Managing Traffic in Multi-Service Designs: Part 1

Rob Munoz, Agere Systems - August 24, 2004

The long-awaited migration from TDM and ATM wide area networks (WANs) to packet networks based on IP and in some cases multi-protocol label switching (MPLS) has clearly begun. These new networks offer the prospect of new revenue-generating converged services, including broadcast video, audio/video conferencing, collaborative editing, video-on-demand (VoD), PPV (pay-per-view movies and content stored in the network and streamed on demand), Internet telephony, Internet radio, premium interactive content, and interactive gaming.

However, this migration is likely to be evolutionary rather than revolutionary. New networking equipment will often interoperate with the installed base during the transition. Interoperating with the installed base of protocols and services while providing services over the new network requires multi-service networking capabilities. Economically and robustly delivering a mixture of high value services over multi-service networks requires paying special attention to traffic management capabilities.

In Part 1 of this two-part set, we'll provide an overview of traffic management and its application in ATM and packet networks. We'll further the discussion by looking at traffic management functions and their associated challenges in multi-service network equipment designs that support both ATM and packet traffic.

In Part 2, we'll examine how traffic management techniques can be employed in a DSLAM design. We'll also discuss traffic management efficiency and effectiveness and the economics of traffic management.

Traffic Management: Why It's Needed

There are essential quality of service (QoS) and other capabilities provided by existing TDM and ATM networks that service providers must still retain in order to economically operate these new networks and reliably deliver new revenue-generating services. High-speed data services and stored audio/video streaming services require bandwidth guarantees while interactive multimedia applications such as video conferencing require both bandwidth and delay guarantees. Network support is also needed for additional control functions to support the various on-demand services. While QoS guarantees and service level agreements (SLAs) are almost taken for granted in TDM and ATM networks, IP networks used for web browsing and email have traditionally only delivered best effort service.

When service providers have attempted to deliver service-level guarantees over their IP-based networks, almost invariably they have over-provisioned their networks with excess bandwidth to do so. In addition to this strategy being completely impractical in bandwidth constrained access networks (e.g. imagine the cost of replacing the existing copper plant with optical fiber), over-provisioning also has questionable economics in many other situations. In addition to the cost of
deploying and operating additional transmission facilities and switching/routing equipment, there is often a reduced incentive for customers to purchase premium services as their best effort traffic also benefits from the increase in network bandwidth.

In access networks, not only is over-provisioning not used, but the exact opposite, oversubscription, is the norm. Service providers do not expect all customers to simultaneously use their broadband service at full bandwidth. Oversubscription is widely used to help keep network costs low and thereby keep services affordable. When congestion inevitably occurs, some portion of the traffic will suffer large delays and/or be discarded. It is the role of traffic management to intelligently determine what traffic should be discarded and what traffic should be scheduled ahead of other traffic. Clearly, without some kind of traffic management, it would be impossible to deliver service-level guarantees when congestion occurs.

Even when service-level guarantees are not required, traffic management is still required to maintain network efficiency. Because communication networks, much like highways, are congestion prone systems, past a certain point increases in offered load result in a decrease rather than an increase in network throughput (Figure 1). Traffic management is required to manage the offered load to keep the network operating at peak efficiency.

![Figure 1: Diagram showing normalized network throughput vs. load.](image)

Traffic management functions include traffic policing, queuing, buffer management, scheduling, and shaping. Proper traffic classification is required in advance to identify traffic that must be treated differently and packet/cell modification is required to appropriately forward traffic and mark it if necessary. While traffic management functions are often important in many network settings, they are particularly critical to service provider networks. Network operators have large amounts of capital invested in transmission facilities and equipment. It is therefore imperative that they maximize the revenue they can extract from these large, relatively fixed cost investments so that they can keep the services that they deliver affordable.

