Designing a Reliable IPTV Network

Robert Doverspike, Guanzhi Li, Kostas Oikonomou,
K. K. Ramakrishnan, Rakesh K. Sinha, Dongmei Wang, Chris Chase
{rdd, gli, ko, kkrama, sinha, mei}@research.att.com; chase@labs.att.com
AT&T Labs, USA

Abstract: Carriers are implementing IPTV that streams real-time content, such as broadcast TV, using IP multicast. While providing networking advantages, IPTV requires stringent Quality of Service (QoS) constraints, such as low latency and loss. The challenge is to design the combination of underlying IP-transport network, restoration and video/packet recovery methods that achieves the necessary QoS. Carriers use link-based Fast Reroute (FRR) as the primary underlying transport restoration method to achieve this goal. Furthermore, we can carefully tune the link weights in the IP routing protocol to avoid traffic overlap from FRR during single link failures. But multiple failures may still cause path overlap in long-distance networks. By establishing an architecture where FRR, IP routing and other packet protocols work in harmony, path overlap can be avoided in all but the rarest of multiple failures. This paper provides an overview of these challenges and the approaches used to solve them.

Keywords: Multimedia, Multicast, Network Restoration, Network Availability.

Introduction

Distribution of real-time multimedia over an IP backbone has been gaining momentum with content and service providers ([1], [2], [3], [4]). Distribution of real-time ‘linear’ (also called broadcast) TV requires that the delay experienced by a user be limited to less than a few seconds. Moreover, loss-recovery mechanisms have a limited capability to recover from burst packet losses. The combination of player loss-concealment algorithms, recovery through retransmissions and packet-level redundancy mechanisms is designed to recover from short burst losses. When we combine these higher-layer loss-recovery mechanisms/protocols with the capability to restore the lower-layer within (roughly) 50 milliseconds after a network failure, then we can achieve the needed QoS to the end-user. To see this, we observe that the vast majority of network (physical layer) failures are single link failures and can be recovered using Fast Reroute (FRR) within approximately 50 milliseconds. Thus, our remaining task is to identify methods to mitigate the effect of multiple failures such that the resulting downtime does not exceed the target network unavailability objectives.

Clearly, to deliver the needed QoS to the end user, one must carefully design an architecture that melds standard IPTV architectures, video/packet recovery mechanisms/protocols, and underlying network design and restoration methods. Consequently, based on our hands-on experience with real networks, in this paper we describe such an implementation. For example, multicast ([5], [6], [7]) enables design of efficient cost-effective network; link-based Fast Reroute (FRR) ([8], [9], [10]) is a common approach to providing the needed high availability. However, if FRR is not coordinated carefully with the placement of network capacity, it can result in congestion and packet loss during network failures. We show that the traffic overlap can be avoided during any single IP link failure by tuning the link weights of the Interior Gateway Protocol (IGP) routing protocol, such as IS-IS or OSPF [11]. But multiple failures may still cause traffic overlap with FRR. By having FRR, IGP and multicast protocols work in harmony and with appropriate link weight assignments, path overlap during multiple failures can be minimized. We summarize studies that show how the proposed method improves service availability and almost completely avoids traffic overlap in all but the very rare cases of simultaneous failures.

