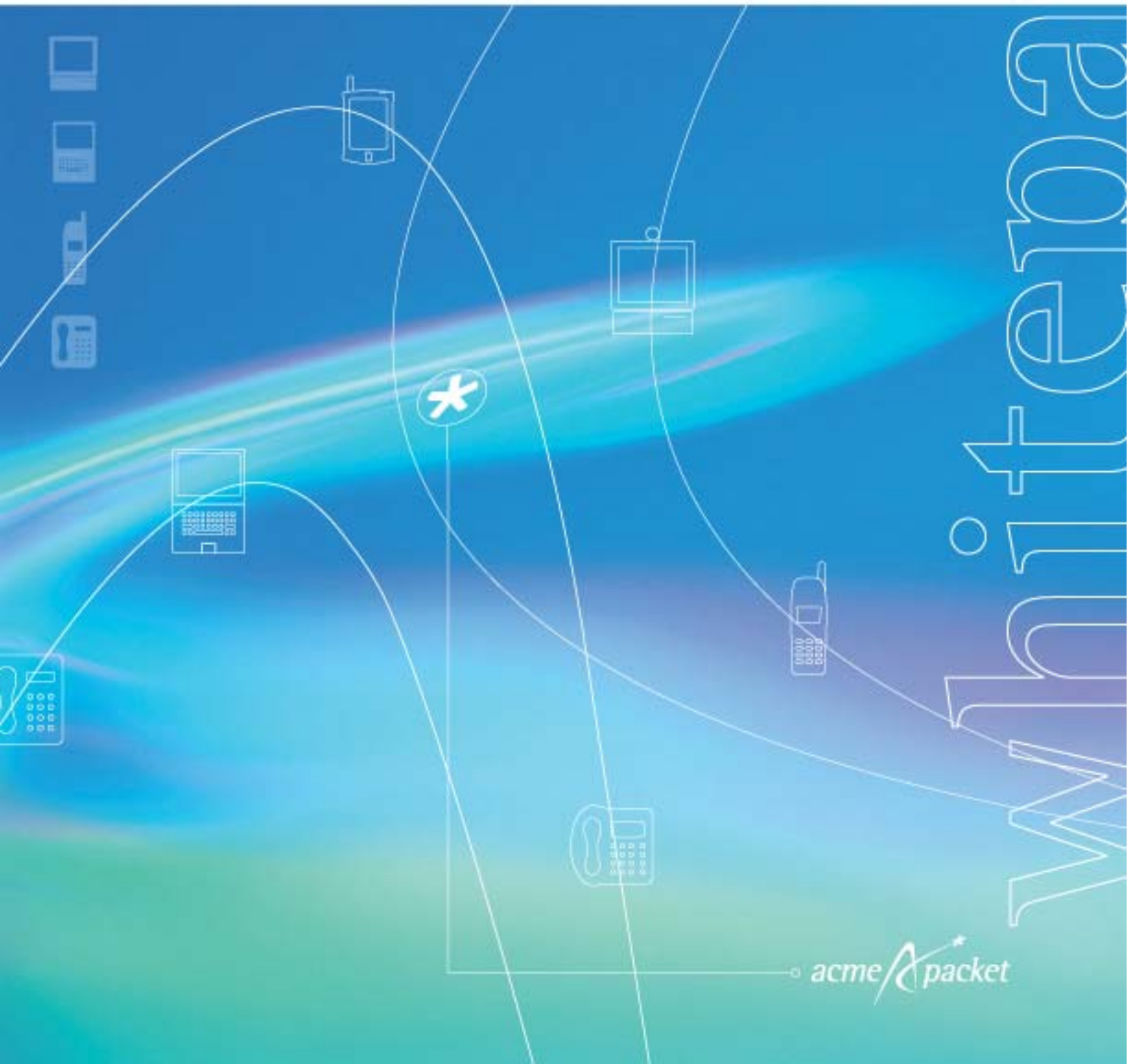
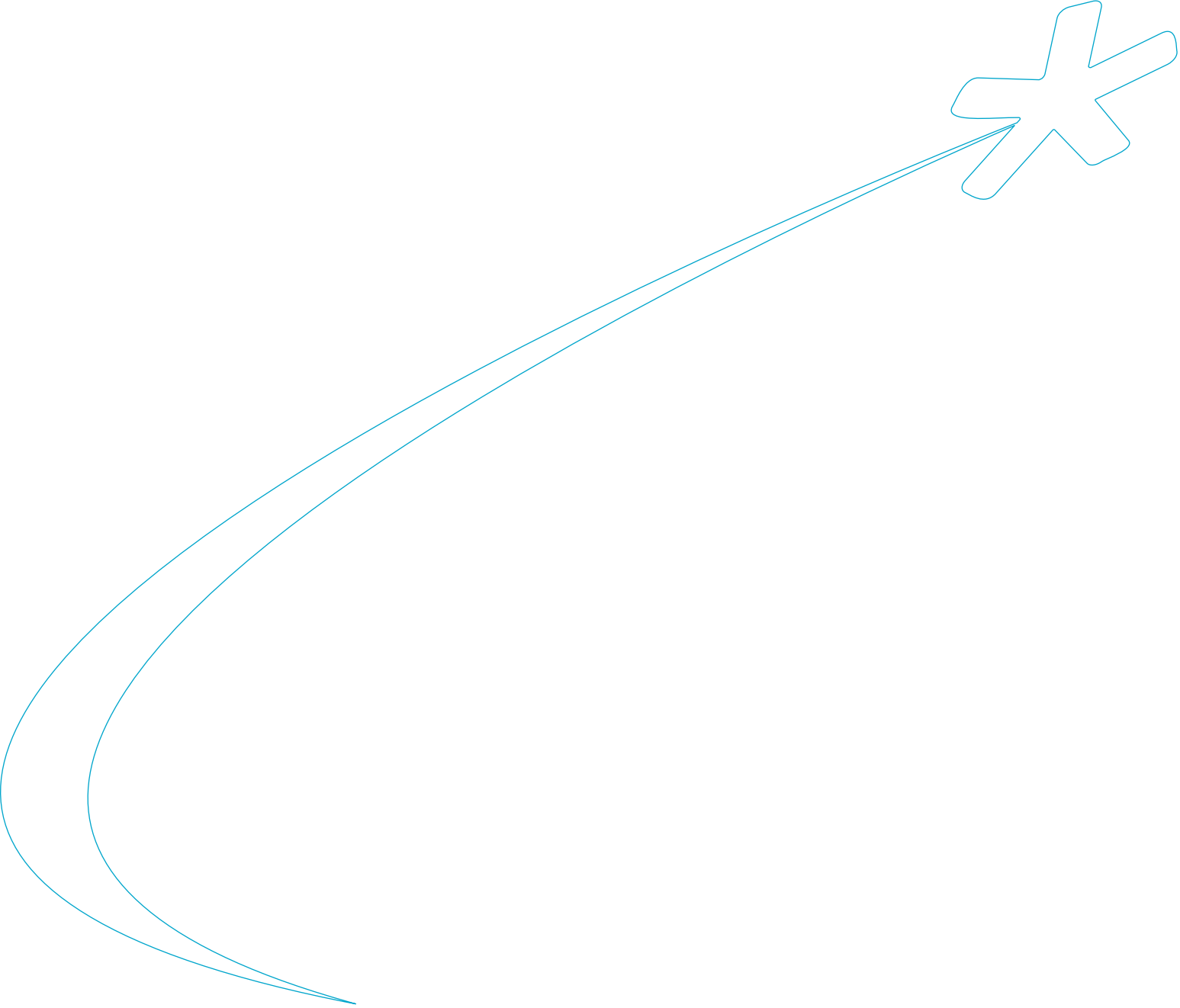
Session border controllers: Delivering interactive communications across IP network borders