**Specifying Traffic Management Treatment**

The goal of traffic management is to manage the flow of traffic so that the revenue that the network supports is maximized. As was mentioned earlier, this requires both that service-level guarantees are delivered to premium traffic and that the overall traffic load on the network is managed to keep the network as a whole operating efficiently. In order to do this, each type of traffic must implicitly or explicitly describe itself and indicate the treatment it expects to receive from the network. The
network in turn must deliver satisfactory treatment and ensure that the traffic only receives the network resources it is paying for and does not impact any quality of service guarantees promised to other traffic.

The framework for doing managing traffic in ATM networks is widely deployed and provides strong QoS guarantees. IP and MPLS networks support several different frameworks for doing managing traffic that typically provide weaker QoS guarantees and have not yet been very widely deployed. As we will see, there are a number of common functions and mechanisms which are required in some form by each of these frameworks for proper traffic management.

**ATM Framework**
The traffic management framework used in ATM networks is defined in the TM4.1 specification. Since ATM traffic is carried over end-to-end virtual circuits (VCs), it is relatively straightforward to associate the traffic management treatment required for end-to-end ATM traffic with the underlying ATM connection.

Service requirements for each connection are specified in an ATM traffic contract. The network uses connection admission control (CAC) to determine whether there are adequate network resources to admit a new ATM connection while satisfying the requirements specified in the traffic contract. An ATM traffic contract includes:

- The service category
- The QoS parameters required for the connection
- A conformance definition which defines how traffic should behave according to the contract

For switched virtual circuit (SVC) ATM connections, the application uses ATM signaling to negotiate the traffic contract with the network. For permanent virtual circuit (PVC) ATM connections, the network management system specifies the traffic contract.

**ATM Service Categories**
The service category provides a way to group together applications with similar treatment requirements. The service categories defined are:

- Constant bit rate (CBR)
- Variable bit rate (VBR)
- Available bit rate (ABR)
- Guaranteed frame rate (GFR)
- Unspecified bit rate (UBR)

CBR service is intended to support real-time applications (such as voice or circuit emulation) that require low delay and low jitter (variation in delay). Typically, bandwidth for the peak transmission rate of the connection (which is called the peak cell rate [PCR] in ATM) is statically reserved to deliver the low delay and jitter required, though in some cases oversubscription may be used as long as loss, delay, and jitter guarantees can still be met with sufficiently high probability.

There are two subcategories of VBR service: real-time VBR (rtVBR) and non-real-time VBR (nrtVBR). rtVBR has tight end-to-end delay requirements making it suitable for interactive video and gaming traffic. nrtVBR does not guarantee delay and is therefore suitable for streaming video and similar applications. Because it is not as delay sensitive, nrtVBR can be more easily oversubscribed as buffering can be used to tolerate transient congestion without discarding cells. The PCR, sustainable cell rate (SCR), and maximum burst size (MBS) traffic parameters characterize the bandwidth requirements of VBR connections.
ABR service uses a rate-based flow-control mechanism to dynamically limit the amount of bandwidth a connection uses to minimize congestion-related cell or frame loss. Due to its complexity and limited effectiveness compared to other ways of efficiently carrying frame-based traffic over ATM networks, ABR has seen very little deployment in production networks.

GFR service is intended for non-real-time frame-based applications (e.g. frame relay) that may desire a minimum guaranteed rate (which is called the minimum cell rate [MCR] in ATM). Since buffering is used to provide low loss, the MCR is provided in the context of a maximum burst size (MBS) and a maximum frame size (MFS) The service also attempts to discard whole packets rather than just individual cells when congestion does occur. Traffic that conforms to the parameters in the service contract will be delivered with minimum loss, but there are no loss guarantees for any traffic above the guaranteed minimum.

UBR service is intended for non-real-time, "best effort" traffic (e.g. email). There are extensions to the baseline TM4.1 specification that allow applications to request a minimum desired cell rate (MDCR) and to specify a behavior class (BCS) with UBR service.

