P802.1Qat Delay and Bandwidth Parameterization

Parameters for delay and bandwidth capacity calculations for IEEE P802.1Qat SRP

Version 1

Norman Finn

Cisco Systems

Disclaimer

- I would be surprised if this whole presentation is not in a textbook, already.
- But, I have not read that textbook.
- In the meantime, here is the information.
- If someone can provide a reference to the textbook, the Task Group, including me, would be grateful.

Introduction

Introduction

- The current revision of the assumptions document says:
  - Maximum Interference Amount per Hop
    - Class A: 1 Max size frame + Sum of the Maximum size of the Class A frames on each of its other ports – Ref 5
    - Class B: 1 Max size frame + 1 Max size Class A burst (based on max Class A BW allocation) + Amount of other Class B frames on each of its other ports
- This presentation will attempt to define what “Max size Class A burst” means, and extend the concept to any number of Classes.
- This will lead us to the appropriate management parameters to use to characterize the per-Class and Per-Port limitations on bandwidth reservations.

Latency Calculations

Worst-case latency contributions

- The worst case latency for a single hop from Bridge to Bridge, measured from arrival of the last bit at Port *n* of Bridge A to the arrival of the last bit at Port *m* of Bridge B, can be broken out into the following components:
  - Input queuing delay. (There are no input queues in the 802.1 architecture, but if present, the implementation must account for them.)
  - Output queuing delay. (The subject of this presentation.)
  - Frame transmission delay. (One maximum frame time at output line rate for non-cut-through architecture.)
  - LAN propagation delay. (Depends on length of output wire, measured by P802.1AS.)
  - Store-and-forward delay. (Includes all forwarding delays, assuming that the input and output queues are empty.)

Store and forward delay

- Store and forward delay includes all delay causes other than those enumerated in the previous slide. This would include, for example:
  - Time needed to pass from the input port to the output port, assuming empty queues.
  - The difference, if any, in the delay incurred by a frame that bypasses an empty queue, vs. that incurred by a frame that must be enqued.
  - Time added by the lengthening of the frame due to additional frame headers such as Q-tags or Sec-tags (may be negative).
  - Time needed to encrypt an 802.1AE frame.

Output queuing delay

- The output queuing delay for frame X, in turn, can be broken out into the following components:
  - The frame that was selected for transmission an arbitrarily small time before frame X arrived (became eligible for transmission selection).

    This is well understood – it is “one max sized frame”.
  - The delay caused by queued-up frames from all 802.1Qat frames with higher priority than frame X’s class (e.g., the “max size Class A burst”).

    This is the tricky part.
  - The fan-in delay caused by other frames in the same class as frame X.

    Fan-in delay is explained in the following slide.

