♦ Ethernet Aggregation and Core Network Models for Efficient and Reliable IPTV Services

Christian Hermsmeyer, Enrique Hernandez-Valencia, Dieter Stoll, and Oliver Tamm

With the growing interest on wireline network architectures for residential triple-play and business Ethernet services there is a renewed demand for efficient and reliable packet-based transport capabilities between the content providers and the end users. Voice and data traffic carried over a variety of access technologies is collected via technology-specific access networks (e.g., digital subscriber line [xDSL], passive optical network [xPON], and wireless fidelity [WiFi]). Metro and core networks need to aggregate the various user flows from different access network nodes and provide scalable and cost-effective distribution of various flow types (e.g., Internet access, voice, video on demand, and broadcast TV services) to the relevant service access points. Varying quality of service and resiliency requirements for these services are being reflected in a new breed of converged Ethernet and optical network elements with capabilities to interwork the bearer-planes of these two networking technologies seamlessly. Network elements based on Ethernet/Optical converged technology are able to select the most fitting mechanisms from each networking technology to meet the transport requirements for each individual service demand better while providing significantly enhanced implementation and operational efficiencies. This paper discusses network architecture models and network elements addressing these goals. © 2007 Alcatel-Lucent.

Introduction

The traditional public switched network (PSN) was designed to transport fixed-rate telephony and private line traffic in a reliable and resilient manner. Over the last decade, service providers have experienced tremendous growth in residential and business data services with the aggregate demand for packet-oriented traffic now exceeding circuit-oriented traffic. Bell Labs internal projections forecast new residential services such as Internet Protocol–(IP)-oriented broadcast video and video on demand, which promise an additional 5- to 10-fold increase in the amount of packet-oriented traffic in the metro/regional transport networks within the next decade [1]. Unlike Internet data traffic, broadcast and near-real-time video applications call for very strict packet loss, packet delay, and service resiliency performance objectives. For service providers who want to expand their offerings beyond vanilla Internet access and IP-virtual private network (IP-VPN) services, the need to compete and differentiate from traditional cable-based data and new video distribution offerings

Bell Labs Technical Journal 12(1), 57–76 (2007) © 2007 Alcatel-Lucent. Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com). • DOI: 10.1002/bltj.20217

Panel 1. Abbreviations, Acronyms, and Terms

3GPP—3rd Generation Partnership Project
ADM—Add/drop multiplexer
ADSL—Asynchronous digital subscriber line
BE—Best effort
BER—Bit error rate
BTV—Broadcast TV
CAPEX—Capital expenditure
COPS—Common Open Policy Service
CoS—Class of service
D&C—Drop and continue
DVB-ASI—Digital video broadcast asynchronous serial interface
DSL—Digital subscriber line
DSLAM—Digital subscriber line access multiplexer
DWDM—Dense wavelength division multiplexing
EPON—Ethernet passive optical network
ETSI—European Telecommunications Standards Institute
FRR—Fast re-route
GE—Gigabit Ethernet
GPON—Gigabit passive optical network
HDTV—High-definition TV
IETF—Internet Engineering Task Force
IGMP—IP Group Multicast Protocol
IMS—IP Multimedia Subsystem
I/O—Input/output
IP—Internet Protocol
IPTV—Internet Protocol television
ITEA—Information Technology for European Advancement
ITU—International Telecommunication Union
ITU-T—ITU-Telecommunication Standardization Sector
L1—Layer 1
L2—Layer 2
L3—Layer 3
LAG—Link aggregation group
LAN—Local area network
LSP—Label switched path
MAN—Metropolitan area network
MPEG—Motion Picture Experts Group
MPLS—Multiprotocol label switching
MTV—Music Television
NACF—Network attachment control function
NAT—Network address translation
NAPT—Network address port translation
NE—Network element
NG—Next-generation
NGN—Next-generation network
NVoD—Near real-time video on demand
OAM—Operations, administration, and maintenance
OAM&P—Operations, administration, maintenance and provisioning
OPEX—Operational expenditure
OTN—Optical transport network
PD-FE—Policy decision functional entity
PE-FE—Policy enforcement functional entity
PIM—Protocol independent multicast
PON—Passive optical network
PSN—Public switched network
QoS—Quality of service
RACF—Resource and admission control function
ROADM—Reconfigurable optical add/drop multiplexer
RPR—Resilient packet ring
RTP—Real-time Transport Protocol
RTSP—Real Time Streaming Protocol
SDH—Synchronous digital hierarchy
SDTV—Standard definition TV
SHE—Super headend
SLA—Service level agreement
SONET—Synchronous optical network
TDM—Time division multiplexing
TRC-FE—Traffic resource control functional entity
TV—Television
UDP—User Datagram Protocol
VDSL—Very high-speed DSL
VHO—Video hub office
VLAN—Virtual LAN
VoD—Video on demand
VoIP—Voice over IP
VPLS—Virtual private line service
VPN—Virtual private network
VSO—Video switching office
WiFi—Wireless fidelity

effectively further increases the necessity to control and arbitrate access to network resources, and hence, to improve the quality of the user experience.

In considering the highly diverse combination of service attributes [2, 8] and the wide range of performance expectations for such bundled services, all these features together make for an extremely challenging mix of requirements on the transport network elements. Moreover, new services must fit into the preexisting operational models in support of well-established private line, business VPNs [7], and circuit-based services as service providers strive to reduce capital as well as operational expenditures (CAPEX and OPEX).

Elaborating on the less well-understood IP-based TV applications it must be realized that when talking about "video," one needs to differentiate between real-time broadcast TV (BTV), near-real-time video on demand (NVoD), and video on demand (VoD) services [9]. While NVoD and VoD share many service attributes such as high bandwidth and point-to-point connectivity between video server(s) and end customer set-top boxes, BTV is dominated by point-to-multipoint multicast connectivity from a single head-end server to a potentially large number of clients/subscribers. In spite of this distinguishing factor, all three services have a very important attribute in common: very strict performance commitments in terms of packet loss, packet delay and delay variation, and reliability. Since, as further shown in **Table I**, video services consume significantly more bandwidth in the transport network than other services, e.g., Voice over Internet Protocol (VoIP), it is paramount that next-generation transport networks, and next-generation transport network elements, be optimized for these services.

The current generation of packet transport solutions have been optimized around unicast traffic patterns and

Table I. Traffic characteristics for wireline services.

Application		Bearer traffic characteristics				
		Flow type	Pattern	Upstream	Downstream	Resiliency needs
Residential	Internet access	2-way unicast	Bursty	64–512 Kbps	1–6 Mbps	Low
	VoIP	2-way unicast	Bi-rate	n x 8 Kbps	n x 8 kbps	Mod
	Broadcast video	1-way multicast	CBR	none[1]	20 Mbps	Very high
	VoD	2-way unicast	CBR	none	2–10 Mbps	High
Business	E-Line	2-way unicast	Bursty	10 * −n * 100 Mbps	10 * −n * 100 Mbps	By SLA
	E-LAN	N-way uni/ broadcast	Bursty	10 * −n * 100 Mbps	10 * −n * 100 Mbps	By SLA
	CES	2-way unicast	CBR	n * 1.5–2 Mbps	n * 1.5–2 Mbps	Very high
	Internet access	2-way unicast	Bursty	1–100 Mbps	1–100 Mbps	Low-mod
	VoIP	2-way unicast	Bursty	n * 8 kbps	n * 8 kbps	Mod-high
	IP-VPN access	N-way unicast	Bursty	10 * −n * 100 Mbps	10 * −n * 100 Mbps	Mod-high
	Backhaul/ wholesale	2-way unicast	Bursty	0.15–1 Gbps	0.15–1 Gbps	Very high

CBR—Constant bit rate
CES—Circuit emulation service
E-LAN—Ethernet LAN
E-Line—Ethernet line
IP—Internet Protocol
LAN—Local area network
SLA—Service level agreement
TV—Television
VoD—Video on demand
VoIP—Voice over IP
VPN—Virtual private network

[1]Upstream traffic consists mostly of control and management traffic (e.g., TV channel selection)

a comparably low level of traffic flooding. Performance expectations for real-time multicast traffic and the growing interest in applications with massive multicast requirements (e.g., 90% of the switching capacity) and with possibly thousands of multicast groups present a major network and system engineering challenge. Multistage multicast in a packet switching network element (NE) is a function that tests the limits of traditional packet forwarding and control performance because of the asymmetric nature of the traffic, high traffic flow fan out, and burstiness. Unless the forwarding and system control architecture of future NEs—which are responsible for performing recalculation of multicast trees and subsequent restoration—is designed around these specific capabilities, service scalability expectations may be hard to achieve.

In areas where awareness of individual multicast streams is not prerequisite to determine the multicast topology, e.g., if all multicast groups follow the same path, dictated by the physical extension and/or nature of the distribution network, then broadcasting in the optical domain offers a simpler and cost-efficient solution: a single optical splitter performs a 1:2 replication of an aggregated signal that can consist of thousands of multicast channels on a bit-by-bit level. Therefore, the idea of combining optical replication with layer 2– (L2)-based multicasting, and having a common control function supporting them, can be seen as a very attractive proposition to optimize the metro/regional portions of the distribution network for BTV flows. If in addition, the layer 2 functions can be used intelligently to segregate between service types at the packet layer and facilitate the mapping of service types into the virtual links so as to help steer service into the most appropriate dense wavelength division multiplexing (DWDM), time division multiplexing (TDM), or packet infrastructure channel, a more efficient and cost-optimized multiservice transport infrastructure can be realized.

