Performance Modelling and Evaluation of Firewall Architectures for Multimedia Applications

Utz Roedig[1] and Jens Schmitt[2]

[1]Mobile & Internet Systems Laboratory (MISL), University College Cork, Ireland
utz@cs.ucc.ie

[2]Distributed Computer Systems Lab (DISCO), University of Kaiserslautern, Germany
E-Mail: jschmitt@informatik.uni-kl.de

Abstract. Firewalls are a well-established security mechanism to restrict the traffic exchanged between networks to a certain subset of users and applications. In order to cope with new application types like multimedia applications, new firewall architectures are necessary. The performance of these new architectures is a critical factor because Quality of Service (QoS) demands of multimedia applications have to be satisfied. We show how the performance of firewall architectures for multimedia applications can be determined. A model is presented which can be used to describe the performance of multimedia firewall architectures. This model can be used to dimension firewalls for usage with multimedia applications. In addition, we present the results of a lab experiment, used to evaluate the performance of a distributed firewall architecture and to validate the model.

1 Motivation and Introduction

Within a global networked environment, security aspects have become more and more important and access control at network borders is considered essential. For this purpose firewalls are used. As an integral part of the network infrastructure, firewalls are strongly affected by the development and deployment of new communication paradigms and applications. Recently, there has been a rise in the use of multimedia applications which, from the perspective of firewalls, differ in many aspects from “traditional” applications. One of the most important aspects is the difference in performance requirements. Existing firewalls are not able to support multimedia applications in an efficient and secure manner [1]. In particular, a traditional firewall may not be able to support the QoS requirements of a multimedia application.

To overcome these deficiencies, new firewall architectures are currently discussed and proposed. Besides many other facets - e.g. security, maintainability, flexibility - these are intended to optimize firewall performance. Of course, all these characteristics have to be optimized simultaneously to meet the given requirements.

Currently, appropriate methods and tools to evaluate the performance of multimedia firewall architectures are missing. Hence, ascertained performance parameters of proposed firewall architectures are also unavailable. To solve these problems the following topics are covered in this paper:

(i) Analysis of performance bottlenecks in multimedia firewall architectures;

(ii) Performance modelling of multimedia firewall architectures;

(iii) Experimental performance evaluation and model validation.

In the remaining paragraphs of this section the terms "multimedia application" and "firewall architecture" are described in detail as they are used in the context of this paper. In Section 2, the parameters which characterize the performance of a multimedia firewall are defined. Further, performance bottlenecks in firewall architectures are analyzed. In Section 3, the performance model is introduced. In Section 4, the lab experiment is described, including measurement methods and tools that were used. In Section 5, the experimental results are compared with the model and the model is validated. Section 6 reviews related work. In the last section, our findings are summarized.

Multimedia Applications. Multimedia applications use a combination of continuous and discrete media data, with the continuous media usually being audio and/or video streams. The discrete media often consist of control data streams for the audio and video data streams and additional information.

In order to describe communication scenarios, the following terms to distinguish the granularity at which an application's data stream is considered are defined. A *flow* is a single data stream, identified by a tuple of characteristic values (e.g. source address, source port, destination address, destination port, protocol number). A *session* describes the association of multiple flows which together constitute an application's data stream.

Firewall Types and Architectures. A firewall examines all network traffic between connected networks. Only data that is explicitly allowed to, as specified by a security policy, is able to pass through it. The tasks of a firewall are well defined, but there are many possible firewall architectures to fulfil them. Firewalls may consist of different firewall components, e.g. filters, stateful filters or proxies. In addition, the applications may interact explicitly with a firewall to support it to fulfil its task.