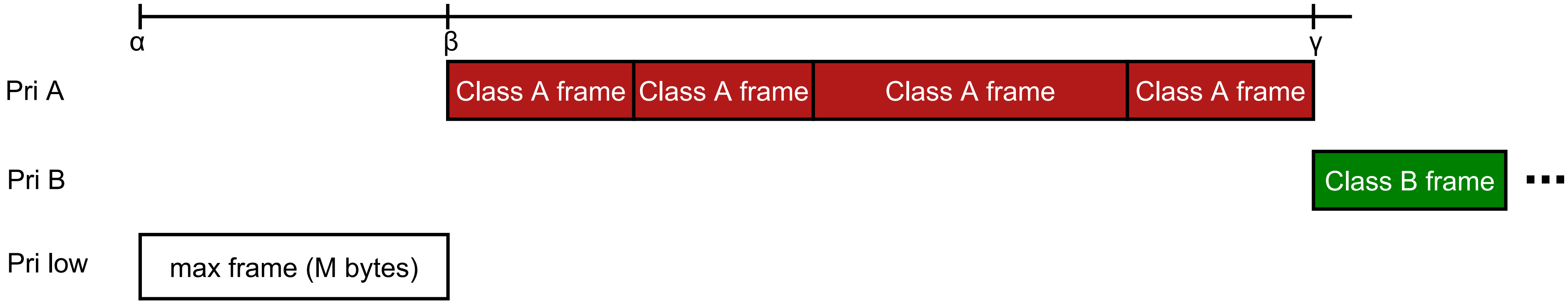
Fan-in contribution

- Both “Sum of the Maximum size of the Class A frames on each of its other ports” and “Amount of other Class B frames on each of its other ports” refer to the same issue:
  - In the worst case for a Bridge with *n* Ports, even if all Talkers are perfectly regulated, the Bridge may get unlucky, and on each Port 1 through Port *n*-1, a frame destined for Port *n* can arrive, all at the same instant.
  - One of those *n*-1 frames has to wait for all of the others to be transmitted before it can be transmitted.
- So, the fan-in contribution for Class Z on Port *x* on a Bridge with *n* Ports (including *x*) is (*n* – 2) * (transmit time for a max-sized Class Z frame).

Max-size higher-class burst

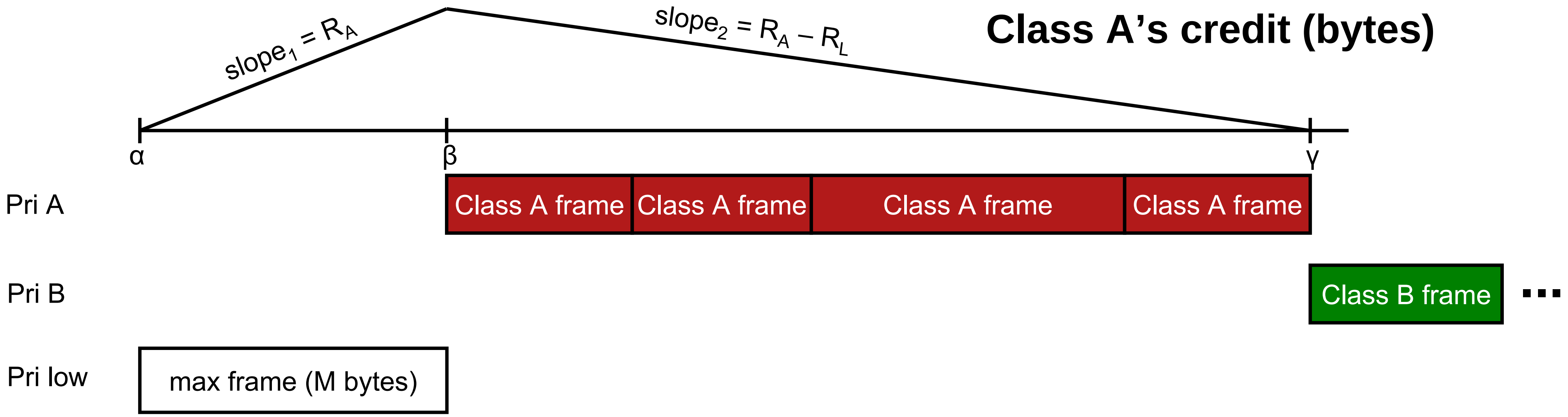
- Suppose that the queue for Class A is full, and has accumulated the maximum amount of credit.
  - Because Class A frames have priority over all other traffic (even BPDUs), the maximum credit for Class A is merely the credit accumulated during the “one max frame transmit time” required to transmit a lower-priority frame.
- Until the that credit is gone, Class B (C, D, ...) frames cannot be transmitted.
  - If Class A were permitted to use 100% of the LAN bandwidth, then the Class A queue would never catch up, because it would use credit as fast as it was gained.
  - If Class A were permitted to use 99% of the LAN bandwidth, then that max accumulated credit would be drained at 1% of the LAN bandwidth, until it is gone.

Max-size higher-class burst



- At point α, the Class A and Class B queues are empty (else, they would be sending, not the low-priority queue), so low-priority frame starts sending.
- Between α and β, fan-in frames arrive for Class A.
- Between α and γ, fan-in frames arrive for Class B.
- Class A starts sending at time β, Class B at time γ.

