dvjUpperClass2003Nov08.fm

November 10, 2003 12:58 pm

Refined fairness

This contribution provides replacement text for Clause 9 of P802.17 RPR D2.6. Two types of changes are proposed:

a) High-class fairness. Queue-depth based flow-control protocols ensure sufficient and verifiable high-class (classA1 and classB) allocations.

The following TBDs are being considered.

a) Curves to illustrate supportable levels of classA1 and classB-CIR would help illustrate concerns.

3.2 New definitions

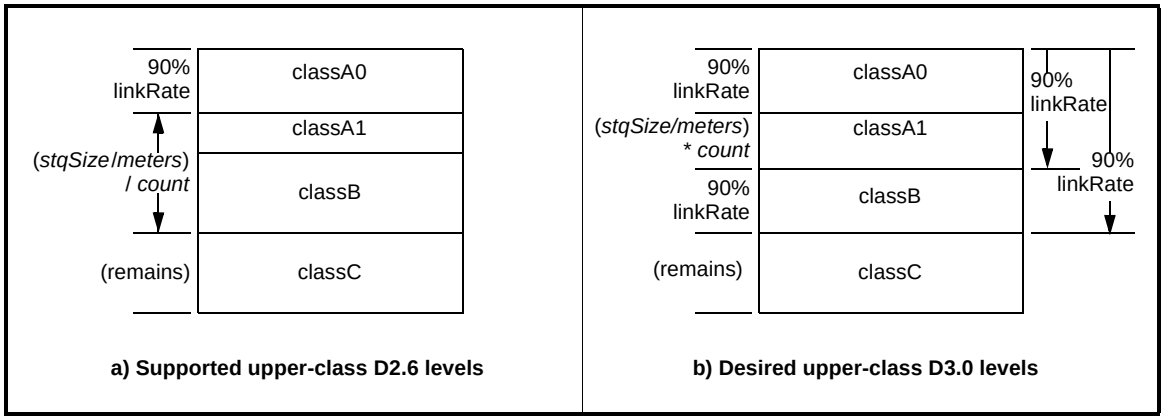
3.2.1 high-class: An adjective that refers to either of subclassA1 or classB-EIR traffic classes.

3.2.2 upper-class: An adjective that refers to either of subclassA0, subclassA1 or classB-EIR traffic classes.

6.3.1 Supported upper-class allocation levels

As of D2.6, with its generalized downstream shaper, any level of classA0 traffic can be supported. Unfortunately, the levels of supportable classA1 and classB levels are still heavily constrained, as illustrated in Figure-Cell 6.12-a, much less than the desirable levels of Figure-Cell 6.12-b.

Figure 6.12—Supported allocation levels (two options)



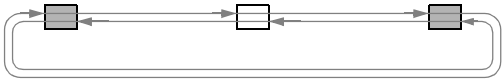
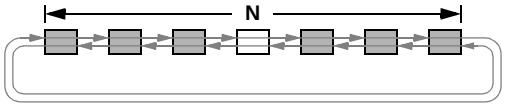
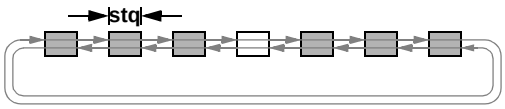
The problems with D2.6 (illustrated in Figure-Cell 6.12-a) can be severe: the *F*(*stqSize*) function can severely constrain levels of supportable subclassA1/classB-EIR bandwidths, even thought a station's STQ has been sized to match the topology's ring round trip time (RRTT).

An apparent WG goal was to support near-arbitrary levels of classB bandwidth, independent of the STQ sizing. The STQ size needn't limit levels of supportable classB traffic, since the classB traffic may have longer latencies (on the order of the RRTT). These longer latencies provide sufficient time to throttle upstream stations, after which the blocked classB transmissions can resume.

6.3.2 STQ depth-control benefits

The D2.6 equations acknowledge the upper-class bandwidth limitations described in 6.3.1. A direct application of these (somewhat optimistic) equations validates that supportable levels of high-class bandwidths decreases as 1/N, where N is the number of stations attached to a fixed-diameter ringlet, as illustrated in Figure-Row 6.13-3. Similarly, the levels of supportable high-class bandwidths decreases when existing stations are upgraded with larger STQs, as illustrated in Figure-Row 6.13-4.

