Congestion Notification

Mick Seaman

This note summarizes and follows up on discussions at the 802.1 Congestion Management interim meeting in January. It concludes with draft text for a first PAR (and five criteria) in the CM area.

1. Summary

The principal goal of the CM activity is support for a broader range of applications†1 by bridged networks simultaneously supporting existing Ethernet oriented applications†2. Any specification of the behavior of the bridges that compose such integrated networks has to address its (real or potential) impact on existing application performance, bridge design requirements†3, and bridged network configuration requirements†4. The developers of the existing bridging specifications reasonably require a definite proposal addressing the devilishly tricky issues of interoperability and coexistence (of both systems and applications). It is suggested that frames (and hence data flows) are subject to CM-control within a closed domain of CM-capable bridges, and are segregated from other application flows within that domain by their use of two†5 dedicated VLAN priority values. This note discusses this mechanism and its consequences in more detail. Its choice (or an alternate) is fundamental to being able to start writing the detailed specification. Since a PAR†6 grants permission to begin writing, rather than permission simply to investigate, it is appropriate that the PAR and supporting documentation incorporate the coexistence and interoperability provisions (rather than just a declaration of good intent†7) as a bound upon the project†8.

This note discusses the further project bounds proposed at the interim. In part these seek to ensure that the project preserves existing important characteristics of the bridging solution, does not impose on existing bridge implementations†9, and does not simply serve as a vehicle for reworking difficult areas that have been settled after much discussion†10. At the same time freedom to change some aspects of the deliberately simple existing bridge architecture is desired, and it is wiser to document that now rather than dispute its acceptability during the course of the project. Again an appropriate place to document such constraints is within the PAR Scope and Purpose, or within the accompanying five criteria and supporting documentation.

New projects are generally attractive and are often presented in terms that are sufficiently abstract to allow those with a wide range of technological aspirations to build project momentum, each hoping that their particular solution will be incorporated or their particular problem addressed. However the time taken to complete a standards project is a strongly non-linear function of its size†11, and there is a considerable risk that a vaguely scoped project will be subject to horse-trading or hostage taking†12. Breaking down proposed ‘projects’ that identify broad market or technological areas into the smallest possible stand alone components is good project discipline†13. Accordingly the interim proposed a project focused on an explicit congestion notification mechanism†14, from a bridge experiencing congestion†15 to an end station†16 with a participating traffic flow. Further projects may follow as the CM work progresses, but it seems important to begin writing with a project that has already been the subject of much technical presentation.

The project constraints described above clarify what the proposed congestion notification project may and may not change. It is also important to be specific about its area of applicability, where it is expected to deliver benefit, and to limit the extent to which the project will address the further consequences or opportunities presented by the standardized protocol†17 †18.

Congestion notification is proposed as an amendment to 802.1Q. To ensure that it is actually possible to execute such

†1 Chiefly those that have been designed in such a way as to make their “goodput” (useful network throughput) very sensitive to frame loss.

†2 Mainly applications using TCP both for data transfer and short control exchanges, and UDP (chiefly for the latter).

†3 Including forcing obsolescence of systems capable of meeting both existing and integrated network requirements.

†4 Including introducing new ways in which misconfiguration could cause network failure.

†5 Or at least not more than two.

†6 Project Authorization Request. One PAR results in (is consumed by) the completion of one standards document.

†7 The carrying out an adequacy of which could be the subject of much dispute in the course of the project.

†8 If the project discovers that the bound is not appropriate, then the PAR could be modified, subject to the approval of the entire working group. The administrative overhead of modifying a PAR is tiny compared to the necessary technical due diligence.

†9 The creation of new options that, despite their utility only in certain circumstances for certain applications, become mandatory for a vendor to implement whatever the applications is an ever present marketing danger.

†10 In other words “first do no harm”.

†11 The time taken to complete projects is a function of the mean time between major transient mistakes and/or the introduction of significant good ideas. When this mean time is less than the expected uninterrupted project duration the actual completion time tend to infinity. Larger projects naturally attract more ideas, good and bad, and ill specified projects even more. Once the adjusted completion time nears that for a significant turnover in the voting base it is unlikely that a standard of any technical integrity, rather than an assemblage of loose contributions, can be produced.

†12 Any project can be trivially halted or diluted by a 25% minority, unless what that minority wants is clearly outside the scope, purpose, or other formally adopted project constraints. Clearly scoped projects encourage significant minorities to build consensus around other projects that do meet their needs.

†13 Essentially this forces the major risk of the project, that everyone has very different ideas of what is acceptable, to be confronted up front and flushes out undisclosed agendas. It also makes the project less attractive to those whose major purpose is to write glowing trip reports.

†14 The mechanism could involve signalling to either the source or the destination of the flow subject to congestion, however the only mechanism advocated in any detail so far is the explicit generation of backward congestion notification frames proposed by Davide Bergamasco.

†15 As indicated by the length of a (per priority per port) output queue.

†16 The term “an end station” is used here in its precise 802.1 sense, i.e. as a MAC Addressed end point (MSAP) of communication using the MAC Service. From the point of view of such communication a bridge port is not an end station, though it could act as one for the purpose of generating related congestion notifications.

†17 These are often the source of unwanted scope creep, with ballot comments like “this standard needs to explain how it is used with X, works with Y, etc.”.

†18 Any technique designed to work within a closed domain invites the design of intriguing but complex gateways designed to propagate its characteristics beyond the explicitly chosen limits, and thus subtly or blatantly circumventing the agreed scope. These are proscribed.

a project in parallel with the already approved amendments in progress some plan as to how the amendment will be constructed is necessary, identifying particularly the scope and nature of changes to existing clauses†1. The soundest foundation for this is a clear technical architecture, expressed in the terms and style already used by the base document. Fortunately it appears that it is possible to develop this to a considerable extent without prejudging much of the detail of the congestion notification protocol†2.

2. Overview

As a prelude to suggesting text for a congestion notification (CN) PAR and 5 criteria, the remainder of this note discusses:

- The applications that are the target of the CM effort.
- The applicability of CM and the mechanisms that it uses.
- The proposed coexistence and interoperability provisions, their consequences and alternatives.
- Aspects of the existing bridge architecture that are to remain unchanged:
  - Queue structures
  - Transmission selection algorithms
- Aspects of the existing bridge architecture that may change, including:
  - Generation of data frames at a rate proportional to the line rate.
- Obvious consequences and opportunities for further specification presented by CM, and their inclusion (or otherwise) within the scope of the proposed PAR.
- A first cut at a CN architecture, and the insights that it provides including the relationship of CN to:
  - VLANs
  - Shortest Path Bridging
  - MAC Security

The suggested PAR and 5 criteria collects up the explicit constraints and freedoms noted in the discussion of the above.

