Valiant Load Balancing, Capacity Provisioning and Resilient Backbone Design

Alejandro López-Ortiz

Cheriton School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1 Canada
alopez-o@uwaterloo.ca

Abstract. The two main alternatives for achieving high QoS on the public internet are (i) admission control and (ii) capacity overprovisioning. In the study of these alternatives the implicit (and sometimes explicit) message is that ideally, QoS issues should be dealt with by means of sophisticated admission control (AC) algorithms, and only because of their complexity providers fall on the simpler, perhaps more cost-effective, yet "wasteful" solution of capacity overprovisioning (CO) (see e.g. Olifer and Olifer [Wiley&Sons, 2005], Parekh [IWQoS'2003], Milbrandt et al. [J.Comm. 2007]). In the present survey we observe that these two alternatives are far from being mutually exclusive. Rather, for data critical applications, a substantial amount of "overprovisioning" is in fact a fundamental step of any safe and acceptable solution to QoS and resiliency requirements. We observe from examples in real life that in many cases large amounts of overprovisioning are already silently deployed within the internet domain and that in some restricted network settings they have become accepted practice even in the academic literature. Then we survey the main techniques currently in use to compute the provisioning capacities required in a resilient high QoS network.

1 Introduction

In the quality-of-service literature (QoS) two main alternatives are given for achieving high QoS on the public internet. These are (i) admission control and (ii) capacity overprovisioning. In the study of these alternatives the implicit (and sometimes explicit) message is that ideally, QoS issues should be dealt with by means of sophisticated admission control (AC) algorithms, and only because of their complexity providers fall on the simpler, perhaps more cost-effective, yet "wasteful" solution of capacity overprovisioning (CO) (see e.g. [22,23,20]). AC researchers often express the hope that this situation will eventually remedy itself and that sophisticated AC algorithms will do away with the need for bandwidth overprovisioning (e.g. [8]). Only recently Menth et al. in a SIGCOMM'06 paper gave evidence that CO might not be as undesirable as previously thought [19].

In the present survey we observe that these two alternatives are far from being mutually exclusive. Rather, for data critical applications, a substantial amount

J. Jansen and P. Prałat (Eds.): CAAN 2007, LNCS 4852, pp. 3–12, 2007.

of "overprovisioning" is in fact a fundamental step of any safe and acceptable solution to QoS requirements. Indeed, a survey of common practices in the field suggests that this observation has been arrived to independently and empirically by network engineers in various settings within the Internet and otherwise, yet the QoS literature so far does not reflect this discovery nor has it attempted to explain its root causes.

We observe from examples in real life that in many cases large amounts of overprovisioning are already silently deployed within the internet domain and that in some restricted network settings they have become accepted practice even in the academic literature. In other words, distate for overprovisioning is not a universally held belief outside the QoS domain. In fact, the telephony network which is considered a classical example of AC is in practice heavily overprovisioned and actual AC policy is rarely relied upon even though it is deployed on the network [21]. Then we survey the main techniques currently in use to compute the provisioning capacities required in a resilient high QoS network. We term this amount rightprovisioning. Lastly, we give reasons why QoS over a rightprovisioned network has different needs and goals than those currently addressed by admission control and other such mechanisms.

2 Internet QoS

The two main mechanisms for achieving a desired level of service on the internet are admission control and capacity overprovisioning. QoS on the network allows the user to make choices as to the level of service it requires. Typical parameters are: data rate (bandwidth), availability, end-to-end delay (latency), variation of end-to-end delay (jitter), and packet loss rate [8].

2.1 Capacity Overprovisioning

Capacity overprovisioning consists in increasing available bandwidth until it is large enough to sustain the vast majority of peaks in demand. Depending on the level of reliability desired this can be as low as 25% above average data rate to handle 95% of all traffic demands without loss, 50% extra bandwidth to carry 99% of traffic, and double the average bandwidth to meet 99.99% or higher of all traffic demands without loss (see e.g [32]). This last choice, meaning 50% utilization of the pipe, is often anecdotically referred to as the upper limit of utilization currently acceptable by large ISPs, with the load on an average link often being well below that [7,8].

