

# Optimizing Request Denial and Latency in an Agent-Based VPN Architecture

Haiyang Qian<sup>1</sup>, Steve Dispensa<sup>2</sup>, Deep Medhi<sup>1</sup>

<sup>1</sup>University of Missouri–Kansas City, USA

<sup>2</sup>Positive Networks, Inc., USA

**Abstract**—Agent-based virtual private networks architecture (ABVA) refers to the environment where a third-party provider runs and administers remote access VPN service for organizations that do not want to maintain their own in-house VPN servers. This environment presents interesting management challenges for an ABVA provider. In this paper, we consider the problem of connecting users of an organization to an optimal VPN server location so that request denial probability and latency are balanced. Because of the bandwidth requirement of a user to be met when connected to a server, this system has the appearance of a standard loss system. However, due to latency perceived by a user from its current location to a VPN server and to allow for servers to be located in a distributed fashion, this environment is not a pure loss system. By considering a finite population, this environment can be approximately represented using the Engset model; however, this does not address the latency issue either. We present a number of strategies regarding which VPN server is to be selected and the number of attempts to be tried so that request denial probability is minimized without unduly affecting latency. Through computational results, we show that the clustering with directional hunting (CDH) strategy provides the best result. However, in the heterogeneous case of users with differing data rates (“traffic classes”), request denial observed by each class is different leading to unfair treatment. We have proposed a reserved capacity based add-on feature with CDH, which allows service classes with different data rates to be treated fairly.

## I. INTRODUCTION

Virtual Private Network (VPN) services have seen significant growth over the years. There are many forms of VPN services. A common one is remote-access VPN in which an employee of an organization can use the VPN service to access the intranet of the organization from an off-location through the public Internet by establishing a VPN tunnel. While such VPN services are often deployed with each organization installing a VPN server (“gateway”) on its premise, this model does not serve well for organizations that do not often have the IT expertise to install and maintain VPN servers on a day-to-day basis. In fact, many organizations would want to have a hassle-free VPN service for their users if it can be provided and maintained by a third party and the service is cost-effective. Such a service is also appealing if it is flexible and adaptable to changing organizational and technological requirements. An *Agent-Based VPN Architecture* (ABVA) is a third-party approach that fits into a need. In this approach, one or more VPN agents (or brokers) are located outside the organizations that serve as *Rendezvous Points* (RP), where users and their associated organizations meet; the

Rendezvous Point (RP) service is provided by a third party (ABVA provider) for multiple organizations while maintaining separate virtual tunnels for each organization. Currently, an agent-based VPN service is deployed by Positive Networks [13]. We will interchangeably use the terms, Rendezvous points and VPN servers (or VPN server locations).

From the perspective of an ABVA provider, there are challenging design and management problems such as: 1) customers wanting a certain quality-of-service as part of service-level agreements (SLAs) such as a user’s connectivity latency, bandwidth guarantee, and acceptable rate of connectivity, and 2) determining the location of such servers if the users are geographically distributed around the globe. In this work, we focus on the first issue. In an earlier work, the overall latency minimization problem was formulated while providing bandwidth guarantee when the location and the number of the users are known and fixed [10]; through this formulation, lower bounds on the overall system latency can be obtained and it was observed that bandwidth guarantee was the dominant factor in a highly loaded environment, while connecting to the closest RP is important in a lightly to moderately loaded environment.

In this work, we relax the requirement that the user arrival is static. When the arrival of users seeking this service is stochastic, *and* bandwidth guarantee during the session and tolerable acceptance rate of service are the primary factors, then this problem looks similar to a loss system. On the other hand, the geographic diversity of server locations as well as the latency perception of users make this problem fundamentally different than a pure loss system. This is further complicated by the fact that a user might travel from one city to another around the globe and logon; thus, the proximity of a server for any user is not known easily. For instance, if a user is connected to a far-off VPN server, then it might be able to provide the bandwidth guarantee while the latency perceived is beyond the SLA-specified value. On the other hand, a user might locate a nearby VPN server, but this RP might be already heavily loaded and not able to accommodate another user, thus denying this user any connectivity. Therefore, a service management problem for an ABVA provider is to trade-off between latency and request denial. Since the software client loaded to a user’s device (such as a laptop computer) is customizable, distributed protocols can be used to push additional information, which can be cached based on previous logon location and so on.

By considering the systems management problem under stochastic arrival of users, we address trade-offs between latency and bandwidth guarantee. In particular, we present analytical approximations that use a hunting scheme to locate a VPN server; through simulation, we show the results on trade-off between latency and acceptance rate for a variety of load conditions. We also briefly discuss how parameters can be tuned for the heterogeneous case in order to maintain overall efficiency and fairness in a distributed environment.