3. Applications

It is worth being very explicit about the applications that are targeted by the CM effort, particularly because there is a considerable conceptual gap between those who view techniques, such as flow control, that can reduce frame loss at the cost of lowered throughput, as being of universal benefit to all applications and those who do not†3. From the point of view of the latter, flow control (as a substitute for buffering) reduces network goodput in large Ethernet networks and constitutes a threat, not an opportunity. Experience in IP networks has resulted in the recommendation that source quench techniques not be used, and the results from congestion notification have not met original expectations†4. It is incredibly difficult to conduct worthwhile and convincing simulations of IP traffic for networks of any significant size, as the response of the applied load to network performance is an important factor. Realistic application behavior simulation therefore becomes a precondition for believable simulation, with the result that significant networks are actually required in order to conduct the proper studies. At the same time TCP implementations have grown ever more sophisticated and do sense and adjust to network behavior, e.g. increases in the time take for a packet to be acknowledged provide a per flow indication of congestion†5 †6.

The wealth of experience, and in some cases disappointment, in studying the behavior of control techniques in large TCP nets does not make it at all plausible that a new congestion scheme for LANs will be of universal benefit, and reasonably likely that pursuing such a scheme on universal benefit grounds will simply lead to permanent deadlock. At the same time abstracting the notion of flow control away from the behavior of particular upper layer protocols that are highly loss intolerant runs the risk of introducing a very non-optimal solution. Or to make the point another way: today's Ethernet seems to be just fine for TCP and UDP applications†7 so what problem(s) are we really trying to solve. There appear to be two:

1) Use of Ethernet to support Fibre Channel applications, with the twin goals of making those applications available to a broader range of users and lowering operating costs for all through the use of a single integrated network for both those applications and TCP oriented applications.
2) Use of Ethernet as a backplane connect within a system.

These are both laudable and different from each other†8, and differ from attempting to make Ethernet†9 generally better. Like all objectives they invite freeloaders —other goals, explicit or otherwise, that are based on 'obvious' solutions to the real goals. Whether they do get carried for free depends on the best design to meet the real goals, and how much they are allowed to compromise that.

CN PAR Purpose: The proposed CN PAR is strongly oriented towards the Fibre Channel application goal and that should be explicitly stated less the project lose sight of that in the face of targets of opportunity.

†1 Which may already be subject to change by the other amendments. The idea that an amendment might be written 'somewhat stand alone' so that the reader would be responsible for appreciating its impact on other clauses, in effect becoming the integrator of last resort is not to be contemplated.

†2 Norm Finn suggested an architecture that draws parallels between CN placement and that for CFM. A "shim" based architecture such as this is very attractive, and facilitates early exploration of some of its more tricky consequences, such as its relationship to security.

†3 It is worth pointing out that the frame loss in bridges in enterprise networks is usually tiny. The most lossy IP network environments are the largest, where it is most difficult to use congestion management within the network.

†4 Worthwhile yes, up to expectations I believe no. There are of course many contributing factors, of which preventing system under the control of many different administrators has been a key difficulty.

†5 Not knowing what version of TCP is being used is an incredibly common simulation deficiency. I am chiefly talking about TCP Vegas here.

†6 It is not true that end to end protocols can only notice congestion when a packet is dropped, however this seems to be almost universally believed by flow control proponents. It is true that TCP Vegas has not supplanted TCP Reno in the way it would in an ideal world, because of the 'gaming' problem, that's less of an issue in 'single administrator' networks.

†7 The motherhood and apple pie of 'you don't want to drop packets, do you' doesn't cut it. Attempting to not drop packets at all will lead us back to the deadlock avoidance studies of the 1970's.

†8 For example the backplane connect is usually loop-free by design, so does not need packet drop as a deadlock avoidance measure. Moreover the timing and delays across a backplane are usually small and much better known than in any network. I believe both these goals are entirely capable of standing alone and solutions do not need to be compromised in order to attract political support from both to get each of them done.

†9 Whatever is meant by the term.

CN 5C Broad market potential: Needs to explicit identify data storage networks and non-TCP/IP applications. †1.

CN 5C Broad market potential & Economic feasibility: Needs to mention equipment and operational costs benefits attributed to use of a single integrated network.

The rest of this note concentrates on the Fibre Channel application goal, with some explicit exceptions.

4. Applicability

Congestion management within a network naturally involves some feedback to the sources of traffic, possibly coupled through intermediate systems†2. As such it works best when:

1) There is little delay in the feedback loop, i.e. the network is small in terms of the number of packet transmission time delays within the loop.
2) There is little noise in the feedback loop.
3) The applied load lasts long enough and is consistent enough (i.e. not noisy itself) for the feedback to work.
4) The network topology is simple enough for any feedback operating in the network to have minimal effect, at least probabilistically, on unrelated flows (network loads)†3.

These characteristics match those of back-end networks and their bulk data transfer intensive applications, so not surprisingly the attached systems and application protocols have evolved to take advantage of them to shift the burden of controlling the offered load to the network itself. In general IP networks have not had this luxury and IP end stations and protocols have had to evolve to work well without it.

Since the prospects for simultaneously replacing all of a stack of protocols and equipment are generally poor, and the characteristics of Fibre Channel applications may indeed allow for superior performance when the network has a greater role to play in determining the offered load, a proposal that they should all be moved to run over TCP leaves a considerable need unsatisfied†4.

CN PAR Scope: Needs to include (a) congestion notification (b) for non-TCP/non-IP flows (c) for long-lived flows (d) for limited diameter networks.

5. Coexistence and interoperability

The proposed approach for coexistence and interoperability, as outlined in the Summary above, is that CM mechanisms (in general, not just CN) operate only within closed domains of CM-capable bridges, and are segregated from other application flows within that domain by their use of (at most) two dedicated VLAN priority values.