The goal of this paper is to demonstrate that by integrating Optical and Ethernet networking technology—DWDM, optical transport network (OTN), and Ethernet—into a single, converged transport system, a highly optimized transport network architecture can be realized to support residential triple-play and business applications efficiently. Reflecting the dominant role that IPTV will play in the architecture of future multiservice networks, this paper will focus specifically on BTV. We will first discuss the technological background and follow with network design considerations and the needs and advantages of converged nodes. We conclude with a summary and closing remarks.

Technological Background

Today multiple networking technologies are available to address transport requirements for residential triple-play services. In order to derive the most appropriate network architecture for a particular technology selection and a given nodal traffic demand, the most stringent services in terms of offered load and performance constraints need to be identified and understood. Then, a reference network scenario must be defined and the available technologies assessed.

Demands to Be Satisfied for IPTV

Table I summarizes traffic characteristics for the most common residential and business applications in terms of connectivity, traffic flow pattern, bandwidth, and resiliency demands. From there the special role played by BTV stands out. Among key requirements, BTV demands support for a large aggregate of high-bandwidth constant bit-rate signals. In addition, it must be delivered with high-quality resilience—attributes it shares with VoD services—and it must be delivered simultaneously to a large number of subscribers. Thus, BTV demands a very efficient traffic distribution mechanism to keep the transport cost low.

Since a typical BTV service may include from tens of channels to upward of several hundred channels, the access aggregation network needs be able to satisfy a total bandwidth demand in the range of tens of megabits per second to several gigabits per second. Ideally, the full bandwidth demand must be transported as closely as possible to the point where customer-driven channel selection takes place. This follows from the desire to support fast interaction with the user for channel selection (zapping), which in turn demands that any associated control plane protocols are handled as closely as possible to the customer location. Since users will be able to select BTV channels from the same locally available set (which may consist

of nationwide, regional, and even local channels), the channels must be distributed, i.e., multicast, over the metro/regional network in the most efficient way. This demand for transport (CAPEX) efficiency, however, makes the network design a multilayer traffic optimization problem, i.e., one that requires steering the client flows in accordance with the server layer topology. The sections that follow provide further details.

Technological Alternatives

Before introducing the proposed reference network, we will address technological alternatives in terms of their capabilities to support the aforementioned network design requirements for BTV services. On this basis, the video flow distribution can best exploit the capabilities of the associated transport nodes. Available alternatives for the realization of BTV distribution network include an IP networking layer, multiprotocol label switching (MPLS)/virtual private line switching (VPLS) pseudo-wires, native Ethernet and/or RPR, or optical transmission. Strengths and weaknesses of these alternatives are discussed in the following.

The IP case. IPTV—as the name indicates—is based on IP multicast capabilities and addressing. In most realistic scenarios, only a fraction of the nodes in the video distribution network will be IP routers; the remaining ones will be just optical or OTN switches without native IP forwarding capabilities but with packet aggregation capabilites. For the multicast traffic distribution, IP routers use protocol-independent multicast (PIM) procedures to determine the connectivity between multicast capable routers and hosts. The main disadvantage in this approach is that the IP network is completely unaware of the underlying layer 1 (L1) transport network topology. Since the IP-routed network topology will in general form a meshed network, routers will appear as adjacent in the IP layer even though they are not on the transport layer. As an example, a routed IP mesh would be superimposed on a physical ring transport infrastructure. There, an IP router would send a separate copy of each TV channel to each router for which it maintained an IP adjacency. Given the underlying ring topology, this design yields wasted bandwidth in every span of the ring where multiple IP links overlap.

When fast transport resiliency is required, it must be realized via MPLS fast reroute (FRR) as no fast reroute capability exists in the IP layer. Although this method is capable of achieving protection times comparable to synchronous optical network (SONET)/synchronous digital hierarchy (SDH), it still cannot address lack of knowledge of the underlying physical transport topology. Note that, in general, an established protection path, even when guaranteed to be node and link disjoint to the protected path in the IP layer, cannot be guaranteed to be node and link disjoint in the optical layer. Even if this issue is resolved—by manual provisioning, or in the future through automated optical control plane procedures—mechanisms to optimize cross-layer multicast flows over label switched paths dynamically are far from mature and thus require vendor proprietary enhancements.

The MPLS/VPLS case. VPLS technology provides a mechanism to emulate a bridged Ethernet network to its clients. VPLS networks often stretch from the metro/regional core to the access network, where they are then terminated, and end user data traffic is then forwarded via native Ethernet. In a VPLS network, however, packet forwarding behaves similarly to a label switched network complemented with an IP routed control plane. The label switched network utilizes MPLS label switched paths (LSPs) as a traffic tunneling mechanism, i.e., pseudo-wires. For LSP setup and maintenance, as well as for traffic engineering of the aggregate traffic, the IP control plane may be invoked together with its traffic engineering extensions.

The native Ethernet portion of the access network will be discussed later. Therefore, we focus here on the VPLS portion of the network. This approach suffers from the same disadvantage as the pure IP network approach: it is oblivious to the underlying transport network topology. Replicated BTV channels must be forwarded over the full mesh of outer tunnels associated with VPLS pseudo-wires for transport from the video source to all the drop nodes. This results in multiple identical data streams over the same physical links. Recently, the use of multipoint LSPs has been introduced to compensate for this deficiency, but it still requires manual intervention (leading to higher OPEX) for the appropriate routing of these LSPs.

Set-up, maintenance, and resiliency procedures for multipoint LSPs are still immature and more research work is required to exploit them to their full potential.

The native Ethernet case. As a proven LAN technology, Ethernet has built-in capabilities for unicast, multicast, and broadcast forwarding. Yet, native Ethernet forwarding is not very scalable for large metro/regional networks in terms of number of nodes and end points in a given network domain. As such, it is most effective in the access portion of the distribution network, that is, for packets flowing along the leaf structure of a video distribution tree. The physical layer topology underneath is, similarly to the native IP case, inefficiently used, as the switching nodes do not align according to the physical transport topology.

Although Ethernet does provide native control plane mechanisms to manage L2 multicast distribution trees, Ethernet switches may also snoop on IP multicast control protocols to improve their own forwarding topology. As an example, IP Group Multicast Protocol (IGMP) join and leave requests between clients and IP routers are typically snooped by Ethernet switches to detect multicast traffic activity on a link. Consequently, unused channels can be removed from the multicast forwarding tables while IGMP join requests to active channels can be answered immediately without the need to wait for a response from the upstream PIM router.

The RPR case. Resilient packet ring (RPR) is a ring-based packet switching technology. Initially introduced as an extension to Ethernet switching for metropolitan area network (MAN) applications, it features traffic management and fast protection/restoration capabilities that are highly adapted to ring network topologies. Currently in an early adoption stage, it is available in metro and access rings, but largely for private enterprise network solutions given scalability limitations with respect to resource sharing and connectivity beyond a single ring. As such, it does not support the full range of capabilities required in the outlined network scenario. As a private access ring solution it can be considered as an alternative to native Ethernet, providing some advantages with respect to protection switch times. Because of this limitation it cannot be considered as a solution to the BTV distribution problem in general.

The optical layer case. Optical technology provides a simple mechanism for distribution of unidirectional high-bandwidth multicast flows, such as BTV. Optical drop-and-continue can be used to efficiently replicate packet flows at each node of a multicast distribution tree. Since the traffic replication occurs at L1, it is also congruent with the physical layer topology. Hence, the transport bandwidth is used only once on each link between neighboring nodes. Another advantage of this approach is that it does not introduce any additional packet delay jitter, an important consideration for delay variation sensitive applications such as TV distribution. A limitation of this approach is the lack of awareness about the various types of packet traffic transported, and thus the lack of support for packet-level flow processing, say, for channel selection and traffic prioritization to the user. Thus, optical transport must be complemented with one or more of the preceding packet transport technologies.

Technology comparison. **Table II** captures a high-level assessment of technology alternatives along with the features and capabilities discussed. Looking at the relative ranking, it can be seen that the combination of native Ethernet together with optical technology provides an attractive, efficient, and cost-effective approach to serve most BTV and IPTV distribution needs. Considering the considerable price differences for subscriber access Ethernet equipment, compared to the other technologies, the advantages of this combination as an end-to-end solution become even more pronounced.

The optimal solution from the previous discussion is thus a distribution network built from nodes that have a combination of L1 and L2 capabilities. In this network, L1 technology is used for bandwidth and network cost-efficient distribution in the core and metro network, and L2 technology is used for efficient multicast to digital subscriber line access multiplexers (DSLAMs), channel selection support, and quality of service (QoS).

Network Design Considerations

Given the selection of technologies specified, the main topics to be addressed in the area of network design include the location and efficiency of multicast capabilities in the transport network. This is

Table II. Technology comparison matrix.

Criterion	IP only	IP over MPLS	VPLS	RPR	"Native Ethernet"	Optical
Multicast support	Yes	Immature	Immature	Yes	Yes	Yes
Protection sub 50ms	No	Yes	Yes	Yes	No	Yes
Core network capable	Yes	Yes	No	No	No	Yes
Metro network capable	Yes	Yes	Yes	No	Yes	Yes
Access network capable	Yes, expensive	Yes, expensive	Yes, expensive	Yes, immature	Yes	No
IPTV channel selection	IGMP	IGMP	IGMP snooping	IGMP snooping	IGMP snooping	No
Triple play support	Yes	Yes	Yes	Yes	Yes	No
VPN support	Yes	Yes	Yes	Yes	Yes	No
Circuit based services	No	No	No	No	No	Yes
Efficient transport usage	No	No	No	Yes, Only Rings	No	Yes

IGMP—IP Group Multicast Protocol
IP—Internet Protocol
IPTV—IP television
MPLS—Multiprotocol label switching
RPR—Resilient packet ring
VPLS—Virtual private line service
VPN—Virtual private network

influenced by the number of channels to be made available, national versus regional program insertion, network provisioning, and channel selection mechanisms. The next important topics driving network design are reaction times on channel selection changes and resiliency considerations.

