Economics of Scalable Network Services

John Chuang[*]
School of Information Management and Systems, University of California, Berkeley

ABSTRACT

This paper explores the economics of scalable network services by posing two simple questions. First, what is the difference between scale economies and scalability? Second, why and how should we scale network services for competition and cooperation? By answering these questions in the context of network services such as multicast, QoS and web caching, we gain some insight into the tradeoffs involved in the design of scalable network services.

1. INTRODUCTION

The Internet is simultaneously characterized by scale and heterogeneity. Hundreds of millions of end hosts are connected to each other via the global Internet, by virtue of the ubiquitous communications protocol IP (Internet Protocol). At the same time, the end hosts are also highly heterogeneous, all the way from the physical connectivity (e.g., voiceband, broadband, wireless) to the application level requirements (e.g., low latency for real-time applications, high throughput for bulk data transfer applications, or both for interactive multimedia applications.) Even the networks are heterogeneous, in terms of capacity, coverage, availability, as well as ownership, revenue model, etc.

In this environment of scale and heterogeneity, IP has established itself as the archetype of a scalable network service. It has successfully scaled from 10^0 to 10^5 networks, and from 10^1 to 10^8 hosts in the last 30 years. Much of the recent research on advanced network services, such as multicast, QoS and web caching, focuses on their scalable provisioning over this global heterogeneous Internet.

It is also in this context of scale and heterogeneity that we want to begin a discussion, among the networking research community, on the economics of scalable network services. Traditionally, Internet economics researchers have focused on the application of microeconomic theory to the optimal pricing of network services [1-4]. This paper poses some broader, and perhaps more fundamental questions, that should be of concern to network architects. For example, how is the economics of network services different from the economics of traditional goods and services? What is the difference between scale economies and scalability? Why do we need to scale our network services for competition and cooperation? These are the questions we will address in the remainder of this paper.

2. SCALE ECONOMIES VERSUS SCALABILITY

Scale economies and scalability are two sides of the same coin of 'scale', though the concepts come from very different fields of study. Making a distinction between the two is a good starting point for discussion. Scale economies, or economies of scale, is an economic term used to describe a supply side condition where the average cost of production decreases as the production volume increases [5]. Infrastructural goods and services (e.g., transportation and communication networks) usually exhibit a high degree of scale economies. A significant up-front capital investment (high fixed cost) is required, but once the infrastructure is complete, the marginal cost of production or provision is often negligible.[1]

Scalability was first applied to computer systems design, specifically to the field of massively parallel computation [6]. An *n*-processor system is scalable if it can execute a program, task or algorithm at a speedup of $S(n)$ over a single-

---

[*] chuang@sims.berkeley.edu; http://www.sims.berkeley.edu/~chuang/; 102 South Hall, Berkeley, CA, USA 94538.

[1] The Internet is said to exhibit strong economies of scale in the demand side as well. Metcalfe's Law states that users value their connectivity to the network according to the square of the number of the users of the network. This condition is more accurately characterized as the presence of positive network externalities, and is outside the scope of this current discussion.

processor system. Scalability has since been applied to the design and evaluation of a wide variety of distributed systems, such as distributed file systems and the world wide web [7-9].

In the context of network services, scale economies provide a measure of how much bandwidth savings can be realized when using a network service over a shared communication medium, while scalability provides a measure of how much cost is incurred to support large scale deployment and provisioning of the service. 'Scale economies' provides the benefit measure, usually in the data plane, of 'why' a service is efficient; 'scalability' provides the cost measure, usually in the control plane, of 'how' to make a service technically and economically feasible.

To make the discussion more concrete, let us consider the scale economies and scalability of three broad categories of network services, namely multicast, QoS, and web caching.

2.1 Multicast

The primary motivation of multicast is to realize data-plane economies of scale when sending data to multiple destinations. Multicast postpones packet duplication until the point of route divergence, thereby avoiding the need of sending duplicate packets across any given link. Several works have quantified the economies of scale savings achievable by multicast [4,10-12]. There appears to be a power law relationship between the normalized multicast tree cost (L_m) and multicast group size (N), expressed as $L_m = N^k$, where k is the economies of scale power coefficient. If the coefficient is equal to unity, then there are no economies of scale, and multicast performs only as well as unicast. Empirical studies have placed the value of k to be in the 0.65 to 0.8 range.

