Congestion Avoidance in S-CSCF Selection in an IMS Network

Plarent Tirana and Deep Medhi
Networking & Telecommunication Research Lab
Computer Science & Electrical Engineering Department
University of Missouri–Kansas City, USA

***Abstract*—In the signaling plane of an IMS (IP Multimedia Subsystem) network, the interrogating Call/Session Control Function (I-CSCF) is required to choose the serving CSCF (S-CSCF) for user requests arriving at the proxy CSCFs (P-CSCF). This requires S-CSCF selection in such a way that the response time is minimized. However, in an overloaded situation, the P-CSCF entry points for arriving requests must balance between rejecting new requests and having an acceptable response time from S-CSCF for those that are allowed to enter. In this setting, we consider a number of controls for congestion avoidance that work in conjunction with S-CSCF selection in an IMS Network. For balancing between rejection of requests and the response time, we also propose a composite metric. Through our studies on grid topologies, we observe that one of the avoidance algorithms, the bang-bang control, generally shows to be the most promising toward striking a good balance between rejection of requests and the response time.**

I. Introduction

IP Multimedia Subsystem (IMS) is a generic open-systems architecture offering converged multimedia services over IP [6], [27] (Fig. 1). Three types of SIP (session initiation protocol) proxies are defined by 3GPP for IMS signaling, commonly known as CSCF (Call/Session Control Function). These proxies are categorized by their signaling functionality as P-CSCF (proxy CSCF), I-CSCF (interrogating CSCF) and S-CSCF (serving CSCF). Note that in an operational environment, multiple instances of different CSCFs may be distributed geographically across the network. Two of these types (P-CSCF and S-CSCF) are assigned to the IMS terminal (user) for the entire registration process, while the I-CSCF acts as a load balancer only during the IMS registration phase and selects the best-fit S-CSCF for the user. This process is defined as the *S-CSCF Assignment* problem by 3GPP. The P-CSCF acts as a liaison for the user in the signaling plane. In fact, it intercepts every SIP message destined/originated to/from an IMS terminal, decides which S-CSCFs to invoke, checks the user profile and authorizes the services.

In our recent work [31], [32], we have investigated S-CSCF assignment algorithms in an IMS network in a moderately loaded environment. Generally, the S-CSCF selection is a load balancing problem. Load balancing is not a new concept in distributed computing. There has been significant work on server load balancing in the Internet [4], [7], [20]. However, little work has been done on server load balancing in the IMS network. For instance, in an IMS network, because of multiple S-CSCFs for handling requests located in diverse geographical regions and with I-CSCF being involved as intermediaries, the knowledge available at each node is not instantaneous. Therefore, the latency may be measured at the receiving end, but feeding them back to the entry points has a time lag. Thus, we take such issues into account to study S-CSCF selection algorithms in our recent work [32]. Despite SIP being transport protocol agnostic, 3GPP specifies that TCP be used for SIP for IMS signaling and for stateful treatment of sessions. With this stateful treatment, the I-CSCF will assign the S-CSCF at the registration time (beginning of the session) and stick with it for the entire session duration for a particular request. The P-CSCF will send the SIP messages directly to the S-CSCF without again involving the I-CSCF for this request.

The focus of our previous work [31], [32] was to study S-CSCF selection in which no incoming requests were dropped when the load is low or moderate; instead, we allowed the latency to be built up, while at the same time identifying the best S-CSCF selection; in other words, our previous work did not consider overloaded conditions. In this paper, we take into consideration, the case of overload for requests arriving at P-CSCFs. In order not to unduly impact the latency in response from the S-CSCF in an overload situation, this case forces some requests to be rejected at the entering P-CSCFs. However, it is important to be proactive in rejecting at the P-CSCFs by taking a congestion avoidance approach. In this regard, we consider a number of access control schemes for congestion avoidance.

There have been significant works on congestion avoidance for a variety of networks (see Section V). On the other hand, the congestion avoidance problem in the S-CSCF selection in an IMS network is quite unique. This is mainly due to the fact that a new SIP request arriving at a P-CSCF has to be routed to an S-CSCF, where all packets for this request that are generated at the P-CSCF are processed through a TCP connection within the IMS network to its destination S-CSCF; this is to be done while keeping the response time as low as possible. That is, for an accepted request, ideally no packets for this TCP connection are to be dropped either at the P-CSCF, I-CSCF, or S-CSCF (unlike active queue management in the Internet) as these packets are going over an IP network that is used exclusively for the IMS architecture so as to ensure that the response time is minimized. On the other hand, new requests arriving at a P-CSCF can be rejected in an overloaded situation. Thus, congestion avoidance in S-CSCF selection in