The Proposed Reference Network

For the sake of this discussion, a reference topology is introduced in **Figure 1**. The proposed network topology consists of two tiers of logical rings. The top ring, or inner ring, represents the nationwide portion of the transport network, while the lower one, or outer ring, represents the regional or metro portion of the transport network. NEs in the metro ring aggregate traffic from the access network, in this case DSLAMs, which in turn connect to customer premise equipment.

A ring topology is used for the proposed reference network since most transport networks today have a low nodal degree of connectivity, and, under these conditions, dual interconnected rings provide a higher degree of resiliency than most partial meshes. For this discussion, we will focus on a single metro and a single national ring, but this limitation need not entail loss of generality, since it just simplifies the description of the data path selection process. The rings themselves consist of several NEs, of which only those relevant to the discussion are shown explicitly. Depending on the technology chosen, these NEs can encompass a multitude of equipment from purely optical switches to a combination of optical, switching, and routing equipment.

Typically, because of the related high cost for off-air video streaming, BTV content is delivered from only one or two locations, while regional programming is inserted into the video stream at various points in the regional and metro networks, depending on content locality, e.g., the regional franchise of major TV stations versus educational stations from community colleges, or community advertising inserts. Hence, even though the video distribution network may cover a large geographical area (e.g., in the case of the United States), content distribution requires the insertion of national, regional, and local BTV streams at various points in the network. This creates the need to combine different traffic streams in a decentralized fashion, and this must occur in the service stream domain, which is packet oriented. As a result,

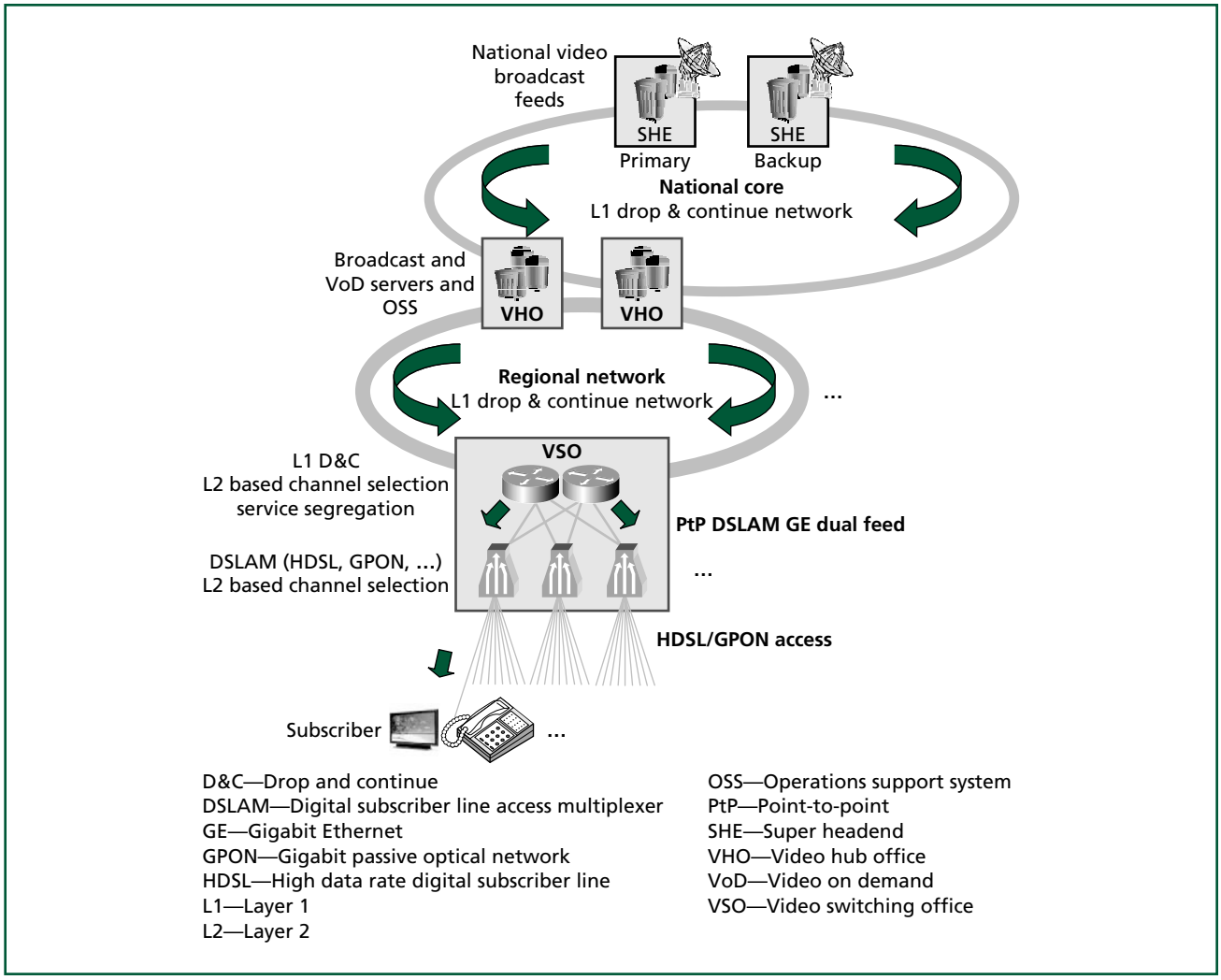

Figure 1.
Proposed reference network.

the network elements at those points in the network must terminate the optical layer and process the data in the packet domain, for further encapsulation and transport via optical fiber to the handoff points. The functional procedures of content interleaving are beyond the scope of the discussion of this paper and will not be detailed any further.

Video Distribution Network Architecture

At the super headends (SHEs), Motion Picture Experts Group (MPEG)-formatted multiprogram streams of broadcast video content are received from the original satellite video feeds. These feeds may be formatted as either MPEG over digital video broadcast asynchronous serial interface (DVB-ASI) or MPEG over Ethernet and, hence, not yet in a format for transmission over an IP network. At the video hub office (VHO), the MPEG multiprogram streams are broken down and reformatted as MPEG single program streams, mapped into IP (either IP/User Datagram Protocol [UDP] or IP/UDP/Real-time Transfer Protocol [RTP]), delayed for transmission as per time-zone programming schedule, and mixed with other IP-encoded regional/local video content and transmitted downstream toward the end customer via an Ethernet-enabled regional/metro distribution network. As a result, each source of video content is now characterized by a dedicated IP (unicast) source address and an IP (multicast) destination address per BTV (i.e., IPTV) channel. IP multicast addresses are mapped into

Ethernet multicast addresses via established procedures [3]. The resulting packet stream is then injected into the video distribution network.

The network elements connected to the distribution network consist of a packet switching system and a reconfigurable optical add/drop multiplexer (ROADM):

- The packet switching system receives a number *m* of individual IPTV streams; replicates all relevant streams via Ethernet multicast to two ports or logical groups of ports, such that each port receives the same set of *m* IPTV streams; and optionally shapes each stream in order to reduce the packet jitter and burstiness to be introduced over the network.
- The ROADMs are used to implement L1 connectivity for the distribution network. At a minimum, two wavelengths are dropped for further processing at this node, one for bidirectional packet transport (e.g., for services such as Internet access and VoIP not addressed in this discussion), and another for the unidirectional IPTV bearer stream. Each port from the packet switching system (actually, an Ethernet router) is mapped in to a separate transport wavelength. Each of the wavelengths carrying the IPTV stream travels in the opposite direction from the ring traffic to provide the require degree of path diversity under link and node failures.

It is important to appreciate that any packet replication at the VHO location happens in the packet domain, rather than in the optical domain, in order to preclude a single handoff point between the packet domain and the optical domain, and hence, a single point of failure. At the VHO location, multiple regional rings may be hanging off the national core. In addition, regional video content (e.g., local programming or advertisements) can be inserted into the video stream from the national core. Once the final video stream is generated, downstream transmission toward the video switching office (VSO) can fully take place in the optical domain.

The unidirectional nature of the IPTV bearer stream provides further options for network design optimizations. In particular, the IPTV bearer stream does not require the provisioning of a separate wavelength for upstream (toward the VHO) traffic. (IPTV control traffic such as IGMP will be discussed later.) Optical drop and continue (D&C) can then be used to replicate IPTV traffic across the metro/regional distribution network and eliminate the need to perform data flow replication in the packet domain, which is processing-intensive and jitter-prone. Specifically, the working and protection wavelength carrying the set of *m* IPTV channels is forwarded from node to node as an optical signal, with each wavelength following a diverse path across the metro/regional subtending rings. At each node, these IPTV bearer wavelengths are either optically replicated for further distribution downstream to the end customer or dropped into the local Ethernet switch for distribution to the subtending DSLAMs. Note that if another tier of access rings were to be added to the reference network, early termination of the optical signal might be required at each of these points to forward the traffic to a local packet switch in order to insert local programming, such as local TV, at this point in the distribution network.

The network location where the IPTV stream leaves the video distribution network and is forwarded to the local access network is referred to as the video switching office (VSO). A large number of DSLAMs, or other access-specific systems, e.g., Ethernet passive optical network (EPON)/Gigabit Ethernet passive optical network (GPON), are typically hosted in these locations. Each DSLAM may support multiple hundreds of subscribers per site, depending on local population density and service take rate. With 50 to 100 DSLAMs per VSO and 500 to 1,000 subscribers per DSLAM, over 50,000 subscribers may be easily handled by a single, very large VSO.