According to this power-law scaling property (Figure 1), the data plane bandwidth savings achievable can be significant, especially for large-sized group applications. For example, for a session with 10 randomly distributed receivers, multicast realizes a bandwidth savings of 37% over unicast, while for a session with 100 receivers, the bandwidth savings increases to 60%. The savings increases further for even larger-sized multicast groups.

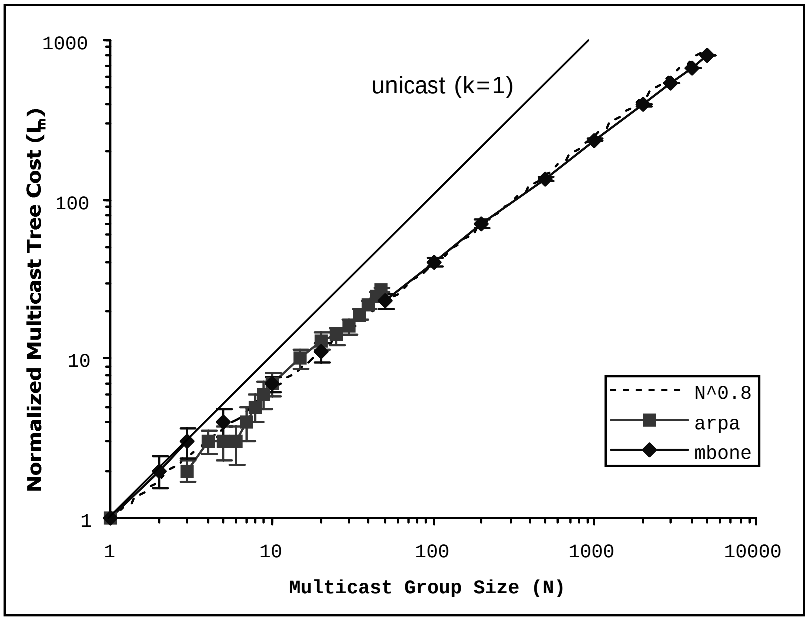

Figure 1: Economies of scale achievable by multicast [4].

The scalability of multicast, on the other hand, is dictated by the amount of control state and message exchange necessary to support multicast data delivery. In layer 3, routers need to maintain multicast forwarding state. Because no CIDR-style aggregation for class-D addresses is possible, one entry is needed for each multicast address. The amount of forwarding state per routing node is O(N), where N is the number of active multicast groups. Furthermore, if source-based multicast trees are used instead of shared multicast trees, and there are M active sources per multicast group, then the forwarding state is O(M*N) [13,14].

For reliable (layer 4) multicast, scalability is centered around the mechanisms for re-transmission/repair of lost packets. Traditional unicast solutions cannot scale to the multicast domain (e.g., ACK implosion). Depending on network, population, and loss characteristics, a variety of solutions (e.g., local repair [15,16], forward error correction [17]) may be appropriate for scalable reliable multicast.

2.2 Quality-of-Service

The primary motivation of Quality-of-Service (QoS) is to provide performance expectations, not to realize scale economies. However, different classes of QoS-aware services do produce different economies of scale outcomes. For example, services with deterministic guarantees have to reserve resources for the worst case, thereby resulting in low resource utilization. Services with statistical guarantees, as well as differentiated services, can realize statistical multiplexing gains and support a larger number of flows from the same amount of resource. Various analytical and empirical studies have shown that this statistical multiplexing gain can be significant, even if the traffic exhibits long range dependence [18-21].

The scalability of a QoS architecture is a function of the amount of state maintained and exchanged between the various QoS components (e.g., reservation manager, admission controller, traffic shaper, classifier, scheduler.) For example, the IETF intserv framework [22] provides per-flow service guarantees, but also requires per-flow state, making it less scalable than the diffserv framework [23], which only requires per-class state. More recent works such as reservation aggregation [24,25] and end-point admission control [26] seek to improve intserv scalability. [27] has also sought to provide service guarantees without per-flow state in the network core.