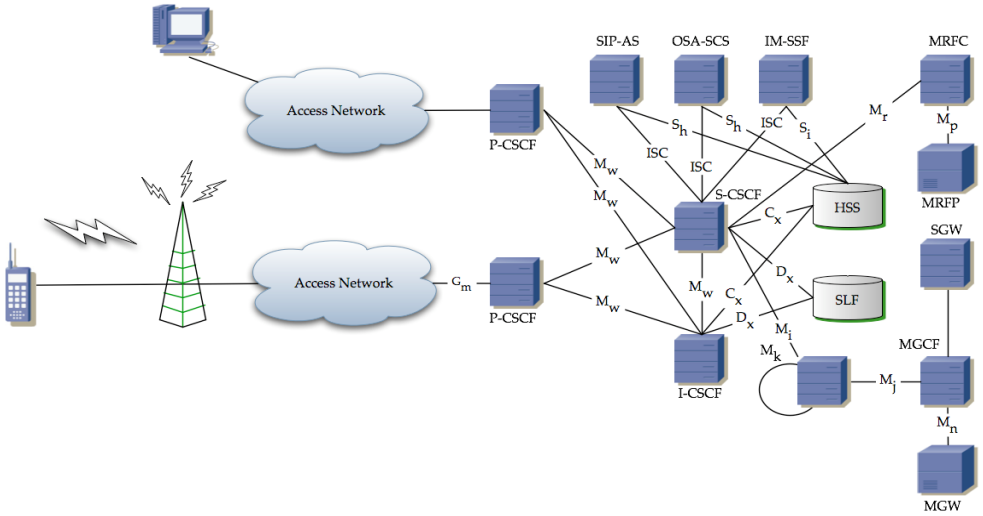

Fig. 1. Reference IMS Architecture

an overloaded environment requires a balancing act between request denial for new arriving requests and the response times of those accepted—this is what we attempt to understand in this work. To our knowledge, this problem has not been addressed in the literature for an IMS network. In this context, we discuss a number of access control schemes for congestion avoidance that can be combined with the S-CSCF selection process.

The rest of the paper is organized as follows. In Section II, we present an overview of the S-CSCF selection process. We then introduce access control algorithms for congestion avoidance in Section III. Then in Section IV, we present simulation studies and analyze the results. In Section VI, we present summary and future work.

II. S-CSCF Selection: A Review

In this section, we present a brief review of the S-CSCF selection environment; this is detailed in [32]. This review is included here to present the context of the access controls schemes for congestion avoidance, which will be presented in the next section.

We consider an IMS environment where a set of K I-CSCFs can share the load with a set of N S-CSCFs while there are M independent P-CSCFs (input sources) as shown in Fig. 2. This environment is more generic than the scenario of a single I-CSCF with N S-CSCFs sharing the load considered in other work [6], [8]. In this environment, the measure of importance is the response time for request routing and processing to understand the user's perception of set-up delay. First, we note that the role of an I-CSCF is primarily that of a lookup function to forward a request to the right S-CSCF; the lookup time (however small) would be reduced with distributed I-CSCFs across the network. However, it is crucial for the fastest response time algorithm that the I-CSCF has some knowledge of the overall delay between any P-CSCF and the S-CSCFs. This information, however, may not be available instantaneously to all entities in the environment. Thus, to account for this issue in an actual operational scenario, we have designed an update protocol to convey the delay information to the K I-CSCFs. Briefly, the update protocol includes both pushes from the P-CSCFs and S-CSCFs toward the I-CSCFs and periodical and/or on-demand pulls from I-CSCF. We use SIP extra-headers to convey the delay information.

In this IMS operational environment of M P-CSCFs, K I-CSCFs, and N S-CSCFs, three load balancing distribution algorithms for S-CSCF selection are considered: 1) *uniform random allocation (UA)* where the S-CSCF is chosen randomly between the available ones, 2) *round robin (RR)* where the next request is routed to the next available S-CSCF, and 3) *lowest response time with feedback (LRT-w-FB).* The first two are based on well known approaches. The third approach ("lowest response time with feedback") is our proposed approach to handle incoming requests by periodically

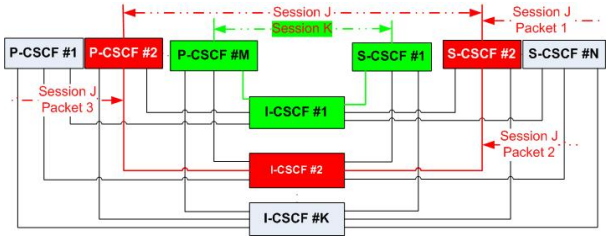

Fig. 2. IMS environment for study

considering the delay estimate using the delay update protocol. In this approach, every load balancer (I-CSCF) has a view of the current delays, Δ ($\Delta_{ps}, 1 \leq p \leq M, 1 \leq s \leq N$), between the P-CSCF and the S-CSCF; this is obtained through a periodically measured feedback system that is invoked using the delay update protocol. For the delay estimate, smoothing was used between the recent data and exponentially weighted moving average for the earlier data. Finally, we note that a request is treated in a stateful manner. That is, all packets arriving for a specific call request at a P-CSCF are routed via the same I-CSCF to the same S-CSCF.

III. Congestion Avoidance: Access Control Algorithms

Congestion avoidance mechanisms in an IMS network must work in concert with S-CSCF selection. While the goal of the S-CSCF selection is to reduce the response time (latency) for any request, this can become significantly high in an overloaded situation if some requests are not denied at the entry points (P-CSCFs). The issue is what information is of importance for the P-CSCFs to know in order for them to make this decision. A possible factor to use is the delay as this impacts the overall response time. On the other hand, the service time of different servers can impact this delay. Because of this, we use the ratio of the current response time to the minimum average response time in empty system. Note that in a network of P-CSCFs and S-CSCFs, the minimum average response time is the average over the minimum response time incurred between different P-CSCFs and S-CSCFs if the system is empty. The schemes we considered are as follows:

- Bang-bang control algorithm (BBC) X-Y: This algorithm has two bounds: X is the lower bound and Y is the upper bound. If the current value goes beyond the upper bound, we reject incoming sessions; if it goes under the lower bound, we start accepting again. Certainly, when the system starts afresh, we allow up to the upper bound. Here, X and Y are based on the response ratio between the current response time and the minimum average response time.
- Occupancy control algorithm (OCC) Z: This algorithm has the control point: Z. If the current value is under Z, we allow any session to be accepted. After crossing the value Z, we make it harder with a varying probability to allow new sessions; i.e., the more we go higher than Z, the tougher it is to admit a new session. For Z, we use the response ratio between the current response time and the minimum average response time.
- Call spacing algorithm (CS) W/T: This algorithm has two control parameters: W and T. If the current value is under W, we allow any sessions. After crossing the value W, we start a timer, T, at which point we do not allow any new sessions for the duration of this timer; after T expires, we admit a new request again. For W, we use the response ratio between the current response time and the minimum average response time. For T, a fixed value is used in this work. Since we use a fixed value in this study, we will refer to this algorithm simply as CS W.
- Level control algorithm (LC) L: This algorithm is based on the single parameter, L. If the current value is under L, we allow any sessions. After crossing the value L, we reject any sessions. For L, we use the response ratio between the current response time and the minimum average response time.