It should be noted that there is a direct relation between the ABVA service and overlay networks since the RPs in an ABVA environment form an overlay network. There has been significant research on overlay networks of various types in recent years (for example, see [1], [9]); however, the ABVA poses its unique challenges. Though there are many works on various aspects about VPN (for example, see [2], [3], [4], [12], [15]), they primarily address the core design of overlay networks. However, to our knowledge, how to efficiently provide user connectivity and services in an agent-based VPN architecture over IP-based networks remains an important problem that has not received attention in the literature.

The remainder of the paper is organized as follows: in Section II, an overview of the ABVA framework is presented. In Section III, the system model and a number of strategies are presented. In Section IV, numerical results are presented and studied. Concluding remarks along with future work are presented in Section V.

## II. OVERVIEW OF ABVA FRAMEWORK

A high-level conceptual view of the ABVA framework with users, Rendezvous Points, and organizations is presented in Fig. 1. The Rendezvous points are marked as RP1, RP2, RP3, and RP4. Here, user U1 is associated with organization ORG1, and user U2 with organization ORG2. The Internet serves as the native network that has IP routers identified as R1, R2, and so on. When user U1 wants to connect to organization ORG1 for remote-access VPN services, it is actually connected to Rendezvous Point RP1, which, in turn, connects to organization ORG1; similarly, for user U2 connecting to organization ORG2. The underlying path from the user to the Rendezvous point is over the Internet; thus, a user can access it from anywhere in the Internet. Certainly, the Rendezvous Points must have presence in the Internet.

The path from a Rendezvous point to an organization can be over either the public Internet or it can be over a private network; for ease of illustration, the figure shows only the former case. Usually, this path is provided based on service level agreements between the ABVA provider and the native Internet service provider, by specifying acceptable quality of service and bandwidth guarantees. For instance, the path from an RP to an organization can be set up as an MPLS tunnel that can provide such guarantees; or, private dedicated circuits may be used between an RP and an organization.

With the above background, we are now ready to provide further clarity to the latency and bandwidth denial aspects of

the problem. Based on an arriving user's request, latency refers to latency for the segment between the user and the RP it is connected to; the part from an RP to the organization is bounded by the SLA and is, therefore, not required to include in this consideration. Certainly, this part must be acceptable to begin with. The latency from a user to the RP is important to consider for two primary reasons: 1) this is entirely over the public Internet, 2) once connected, during the entire duration of the session, this latency (with some acceptable fluctuation) is perceived by the user; for example, when downloading a file or accessing email on the intranet. It should be noted that in general, a user location is not fixed; a user may move from a particular location to another overnight and access from a different location (e.g., hotel instead of home), and thus, a particular user should not always be tied to a specific RP. Since the bandwidth guarantee is to be ensured on the path from an RP to the organization, this means that if a specific RP does not have the ability to handle a request, then it will be denied. That is, bandwidth guarantee is not factored in directly in the public Internet part (i.e., from the user location to an RP); however, a user (and the organization it belongs to), if accessing from a particular access technology, might have certain expectations. For instance, if the access technology is DSL, the user would expect/perceive the bandwidth to be at a certain rate, as opposed to a dial-up connection. Based on the above discussion as well as agent-based VPN service offering of [13], it is reasonable to consider the system model by considering latency from a user to an RP (VPN server) while bandwidth guarantee at the entrance of the RP to provider service. Thus, starting in the next section, we focus on this system modeling framework.