2.3 Web caching and replication

Replication is a well-known technique for improving the scalability of distributed systems. Replication removes single points of failure, and facilitates the distribution and balancing of load among the multiple replicas.

Web caching can be considered a form of replication, and it has significantly improved the scalability of the world wide web [28]. Web caching accomplishes economies of scale in the temporal dimension. By storing local copies of data objects in close proximity to the clients, these caches may satisfy subsequent requests for the same objects, thereby avoiding the cost of re-transmitting the objects. Studies have shown that web access exhibits a strong degree of reference locality [29,30], thus providing justification for the use of web caching. Various cache management and object replacement strategies are proposed and evaluated to improve the hit rate of, and therefore scale economies realizable by, individual cache nodes and caching systems [31,32].

The scalability of individual standalone cache nodes is directly related to the underlying data structure used for object management. For example, a least-recently-used (LRU) replacement algorithm can be implemented using a linked list, while a least-frequently-used (LFU) replacement algorithm can be implemented using a heap structure. The memory and processing complexity of these data structures are well understood. Derivative algorithms such as SIZE, LRV, GDS seek to improve hit rate by storing and computing additional state information on object size, object fetch cost, etc [31,32].

The scalability of multi-node caching systems is of greater importance and interest. A large number of proposals exist for the organization of cache nodes to maximize overall hit rate. One important question is: in the event of a local cache miss, to which cache shall the request be forwarded? Caches may be organized into some form of a hierarchy, mesh, or co-operative [33,34], and may maintain and exchange state information to assist in request routing [35-38].

Another important scalability concern is cache consistency. Different solutions offer different degrees of consistency at different costs. For example, invalidation provides a stronger consistency model than TTL-based polling, but it is unclear if it does so at the expense of scalability [39,40]. On the other hand, it is clear that multicast can be employed to improve the scalability of cache consistency mechanisms [41,42].

2.4 Tradeoff between scale economies and scalability

The design of network services often involves the tradeoff between scale economies and scalability. While network architects may not always think in these terms, successful network services are the ones that achieve a good balance between scale economies and scalability.

Once again, there are ample examples in the real world to illustrate this tradeoff. In IP multicast, shared trees require less forwarding state and are therefore more scalable than source-based trees. However, shared trees result in a lower routing efficiency because the delivery tree may be sub-optimal for senders far away from the root. In reliable multicast, the use of forward error correction (FEC) techniques may require the initial transmission of additional, redundant data, but is more scalable than the traditional ARQ (automatic repeat request) technique of data repair used in unicast TCP. In QoS, there is a similar tradeoff between scalability and achievable resource utilization [43]. For example, reservation aggregation [24] can be used to improve the scalability of intserv, but at the potential cost of reduced resource utilization [25]. Finally, in web caching, much of the recent research activity has focused on the improvement of system-wide hit rate without making the system non-scalable.

Of course, tradeoffs can be made between scale-economies/scalability and other factors as well. The premise of recent end-system multicast proposals is to trade off scale economies for deployability [44,45]. The intserv/diffserv dichotomy is between scalability and service guarantees. The tradeoff between strong and weak web cache consistency is that of scalability/scale-economies and data freshness.

3. SCALING FOR COMPETITION AND COOPERATION

When we talk about scalable network services, we usually think about scaling in terms of number of nodes, links, flows, bytes and/or timescales. We consider the amount of state that needs to be stored, exchanged and updated, and the locations at which the state must reside. There is, however, an additional and important dimension of network service scalability -- scaling for competition and cooperation. Along this dimension, we want to design and deploy network services to scale from one to multiple service providers, from intra-domain to inter-domain, from homogeneous to heterogeneous environments, from standalone network to interconnection and peering, and from monopoly to sustainable competition and cooperation.

The Internet Protocol has successfully scaled from 10^0 to 10^5 networks, and from 10^1 to 10^8 hosts in the last 30 years. In fact, IP service is now traded as a commodity over bandwidth markets [46], offering a strong testimony to the robustness of the competition in this service.