Max-size higher-class burst



$\text{slope}_1 = R_A$

$\text{slope}_2 = R_A - R_L$

Class A's credit (bytes)

α β γ

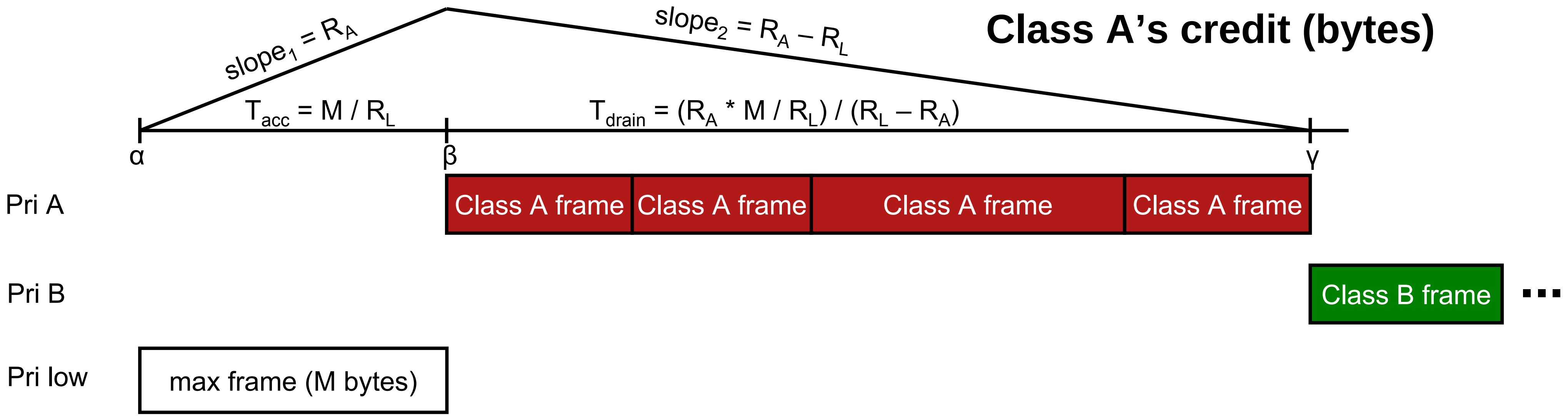
Pri A: Class A frame | Class A frame | Class A frame | Class A frame

Pri B: Class B frame ▪▪▪

Pri low: max frame (M bytes)

- Let R_L be the LAN data rate (bytes per second), R_A be Class A's maximum data rate, R_B for Class B, etc.
- Class A accumulates credit at the rate $\text{slope}_1 = R_A$ during the max frame transmission.
- This credit is drained at the rate $\text{slope}_2 = (R_A - R_L)$ [which is a negative value] during the burst.

Max-size higher-class burst



Class A's credit (bytes)

$\text{slope}_1 = R_A$

$\text{slope}_2 = R_A - R_L$

$T_{acc} = M / R_L$

$T_{drain} = (R_A * M / R_L) / (R_L - R_A)$

α

β

γ

Pri A

Class A frame

Class A frame

Class A frame

Class A frame

Pri B

Class B frame

…

Pri low

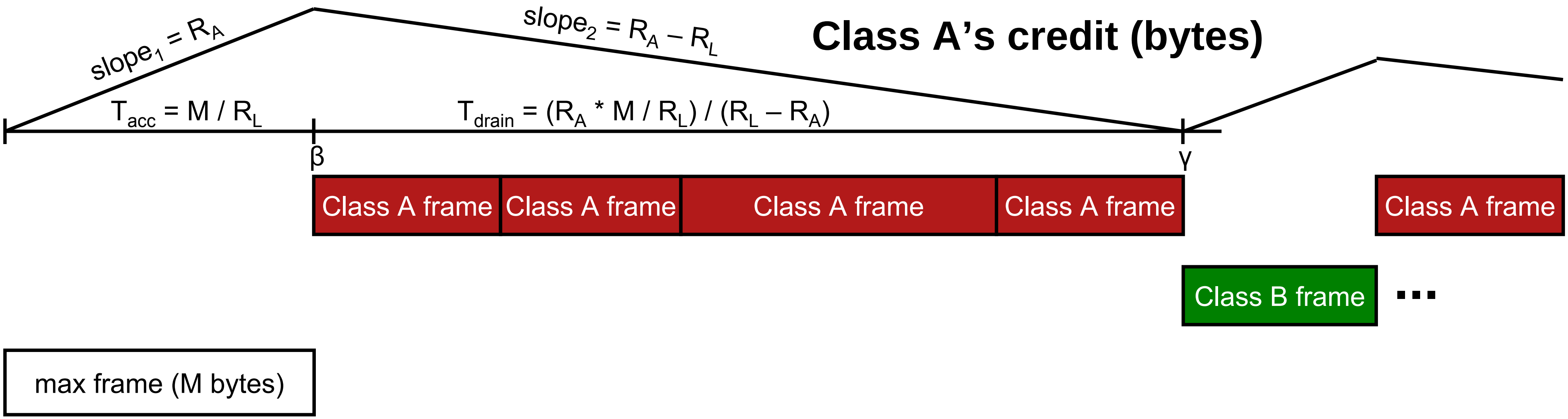
max frame (M bytes)