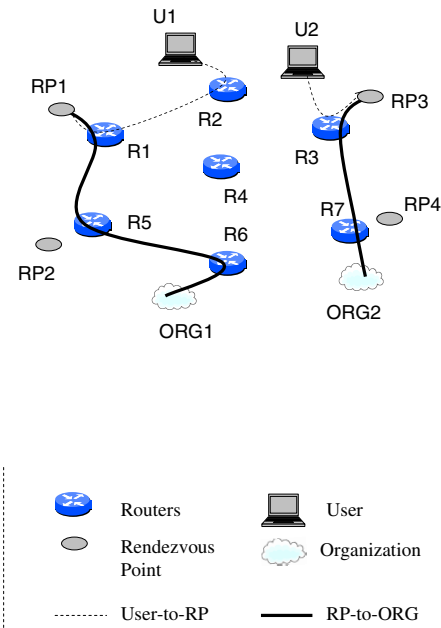


Fig. 1. ABVA: Basic conceptual framework

### III. SYSTEM MODEL AND STRATEGIES

Consider an ABVA provider with  $N$  Rendezvous points and  $U$  users. Assume that each RP has capacity  $C_n$ ,  $n = 1, 2, \dots, N$  for providing bandwidth guarantee to users; for simplicity, we assume that all servers have the same capacity, i.e.,  $C_n = C$ . Without loss of generality, we assume also that all users belong to a single organization. Note that in the multiple organization case, the capacity at RPs can be partitioned for each organization; thus, each organization can be managed independently, which is preferable since each can have separate SLA with the ABVA provider. In other words, it is sufficient to consider a single organization for our study. A user connection request is accepted if the VPN server attempted has the capacity to accommodate this connection; otherwise, it is a candidate for rejection. We do not rule out the possibility that a user may try one or more servers if it does not go through at the first one. This will be further discussed later on.

We assume that the connection request arrival process of each user is Poisson with rate  $\lambda$  and that the average duration of a session is assumed to be exponentially distributed with unit mean. In general, each user's arrival request has a different data rate requirement for service from the system. For simplicity, we consider two cases: 1) all users have the same data rate requirement (which is set to unity), 2) users may avail a data rate from a set of allowable, but differing data rates for bandwidth guarantee at an RP; the former will be referred to as the *homogeneous* case and the latter as the *heterogeneous* case.

We consider first the homogenous case. If all the VPN servers were located at a single RP, then the capacity would be  $N \cdot C$ . In this case, there is no possibility to optimize latency for any user. On the other hand, the request denial probability can be analytically calculated since it can be thought of as an Engset system ( $M/M/m/m/K$ ) for which the request denial probability,  $B$ , is given by [5], [7]

$$B(\lambda, U, NC) = \frac{\binom{U-1}{NC} \cdot \lambda^{NC}}{\sum_{\ell=0}^{NC} \binom{U-1}{\ell} \cdot \lambda^{\ell}}. \quad (1)$$

In this case, we denote the average latency for a user by  $T(\lambda, U, NC)$ . The case of all server capacity being located at a single RP will be referred to as the centralized-all scheme (CAS).

Now consider the case where only one RP is associated with a VPN server with capacity  $C$ . If any user now attempts this RP, then the request denial probability is given by

$$B(\lambda, U, C) = \frac{\binom{U-1}{C} \cdot \lambda^C}{\sum_{\ell=0}^C \binom{U-1}{\ell} \cdot \lambda^{\ell}}. \quad (2)$$

Clearly,

$$B(\lambda, U, NC) < B(\lambda, U, C) \quad \text{for } N > 1. \quad (3)$$

That is, the request denial probability with (2) will be higher than (1).

Since the client software associated with a user's computer is programmable, a distributed caching concept can be used. For instance, if a subset of users is likely to stay within a region, then they can be assigned to the RP in this region. Thus, we can consider a cluster-based approach in which each user is associated with a primary RP. For simplicity, we assume that all users are uniformly divided to each RP—thus, each cluster group associated with a RP will then have  $U/N (= M)$  users. That is, for the cluster based approach, the population is  $M$  users, which is served by a server of capacity  $C$ . Thus, in this case, the request denial probability is given by

$$B(\lambda, M, C) = \frac{\binom{M-1}{C} \cdot \lambda^C}{\sum_{\ell=0}^C \binom{M-1}{\ell} \cdot \lambda^{\ell}}. \quad (4)$$

It is easy to see that the CAS scheme has lower blocking than the above, i.e.,

$$B(\lambda, NM, NC) < B(\lambda, M, C) \quad \text{for } N > 1. \quad (5)$$

However, in this case, due to clustering, the average latency,  $T(\lambda, M, C)$ , is likely to be lower than the average latency,  $T(\lambda, NM, NC)$ , compared to the case in which all VPN servers are centralized at a single RP. Thus, the cluster option presents an important trade-off between the request denial probability and the average latency.

Expanding on the clustering notion, and in order to reduce request denial probability, a hunting feature can be added. This means that if a user's request is denied by the VPN server to which it currently belongs to, then the user's request is automatically forwarded to a VPN server at another RP. In order to do that, the user's software client would need to store the location information of possible RP locations; out of these, one RP is randomly chosen and attempted. Note that this introduces additional delay in set up due to this hunting feature. For modeling approximation, we assume that this is instantaneous and retry time is negligible. For brevity, this hunting strategy for a second RP randomly beyond its cluster will be denoted CRH-2. In this case, blocked users,  $\widehat{M}$ , from its primary server would randomly try one out of the remaining  $(N-1)$  servers; therefore, they will be equally distributed as  $\widehat{M}/(N-1)$ . From a receiving point of view, a server will receive such equally distributed redirects from the rest  $(N-1)$  servers; therefore, the net effect is that each server will receive  $\widehat{M}$  such secondary users. This entire situation can be simply thought of as a two-cluster environment where primary users blocked from its cluster ("first cluster") are directed to the second cluster, and reciprocally, the primary users blocked from the second server are directed to the first server. In an approximate sense, there are then  $2M$  users vying for  $2C$  units of capacity except for this ordering and reciprocity. Therefore, in an approximate sense, the request denial probability for CRH-2 is expected to be close to  $B(\lambda, 2M, 2C)$ . Therefore, going from clustering with no hunting to clustering with hunting for a second server randomly decreases the request denial probability as is evident from (5), while upwardly impacting latency. Later, we shall discuss this approximation with simulation results.