Network Architecture for IPTV Service

Figure 1 shows an example, simplified network topology for the long-distance-backbone, metro, and access segments to deliver IPTV service for the continental USA. The Super Hub Office (SHO) gathers content from the national video content-providers, such as TV networks and cable networks (mostly via Satellite today), and distributes it to a large set of receiving locations, called Video Hub Offices (VHOs). Each VHO will in turn feed a metropolitan area. IP routers are used to transport the IPTV content in the SHO and VHOs. The combination of SHO and VHO routers plus the links that connect them comprise the long-distance IPTV backbone. The VHO combines the national feeds with local content and other services (described later) and then distributes the content to Video Serving Offices (VSOs) via the routers in the metro IPTV backbone. Finally, the routers or Ethernet Switches (or hybrids thereof) in the VSOs deliver the content down the feeder plant to the Digital Subscriber Line Access Multiplexers (DSLAMs) located in environmentally-controlled vaults, sometimes called Video Ready Access Devices (VRADs), or huts arising from lawns or easements or common space in apartment buildings.
A DSLAM often serves from 100-200 Residential Gateways (RGs) that are attached to the outside of a residence. There are a variety of techniques to transport the signal between the DSLAM and RG. Very High Bitrate Digital Subscriber Line (VDSL) is the typical technology used for copper and Broadband Passive Optical Network (BPON) [ITU-T G.983] and Gigabit PON (GPON) [ITU-T G.984] are the most frequent technologies used for fiber transport. Note the “hub-and-spoke” architecture in the metro/access segment, wherein smaller switches home/concentrate on larger switches. The motivation for this architecture is that as we traverse the network (VHO→VSO→DSLAM→RG) we find that the proportion of the cost of that segment of the IPTV-network increases. To further reduce costs, at each stage the switch more closely resembles a pure Ethernet switch. For example, the packet-switch in the VSO is an Ethernet switch with some additional capabilities found in carrier-grade routers; the switch in the DSLAM closely resembles a pure Ethernet switch. Furthermore, when one considers that the DSLAM switch is most often deployed outdoors, cost savings of the switch itself is imperative.

In the IPTV environment the VHO is envisaged as the nexus for serving the bulk of residential entertainment and telecommunications. In addition to broadcast video illustrated above, this includes Video-on-Demand (VoD), voice, broadband Internet service, and other multi-media services, such as music. For example, the RG delivers video content to the Set Top Box (STB); voice services are either 1) converted to Voice-over-IP (VoIP) and transported as IP flows to the VHO and then to a VoIP gateway or 2) multiplexed in analog form over VDSL. In the latter case, the voice signal is de-multiplexed at the DSLAM and connected to pre-existing Remote Terminal (RT) architectures that use traditional Time Division Multiplexing (TDM) transport technologies, such as DS-0, DS-1 and SONET.

When we examine the long-distance backbone network of Figure 1 in more detail, we see that the SHO and each VHO have two routers to provide redundancy in case of router component failure, router hardware maintenance, or software upgrade. This redundancy is needed to meet the network availability objective discussed previously. The links of the long-distance IPTV backbone are usually 2.5 Gb/s or 10 Gb/s Ethernet or SONET. Because of the use of cost-efficient tree-like topologies, IP multicast provides economic advantages for the delivery of IPTV services. Red arrows illustrate the directed (unidirectional) links of the example long-distance multicast tree in Figure 1. Indeed, note the scarcity of links between SHO routers and its neighbor VHO routers needed to form the broadcast tree (a 1-connected network). However, note that some links designated with dashed lines are not part of the multicast tree. These extra links are provided to enable alternate paths needed for the restoration capabilities described previously. Because of the less-connected nature of most metro telecommunications networks, the topologies of the metro IPTV backbones resemble rings. The links between the VSO switch and metro intermediate router are high speed (e.g., 10Gb/s Ethernet).
and automatic restoration is provided via the establishment of an alternative path between each VSO switch and IO router, generally through another intermediate router. Typically, the access segment of the network (between the VSO and RG) is only 1-connected.

The dashed line in Figure 2 gives an alternate, simplified view of the journey of the broadcast video content that originates at the SHO and ends at the set-top-box. In the SHO and VHO we see Content Acquisition devices (also called A-servers) and Content and Distribution devices (also called D-servers). The live-feed video and other multi-media flows are buffered in the D-server for re-transmission and other applications. The average I/O capacity of a D-server is on the order of several hundreds of Megabits per second. The purpose of the A-servers at a VHO is to gather local content, such as local television stations or community programs, to add to the national content gathered at the SHO. In addition to the broadcast feed, Figure 2 shows Video-on-Demand (VoD) servers. Typically, VoD is gathered at the SHO or other centralized location and distributed in the background (i.e., not usually live) to local VoD servers in the VHO which, because of the need for individual user control, then interact with the STBs via IP connection-oriented (point-to-point) unicast flows. However, research is occurring on more capacity-efficient methods of VoD distribution [12]. Note that servers and routers in the VHO are usually inter-connected via Ethernet switches in a hub-and-spoke topology. The router in the VHO splits out all these IP flows to the various video servers and edge routers of Internet Service Providers (ISPs).