Managing ITPV flows from the VHO to the VSO. The VSO acts as the final distribution point of IPTV channels to the access network and its multiplicity of end customers. Whenever the VSO population size is large enough, it proves far more efficient to send all IPTV channels directly from the VHO to the VSO and provide channel selection capabilities directly at the VSO [9]. For instance, with a local active population of 1,800 consumers, and assuming a flat/Zipf-distributed channel usage preference, the average number of simultaneously watched IPTV channels would be close to 90% of the available channels for video systems with a capacity of between 100 and 400 IPTV channels. Larger VSOs make this percentage even higher.

In a different scenario, with only a few users who are relatively widespread across a large geographical area, channel selection should be done closer to the video servers, to help amortize the cost of the video system infrastructure and reduce wasted bandwidth due to the transport of IPTV channels that are not being watched. Here, a mostly static wide distribution approach might not be economically favorable. In situations where the active user population of a video system with 250 to 500 channels falls below 100 subscribers, as few as 50 of those channels may be watched simultaneously [9].

Managing IPTV flows within the VSO. The very large number of access nodes (e.g., DSLAMs) and subscribers supported in a VSO demands a strong resiliency strategy for components of the distribution network. In addition, channel change requests (e.g., IGMP join or leave requests) are signalled via control plane protocols between the end user and the video server systems. Ideally, processing of channel selection requests is best handled as closely as possible to the subscriber (i.e., the access nodes). This approach helps to achieve better control plane load sharing and to reduce the reaction times to these end user requests. It also validates the strategy of delivering as many IPTV channels as possible to the VSO, as the high subscriber concentration already makes the number of concurrent requests for different channels quite high.

Figure 2 illustrates the implementation of a VSO via a hybrid ROADM/Ethernet system. The packet switching subsystem receives redundant feeds of IPTV channels from the ROADM subsystem, typically via two 10-gigabit Ethernet (10 GE) links. It then selects the proper IPTV channel signal from either the red or the blue wavelength feed, by monitoring the failure status of the physical signal, or of a signal group in the case of a unidirectional link aggregation group (LAG). The NE that feeds the original video signal at the VHO may use a variety of mechanisms such as in-band keep-alive messaging, delay information, and/or signal integrity checks, to allow for local failure-tolerant signal selection.

It is the task of the packet subsystem in Figure 2 to replicate a predefined or selected set of IPTV channels to the attached access nodes. This requires the packet forwarding subsystem efficiently to produce and control replicas of up to thousands of IPTV channels. Typically, packet switches offer dedicated multicast replication stages with strong bearer plane bandwidth limits. In the scenario described, however, multicast may be to a large extent the predominant traffic on the system. Looking at Figure 2, an incoming bandwidth of 10 GE needs to be replicated selectively to a number of DSLAMs, ranging between 10 and 100, each potentially receiving 1 Gbps to 4 Gbps of bandwidth.

Channel Selection

Even with the new generation of broadband access technology, there still may be a need to manage the traffic an access node receives from its homing broadband aggregation switch. For example, a 200- to 400-channel video system would require at least 4 Gbps to 8 Gbps of bandwidth to each access node just to support the needs of high-definition TV traffic alone, not to mention the needs of other residential applications such as VoD, VoIP and Internet access. On the subscriber link side (e.g., ADSL or VDSL) the system capacity is further reduced to a handful of channels at most, depending on signal quality and bandwidth availability.

In an IPTV video system, the IGMP protocol is used to manage BTV channel access via the customer's set-top box.

The edge PIM router at the VHO processes any incoming IGMP join/leave requests. If the IPTV channel is not already available, the IGMP joint request is further propagated to the source video server, which forwards the desired IPTV channel to the requesting PIM router. Once the target IPTV channel becomes available, any further requests to an active IPTV channel are handled directly by the edge PIM node. In the proposed reference network, however, since all channels are made available at the edge of the distribution network, the IGMP protocol messages can be snooped by packet processing devices at the VSO to help optimize channel selection. The most natural place to perform this optimization is at the first aggregation device in the VSO: the access node. In a manner similar to IGMP processing by edge PIM routers, the access node can snoop the IGMP messages and directly forward the requested IPTV channel if it is already available in the access node. Similarly, if the

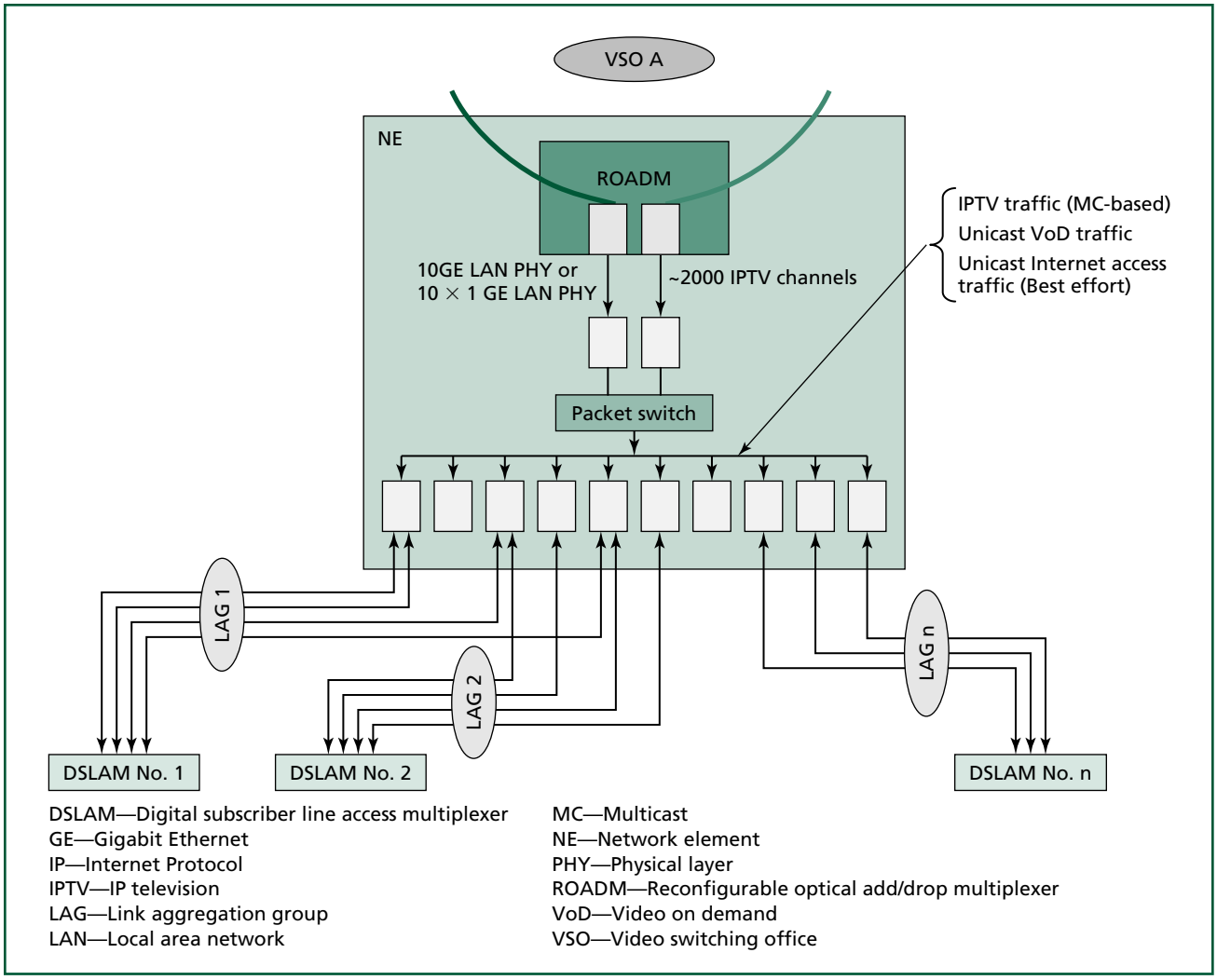

Figure 2.
IPTV flow replication strategy at VSO.

requested IPTV channel is not available at the downstream access node, the broadband aggregation switch can also snoop IGMP messages forward of the IPTV channel, if available, to the access node for further distribution downstream to the intended subscriber.

A further optimization is possible under the proposed reference network architecture. IGMP control messages need not be further propagated upstream to the video server (which also resides in the VHO location) but filtered in the aggregation device. This is possible as aggregation devices at the VSO are already supplied with all the IPTV channels. In such a scenario, the video server and the PIM routers can be statically provisioned to feed all the IPTV multicast groups, offloading the network from unnecessary control traffic.

Transport Resiliency at the VSO

As outlined earlier, the expected high number of broadband IPTV channels demands a careful approach to network resiliency to ensure both transport efficiency and associated lower CAPEX. Therefore, the L3/L2 forwarding topology cannot be treated in isolation from the underlying L1 transport network topology, as network engineering is a multilayer multiflow optimization problem. Of special value, in this context, are converged nodes that can provide integrated L1 and L2+ forwarding functionality, which makes the

required information from the L1/L2+ network topology readily available, and hence, more easily exploited to provide efficient forwarding of the client IPTV signals

Reliable, fast, and inexpensive protection of the access link to the access node can be addressed via native Ethernet's LAG capabilities [4], which support dynamic forwarding of packet flows across multiple parallel links. Since these links must support a multiplicity of residential subscriber services including IPTV, VoD, Internet access, and VoIP, traffic distribution across these aggregated links cannot be optimized to satisfy the needs of IPTV traffic alone. Although sophisticated flow-based traffic distribution functions could be proposed, it will be simpler to aggregate flows on a per class of service (CoS) basis and treat the vulnerable valuable services (such as IPTV and VoIP) with a higher level of resiliency than best effort services (such as residential Internet access).

