Attaining VoIP–grade QoS via Deflection: A Buffer Space Tradeoff Study

André Muezerie[1]*, Ioanis Nikolaidis[2], and Pawel Gburzyński[2]

[1] Physics Institute of São Carlos - University of São Paulo
Caixa Postal 369, São Carlos - SP, 13560-970 Brazil
andremuz@if.sc.usp.br
[2] Department of Computing Science - University of Alberta
Edmonton, Alberta, Canada T6G 2H1
{yannis, pawel}@cs.ualberta.ca

Abstract. We consider the issue of QoS (Quality of Service) in a network catering to isochronous VoIP (Voice over IP) sessions. Our simulation experiments strongly suggest that the single–path paradigm of implementing network–layer virtual circuits neither yields the best utilization of network resources (buffer space) nor is it able to provide the best end-to-end service in terms of loss rate and delay. Instead, it turns out that the dynamic exploration of alternative paths at the packet level with small buffer spaces at the routers results in a more efficient and economical delivery.

Keywords: routing, multiple path routing, deflection, VoIP, QoS

1 Introduction

Voice-over-IP (VoIP) is quickly becoming a popular application that demonstrates how packet–switched networks can compete in providing services historically associated with circuit–switched reservation–based networks. VoIP sessions pose a wide range of Quality of Service (QoS) requirements, such as low delay, jitter and packet loss, but the anticipated scale of VoIP suggests that fine–grained admission control of the data stream, i.e., at the level of individual calls, would result in a substantial overhead and should be avoided.

A design principle that is "common wisdom" of many QoS solutions is that the most QoS friendly implementation of an end-to-end session should involve a network-layer virtual circuit, whereby all packets of the session follow the same path from source to destination. Intuitively, the deterministic character of such a connection makes it easier to set aside the right amount of resources at every intermediate node and predict what is going to happen when several virtual circuits cross at the same router. Consequently, most work on QoS-driven resource allocation has focused on path selection algorithms [1, 2], assuming that,

* André Muezerie is grateful to CAPES for the generous support of his study leave.

once selected, the (single) path will be followed by all packets of the session. This approach equates a transport-layer session with a network-layer virtual circuit, even if (as in the Internet) the network-layer virtual circuit is not explicit. Thus, it does not exploit degrees of freedom afforded by datagram routing.

According to recent results [3], the usefulness of buffering in the core routers for the sake of congestion–relief is highly debatable. In particular, [3] demonstrates that even TCP can live comfortably with a 99% reduction of the buffer space at the routers. These observations indirectly suggest that the intuitions behind the contemporary prevalent trends in QoS provisioning via core router buffer dimensioning are not necessarily correct.

We demonstrate that deflection routing [4] is a viable alternative to providing QoS, due to deflection's flexible ("as needed") use of alternate routing paths compared to routing across pre-established source-destination paths. The results also indicate that real–time isochronous traffic sessions, such as VoIP, *do not* benefit from buffering in the core.

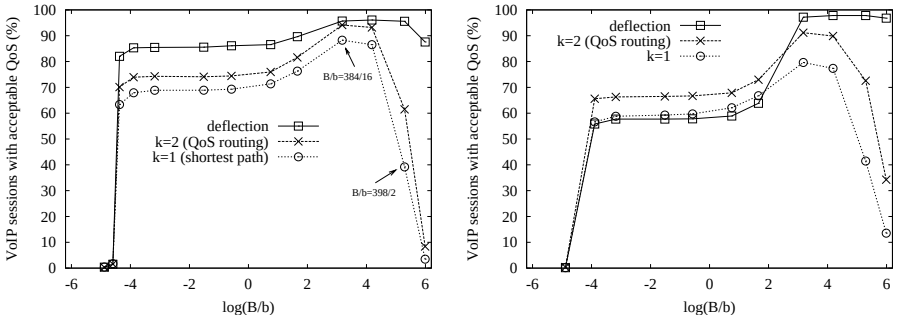
2 The Model

Three basic routing models are considered. In the traditional single path model, packets are forwarded along a single shortest path connecting the source to the destination, or dropped if the output buffer is full. As a variant of traditional routing (and a representative of alternative path routing schemes), we consider a simple strategy whereby the source is aware of a second-best route (disjoint from the first at least on the first hop) and utilizes it whenever the first route appears congested. With deflection routing, each router along the source-destination path is allowed to forward any packet over any of its output links, which are ordered according to the length of the shortest path to the destination offered by the next hop node. The packet is queued for forwarding on the best (i.e., shortest path) link whose buffer is not full.