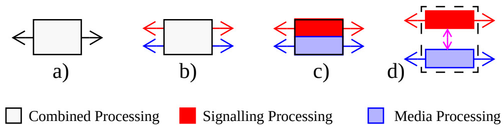

Figure 1 Firewall Types

To select a useful architecture for the usage in conjunction with multimedia applications the following basic evolution of firewall types - illustrated by Figure 1 - has to be taken into account [1]. Figure 1a) abstractly describes the behavior of a "standard firewall". All traffic is sent through the firewall component which is responsible to apply the security functionality. In this case the specific characteristics of multimedia applications' traffic are not taken into account. If these specific characteristics (as shown in Figure 1b)) are regarded it is obvious that the same firewall component has to take care of different traffic types of the different traffic flows (control and media flows). In this case, it is not possible to adapt the one firewall component to the needs of the two different flow types. This results in many problems, in particular performance problems [2]. To overcome this weakness, two different firewall components for the processing of the two different flows can be used (Figure 1c)) [1]. This additional degree of freedom allows specific component optimizations for the different flow types. To maintain session state within the firewall, information exchange between the components is necessary. If the separation between signalling and media processing is further extended by even physically distributing them (Figure 1d) additional optimizations are possible [1], [3]. In this case the information exchange between the components has to be realized by an appropriate network protocol [4]. The implementation of the useful firewall types shown in Figure 1c) and Figure 1d) lead to different multimedia firewall architectures which are currently proposed. The focus is on these architectures in the remaining paper:

- **Architecture AI** (implementation of firewall type c)): The firewall consists of a single computer system containing a signal and media flow processing component. Well known firewalls following this design principle include firewall products like CISCO's PIX and Checkpoint's Firewall-1.
- **Architecture AII** (implementation of firewall type d)): The firewall consists of several computers. A well defined interface between signalling and media processing component(s) is used. A practical implementation of such an architecture is the Netscreen 500 firewall for SIP based IP-telephony applications [5].
- **Architecture AIII** (implementation of firewall type d)): In this case, the available signalling processing component within multimedia applications in end systems is used. By choosing this architecture, the need of centralized signalling processing components is avoided. These systems are not used today, but theoretical work exists [4].

To select one of the architectures, one has to consider the advantages and disadvantages and rate how important they are in the considered target scenario. Independent from these considerations, the firewall system has to be dimensioned to meet the QoS requirements of multimedia applications. It is necessary to know how many signalling and media processing units are necessary and what capacity they should have.

2 Firewall Performance

To determine the performance of a multimedia firewall architecture it is necessary to define the term performance in this context first. The performance of a firewall, respectively of a firewall architecture, is defined by:

(i) its influence on applications' QoS parameters

(ii) its total capacity

The influence on QoS parameters of multimedia applications by a firewall within the communication path should be low and predictable. The maximum possible throughput, its capacity, should be as high as possible.

2.1 Quality of Service Parameters

To be able to rate the performance of a multimedia firewall, useful quality parameters have to be defined. These quality parameters should allow the objective validation of a firewall's performance. In the following, the necessary QoS parameters of multimedia applications are described. From these parameters quality parameters necessary to rate firewalls are derived.

Signalling Flow. The quality of the signalling plane is mostly influenced by the session setup delay. If the necessary time for a session setup is too long, a user of a multimedia application will feel disturbed or will regard the connection's quality unacceptable. The following definition is used:

> *The* **session setup time** T_S *is the time from the setup of the control flow till the start of the first media flow.*

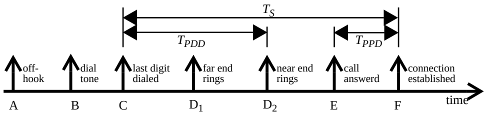

Figure 2 Session setup steps

The determination of boundary values and an exact definition depends on the type of investigated application. The session setup time can also be divided in substeps, which might be subject to different requirements. The requirements for the session setup time for IP-telephony applications are described below, because these applications are used in the experiment described in Section 4. Figure 2 describes the substeps within the session setup as used in H.323 based IP-telephony applications [7]. In this case, the session setup time is given by $T_S = F - C$. In addition, the post dial delay $T_{PDD} = D_2 - C$ and post pickup delay $T_{PPD} = F - E$ can be defined. The post pickup delay is particulary critical. If the latter value is too high, the first words of the conversation are lost because the media channels are not yet established. Boundary values can be derived from values given for ISDN networks [6]. The post dial delay should be between 2 and 7 seconds, the post pickup delay should be between 0.75 and 2 seconds.