Unfortunately, no subsequent network service has been able to duplicate, or even come close to, the successful scaling of IP. While intra-domain multicast and QoS are well researched and standardized, widespread deployments of multicast and QoS services have so far been impeded by the lack of scalable inter-domain solutions. Web caching has been inter-domain from the beginning, but has significant shortcomings, and is therefore facing serious competition from proprietary, intra-domain caching solutions (e.g., content delivery networks.) Even the interconnection of IP networks is showing signs of strain, as large networks begin to exercise their market power to dictate the terms of peering and interconnection (see [47] for an example). In this section we will explore the tradeoffs and strategies involved in scaling for competition and cooperation.

3.1 Scaling for competition

From an economic efficiency and public policy perspective, it is desirable for network services to be offered by multiple competing service providers. However, due to the economic characteristics of the Internet, scaling for competition may imply a penalty in terms of loss in economies of scale.

Consider the business of network service provision. Like the provisioning of other communications services, it involves a very high fixed cost and a very low marginal cost, resulting in strong economies of scale. The service provider thus faces a declining average cost curve as shown in Figure 3. The greater the quantity Q produced, the lower the average per-unit cost. This is in contrast to the production of traditional goods (Figure 2), where the average cost curve faced by a firm reaches a minimum at Q^*, and begins to rise with further increases in Q, due to management costs, etc. This firm will choose to produce at the optimal quantity Q^*. If the market size is $Q_{TOT} = NQ^*$, then the market can support N competing firms efficiently.

Returning to Figure 3, the entire market Q_{TOT} is most efficiently served by a single provider, rather than by multiple providers. Economists call this scenario a 'natural monopoly'. A natural monopoly behaves like regular monopolies -- constrain the production level and set prices to maximize profits, not social surplus. Therefore, whatever gains realized from economies of scale will be reaped by the monopolist, but not passed on to the consumers.

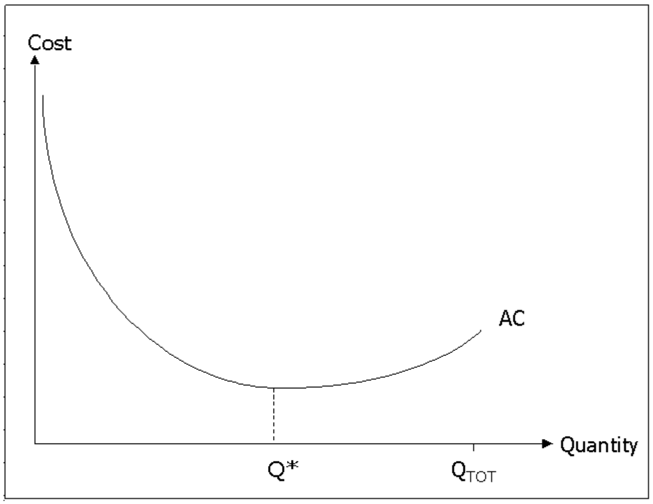

Figure 2: Average cost curve for traditional goods and services.

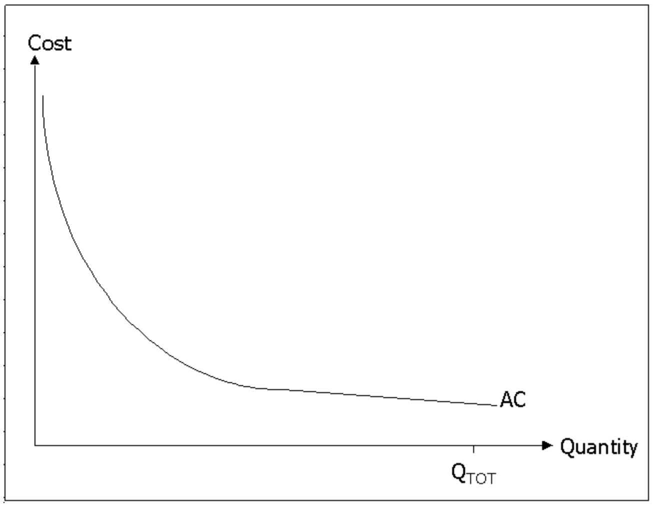

Figure 3: Average cost curve for network service provision.