A LAG distribution function with a CoS-based overlay protection scheme can be implemented by establishing, for example, via preprovisioning, separate working and protecting component links out of the links in the LAG in a N:M protection arrangement, where N refers to the number of working links and M refers to the number of protection links. **Figure 3** illustrates considerations for the LAG distribution function implementation. Here, the NE acts as an aggregation device for three DSLAMs. There are two distinct IPTV streams set up, TV1 and TV2, both of which are to be multicast to all DSLAMs. On DSLAM 3, TV2 is received from port 9 of the system. Should the physical link fail, the LAG distribution function inside the NE needs to rearrange the multicast flows such that a protection port can be used to carry TV2 (port 10), without changing the forwarding on all other LAGs. This scenario highlights the operational complexity

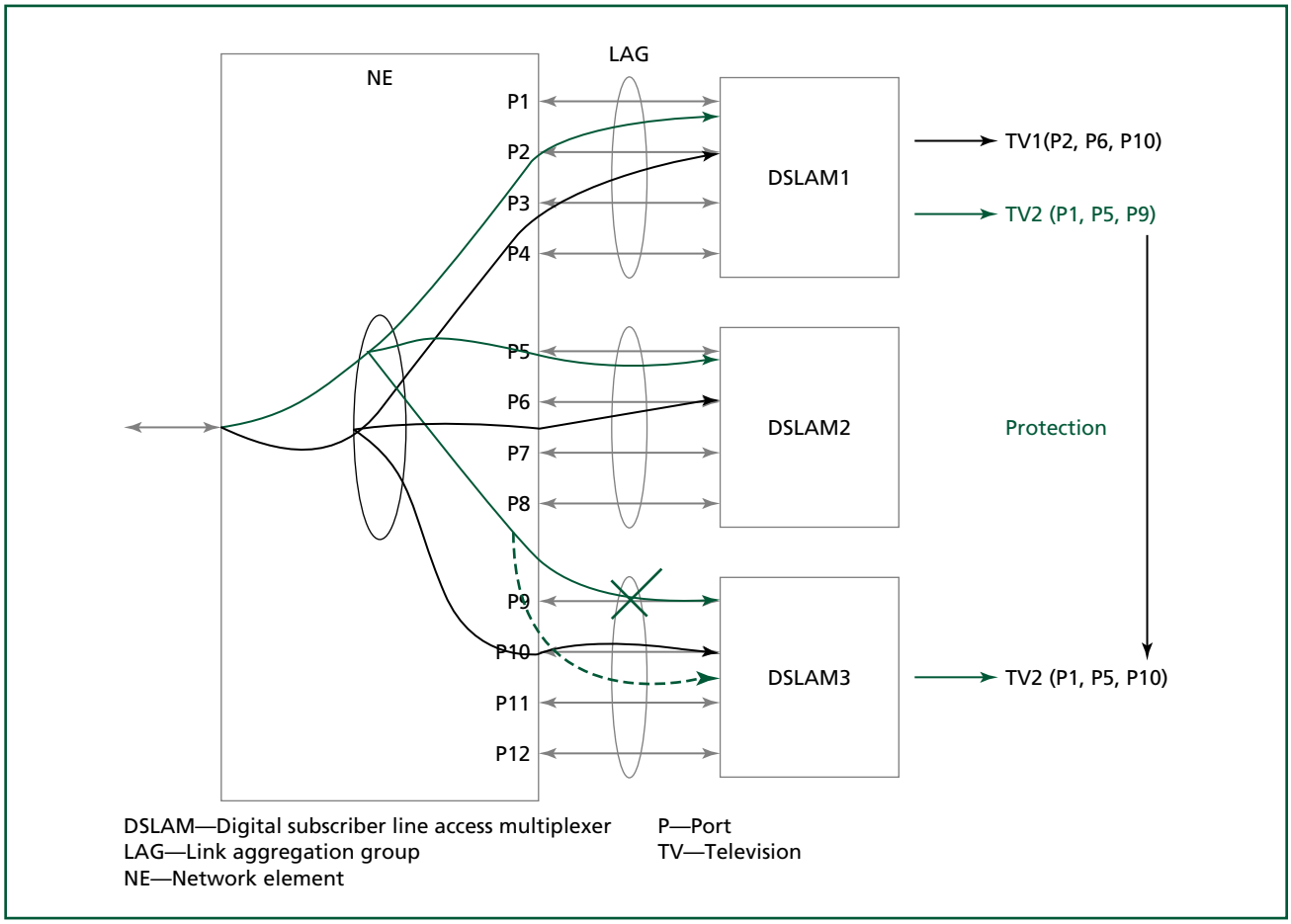

Figure 3.
LAG and distribution function towards the DSLAM.

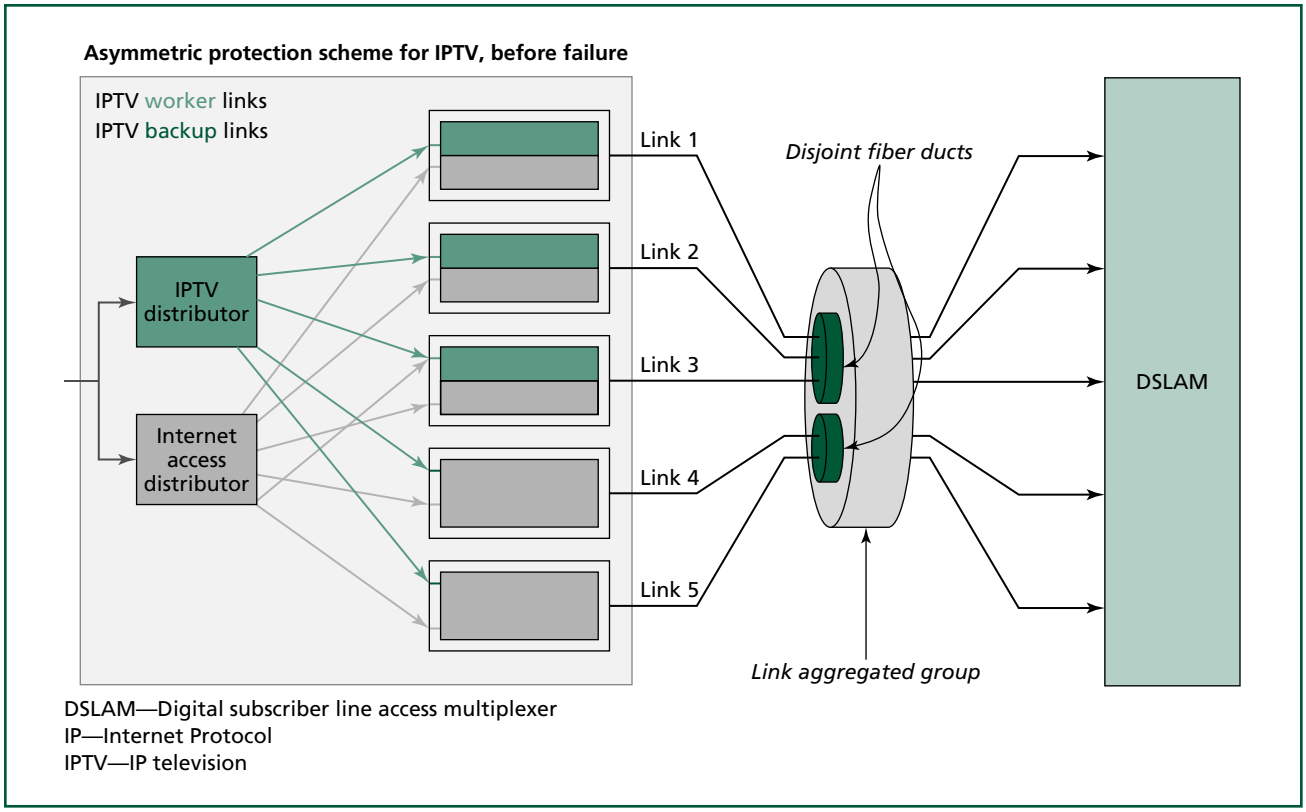

Figure 4.
A link protection strategy for mixed services.

involved in the aggregation NE: there may exist thousands of multicast groups (i.e., IPTV channels) that are being distributed each to an arbitrary subset of up to 100 DSLAMs, connected to link aggregated groups of several (e.g., two to eight) physical interfaces.

IPTV protection overlay. A LAG distribution function that implements such a CoS-based overlaying protection scheme for the premium traffic, such as BTV and VoD traffic, while it provides a best effort protection and scheduling scheme for Internet access traffic is shown in **Figure 4.** When IPTV multicast traffic is forwarded to multiple DSLAMs, IPTV channels become allocated to one of the working links of the LAG as long as there is enough free bandwidth available to handle the bandwidth demand. A threshold may be defined such that only, for example, half of the link capacity can be allocated for video distribution. The rest of the link bandwidth is used for best effort traffic. The LAG distribution function further ensures that all the configured IPTV channels are only allocated to working links; protection links are not utilized for any premium services but can carry any best effort traffic.

