Network Working Group
Request for Comments: 3248
Category: Informational

G. Armitage
Swinburne University of Technology
B. Carpenter
IBM
A. Casati
Lucent Technologies
J. Crowcroft
University of Cambridge
J. Halpern
Consultant
B. Kumar
Corona Networks Inc.
J. Schnizlein
Cisco Systems

March 2002

A Delay Bound alternative revision of RFC 2598

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Abstract

For historical interest, this document captures the EF Design Team’s proposed solution, preferred by the original authors of RFC 2598 but not adopted by the working group in December 2000. The original definition of EF was based on comparison of forwarding on an unloaded network. This experimental Delay Bound (DB) PHB requires a bound on the delay of packets due to other traffic in the network. At the Pittsburgh IETF meeting in August 2000, the Differentiated Services working group faced serious questions regarding RFC 2598 - the group’s standards track definition of the Expedited Forwarding (EF) Per Hop Behavior (PHB). An ’EF Design Team’ volunteered to develop a re-expression of RFC 2598, bearing in mind the issues raised in the DiffServ group. At the San Diego IETF meeting in December 2000 the DiffServ working group decided to pursue an alternative re-expression of the EF PHB.

Specification of Requirements

This document is for Informational purposes only. If implementors choose to experiment with the DB PHB, key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are interpreted as described in RFC 2119 [3].

1 Introduction

RFC 2598 was the Differentiated Services (DiffServ) working group’s first standards track definition of the Expedited Forwarding (EF) Per Hop Behavior (PHB) [1]. As part of the DiffServ working group’s ongoing refinement of the EF PHB, various issues were raised with the text in RFC 2598 [2].

After the Pittsburgh IETF meeting in August 2000, a volunteer ’EF design team’ was formed (the authors of this document) to propose a new expression of the EF PHB. The remainder of this Informational document captures our feedback to the DiffServ working group at the San Diego IETF in December 2000. Our solution focussed on a Delay Bound (DB) based re-expression of RFC 2598 which met the goals of RFC 2598’s original authors. The DiffServ working group ultimately chose an alternative re-expression of the EF PHB text, developed by the authors of [2] and revised to additionally encompass our model described here.

Our proposed Delay Bound solution is archived for historical interest. Section 2 covers the minimum, necessary and sufficient description of what we believed qualifies as ’DB’ behavior from a single node. Section 3 then discusses a number of issues and assumptions made to support the definition in section 2.

2. Definition of Delay Bound forwarding

For a traffic stream not exceeding a particular configured rate, the goal of the DB PHB is a strict bound on the delay variation of packets through a hop.

This section will begin with the goals and necessary boundary conditions for DB behavior, then provide a descriptive definition of DB behavior itself, discuss what it means to conform to the DB definition, and assign the experimental DB PHB code point.

2.1 Goal and Scope of DB

For a traffic stream not exceeding a configured rate the goal of the DB PHB is a strict bound on the delay variation of packets through a hop.

Traffic MUST be policed and/or shaped at the source edge (for example, on ingress to the DS-domain as discussed in RFC 2475 [5]) in order to get such a bound. However, specific policing and/or shaping rules are outside the scope of the DB PHB definition. Such rules MUST be defined in any per-domain behaviors (PDBs) composed from the DB PHB.

A device (hop) delivers DB behavior to appropriately marked traffic received on one or more interfaces (marking is specified in section 2.4). A device SHALL deliver the DB behavior on an interface to DB marked traffic meeting (i.e. less than or equal) a certain arrival rate limit R.

If more DB traffic arrives than is acceptable, the device is NOT REQUIRED to deliver the DB behavior. However, although the original source of DB traffic will be shaped, aggregation and upstream jitter ensure that the traffic arriving at any given hop cannot be assumed to be so shaped. Thus a DB implementation SHOULD have some tolerance for burstiness - the ability to provide EF behavior even when the arrival rate exceeds the rate limit R.

Different DB implementations are free to exhibit different tolerance to burstiness. (Burstiness MAY be characterized in terms of the number of back-to-back wire-rate packets to which the hop can deliver DB behavior. However, since the goal of characterizing burstiness is to allow useful comparison of DB implementations, vendors and users of DB implementations MAY choose to utilize other burstiness metrics.)

The DB PHB definition does NOT mandate or recommend any particular method for achieving DB behavior. Rather, the DB PHB definition identifies parameters that bound the operating range(s) over which an implementation can deliver DB behavior. Implementors characterize their implementations using these parameters, while network designers and testers use these parameters to assess the utility of different DB implementations.