Figure 6.13—Relative levels of supportable classA1 bandwidth

Current D2.6 limitations	Row	Current D2.6 spec	Proposal revisions
	1	☹My classA1 allocation depends on others' STQs	☺My classA1 allocation ensured by my STQ
	2	☹My classB allocation depends on my STQ	☺My classB allocation is nearly the link bandwidth
	3	☹My (classA1+classB) allocation limit degrades as O(1/N)	☺My classA1 allocation limit improves as O(N)
	4	☹My (classA1+classB) allocation limit degrades as others' STQs increase	☺My classA1 allocation limit improves as others' STQs increase

The baseline problem with D2.6 upper-class guarantees is the lack of STQ-depth controls, as follows:

a) Delayed indications. Upstream congestion conditions are delayed by the dynamics of the fairness protocols. (The fairness protocols typically have time constants on the order of several LRTTs.)

b) Uncontrolled release. Upstream stations do not limit the rates at which their STQs empty. This allows the STQ to empty at high (linkRate-classA1) rates, thereby swamping transmissions of downstream classA1 and classB traffic.

Simple solutions can address each of these problems, as follows:

a) Immediate indications. Upstream congestion conditions are sent when STQ reach critical threshold values, without regard to the fairness control protocol dynamics.

b) Throttled release. Upstream stations limit the rates at which their STQs empty, while downstream congestion indications remain asserted.

6.4 Upper-class congestion management

6.4.1 Spillover failures

FIFO spillover controls involve monitoring the STQ depth, as illustrated in Figure 6.14. Each station communicates its STQ depth conditions to its upstream station, which concurrently limiting its downstream transmissions based on its congestion condition relative to its downstream station. While the specification is based on precise FIFO depth, coarse granularity of FIFO-depth measurements are allowed.

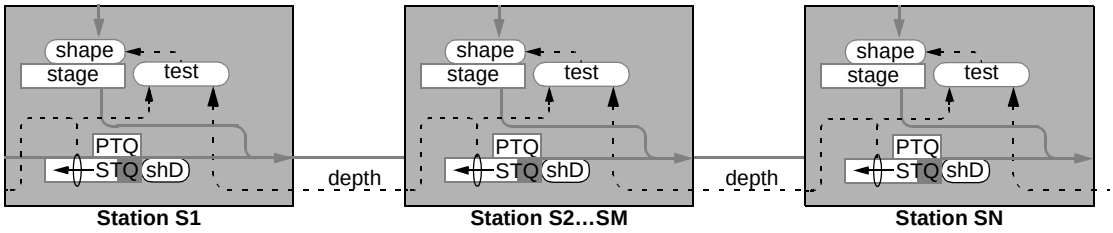


Figure 6.14—Spillover failures

Within such topologies, STQ-based failure scenarios are based on transient high-class transmissions, as follows:

a) Leftside station S1 sends constant near-line-rate classC traffic.

b) Center stations S2 through SM send transient classB traffic, thereby partially filling their STQs. Until these queues empty, STQ-based transmissions continue at a high (linkRate-subclassA0) rate.

c) Rightside station SN sends transient subclassA1 and classB traffic, which fills its STQ. A priority inversion occurs when its STQ approaches full: bandwidth and latency guarantees are violated.

The STQs are effective in relieving transient excesses, but their uncontrolled release causes sustained flooding (and hence transportation delays) within downstream stations. Increasing STQ sizes is an ineffective solution: the additional station SN capacity is largely consumed by the increased duration of STQ releases in upstream station S2-to-SM.

6.5 Dual-queue MAC

The transmit behavior of a dual-queue MAC is described by its transmit-selection protocols (described in this subclause) and shaping functions. A dual-queue MAC uses two transit queues, the primary transit queue (PTQ) for classA traffic, and the secondary transit queue (STQ) for classB and classC traffic. The size of the both transit queues is left to the implementations. The size of the secondary transit queue determines its flow-control threshold values. The dual-queue design is described in following subclauses.