![Figure 2: Flow of broadcast video through IPTV network](image)

It is beyond the scope of this paper to discuss all the protocols used in IPTV transport. However, we make a few salient observations that are relevant to the theme of our paper. The video content uses a differential video compression technology, such as the Moving Pictures Expert Group (MPEG) [13] or ITU-T H.264 standard [14]. The video frames are packetized and carried typically over Real-Time Transport Protocol [15] over User Datagram Protocol (UDP) over IP [16]. Protocol Independent Multicast, Source Specific Multicast (PIM-SSM) [7] is used to multicast the TV content over the IP network. Each channel from the national live-feed at the SHO is assigned a unique multicast group. There are typically hundreds of channels assigned to Standard Definition (SD) (1.5 – 3 Mb/s) and High Definition (HD) (6 – 10 Mb/s) video signals plus other multi-media signals, such as music. Therefore, the live feed can be multiple gigabits-per-second in total bandwidth. This total is expected to grow significantly as the numbers of HD channels grows. As the broadcast signals enter the VSO, since the VSO switch does not have full capabilities of a carrier-grade router, Internet Group Management Protocol [17] is used for the STBs to join the multicast group at the data link layer. In addition, to manage the potential explosion of control messages to the metro routers that could occur from
many switches in the hub-and-spoke architecture, a technique called “IGMP snooping” is implemented wherein the intermediate switches between the STB and the metro intermediate router filter the IGMP messages and locally replicate the channels to those users who request them. This concentration enabled by IGMP snooping also allows the link between the DSLAM and VSO switch to have less bandwidth requirements than that between the VSO and IO router.

Robust Restoration Mechanisms for IPTV Backbones

The previous section described the general network architecture and design for distribution of broadcast video and other services. As mentioned previously, we now will describe the combination of video/packet recovery mechanisms/protocols and underlying network design and restoration methods used to deliver the needed video QoS to the end user. Relatively infrequent and short bursts of loss may be recovered using protocols and mechanisms, such as the Society of Motion Picture and Television Engineers (SMPTE) 2022-1 Forward Error Correction (FEC) [18], retransmission approaches based on RTP/RTCP [15]) and Reliable UDP (R-UDP) [19], and video player loss-concealment algorithms in conjunction with STB buffering. The desired functionality of R-UDP is primarily retransmission-based packet loss recovery. This aspect of R-UDP has been implemented by vendors without incorporating the other mechanisms specified in RFC 1151 [19], such as congestion control algorithms. Besides preventing miscellaneous video impairments due to transmission problems, such as server hiccups or last-mile facility degradation, some combination of these methods can basically recover from any network failure of 50 milliseconds or less. The repair of network failures usually takes far more than 50 milliseconds, but when combined with link-based FRR, this could provide a restoration methodology to meet the stringent requirements needed for video against single link failures. We illustrate how link-based FRR might be implemented in an IPTV long-distance backbone in Figure 3, which depicts a network segment with four node pairs that have virtual links, called pseudowires, defined. For example, node pair E-C has a physical (Layer-1 PHY) link in each direction and a pseudowire in each direction (a total of four directed links) used for FRR restoration. We depict with a red dashed line the FRR backup path for the pseudowire E→C. Note that links, such as E-A, are provided for restoration and, hence, have no pseudowires defined. Pseudowire E→C routes over a primary path which consists of the single PHY link E→C (solid line in Figure 3). If a failure occurs to a PHY link of the primary path, such as illustrated for C-E, then the router at node E attempts to switch to the backup path using FRR. The path from the root node A now switches to the backup path at node E (E-A-B-C), reaches node C, and then continues on its previous (primary) path to node A (C-B-F-A). Note that during the failure, although the path retraces itself between the routers B and C, because of the unidirectionality of the links the multicast traffic does NOT overlap. Also, although the IGP view of the topology realizes that the PHY links between E-C have gone “down”, because the pseudowire from E→C is still “up” and has the least weight, the shortest path tree remains unchanged and, consequently, the multicast tree remains unchanged. The IGP is unaware of the actual routing over the backup path.