Media Flow. The media flows also have to meet specific requirements. Possible effects if specific bounds are violated might be for example echo or noise. The characteristic parameters to describe the quality of a media stream are delay $T_{D(i)}$, jitter $T_{J(i)}$ and loss L. As the experiments described in this paper target the control plane, we refer to [1] for a detailed definition and explanation of theses parameters.

Quality Index. Firewall quality indices can be derived from the previously described QoS parameters of multimedia applications. The following definition for quality indices is used:

> *The quality index G_X defines the percentage of the upper bound of a QoS parameter X of a specific multimedia application that is consumed by the firewall.*

The different quality indices may depend on the number n of similar active application sessions that are handled by the firewall. The quality indices are then given by:

$$G_X(n) = \frac{\Delta X(n)}{X_{max}} \qquad X \in \{T_S, T_D, T_J, L\} \tag{1}$$

with $\Delta X(n)$ describing the value consumed by the firewall and X_{max} representing the selected upper bound of the investigated QoS parameter.

2.2 Capacity of Firewall Architectures

The capacity of a firewall can be determined by the definition of upper bounds G_{Xmax} for the four different quality indices. The capacity is defined as:

> *The capacity N of a multimedia firewall is given by the number of concurrent active sessions such that*

$$G_X(n) \le G_{Xmax} \qquad \forall n \le N, \forall X \tag{2}$$

In the following section bottlenecks in firewall architectures, their influence on the capacity and also the resulting impact on the dimensioning of firewalls is discussed.

Filter Bandwidth. The media flow processing within the firewall architectures described in Section 1 is normally implemented as a packet filter. For these filters, the maximum bandwidth B, which normally depends on the packet size s, is known. If the number of media flows r that are used for a specific multimedia application and the bandwidth b of these flows is also known, the upper bound on the capacity N_B of the firewall can be calculated:

$$N_B = \frac{B(s)}{r \cdot b} \qquad (3)$$

The bandwidth used for the signalling and media control flows are not taken into account because they are small compared to the bandwidth of the media flows. In addition, it is assumed that the quality indices are within the boundary values according to equation (2).

Session Setup. The component used to process the signalling flow is limited in the amount of packets that can be processed in a certain time period. Therefore, a limit on the amount of session setups per second μ that can be handled exists. If it is assumed that all applications have duration T and further that the session setups are uniformly distributed, the upper bound on the capacity is given by:

$$N_S = \mu \cdot T \qquad (4)$$

In firewalls used today, the capacity of a firewall is mainly constrained by the signalling processing component, not by the available filter bandwidth (N_B » N_S). To overcome this shortage, several signalling processing components p might be used. For each additional component the gain might be reduced (given by the parallelization efficiency α) due to the distribution overhead:

$$N_S = p \cdot \alpha(p) \cdot \mu \cdot T \qquad (5)$$

Summary. As shown, it is necessary to regard both factors, filter bandwidth and session setup, to determine the capacity of a multimedia firewall architecture. Especially in firewalls used today (implemented according to architecture AI) the session setup factor is not taken into account. This might lead, depending on multimedia applications characteristics, to a waste of resources and a lower than expected performance.

3 Performance Model

Today the performance behavior of media processing components (e.g. packet filters) is generally well understood and manageable. For the various available components, characteristic curves for the media flows related quality indices are directly or indirectly available. It is also possible to obtain media processing components suited for high bandwidths, so that desired bounds for the media flow quality indices can be met.

In contrast, the performance of signalling processing in multimedia firewalls has not been investigated in depth. To be able to state and predict the performance behavior of the signalling processing, a generic performance model is necessary. In this section, a performance model for the signalling processing using queuing theory is developed.

In Section 4, a lab experiment is carried out which is used to verify the developed performance model.

3.1 Modelling of Multimedia Firewall Types

The signalling processing component of the multimedia firewall types (described in Section 1, Figure 1) can be modelled as shown in Figure 3.