6.5.1 Dual-queue MAC datapaths

A dual-queue MAC makes use of two transit queues, as shown in Figure 6.15.

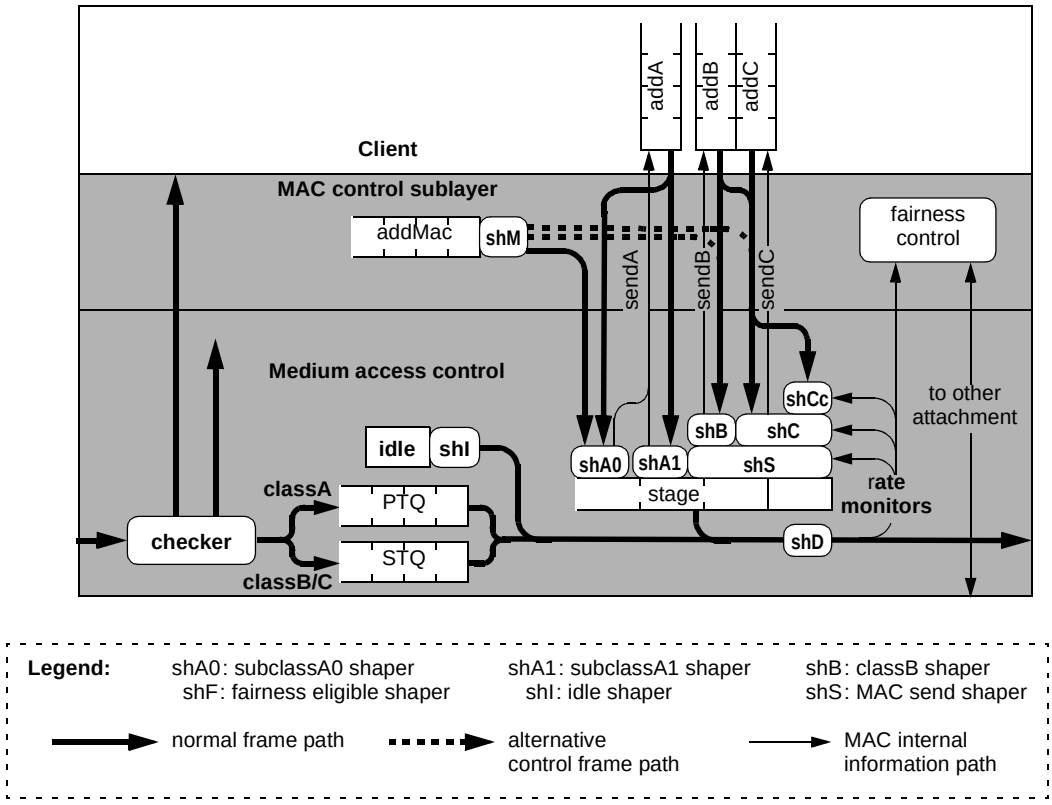


Figure 6.15—Dual-queue MAC datapath

NOTE—The source shaper has been modified slightly from D2.4, to include classB as well as classC traffic. This is necessary to avoid STQ starvation once the classB allocation limits no longer depends on the STQ depths.

NOTE—The downstream shaper has been modified slightly from D2.4, as described in 6.5.2.

6.5.2 Queue-depth indications

The behavior of the downstream shaper depends on the depths of the STQs in this station and its downstream neighbor, called *myLevel* and *rxLevel* respectively, as shown in Figure 6.16.