The protection overlay state machine for the LAG distribution function operates on the granularity of the traffic class per physical link. The LAG distribution function handles the premium traffic classes, such as a group of IPTV channels, as a single protected entity that is switched in its entirety in case of a working link failure. This protected entity is referred to as a bouquet. For example, a bouquet on a 1-Gbps physical link may consist of 100 channels at 5 Mbps each. In Figure 4, three individual bouquets have been defined and allocated to the working links of the LAG. Upon failure of a working link (e.g., link 2), the complete bouquet of the failed link will be transferred to one of the protection links, as shown in **Figure 5.** To minimize failure detection and processing delay associated with recalculating internal multicast replication and redistribution of potentially hundreds of channels when failures occur on different LAGs to different DSLAMs, the protection links can be

preassigned for each working link. This does not imply that flow distribution symmetry must be enforced between working and protection links. However, this preassignment allows for precalculation of the protection scenario and thus dramatically improves switchover times.

Equipment protection is another design objective that is nicely supported by this traffic engineering mechanism: either the working and protection links for video and associated optical signals/fibers are routed via a physically disjoint path, and/or the physical links may be terminated in the converged optical node on different line cards, i.e., all working links on one line card and all protection links on another. In this way, the protection groups can easily cope with line card equipment failures and breakage of a complete fiber duct, as shown in **Figure 6.**

Best effort traffic, e.g., Internet access traffic, is handled in a totally different manner. For this traffic type, all operating links of the LAG are used for packet distribution. Priority scheduling per physical port ensures that on working links, best effort traffic cannot burst into the reserved bandwidth for premium traffic (e.g., IPTV). Failure handling for best effort traffic simply removes the failed link from the LAG distribution scheme, and hence, gracefully degrades capacity available for best effort services illustrated in Figure 6.

Should the available protection capacity not be sufficient (e.g., as a result of multiple link failures), a decision might be made independently as to which of the configured bouquets should be dropped completely. A more sophisticated mechanism, depicted in **Figure 7,** may even prefer to protect high-definition television (HDTV) channels over standard television (SDTV) channels as these consume less bandwidth, or protect national channels over regional ones, or pay TV channels over free TV channels. This requires the distribution mechanism to have further knowledge of the channel priorities or service types.

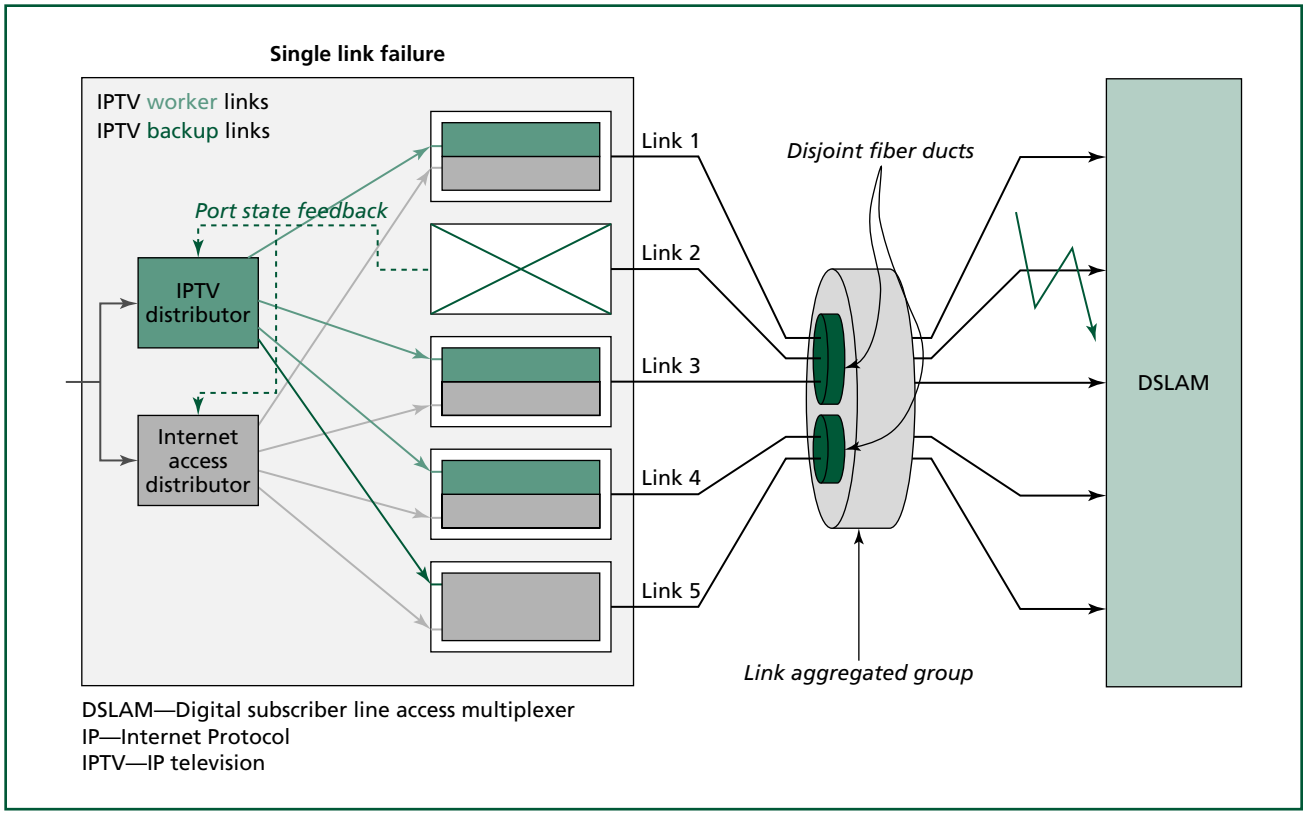

Figure 5.
Single physical link failure and dual-service protection strategy.

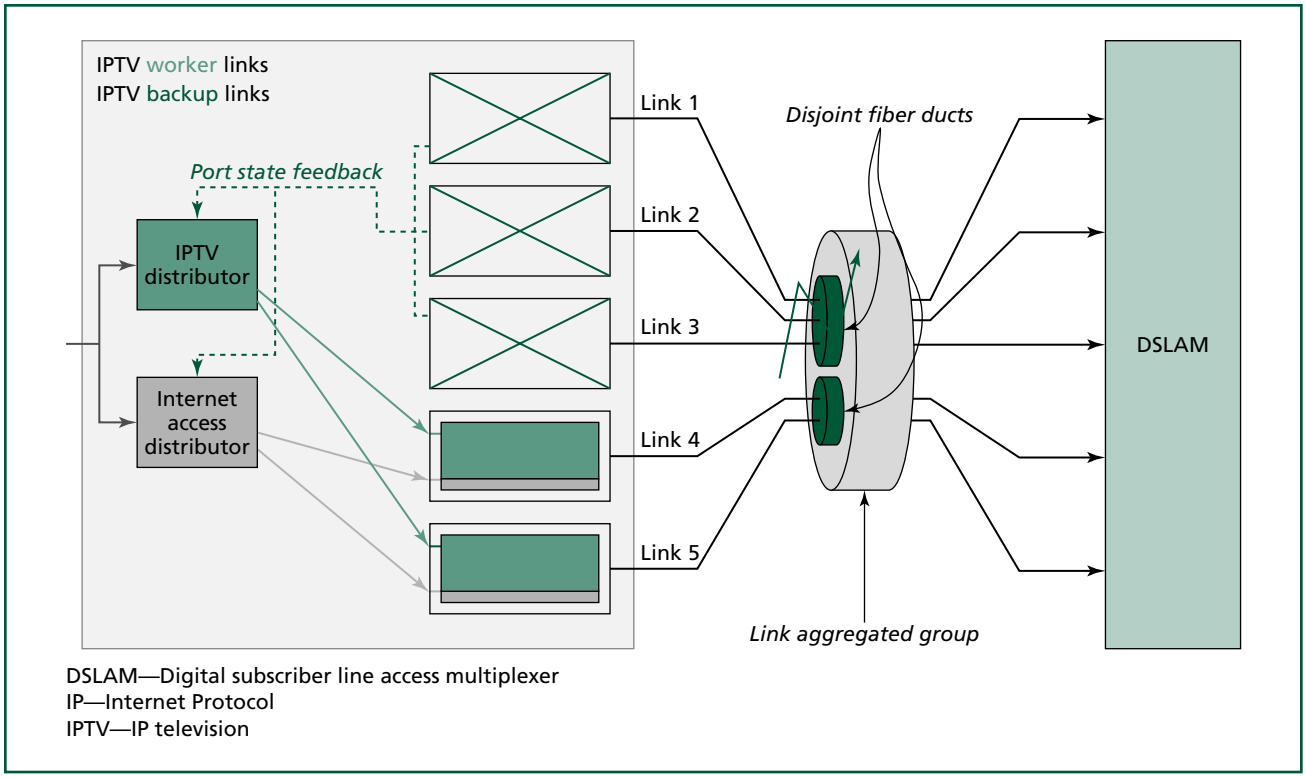

Figure 6.
Fiber duct failure.

Summary: Role of DSLAM and Transport Nodes

The function of each of the nodes in the metro/regional distribution network can be summarized as follows: The feeding switch in the VHO location is a converged Optical/Ethernet system (converged ROADM/Ethernet switch) made up of a packet subsystem and an optical subsystem. The packet subsystem receives video feeds from a PIM router via one or more interfaces. It gathers all the BTV channels and multiplexes them onto the drop and continue optical ring via the optical subsystem.

To the PIM router, the converged Optical/Ethernet system can either statically request all provisioned BTV channels or request them dynamically via an IGMP join message. The NEs at the VHO location act as the source of the video feed for a regional subtending optical ring. They may aggregate national video content with regional streams in the packet domain and multiplex both onto the optical domain. In other scenarios, they may just operate in the optical domain as virtual fiber ROADM. For operations, administration, and maintenance (OAM) considerations, they may terminate control messages from the VHO and generate control messages to the VSO location.