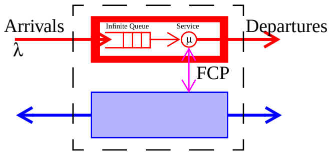

Figure 3 Queueing System

With a rate of λ and a certain statistical distribution, new sessions arrive at the signalling processing component. In order to keep the model tractable but also due to many empirical studies on session arrival characteristics it is assumed that the session inter-arrival time is exponentially distributed. The queue is assumed to be infinite. This means the space (available memory) for waiting sessions is assumed to be sufficient at all times. The processor within the signalling processing component is able to process session setups with a rate of μ. The service time has a general distribution with average $1/\mu$ and variance σ^2. The variance of the service time is caused by the necessary communication between signalling and media processing components using a Firewall Control Protocol (FCP). The necessary processing time T_P for each session setup is composed of the following time segments:

$$T_P = i \cdot T_{sig} + k \cdot T_{FCP} \tag{6}$$

First T_P comprises the necessary and constant processing time T_{sig} for the i exchanged signalling messages used for session setup, second the time T_{FCP} necessary to submit and process k FCP messages (e.g. containing flow specifications) is included. T_{FCP} might have a statistical distribution if T_{FCP} is strongly influenced by queueing effects within the FCP message handling in the signalling or media processing component or by the characteristics of the network used to transport the FCP messages.

The resulting queueing system to model the behavior of one signalling component is therefore an M/G/1 queue according to Kendall's notation. If p signalling processing components are used, the arrival rate λ' for each queueing system is:

$$\lambda' = \frac{\lambda}{p} \tag{7}$$

If T_{FCP} can be considered small compared to T_{sig} or shows little fluctuations, the service time can be assumed to be constant. In this case, the resulting queueing system to model the behavior of one signalling component is an M/D/1 queue.

To be able to predict the session setup time, the expected queueing delay (= expected session setup time introduced by the firewall) $E(\Delta T_S)$ for the queueing system has to be calculated. The expected queueing delay in an M/G/1 system is given by [8]:

$$E(\Delta T_S) = \frac{1}{\mu \cdot (1-\rho)} \cdot \left(1 - \frac{\rho}{2}(1 - (\mu^2 \cdot \sigma^2))\right) \text{ with } \rho = \frac{\lambda}{\mu} \tag{8}$$

For the special case of a deterministic (constant) service time, the variance of the service time is zero ($\sigma^2 = 0$). In this case, (8) gives the expected queueing delay in an M/D/1 system.

3.2 Performance Models for Firewall Architectures

To model the architectures presented in Section 1, the number of signalling processing components has to be taken into account.

Architecture AI. For the hybrid architecture, where only one processing component is available, (8) can be used directly to give a model for the firewall's session setup time ΔT_S in relation to the number of sessions n. If the duration of the sessions is assumed to be constant, $\lambda = n/T$ is obtained. With (8) the following model is obtained:

$$\Delta T_{S_{AI}}(n) = \frac{1}{\left(\mu - \frac{n}{T}\right)} \cdot \left(1 - \frac{n}{2 \cdot \mu \cdot T}(1 - (\mu^2 \cdot \sigma^2))\right) \tag{9}$$

Architecture AII. To model the locally distributed architecture, p signalling processing components have to be taken into account. Each signalling processing component comprises an M/G/1 queue. Therefore the arrival rate is split among the processing components and depends on their number p. Using (8) the following model is obtained:

$$\Delta T_{S_{AII}}(n) = \frac{1}{\left(\mu - \frac{n}{T \cdot p}\right)} \cdot \left(1 - \frac{n}{2 \cdot \mu \cdot T \cdot p}(1 - (\mu^2 \cdot \sigma^2))\right. \tag{10}$$

Architecture AIII. Within the totally distributed firewall architecture, for each session a distinct signalling processing component is available. Using (10) and $p = n$ the following model results:

$$\Delta T_{S_{AIII}}(n) = \frac{1}{\left(\mu - \frac{1}{T}\right)} \cdot \left(1 - \frac{1}{2 \cdot \mu \cdot T}(1 - (\mu^2 \cdot \sigma^2))\right) \tag{11}$$

Therefore, $T_{S_{AIII}}(n)$ is constant and does not depend on the number of concurrent active sessions. If it is also assumed, that the session duration T is long and the service rate μ is high, $\Delta T_{S_{AIII}} \approx 1/\mu$ is obtained.