In contrast, in the QoS literature overprovisioning is considered a simple but wasteful solution to QoS demands. For example, to quote from a computer networks textbook [22]:

> Overprovisioned services keep the network infrastructure simple (no additional tools and configurations) but are wasteful as 60-70% of potential network resources are not in use. Under such conditions the best-effort

service on a standard IP network turns out to be good enough for all network applications including time-sensitive ones.

Indeed the term "overprovisioning" itself has the implication that more capacity than what was required was provisioned and hence it ends up being wasted. Yet, subutilization of a resource alone does not imply it was *over*provisioned. In fact, most mission critical applications such as avionics routinely rely on highly redundant configurations, which under normal operational procedures are not used. For example an ocean liner arriving safely to port did not utilize its life boats, yet no one would argue that they were thus "overprovisioned".

2.2 Admission Control

Admission control is mostly about using resource reservation and limits on traffic volume to preven overload on the network. It is predicated on the basis that not all network traffic is time-sensitive and mission critical. The AC alternative to overprovisioning is denying resources to non crucial flows. Typical examples of time-sensitive traffic are real time flows (e.g. video/audio streaming, IP telephony) and high value transactions (stock trades, last bid at an online auction). Packets are assigned a priority value with higher priority packets being given preferencial service. Yet a look at the historical development of the internet suggests that, over the years, the majority of the traffic overtime has become more time sensitive and mission critical. Recall that in the original internet the majority of traffic was smtp (email) and nntp (usenet) based. These protocols have acceptable delay tolerances from several minutes to as long as days. Web traffic which is served interactively has acceptable delays in the 10 second or less range. VoIP and other streaming traffic have subsecond delay tolerances.

As more of the nation infrastructure migrates to the public internet, a disruption in the network has larger consequences. The financial, defense, telephone, commerce, government, and business infrastructure now rely on the availability of the Internet to operate properly. Even a seemingly non-mission critical application such as a standard home network connection which might have been initially deployed for one parent's non-time sensitive email (smtp) traffic later on became used by the kids for highly time-sensitive gaming and audio streaming as well as by a parent bidding in online auctions for objects worth thousands of dollars, and as of recently is being used as a carrier for VoIP services which means that emergency calls (911 or to the family doctor) are routed over it. These last type of calls are both time-sensitive *and* mission critical. Thus, it is not far-fetched to envision a world in which the majority of the traffic will be labeled as time sensitive and hence the savings from AC would be minimal, since not many flows can be dropped. This would make packet classification schemes at admission control points progressively more difficult and less useful, the majority of the traffic is critical to start with.

This suggests that as more data exchanges migrate to the Internet infrastructure, the need for higher reliability will further increase while the ability to differentiate between types of traffic will continue to decrease.

3 Rightprovisioning

Capacity overprovisioning is common place in the current internet [1,8,18,12]. AC based solutions remain unused while anecdotal evidence suggests that CO is the prefered method for QoS delivery in the commercial internet. Currently QoS due to CO is such that no packets are dropped in the backbones [8,15,4]. Packet loss occurs mostly in the interface between the end points of the network and the large ISP providers. As providers have focused on ensuring that there is sufficient deployed capacity rather than on implementing admission control solutions. ISPs will go to the extent of delaying by several months the start of connectivity for a new customer to ensure that there is enough capacity on the network to support the bandwidth demands of the new customer (this can be argued is a crude form of admission control). In other words, currently ISPs find that CO is a cost effective way to achieve QoS.

While most of the literature is critical of CO as a solution of QoS, recent developments suggest that even in theory its performance is better than originally thought. Bhagat observes that in certain settings overprovisioning seems to be a better answer to the performance needs from users, and indeed he goes as far as questioning the need for admission control based QoS solutions [6]. In a recent breakthrough paper in SIGCOMM'06 Menth et al. [19] show that if overprovisioned capacity is also used to achieve resilience against network failures, then the demands in terms of bandwidth of failure-resilient AC and CO schemes are comparable, as the overprovisioned capacity can be deployed for various uses depending on the type of congestion and/or failure detected. In sum, so far we have argued that

1. selective admission as required by AC is becoming increasingly less of an option at the backbone level since traffic is increasingly time and mission critical,
2. that CO in large trunks is already in place and provides excellent QoS within the core of the network,
3. that as such its effectiveness is well supported by established practice, and that
4. the academic literature has started to explain why CO is such an effective solution.

