Availability Planning of the Network Access Server

Dr. Olga I. Bogoiavlenskaia

Department of Computer Science, University of Petrozavodsk

Lenin St., 33, Petrozavodsk, Republic of Karelia, 185640, Russia

E-mail: olbgvl@mainpgu.karelia.ru

Abstract

The paper presents an approach for analyzing and planning the availability of a Network Access Server (NAS). The approach is based on the heterogeneous queuing model with loss [8, 9] developed with a respect to the key features of the NAS and its traffic. The model also allows evaluating the characteristics of the NAS concerning the details of the data transfer process. The last part of the manuscript treats the issues of the queueing model parameters estimation.

Contents

1 Introduction

The Network Access Servers (NAS) are one of the most important parts of the modern Internet structure. A NAS is the first network device to provide the end-user service and is a gateway for further services. Modern NAS support a big variety of services and are implemented using a high diversity of technical bases. The area is demonstrated by the high growth of the number and complexity of using NAS leading to the big intensity of the research related to software and hardware implementation (See for example [1, 2]).

The exhibited tendencies imply two motivations for performance related research in the area. The first one is that the NAS developers need to predict the performance of the systems they create. The second one is that providers using NAS in a commercial manner have to define parameters of their NAS instances to provide appropriate service and cost-efficiency levels. Providers also need strategies for NAS managing and capacity planning.

In this paper we propose an approach to evaluating NAS performance which is based on the queueing model. Mainly we concentrate on two metrics of NAS performance key interest. Those are availability and utilization. We understand the availability as a fraction of arrived remote calls which have got access to the NAS related with the total number of remote calls. Availability may be related to well-known performance metrics such as throughput and blocking probability. For the simplest cases blocking probability is a good measure of availability. Complex systems create the necessity to consider a set of values describing the behavior of different parts of the system (for instance the set of partial blocking probabilities). In the terms of throughput, availability is understood as relation of actual throughput to the maximal one.

Considering the NAS channel as a server and a data transmission session as a service one may model NAS as a classical queue $M|G|m|0$. For the case the distribution of busy channels is given by the Erlang formula

$$p_n = \frac{\rho^i}{i! \sum_{j=1}^{m} \frac{\rho^j}{j!}} , \tag{1.1}$$

where m is the number of NAS channels and ρ is the relation of the arrival

rate to the service rate. Although modern telecommunication systems in most cases demonstrate behavior which does not fit the Poisson flow assumption [4] that is acceptable in the NAS case. Here we treat the flow of users' single calls but not the flow of the data packets they generate and hence we may appeal to the flows central limit theorem [5]. We also refer to [6] which notes Poisson's property of single task flow and to our investigations of the Erlang model validity for NAS's simplest analysis [8].

This point looks like the end of the story. The probability p_m gives us the measure of availability and $1 - p_0$ defines utilization. Unfortunately these simple considerations are not enough for many practical cases due to the complex heterogeneous nature of most modern implementations. As the NAS provide different classes of services for the different classes of customers, its performance metrics must be evaluated correspondingly to the actual configurations. Due to the NAS heterogeneity general figures given by (1.1) may have a different meaning for different classes of customers or services. Therefore characteristics related to the real configuration are very important as they allow precise managing of the NAS including the managing of the structure of its heterogeneity.

The rest of the paper is organized as follows. Section two describes the possible sources of NAS heterogeneity, briefly presents the general model and gives some examples of its usage for an availability evaluation. Section three considers simple examples of capacity planning. Section four presents the extension of the general model which allows considering the details of the data transmission process and Section five considers the problem of the model's parameters estimation.

2 Availability and utilization for the heterogeneous NAS

We concentrate on the base sources which determine the NAS heterogeneity. The first one comes from the NAS itself, as it may consists of different parts with different service abilities. These parts provide different service rates inside one NAS instance. The second one is the workload which typically has a very complicated structure. We have to note that in most cases NAS workload characterization is a very complex problem and it is out of the scope of this work. At this point we suppose that workload is

divided into several classes which are characterized by unique arrival and service rates.