It is worth looking closely this proposal's attributes, and comparing it with alternatives, with particular attention to:

1) Why the domain is closed, excluding non-CM bridges, and not just limited in size.
2) How the domain is closed, and how traffic enters and leaves the domain.
3) The benefit of allocating dedicated VLAN priorities to CM-controlled flows.

It is also worth mentioning:

4) Why we believe we can spare two VLAN priority values, and why we might need two†5.

In considering interoperability with existing 802.1Q bridges the questions arises as to how any new frame formats might be treated by the VLAN ingress rules. If a new frame format is defined for frames that are not addressed to one of the bridge's Reserved Addresses†6 then the assumption has to be that that frame will be VLAN-tagged. This effectively rules out adding a new sort of tag (such as might be required to carry CM information) in front of the VLAN tag if the frame (with the new tag) can leave and subsequently reenter a CM capable domain†7.

A requirement not to introduce any CM-tag before the VLAN tag, or to introduce any non-taggable frames that are intended to pass through CM-capable bridges appears unduly restrictive†8. By only adding CM-tags within a closed domain and removing them prior to transmission to any non-CM capable bridge†9 the possibility of stacking tags for ever is avoided.

†1 Note that CN should make Ethernet more technically attractive for use in those environments and improve application performance, but it cannot be claimed that CN will make Ethernet loss free. Anyone who thinks of the CM effort as having principally a marketing benefit, countering any (supposed) Fibre Channel proponent with a claim that "the problem with Ethernet is that it loses packets", is going to have to do more than wave the CN PAR around.

†2 As in back-pressure flow control.

†3 This last point is not a concern for CN, operating as it does directly between the point of congestion and the point where the load is applied. In hop-by-hop flow control it can result in congesting spreading and even network deadlock.

†4 Otherwise we wouldn't be studying the current work, because (as I understand it) the specifications required to use TCP have been complete.

†5 Omitted from this first draft, but not difficult.

†6 Always filtered.

†7 Thus creating a possible loop in which each packet loops with the repeated addition of tags until it exceeds the maximum acceptable frame size. We managed to get around this with MAC Security, since security is necessarily not permissive and explicitly prevents misconfigured communication. While it is in principle possible to get a MACsec tagged frame to get tagged by a non-MACsec capable bridge (using shared media) and have that tagged frame reappear at the tagging bridge, that bridge would only accept the frame once more if the security attributes of the ports on that single bridge are inconsistently configured in a way that would subvert the security. The same defence is not available for other protocols because it is not possible to mandate a procedure analogous to the key agreement protocol, and even if it were it is not possible to prevent communication if the mandated procedure is ignored. The need for shared keys in security provides some defence here.

†8 Note however that the proposed protocol stack architecture for CN places the CN tag after the VLAN tag. This doesn't negate the other advantages of the closed domain approach however, and the question remains as to whether it is wise to assume that CM will be restricted to tagging after the VLAN tag in the future. Of course in the unlikely event of CM being deployed in a provider network there is no "out" as a tag after the S-VLAN tag will appear before the C-VLAN tag.

†9 The usual care has to be taken with shared media or virtual shared media. Any LAN with a non-CM capable bridge attached lies outside the domain.

Of course insisting on a closed domain makes it easier to ensure that the size of the domain is indeed limited. In order to simplify end station behavior, or at least reduce its potential cost, it is suggested†1 that rate controls can be applied to a queue or queues that contain traffic for multiple destinations. This is based on the observation that it is unlikely that there is more than one congestion point at a time†2.

CN PAR Scope: Don't forget the end station part. This needs to be part of the same standards project (and therefore finally part of 802.1Q) to ensure that the behavior of the entire system - bridges and end stations is properly specified and evaluated.

Closing the domain makes it much more likely that the congestion from two different flows, either at the same congestion point or at different points, is subject to roughly the same control loop delays†3.

The domain is closed by ensuring that a CM-capable bridge port has, as its immediate LAN attached neighbour, another CM-capable bridge or CM capable end station. If that is not the case the bridge port discards CM specific frames that it receives or would otherwise transmit, and removes CM specific tags from frames received or transmitted. Closing the domain in this way simplifies both present and future interworking. It means that systems outside the domain do not have to understand the format of CM frames and tags in order to support protocols subject to CM within the domain. It also means that it should be possible to allow different or differently parameterized CM schemes (if necessary) in the future, each within its own domain. In an area that is as difficult to analyze as CM, that seems prudent insurance against design risk.

CN PAR dependency: Reliably closing the domain in this way requires knowing the identity of the neighbouring stations, detecting the existence of non-standard bridge like intermediaries such as the "buffered repeaters". This requires some help from the media access control method, and is known to be useful to other projects.

In addition to ensuring that CM information neither enter or leave the domain, frames entering the domain with a priority used locally for CM controlled traffic would have that priority remapped. This ensures that the CM managed queues are not subject to external noise. Frames that leave the domain with a CM controlled priority can however retain that exact priority.

CM control thus operates between bridges and end stations that are within the same domain. The design point is that both end stations be within the domain, depending on the mechanism it may be that there is some benefit to controlled traffic within the domain if one of the end stations is in the domain†4.

The proposal to dedicate up to two VLAN priorities to CM controlled traffic within a CM domain can be contrasted with two obvious coexistence and interoperability alternatives, each at an extreme of a potential spectrum of choices.

The first of these alternatives is to simply allocate Ethertypes†5 and sub-types to identify Fibre Channel messages on a 1 for 1 basis†6. The specification of the resulting bridge within the CM domain involves multiplexing and demultiplexing frames with these types to a forwarding function that behaves as specified by the Fibre Channel specifications. How options should be selected from those specifications would be a Fibre Channel not an 802 problem, and it is not even clear that there need be any 802 involvement in the specification at all.

A problem with this first alternative is that there is no standard way of carrying the Fibre Channel traffic to and from the CM domain. A potentially complex non-standard gateway might be constructed. In that respect it falls far short of the 'just fix Ethernet so that it doesn't lose frames' marketing requirement.

The second alternative is to attempt to provide that last supposedly obvious fix, by placing flow control under all traffic types within the CM domain. However much experience in connectionless packet networks indicates that this would degrade the performance of protocols such as TCP/IP even in modestly complex networks such as mid-size enterprise cores, while the fact that most of the problems it can introduce were explored long ago and are not commonly discussed†7 would cause the flow control to be more widely deployed than advisable. All that adds up to increase equipment and operational cost with vendor and user confusion. It is very unlikely that a new lossless flow control scheme could be proved to be harmless, much less of significant benefit, without widespread deployment, while the cost of that experiment is considerable.