The question then remains what is the proper level of overprovisioning, i.e. rightprovisioning. Currently the model most commonly in use is a statistical guarantee of the probability of connection denial. We argue that the right metric is to provide enough capacity so that any valid traffic matrix can be realized.

Definition 1. *Formally, let $e_1, \ldots, e_n$ be n end points in the network each with a send and receive capacity s_i and r_i respectively. A* traffic matrix $A = [a_{ij}]$ *contains in entry a_{ij} the instantaneous amount of traffic from node e_i destined to e_j.*

Definition 2. *A given traffic matrix is said to be* valid *if* $\sum_{j=1}^{n} a_{ij} \leq s_i$ *and* $\sum_{i=1}^{n} a_{ij} \leq r_j$. *That is no node is attempting to send more data than it has uplink provisioned capacity for and no node is being sent more data than it has contracted capacity to receive.*

In the past providers have deployed enough capacity to handle the average traffic matrix or a percentage of traffic matrix configurations (say 95% of the time the traffic matrix should cause no loss in traffic). Since the aim is to provide connectivity for the worst case traffic matrix we need to determine what is the minimum or most cost efficient capacity that satisifies this requirement. We could simply consider the sum of all contracted capacity by users, however this does not take into account that currently connectivity is provided in an average fashion, typically at a certain average rate per month with a maximum burstable rate.

In the new regime, two types of traffic would be provisioned. Traffic of type A, which is mission critical and always available at the contracted capacity and traffic of type B, at an average contracted capacity but rate-controlled depending on connectivity characteristics. In essence this could be thought as rate modulation over a pipe carrying type B traffic, not unlike in nature and effects to that performed by a modem in the presence of high levels of line noise. Observe that this establishes a very simple form of admission control. Traffic of type A would be unavailable at most on the order of subsecond to few seconds per year range (seven to eight nines of reliability). At the same time the entire contracted capacity should be generally available, with traffic of type B being flow rate controlled in the order of a half a minute to a few minutes a year (five to six nines range). This last is the current level of service reliability that the telephone network claims to have, even though arguably telephone traffic is less time critical than many of the current uses of the network. It is worthy of note that the telephone network operates at 33% capacity [21] and that the amount of admission control is minimal. For example "on Monday, Dec. 2, 1991, which was the busiest day for the AT&T network until then, of 157.5 million calls, only 228 were blocked on intercity connections" (from [3] as quoted by [21]). Our proposal parallels this design choice.

Interestingly enough, worst-case traffic matrix $n \times n$ capacity already exists in certain network settings. In the LAN the proper amount of overprovisioning has evolved to be such that, given n nodes on an Ethernet, a complete set of $n/2$ disjoint pairs can communicate at full speed. Recall that this was not always the case, as the original co-axial ethernet only had sufficient capacity for a single pair to communicate freely at full capacitywithout collision; eventually star switches with higher capacity buses became commonplace, and currently common $n \times n$ crossbar or Beneš network switches have the ability to sustain $n/2$ disjoint pairs of communication [2]. Similarly Network Access Points (NAPs) as well as cores of large corporate networks often consist of an optical ring providing enough capacity for all possible crossconnects. This is not unique to the internet. In the 1970s telephone networks deployed switches with $n \times n$ capacity at certain critical points of the infrastructure [26].

For statistical guarantees the law of large numbers can be used to determine the maximum simultaneous demand that may originate, on the aggregate, from a neighborhood of nodes sharing an entry point to the ISP backbone. This is repeated for all entry points into the backbone and then a full $n \times n$ bandwidth capability over those averages can be deployed. The size of such an $n \times n$ network is well understood. We discuss in detail the various known alternatives in Section 5.