Interactive communications over IP networks

How many times have you heard "IP networks don't make any money!" Probably way too many! Compared to the PSTN, IP networks are big zeroes in terms of financial appeal. Today, while data consumes more than half of network bandwidth, ordinary telephone calls generate about 80% of total earnings.

There definitely is money in interactive communication services - real-time, high quality voice and video communications between people. Businesses and consumers have been paying good money for these services for over 100 years.

Interactive communications over IP networks opens up several business opportunities:

- Transport only services
- HIP (hosted IP) voice services (aka IP Centrex or CLASS 5 services) including unified messaging, conferencing, etc.
- New services not possible in the PSTN like presence with instant calling or video conferencing from Windows XP PCs, multimedia customer care web sites, distance learning with real-time Q&A capabilities and others.

Regardless of opportunity, interactive communication over IP networks must be able to reach anyone, anywhere, anytime to maximize its value. To paraphrase Metcalfe's Law: the usefulness, or utility, of interactive communication equals the square of the number of users. New interactive communication services and applications, therefore, must ultimately span business and consumer, wired and wireless networks.

Consequently, simply building standalone voice, video and multimedia over IP network islands is not sufficient. They must be built and interconnected in a way that ensures security and peak performance end-to-end. (See Figure 1) Businesses and consumers will be satisfied with-and pay good money for-nothing less.

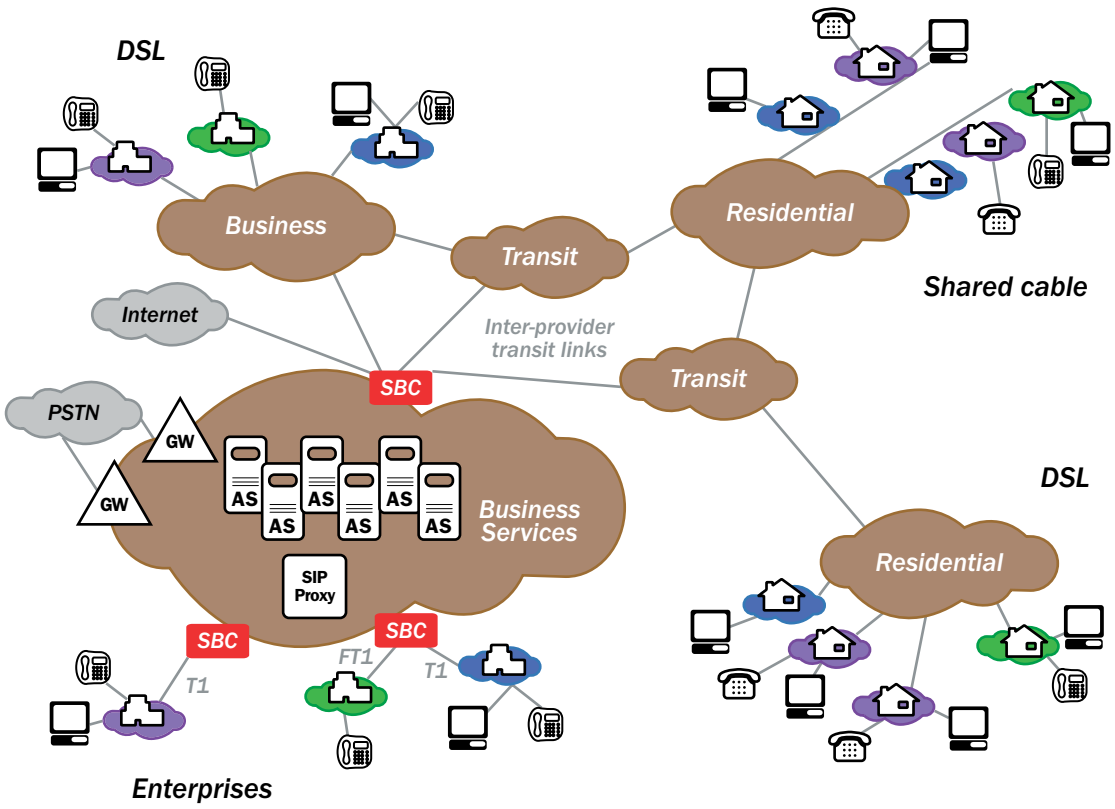


FIGURE 1: Using session border controllers in and among networks.

The new needs at the edge

Connecting even just two IP networks, such as an enterprise and a provider's network, introduces new edge requirements in three major areas - security, service assurance and law enforcement. Because these requirements cannot be satisfied by existing products, they have spawned a new product category called session border controllers (SBCs).