Some router/switch vendors have demonstrated an FRR mechanism to switch to the backup path within 50ms [10]. Because the primary path and its backup path are disjoint at the physical layers, no virtual link will fail as a result of any single fiber or WDM lower layer link failure if the hold-down timers are specified appropriately. Thus, for any single link failure, IGP will not detect any change in the IP topology and, therefore, routing and multicast tree will remain unaffected, reducing the impact of the single link failure from tens of seconds to the order of 50 ms.

Because IGP is unaware of the FRR back-up paths during network failures, traffic overlap could potentially occur. Here, “traffic overlap” means that the packets of the same multicast flows travel over the same link (in the same direction) two or more times. Because of the need to deploy economically efficient networks and increasing SD and HD required bandwidth, high link utilization might occur during the normal (no-failure) state; thus, congestion would occur during traffic overlap. For real-time video services, loss from even mild congestion often has the same effect on the customer’s perception of service as an unprotected link failure. Therefore, to exploit the cost reduction made possible by multicast, it is desirable to avoid this potential traffic overlap. Reference [20] presents a scheme for preventing traffic overlap from a single link failure by constructing the multicast tree to be disjoint from the backup path of each link.

The methodology described above, namely, the combination of small-loss recovery mechanisms, FRR, and intelligent selection of non-overlapping paths (such as [20]), would solve the problem of providing high network availability to achieve the stringent video QoS if only single link failures occurred in the network. However this does not mitigate the problem of multiple link failures. The backup path for a given link is typically pre-calculated and there is no real-time (dynamic) accommodation for different combinations of multiple link failures. For example, if another link failure occurs during the outage period of the first failure, then the backup path of a virtual link might fail and thus the IGP routing protocol would reconverge and generate a new multicast tree.
An alternative approach builds on the FRR mechanism, but effectively limits its use to a short period. After a single failure occurs and the primary path of a virtual link fails, the traffic is rapidly switched over to the backup path. However, soon afterwards the router sets the virtual link weight to a very high value and thus triggers the IGP re-convergence process (colloquially called “costing out” the link). Once IGP routing converges, a new PIM tree is rebuilt automatically. This achieves the benefit of rapid restoration from single link failures, yet allows the multicast tree to dynamically adapt to any additional failures that occur during a link outage period. It is only during this short, transient period when another failure could expose the network to potential path overlapping on the same link. However, the potential downside of this approach is that it incurs two more network re-convergence processes, i.e., the period right after FRR has occurred and before repair of the first failure and then again when the failure is repaired (network normalization). Thus, if it is not carefully executed to handle potential small “hits” due to re-convergence, this alternative approach can cause small video interruptions for the more frequent single failures.

To prevent this potential downside, a key component of the method is the make-before-break change of the multicast tree, i.e., the requirement to switch traffic from the old multicast tree to the new multicast tree with minimal loss of traffic. The details of this technique are too complex to describe here, therefore we summarize its steps:

1. Use the algorithm in [20] to set link weights.
2. For each (one-hop, unidirectional) IP-layer link, set the primary path equal to that same (single-hop) link. Pre-compute a FRR backup path for each IP-layer link. The backup path should not overlap with the multicast traffic flow over other primary paths. The link weights generated in Step 1 guarantee that such a backup path exists.
3. When a primary path failure occurs, invoke the FRR mechanism to reroute the traffic to the backup path, provided the backup is operational.
4. Send out an IGP Link State Advertisement (LSA) with a high weight (“cost-out”) for the associated IP-layer link. During the time of IGP re-convergence, traffic is forwarded along the backup path.
5. After IGP re-convergence is completed and its shortest path tree is constructed, PIM rebuilds its multicast tree with join and prune requests, but in such a way that the branches of the new tree are created before the branches of the original tree are pruned and that downstream nodes are not joined to the new tree until traffic flows to their parent nodes. This method incurs a virtually “hitless” multicast reconvergence process.
6. After the failure is repaired, the IGP re-convergence and PIM tree rebuilding process is executed similarly as in step 5.
The make-before-make is similar to the steps used to switch from a shared tree to a source specific tree with PIM as described in [6]. In this way, each receiver is guaranteed to continuously receive packets during the period when switching from one tree to the other.