Lastly, as Menth et al. observed, redundant equipment can be deployed for multiple purposes, so long as the probability of failure of such equipment is independent [19]. This amortizes the additional cost of redundant equipment. In particular redundant capacity can be used to circumvent router and link failures (digging). This has been observed to reduce the amount of apparent "subutilization". As well, secondary sources of traffic which can be quenched at the source point can be sent over the spare capacity. Examples of this are CDN content and remote backup data which are resilient under short time delays. Anecdotal evidence suggests that spam traffic is delivered at off-peak times by certain ISPs using deployed overcapacity.

4 QoS and AC in a Rightprovisioned World

Observe that we do not claim that overprovisioning at the backbone is sufficient to achieve all QoS requirements, nor would it make AC trivial. This is in contraposition to claims to that end in the literature, e.g. "only when the ratio of resources at the edges of a network to those available in the core of a network becomes high is the problem of service differentiation interesting, [...] when this ratio is low, any QoS mechanism appears redundant as most users receive the service they require anyway, and so the cost introduced by a QoS scheme appears unjustified, and research into QoS mechanisms appears unnecessary" [8].

For one, as the network is used for more life-critical operations such as VoIP phone calls (911), financial transactions (stock exchange), remote surgery, and air traffic system, perhaps even carrier grade reliability is not good enough. It is not hard to envision demands for reliability reaching into the 99.999999% range (in fact today it is possible to provision bandwidth with a stated 100% reliability guarantee in the sense that *any* amount of downtime is contractually heavily penalized). Such high levels of reliability will require overprovisioning, multihoming, redundancy, admission control and intelligent routing, though the types of solutions required, their price/performance ratio and their goals change. As well, end users will still, on occasion, attempt to send or receive more time critical data that is feasible given their available network connectivity. Admission control in such situations will be needed to prioritize say, a 911 VoIP call (type A traffic) over downloading email (type B traffic).

Admission control starts from the assumption that congestion will always take place at the edge given the reduced capacities of the endpoint as compared to the capacity of the entire network (i.e. the need to send or receive more data than

what we have capacity for). What this work argues is that congestion should only take place at the edge and that CO is the way to ensure this.

The model we propose assumes that all packets reaching the network core are assumed to be critical and hence failure of delivery is not an option. Within the core there would be no differentiated services with AC taking place as a weak and simplified form of resource reservation: if the packet is admitted, it can be delivered. The end node would send data in one of two modes: normal mode in which all traffic is accepted without need for any AC intervention and exceptional mode in which the application/user is alerted of a temporary service disruption and given the choice to proceed with the communication at full speed or throttle down for a few seconds (type A or B classification). Incentives such as price differentials can be built in to ensure that the user delays non-essential traffic.

Given the reliability needs detailed above this would occur with a very low probability, in the range of thirty seconds to a few minutes of service disruption per year. Such a rare occurrence means that only the simplest of differentiated services and admission control policies can be justified from the perspective of economic viability. As it has been observed [8] a weak form of AC already takes place in the edges in that providers delay customer activation to ensure that enough capacity is present to satisfy demand. This is a crude yet effective form of denying a transmission request.

As well routing in an overprovisioned network is more complicated as the multiplicity of paths allows for an intelligent choice. This determination does not involve the end point as the network makes best effort for all packets.

5 Valiant Load Balancing and Beneš Networks

Claude Shannon pioneered the study of networks that support $n \times n$ communication pairs. He proved that if the proving that a fabric of $n \log n$ switches is necessary so long as total deployed capacity is linear. Beneš introduced the later termed Beneš networks which match Shannon's lower bound switch [2,10,11,26]. Arora et al. combined Beneš networks with the butterfly network to obtain a similar topology that supports all cross connects in an online fashion, while preserving the efficiency in terms of deployed capacity.

If there are no restrictions in the total deployed capacity then other alternative realizations are possible. Two of the most common being a central high capacity ring and the $n \times n$ crossbar. Pippenger extensively studied the topology of telephone switches that support $n \times n$ connection patterns [24,25,26,27,28].