Each node is capable of acting as a router and a host at the same time, i.e., it can be a source and/or destination of a VoIP session. To express the buffer space tradeoff, we represent the total amount of buffer space in the network as $B + b$, where B denotes the space to be used for reassembly buffers and b stands for the amount of storage equally partitioned among the output links of all routers. The varying ratio B/b determines the adjustable balance between the two categories of storage. While each node implements both types of functionalities, its router and host operations are isolated and they use two separate buffer pools.

3 Results

In the simulated networks, all links have the same bandwidth of 1 Mbps and an equivalent *bandwidth* $\times$ *delay* capacity for 2 packets (around 640 km). We set the *playout threshold*, i.e., the delay elapsing between the reception of the first packet of a talkspurt and the actual commencement of the playout to the equivalent of four packet transmissions. We define "acceptable QoS" as a loss



(a) 6x5 rectangular topology, 350 simultaneous sessions

(b) triangular topology, 100 simultaneous sessions

Fig. 1. Fraction of sessions with acceptable QoS. VoIP sessions: random source/dest. pairs, bursty talkspurts (mean 1.2 s) and silences (mean 1.8 s), packetization every 20ms, 200 bytes (160 voice + 40 headers)

rate below 5%, which is tolerated by most audio decoders with no appreciable degradation in voice quality.

Interestingly, it turns out that extending the reassembly buffer at the destinations in proportion to the reduction of the buffer space at the routers is an over-compensation. Consider Fig. 1(a), showing the percentage of sessions with acceptable QoS regardless of the actual size of the reassembly buffer at the destination for a rectangular topology (similar to the bi-directional Manhattan Street Network [5], except for the absence of the links closing the rectangle onto a torus). The curves $k = 1$ correspond to the classic single path routing and $k = 2$ represents the case where the source has two distinct shortest paths to the destination.

With deflection, we observe a trend of improving QoS as buffer space is removed from the routers and assigned to the destinations. In traditional circuit-style routing, in the absence of deflection, the routers require storing the packets that cannot be immediately forwarded on the single congested output link. Figure 1(a) in fact demonstrates that the benefits (flexibility) of deflection more than compensate for the reduced buffer space at the routers. At the area of large B/b, where the traditional routing breaks down, deflection is still able to provide reasonable service to a large population of sessions. In agreement with the observations made in [5], deflection can take advantage of some limited router buffer space. In comparison to a completely buffer-less case, small buffers bring considerable improvement, while large buffers yield no appreciable gains.

Even small networks can benefit from deflection. In the particular topology shown in Fig. 2, the best (shortest) routes for all nodes pass through the even nodes ("hotspots"), and when the links between these nodes become congested, they aggressively discard packets under classical routing (Fig. 1(b)). With de-

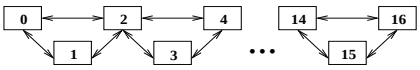


Fig. 2. Triangular topology

flection the links between even and odd nodes are used to distribute the load and achieve higher throughput and therefore, lower losses. At higher loads the effect is even more pronounced.

4 Conclusions

We analyzed the QoS perceived by concurrent VoIP sessions in networks operating under three routing policies: 1) classic shortest path best effort routing, 2) QoS routing with two alternative shortest paths, and 3) asynchronous deflection. Our experiments show that, especially under high loads, deflection performs at least as well as the other two alternatives, and frequently outperforms them with respect to the overall QoS measures perceived by the application.

The better load distribution over the network brought about by deflection translates into better buffer space utilization as a global network resource, lower losses and end-to-end delays. Contrary to common belief, deflection is not necessarily inferior (from the viewpoint of service guarantees) to the more deterministic alternatives involving pre-arranged predictable routes. Despite the difference in appearance, neither paradigm is in fact deterministic and predictable, and the application-level perception of its operation does not necessarily endorse determinism within the core.

References

1. Apostolopoulos, G., Guerin, R., Kamat, S., Tripathi, S.K.: Quality of service based routing: a performance perspective. In: Proceedings of the ACM SIGCOMM '98 conference on Applications, technologies, architectures, and protocols for computer communication, ACM Press (1998) 17–28
2. Xiao, W., Soong, B.H., Law, C.L., Guan, Y.L.: Evaluation of heuristic path selection algorithms for multi-constrained qos routing. In: IEEE International Conference on Networking, Sensing and Control. Volume 1. (2004) 112–116
3. Appenzeller, G., Keslassy, I., McKeown, N.: Sizing router buffers. In: ACM SIGCOMM'04, Portland, Oregon (2004)
4. Olesinski, W., Gburzynski, P.: Service guarantees in deflection networks. In: Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'01). (2001) 267–274
5. Borgonovo, F., Cadorin, E.: Locally-optimal deflection routing in the bidirectional Manhattan network. In: Proceedings of IEEE INFOCOM'90. (1990) 458–464