There is a subtle but important difference between BBC and LC. In the case of BBC, there is also a lower bound, i.e., no new request is accepted unless the response ratio goes below the lower bound X. The notion of the CS W algorithm is essentially based on the classical call gapping algorithm used in the telephone network [11], [23, Chapter 11], [33]; in our case, the response ratio control parameter is used to trigger the control while the timer provides the space between new session requests to be admitted when this response ratio control level is reached. The bang-bang control algorithm has been addressed in regard to SIP servers [12], [25] and is adapted in our framework.

IV. Results

We have developed a simulation model to study control algorithms in the presence of S-CSCF selection. In our simulation, the inbound requests to proxies are assumed to follow the Poisson process with a mean call arrival rate $\lambda_p, p = 1, ..., M$. Call arrival (SIP messages in the IMS scenario) is often modeled as a Poisson process; therefore, this is a reasonable assumption. We assume that P-CSCFs and S-CSCFs operate with exponentially distributed service times, while I-CSCF merely works as a forwarding function without incurring any delay compared to the P-CSCFs and S-CSCFs.

The input rates, λ_p, for different P-CSCFs are randomly chosen within a range. Each scenario in our simulation is completed for a million incoming requests to be routed to the N S-CSCFs. Since the packets related to a request are to be statefully treated, we keep track of the session (request) and route all the subsequent packets for this same session to the same S-CSCF via the same I-CSCF.

In all our studies, we have used a number of grid topologies by changing the number of P-CSCFs (M) and S-CSCFs (N), while keeping the actual number of P-CSCFs and S-CSCFs the same for each case. In this work, we will discuss results for 4×4 and 8×8 grid topologies. The simulation code was validated against an analytical model assuming uniform random allocation for stateless requests. As discussed in our previous work [31], this special case can be analytically modeled as a Jackson queue [19]; the results between the analytical model and the simulation model were found to differ by only 1% to 2%.

A. Temporal Behavior of the Access Control Schemes

Recall from Section III that the access control schemes are triggered based on the system response ratio observed at each P-CSCF at the time of arrival of a new request. To illustrate the temporal behavior, we limit here to four access control

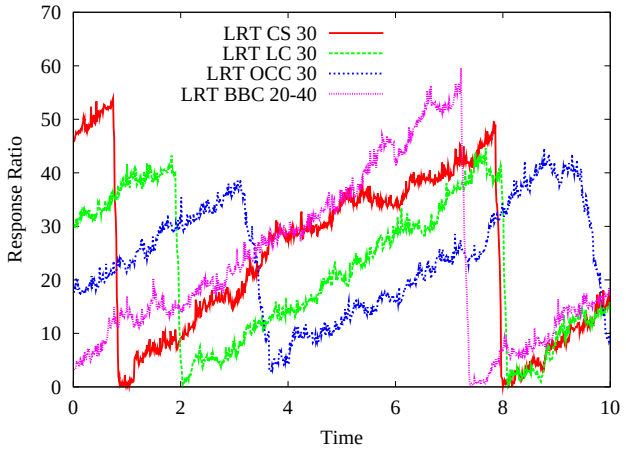

Fig. 3. Temporal behavior of access control schemes

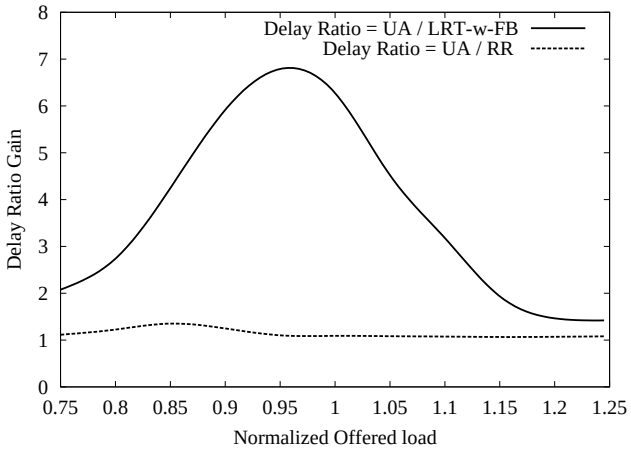

Fig. 4. Delay ratio gain comparison of S-CSCF selection algorithms (4x4 topology)

cases, CS 30, LC 30, OCC 30, and BBC 20-40, assuming all of them use the S-CSCF selection based on LRT-w-FB in an overloaded situation; the graph is shown in Fig. 3. Consider, for example, BBC 20-40; it may be noted that while the control is triggered at the response ratio of 40 (the upper bound) at time 5.5 sec, the system response continues to rise for a while (till around 7.3 sec) to serve the backlog of messages that needed to be processed for already admitted requests by each P-CSCF. This lag time cannot be completely avoided due to the distributed locations of S-CSCFs as much as the load balancing algorithms tries to minimize the response time. We do, however, note differences between different schemes. For example, the lag time for BBC 20-40 is less than that for CS 30 since for CS 30 the control is triggered at time 5 sec while the drop occurs at 8 sec. This, in turns, impacts on how many new requests can be admitted over a time horizon. To understand this better, in the following sections, we focus on the overall impact on average response and request rejects at different normalized offer load cases.

