Communication Research at Dalarna University

Ernst Nordström[a], Ming Fan[b]

[a]Department of Culture, Media and Computer Science,
[b]Department of Mathematics, Natural Sciences and Engineering,
Dalarna University,
SE-781 88 Borlänge, Sweden
eno@du.se, fmi@du.se

1 Introduction

This paper gives a summary of the research at Dalarna University on resource management in multi-service communication networks. The project deals with mathematical modeling and simulation of resource allocation at the call layer and the physical/virtual nework layer. The project comprises five sub projects:

- Call traffic models,
- Link GoS models,
- Network GoS models,
- CAC and routing,
- Network dimensioning.

The network is modeled as a circuit-switched network, representing ATM and STM networks, and IP networks, provided they are extended with resource reservation capabilities. Call requests are subject to on demand or delayed call set up. In the first case, call requests that find the network busy are lost, while in the second case they are put in call queues to wait for sufficiently free capacity to become available in the network.

Two ideas are central to the project:

- resource management should be based on realistic traffic models in order to maximize the revenue and maintain the Grade of Service (GoS),
- delayed call set up can provide revenue improvement and efficient GoS control.

2 Modeling of call traffic to OD pairs

The network is offered traffic from K call classes. Each call class is associated with one origin-destination (OD) node pair and one call category.

2.1 Connection arrival process

Since the days of Erlang the Poisson model has commonly been used to describe the random arrivals of call requests to the OD pairs of a telephone network. Although the Poisson model serves its purpose in telephone networks, it lacks descriptive power in the case of Internet where a substantial portion of traffic is WWW, NNTP and SMTP connections transported by TCP. The WWW, NNTP and SMTP services produce connection arrivals which are different in nature from the telephone service; For example, a person using the WWW service is more likely to initiate additional downloads after the first download. A person using the telephone service is more likely to initiate independent calls.

Measurements on real WWW, NNTP and SMTP connection arrivals in the Internet have revealed that the arrival process shows burstiness over many time scales, ranging from seconds to hours. Paxson and Floyd found that actual WAN traffic is consistent with statistical self-similarity for sufficiently large time scales [8]. These findings have been verified by Feldmann *et al.* [6].

Crovella has proposed an ON/OFF model for the downloading of web documents [1]. A single TCP connection is invoked in each ON period. The duration of the TCP connection follows a heavy-tailed distribution since the distribution of WWW document sizes on Internet is heavy-tailed. The OFF period corresponds to the user thinking time. Crovella argues that also the OFF period is heavy-tailed.

Deng also developed an ON/OFF model for the web service [2]. During the ON period, the user makes multiple web requests each resulting in a new TCP connection. The OFF period is the time between two ON periods while the user views the page. Deng proposed distributions for three parameters: the duration of the ON period was found to be Pareto distributed, the duration of the OFF period was found to be Weibull distributed as well as the inter-arrival time of web requests during the ON period.

Feldmann proposed to model the arrivals of WWW connections by a Weibull inter-arrival time distribution [5]. The Weibull distribution has finite moments, including finite mean and variance of the inter-arrival time. For this reason the Weibull distribution is considered to have a light tail. Since its variance is finite the Weibull distribution does not give rise to a self-similar arrival process.

2.2 Connection holding time distribution

The traditional model of call holding times B_k is the (negative) exponential distribution with rate parameter μ. The exponential distribution matches the actual holding times in case of telephony among other services. However, for the WWW and FTP service, the connection holding time is more heavy tailed [1, 2, 8].

3 Modeling of call traffic to individual links

3.1 Connection arrival process

In general, each link in the network is offered renewal call arrival processes from many, perhaps all, of the call classes. The arrival process offered to a given link is a result of splitting and merging of component arrival processes. First, the per-class j arrival process is *splitted* over many alternative paths. Second, at each link the per-path arrival processes are *merged* to form a superposed arrival process to the link.

The Palm-Khintchine theorem states that when the number of normalized renewal processes goes to infinity and the inter-arrival distribution is light-tailed, the limit process becomes Poisson. On the other hand, if the distribution is heavy-tailed, the limit process is fractional Brownian motion (FBM).

The research focuses on modeling the superposed arrival process to individual links. The rate of convergence to the limit processes (Poisson and FBM) will be investigated theoretically and by simulation.

3.2 Connection holding time distribution

In case of exponentially distributed call holding times we exploit the memoryless property of the exponential distribution. The mean link call holding time for a category is computed as the weighted sum of per-class mean holding times, with weights given by the probability that the call on the link is from the given class. The mean departure rate for the category is the reciprocal of the mean call holding time.

In case of general holding time distributions, e.g. Pareto, the modeling task is similar

to the modeling of the superposed call arrival processes. We expect techniques such as moment matching are useful for describing both the arrival and departure process.