Separating CM-controlled traffic by VLAN priority provides a "ships in the night" approach. It allows existing IP oriented traffic that would receive a marginal, negative, or at least endlessly disputable benefit from CM to be conveyed with no change†8. It allows CM to be applied to just those applications that depend on explicit signalling

†1 I believe that this is part of the current BCN proposal, following Hugh's presentation at the interim.

†2 In my personal comments in the interim I probably over emphasized the case for a single congestion point. It is true that it is overwhelmingly likely that any given flow will have a single point of congestion (if it is congested at all) and the analogy to the well known rate determining step of chemical reactions is appropriate. However it is possible that two distinct flows (two distinct reactions if you like) have different congestion points, as long as any common part of their respective parts is uncongested. Assuming that the source of each flow is capable of matching the local transmission bandwidth and there is some competition for the destination or part of the path then there will be congestion at a point along each flow - or else the congestion avoidance algorithm used makes less than best use of available bandwidth. The likelihood of the congestion points for different flows being distinct is related to the complexity of the topology, simple star topologies being most likely to share congestion points and rich mesh topologies being least likely. In the latter case the use of end station shared queues for CN is likely to mimic the undesirable congestion spreading effects of hop by hop flow control. Use of fewer end station queues than flows is probably suboptimal while trying to take advantage of emerging multi-path approaches, such as shortest path bridging.

†3 Consider the possible effects of the alternative, with the congestion contribution to a single queue being made up of a flow whose end points are very close together and a flow with a very distant end point. The resulting phase differences of the effects of the control are very likely to affect its stability. I am not sure whether this has been simulated or not.

†4 For example the proposed BCN mechanism could control intra-domain congestion from a source within the domain even if the destination lies out side it.

†5 And potentially a few MAC Address (Reserved Addresses or otherwise) for various control and configuration functions.

†6 To a first approximation. Alternately the Fibre Channel applications over IP work could be used to derive a straight mapping to Ethernet frames. I realize I am oversimplifying here, but have no wish to spend too long on an option that seems to have been ruled out by those closer to the work.

†7 Despite the fact that we seem to have a failed flow control project every five years or so.

†8 Of course this approach does not meet the goals of anyone who has hopped onto the CM project with the aim of saving buffers in existing bridges for existing applications. It encourages us to move away from attempting the very minimal buffering approach, since a bridge implementation for an integrated network will necessarily include the requisite buffering for the bursty IP applications.

from the network†1. As compared to identifying that application traffic with a new set of Ethertypes over any VLANs present it allows the existing queue structures, and other cost oriented aspects of bridge designs to be retained. Further it allows traffic to be carried to and from (or between) CM domains quite simply. One way this might be done is through a provider Service Instance that operates no CM mechanisms itself but uses the mechanisms available to the operators of provider bridged networks to offer a policed bandwidth guarantee. By matching to that guarantee, and using the CM mechanisms to trim traffic to live within the guarantee, CM capable bridges could provide suitably low loss characteristics†2.

6. Frame queuing

It is proposed that neither the CN project, or CM work in general, modify the queuing of frames, as currently described in 802.1Q-2005 clause 8.6.6. It is worth spelling out what that means. Whatever the internal detail of an implementation that conforms to the standard, the external behavior is that of a set of FIFO queues, one per traffic class per port. Further the only information†3 relied upon for those queues are the number of frames and number of octets in the queue, and the (very rough) age of the frame currently at the head of the queue†4. Note particularly that an output queue model is assumed. Though this can be clearly implemented by a larger number of input queues with transmit time selection from the head of those queues, it is important that we do not mandate such an implementation†5.

Explicitly preserving the existing queue structures addresses two issues:

1) Clarification of the nature and extent of the additions and changes for the proposed CN amendment, making it possible to construct an initial draft as a framework for the technical detail.
2) The fear that CM will require and thus degenerate into per flow queuing.

†1It is profoundly to be hoped that the likelihood of these applications changing to become more network tolerant within the next three to five years has been seriously considered. It is rare that a significant standards effort with multiple facets is completed in less than three years, and there is some time lag after that before truly conformant products are deployed. Providing marketing cover for short term products is therefore not a good use of standards development time. It is easy to envisage that initial CM products would attempt to look like pure Fibre Channel adaptors to their end stations, with higher layers of software unchanged, and the question is whether that architecture will be permanent or the only possible way of initiating a very long term migration to Ethernet. By "more network tolerant" I mean of course better sensing that more frames are being queued in flight, thus indicating congestion on the path to the destination and not adding to that congestion: I am not suggesting that coping with actual frame loss is easy or indeed the right target, the point is to keep the network queues well away from the "drop tail", for simple networks with known (or determinable diameter) and modest numbers (few thousands) of system that is certainly possible.

†2Fairly obviously this is a more expensive way of providing a given network throughput over the long term than use of a more best effort oriented service, but a more cost effective goodput is expected if the upper layer protocols are extremely sensitive to packet loss.

†3Just to cut out any idea that agreement might be reached that the queue structures will be preserved, but enhanced by retaining information on how many frames of a certain type are in a given queue, together with new methods for extracting frames

†4This does not have to be at all accurate, what is mandated that frames over a certain age can be removed from the queue and not transmitted, and that frames over a greater age will be so removed. Retention of an accurate time for congestion management purposes would be a definite change top the queue.

†5In particular simulation of any CM mechanism needs to be carried out with the simpler (less information rich) basic bridge model.

CN 5C Economic feasibility: Retention of the existing cost characteristics of bridges, including simplicity of queue structures. The amendment will require no per flow queuing or state in the bridge.

7. Transmission selection algorithms

802.1Q mandates implementation of a very simple algorithm for selecting which frame to send next on a port, the traffic class queues are serviced in strict priority order. It also allows implementation of other, unspecified, algorithms as long as the simple default remains available to the user. Though there has been a fairly constant level of background criticism of this state of affairs, the strict priority approach works well in most bridged networks, particularly enterprise campus networks that are on average lightly loaded. However the reason for not specifying a more sophisticated algorithm, with better performance over a wider range of network loads, is not our failure to appreciate that such algorithms exist but the difficulty in picking one from those on offer†6.