B. Picking an S-CSCF Selection algorithm

To focus on our study, we first considered all three algorithms for S-CSCF selection in order to identify which one to use with congestion avoidance. For this consideration, the network-wide normalized offered load in the 4×4 grid topology is varied from 0.75 to 1.25. It is known that with the increase in the normalized offered load, the response time increases. While typically the average delay is plotted against the offered load, we present a new way to look at the average delay since our interest is to *comparatively* determine how one scheme compares against another. We use the uniform allocation scheme as the baseline as we found this to have the highest average response, and then, we present the *Delay Ratio Gain* as the ratio of the average delay due to the uniform allocation scheme divided by the average delay due to the other S-CSCF selection algorithms. In Fig. 4, we present this delay ratio as gain while the normalized load is varied. As we can see, the ratio value remains close to one for the ratio

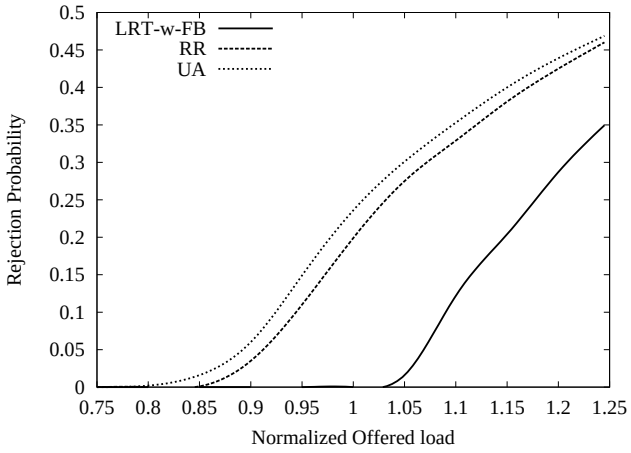

Fig. 5. Rejection probability of S-CSCF selection algorithms (4x4 topology)

between the round robin (RR) and the uniform allocation (UA) schemes—this means that their delays are comparable, irrespective of change in the normalized offered load. On the other hand, the least response time with feedback (LRT-w-FB) scheme results in a delay ratio gain of close to 7 when the normalized load is about 1.0; this means that its delay is about 1/7-th of the delay with the uniform allocation scheme. As the load increases further, this gain slows down; this means that when the normalized load is excessively high (say beyond 1.25), the delays observed by different schemes are about the same. Nevertheless, LRT-w-FB maintains a significant delay ratio gain compared to the round robin scheme in most of the operating region of interest to us.

Now consider the request rejection probability for the same range of normalized offered load (see Fig. 5). Naturally, as the normalized load reaches around 1.0, some new requests are rejected at P-CSCFs, which grow as the load increases. However, the rejection probability is much higher for the round robin and uniform allocation schemes, compared to the least response time with the feedback scheme. Therefore,

considering both rejection probability and the response time, we choose the least response time with feedback (LRT-w-FB) scheme for the S-CSCF selection in the remaining study on congestion avoidance, which is discussed in the next section.

C. Performance of Congestion Avoidance Schemes

Having settled on the S-CSCF selection algorithm to be the least response time with network feedback, we now focus on the access controls schemes for congestion avoidance. As we did above, instead of presenting the delay in graphs, we will present the delay ratio gain using UA as the baseline algorithm since this gives us a better perspective of gain when the normalized load is varied.

For our study, we considered both a 4x4 and an 8x8 grid topology. For the proposed control algorithms, we tried a number of different values for associated parameters for each control scheme. As can been seen with the access control schemes described in Section III, there are one or more parameters with each schemes. We studied many different values of parameters for each of the schemes on the grid topologies; for ease of discussion and due to the notable difference in performances, we limit to the following seven distinct cases: BBC 20-40, BBC 60-80, OCC 30, OCC 50, CS 50, LC 30, and LC 70.

First we consider the 4×4 topology. In Fig. 6, we presented the results of these seven control cases in terms of the delay ratio gain. This clearly shows that BBC 60-80 has the best gain almost throughout except when the normalized load goes beyond 1.15. Consider next the rejection problem for the same cases, shown in Fig. 7. We found that OCC 50 has the lowest rejection probability. On the other hand, this scheme does not have the best delay ratio gain.

In order to understand the balance between delay ratio gain and the rejection probability, it is clear that we need a metric that can capture both these factors. We developed the following metric for this purpose:

$$M(a) = \text{DelayRatioGain} + a * (1 - \text{RejectionProbability}), \quad (1)$$

where $a \geq 0$ is a weight parameter to weigh the importance of these two factors.

In Fig. 8 and Fig. 9, we present metriced M for $a = 0.5$ and $a = 2.0$, respectively, where the smaller value of a gives a low weight to acceptance probability and the higher value gives a high weight to acceptance probability. We found that in both instances BBC 60-80 gives the highest metric value without any significantly different observations among other cases.