Summary. With $\sigma^2 = 0$ all three analytical models can also be adapted to the assumption of constant service times (M/D/1). Using equation (1), the session setup time given by the models can be used to determine the firewall's session setup quality index.

4 Performance Evaluation

To gain realistic performance numbers for the session setup quality index of multimedia firewall architectures, a lab experiment has been conducted. The results of the experiment are used in Section 5 to validate the performance models developed above.

4.1 Measurement Tool

To be able to determine the quality indices of firewalls, the traffic generator and a measurement tool *KOMtraffgen* [1] is used.

Core. The KOMtraffgen tool can be used to generate traffic of concurrently running multimedia applications. The exact behavior, control and media flows of each individual application, is modeled. The software is divided into two parts, the core and the application specific part. The core carries the generic parts, e.g. measurement facility, timer and hooks to include the application specific parts. The application specific part carries the state machine (client or server side) of the emulated application.

Application. To carry out the experiments, an application with IP-telephony like characteristics was implemented (see Figure 4).

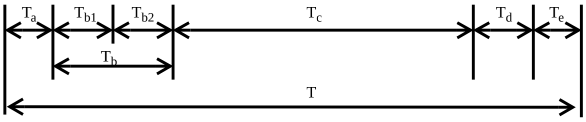

Figure 4 Test application - Time chart

At the beginning of the communication a TCP control flow between both endpoints is set up (T_a). On the control channel, the parameters for the subsequent audio communication are negotiated (T_b; T_{b1} is the post dial delay, T_{b2} is the post pickup delay). Then the audio flows are initiated and media packets are exchanged. The session setup time as well as the media QoS parameters are measured. When the session time is exceeded (T_c), the session teardown is initiated. Appropriate messages are exchanged on the control channel and the media channels are closed (T_d), finally the control flow is closed (T_e). The session setup time according to the definition in Section 2.1 is given by: $T_s = T_a + T_{b1} + T_{b2}$.

Configuration. The *KOMtraffgen* system has to be configured by specifying the number n of concurrent active application sessions and the session duration T_c. Also the specification of the media flows has to be given (packet rate, packet size).

The time between session setups is exponentially distributed which generates a Poisson process of session setups. The setup rate is implicitly specified by: $\lambda = n / T_C$.

Calibration. Before the quality index of the firewall can be determined, a calibration measurement without any firewall intervention is necessary. Two computers, one running the client part of *KOMtraffgen*, the other one running the server part of *KOMtraffgen* are connected via a 100 Mbit Ethernet switch and an intermediate router (see Figure 5). Then the session setup times for different setup rates are measured. The setup rate λ is adjusted by varying T_c with a fixed n. The calibration curves are later used to determine the difference in the session setup time introduced by the analysed firewall.

4.2 Experiment Setup

For the experiment, two different firewall systems - shown in Figure 5 - have been used. The first firewall system (FWa) is an implementation of architecture AI (see Section 1), the second firewall system (FWb) is an implementation of architecture AII.

Both firewall systems are based on firewall components, called the *KOMproxyd* system implemented by ourselves [1]. Our own firewall implementation was necessary for two reasons. First, a locally distributed firewall with several signalling processing units (according to FWb) is not available. Second, it is necessary to be able to compare the measurement results of the two firewall systems. This is only possible if both systems only differ in the interaction between signalling and media processing. If both systems are internally structured differently it is nearly impossible to determine performance differences caused by the architectural changes.