All in all the inclusion of a further transmission selection algorithm is not one that any amendment should take on as an adjunct to its main business, and it is proposed that CM and CN in particular not make changes in this area. The point of mentioning transmission selection in the CM context is that long-lived data flows, i.e. those most amenable to CM, are often not the most time critical †7. This means that higher priority bursty traffic will take transmission bandwidth when it is forwarded by a bridge, and as far as CM is concerned will look like noise or uncontrolled traffic in the CM queue. This effect, and of course the value placed on the priority service, could be diminished, or at least tinkered with, by adopting a a more sophisticated transmission selection algorithm. However the possible gain is small unless higher priorities are to be severely impacted. Using strict prioritization as the standard case for simulation means that CM algorithms adopted should be robust against the use of any sophisticated proprietary transmission selection mechanism†8.

Similar comments could be made about the introduction of transmission rate controls into bridges. The demand for and nature of such controls is likely to be dictated by the needs of other types of networks†9 and coupling CM to those future discussions is unlikely to cause anything to proceed smoothly.

CN PAR 5C Compatibility & Economic feasibility: Will not constrain compatibility with or impose costs on existing or future bridges by introducing new transmission selection algorithms, including rate controls.

At the same time control of transmission rate at the original source is the fundamental method by which the CN mechanism (and in the end any CM mechanism) operates.

†6Some are advertised as being proprietary, the advantages of others seems to be more in thesis than in than in practice, and of course there is always a body of opinion centered around quite different cost points that per-flow queuing is the way to go. Personally I favour DWRR (single-level deficit weighted round robin), but others can very reasonably disagree. With the wide range of networks that are supported by bridges it is hard to show that any particular algorithm has a definite cost benefit as compared to any other, while the likelihood of any vendor discontinuing use of an algorithm more sophisticated than that chosen for the standard is negligible.

†7Having the bulk of data travel at the highest priority doesn't make good use of a priority scheme.

†8Provided realistic amounts of higher priority bursty traffic are incorporated in the simulation. CM algorithms should at least do no harm when most of the traffic is bursty and at a higher priority for some period.

†9Such as interfaces to service provider networks.

CN PAR 5C Economic feasibility and technical feasibility: Explicitly include end station transmission rate limiters.

8. Frame generation by bridges

The current bridge architecture deliberately does not introduce new frames into the data transmitted by the bridge at any rate proportional to the forwarded data rate†1. To permit a wide range of cost effective implementations the standard does not mandate tight timing relationships between relayed frames and those introduced by a bridge for any bridge protocols so far specified†2.

Admitting the proposed BCN mechanism as a candidate for CN would break this rule, and if (as I believe) that is intended, should be explicitly permitted by the PAR documentation†3. Whether this change has a cost impact on any particular implementation depends on that implementation, but very low cost bridging silicon implementations and architectures that rely on a supervisory or management processor to inject any such frames certainly exist, and would not be capable of performing such a function†4.

CN 5C Economic feasibility: Introduction of the CN function may have a cost impact on bridges†5 but the point is more to the overall cost savings of an integrated network, both as overall equipment cost and as operational cost.

CN 5C compatibility: CN may require some functions, specifically the generation of explicit congestion notification frames, that can not be supported by some bridge implementation architectures that are otherwise capable of full standards conformance.

9. Specification opportunities

There are obvious opportunities to extend the scope of the project, to the detriment of timely completion. These include:

1) Specification of CM “gateways” or interworking units, whose purpose would be to transfer the CM information to some other congestion management or avoidance protocol —the IP congestion experienced bit is a clear example— or to another congestion management domain. Such gateways make a clear mockery of the premise used to advance the CM work, viz that it will perform satisfactorily because the control loop is of limited diameter. In this case the complexity of dealing with that untruth has been dumped on the gateway with the risk of placing cost on all bridge implementations that might be called upon to function as such standard gateways. The effort involved in specifying gateway functions in general is considerable, and can be made to drag on as one plausible, enticing, and non-trivial proposal follow another†6. We should neither aim to undertake such work, nor be prey to those who might vote to force it upon us.
2) Specification of front-end or off-load units whose purpose is to stand directly between CM capable Ethernet and Fibre Channel attached end stations, or between CM capable Ethernet and not-quite-CM-capable Ethernet attached systems. It is unclear whether a specification of such devices would be required in the timescale of standards development, and if proved to be not a burden on the standards project would equally prove trivial as a nonstandard but standards compatible offering. Within any standard it would constitute a hard to manage exception to the rule that CM would only operate to and between source and destination end stations in the closed CM domain.

CN PAR Scope: CN only operates between bridges and end stations within the same CM domain, and is not communicated to or from systems outside the domain. The project does not include specification of devices design to communicate or receive congestion information to or from stations outside the domain.

3) Specification of encapsulation of protocols so that they can be carried across a CM domain, with CM operating at the end stations enclosing that domain. The potential for being hit by this as a requirement is more obvious from the point of view of experience with provider bridged networks than it is from a CM viewpoint, and more likely to occur with use of CM to support an Ethernet backplane than it is in any Fibre Channel context. Consider an Ethernet backplane telco oriented chassis offering IP routing and VPN services. Each blade in the chassis might naturally function as an end station on the backplane Ethernet, and since IP on a LAN expects to operate above a MAC addressed end station it is fairly easy to construct an ‘open’ system that blade vendors can agree upon. Now consider the same scenario with the Ethernet backplane supporting a Provider Bridge or Provider Backbone Bride. A further level of as yet unstandardized encapsulation is required, and it is not at all obvious how that should be done and what new Ethertypes should be allocated. Since 802.1 specifies

†1The two arguable exceptions to this case are frames that are generated according to the specification of the media access control method used by an individual bridge port, and the necessary replication of multicast frames. The former do not traverse the bridge at all, and are conveniently handled locally. The latter have a strong effect on good implementations and have historically proved a severe test for router implementations masquerading as fully capable bridges.

†2Nor does it specify tight timing for operation of the Learning Process, which could lead to a specification problem if a frame is to be transmitted back to the source of a received frame. The Filtering Database may not yet know through which port the source is to be reached, and the current specification does not mandate that source port information be associated with a queued frame. Fortunately this can be finessed in a not unreasonable way so long as a backward notification frame is generated before the queuing is done - which is the best place to generate any such backward frame in nay case. I note in passing that flow metering (8.6.5) in 802.1Q-2005 takes place after egress (8.6.4) which is curious since the context for flow metering is stated to be the receiving port. This was changed after 802.1Q-REV D2.0 which has these procedure in the opposite order.

