ATM Based Transport Architecture for Multiservices Wireless Personal Communications Networks," *Proc. IEEE '94*,pp. 559-565, 1994.