There are several possible 'remedies' to this form of 'market failure'. First, the operation can be performed by a public, non-profit entity such as the government (e.g., municipal water and sewage treatment). Second, the monopoly can be subject to regulation to protect consumer interests (e.g., rate regulation for residential local phone service). Third, legislative and/or regulatory actions to encourage competition may be put in place (e.g., the Telecommunications Act of 1996 and its intention to introduce competition into the local loop). It is also worthwhile to note that natural monopolies may not last forever. Technological advances may reshape the cost curve such that economies of scale are no longer applicable for the given size of the market (Figure 4). This market can now be efficiently supported by multiple competing providers.[2]

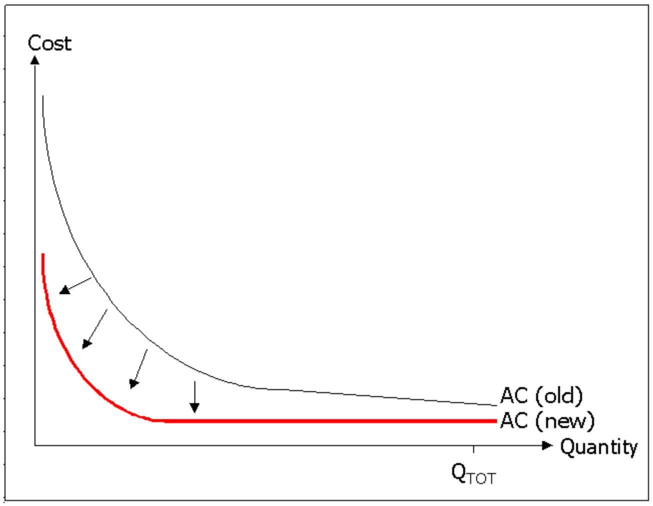

Figure 4: Shifts in cost curve due to technological advances.

3.2 Scaling for cooperation

Given the geographic scale and scope of the Internet, no network service provider can possibly achieve global, end-to-end coverage. Therefore it is of practical necessity that network services are designed to scale for cooperation. The technical challenge is the seamless deployment and provisioning of network services across multiple network domains.

Inter-domain network service design is very different from, and considerably more difficult than, its intra-domain counterpart. Heterogeneity is assumed in all possible dimensions: ownership, capabilities, objectives, etc. Therefore, in addition to performance and cost considerations, the designer has to worry about issues like incremental deployability, accountability, security and trust as well. Furthermore, while intra-domain routing (of IP datagrams and HTTP requests/replies alike) is usually metrics-based, inter-domain routing is usually policy-based, driven by business contracts and agreements. Interconnection and peering can be established at layers 3 and above, and mechanisms must be put in place to support the establishment, monitoring, verification, and potentially, financial settlement of these contractual relationships [48].

It is unsurprising, therefore, that the designs of inter-domain and intra-domain network services have historically been tackled as separate problems, with the inter-domain problem receiving far less attention and priority in the research community.

3.2.1 Multicast

In IP multicast, many candidates have been proposed for intra-domain multicast routing, including DVMRP, MOSPF, PIM-DM, PIM-SM and CBT. The Mbone [49], for example, is an intra-domain multicast backbone that uses DVMRP as the routing algorithm, and uses tunneling to connect the various multicast-aware islands over the largely multicast-unaware Internet.[3] A small number of commercial networks (e.g., UUNET, Sprint) have begun to offer intra-domain multicast in the last few years. Inter-domain protocols such as MASC/BGMP [51] have been proposed, but inter-domain multicast remains an elusive goal.

---

[2] This was essentially the case for long-distance telephony in the early 1980's and the motivation behind the divestiture of AT&T in 1984.

[3] QBone offers a similar story -- researchers want to build a worldwide experimental network that supports QoS, but inter-domain QoS is too difficult to achieve. Therefore, QBone is built, as an intra-domain network, by using tunneling [50].

End-system multicast [44] is a recent proposal that offers an interesting alternative to inter-domain multicast. It attempts to overcome the deployability problem by moving the multicast functionalities to the end systems (at the application layer). This protocol assumes no underlying network support beyond simple IP unicast, and self-organizes the multicast group participants into an overlay network. When the topology of this overlay network differ from that of the underlying IP network, the protocol pays a penalty in routing inefficiency. Therefore, the protocol is more suited for small-group multicast applications. In the context of this paper, the designers are making an explicit tradeoff between scale economies and scalability (in terms of deployability).