†3This is not the same as saying that the proposed BCN mechanism is the guaranteed outcome of the project, only that it appears to be the desire of the group that it be a permitted outcome. It would be nice if someone would spend as much effort as the BCN proponents to consider the use of the Drop Eligible coding of the priority bits to signal forward congestion, or the use of the CFI bit (in the VLAN tag) to signal backward congestion. While the use of forward congestion signalling notionally takes longer for communication to reach the source its use has been extensively investigated, first as originally proposed by Raj Jain as the DECbit (in frame relay and elsewhere) and then as the “congestion experienced” bit in IP. As I recall the original DECbit implementation tended to equilibrium (more or less) after seven round trip times, and I still doubt whether BCN can approach this, as it necessarily sends notifications based on a very low sample rate (1 in 100). Forward signalling using DE encoding would fall naturally within the mainstream trajectory of bridge architecture development, although of course the destination end station adaptor would have to mark or generate reverse direction traffic if the upper layers cannot use a CE notification.

†4Of course such silicon usually requires modification for any forwarding path changes, but some changes —such as setting or clearing (a) bit(s) in a forwarded frame as is done to signal drop eligibility— are easily made because they have little overall architectural impact.

†5Chiefly by excluding the most cost sensitive designs, everyone has some big stuff that can do anything (including a side order of fried eggs) that won’t change cost.

†6For those who remember the SR-TB effort.

bridging there might be an expectation that such a specification accompany the CM work†1.

CN PAR Scope: Specification of protocol encapsulation across CM domains is outside the scope of the project.

There is one specification responsibility that we cannot, and should not, avoid —a suitable MIB.

CN 5C Compatibility: Consistent with recent 802.1 decisions, the development of a MIB is an integral part of this project.

10. CN Bridge Architecture

Norm Finn suggested that CN†2 be architected in a rather similar way to CFM, i.e. it would appear as a protocol shim in bridges, using and providing the EISS (.1Q clause 6.6). Whether it is formulating the specification as a shim is the best approach depends on the details of the mechanism, and might cause some unnecessary architectural rearrangement in 802.1Q†3, but the proposal to use the EISS rather to reinterpret encoded data lower in down in an interface stack is surely right †4.

If an explicit congestion notification mechanism is to be used, as in the current BCN proposal, then the notification should be generated when the relayed frame that causes the notification is queued. If a congestion experienced bit is to be set in the frame for forward notification then the bit should be set based on queue length (or a historic function of queue length) when the frame is dequeued, i.e. after transmission selection and immediately prior to transmission. If a backward congestion experienced bit is to be set in a frame going in the reverse direction then that bit should be set†5 when the frame is received on a port based on the output queue at that port.

One of the virtues of using a strict interface stack formulation for the BCN mechanism is that it makes it clear to which of the forwarding process functions the returned BCN frame would be subject. Figure 1 shows the interface stack, including MAC Security and Link Aggregation, with †6 the queuing moved from the Relay Entity to the stack and the BCN function shown as a shim.

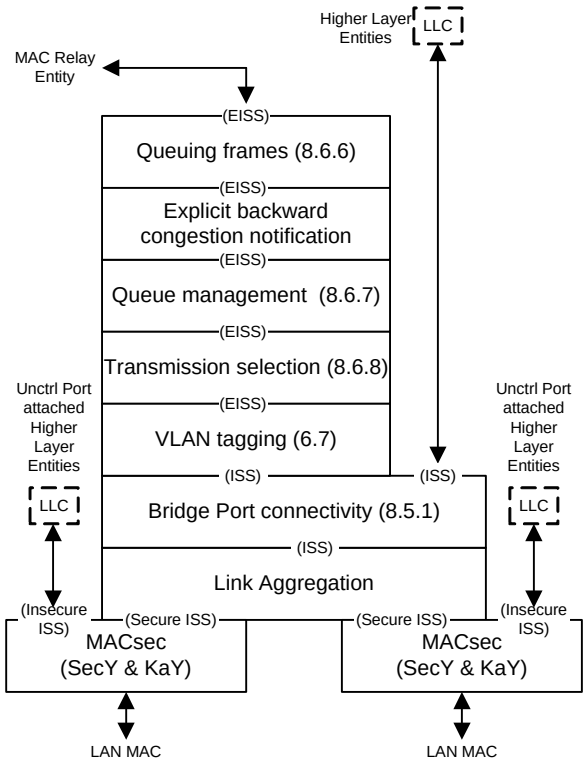

Figure 1—BCN interface stack

Figure 2 is an alternative description, identical from the point of view of externally observable behavior. It is based on 802.1Q-2006 Figure 8-9, extended to illustrate frame forwarding in both directions between two ports. Conceptually the backward notification passes through active topology enforcement, the ingress rules, and other forwarding processes before being queued on the appropriate priority queue for transmission back to the source of the frame that triggered the notification. If the original frame has indeed been queued for transmission, the active topology enforcement should pass the BCN, so the operation of that function is purely nominal. The passage of the BCN through the ingress function is a reminder that the BCN should have an appropriate VID assigned, though changes that were made in 802.1Q-2006 moved the responsibility of classifying untagged frames to the VLAN tagging function (Clause 6.7). While that was clearly the correct move, consistency demands that the BCN VID be the PVID for the transmission port if the VID of the original frame was in the untagged set. Further configuration information will be required if shared VLAN Learning is being used †7, while Shortest Path Bridging requires the choice of VID that would be used by the bridge to add traffic to the same VLAN that is selected by the original frame. It is important to note, and simulate, the queuing of the BCN in a standard queue prior to transmission —the use of additional 'ultra high priority' queues would impact the current (and future) effect of transmission selection algorithms†8, and contradict the explicit project constraint against new queue structures.

†1The interim assured me that this simply wasn't a problem, and that nobody would ever want to use an Ethernet backplane to create a standard open architecture Provider Bridge or Provider Backbone Bridge platform and expect 802.1 (as the standards group for CM and for bridges) to fill the specification gap. However there was no opposition to ensuring that such a remote possibility didn't become a problem.