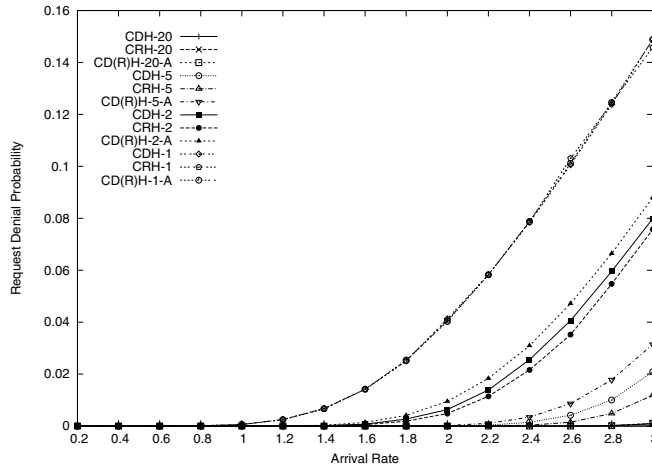


Fig. 3. Comparison of request denial probability between analytic modeling and simulation (“CD(R)H-k A” stands for analytical with  $k$  steps)

the same), clustering with randomized hunting for second RP (CRH-2), clustering with randomized hunting for multiple RPs (CRH- $k$ ), clustering with directional hunting for second RP (CDH-2), and clustering with directional hunting for multiple RPs (CDH- $k$ ). Note that CDH-20 (CRH-20) means that all RPs have been attempted.

First, we comment on the request denial probability between the analytical approximation and simulation; this is shown in Fig. 3. From this figure, we can see that the analytical model,  $B(\lambda, M, C)$ , is accurate for CDH-1/CRH-1 and  $B(\lambda, 20M, 20C)$  is accurate for CDH-20/CRH-20 (i.e., the two extreme cases). With regard to CDH- $k$  and CRH- $k$  when  $k = 2, 5$ , we note that the analytical approximation,  $B(\lambda, kM, kC)$ , overestimates request denial probability compared to simulation while the gap is marginal for smaller values of  $\lambda$ ; thus, the analytical approximation serves as an upper bound for these two schemes. Furthermore, from simulation, we can see that the request denial probability for CRH- $k$  is consistently smaller than that for CDH- $k$  for  $k = 2, 5$ .

In Fig. 4, we present simulation results as load is varied with regard to the performance measures, request denial probability and average latency. Furthermore, for comparison, we also report the average number of RPs (“mean number of steps”) visited; note that the mean number of steps includes counting the original cluster server a user has to visit first. With regard to request denial probability, we note that clustering strategy that is limited to just attempting its cluster VPN server and no additional ones (i.e., CDH-1 and CRH-1) results in the worst performance, which is not surprising as reflected through (5). Note that CDH-1 and CRH-1 are fundamentally the same. Not surprisingly, when all RPs are checked, there is no difference between randomized hunting and directional hunting (compare CRH-20 and CDH-20) as far as request denial probability is concerned. If we now consider limiting the number of RPs to be attempted, then we observe that the request denial probability is lower for clustering with randomized hunting compared to clustering with directional hunting, especially at a higher traffic load. This can be

observed by comparing request denial probability between CRH-5 and CDH-5. On the other hand, when we compare average latency, we note that CDH-5 performs significantly better than CRH-5 as the load increases. However, the trend for the mean number of steps is not similar to that of latency. We note that the mean number of steps is higher with CDH-20 (CDH-5) compared to CRH-20 (CRH-5), but when  $k = 2$ , the mean number of steps with CRH-2 is higher than that of CDH-2. This variation in the number of steps appears to rather depend on the stochastic freedom, not necessarily on CRH or CDH.