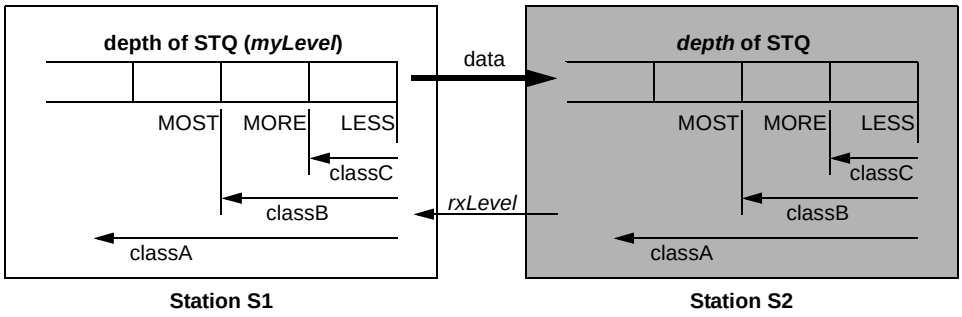


Figure 6.16—Dual-queue MAC datapath

A field in the fairness message reports the depth of the downstream station's STQ, with ranges listed below:

MOST The queue depth allows only classA traffic to be added.
MORE The queue depth allows classA and classB-CIR traffic to be added.
LESS The queue depth allows classA, classB, and classC to be added.
(Fairness protocols may independently throttle classB-EIR and classC transmissions.)

6.5.3 Depth-dependent creditD values

The rateD rate limit of the downstream shaper depends on the queue depth of this station, as described by Table 6.1. The system dynamics and the selected *loLimitD* value avoids having *creditD* decrease below zero. The top-through-bottom rows have high-through-low precedence.

Table 6.1—Depth-dependent rateD values

STQ level conditions	Row	Results	
		rateD	non rateD contributions
myLevel==MOST	1	ringTotals - ringAllocA0 - thisAllocA1;	myAddsA0 ringThruA0 myAddsA1
myLevel<rxLevel	2	ringTotals - ringAllocA0 - ringAllocA1 - thisAllocB	myAddsA0 ringThruA0 myAddsA1 myAddsB
—	3	ringTotals - ringAllocA0	myAddsA0 ringThruA0

#define ringTotals (ringAllocA0 + ringAllocA1 + ringAllocB + ringUnallocatedC)

Row 6.1-1: When this station is heavily congested, *rateD* is increased to restrict its STQ fill rate to what would occur when the maximum subclassA1 traffic is sent. The intent is to delay STQ overflow until relief results from sending upstream queue-level indications, that throttle upstream lower-class traffic.

Row 6.1-2: When the downstream station is more congested, *rateD* is decreased to a level that allows downstream classB transmissions without further filling of the downstream station's STQ.

Row 6.1-3: When queues are not threatened, *rateD* is only restricted by *ringAllocA0* reservations.

6.5.4 ClassC tapering

While within the LESS range, the intent is to allows classC transmissions at low-depth levels while tapering the allowed classC rates as the LESS limit is approached, as illustrated in Figure 6.17. (Or, whatever else that Clause 9 evolves to, with a specific shut-off limit. This is not a fundamental part of this proposal.).

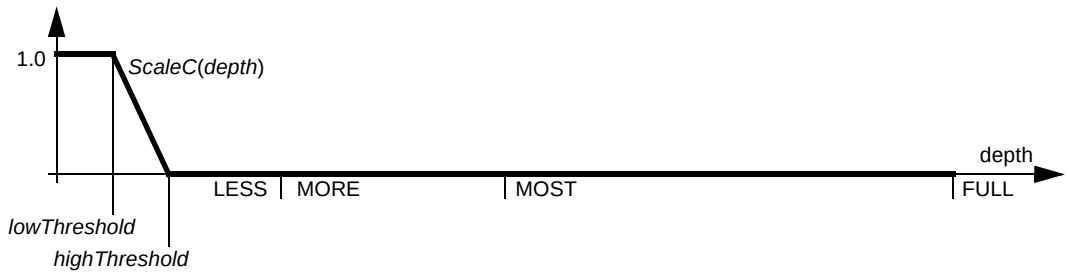


Figure 6.17—*ScaleC*(*depth*)

6.5.5 STQ level normalization

The level normalization of STQ-depth information differentiates between LESS, MORE, and HIGH levels, while providing finer depth granularity within the MORE depth interval, as illustrated in Figure 6.18. The finer granularity is intended to improve the efficiency of classC transmissions by reducing the magnitude of classB bandwidth compensations.