In case there is any confusion with this acronym and the US ILEC, let us clear that up by decomposing the term and understanding the role of a SBC:

- "*Session*" - any real-time, interactive voice, video or multimedia communication using layer 5 IP session signaling protocols such as SIP, H.323, MGCP or Megaco/H.248
- "*Border*" - any IP-IP network border such as those between service provider and customer/subscriber or between two service providers
- "*Control*" - functions that provide security to protect service infrastructure and customer and supplier identities, and service assurance to guarantee SLAs, maximize revenue and minimize costs

Understanding IP interactive communications

To understand the need for SBCs, it's critical to understand how IP-based interactive communications actually works. Every call or session includes three sets of bidirectional communications flows between two endpoints (see Figure 2):

1. Session signaling messages such as SIP, H.323, or MGCP (layer 5 protocols in the OSI model) in TCP packets which are used to initiate, monitor, modify and terminate a session
2. Media streams using RTP in UDP packets which contain the digitized and packetized voice and/or video bits
3. Media control messages using RTCP in UDP packets which contain latency and jitter information for the session

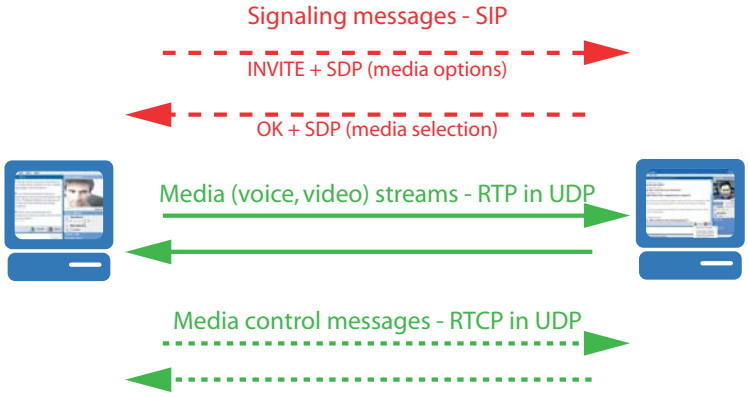


FIGURE 2: Interactive communications on IP networks - how it works

Before a session can begin, it must be set-up using a signaling protocol. The signaling protocol is used to establish a virtual connection between the participants' endpoint devices and negotiate the codec and the IP ports that will be used for the session's media streams and control messages. SIP endpoints use SDP to describe their parameters in this negotiation. Different codecs are required for voice vs. video and offer trade-offs between quality and bandwidth efficiency. For example, G.711 supports "toll quality" voice and requires 64 Kbps bandwidth (not including IP headers) while G.729 requires only 8 Kbps but offers lower quality. Once the call is set-up, the RTP media streams and RTCP control messages flow in both directions. The signaling messages typically use well-known IP ports such as 5060 for SIP. The IP ports for the RTP media streams and RTCP messages, however, are dynamically assigned by each endpoint.

The new edge requirements are extensive, complex, and require tight integration between signaling and media control. Today's products, while necessary, are completely insufficient in terms of providing the required control functions.

Today's installed layer 3 or 4-aware firewall or router can't control the media streams in terms of access control, network address translations, routing or QoS marking since they do not have any signaling intelligence that allows them to identify the IP ports being using for an interactive communication session between two endpoints. Similarly, without signaling intelligence, a firewall or router can't accept/block/route sessions based upon layer 5 information including telephone number or SIP URI, or gracefully reject a call by generating a "network busy" signal.

Softswitches, SIP proxies, H.323 gatekeepers and media gateway controllers, on the other hand, are all focused on signaling and call set-up. Some can apply basic routing policies for signaling messages based upon to/from telephone number or SIP URI and time-of-day. But these products can't control the media streams.

QoS monitoring devices report on service quality, which makes them useful for SLA reporting, but there is no way this information can be used to perform real-time admission control and routing decisions.

Today's products.

Necessary but insufficient.

SBCs sit at the edge of the provider's network and complement existing routers. They perform required control functions by tightly integrating session signaling and media control. SBCs operate as SIP back-to-back user agents, MGCP proxy/NATs, and/or H.323 back-to-back gateways/gatekeepers. What all of this means is that SBCs are the source and destination for all signaling messages and media streams coming into and leaving the provider's network.

Let us now take a more detailed look at the new requirements in the areas of security, service assurance and law enforcement, and find out how SBCs satisfy these requirements.

Security

The security agenda is driven by the fact that in today's world no one trusts anyone else - especially their IP network. There are two basic vantage points - the service provider and the end user customer.

All businesses and an ever-increasing number of residential users use firewalls with network address translation (NAT) to protect their IP networks and computers from external attack. A firewall only allows traffic into a network if it has been requested from the inside and presents a single global IP address to the outside world for all the PCs and phones behind it. While this works fine for requests to web, email, IM and other servers, it is a huge roadblock to inbound signaling and media for voice, video or other peer-peer communications.