**ATM QoS Parameters**
The TM4.1 specification identifies six QoS parameters to measure network performance for a particular ATM VC. The first three of these are negotiated and intimately associated with traffic management. These three parameters are:

1. **Maximum Cell Transfer Delay (maxCTD):** The maximum delay tolerated between sending a cell at the originating end of the virtual circuit and receiving the cell at the receiving end. Cells that experience a delay higher than maxCTD are assumed to be unusable.
2. **Peak-to-Peak Cell Delay Variation (CDV):** The difference between maxCTD and the minimum cell transfer delay (i.e. jitter).
3. **Cell Loss Ratio (CLR):** The ratio of cells lost to total cells transmitted.

Table 1 summarizes the attributes of interest to each of the ATM service categories.

**Table 1:** Table 1: ATM Service Category Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>CBR</th>
<th>rt-VBR</th>
<th>nrt-VBR</th>
<th>UBR</th>
<th>ABR</th>
<th>GFR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traffic Parameters:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCR and CDVT</td>
<td>Specified</td>
<td>Specified</td>
<td>Specified</td>
<td>Specified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCR, MBS, CDVT</td>
<td>n/a</td>
<td>Specified</td>
<td>Specified</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCR</td>
<td>n/a</td>
<td>Specified</td>
<td>n/a</td>
<td>Specified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCR, MBS, MFS, CDVT</td>
<td>n/a</td>
<td>Specified</td>
<td>n/a</td>
<td>Specified</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>QoS Parameters:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak-to-peak CDV</td>
<td>Specified</td>
<td>Specified</td>
<td>Unspecified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxCTD</td>
<td>Specified</td>
<td>Specified</td>
<td>Unspecified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>Specified</td>
<td>Unspecified</td>
<td>See Note 1</td>
<td>See Note 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congestion Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td>Unspecified</td>
<td>Specified 6</td>
<td>Unspecified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Attributes:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDCR</td>
<td>n/a</td>
<td>Optional</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCS</td>
<td>Unspecified</td>
<td>Optional</td>
<td>Unspecified</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes for Table 1:**

1. CLR is low for sources that adjust cell flow in response to control information. Whether a
quantitative value for CLR is specified is network specific.
2. Might not be subject to CAC and UPC (usage parameter control) procedures.
3. Represents the maximum rate at which the ABR source may ever send. The actual rate is subject to the control information.
4. These parameters are either explicitly or implicitly specified for PVCs or SVCs.
5. CDVT refers to the Cell Delay Variation Tolerance (see Section TM4.1 4.4.1). CDVT is not signaled. In general, CDVT need not have a unique value for a connection. Different values may apply at each interface along the path of a connection.
6. See Section 2.4 of TM 4.1 for further information.
7. CLR is low for frames that are eligible for the service guarantee. Whether a quantitative value for CLR is specified is network specific.
8. See Section 5 of TM4.1 for details of ABR feedback and other congestion control mechanisms.
   The MDCR parameter is not considered a traffic parameter because AF-TM-0150 does not define a service commitment based on MDCR.

**ATM Traffic Conformance Management**

To guard against connections using more than their agreed to share of network resources (which would potentially impact the network's ability to guarantee QoS to other more well-behaved connections), it is important to be able to ensure that a connection conforms to its traffic contract. This role falls to the traffic policing and the traffic shaping functions. Both traffic policing and traffic shaping monitor or meter the rate of arriving traffic relative to the traffic contract (or other desired traffic profile).

The generic cell rate algorithm (GCRA) provides a standard way to analyze conformance with the traffic contract. As shown in Figure 2, there are two equivalent forms of the GCRA algorithm: the virtual scheduling algorithm and the continuous state leaky bucket algorithm. Each of these takes two parameters: an increment I and a limit L, where the bucket "leaks" at a rate of 1/I and has a depth of L. Analyzing VBR conformance requires monitoring of both the peak rate and the sustained rate. This calls for use of a dual-leaky bucket (or its virtual scheduling equivalent).
Shaping delays the cells of a connection if needed to make them conform to the desired traffic profile. This controlled delay of cells also reduces the cell delay variation (CDV) that the traffic will experience. Buffer space is required to hold the cells being delayed. If the arrival rate exceeds the shaped rate for an extended period of time, the shaping buffer will fill up and cells must be discarded.