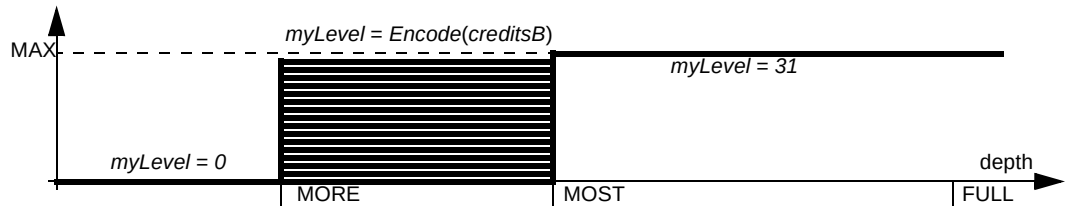


Figure 6.18—Communicated myLevel indication

When below the MORE level, a *txLevel* value of 0 is reported, since there is sufficient STQ queue space to send significant classB traffic. Within the MORE interval, the communicated *txLevel* value is derived from a nonlinear measure of excess classB credits, *myLevel*, so that excess classB credits can be equally shared. When within the MOST interval, the largest *txLevel* value is reported, since classA traffic is threatened.

The *myLevel* value is a nonlinear function of the classB credits, to handle a wide range of accumulated credit values, as specified in Equation 6.1. Each power-of-two increase in accumulated classB credits corresponds to one-larger *myLevel* value.

$$myLevel = \mathrm{Min}(30, \log2(creditsB)); \quad (6.1)$$

Where:

creditsB corresponds to the number of classB bytes that the shaper allows to be sent.

9.3 Fairness frame format

NOTE—This proposal requires small changes to the fairness frame contents, as illustrated below. The reserved bits are sufficient (from all of the author's previous investigations) to also scope the validity of the congestion indication, as might be considered in the future to enhance fairness bandwidth utilization.

The fairness frame payload contains a 16-bit *fairnessHeader* and a 16-bit *fairRate* as illustrated by Figure 9.19