We next considered the 8×8 topology. The delay ratio gain and rejection probability are presented in Fig. 10 and Fig. 11, respectively. What is noticeable is that the delay ratio gain is not as uniform with the 8×8 topology as was in the case of the 4×4 topology. While BBC 60-80 has the edge in general, in certain normalized load cases, OCC 50 has the edge over BBC 60-80. Regarding rejection probability, the loss does not start to show up until much higher normalized offered load values. This is mainly because of more routing paths between P-CSCFs and S-CSCFs in the 8×8 topology compared to the 4×4 topology. In terms of comparison of

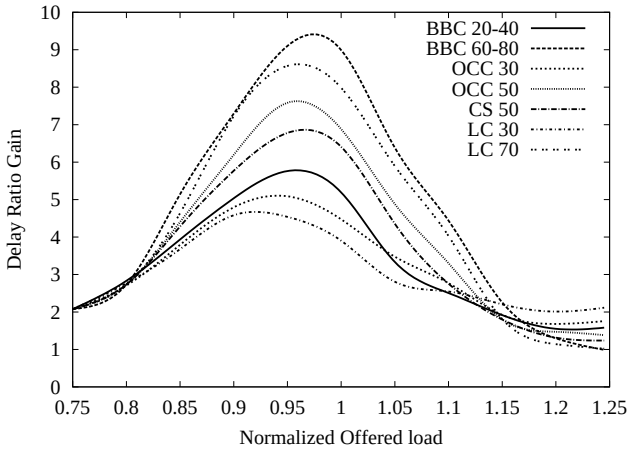

Fig. 6. Delay ratio gain for different control cases (4x4 topology)

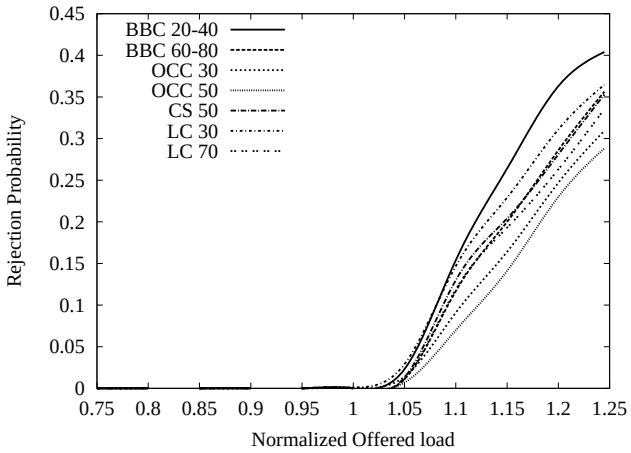

Fig. 7. Rejection probability for different control cases (4x4 topology)

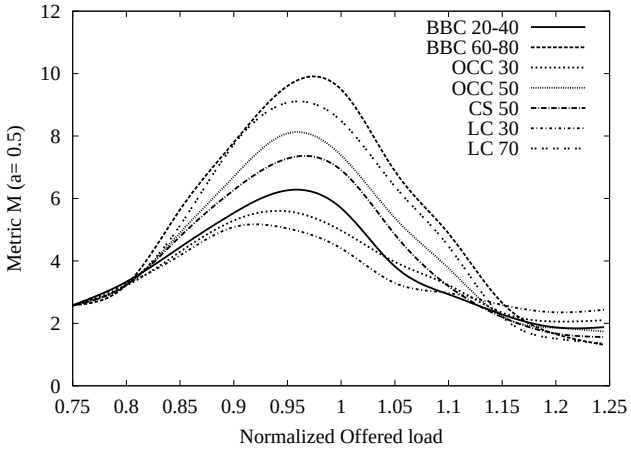

Fig. 8. Metric value $M(a)$ with $a = 0.5$ for different control cases (4x4 topology)

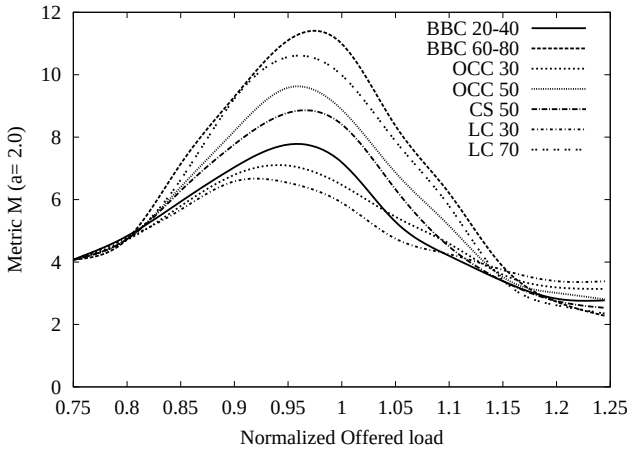

Fig. 9. Metric value $M(a)$ with $a = 2.0$ for different control cases (4x4 topology)

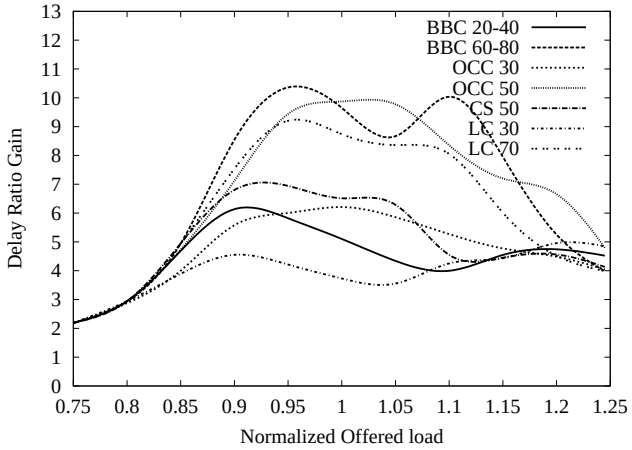

Fig. 10. Delay ratio gain for different control cases (8x8 topology)

the cases, OCC 50 results in the lowest rejection probability. We next considered the metric value M using $a = 5.0$ and $a = 20.0$, respectively (Fig. 12 and Fig. 12). We note that the metric M shows essentially the same behavior as the delay ratio gain while the difference lessens when we go from a smaller value of a to a higher value of a.