It is, however, important to note that CRH-1 (CDH-1) has the least latency; this is, however, misleading since this comes at the price of a very high request denial probability. In order to compare the request denial probability and the average latency together, we also calculated a composite score as shown below:

$$F = \eta * \text{request denial probability} + \text{latency} \quad (6)$$

where  $\eta$  is the weight given to balance the two factors. In Fig. 5, we plotted the composite score for four different values of  $\eta$  ( $\eta = 100, 1000, 5000, 10000$ ). Note that the low value of  $\eta$  means that the dominant factor is latency while the very high value of  $\eta$  means that the dominant factor is the request denial probability. From the composite metric viewpoint, we also observe that CDH is better than CRH. While CDH-20 gives the best performance as load is increased, it is worth noting that CDH-5 gives an excellent performance and it is very close to the result of CDH-20. This means that in practice, it is not necessary to attempt all servers; limiting the number of attempts to a small number, say, five, through directional hunting is sufficient to give almost optimal result. Furthermore, from the results, we can conclude that the weight factor set to either 1000 or 5000 gives a very good understanding of the balance between the two factors.

### B. Heterogeneous Case

Next, we consider the heterogeneous case. For simplicity, we consider only two data rate requirements (“traffic classes”):

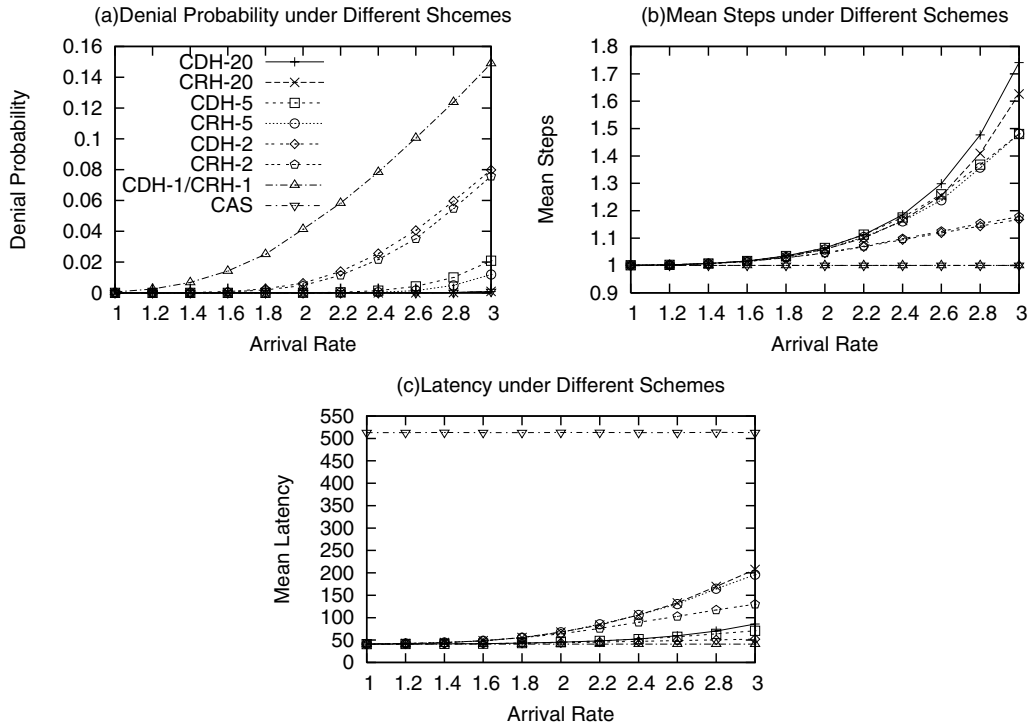


Fig. 4. Performance of Strategies (legends shown only with the first plot; same applies to others)