†2And possibly the CM mechanisms in general, though I am not sure Norm intended to go that far.

†3Norm himself pointed out that the BCN mechanism would naturally place the shim between the beginning of the forwarding function and the queues maintained by the forwarding process. On reflection this seems gratuitous. Use of a shim for BCN also runs into the problem that the backward frame that the shim might generate is not guaranteed to go back through the reception port of the frame that provoked its generation, since the learning process functions asynchronously to forwarding.

†4In fact the EISS is really the interface to the transmit queues already, though all but the most careful reader may have missed the implication of the NOTE in clause 8.6.4 —"The Forwarding Process is modelled as receiving a frame as the parameters of a data indication and transmitting through supplying the parameters of a data request. Queueing a frame awaiting transmission amounts to placing the parameters of a data request on an outbound queue". The queuing process also acts as if the MAC status parameters were present.

†5Strictly speaking this would be done by modifying the priority parameters, taken to include the CFI/DE bit, the DE bit being the obvious one to reuse for this purpose since DE would not be wanted in this application environment. This strict formulation shows how a 'congestion experienced' bit would be used to signal (forward or backward) through the EISS.

†6The usual "test case shims".

†7The standard feature is used in to support features such as "private VLANs".

†8A 1% 'rogue' frame rate is not quite negligible so far as its impact on existing highest priority traffic is concerned, but the bigger issue is where does one stop special casing new traffic types. The traffic transmitted by a bridge in its participation in end station protocols is exceedingly small, and its timeliness guarantees (as opposed to hopes) very modest.

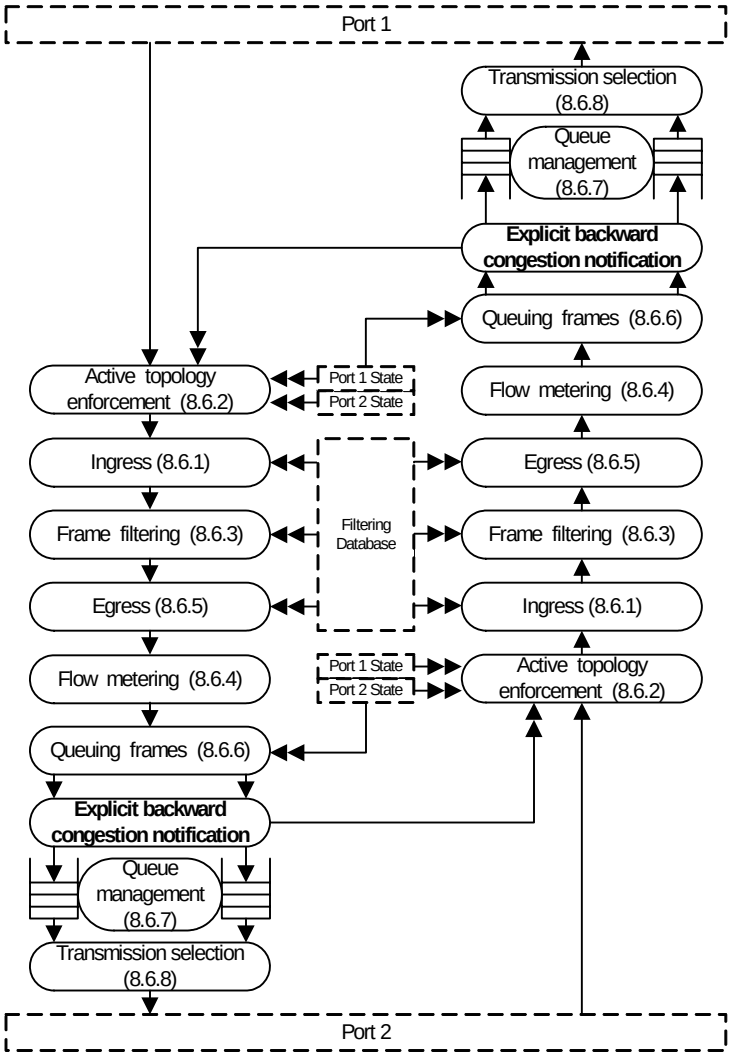

Figure 2—The Bridge Forwarding Process with BCN frame generation

It should also be clear from Figure 1 that congestion notification information in a frame (whether comprising the rest of the frame or being merely a tag header) would occur **after** the VLAN tag†1 if that is present. The CN information would be in the data of the EISS (or in the priority parameter).

11. A suggested CN PAR

Please note that the suggested text in this note for a CN PAR and following five criteria is my own†2, and is an attempt to facilitate the timely development of a standard along the lines of the major contribution to the CM study group, within constraints discussed in the interim meeting†3 and intended to preserve important characteristics of the bridging solution and its existing application.

The proposed project is for the development of a full standard, as an amendment to IEEE Std 802.1Q †4.

11.1 Title

Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks – Amendment 9: Congestion Notification.

11.2 Scope

This standard specifies both bridge and end station participation in an optional congestion notification protocol that supports long-lived data flows for highly loss sensitive higher layer protocols within network domains of limited bandwidth delay product by enabling bridges to signal to end stations capable of rate limiting transmission to avoid frame loss. Specific VLAN tag encoded priority values are allocated to segregate frames subject to congestion control, allowing simultaneous support of both congestion controlled and other higher layer protocols.

This standard does not specify communication or reception of congestion notification information to or from stations

†1Specifically after the VLAN tag used to encode the VID and priority parameters of the EISS shown if anyone is concerned about provider bridges (not really relevant in the contest of this project).

†2Others may have better ideas for PAR text, and I have other preferences for direction of the work, the latter is not the point of this note.

†3As reiterated in the first part of this note, and also documented in Hugh Barass' closing presentation to the interim meeting (in www.ieee802.org/1/files/public/docs2006/).

†4With 802.1Q-2005 and other amendments, completed or already approved, serving as the base text.

outside the congestion controlled domain or encapsulation of frames from those stations across the domain.

11.3 Purpose

This amendment will support the use of a single integrated bridged local area network for Fibre Channel applications as well as IP-oriented applications.

11.4 Reason

At present separate networks are used for IP-oriented connectivity and data storage oriented applications. Use of a single integrated network will realize operational and equipment cost benefits, while making data storage networks available to a wider range of users.