At the VSO location, the optical subsystem extracts the Ethernet signals from one of the selected optical feeds and forwards them to the packet subsystem for selective replication to the attached access nodes. It filters IGMP control traffic from the downstream network elements and uses IGMP snooping to set up its internal forwarding database to provide each access node with the desired subset of IPTV channels. In another scenario, local video content could be inserted at the VSO location, which could be considered as a collapsed VHO/VSO functionality.

The access node receives a subset of statically configured or dynamically requested IPTV channels. On a per subscriber basis, it listens to the IGMP control traffic received from the subscriber line and acts as the final stage of replication for the IPTV channels. It may also implement an IGMP proxy function, in which case it essentially aggregates the IGMP join/leave requests and filters those requests it can handle on its own. The access node would also control the ingress

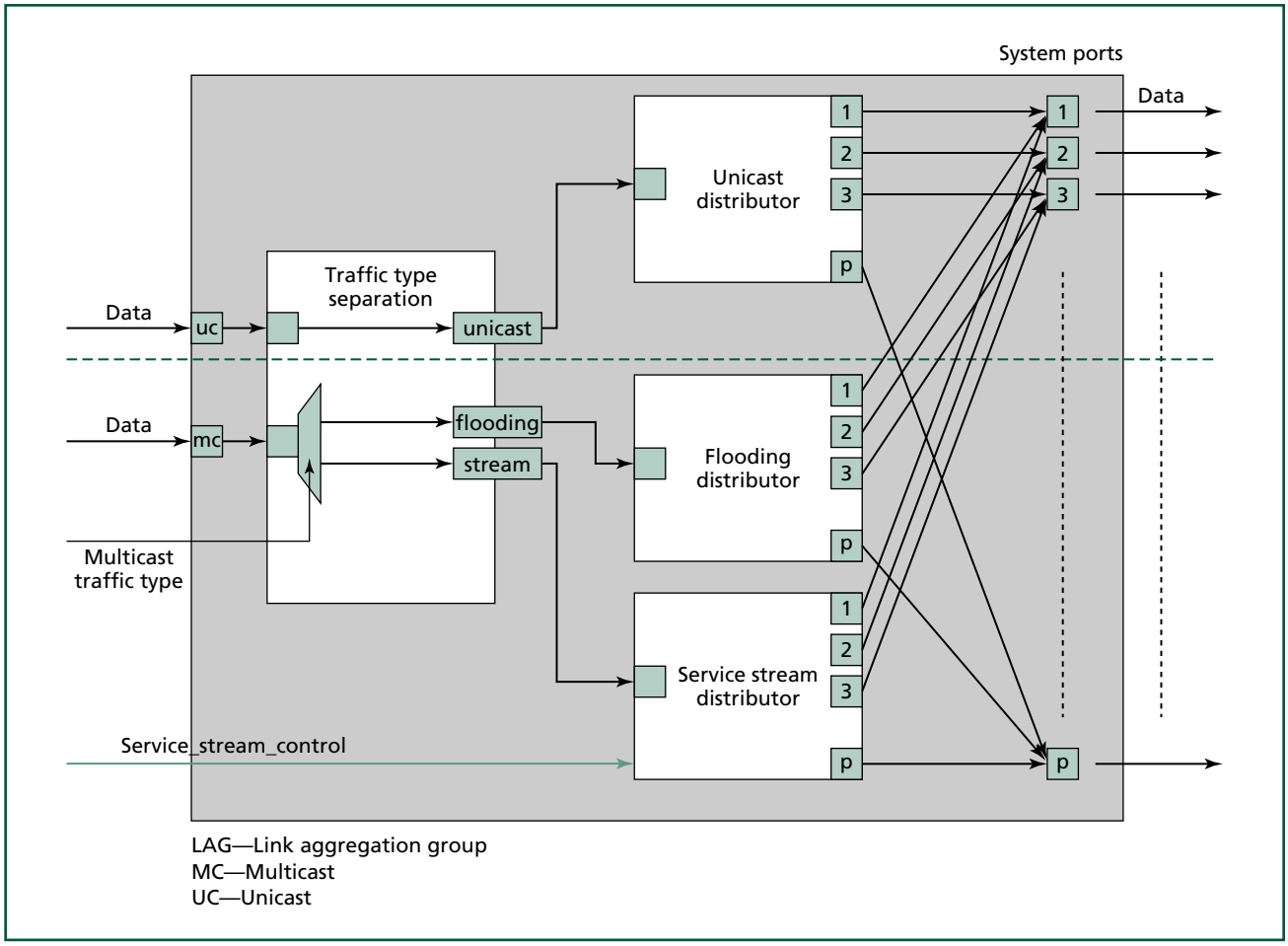

Figure 7.
LAG service aware distributed functions for various data flows.

service parameter so as to support the service level agreement (SLA) per end subscriber by policing, classifying, and marking the ingress traffic per service type and by shaping the egress traffic.

Service Considerations

The fact that there are well-defined standards to invoke and deliver services is not sufficient to guarantee the quality of the user experience. The end user services need to be in harmony with the capabilities of the network infrastructure and properly matched, and accounted for, by a consistent set of network resource management and OAM&P capabilities. The end user may desire guaranteed QoS for specific services (e.g., premium sporting events such as the baseball World Series or soccer World Cup finals) while others may be content with legacy best effort delivery (e.g., viewing the latest MTV* video clip). On the other hand, service providers may want to provide graded levels of QoS, but only to customers with special service contracts. Furthermore, service providers may want to enforce certain security measures, such as blocking distrusted traffic or performing Network Address Translation/Network Address Port Translation (NAT/NAPT) to hide details about their networks and services. All these services compete for the same set of network resources.

All this means that service providers must not only supervise the network infrastructure and correlate alarms (e.g., fault type, location, severity) and threshold crossings (e.g., bit error rate [BER] and offered load), but also be able to monitor and correlate

service-level performance indicators (e.g., usage, delay, and loss) to be able to link these data to the QoS delivered in each particular service instance to validate any QoS commitments and to support proactive traffic engineering needs. Link and network-level performance and supervision functions are technology specific but relatively well understood (see, for instance, ITU-T references [5, 6] for SDH and Ethernet equipment specifications). Service-level performance and supervision functions for IP-based services are not as well documented. Relevant parameters may be extracted from related applications for business-oriented services, typically in terms of packet loss, delay, and jitter, although additional parameters to represent the higher-level impairments, such as the Mean Opinion Score (MOS) metric for voice services, may be required for video-oriented services.

Fourth-Generation Transport Networks: IMS and RACF

Several standards organizations are working on generic QoS frameworks, including the European Telecommunications Standards Institute (ETSI), 3rd Generation Partnership Project (3GPP*), and Internet Engineering Task Force (IETF). Yet most of these organizations have taken a limited application-specific view. ITU-T SG13 has taken an all-encompassing approach to address a resource management framework for next-generation networks (NGNs). Their functional architecture is depicted in **Figure 8.** The ITU-T NGN model is a generalization of the ETSI/3GPP IP Multimedia Subsystem (IMS) and the IETF Common Open Policy Service (COPS) frameworks.

In the ITU-T NGN model, the resource admission and control function (RACF) is a distributed control element that mediates between service control functions and the network infrastructure. Such service control functions may be part of IMS or non-IMS protocols. Video on demand middleware is an example of a non-IMS service control function. The service control functions interface with RACF to request resources and controls for service-related traffic flows. Within the network infrastructure a distinction is made between policy enforcement functional entities (PE-FEs) and the network segments through which they are connected, called the interconnection functions. In the PE-FEs functions such as policing, filtering, QoS marking, usage recording, and NAT are performed. The interconnection functions are responsible for transport, switching, and routing. Interconnection functions can range from a single link to an IP/MPLS core network.

The RACF itself consists of policy decision functional entities (PD-FEs) and transport resource control functional entities (TRC-FEs). The PD-FE is the

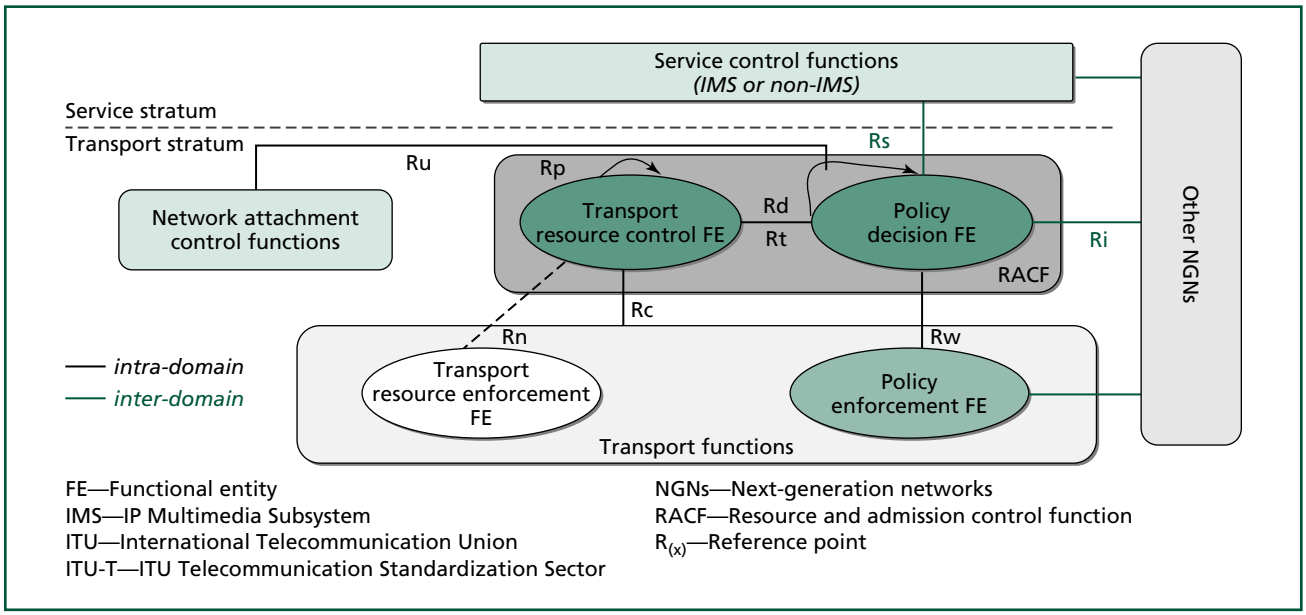

Figure 8.
ITU-T next-generation network RACF functional architecture.