The last important point of the configuration is the rule of arrival, as the NAS may have explicit or hidden access limitation or preferences. Such rules also may serve as a tool for NAS management. Thus starting from those three points we have proposed a queueing system which models key features of the NAS heterogeneity structure.

The queuing system to be considered consists of several groups of servers. All servers in a group are of equal capacity. The number of servers in a group is finite.

Let us define the system state vector $N = (n_1, \ldots, n_s)$, where s is the total number of groups and n_i, $(i = 1, \ldots, s)$ is the number of busy servers in the group i. If, at a time t, the system is in the state N then the probability of a new customer arrival into the group i in the time $t + \Delta t$ is $\lambda_i(N)\Delta t + o(\Delta t)$. The probability of two or more arrivals during the period Δt is of the order $o(\Delta t)$. Let m_i be the number of servers in a group. An arriving customer chooses the server it will occupy randomly, there are no priorities associated with the servers.

If all servers in a group are busy a customer arriving in the group is lost. Let us denote a queuing system specified above with Σ.

The customer service time in the group i is a realization of the random variable η_i with finite expectation and arbitrary distribution function $G(x)$.

Unfortunately the analytical general solution for the model is either beyond reach or extremely cumbersome. But there exists a wide subclass of models which allow to obtain explicitly the distribution of the vector N in the product form. In this case the distribution possesses the invariance property (i.e. depends on $G(x)$ only through its mean). The subclass is defined by the restriction on functions $\lambda_i(N)$ and nevertheless includes most practically important cases. See for details [3, 8]. Let us now consider some simple examples which illustrate how to evaluate availability and utilization using the model described above.

Example. Two types of calls

Suppose that NAS serves two types of calls (customers) with two different traffic intensities (i.e. the ratio of arrival rate and service rate) ρ_1 and ρ_2. If m is the total number of NAS channels then the (1.1) gives us the general loss probability for $\rho = \rho_1 + \rho_2$. If we know the size of

the user population fraction which essentially contributes each type of workload then the important question of availability is: does the NAS customer share its resource in a proportion which may be treated as fair or desirable? Using the presented model one may calculate

- The probability that the system serves only the customers of the first class

$$p_1 = \sum_{n_1=1}^{m} p(n_1, 0) = G \sum_{i=1}^{m} \frac{\rho_1^i}{i!} , \tag{2.1}$$

  where G is the normalizing constant.

- The probability that the system serves only the customers of the second class

$$p_2 = \sum_{n_2=1}^{m} p(n_2, 0) = G \sum_{i=1}^{m} \frac{\rho_2^i}{i!} , \tag{2.2}$$

- The probability that NAS serves equal number of each class of customers

$$p(n_1 = n_2) = \sum_{i=1}^{m/2} \frac{\rho_1^i}{i!} \frac{\rho_2^i}{i!} . \tag{2.3}$$

Now if the essential contribution of the second type of calls flow is provided by 10% of the whole user population, $\rho_2 >> \rho_1$ and $p_2 = 0.6$ an administrator may conclude that the resource sharing is unfair or decide to wait while p_2 will reach 0.8.

Now let $\rho_1 \approx \rho_2$ but $\lambda_1 >> \lambda_2$ and $\mu_1 >> \mu_2$, i.e. there is an intensive flow of short calls and light flow of long calls. In the case $p^1 \approx p^2$, but since the arrival rate of short calls is essentially bigger than that of the long calls, the number of rejected short calls is bigger (in average) for the fixed time period. Thus in this case the NAS must provide different levels of availability for the different classes of the customers. The availability level diversification is also important if NAS serves some critical applications with high requirements for availability due to its nature or importance.

The general utilization of the described system is trivially calculated from the model as $U = 1 - p_0$. The model also allows to obtain partial

utilization related to different types of workload. The measure of system utilization by the first class of customers is

$$U_1 = 1 - p_2 - p_0 \tag{2.4}$$

and by the second class of customers is

$$U_2 = 1 - p_1 - p_0 \ . \tag{2.5}$$

3 Two examples of the availability control policies

3.1 Restrictions of the connection time

Let the NAS administration observe that several users establish very long sessions and hence create a big workload increasing the blocking probability. We consider the following question. If one decides to set a margin for the maximal remote connection time with the aim to reach a certain level of the NAS availability then the question is: what is the marginal value providing the given availability level? Here we consider a pure Erlangian case as it illustrates the approach and allows to avoid unnecessary complications. The following two steps solve the problem.