- Worst-case queuing delay for Class B is $T_{acc} + T_{drain}$.
- T_{acc} is the credit accumulation time, M bytes at rate R_L.

  Bytes of credit accumulated $= R_A * T_{acc} = R_A * M / R_L$.
- T_{drain} is the time needed to exhaust that credit.

  $T_{drain} = -(\text{credit}) / (\text{slope}_2) = (R_A * M / R_L) / (R_L - R_A)$.

How fast does Class B drain?



Class A's credit (bytes)

$slope_1 = R_A$

$slope_2 = R_A – R_L$

$T_{acc} = M / R_L$

$T_{drain} = (R_A * M / R_L) / (R_L – R_A)$

β

γ

Class A frame

Class A frame

Class A frame

Class A frame

Class A frame

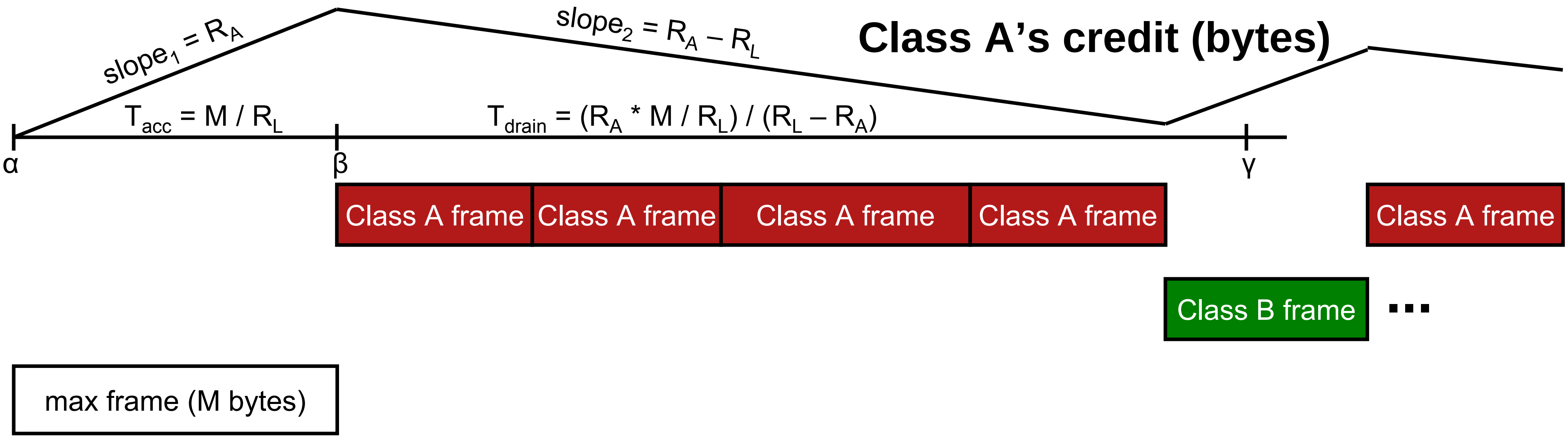
Class B frame

…

max frame (M bytes)

- Class A frames will continue to be transmitted at rate R_A, even while Class B is "bursting".
- In effect, Class B "bursts" at the rate $R_L – R_A$.
- Class B's transmission rate R_B corresponds to the credit drain rate $slope_3 = R_L – R_A – R_B$. (B's credit curve not shown, above.)