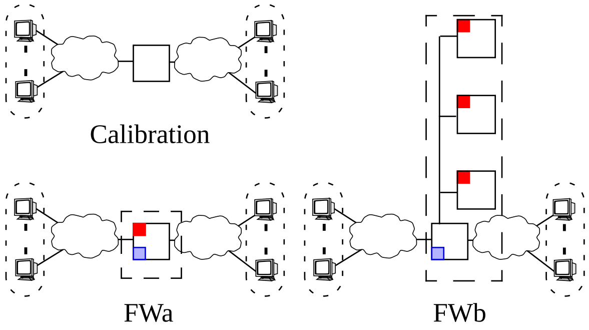

Figure 5 Experiment Setup

In the first scenario (FWa), the interaction between the signalling and media processing component is implemented as I/O-controls. In the second scenario, the exchanged information between the components is transported by a reliable UDP-based Firewall Control Protocol (FCP). All machines used are PIII 450 MHz with a FreeBSD 4.5 operating system. All links are 100 MBit full dupex switched Ethernet.

4.3 Experiment Results

First the monolithic and centralized firewall system FWa is tested. *KOMtraffgen* is parameterized with $n = 50$ concurrent sessions. The setup time required for small session setup rates is nearly constant with $T_{S_{FWa}} = 24ms$. As the load increases, the setup times rise steeply. If $T_{S_{max}} = 6s$ is defined (according to boundary conditions for telephony calls as stated in Section 2.1) and a quality index of $G_{T_S}(n) \leq 2.5\%$ is recommended, we obtain using (1) with $T_c = 180\text{s}$ (standard phone call duration) a total capacity of $N_{S_{FWa}} = 7112$.

Second, the firewall system FWb with 1, 2 and 3 processing units is tested. For the measurement of FWb with one processing unit ($p = 1$), *KOMtraffgen* is parameterized with $n = 50$ concurrent sessions. For the measurement of FWb with $p = 2$ resp. $p = 3$ processing units $n = 100$ resp. $n = 150$ concurrent sessions are used.

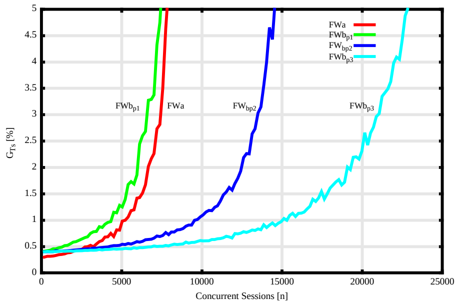

Figure 6 Quality index $G_{T_S}(n)$ for FWa and FWb

Using (1), the calibration measurements and the measurement results, the quality indices $G_{T_S}(n)$ as shown in Figure 6 result.

The quality index for the FWb system with one processing unit ($p = 1$) is always higher than the quality index of FWa. The setup time required for small session setup rates is nearly constant with $T_{S_{FWbp1}} = 31ms$. The difference between FWa and $\text{FWb}_{\text{p=1}}$ is caused by the difference in the communication between signalling and media processing component. The transportation of necessary information (e.g. flow specifications to adjust the filter configuration of the media processing component) over the network accounts for an additional $7ms$. Therefore, the total capacity is $N_{S_{FWbp1}} = 6188$. For the measurements with multiple signalling processing components the following values have been obtained: $N_{S_{FWbp2}} = 13073$, and $N_{S_{FWbp3}} = 20077$.

4.4 Discussion

The experiment shows, that the distributed firewall architecture (AII) with $p > 1$ signalling components can be used to overcome the limits of a hybrid system (AI). Therefore, the trend towards distributed firewalls as currently discussed is justified.

Example. The measurement results obtained for the session setup delay can be used to dimension a firewall system. If an application with $r = 2$, $b = 87.2\frac{Kbit}{s}$, $T = 180\mathrm{s}$ and a media processing component with $B = 2\frac{Gbit}{s}$, $N_D = 11468$ is assumed, architecture AII with $p > 1$ as used in the experiment is necessary to be able to fully utilize the available media processing capacity.

Comparison. If the total capacity of FWa and FWbp1 is compared, we see that 12.9% of the processing capacity of the signalling component has to be spent to implement the FCP communication. Therefore, architecture AII is only useful regarding performance optimization if used with $p > 1$.