ultimate decision point regarding resource control. It bases its decisions on a set of policy rules and on information retrieved from other elements such as the TRC-FEs and network attachment control functions (NACFs). The NACF maintains information such as the binding between a subscriber's IP address and the access node through which that IP address is reachable. The PD-FE directly controls the PE-FEs. It pushes the policing parameters, filtering and QoS-marking rules, and other parameters to the enforcement function. TRC-FEs are linked with network segments. They are technology specific. For example, the TRC-FE responsible for an Ethernet-based metro network differs from a TRC-FE responsible for an MPLS core network. In some sense, TRC-FEs can be considered admission control and traffic engineering functions that may be located within network elements or in servers that control specific network domains.

RACF Usage Scenarios

In the following we present potential implementation of session initiation scenarios for two key services, IPTV and VoD, to illustrate the NGN RACF end-to-end architecture.

IPTV. When a user switches to a new channel, the set-top box in the subscriber home sends an IGMP join request to the network. The access node intercepts the IGMP message. The IPTV service control function (PE-FE) in the access node checks with the TRC-FE (which may be implemented locally in the access node) and verifies there are sufficient resources (e.g., bandwidth) on the subscriber line to carry the new traffic flow. If resources are available, it starts forwarding the requested channel to the user and the IGMP message may be further propagated to the PIM servers for accounting purposes. If not, the IGMP message may be blocked and a notification sent back to the PD-FE. Since IPTV is a well-defined service, most of the service policy attributes are likely to be pushed "down" by the PD-FE during service activation.

An alternative implementation would allow for a proxy PD-FE instance in the access node. This is essentially a shim function between the IPTV service control function and the TRC-FE. Instantiating a PD-FE in the access node guarantees that the processing of IGMP messages remains local and does not require communication with outside servers.

VoD. When the user uses the program guide on his or her TV to start a VoD session, it results in a Real Time Streaming Protocol (RTSP) message exchange with VoD middleware. To reserve resources, the VoD middleware will send a resource request to the RACF. The RACF communicates with the NACF to find out where the user is located. It then communicates with the TRC-FE for the access node to check whether there is enough bandwidth available on the subscriber line. It may also communicate with other TRC-FEs in the connection path if intermediate bottleneck points may be encountered. It then communicates with the PE-FE in the access node to push down the relevant policy parameters and QoS marking rules, according to the requested service type, so that the traffic flow related to the VoD call will be handled appropriately on the outbound ports.

Conclusion and Outlook

Next-generation transport networks are slated to include functions that have been traditionally provided by separate network elements in separate network layers as network operators seek to simplify and reduce complexity on the transport and service network infrastructure. By combining DWDM and packet functionality and by utilizing the strength of each networking layer, a significant new value can be created by the operators. High-capacity optical drop and continue core rings with L2 switch function can selectively pick up BTV channels on the basis of the quality of the signals as well as service policy. Using the L2 switch functions for BTV channel (multicast) selection creates a highly optimized network architecture for BTV distribution. By using an embedded DWDM subsystem, multiple wavelengths (links) can be connected to a L2 switch subsystem, with each color capable of supporting a different optical network topology. Hence, service-aware traffic segregation can be implemented in a highly efficient fashion and each service can be carried via an individually optimized network topology. For instance, different service types can be preclassified by identifying the individual services via virtual local area network (VLAN) tags and by providing per

service type QoS mechanisms. Further optimizations for BTV can be envisioned, including the following:

- Additional service selection criteria, e.g., inserting OAM information into the video distribution stream and making protection switching depending on this information
- Integrating optical front ends like DWDM transponders and packet input/output (I/O) function into single packs, saving additional hardware costs

Acknowledgments

We acknowledge gratefully that part of this work has been supported by the German Ministry for Research and Education (BMBF) under grant AK067A supporting the European ITEA ENERGy project.

*Trademarks

3GPP is a trademark of the European Telecommunications Standards Institute.

MTV is a registered trademark of Viacom International, Inc.

References

[1] A. Bodzinga, S. White, and M. Weldon, "Enhancing the IPTV Service Architecture to Enable Service Innovation," Delivering the Promise of IPTV: Comprehensive Report, International Engineering Consortium (IEC), Chicago, 2006, pp. 43–60.

[2] A. Cohen and E. Shrum (eds.), Migration to Ethernet-Based DSL Aggregation, DSL Forum, TR-101, Apr. 2006.

[3] S. Deering, "Host Extensions for IP Multicasting," IETF RFC 1112, Aug. 1989, <http://www.ietf.org/rfc/rfc1112.txt?number=1112>.

[4] Institute of Electrical and Electronics Engineers, "Part 3: Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications," IEEE 802.3–2002, Mar. 8, 2002, <http://www.ieee.org>.

[5] International Telecommunication Union, Telecommunication Standardization Sector, "Characteristics of Transport Equipment—Description Methodology and Generic Functionality," ITU-T Rec. G.806, Feb. 2004, <http://www.itu.int>.

[6] International Telecommunication Union, Telecommunication Standardization Sector, "Characteristics of Ethernet Transport Network Equipment Functional Blocks," ITU-T Rec. G.8021/Y.1341, Aug. 2004, <http://www.itu.int>.

[7] M. Lasserre and V. Kompella (eds.), "Virtual Private LAN Services Using LDP," IETF Internet Draft, June 2006, <http://www.ietf.org/internet-drafts/draft-ietf-l2vpn-vpls-ldp-09.txt>.

[8] Metro Ethernet Forum, "Ethernet Services Attributes Phase I," Tech. Spec. MEF 10, Nov. 2004, <http://www.metroethernetforum.org/PDFs/Standards/MEF10.doc>.

[9] D. T. van Veen, M. K. Weldon, C. C. Bahr, and E. E. Harstead, "An Analysis of the Technical and Economic Essentials for Providing Video Over Fiber-to-the-Premises Networks," Bell Labs Tech. J., 10:1 (2005), 181–200.

(Manuscript approved November 2006)

CHRISTIAN HERMSMEYER is a distinguished member of the technical staff in the Systems Engineering and Architecture Department of Alcatel-Lucent's Multimedia Network Solutions division in Nuremberg, Germany. After working in the areas of hardware and system architecture within PHILIPS Kommunikations Industrie AG in Nuremberg and Glasgow, Scotland, Mr. Hermsmeyer moved to the ASIC Department at Lucent Technologies, where he developed SONET/SDH and data devices. He further contributed to OTN transmission devices and 40-Gbps transmission systems. In his current work in Alcatel-Lucent's Systems Engineering Department, he focuses on the architecture and evolution of packet and transport integrating systems. He holds an M.S. degree in electrical engineering from the University of Dortmund, Germany, and three patents for transmission networks.

ENRIQUE HERNANDEZ-VALENCIA is a Bell Labs Fellow and a consulting member of technical staff at Alcatel-Lucent in Holmdel, New Jersey. He received his B.Sc. degree in electrical engineering from the Universidad Simon Bolivar, Caracas, Venezuela, and his M.Sc. and Ph.D. degrees in electrical engineering from the California Institute of Technology, Pasadena. He has over 15 years of experience in the design and development of architectures, systems, and protocols for high-speed communications networks. Dr. Hernandez-Valencia is a member of the Institute of Electrical and Electronics Engineers, Association for Computing Machinery, and Sigma Xi societies.

DIETER STOLL is a distinguished member of technical staff in the Software Engineering Department of the Optical Networking Group at Alcatel-Lucent in Nuremberg, Germany. He graduated from the science faculty and received a Ph.D. in theoretical physics from the University of Erlangen, Germany, and he has held postdoctoral research positions at the University of Tokyo, Japan, and the University of Erlangen, Germany. His work at Alcatel-Lucent has spanned the areas of performance engineering, embedded software architecture, and system definition, and he has managed research projects in the area of automatically switched optical networks (ASONs), generalized multiprotocol label switching (GMPLS), multilayer networks, and QoS and service management. Dr.Stoll is coeditor of two books and a contributor to more than 30 books, journals, and conference proceedings.

OLIVER TAMM is a technical manager in the Systems Engineering and Architecture Department of the Multimedia Network Solutions Group at Alcatel-Lucent in Nuremberg, Germany. He was lead architect for the LambdaUnite® MultiService Switch and Universal Packet Mux and is currently heading a team defining Alcatel-Lucent's next-generation data convergence and optical cross-connect and multiservice provisioning platform architectures. He holds a M.S. degree in electrical engineering from Technical University of Darmstadt, Germany, and patents for transmission networks. ◆