SBCs support a hosted NAT traversal feature that eliminates this roadblock without any new premise-based hardware or software and also without any firewall configuration changes, preserving existing security policies. This feature exploits periodic endpoint registrations to keep a signaling port open in the firewall for incoming signaling messages. As registrations pass through the SBC, it maps the layer 3 IP address/port on the firewall to the layer 5 user name/phone number behind the firewall. When an incoming signaling message is received, the SBC sends it to the right address and port on the firewall for the callee. During call set-up, the ports for the bi-directional media flows are dynamically established. Since the media flows also pass through the SBC, it can identify the IP address/port on the firewall used for the outgoing media coming from a user name/phone number. It then uses that same firewall IP address/port for sending the incoming media to the correct user name/phone number behind the firewall. For additional security and control, the firewall can be configured to only allow incoming traffic from the IP address of the SBC.

Similarly, a service provider must only allow authorized users access to their high quality services while protecting internal service infrastructure from denial of service attacks. This

infrastructure (see Figure 1) includes softswitches, SIP proxies or H.323 gatekeepers for session authorization and routing; application and media servers for unified messaging, conferencing, presence and instant communications, etc; and media gateways for providing PSTN connectivity.

The geographically distributed nature of this infrastructure makes the security problem even more difficult. There is also a requirement to conceal valuable route information from inquisitive customers and competitors. If a provider is providing transit or termination services through another provider, a knowledgeable large enterprise customer might approach that provider directly for a better price.

SBCs work with the provider's signaling infrastructure to perform access control based upon layer 5 signaling messages to support user mobility, not layer 3 IP addresses used by firewalls or routers. For authorized communications, SBCs let the media streams into the network by opening and closing firewall pinholes. SBCs hide network topology by performing NAPT (network address and port translations) on all signaling and media IP packets. However, layer 3-only NAPT is simply not enough. Internal IP addresses can also be exposed in signaling messages, including error messages. Consequently, signaling and error messages are inspected by SBCs for embedded IP addresses and rewritten if present. These layer 3 and 5 NAPT features can also be used to preserve IP addresses by enabling the use of private addresses for CPE (customer premise equipment).

To protect against infrastructure overloads, incoming signaling messages may also have to be intelligently throttled. If a softswitch can only handle 50 calls per second before it keels over, the SBC must gracefully reject new call requests when activity reaches this threshold. SBCs must also be self-resilient to IP DoS attacks and perform these security functions at gigabit Ethernet wirespeed with sub-millisecond latency in order to minimize end-to-end call set-up and media stream latency.

Service assurance

The service assurance agenda divides into two major areas. The first area, SLA assurance, is concerned with guaranteeing session capacity and quality for customers. The second area, revenue and profit assurance, is focused on maximizing service provider revenue and minimizing costs.

The biggest SLA assurance challenge today entails converging premium revenue-generating voice, video and multimedia with data traffic - email, IM, Internet and corporate data applications - on the constrained and oversubscribed access links that connect enterprise or residential customer locations. These access links (See Figure 1) include low bandwidth T1 or DSL connections or the shared bandwidth cable HFC network. None of today's products or IP QoS mechanisms (including DiffServ, MPLS and RSVP) have the capability to understand access link capacity and utilization and make call admission control decisions based upon that intelligence.

Admission control policies implemented at the signaling level within the SBC can guarantee the total number of calls (measured on a bandwidth basis by looking at the codec), the number by type of call (voice vs. video), and the ability to make pre-emptive calls such as emergency 911 calls. If the access link is at capacity, new call set-up requests will be rejected (except for that 911 call). Adding just one more call will deteriorate the quality of each and every active call.

These policies are configurable based upon the bandwidth of the access link and the customer's requirements so that some amount of data traffic can always get through. In DSL and frame access networks, these policies must also be established for aggregated links between DSLAMs or frame switches and the edge router within the POP. Because they participate in both signaling messages and media streams at the edge of the network, SBCs don't have to rely on mapping endpoint IP addresses to access links. This scheme, offered by some sofswitches and SIP proxies, does not scale and becomes completely unmanageable. SBCs can also easily identify calls between endpoints within a single customer site and does not count that call against the call bandwidth limits for the access link.