Efficiency. If the total capacity of the firewall system FWb is compared using equation (5) we obtain: $\alpha(p = 1) = 1 \Rightarrow \alpha(p = 2) = 1.06$; $\alpha(p = 3) = 1.08$

At first glance it is surprising that the efficiency α is slightly greater than 1 and that this factor is nearly independent from the degree of distribution. Yet, according to the performance model introduced in Section 3, this behavior has to be expected. A detailed comparison of the model and the experimental results is given in the next section.

5 Comparing Model and Experiment

For the comparison of the experimental results and the models introduced in Section 3, values for the variables μ and σ reflecting the experiment have to be determined.

Adaptation. To determine the service rate μ and the variance of the service time σ the appropriate model curve is fitted to the measurement curve using:

$$\begin{bmatrix}\mu\\ \sigma\end{bmatrix} = \underset{0 \le \mu,\, \sigma \le \infty}{\arg min}\left\{\sum_{\aleph \le N} (\Delta T_{S_A}(n, \mu, \sigma) - \Delta T_{S_{FW}}(n))^2\right\} \qquad (12)$$

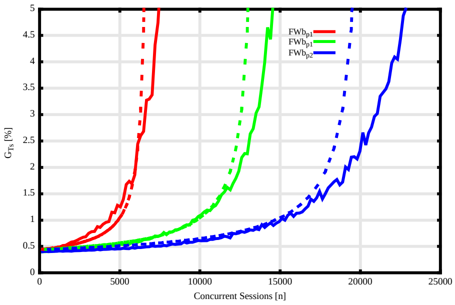

Figure 7 Quality index $G_{T_S}(n)$ FWb

For the model AI fitted to FWa $\mu = 42.5$ and $\sigma = 0.01$ is obtained. If the model of AII is fitted to the measurement curve of FWbp1 $\mu = 37.9$ with $\sigma = 0.01$ is obtained. In both cases, the variance of the service time σ is very close to 0. Thus, this gives evidence that the investigated firewalls process sessions with nearly constant service time. Therefore, the simplified models based on a M/D/1 queue are applicable (see Section 3.1).

Comparison. For the comparison, the values μ and σ obtained from the fitting described before are used. Again the quality index $G_{T_S}^{AII}(n)$ for the session setup delay is determined. With $T_{S_{max}} = 6s$ and $T_c = 180\mathrm{s}$ using (1) the results shown in Figure 7 are obtained. Figure 7 also shows the measurement results $G_{T_S}^{FWb}(n)$ of Section 4.3.

Quality of Prediction. The model for FWa and FWbp1 can be used to calculate the FCP communication overhead. This number can be compared with the communication overhead determined by the experiment (Section 4.4). For the measurement an overhead of 12.9%, for the model an overhead of 10.8% is obtained (16% deviation).

If the model to determine the total capacity of the system assuming a recommended quality index of $G_{T_S}^{AII}(n) \leq 2.5\%$ is used, the results shown in Table 2 are obtained.

Table 1: Total Capacity for $G_{T_S}^{AII}(n) \leq 2.5\%$

Measurement	Model	Deviation
$N_{S_{FWbp1}} = 6188$	$N_{S_{AIIp1}} = 6156$	0.05 %
$N_{S_{FWbp2}} = 13073$	$N_{S_{AIIp2}} = 12312$	5.8 %
$N_{S_{FWbp3}} = 20077$	$N_{S_{AIIp3}} = 18468$	8.0 %

As it can be seen (Figure 7), the prediction of the the model regarding the total capacity tends to be more precise in the area where the signalling processing components are not stressed by heavy load. Compared with the experiments described in Section 4 the model allows us to predict the quality index curve (with $G_{T_S}^{AII}(n) \leq 2.5\%$) with a deviation of at most 8%.

6 Related Work

The performance of firewalls has always been a critical issue. Therefore, much research work has been carried out in the past regarding this topic. For basic firewall performance tests, standardized methods exist [9]. However, none of the previous work covered the investigation of the performance of *multimedia* firewalls and especially of perfomance bottlenecks on the signalling path.