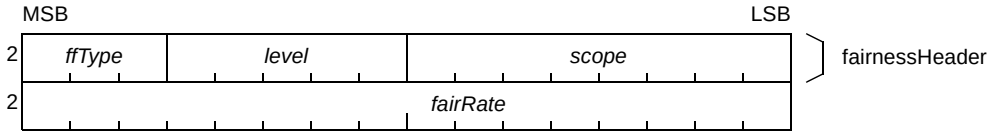


Figure 9.19—Fairness frame payload

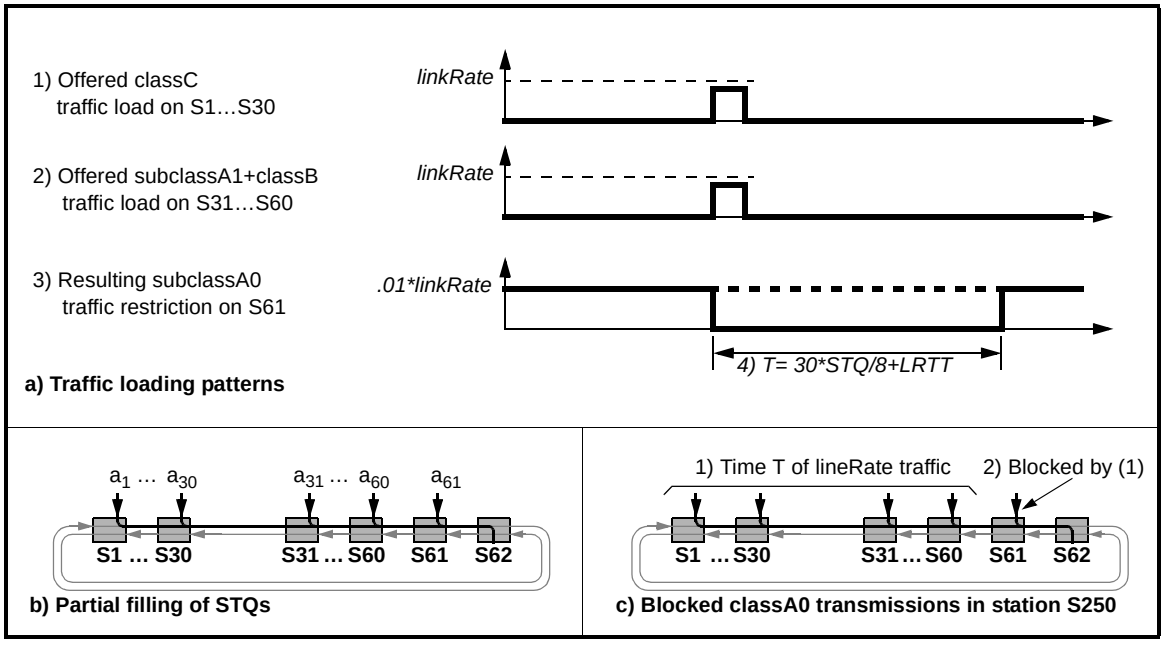
9.3.1 ClassA0 blockage scenario

The purpose of the secondary transit queue (STQ) is to queue incoming traffic until transient congestion conditions can be communicated and resolved. Since the STQs can fill in normal operations, the allocation and fairness protocols should function correctly when the STQs are empty, slightly filled, or nearly full.

To illustrate potential STQ filling problems, consider the traffic loads of Figure-Cell 9.20-a applied to topology of Figure-Cell 9.20-b, leading to the classA0 bandwidth guarantee failure of Figure-Cell 9.20-c. For this illustration, assume:

a) Stations S1-to-S30 generate cumulative classC traffic loads of .99**linkRate*.
b) Stations S30-to-S31 simultaneously burst subclassA0+classB traffic at 3% of link rate.
c) Stations S1-to-S61 transmissions are destined for the S62 station.
d) Stations S31-to-S60 have large STQs, with high&low thresholds of STQ/4 and STQ/8.
e) Single-queue station S61 is allocated .01**linkRate* classA0 capacity.
f) The ring latencies are dominated by a loop round trip time (LRTT) of the S30-to-S31 link.

Figure 9.20—ClassA0 blockage scenario



With an unfortunate timing of the offered *a31…a60* add traffic, their STQs can fill to the low threshold (1/8 of the STQ) before a congestion condition is communicated. An additional delay of LRTT occurs before the cumulative upstream classC traffic from *a1…a30* can be stopped. As a result, the cumulative fill levels of the S31…S60 queues equals a time duration *T=30*STQ/8+LRTT*.

Because there is no downstream shaper on the STQs of stations S31…S60, these stations continue to transmit at the *linkRate* until their STQs have emptied. Thus, the downstream single-queue S61 station is blocked for at least the duration *T*, which directly effects the worst case classA0 traffic jitter. The LRTT delays are significant; for large buffer sizes, the STQ related delays become intolerable.

9.3.2 ClassA0 blockage avoidance

The classA0 blockage is caused by the partial filling of upstream STQ buffers, allowing the upstream stations to effectively exempt the transiting traffic from downstream shapers. Two types of solutions are therefore possible:

a) Revise the transmission protocols, so that upstream STQs are never filled.

b) Improve the effectiveness of the downstream shaper, so that STQ traffic is no longer exempted.

The viability of a type (a) solution is unlikely, so a type (b) solution was proposed. The downstream shaper is applied to all conflicting traffic: STQ and added classB, in addition to added classC. There is no need to limit the added classA1 traffic, since subclassA0 and subclassA1 traffic have equal precedence.

9.3.3 ClassA1 blockage avoidance

The same problem exists with classA1 blockage, in that the burst of STQ-sourced traffic from upstream stations can throttle a downstream station attempting to send classA1 traffic, because its STQ fills.

The solution is to limit the burst rate of STQ transmissions when the downstream station is threatened.

9.3.4 ClassB blockage avoidance

The same problem exists with classB blockage, in that the burst of STQ-sourced traffic from upstream stations can throttle a downstream station, because its STQ fills. Note that large levels of allocated classB traffic cause problems, when the STQ design is limited and only much smaller levels of classA1 traffic can be supported.

The solution is to limit the burst rate of STQ transmissions when the downstream station accumulated creditsB are larger than those of the upstream stations.