Ensuring call quality requires guiding routers on both ends of the access link to correctly prioritize traffic. Any accepted session must be given top priority. Since the SBC is the destination for all signaling and media streams, customer access routers and the provider's edge routers can easily prioritize authorized session traffic based upon the IP address of the SBC. No CPE is required. If the access link ever becomes congested, the router will only drop data packets. This will not cause problems since the TCP-based date packets will be retransmitted.

In cable networks, DQOS, a layer 2 mechanism, manages the shared bandwidth in the HFC access network. DQOS utilizes a reservation-based approach between the headend CMTS managing the bandwidth pool and the MTA at the subscriber's location. The problem, howev-

er, is these standards today only support MGCP. SBCs with a communications interface to the CMTS can enable high-quality, revenue generating SIP-based communication including instant calling and video conferencing from Windows XP PCs.

These admission control issues also apply to transit links between providers or administrative domains within a single carrier. Because every transit link has a finite capacity, the number of active sessions must be actively managed to prevent that 'one more call' from deteriorating the quality for all calls.

For optimal routing across the backbone network, SBCs can explicitly assign QoS markings - ToS bits, DiffServ code points or MPLS labels - off-loading this task from over-worked edge routers. Alternatively, the router can easily do this since all signaling and media traffic is going to/coming from the SBC.

Explicit QoS marking also protects against QoS theft by eliminating the requirement to trust any packet markings coming from another network. Otherwise, any "enterprising" network administrator or sophisticated user could set the ToS bit markings on IP endpoints to steal QoS from the provider. Only media packets authorized by the service provider's signaling infrastructure and allowed into the network by the SBC will be given high priority across the backbone.

SBCs also protect against bandwidth theft by policing bandwidth usage. This feature ensures, for example, that a low-bandwidth voice call between two SIP endpoints cannot autonomously turn into a high-bandwidth video call without explicit authorization (e.g. paying money). The SBC will rate limit the media streams based upon the codec specified in the SDP.

Real-time QoS reporting capabilities are required for SLA reporting, problem alerting and isolation, and session admission control and routing. SBC can measure end-to-end jitter, latency and packet loss by observing actual RTP and RTCP packets and generating RTCP packets if they are not generated by the endpoints. Quality can also be measured and problems isolated by network domain in terms of the origination network(s), the service provider network using SBCs and the termination network(s).

In any network topology where there are two paths to a destination, QoS and cost-based routing and admission control is also required to assure SLA performance and profits. If one network is providing poor quality, signaling messages and media streams must be routed to the other network. Similarly, if both networks provide acceptable QoS but have different

costs for transit or termination, the less expensive network should be selected. Routing the signaling messages entails simply selecting the next-hop session routing element in the preferred network. However, layer 3 routers cannot be trusted to route the media streams correctly. SBCs perform "packet steering" by sending the media streams to the SBC at the egress point to the preferred network.

Simultaneous support of multiple signaling protocols and interworking between protocols, protocol versions or configurations is another function performed by some SBCs, maximizing network reach and revenues while minimizing costs. By supporting SIP-H.323 interworking, for example, a provider can build one SIP service backbone yet support both SIP and H.323 customers or visa versa. Similarly, H.323 version and configuration interworking enables the provider to build a backbone supporting just one H.323 version and configuration while supporting customers using different H.323 versions or configurations.