3.2.2 Web caching and content peering

Finally, let us move to the application layer to consider the case of web caching, content delivery networks (CDNs) and content peering. The global web caching infrastructure is also characterized by heterogeneity. A large number of cache nodes are deployed and operated by independent entities (e.g., ISP's, corporations, universities and other institutions) throughout the world. The caches have been deployed as proxies, surrogates (reverse proxies), transparent caches, etc., and may belong to some caching hierarchy, mesh or co-operative [33,34]. Various cooperative caching proposals call for the nodes to maintain and exchange state information regarding their cached contents, such that requests can be forwarded to nodes where the objects may actually reside [35-38].

While these proposals seek acceptance and adoption by the caching community, several commercial content delivery networks (e.g., Akamai, Digital Island) have been successfully deployed. They are offering, in essence, an intra-domain caching service to their paying customers (e.g., content providers). The largest CDNs have each deployed, in an overlay fashion, thousands of cache nodes throughout the Internet, and dynamically routes object requests to nodes that are in close proximity of the request originators. Because all the cache nodes are under a single domain of administrative control, the CDNs can easily manage the placement, replacement and update of objects, as well as provide usage and SLA reports to their customers. These are difficult to accomplish in the traditional, inter-domain caching environment.

While intra-domain overlay caching has significant advantages in deployability and control, it is difficult for a single CDN provider to achieve global coverage. The existing web caching infrastructure is at least an order of magnitude larger (in terms of number of nodes) than the largest CDN. Therefore, several content peering initiatives have been submitted to the IETF [52,53]. These initiatives aim to interconnect the CDNs with one another and with the open caching infrastructure. It remains to be seen if these standardization efforts will be successful, but it is clear that the existing web caching and content delivery solutions must eventually be scaled for cooperation and competition.

4. CONCLUSION

This paper raises more questions on the economics of scalable network services than it attempts to answer. However, it is important for network researchers and architects to be cognizant of the economic aspects of their designs. First, we need to recognize that there is a distinction between scale economies and scalability of network services, and often times, network design involves some fundamental tradeoffs between these two factors. Second, network services should be designed to scale for competition and cooperation, even though this may come at the expense of efficiency due to the loss of scale economies.

The balance between scale economies and scalability is likely to be dependent on the characteristics of the specific service in question. In some instances, a ‘one-size-fit-all’ solution might be desirable or achievable. In other instances, a tiered approach (e.g., inter-domain and intra-domain) might be more appropriate. The application layer overlay approach provides the ability to trade off scale economies for deployability. Finally, it may be worthwhile to explore the design of ‘tune-able’ network services, where scalability and/or scale economies are tune-able service parameters that can be controlled by the service provider and/or consumer to achieve targets of cost-effectiveness.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Cooperative Agreement Number ITR-0085879. Views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of NSF or the U.S. Government.

REFERENCES

[1] R. Cocchi, S. Shenker, D. Estrin and L. Zhang. Pricing in computer networks: motivation, formulation and example. IEEE/ACM Transactions on Networking 1(6): 614-627, 1993.

[2] S. Shenker, D. Clark, D. Estrin and S. Herzog, Pricing in computer networks: reshaping the research agenda. Telecommunications Policy 20(3) (April 1996).

[3] Clark, D. Internet cost allocation and pricing. In *Internet Economics*, L. McKnight and J. Bailey, eds., pp. 215-252. MIT Press, 1997.

[4] J. Chuang and M. Sirbu, "Pricing multicast communication: A cost based approach," in Proceedings of INET'98, Geneva Switzerland, July 1998.

[5] G. Stigler. 1958. The economies of scale, Journal of Law and Economics 1, 54-71.

[6] M. D. Hill. What is Scalability? Computer Architecture News, 18(4):18--21, December 1990.

[7] P. Jogalekar, M. Woodside, Evaluating the Scalability of Distributed Systems, IEEE Transactions on Parallel and Distributed Systems Vol. 11, No. 6: June 2000, pp. 589-603.