1. Let $\tilde{p}$ be the desired value of the blocking probability. The model yields the equation to find the service rate μ providing $\tilde{p}$ under the given arrival rate λ.

$$f(\mu) = \tilde{p} \ , \tag{3.1}$$

   where $f(\mu)$ is defined according to (1.1). Let $\tilde{\mu}$ denote the solution of (3.1).

2. Let $G(x)$ be the current service time distribution with the expectation $1/\mu_0$. Setting the connection time's margin t_0 means that we set $\mathbf{P}(t > t_0) = 0$ and $\mathbf{P}(t = t_0) = 1 - G(t_0)$. Since

$$\frac{1}{\mu_0} = \int_0^\infty x dG(x) \ , \tag{3.2}$$

the case of restricted connection time yields the equation

$$\frac{1}{\tilde{\mu}} = \int_0^{t_0} x dG(x) + t_0(1 - G(t_0)) \ , \tag{3.3}$$

The solution of (3.3) is the upper margin of the connection time.

The precise solution of (3.1) and (3.3) look intricate, but both functions are monotonously increasing and hence fit for numerical methods. Using equation (3.3) needs an assumption on the view of service times distribution.

3.2 Flow control

At this point we present a simple example to illustrate how to reconfigure the NAS structure and to evaluate the characteristics of the new configuration using the proposed queueing model.

Let us consider NAS with the m channels which serves two types of the customers creating the workload ρ_1 and ρ_2. Let the first type of the customers be of big importance for NAS users and need a higher level of availability. Thus, NAS is extended by the m_1 channels which serve the customers of the first class and reject the customers of the second class. The queuing model allows obtaining the following distribution

$$p(n_1, n_2) = G \frac{\rho_1^{n_1}}{n_1!} \frac{\rho_2^{n_2}}{n_2!} \ , \tag{3.4}$$

where n_1 and n_2 are the number of first and second types of customers in the system $n_1 + n_2 < m + m_1$ and $n_2 < m$. Here the blocking probability for the first class of customers is

$$p^1_{\text{loss}} \quad = \sum_{n_1+n_2=m+m_1} p(n_1 + n_2) \ . \tag{3.5}$$

Unfortunately in this case the distribution (3.4) does not let us calculate the exact value of the loss probability for the second class of the customers, but it gives the estimation:

$$p^2_{\text{loss}} \geq p^1_{\text{loss}} \ . \tag{3.6}$$

Another possible solution is to split NAS and let each type of the customers address a predefined set of channels. This solution allows to keep the desirable value of the blocking probability for each type of customer by choosing an appropriate number of the channels. The distribution (3.4) stays true for this case also. The arrival rule participates in (3.4) through normalizing constant G. Besides the blocking probabilities the distribution (3.4) also yields the probabilities of being idle (for the whole system or its part), the values of utilizations, means, variances, etc.

Note, that under these structures NAS may reject calls even if there are free channels.

4 The detailed analysis of the data transmission process

Let us consider the following extension of the queueing model Σ. Let us suppose that the customers have a compound structure. Each customer consists of a finite number of independent units. The number of units in a customer is denoted by the random variable ξ, with the distribution

$$\mathbf{P}\{\xi = k\} = \phi_k \,, \qquad k = 1, 2, \ldots \,. \tag{4.1}$$

Obviously,

$$\sum_{k=1}^{\infty} \phi_k = 1$$

and $\phi_0 = 0$. Let us assume that the expectation of ξ is finite. All units of a customer are served without delay one by one, in the order of arrival.

For the extended model we have obtained the joint distribution of the following event—the queue is in the state N and the jth busy channel in group i serves the r_i^jth unit of the customer. This distribution may be treated as a discrete analog of the supplementary variables method [7]. The joint distribution investigates the discrete interpretation of the data transfer process since it corresponds with the original objects and the terminology accepted in the area. Telecommunication systems typically process bytes, packets, files, application, etc. which obviously are measured using discrete units.