When the policing function detects that a cell is non-conforming or does not fit the desired traffic profile, the traffic management function can elect to:

- Discard the non-conforming cell.
- Tag the congestion loss priority (CLP) bit in the non-conforming cell's header. This indicates to downstream systems that the cell was non-conforming and can be discarded later if needed.
- Shape the connection (which is termed soft policing when it is done as part of the policing function) so that cells are sent no faster than the traffic contract or other traffic profile specifies.
- Do nothing (as long as the network can still guarantee that the QoS guarantees of other traffic will not be violated).

Queuing allows each group of traffic to wait in a separate line while it awaits transmission. Buffer management controls how much buffer space each group of traffic is allowed to occupy while it is waiting and decides which individual cells must be discarded as buffer occupancy grows. The scheduling mechanism determines when each queue is allowed to transmit a cell. Shaping can be viewed as a form of scheduling which carefully spaces out cell transmissions to fit the desired traffic profile. The ATM standards do not specify how these mechanisms are implemented, leaving it open to vendor differentiation.

Figure 2: Equivalent versions of the generic cell rate algorithm.⁸
Specifying Traffic Management Treatment in IP Networks

IP networks support several different frameworks for specifying traffic management treatment. Some prominent examples are integrated services (IntServ),
differentiated services (DiffServ), and multi-protocol label switching (MPLS). Commonality between these approaches includes:

- A distinction between edge and core routers
- Edge routers accept traffic from outside the network
- Core routers provide transit forwarding service between other core routers and/or edge routers
- Edge routers are responsible for characterizing, policing, shaping, and/or marking incoming traffic
- Edge routers may use admission control to decline requests for network service
- Core routers differentiate traffic when necessary to manage transient congestion within the network

IntServ defines two service models to augment conventional best effort internet service. The first is a guaranteed service that ensures datagrams will arrive within the guaranteed delivery time and will not be discarded due to queue overflows, provided the flow's traffic stays within its specified traffic parameters. The guaranteed service does not attempt to minimize the jitter but merely controls the maximal queuing delay.

The second is a controlled load service. This service approximates the behavior visible to applications receiving best-effort service under unloaded conditions from the same series of network elements.

An IntServ application would typically use RSVP to signal its QoS requirements to the network. A source characterizes the traffic it will offer using a SENDER_TSPEC element in an RSVP PATH message with the following parameters:

- Token bucket rate $r$ in bytes per second
- Token bucket size $b$ in bytes
- Peak data rate $p$
- Minimum policed unit $m$
- Maximum packet size $M$

A receiver describes the desired QoS to be applied to the sender's traffic using a FLOWSPEC element in an RSVP RESV message. Multi-field classification (covering source and destination IP address, source and destination TCP/UDP port, and protocol type) is used to identify flows and provide per-flow policing and queuing.

The operation of the token bucket is somewhat similar to ATM's leaky bucket but must account for variable length packets versus fixed length ATM cells. Tokens are added to the bucket at a rate of $r$ tokens per second up to a maximum of $b$ tokens in the bucket. When a packet of length $L$ bytes arrives, if the bucket has at least $L$ tokens the packet conforms to the token bucket and $L$ tokens are removed from the bucket. If the packet doesn't conform no tokens are removed from the bucket.

Owing to scalability concerns around maintaining per-flow state information in all network elements, IntServ has seen scant deployment in public networks. However, the lessons learned developing IntServ have clearly influenced subsequent IP QoS approaches.

DiffServ

DiffServ provides a class of service approach to IP quality of service. It redefines six bits of the type-of-service (TOS) byte in the IP header as the differentiated service code point (DSCP).
Edge routers perform multi-field classification (including examination of the incoming DSCP) to identify different types of traffic and match traffic flows with their associated traffic conditioning agreement (TCA). The TCA indicates both what conditioning (policing and/or shaping) is required as well as how to mark the DSCP in the packet header. The token bucket-based trTCM and srTCM algorithms are often used to meter and mark DiffServ traffic. Core routers need only examine the DSCP to determine QoS treatment. By doing this, the core routers do not have to maintain per-flow state, greatly facilitating scalability.