Valiant proposed an elegant network topology in the context of interprocessor communication networks for parallel computers [33,34]. The network starts with the complete graph on n nodes which trivally can support all independent connection pairs. However it is an inefficient solution as it requires n^2 contracted capacity. Valiant's key observation is that a two phase communication protocol on a complete graph in which every link has capacity $2/n$ suffices. This reduces the total deployed capacity to $2n$ which is a constant times the deployed capacity.

Extensions and generalizations of both Pippenger's and Valiant's work have been the subject of intense study within theoretical computer science [9,14,2] as well as the networks community [17,30,31,29]. The field is now refered to as a Valiant Load Balancing Network and/or as a Virtual Private Network load balancing. The term VPN comes from the fact that VPNs were one the earliest users of the internet requiring high degree of reliability. Shepherd et al. and Prasad et al. have run simulations to determine the effect of VPN load balancing in existing networks, and have observed that peak traffic loads are lowered down while resilience is improved. Many open questions remain, among them

- how to efficiently design an overprovisioned network under realistic cost measures?
- design an overprovisioned network which readily scales under incremental growth?
- given a pre-existing network infrastucture compute the lowest cost links that must be added for the network to support the worst case traffic matrix
- how to add links to an existing network infrastucture in a way that the can serve the dual purpose of worst case traffic matrix provisioning and resiliency under link cuts?
- how to implement the desired routing patterns using the current routing protocols (BGP/IGP/OSPF)?

6 Conclusions

We have argued that given the evolution path of internet traffic, higher levels of reliability will be required. As such admission control schemes which refuse connections are no longer feasible. At the same time we give evidence that capacity overprovisioning with a high probabilistic guarantee of delivery for $n \times n$ traffic is already in place in the internet, though not generally recognized. We also observed that in other network settings such large capacity has been openly, purposely deployed with the full acceptance of theory and practice. We noted that "overprovisioned" capacity can be put to other uses as others have shown [19], and CO is more efficient than generally believed. We give general bounds on the amount of traffic that is required for service guarantees and we term this *rightprovisioning* the network. Lastly we argued that there is still need for QoS and AC policies at the network edge.

References

1. Armitage, G.J.: Revisiting IP QoS: Why do we care, what have we learned. ACM 2003 RIPOS Workshop Report, ACM SIGCOMM Comp. Comm. Rev. vol. 33(5) (October 2003)
2. Arora, S., Leighton, F.T., Maggs, B.M.: On-line Algorithms for Path Selection in a Nonblocking Network. In: Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, pp. 149–158 (May 1990)