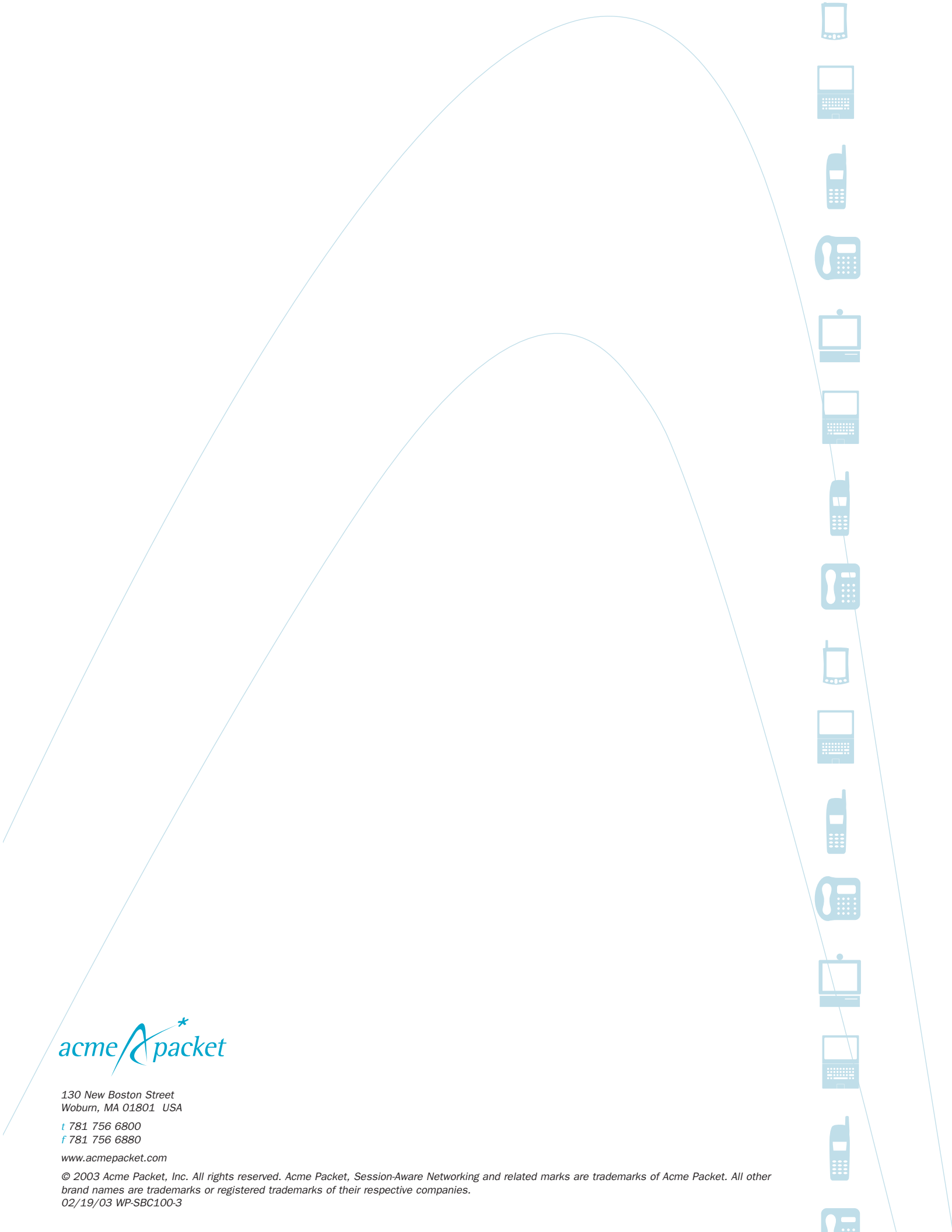
Lastly, session accounting for capacity planning or billing purposes is required for all or just selected session types (e.g. video only). Call detail records (CDRs) are produced, which identify the caller, callee, call duration, time of day, and QoS metrics for the call. Session timers within the SBC end sessions that do not terminate correctly. The SBC will generate a termination message and associated CDR entry if no signaling message or media packet is received within a configurable time window. This eliminates the huge customer care costs associated with reconciling accounting records for open calls.

Law enforcement

It is inevitable that IP networks supporting interactive communications must support lawful intercept capabilities (e.g. wiretapping) as required by governments around the world for the PSTN. For example, the US Communications Assistance for Law Enforcement Act (CALEA) demands that it must be possible to replicate and route target calls, including both signaling messages and media streams, to multiple law enforcement agencies (LEA) simultaneously and transparently. Transparently means that neither the caller nor callee should be aware of the wiretap, nor should any LEA be aware that other LEAs are involved. Since SBCs actively filter both signaling messages and media streams with negligible latency, they are very logical devices to implement this wiretapping capability.

The Net-Net: session border controllers - a smart investment

Session border controllers are an attractive and wise investment for service providers delivering revenue-generating interactive communications across IP network borders securely with premium performance. They provide the financial appeal currently missing in IP networks. Session border controllers - a wise investment at the edge



130 New Boston Street
Woburn, MA 01801 USA

t 781 756 6800
f 781 756 6880

www.acmepacket.com

02/19/03 WP-SBC100-3