How fast does Class B drain?



Class A’s credit (bytes)

$\text{slope}_1 = R_A$

$\text{slope}_2 = R_A - R_L$

$T_{acc} = M / R_L$

$T_{drain} = (R_A * M / R_L) / (R_L - R_A)$

α

β

γ

Class A frame

Class A frame

Class A frame

Class A frame

Class A frame

Class B frame

…

max frame (M bytes)

- Note that this is a worst-case scenario. If the Class A frames available do not *exactly* fill up the T_{drain} time, then the Class B frame will jump in earlier than point γ, and so suffer less delay than the worst case.

Total Class A delay

- So, what is $T_{acc} + T_{drain}$ = Class A's delay?

$T_{acc} + T_{drain} = M / R_L + (R_A * M / R_L) / (R_L – R_A)$

$= M / R_L + M * R_A / (R_L * (R_L – R_A))$

$= M*(R_L – R_A)/(R*(R_L – R_A)) + M*R_A/(R*(R_L – R_A))$

$= (M * (R_L – R_A) + M * R_A) / (R_L * (R_L – R_A))$

$= (M * R_L – M * R_A + M * R_A) / (R_L * (R_L – R_A))$

$= (M * R_L) / (R_L * (R_L – R_A))$

Class A delay $= M / (R_L – R_A)$

- (: An encouragingly simple result! :)

And Class B's total delay?

- During the delay, Class B accumulates this much credit:

  $(\text{time}) * R_B = M * R_B / (R_L - R_A)$

- So, Class B takes this long to drain:

  $(M * R_B / (R_L - R_A)) / (R_L - R_A - R_B)$

- Adding Class B's accumulation and drain times, we get:

  $$M / (R_L - R_A) + (M * R_B / (R_L - R_A)) / (R_L - R_A - R_B)$$
  $$= (M{*}R_L - M{*}R_A - M{*}R_B + M{*}R_B) / ((R_L - R_A) / (R_L - R_A - R_B))$$
  $$= M * (R_L - R_A) / (R_L - R_A) / (R_L - R_A - R_B)$$
  $$= M / (R_L - R_A - R_B) = \text{Class B's worst-case delay.}$$

- And, of course, Class B's accumulation + drain times equals Class C's accumulation time.

What about Class C delay?

- Class C's total credits = (accumulation time) * R_C.

  credits = $M * R_C / (R_L – R_A – R_B)$

- Class C's drain time is therefore:

  credits / drain = $M * R_C / (R_L – R_A – R_B) / (R_L – R_A – R_B – R_C)$

- And so on ...

Class A vs. Class B vs. Class C

- Class X’s accumulation time = Class (X-1)’s total delay.
- Accumulation time before draining:
  - Class A: $M * R_L$
  - Class B: $M / (R_L – R_A)$
  - Class C: $M / (R_L – R_A – R_B)$
  - ...
- Total delay:
  - Class A: $M / (R_L – R_A)$
  - Class B: $M / (R_L – R_A – R_B)$
  - Class C: $M / (R_L – R_A – R_B – R_C)$
  - ...

Max credits and buffer sizes

- Over and above fan-in issues, a Class can be collecting frames and accumulating credits during its worst-case delay scenario, while the better-delay Classes are bursting (after *their* worst-case scenarios).
- Obviously, if the buffer overflows during this time, frames will be lost. So, the minimum buffer size required for each Class is a function both of fan-in and worst-case delay. Or, if you prefer, the minimum buffer size is a function of the worst-case delay.
- Also, max_credits must be large enough to absorb that worst-case buffer size, or successive worst-case events will cause frame loss.