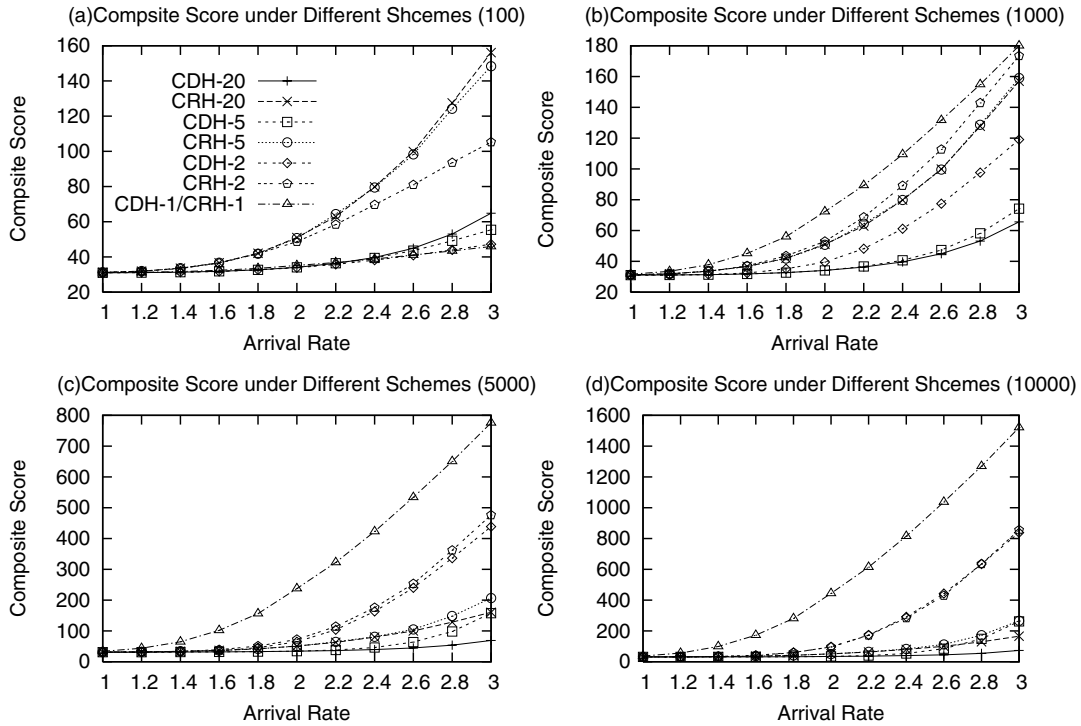


Fig. 5. Composite scores of different strategies for  $\eta = 100, 1000, 5000, 10000$  (legends shown only with the first plot; same applies to others)

we label them as 'low data rate' and 'high data rate' traffic classes. Specifically, the low data rate traffic class has a unit bandwidth requirement (same as the homogeneous case), while the high data rate class has a bandwidth requirement of 2 units per request. Thus, for a server with 20 units of capacity, the effective capacity for the high data rate traffic class is 10. For

this study, we assume that 10% of the users are high data rate users. Since CDH was found to be better than CRH for the homogeneous case, we chose to concentrate our study for the heterogeneous case only for the CDH strategy.

As stated earlier, performance evaluation in terms of request denial probability for a loss system with heterogeneous

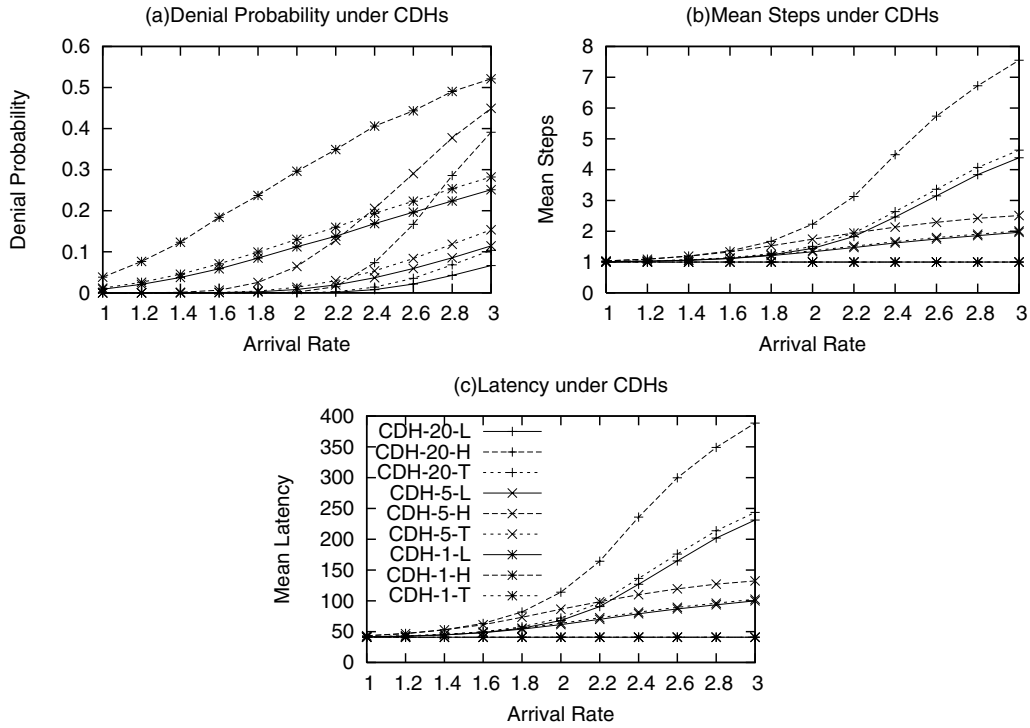


Fig. 6. Performance of CDH strategies for the heterogeneous case (legends shown only with the last plot; same applies to others. L refers to low data rate, H refers to high data rate; T refers to combined performance)