Rather than defining end-to-end service classes, the DiffServ effort has instead focused on defining the traffic management treatment at each node in the network in the form of per-hop behaviors (PHBs). Four standard PHBs have been defined:

1. Default PHB: provides traditional best-effort service.
2. Class selector PHB: provides backward compatibility with previous uses of the TOS byte.
3. Assured Forwarding (AFxy) PHB.
4. Expedited Forwarding (EF) PHB.

The behavior AF provides is similar to IntServ's controlled load service. AFxy actually refers to a group of PHBs with scheduling class x (1-4) and drop precedence y (1-3) for a total of 12 possible combinations. Conceptually, traffic from each service class is queued in its own queue. The AF PHB requires active queue management such as weighted RED (WRED) or similar algorithms would provide so that short-term traffic bursts are buffered without loss while long-term congestion results in selectively dropping packets.

The EF PHB provides the equivalent of a virtual leased-line service. The departure rate of EF traffic from any DiffServ node must be greater than or equal to a configurable rate. The EF traffic should obtain this rate independent of any other traffic. If this behavior is implemented by a mechanism that allows unlimited preemption of other traffic (e.g. strict priority queuing), the node must rate limit (e.g. via policing and/or shaping) EF traffic to a configured maximum rate and burst size.

MPLS
In MPLS, packets entering the network are assigned to a particular forwarding equivalence class (FEC) and class of service by an ingress MPLS label edge router (LER). Since a packet is assigned to a FEC and class of service when it enters the network, the ingress LER may use any information it has about the packet (including incoming port or information from headers in the arriving packet) to make the assignment.

The FEC and class of service to which the packet is assigned are encoded as a short fixed-length value known as a "label". An MPLS label is similar to the ATM virtual path indicator/virtual channel indicator (VPI/VCI). The path through the MPLS network that a particular FEC will take is referred to as a label switched path (LSP) and is analogous to an ATM connection.

A leading approach to provide QoS using MPLS is to support DiffServ PHBs over an MPLS network. In this approach, there are two ways of specifying scheduling class and drop precedence. The first is to use the 3-bit EXP field in the MPLS shim header to distinguish between classes with the same label. The mapping from the EXP field to the service class and drop precedence for a given such LSP is either explicitly signaled at LSP set-up or relies on a pre-configured mapping.

The second is to use distinct label switched paths (LSPs) for each class. With this approach, the scheduling class is explicitly signaled at the time of label establishment, so the LSR can infer exclusively from the label value the scheduling class that should be applied. When the MPLS shim header is used, the drop precedence is conveyed using the EXP field. When the shim header is not
used (e.g., MPLS over ATM), the drop precedence to be applied by the LSR to the labeled packet is conveyed inside the link layer header encapsulation using link layer specific drop precedence fields (e.g., ATM CLP).

On to Part 2
That wraps up Part 1 in our two-part series on managing traffic in multi-service networking equipment designs. In Part 2, we'll further the discussion by looking at the implementation of traffic management techniques in a DSLAM design. We'll also examine the effectiveness, efficiency, and costs associated with various traffic management approaches. To view Part 2, click here.

References
1. The last three sections of this paper are based on Deepak Kataria and Rob Munoz's paper and presentation "Quantifying the Economics of Traffic Management" at the 2004 Communications Design Conference.
7. AF-TM-0121.000 table 2-1, p. 6 extended with information from AF-TM-0149.000 and AF-TM-0150.000.
8. AF-TM-0121.000 figure 4-1, p.25.
11. IETF RFC 3031, "Multiprotocol Label Switching Architecture", 1/01.

**About the Author**

**Rob Munoz** is a systems engineer at Agere Systems. Rob holds an MS degree from the University of Texas, Austin, and a BS degree in computer science from Virginia Tech. Rob can be reached at rjmunoz@agere.com.