[8] M. Satyanarayanan, The Influence of Scale on Distributed File System Design, IEEE Transactions on Software Engineering, Vol. 18, N. 1, January 1992, pp. 1 - 8.

[9] C. Yoshikawa, B. Chun, P. Eastham, A. Vahdat, T. Anderson and D. Culler. Using Smart Clients to Build Scalable Services. In Proceedings of the 1997 USENIX Technical Conference, Jan. 1997.

[10] G. Phillips, H. Tangmunarunkit, and S. Shenker. Scaling of Multicast Trees: Comments on the Chuang-Sirbu scaling law. ACM SIGCOMM 1999.

[11] R. Chalmers and K. Almeroth, "Modeling the Branching Characteristics and Efficiency Gains of Global Multicast Trees", IEEE Infocom 2001, Anchorage, Alaska, USA, April 2001.

[12] R.W. van der Hofstad, G. Hooghiemstra, P. Van Mieghem, "On the Efficiency of Multicast", under submission.

[13] D. Thaler and M. Handley. On the Aggregatability of Multicast Forwarding State. In Proceedings of IEEE INFOCOM 2000, Tel Aviv, Israel, March 2000.

[14] T. Wong and R. Katz, An Analysis of Multicast Forwading State Scalability, In the 8th International Conference on Network Protocols (ICNP), November 2000, Osaka, Japan.

[15] H. Holbrook, S. K. Singhal, and D. R. Cheriton, "Log-based receiver-reliable multicast for distributed interactive simulation, " in Proc. ACM SIGCOMM'95, pp. 328--341, August 1995.

[16] S. Floyd, V. Jacobson, C. Liu, S. McCanne, and L. Zhang. A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing, IEEE/ACM Transactions on Networking, December 1997, Volume 5, Number 6, pp. 784-803.

[17] L. Rizzo, L. Vicisano, "A Reliable Multicast data Distribution Protocol based on software FEC techniques", Proceedings of The Fourth IEEE Workshop on the Architecture and Implementation of High Performance Communication Systems (HPCS'97), Sani Beach, Chalkidiki, Greece June 23-25, 1997.

[18] A. Erramilli, O. Narayan, and W. Willinger. 1996. Experimental queueing analysis with long-range dependent packet traffic. IEEE/ACM Transactions on Networking 4:209--223.

[19] E. W. Knightly, D. Wrege, J. Liebeherr, and H. Zhang. Fundamental limits and tradeoffs of providing deterministic guarantees to VBR video traffic. ACM SIGMETRICS'95 Conference, pages 98--107, May 1995.

[20] I. Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks", IEEE Journal on Selected Areas in Communications, Vol. 13, No. 6, pp. 953962, August 1995.

[21] N. Duffield, "Economies of scale for long-range dependent traffic in short buffers," Telecommunication Systems, 7 (1997) 1-3, pp. 267-280

[22] R Braden, D. Clark & S.Shenker. "Integrated Services in the Internet Architecture: an Overview", June 1994, RFC 1633.

[23] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss. An Architecture for Differentiated Services RFC 2475, Dec 1998.

[24] F. Baker, C. Iturralde, F. Le Faucheur, and B. Davie. Aggregation of RSVP for IP4 and IP6 Reservations. Internet Draft, draft-ietf-issll-rsvp-aggr-02.txt, March 2000.