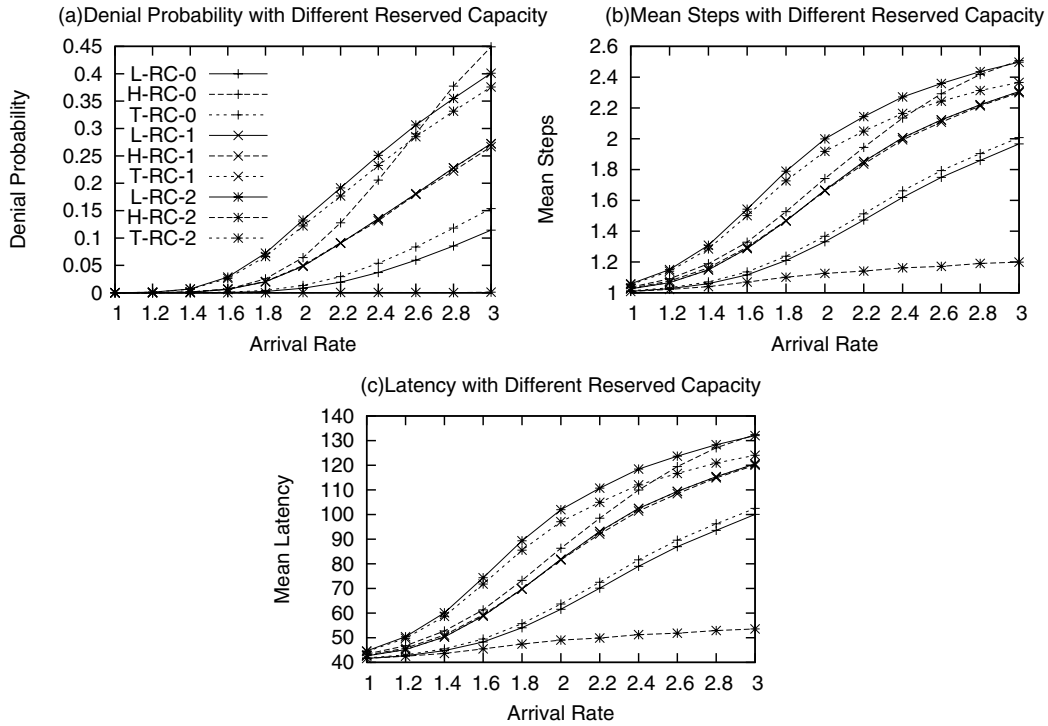


Fig. 7. Performance of CDH-5 for the heterogeneous case in the presence of the reserved capacity feature (legends shown only with the first plot; same applies to others. L refers to low data rate, H refers to high data rate; T refers to combined performance)

traffic can be obtained by extending Kaufman-Robert formula. It is also known that in a heterogeneous environment, traffic classes are treated unfairly; in particular, the high data rate class faces higher request denial probability than the low data rate class in a completely shared environment [8], [11].

Through our simulation, we have found that this unfairness in request denial probability continues to persist in a distributed ABVA environment regardless of the number of tries in the CDH strategy (see Fig. 6); the difference is found to be more pronounced for CDH-1 and becomes less pronounced

for CDH-20. On the other hand, we observe the opposite effect with regard to latency; the difference in latency is more pronounced for CDH-20 compared to CDH-1. This reverse effect can be understood by following the results for the mean number of steps; a high data rate service, on average, requires more steps than a low data rate service. Since the data rate requirement is higher, this class is not able to find the nearest neighboring clusters compared to a low data rate service—this is partly due to the effect of the unfairness phenomenon of request denial, which, in turn, affects the mean number of steps, thereby increasing latency. In this regard, a good balance on both request denial and latency is observed for CDH-5, although unfairness still persists.

This unfairness behavior is not acceptable from the customer point of view since all users, regardless of the rates, belong to the same organization. Thus, it is important to be able to provide equal service perception to all users, regardless of the data rate. We can achieve this by employing a reserved capacity (RC) scheme at the VPN servers; this scheme is an extension of a scheme for the multi-rate loss system discussed in [11]. Briefly, this scheme says that if the available capacity of a server at an RP drops below a certain threshold at a certain instant, then only the high data rate service is accepted to use the reserved capacity (provided there are at least two units of reserved capacity, as required by an arrival of a high-data rate service). In Fig. 7, we plotted the CDH-5 strategy with three reserved capacity scenarios: RC-0 (as in the earlier case), RC-1 (reserve equivalent of one service unit of capacity for high data rate service), and RC-2 (reserve equivalent of two service units of capacity for high data rate service). We can see that CDH-5 with RC-1 results in almost equal treatment for both service classes. Thus, such a reserved capacity feature can be added to the clustering with directional hunting strategy for the heterogeneous case. Certainly, the exact amount of reserved capacity to be assigned depends on the number of users, load, capacity of the server at each RP location, and data rates.