2.2 Description of DB behavior

For simplicity the definition will be explained using an example where traffic arrives on only one interface and is destined for another (single) interface.

The crux of this definition is that the difference in time between when a packet might have been delivered, and when it is delivered, will never exceed a specifiable bound.

Given an acceptable (not exceeding arrival rate limit R) stream of DB packets arriving on an interface:

> There is a time sequence E(i) when these packets would be delivered at the output interface in the absence of competing traffic. That is, E(i) are the earliest times that the packets could be delivered by the device.
>
> In the presence of competing traffic, the packets will be delayed to some later time D(i).

Competing traffic includes all DB traffic arriving at the device on other ports, and all non-DB traffic arriving at the device on any port.

DB is defined as the behavior which ensures, for all i, that:

$$D(i) - E(i) <= S * MTU/R.$$

MTU is the maximum transmission unit (packet size) of the output. R is the arrival rate that the DB device is prepared to accept on this interface.

Note that D(i) and E(i) simply refer to the times of what can be thought of as "the same packet" under the two treatments (with and without competing traffic).

The score, S, is a characteristic of the device at the rate, R, in order to meet this defined bound. This score, preferably a small constant, depends on the scheduling mechanism and configuration of the device.

2.3 Conformance to DB behavior

An implementation need not conform to the DB specification over an arbitrary range of parameter values. Instead, implementations MUST specify the rates, R, and scores S, for which they claim conformance with the DB definition in section 2.2, and the implementation-specific configuration parameters needed to deliver conformant behavior. An implementation SHOULD document the traffic burstiness it can tolerate while still providing DB behavior.

The score, S, and configuration parameters depend on the implementation error from an ideal scheduler. Discussion of the ability of any particular scheduler to provide DB behavior, and the conditions under which it might do so, is outside the scope of this document.

The implementor MAY define additional constraints on the range of configurations in which DB behavior is delivered. These constraints MAY include limits on the total DB traffic across the device, or total DB traffic targeted at a given interface from all inputs.

This document does not specify any requirements on DB implementation’s values for R, S, or tolerable burstiness. These parameters will be bounded by real-world considerations such as the actual network being designed and the desired PDB.

2.4 Marking for DB behavior

One or more DiffServ codepoint (DSCP) value may be used to indicate a requirement for DB behavior [4].

By default we suggest an ’experimental’ DSCP of 101111 be used to indicate that DB PHB is required.

3. Discussion

This section discusses some issues that might not be immediately obvious from the definition in section 2.

3.1 Mutability

Packets marked for DB PHB MAY be remarked at a DS domain boundary only to other codepoints that satisfy the DB PHB. Packets marked for DB PHBs SHOULD NOT be demoted or promoted to another PHB by a DS domain.

3.2 Tunneling

When DB packets are tunneled, the tunneling packets must be marked as DB.

3.3 Interaction with other PHBs

Other PHBs and PHB groups may be deployed in the same DS node or domain with the DB PHB as long as the requirement of section 2 is met.

3.4 Output Rate not specified

The definition of DB behavior given in section 2 is quite explicitly given in terms of input rate R and output delay variation D(i) - E(i). A scheduler’s output rate does not need to be specified, since (by design) it will be whatever is needed to achieve the target delay variation bounds.

3.5 Jitter

Jitter is not the bounded parameter in DB behavior. Jitter can be understood in a number of ways, for example the variability in inter-packet times from one inter-packet interval to the next. However, DB behavior aims to bound a related but different parameter - the variation in delay between the time packets would depart in the absence of competing traffic, E(i), and when they would depart in the presence of competing traffic, D(i).

3.6 Multiple Inputs and/or Multiple Outputs

The definition of ’competing traffic’ in section 2.2 covers both the single input/single output case and the more general case where DB traffic is converging on a single output port from multiple input ports. When evaluating the ability of an DB device to offer DB behavior to traffic arriving on one port, DB traffic arriving on other ports is factored in as competing traffic.

When considering DB traffic from a single input that is leaving via multiple ports, it is clear that the behavior is no worse than if all of the traffic could be leaving through each one of those ports individually (subject to limits on how much is permitted).

3.7 Fragmentation and Rate

Where an ingress link has an MTU higher than that of an egress link, it is conceivable packets may be fragmented as they pass through a Diffserv hop. However, the unpredictability of fragmentation is significantly counter to the goal of providing controllable QoS. Therefore we assume that fragmentation of DB packets is being avoided (either through some form of Path MTU discovery, or configuration), and does not need to be specifically considered in the DB behavior definition.