**Network Performability Evaluation Studies**

Performability (performance + reliability) evaluation analyzes the performance of a network in the presence of failures. The nperf performability analysis tool [21] represents the network under study by a multi-level model. In particular, the “component” level, which is closest to the real network's physical layer, is where all failure mechanisms are modeled. E.g., there are line card components in routers and there are optical amplifier and fiber bundle components, and the failure of any of these affects the graph's edges. Each class of component (e.g., all line cards of a certain type) is represented by a Mean Time Between Failures (MTBF) and a Mean Time To Repair (MTTR). nperf generates network states (failure scenarios) systematically by assigning a “working” or “failed” mode to each of the components defined at the component level. A network state can be conveniently thought of as a subset of failed components. The probability of a component being in each of its modes can be found from the MTBFs and MTTRs specified at the reliability level, and since the components are assumed to be independent, the probability of a network state can be found simply by multiplying together the appropriate component mode probabilities. nperf generates the states in order of decreasing probability of the most likely states. Since the total number of potential states is intractable (exponential in size), nperf uses bounding mechanisms to determine when the objective probabilistic coverage (say, a total probability of 0.9999) is reached and to stop examining further failure states. In each network state, nperf calculates the values of two performance measures: traffic lost because of no path, and traffic lost because of congestion. Both of these measures take into account various protocol-related timing parameters (see Table 1 below).

nperf was used to evaluate the performability of a hypothetical US backbone network with 28 VHOs and 45 links. As already mentioned, each location contains two backbone routers for redundancy, and we consider a VHO reachable as long as at least one of its backbone routers is reachable from the source. Additional details of the network, analysis and the algorithm for weight setting are described in [21]. Table shows the essential reliability and convergence time parameters for our network model.

<table>
<thead>
<tr>
<th>Table 1: Reliability and protocol parameters for nperf model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link MTBF (including the two router interfaces and a fiber)</td>
</tr>
<tr>
<td>Router MTBF</td>
</tr>
<tr>
<td>All MTTRs</td>
</tr>
<tr>
<td>FRR convergence time</td>
</tr>
<tr>
<td>IGP convergence time</td>
</tr>
<tr>
<td>PIM convergence time</td>
</tr>
</tbody>
</table>

With these parameters, nperf is used to generate the most probable network states until the total probability of the evaluated state space reaches the objective tolerance of 0.99999. The traffic impact due to no-path, congestion, and routing protocol convergence, as well as the number of events lasting for more than 1 second per year per VHO are calculated, including all failure events that cause traffic loss due to no path, congestion, and IGP convergence.

Three methods were evaluated for dealing with failures: 1) use only IGP but not FRR, 2) apply FRR whenever backup paths are available 3) use FRR, but only over the interval until the IGP/PIM re-convergence process completes. For each VHO, the total traffic impact in minutes per year is shown in the top graph and the total number of service impacting events per year is shown in the bottom graph of Figure 4. From these graphs, we observe the following: (1) Method 1 would result in a very large number of service impacting events per year for all nodes, since any failure requires IGP routing convergence to restore connectivity; (2) Method 2 reduces the total number of service impacting events per year significantly; however the total traffic impact from congestion due to multiple failures is still large. (3) Method 3 leads to both the smallest number of service impacting events and the least total traffic impact for all VHOs. Detailed investigation of the protocols needed for Method 3 is the subject of our current work.
Figure 4: Comparison of total traffic impact (minutes/year) and service impact (events/year)

Summary

To deliver the needed QoS for distribution of broadcast TV over an IPTV network, one must carefully design an architecture that melds standard IPTV architectures, video/packet recovery mechanisms/protocols, and underlying network design and restoration methods. We described such an implementation based on our hands-on experience with real networks. This implementation uses link-based Fast Reroute (FRR) in the underlying IP transport network to provide the needed high availability. However, if FRR is not coordinated carefully with the placement of network capacity, it can result in congestion and packet loss during network failures. We show that traffic overlap can be avoided during any single IP link failure by tuning the link weights of the routing protocol. In addition, by having FRR, IGP and multicast protocols work in harmony and with appropriate link weight assignments, path overlap during multiple failures can be minimized. We presented network studies that demonstrate how the proposed methods improve service availability and almost completely avoids traffic overlap in all but the very rare cases of simultaneous failures.

References


[18] The Society of Motion Picture and Television Engineers (SMPTE), http://www.smpte.org/standards/.