Finally, it may be noted that results for the delay ratio gain discussed so far were based on using the average delay values. In order to see whether there is any difference, we also computed the delay ratio gain by comparing the 95-th percentile of the delay; this is plotted for the 4×4 and the 8×8 topologies in Fig. 14 and Fig. 15, respectively. When we compare the delay ratio gain based on the 95th-percentile delay compared to the average delay, we do not notice much difference in the case of the 4×4 topology (refer to Fig. 6 and Fig. 14). On the other hand, for the 8×8 topology, we note that the OCC 50 does edge out BBC 60-80 over a broader range of the normalized offered load (refer to Fig. 10 and Fig. 15).

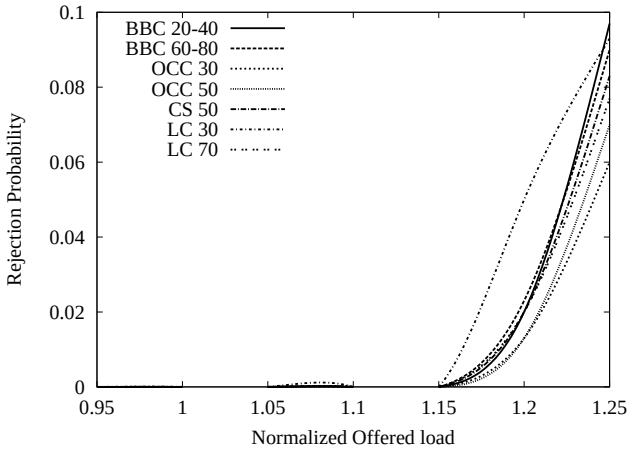

Fig. 11. Rejection probability for different control cases (8x8 topology)

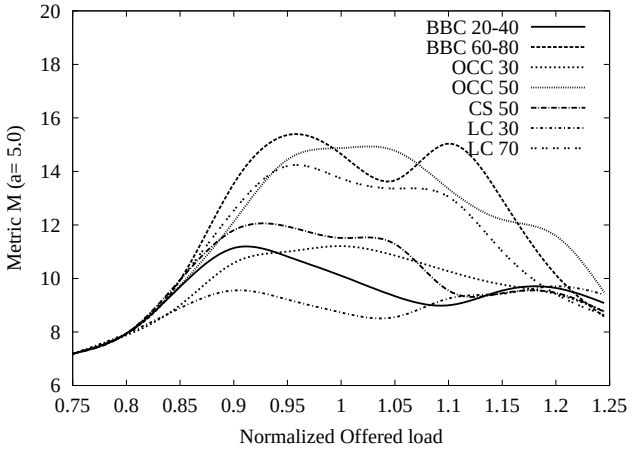

Fig. 12. Metric value $M(a)$ with $a = 5.0$ for different control cases (8x8 topology)

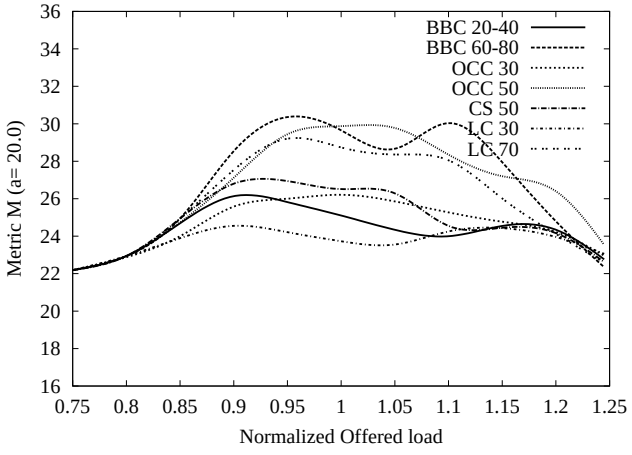

Fig. 13. Metric value $M(a)$ with $a = 20.0$ for different control cases (8x8 topology)

V. Related Work

An important focus of our work is when the message size is more than one that may be generated due to SIP signalling of

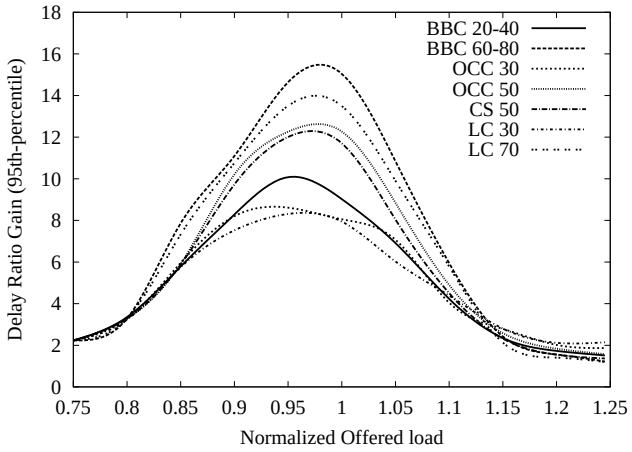

Fig. 14. 95th-percentile-based Delay ratio gain for different control cases (4x4 topology)