12. Five criteria for the suggested CN PAR

12.1 Broad market potential

A standards project authorized by IEEE 802 shall have a broad market potential. Specifically, it shall have the potential for:

a) Broad sets of applicability.

Mechanisms to avoid frame loss, of which congestion notification is one, are essential for support of the highly loss sensitive (non-IP and non-TCP) higher layer protocols, which are prevalent in the important data storage market.

Back-end data storage networks are typically limited in size, making them amenable to a congestion control mechanism that is most effective with a limited network bandwidth delay product. Each network is typically under the control of a single administrator, so a control technique that could otherwise be ‘gamed’ by separate organizations attempting to acquire an unfair share of the bandwidth is applicable.

The data traffic to be controlled by the proposed congestion notification mechanism will be segregated using a VLAN-based technique, thus ensuring that traffic types already support by VLAN Bridges are not affected and that there is no diminution of applicability to integrated networks.

b) Multiple vendors and numerous users.

Multiple equipment vendors have expressed interest in the proposed project. There is strong and continued user interest in converting existing networks to Ethernet, and in the realization of operational and equipment cost savings through use of a single integrated network. Further there it strong interest in increased use of data storage networks, provided that they can be realized with familiar technology and an integrated network.

c) Balanced costs.

The introduction of congestion notification is not expected to materially alter the balance of costs between end stations and bridges. While the introduction of the congestion notification option may constrain bridge implementation, significant equipment and operational costs savings are expected as compared to the use of separate networks for IP connectivity and for data storage.

12.2 Compatibility

IEEE 802 defines a family of standards. All standards shall be in conformance with the IEEE 802.1 Architecture, Management and Internetworking documents as follows: 802. Overview and Architecture, 802.1D, 802.1Q and parts of 802.1f. If any variances in conformance emerge, they shall be thoroughly disclosed and reviewed with 802.

Each standard in the IEEE 802 family of standards shall include a definition of managed objects which are compatible with systems management standards.

The proposed standard will be an amendment to 802.1Q, and will interoperate and coexist with all prior revisions and amendments of the 802.1Q standard. The data traffic to be controlled by the proposed congestion notification mechanism will be segregated using a VLAN-based technique, thus ensuring that traffic types already support by VLAN Bridges are not affected.

Congestion notification frames and frame headers are confined to a domain composed solely of congestion notification capable bridges and end stations, thus preventing interoperability or compatibility problems from arising with either existing end stations and bridges, or with future systems using possible different techniques.

The proposed amendment will not introduce new bridge transmission selection algorithms or rate controls. Proposed end station controls on transmission rate and queueing are intended for use with 802.3 end stations and will be compatible with transmission control mechanisms already developed or under development by 802.3 and subject to liaison with 802.3 using the already established procedures. Such end station controls will be independent of the details of the 802.3 media access control technology and will make use of the existing interface (jointly standardized with 802.3) used by bridges.

The proposed amendment will contain MIBs, or additions to existing MIBs, to provide management operations for any configuration required together with performance monitoring for both end stations and bridges.

12.3 Distinct identity

Each IEEE 802 standard shall have a distinct identity. To achieve this, each authorized project shall be:

a) Substantially different from other IEEE 802 standards

IEEE Std 802.1Q is the sole and authoritative specification for VLAN-aware Bridges and their participation in LAN protocols. No other IEEE 802 standard addresses congestion notification by bridges.

b) One unique solution per problem (not two solutions to a problem).

Congestion notification is a reactive (not prescriptive) mechanism, and has not been anticipated by any other IEEE 802 specification. It does not require or restrict the use of admission control techniques. It signals congestion through bridges, unlike mechanisms that are specific to individual media access control methods.

The proposed congestion notification mechanism addresses the needs of non-TCP/IP protocols so the existing explicit congestion notification (ECN) mechanism specified by the IETF is not directly applicable.

c) Easy for the document reader to select the relevant specification.

IEEE Std 802.1Q is the natural reference for VLAN bridging technology, which will make the capabilities added by this amendment easy to locate.

12.4 Technical feasibility

For a project to be authorized, it shall be able to show its technical feasibility. At a minimum, the proposed project shall show:

a) Demonstrated system feasibility.

Congestion notification techniques have shown to be useful even in networks that are as difficult to control as the Internet. The proposed amendment will be applied only in networks of limited bandwidth delay product and where both bridges and end stations are typically under the control of a single administration, reducing the risk that the benefits of the technique will be eroded by over extended control loops or by some of the end stations 'gaming the system'.

The amendment will specify a one way bandwidth delay product across the congestion controlled domain of 8 typical frame sizes (or less) and simulation and analysis will verify performance characteristics up to the advertised bandwidth delay product.

It has been shown that end station rate limiting capabilities, suitable for use with congestion notification, can be implemented in hardware at acceptable cost.

b) Proven technology, reasonable testing.

The proposed amendment is based on extensive simulation and analysis in an area that has been studied for over 20 years.

c) Confidence in reliability.

In keeping with best practice in this technical area, both end station and bridge behavior will be specified, and the performance, stability, and fairness of the congestion control algorithm and resulting network throughput simulated and analyzed to the bounds of the specification.

d) Coexistence of 802 wireless standards specifying devices for unlicensed operation.

Not applicable.

12.5 Economic feasibility

For a project to be authorized, it shall be able to show economic feasibility (so far as can reasonably be estimated), for its intended applications. At a minimum, the proposed project shall show:

a) Known cost factors, reliable data.

The proposed amendment will retain existing cost characteristics of bridges including simplicity of queue structures and will not require maintenance of additional queues or queue state beyond the existing per traffic class (priority) queues for conformance to either its mandatory or optional provisions. In particular per flow queuing will not be required.

The proposed amendment may require some functions, specifically the generation of congestion notification frames, at a rate and within a time not practical for some existing and otherwise conformant bridge implementation architectures. However these functions can be performed by some existing bridges with known implementation costs.

The proposed amendment is technically feasible, in the envisaged application environment, with minimal flow state in end stations and will allow for complexity/throughput optimization trade-offs.

b) Reasonable cost for performance.

The proposed technology will reduce overall network costs where a separate storage network is currently required, and overall system costs by allowing the use of a data storage network where the cost of provide a separate network is excessive.

c) Consideration of installation costs.

Installation costs of VLAN Bridges are not expected to be significantly affected, any increase in network design costs is expected to be more than offset by a reduction in the number of separate networks required.