The compound customers approach demonstrates high flexibility. Since it allows a large variety of the customer unit interpretations, it promises to be fruitful for the large diversity of the performance problems. Besides the units transfered analysis the compound customers approach can evaluate the case of unreliable connections. In this case the transfer period and the period of recovery are treated as independent units of a single customer.

Let us consider the NAS with unreliable channel. We assume that during the connection time the line breaks and then recovers i times with the probability p_i. Let B be the expectation of the random variable i. The connection starts with the period of data transfer which is followed by the period of line recovering. Suppose that a session always ends by the data transfer period. Here one may treat the transfer and recovering periods as units of a customer. The probability ϕ_i of there being i units in the customer is

$$\phi_i = \begin{cases} p_{\lfloor \frac{i}{2} \rfloor}\,, & \text{if } i \text{ is odd} \\ 0\,, & \text{otherwise} \end{cases} \tag{4.2}$$

Let $F_i = \sum\limits_{j=i}^{\infty} \phi_j$. According to the definition of ϕ_i, $F_i = F_{i+1} = \sum\limits_{j=\lfloor \frac{i}{2} \rfloor}^{\infty} p_j$ for every even i. In this model the transmission periods correspond to the odd units and the silence periods correspond to the even one. Hence the extended Σ model derives the probability of the following event— k channels are broken among n busy channels. We again consider the Erlang queue for simplicity. The probability of k particular channels are silent (i.e. are recovering after a break) if $n > k$ of them are busy is

$$p^{n,k} = p_0 \left(\frac{\lambda}{\mu}\right)^n \left(\sum_{\text{odd } i_j} \prod_{j=1}^{k} F_{i_j} \right) \left(\sum_{\text{even } i_s} \prod_{j=1}^{n-k} F_{i_s} \right) . \tag{4.3}$$

Note that

$$\left(\sum_{\text{odd } i_j} \prod_{j=1}^{k} F_{i_j} \right) = B^k \tag{4.4}$$

and

$$\left(\sum_{\text{even } i_s} \prod_{j=1}^{n-k} F_{i_s} \right) = (B + \sum_{i=0}^{s} p_i)^{n-k} = (B+1)^{n-k} . \tag{4.5}$$

Hence the probability of any k channels being silent if n channels are busy is

$$p_n^k = p_0 \left(\frac{\lambda}{\mu}\right)^n \binom{n}{k} B^k (B+1)^{n-k} \tag{4.6}$$

and the probability of n channels being busy in spite of the type of their activity (transmission or recovery) is

$$p_n^k = p_0 \left(\frac{\lambda(B+B+1)}{\mu}\right)^n = p_0 \left(\frac{\lambda(2B+1)}{\mu}\right)^n . \tag{4.7}$$

Now let us suppose that in the previous model the session ends by the transmission period with probability P and by the silence period with probability $1-P$. In this case

$$\phi_i = \begin{cases} (1-P)p_{\lfloor \frac{i}{2} \rfloor} & \text{if } i \text{ is odd} \\ P p_{\lfloor \frac{i}{2} \rfloor} & \text{otherwise} \end{cases} \tag{4.8}$$

Here the summation by the odd indices yields the same result as the formula (4.4) does. For the even indices one may obtain

$$\sum_{\text{even } i_s} \prod_{j=1}^{n-k} F_{i_s} = (B+1-P(1-p_0))^{n-k} . \tag{4.9}$$

Now to calculate p_n^k and p_n one needs to replace $(B+1)$ by $(B+1-P(1-p_0))$ in (4.6) and (4.7).

5 The parameters estimation for the NAS models

The model Σ presented in Section 2 and the extended Σ (Section 4) aim at representing the real NAS in terms of their performance. The models clearly extract a set of variables which are meaningful for evaluation of NAS key performance characteristics. For each NAS, in spite of its structure and implementation, the model parameters form three groups: the set of the service rates, the set of the arrival rates and the arrival rule which sets the correspondence between arriving customers and the NAS channels.

The source of the information for obtaining parameters mentioned above is supposed to be the NAS registration system. Nowadays most of them keep information about connection start and stop time, the amount of data transfered, the protocols and the services used, etc. The information of the registration system combined with several general facts about the implementation under analysis are sufficient to extract the input parameters for the proposed model. But in some practical cases the specialty of the NAS activity creates the lack of input information. We briefly discuss most typical cases.