3.8 Interference with other traffic

If the DB PHB is implemented by a mechanism that allows unlimited preemption of other traffic (e.g., a priority queue), the implementation MUST include some means to limit the damage DB traffic could inflict on other traffic. This will be reflected in the DB device’s burst tolerance described in section 2.1.

3.9 Micro flow awareness

Some DB implementations may choose to provide queuing and scheduling at a finer granularity, (for example, per micro flow), than is indicated solely by the packet’s DSCP. Such behavior is NOT precluded by the DB PHB definition. However, such behavior is also NOT part of the DB PHB definition. Implementors are free to characterize and publicize the additional per micro flow capabilities of their DB implementations as they see fit.

3.10 Arrival rate ’R’

In the absence of additional information, R is assumed to be limited by the slowest interface on the device.

In addition, an DB device may be characterized by different values of R for different traffic flow scenarios (for example, for traffic aimed at different ports, total incoming R, and possibly total per output port incoming R across all incoming interfaces).

4. IANA Considerations

This document suggests one experimental codepoint, 101111. Because the DSCP is taken from the experimental code space, it may be re-used by other experimental or informational DiffServ proposals.

5. Conclusion.

This document defines DB behavior in terms of a bound on delay variation for traffic streams that are rate shaped on ingress to a DS domain. Two parameters - capped arrival rate (R) and a ’score’ (S), are defined and related to the target delay variation bound. All claims of DB ’conformance’ for specific implementations of DB behavior are made with respect to particular values for R, S, and the implementation’s ability to tolerate small amounts of burstiness in the arriving DB traffic stream.

Security Considerations

To protect itself against denial of service attacks, the edge of a DS domain MUST strictly police all DB marked packets to a rate negotiated with the adjacent upstream domain (for example, some value less than or equal to the capped arrival rate R). Packets in excess of the negotiated rate MUST be dropped. If two adjacent domains have not negotiated an DB rate, the downstream domain MUST use 0 as the rate (i.e., drop all DB marked packets).

Since PDBs constructed from the DB PHB will require that the upstream domain police and shape DB marked traffic to meet the rate negotiated with the downstream domain, the downstream domain's policer should never have to drop packets. Thus these drops (or a summary of these drops) SHOULD be noted (e.g., via rate-limited SNMP traps) as possible security violations or serious misconfiguration.

Overflow events on an DB queue MAY also be logged as indicating possible denial of service attacks or serious network misconfiguration.

Acknowledgments

This document is the product of the volunteer 'EF Resolve' design team, building on the work of V. Jacobson, K. Nichols, K. Poduri [1] and clarified through discussions with members of the DiffServ working group (particularly the authors of [2]). Non-contentious text (such as the use of DB with tunnels, the security considerations, etc.) were drawn directly from equivalent text in RFC 2598.

Intellectual Properties Considerations

To establish whether any considerations apply to the idea expressed in this document, readers are encouraged to review notices filed with the IETF and stored at:

http://www.ietf.org/ipr.html

References

[1] Jacobson, V., Nichols, K. and K. Poduri, "An Expedited Forwarding PHB", RFC 2598, June 1999.

[2] Davie, B., Charny, A., Baker, F., Bennett, J.C.R., Benson, K., Le Boudec, J.Y., Chiu, A., Courtney, W., Davari, S., Firoiu, V., Kalmanek, C., Ramakrishnan, K. and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[3] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[4] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[5] Black, D., Blake, S., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

Authors (volunteer EF Design Team members)

Grenville Armitage
Center for Advanced Internet Architectures
Swinburne University of Technology,
Melbourne, Australia
EMail: garmitage@swin.edu.au

Brian E. Carpenter (team observer, WG co-chair)
IBM Zurich Research Laboratory
Saeumerstrasse 4
8803 Rueschlikon
Switzerland
EMail: brian@hursley.ibm.com

Alessio Casati
Lucent Technologies
Swindon, WI SN5 7DJ United Kingdom
EMail: acasati@lucent.com

Jon Crowcroft
Marconi Professor of Communications Systems
University of Cambridge
Computer Laboratory
William Gates Building
J J Thomson Avenue
Cambridge
CB3 0FD
Phone: +44 (0)1223 763633
EMail: Jon.Crowcroft@cl.cam.ac.uk

Joel M. Halpern
P. O. Box 6049
Leesburg, VA 20178
Phone: 1-703-371-3043
EMail: jmh@joelhalpern.com

Brijesh Kumar
Corona Networks Inc.,
630 Alder Drive,
Milpitas, CA 95035
EMail: brijesh@coronanetworks.com

John Schnizlein
Cisco Systems
9123 Loughran Road
Fort Washington, MD 20744
EMail: john.schnizlein@cisco.com

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.