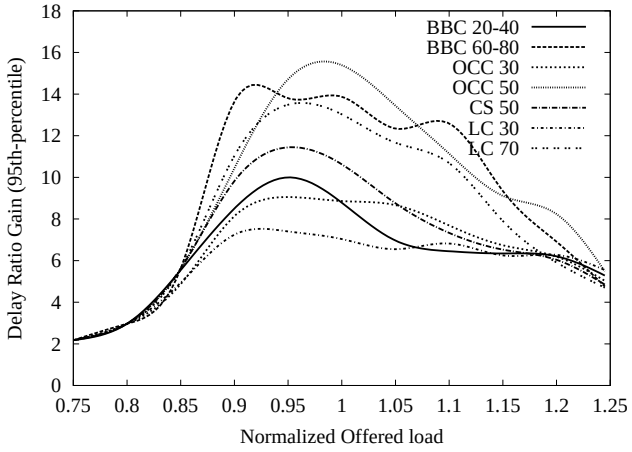

Fig. 15. 95th-percentile-based Delay ratio gain for different control cases (8x8 topology)

a request. Since TCP is used for transport for SIP signalling in an IMS environment, it seems to give the impression that TCP modeling could be applicable here. On the other hand, while there has been much work on TCP modeling, there are some important differences here. Most TCP modeling studies consider a dumbbell topology to understand active queue management, i.e., a number of connections are tunneled through just a single congested link (see, for example, [5], [9], [10], [15], [21], [24], [35]). Note that load distribution through nodes in the IMS network results in request routing that is not possible in a dumbbell topology with a single congested link. Furthermore, due to the importance of the real-time requirement of the signalling responses, this part of the IMS network is provided through a private IP network that is engineered so that this does not become the dominant factor in latency in the SIP signal communication for a request in the IMS network (as opposed to the notion of congested link in most TCP modeling study).

For congestion control and avoidance in the Internet, a number of mechanisms have been developed such as TCP slow start at the hosts and active queue management at routers; for example, see [5], [10], [16], [23, Chapter 22]. It may be noted that the basic notion remains that all TCP requests are accepted at the entry points (hosts), while the TCP flows are regulated. On the other hand, the situation considered here for the IMS network requires balancing between request denial and response time of accepted requests.

For the telephone network, a number of features such as code control and dynamic overload control have been implemented that prevent new call requests from being accepted at the entry points in an overloaded situation; for example, see [11], [23, Chapter 11], [26], [33]. For SS7/Intelligent networks there have been several studies on congestion control [2], [3], [17], [18], [22], [29], [30]. In a PSTN-SS7 scenario, the call rejections in PSTN and the overload due to signaling impact in the SS7 are considered independently. An important difference here is that in an overloaded situation in an IMS network, we need to balance request rejection while minimizing the response time through load balanced CSCF servers for signaling requests that are set up using TCP connections.

For the IMS architecture, there has been work on the scheduler and improvement of the IMS presence service [1], [34] and on an application server workload [28]. However, these works are orthogonal to our work. In a somewhat related work, SIP server overload has been studied [12], [13], [14], [25]; however, this is not directly applicable to the IMS environment where multiple CSCFs are present.

VI. Summary and Future Work

An operational IMS network can have a number of P-CSCFs, I-CSCFs, and S-CSCFs to entertain request arrivals. In this work, we consider the problem of overload in traffic with S-CSCF selection when a request weaves through the network of P-CSCFs and S-CSCFs in an IMS network while some requests are dropped from entry due to overload. We discussed the basics of this environment and identified where and why this is different than congestion avoidance approaches used in the Internet or the telephone network. A fundamental difference is that congestion avoidance requires a good balance between rejecting requests at the entry point and maintaining good response times from the S-CSCFs for the ones that are admitted. To our knowledge, this problem for the IMS network has not been addressed before.

We presented a number of access control schemes for congestion avoidance to work with S-CSCF selection in an IMS network. In order to do that, we briefly presented three S-CSCF allocation algorithms: uniform random allocation, round-robin, and least response time with feedback. We then discussed how they are used in conjunction with access control schemes for congestion avoidance. For our main study, we used the least response time with feedback with the S-CSCF selection. Due to the parameters involved with the access control schemes, a number of cases arise; we limit our study to seven distinct cases out of many possible ones.

Our study then involved considering these scenarios on 4×4 and 8×8 topologies. We varied the network-wide normalized offered load to consider both the non-congested region and congested region where new requests needed to be dropped at the entry points. From our analysis, we found that the bang-bang control scheme with a lower bound set to 60 and an upper bound set to 80 (BBC 60-80) is the best performing approach in most cases, for balancing between request denial and the response time. We also presented a metric that combines these two factors. A difference between two different topologies is that in the larger topology, there are more paths between P-CSCFs and S-CSCFs, resulting in a lower request denial probability at the higher offered load.

Our future work will consider topologies other than grid topologies and also try to develop additional composite metrics that are helpful in balancing between request denial and response time with additional information such as the 95%-percentile of the response time. Furthermore, we plan to consider developing an adaptive approach that can use a combination of the schemes discussed here that can be triggered as appropriate based on the network situation.

REFERENCES

[1] M. T. Alam, "Design and analysis for the 3g ip multimedia subsytem," Ph.D. dissertation, Bond University, Australia, 2007.

[2] A. Arvidsson, L. Angelin, and S. Pettersson, "Profit optimal congestion control in intelligent networks," in *Proc. of 15th International Teletraffic Congress (ITC15)*, 1997, pp. 911–920, V. Ramaswami and P. E. Wirth (eds.).

[3] Y. Bakshi, A. H. Diaz, K. Meier-Hellstern, R. A. Milito, and R. Skoog, "Overload control in a distributed system," in *Proc. of 15th International Teletraffic Congress (ITC15)*, 1997, pp. 571–582, V. Ramaswami and P. E. Wirth (eds.).

[4] T. Bourke, *Server Load Balancing*. O'Reilly, 2001.

[5] B. Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang, "Recommendations on queue management and congestion avoidance in the internet," *Internet Engineering Task Force – RFC 2309*, April 1998.