4 Link GoS models

The queuing systems studied in our project can be classified into pure loss, pure delay and mixed loss-delay systems. In a loss system, customers are accepted in order of arrival (first come, first served) and customers are lost when no free server is available. In a delay system customers are served on a FCFS basis and when no free server is available, the customer is delayed in a finite or infinite waiting room. In a mixed loss-delay system we have two types of customers [9]. Calls of type I have access the link with no restriction, but will be blocked if all servers are busy. Calls of type II have restricted access to the service facility; the cutoff parameter r_0 specifies the maximum number of type II calls that can be in service at the same time. Therefore, a queue of type II calls forms as soon as a type II arrival finds r_0 type II calls already in service or otherwise if there are not enough free servers.

The current research focuses on the mixed loss-delay system with general renewal arrival processes and/or general holding time distributions. We intend to combine the embedded Markov chain method with the matrix analytic technique to establish the queueing model on the above mentioned system.

5 Network GoS models

The loss network operating under fixed or load sharing routing with Poisson call arrival processes to the OD pairs, and generally distributed call holding times, is the normal assumption in the work on network GoS evaluation [7]. The network GoS measures are obtained by solving a set of Erlang fixed-point equations, also called reduced load approximation [7].

Besides the call traffic model and the network capacity model, the routing algorithm strongly affects the network GoS model.

On the short term, the research focuses on GoS evaluation of loss networks with LLR and MDP routing. On the long term, the research focuses on GoS evaluation of networks with mixed loss-delayed call set up, and general call arrival processes and/or general call holding time distributions.

6 CAC and routing

Our work on Call Admission Control (CAC) and routing focuses on the Markov Decision Process (MDP) method. MDP theory was developed during the 60's and 70's by Bellman, Howard and others, and matured during the 80's and 90's.

MDP-based CAC and routing has two major advantages over conventional routing methods such as Least Loaded Routing (LLR) and sequential routing. First, the MDP method is able to take CAC and routing decisions which yield *near-optimal average revenue per time unit.* Second, the GoS distribution among the call categories and OD pairs can be continuously controlled by varying the reward parameter associated with each call class.

The MDP method have been applied to both loss networks [4] and mixed loss-delay networks [3].

The main critique towards the MDP method is its numerical complexity which can be prohibitive if the link is offered traffic from many call categories. A suboptimal solution with manageable computational complexity is achieved by applying MDP theory together with a three-level approximation. First, the network is decomposed into a set of links assumed to have independent Markov and reward processes, respectively. Second, the dimensions of the link Markov and reward pro-

cesses are reduced by aggregation of the call classes into call categories. Third, by applying an approximative link MDP model the link MDP tasks are simplified considerably.

The research focuses on the following tasks:

- Development of approximate link MDP models for Poisson arrival processes and exponential holding time distributions,
- Development of exact and approximate link MDP models for general arrival processes and/or general holding time distributions,
- Development of a MDP framework with periodic state information.

7 Network dimensioning

The objective used in dimensioning of link capacities in virtual and physical networks can be of several types. Two common examples are maximization of the average revenue rate and minimization of total network link cost. The GoS constraints for each call class are expressed in an absolute or relative manner. The relative GoS constraints can be expressed by preference functions from cooperative game theory.

The problem of finding vector of link capacities that optimizes the non-linear objective function under the given set of non-linear constraints can be solved using Sequential Quadratic Programming (SQP). The SQP method can be viewed as the natural extension of Newton and quasi-Newton methods to the constrained optimization setting.

On the short term, the research focuses on dimensioning of loss networks with LLR and MDP routing. On the long term, the research focuses on dimensioning for networks with mixed loss-delayed call set up, and general call arrival processes and/or general holding time distributions.

References

[1] M. Crovella, A. Bestavros, "Self-similarity in world wide web traffic – evidence and possible causes",*IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, pp. 835-846, 1997.

[2] S. Deng, "Empirical model of WWW document arrivals at access link", In *Proc. of International Conference on Communication*, ICC'96, 1996.

[3] Dziong Z., Liao K-Q., Mason L.G., "Flow control models for multi-service networks with delayed call set up", In *Proceedings of IEEE INFOCOM'90*, pp. 39-46, IEEE Computer Society Press, 1990.

[4] Z. Dziong, L. Mason, "Call admission and routing in multi-service loss networks", *IEEE Transactions on Communications*, Vol. 42, No. 2, 1994.

[5] A. Feldmann, "Characteristics of TCP connection arrivals", Tech. Rep., AT&T Labs Research, 1998.

[6] A. Feldmann, A. Gilbert, A. Willinger, T. Kurtz, "The changing nature of network traffic: scaling phenomena", *Computer Communication Review*, Vol. 28, No. 2, pp. 5-29, April 1998.

[7] F. Kelly, "Routing in circuit-switched networks: optimization, shadow-prices and decentralization", *Adv. Appl. Prob.*, No. 20, 1988.

[8] V. Paxson, S. Floyd, "Wide-area traffic: the failure of Poisson modeling", *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp. 226-244, June 1995.

[9] Y. Serres, L. Mason, "A multi-server queue, with narrow- and wide-band customers and WB restricted access", *IEEE Transactions on Communications*, Vol. 36, No. 6, June 1988.