It is natural to accept the average number of the calls arriving in the time unit as an estimation for the arrival rate. Nevertheless a registration system typically does not make notes about rejected calls arriving in a blocked system. Hence at least at the peaks the arrival rate observed at NAS forms just a part of the whole arrival rate.

The estimation of the service rate typically is set as average number of the customers (calls) departing from the NAS in the time unit. In the NAS case this value integrates many technical aspects in one figure. Also it may depend on the number of busy channels. If so the average is useless.

The rule of arrival sometimes is given explicitly. It happens if the NAS administration directly regulates the arrival flow. But typically a user preferences also contribute to the rule of arrival. This contribution is latent and the analyst is supposed to have an assumption basing on extra information, common sense and personal experience.

In the cases of incomplete input information we set the reverse problem, i.e. to reestablish full input information on the base of the model and observing performance characteristics. Thus besides predicting the performance for the known system state and workload the model allows obtaining the description of the current NAS instance. For example, planing NAS availability needs to forecast its workload. The forecast needs precise information on the current workload but the whole arrival is not registered at the system as we mentioned above. The model rebuilds the actual arrival rate.

We again address the Erlangian queue for simplicity. Let p_n be the probability of the number of busy channels, T is the observation period and T_n is the amount of time when any n channels was busy simultaneously. We accept $q_n = \frac{T_n}{T}$ as an estimation of p_n and then consider the

least square function.

$$\sum_{i=1}^{m} \left(p_0(\rho) \frac{\rho^i}{i!} - q_n \right)^2 \longrightarrow \min . \tag{5.1}$$

Let $\tilde{\rho}$ be the value minimizing the least square function. Hence the total arrival flow is estimated as $\tilde{\lambda} = \tilde{\rho}\mu$. If the service rate information is also incomplete then μ should be estimated from another set of data. For instance one may use periods of light load where the arrival flow is fully observable.

The case of incomplete information about the arrival rule needs to estimate the functions $\lambda_i(N)$.

6 Conclusion

Due to the increasing role of the NAS for Internet communications we present an approach which aims at evaluation of the key performance features of the NAS. The approach bases on our recent results concerning the heterogeneous queuing model with loss. The model is developed to reflect the main features of the NAS. We present several examples which illustrate how to use the model to evaluate availability and utilization of the current NAS implementations and how to plan future modifications. We also present the extension of a queueing model which treats the details of the data transmission process. The extended model is illustrated by the example of NAS with unreliable channels. The rest of the paper considers the issues of the model input parameters estimation.

References

[1] Mitton D., Beadles M., *Network Access Server Requirements Nest Generation.* NAS Model, RFC 2881, July 2000.

[2] Mitton D., *Network Access Servers Requirements: Extended RADIUS Practices.* RFC 2882, July 2000.

[3] Gnedenko B. V., Kovalenko I. N., *Queuing theory.* Jerusalem: Israel Program for Scientific Translation . 1968.

[4] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, *On the Self-Similar Nature of Ethernet Traffic.* IEEE/ACM Trans. on Networking, vol. 2, N 1, 1994, pp. 1–15

[5] Khinchin A.J., *On the Poisson flows of random events.* Theory Prob. Appl. Vol. 1, N 3, 1965, pp. 320–327.

[6] Willinger W., Paxon V., *Where Mathematics meets the Internet.* Notices of the American Mathematical Society, 45(8), pp. 961–970, Sept., 1998.

[7] Schassberger R. *Insensitivity of steady state distributions of generalized semi-Markov process with speeds.* Adv. Appl.Prob. 6, 1978, pp. 836–851.

[8] Bogoiavlenskaia O. I., *Access System Modeling by Queues with Compound Customers.* Proceedings of FDPW'97-98, vol. 1, University of Petrozavodsk, 1998. pp. 88–111.

[9] Bogoiavlenskaia O. I., *The Decomposition Property of the Blocking Queueing Model in a Random Environment.* Proceedings of FDPW'99, vol. 2, University of Petrozavodsk, 1999. pp. 46–56.