## V. CONCLUSION AND FUTURE WORK

In this paper, we consider the offering of an agent-based VPN service where organizations and their users utilize Rendezvous Points for remote-access VPN service. In particular, we consider the systems management problem of an ABVA provider to balance request denial probability and latency perceived by users. To consider the case of VPN servers being located at multiple Rendezvous Points, we have identified a number of strategies for server selection, which includes a concept that is based on clustering. In particular, we found that clustering with directional hunting is the best option if the goal is to balance request denial probability and latency. For the heterogeneous case, we also proposed an add-on feature that uses the reserved capacity feature along with the CDH strategy to provide fair treatment to traffics of different classes with differing data rates.

Our future research goal is to consider several different directions. First, we plan to consider a number of different topologies (beyond the ring topology) to do an extensive

assessment of the strategies. Beyond that, a critical direction to consider is scalability. For instance, if the number of users and the size of servers are scaled up, how does the system impact changes with the different strategies; secondly, it is well-known that when the size of the population is very large, a loss system can be modeled as  $M/M/m/m$ , instead of using the Engset-model; the question remains: how does scalability affect the accuracy of models? Such models are important to understand since they are useful for a what-if analysis of the overall system. Thirdly, we want to collect measurement data from Positive Networks' operational environment, and verify if Poisson arrival model is applicable. Another important direction is the generalization of the reserved capacity feature for the heterogeneous case when more than two service classes are available.

## REFERENCES

- [1] Z. Duan, Z. Zhang, and Y. Hou, "Service overlay networks: SLAs, QoS, and bandwidth provisioning," *IEEE/ACM Trans. Networking*, vol. 11, pp. 870–883, 2003.
- [2] N. G. Duffield, P. Goyal, and A. Greenberg, "A flexible model for resource management in virtual private networks," in *Proc. ACM SIGCOMM*, 1999.
- [3] N. G. Duffield, P. Goyal, A. Greenberg, P. P. Mishra, K. K. Ramakrishnan, and J. E. van der Merwe, "Resource management with hoses: point-to-cloud services for virtual private networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 5, pp. 679–692, 2002.
- [4] B. Gleeson, A. Lin, J. Heinanen, G. Armitage, and A. Malis, "A framework for IP based virtual private networks," *Internet RFC 2764*, February 2000.
- [5] V. B. Iversen, *Teletraffic Engineering and Network Planning*, manuscript, Technical University of Denmark, January 2007.
- [6] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. on Communications*, vol. COM-29, pp. 1474–1481, 1981.
- [7] L. Kleinrock, *Queueing Systems: Volume I*, Wiley Interscience, New York, 1975.
- [8] B. Kraimeche and M. Schwartz, "Analysis of traffic access control strategies in integrated service networks," *IEEE Trans. on Communications*, vol. COM-33, pp. 1085–1093, 1985.
- [9] K. Liu, J. Lui, and Z. Zhang, "Distributed Algorithm for Service Replication in Service Overlay Network," in *Proc. IFIP TC6 Networking Conference*, May 2004.
- [10] D. Medhi, B.-Y. Choi, C. Scoglio, S.-J. Song, and S. Dispensa, "Agent-based VPN architecture (ABVA): A Framework and The Optimal User Connectivity Problem (static case)," *Proc. of 15th International Conference on Advanced Computing & Communication (ADCOM2007)*, pp. 138–143, Guwahati, India, December 2007.
- [11] D. Medhi, A. van de Liefvoort, and C. S. Reece, "Performance analysis of a digital link with heterogeneous multislot traffic," *IEEE Trans. on Communications*, vol. 43, pp. 968–976, March 1995.
- [12] A. Nagarajan, Ed., "Generic requirements for provider provisioned virtual private networks (PPVPN)," *Internet RFC 3809*, June 2004.
- [13] Positive Networks, <http://www.PositiveNetworks.com/>
- [14] J. W. Roberts, "A service system with heterogeneous user requirements: application to multi-services telecommunications systems," in *Performance of Data Communication Systems, and Their Applications*, G. Pujolle (Ed.), North-Holland, pp. 423–431, 1981.
- [15] E. Rosen and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," *Internet RFC 4364*, February 2006.