[6] G. Camarillo and M. A. García-Martín, *The 3G IP Multimedia Subsystem (IMS), 2nd Edition*. John Wiley & Sons, 2006.

[7] M. Castro, M. Dwyer, and M. Rumsewicz, "Load balancing and control for distributed world wide web servers," in *Proceedings of the 1999 IEEE International Conference on Control Applications*, August 1999, pp. 1614–1619.

[8] B. Chattopadhayay and M. A. Muñoz de la Torre, "S-CSCF load balancing," Motorola, Inc., April 17, 2006, http://www.priorartdatabase.com/IPCOM/000136381/.

[9] W. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, "Stochastic fair blue: A queue management algorithm for enforcing fairness," in *Proc. IEEE INFOCOM'2001*, Anchorage, AK, 2001, pp. 1520–1529.

[10] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. on Networking*, vol. 1, pp. 397–413, August 1993.

[11] D. Haenschke, D. A. Kettler, and E. Oberer, "Network management and congestion in the U.S. telecommunications network," *IEEE Trans. on Communications*, vol. COM-29, pp. 376–385, 1981.

[12] V. Hilt and I. Widjaja, "Controlling overload in networks of SIP servers," in *Proc. of IEEE ICNP'2008*.

[13] V. Hilt, I. Widjaja, and H. Schulzrinne, "Session initiation protocol (sip) overload control," March 7, 2009. http://www.ietf.org/internet-drafts/draft-hilt-sipping-overload-06.txt

[14] V. Hilt (Ed.), "Design considerations for session initiation protocol (SIP) overload control," March 7, 2009. http://www.ietf.org/internet-drafts/draft-ietf-sipping-overload-design-01.txt

[15] C. V. Hollot, Y. Liu, V. Misra, and D. Towsley, "Unresponsive flows and AQM performance," in *Proc. IEEE INFOCOM'2003*, San Francisco, CA, 2003.

[16] V. Jacobson, "Congestion avoidance and control," *Computer Communication Review*, vol. 18, no. 4, pp. 314–329, August 1988.

[17] B. Jennings, A. Arvidsson, and T. Curran, "A token-based strategy for co-ordinated, profit-optimal control of multiple IN resources," in *Proc. of 17th International Teletraffic Congress (ITC17)*, 2001, pp. 245–258.

[18] S. Kasera, J. Pinheiro, C. Loader, M. Karaul, A. Hari, and T. LaPorta, "Fast and robust signaling overload control," in *Proc. of IEEE International Conference on Network Protocols (ICNP 2001)*, November 2001.

[19] L. Kleinrock, *Queuing Systems, Vol. 1: Theory*. Wiley-Interscience, 1975.

[20] C. Kopparapu, *Load Balancing Servers, Firewalls and Caches*. John Wiley & Sons, 2002.

[21] D. Lin and R. Morris, "Dynamics of random early detection," in *Proc. ACM SIGCOMM'97*, Cannes, France, September 1997, pp. 127–137.

[22] D. R. Manfield, G. K. Millsteed, and M. Zukerman, "Performance analysis of SS7 congestion controls under sustained overload," *IEEE Journal on Selected Aread in Communications*, vol. 12, pp. 405–414, 1994.

[23] D. Medhi and K. Ramasamy, *Network Routing: Algorithms, Protocols, and Architectures*. Morgan Kaufmann Publishers, 2007.

[24] V. Misra, W.-B. Gong, and D. Towsley, "Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED," in *Proc. ACM SIGCOMM'2000*, Stockholm, Sweden, August-September 2000, pp. 151–160.

[25] M. Ohta, "Overload control in a SIP signaling network," *Enformatika Transactions in Engineering, Computing and Technology*, March 2006.

[26] R. R. Pillai, "A distributed overload control algorithm for delay-bounded call setup," *IEEE/ACM Transactions on Networking*, vol. 9, no. 6, pp. 780–789, 2001.

[27] M. Poikselka, A. Niemi, H. Khartabil, and G. Mayer, *The IMS: IP Multimedia Concepts and Services*. John Wiley, 2006.

[28] N. Rajagopal and M. Devetsikiotis, "Modeling and optimization for the design of IMS networks," in *Proc. of 39th Annual Simulation Symposium, 2006 (ANSS'06)*, April 2006, 7 pages.

[29] M. P. Rumsewicz, "Analysis of the effects of SS7 message discard schemes on call completion rates during overload," *IEEE/ACM Trans. on Networking*, vol. 1, pp. 491–502, 1993.

[30] M. P. Rumsewicz and D. E. Smith, "A comparison of SS7 congestion control options during mass call-in situations," *IEEE/ACM Trans. on Networking*, vol. 3, pp. 1–9, 1995.

[31] P. Tirana and D. Medhi, "The effects of load distribution algorithms in application's response time in the IMS architecture," in *Proc. of 18th ITC Specialist Seminar on Quality of Experience*, pp. 173–181, Karlskrona, Sweden, May 2008.

[32] P. Tirana and D. Medhi, "Distributed approaches to S-CSCF selection in an IMS network," in *Proc. of 2010 IEEE Network Operations and Management Symposium (NOMS 2010)*, pp. 333–340, Osaka, Japan, April 2010.

[33] D. M. Tow, "Network management—recent advances and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 732–741, 1988.

[34] C. Urrutia-Valdés, A. Mukhopadhyay, and M. El-Sayed, "Presence and availability with IMS: Applications architecture, traffic analysis, and capacity impacts," *Bell Labs Technical Journal*, vol. 10, no. 4, pp. 101–107, 2006.

[35] L. Zhang, "A new architecture for packet switching network protocols," Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, Tech. Rep. MIT-LCS-TR-455, 1989.