3. Ash, G.R.: Dynamic Routing in Telecommunications Networks. McGraw-Hill, New York (1998)
4. Atkinson, R.: QoS vs Bandwidth Overprovisioning. End-to-End mailing list (April 2001)
5. Beneš, V.E.: Optimal rearrangeable multistage connecting networks. Bell System Technical Journal 43, 1641–1656 (1964)
6. Bhagat, S.: QoS: Solution Waiting for a Problem, position paper, Dept of Comp. Sci. Rutgers University
7. Casner, S., Alaettinoglu, C., Kuan, C.-C.: A Fine-Grained View of High-Performance Networking, NANOG 22, http://www.nanog.org/mtg-0105/casner.html
8. Crowcroft, J., Hand, S., Mortier, R., Roscoe, T., Warfield, A.: QoS's Downfall: At the bottom, or not at all! In: Proceedings of the Workshop on Revisiting IP QoS (RIPQoS), at ACM SIGCOMM 2003, August 27, 2003, Karlsruhe, Germany (2003)
9. Dellamonica Jr., D., Kohayakawa, Y.: An algorithmic Friedman–Pippenger theorem on tree embeddings and applications to routing. In: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, pp. 1038–1044 (2006)
10. Feldman, P., Friedman, J., Pippenger, N.: Non-blocking networks. In: Proceedings of the 18th Annual ACM Symposium on Theory of Computing, pp. 247–254 (May 1986)
11. Feldman, P., Friedman, J., Pippenger, N.: Wide-sense nonblocking networks. SIAM Journal of Discrete Mathematics 1, 158–173 (1988)
12. Fraleigh, C., Tobagi, F., Diot, C.: Provisioning IP Backbone Networks to Support Latency Sensitive Traffic. In: Proceedings of IEEE Infocom 2003, San Francisco, USA (2003)
13. Gibbens, R., Kelly, F.: Resource pricing and the evolution of congestion control. Automatica 35 (1999)
14. Gupta, A., Kleinberg, J.M., Kumar, A., Rastogi, R., Yener, B.: Provisioning a virtual private network: a network design problem for multicommodity flow. In: Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, pp. 389–398 (2001)
15. Van Jacobson: A New View of Networking, Google Tech Talk (2007)
16. Keshav, S.: An Engineering Approach to Computer Networking. Addison-Wesley, Reading
17. Keslassy, I., Chang, C.-S., McKeown, N., Lee, D.-S.: Optimal load-balancing. In: Proceedings of IEEE Infocom, pp. 1712–1722 (2005)
18. Martin, R., Menth, M., Charzinski, J.: Comparison of Border-to-Border Budget Based Network Admission Control and Capacity Overprovisioning. In: Boutaba, R., Almeroth, K.C., Puigjaner, R., Shen, S., Black, J.P. (eds.) NETWORKING 2005. LNCS, vol. 3462, pp. 1056–1068. Springer, Heidelberg (2005)
19. Menth, M., Martin, R., Charzinski, J.: Capacity Overprovisioning for Networks with Resilience Requirements. In: Proceedings of SIGCOMM 2006, September 11-15 (2006)
20. Milbrandt, J., Menth, M., Junker, J.: Experience-Based Admission Control in the Presence of Traffic Changes. Journal of Communications 2(1) (January 2007)
21. Odlyzko, A.: Data Networks are Lightly Utilized, and will Stay that Way. The Review of Network Economics 2 (2003)
22. Olifer, N., Olifer, V.: Computer Networks: Principles, Technologies and Protocols for Network Design. John Wiley & Sons, Chichester (2005)

23. Parekh, A.: Why there is no QoS and what to do about it. In: Jeffay, K., Stoica, I., Wehrle, K. (eds.) IWQoS 2003. LNCS, vol. 2707, Springer, Heidelberg (2003)
24. Pippenger, N.: Information Theory and the Complexity of Switching Networks. In: Proceedings of FOCS, pp. 113–118 (1975)
25. Pippenger, N.: On Rearrangeable and Non-Blocking Switching Networks. Journal of Computer Systems and Sciences 17(2), 145–162 (1978)
26. Pippenger, N.: Telephone Switching Networks. In: Proceedings of Symposia in Applied Mathematics, vol. 26, pp. 101–133 (1982)
27. Pippenger, N., Valiant, L.G.: Shifting Graphs and Their Applications. Journal of the ACM 23(3), 423–432 (1976)
28. Pippenger, N., Yao, A.C.: Rearrange-able networks with limited depth. SIAM Journal Algebraic Discrete Methods 3(4), 411–417 (1982)
29. Prasad, R.S., Winzer, P.J., Borst, S., Thottan, M.K.: Queuing Delays in Randomized Load Balanced Networks. In: Proceedings of IEEE Infocom 2007 (2007)
30. Rui, Z.-S., McKeown, N.: Designing a Predictable Internet Backbone with Valiant Load-Balancing. In: de Meer, H., Bhatti, N. (eds.) IWQoS 2005. LNCS, vol. 3552, pp. 178–192. Springer, Heidelberg (2005)
31. Shepherd, F.B., Winzer, P.J.: Selective randomized load balancing and mesh networks with changing demands. J. Opt. Netw. 5, 320–339 (2006)
32. Telkamp, T.: Traffic Characteristics and Network Planning. In: ISMA 2002 (October 7-11, 2002)
33. Valiant, L.G., Brebner, G.J.: Universal Schemes for Parallel Communication STOC 1981, pp. 263–277 (1981)
34. Valiant, L.G.: A Scheme for Fast Parallel Communication. SIAM J. Comput. 11(2), 350–361 (1982)