[25] H. Fu and E. Knightly, ``Aggregation and Scalable QoS: A Performance Study," in Proceedings of IWQoS'01, Karlsruhe, Germany, June 2001.
[26] L. Breslau, E. Knightly, S. Shenker, I. Stoica, and H. Zhang. Endpoint Admission Control: Architectural Issues and Performance. ACM Sigcomm 2000.
[27] I. Stoica and H. Zhang. Providing Guaranteed Services Without Per Flow Management. In Proceeding of SIGCOMM '99, 1999.
[28] J. C. Mogul. Squeezing More Bits Out of HTTP Caches. IEEE Network 14(3):6-14, May/June, 2000.
[29] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. "Web Caching, and Zipf-like Distributions: Evidence, and Implications," Proceedings of IEEE INFOCOM, 1999.
[30] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing Reference Locality in the WWW. In Proceedings of 1996 International Conference on Parallel and Distributed Information Systems (PDIS'96), December 1996.
[31] L. Rizzo and L. Vicisano Replacement policies for a proxy cache. IEEE/ACM Transactions on Networking Volume 8, No. 2 (Apr. 2000) Pages 158 - 170.
[32] P. Cao and S. Irani. Cost-aware WWW proxy caching algorithms. In Proceedings of the 1997 USENIX Symposium on Internet Technology and Systems, pages 193-206, December 1997.
[33] I. Cooper, I. Melve, G. Tomlinson, Internet Web Replication and Caching Taxonomy, RFC 3040, January 2001.
[34] S. Michel, K. Nyugen, A. Rosenstein, L. Zhang, S. Floyd, and V. Jacobson, “Adaptive web caching: towards a new global caching architecture,” Third International WWW Caching Workshop, Manchester England, 1998.
[35] D. Wessels and K. Claffy, "Internet Cache Protocol (ICP), Version 2", RFC 2186, September 1997.
[36] L. Fan, P. Cao, J. Almeida and A. Z. Broder, "Summary Cache: A Scalable Wide-area Cache Sharing Protocol" in Proceedings of the ACM SIGCOMM'98, October 1998.
[37] V. Valloppillil and K. Ross, "Cache Array Routing Protocol", Work in Progress.
[38] M. Hamilton, A. Rousskov and D. Wessels, "Cache Digest Specification - version 5", December 1998.
[39] J. Gwertzman and M. Seltzer. World Wide Web cache consistency. In Proceedings of 1996 USENIX Technical Conference, pages 141--151, San Diego, CA, January 1996.
[40] C. Liu and P. Cao "Maintaining Strong Cache Consistency in the World Wide Web." IEEE Transactions on Computers, 47(4): 445-457, Apr. 1998.
[41] H. Yu, L. Breslau, and S. Shenker. A Scalable Web Cache Consistency Architecture. In Proceedings of the ACM SIGCOMM'99, Boston, MA, September 1999.
[42] D. Li and D. R. Cheriton. Scalable web caching of frequently updated objects using reliable multicast. In Proceedings of the Second USENIX Symposium on Internet Technologies and Systems, pages 92--103, Oct. 1999.
[43] J. Liebeherr, S. D. Patek and E. Yilmaz, Tradeoffs in Designing Networks with End-to-End Statistical QoS Guarantees, Proc. IEEE/IFIP Eighth International Workshop on Quality of Service (IWQoS '2000), pp. 221-230. June 2000.
[44] Y-H Chu, S. Rao, and H. Zhang. A case for end system multicast. In ACM SIGMETRICS, June 2000.
[45] Y. Chawathe, S. McCanne, and E. Brewer. RMX: Reliable Multicast in Heterogeneous Networks. In Proc. IEEE INFOCOM, March 2000.
[46] J. L Mindel and M. A. Sirbu. Taxonomy of Traded Bandwidth. Presented at MIT Internet and Telecoms Convergence Consortium Meeting (ITC) (January 2001).
[47] UUNET. WorldCom Policy for Settlement-Free Interconnection with Internet Networks, January 2001. http://www.uu.net/peering/
[48] P. Srinagesh, Internet Cost Structures and Interconnection Agreements, in Internet Economics, L. McKnight and J. Bailey, Eds.:MIT Press, 1997.
[49] S. Casner and S. Deering, First IETF Internet audiocast, ACM Computer Communication Review (July 1992) 92-97.
[50] B. Teitelbaum, S. Hares, L. Dunn, R. Neilson, R. V. Narayan, and F. Reichmeyer. Internet2 Qbone: Building a Testbed for Differentiated Services. IEEE Network, 13(5):8--16, September /October 1999.
[51] S. Kumar, P. Radoslavov, D. Thaler, C. Alaettinoglu, D. Estrin, and M. Handley, "The MASC/BGMP architecture for inter-domain multicast routing," in Proceedings of SIGCOMM '98, Vancouver, Canada, Sept. 1998.
[52] Content Alliance. http://www.content-peering.org/
[53] Content Bridge. http://www.content-bridge.com/