Parameterization

Per-Class Parameters

- The obvious choice for Bandwidth parameters to use are the W_A values:
  1. Reserved for worse-than-A traffic $W_A = (R_L – R_A)$
  2. Reserved for worse-than-B traffic $W_B = (W_A – R_B)$
  3. Reserved for worse-than-C traffic $W_C = (B_B – R_C)$
  4. ...
- Then, computing the total worst-case output queue delay is trivial:
  – Delay for Class X = M / W_X

Per-Class Parameters

- Using W_X, it is natural to allow Class X to use Class (X-1)'s bandwidth, if there is no Class X-1 traffic.
  - Natural, because Class X doesn't care which of the better-delay Classes can source the frames that contribute to its worst-case delay.
  - If we use R_X, instead of W_X, then any bandwidth not reserved by Class X streams can only be used to carry non-SRP traffic.
  - If we do use W_X, then any bandwidth not reserved by Class X or better can be reserved for and used by all worse-Class traffic, whether SRP or non-SRP.
- If we use W_X, however, we also need a minimum allocable bandwidth G_X for each Class X when Class X+1 is configured; otherwise, worse-delay classes can hog the network. ($G_X = 0$ is the same as no G_X.)

Per-Port Parameters

- In some environments, e.g. in an enterprise with 500-port Bridges, but only one SRP Class, the fan-in component can contribute more to buffers size and delay than the burst component.
- Also, some Ports can have different buffer capacities, relative to their speeds.
- It would therefore be useful to define a maximum fan-in F_P for each Port P, that can be less than the physical fan-in. The fan-in limitation could cause a reservation to be rejected (or rescinded) because the number of Talker Ports sending traffic to some Listener Port P would exceed the allowable F_P for that Port.

Per-Port, Per-Class Parameters?

- Home Bridges are typically simple, with some number of identical Ports, and perhaps a few “uplink” Ports.
- There also exist complex Bridges for the enterprise environment that have wide ranges of optional line card and Port capabilities and/or media types.
- W_X, G_X, and F_P may not be able to allocate complex Bridges’ resources to support a reasonable range of applications’ demands.
- It may therefore be worth our while to:
  - Allow W_X and G_X to depend on the Port, as well as the Class ($W_{X,P}$ and $G_{X,P}$); and/or
  - Allow F_P to depend on the Class, as well as the Port ($F_{X,P}$).

Relationships among parameters

- Each W_X must be smaller than the next-better-delay W_{X-1}.
- Each W_X must be greater than or equal to the sum of all of the same-or-better Classes' G_X values.
- Each difference (W_X – W_{X-1}) must be greater than or equal to G_X.
- "All $G_X = 0$" means that all of the reservable bandwidth (the worst-delay W_X) is available for reservation by all Classes on a highest-Rank-wins basis.
- "Each W_X = sum of all same-or-better Classes' G_X values" means that the part of G_X not actually reserved by Class X is unavailable to any other Class; it can only be used by non-SRP traffic. That is, $W_X = R_X$.
- The simple relationships among W_X, max delay, and max credits makes it relatively easy to figure out which ones limit a given Bridge, Port, or Class, and to compute the other parameter(s).

One last observation to be discussed

- A reasonable way to optimize multiple parameters (namely, F_X, G_X, W_X, bandwidth, and Rank) is to:
  1. Allocate resources to requests in order of one of those parameters, namely Rank, until the first request fails;
  2. Reject (or rescind) all remaining requests.
- If less-demanding requests are granted after a failure, then unless some complex utility function $U(F_X, G_X, W_X, \text{bw}, \text{Rank})$ is used to select the “best” requests, the set of requests granted can be sensitive to the order in which they are processed, e.g., after a topology change.
- Any such utility function would likely cause more controversy and confusion than it would provide gains in network utilization.

Summary

Summary

- Parameters are:
  - W_X, the total bandwidth allocated for use by all Classes that have worse latency than Class X, including non-SRP traffic;
  - G_X, the minimum bandwidth guaranteed to be available for reservation for Class X, alone; and
  - F_P, the fan-in allowed for any Port.
  - (Or, probably, $W_{X,P}$ and $G_{X,P}$, and perhaps $F_{X,P}$, as well.)
- The worst-case output queuing delay = $M * W_X$, where M is the size of the largest frame transmissible on the port (from start of frame to start of next frame).
- Both the buffer size required and the maximum credits allocated for a Port's shaper must be sufficient to handle the worst-case delays for all Classes that can be transmitted on the Port.
- A "utility function" is probably not desirable.

Next steps

Next steps

1. Discuss, correct, validate, rewrite, or discard this presentation.
2. Re-examine the assumptions list in light of the results.