Many firewall vendors provide performance evaluations of their firewalls (e.g. [10]). These evaluations do not give an exact description of the performed measurements. In addition, these evaluations focus on other protocols like HTTP or FTP and so the results cannot be transferred to describe the behavior of a firewall in interaction with multimedia applications. Some firewall vendors provide information about the performance evaluation in conjunction with multimedia applications resp. UDP processing [5]. Yet, these investigations only cover the media processing and make no statements about the signalling processing.

Beside the performance evaluation of firewalls, performance evaluations of multimedia components are available (e.g. performance evaluation of IP-telephony components [11]). These results also cannot directly be transferred to firewall architectures.

7 Summary

The work presented allows a rating and selection of firewall architectures for multimedia applications regarding performance issues. Therefore, the work clarifies many questions regarding firewall architectures that had been recently discussed (e.g. in the IETF). The contributions of the paper can be summarized as follows.

Bottlenecks. In the paper bottlenecks of multimedia firewalls were identified and analytically described. Lab experiments verified their existence. In particular, bottlenecks caused by the signalling processing component of a multimedia firewall were investigated.

Evaluation. Measurement methods that can be used to rate the performance of multimedia firewalls were developed and described. In addition, publicly available measurement tools are provided that can be used to perform firewall performance evaluation.

Modelling. In the paper a queueing model to describe the performance behavior of multimedia firewalls was introduced. This model was validated by a lab experiment.

Application. The above summarized results of the presented work allow two main applications. First, it is possible to use the analytical model to dimension multimedia firewalls. With the now available methods an unnecessary waste of resources can be avoided. Second, the model can be used to integrate a firewall actively in a network providing some form of QoS assurances. The model can be used to predict the behavior of a firewall and thus allows the derivation of information necessary for a dynamic admission control in a QoS-supporting network.

References

[1] U. Roedig. Firewall Architectures for Multimedia Applications. PhD thesis, Darmstadt University of Technology, November 2002.

[2] R. Knobbe, A. Purtell, and S. Schwab. Advanced security proxies: an architecture and implementation for high performance network firewalls. In Proceedings of DARPA information survivability conference and exposition 2000, pages 140–148, 2000.

[3] P. Srisuresh, J. Kuthan, J. Rosenberg, A. Molitor, and A. Rayhan. Middlebox communication architecture and framework. Internet Engineering Task Force, RFC 3303, August 2002.

[4] U. Roedig, M. Görtz, M. Karsten, and R. Steinmetz. RSVP as Firewall Signalling Protocol. In Proceedings of the 6th IEEE Symposium on Computers and Communications, Hammamet, Tunisia, pages 57–62. IEEE, July 2001.

[5] NetScreen. NetScreen-500 System Product Description. P.Num.: 2002.6.50.1.500, 2002.

[6] International Telecommunication Union. Network grade of service parameters and target values for circuit-switched services in the evolving ISDN. Recommendation E.721, Series E: Overall Network Operation, Telephone Service, Service Operation and human factors. Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 1999.

[7] European Telecommunications Standards Institute. End-to-End Quality of Service in TIPHON Systems; Part 2: Definition of speech Quality of Service (QoS) classes. Draft, Telecommunications and Internet Protocol Harmonization over Networks, ETSI, 2000.

[8] L. Kleinrock and R. Gail. Queueing Systems: Problems and Solutions. John Wiley & Sons, 1996.

[9] B. Hickman, D. Newman, S. Tadjudin, and T. P. Martin. Benchmarking Methodology for Firewall Performance. Internet Engineering Task Force, RFC 3511, April 2003.

[10] The Tolly Group. Test summary NetScreen-5200 versus Nokia IP740 and Cisco Systems Inc. PIX 535. Document No. 202121, March 2002.

[11] T. Eyers and H. Schulzrinne. Predicting Internet Telephony Call Setup Delay. In Proceedings of the 1st IP-Telephony Workshop (IPtel 2000), Berlin, Germany, April 2000.