TSVWG K. Chan
Internet-Draft J. Babiarz
Expires: July 22, 2006 Nortel Networks
F. Baker
Cisco Systems
January 18, 2006

Aggregation of DiffServ Service Classes
draft-chan-tsvwg-diffserv-class-aggr-03

Status of this Memo

Copyright Notice

Abstract

In the core of a high capacity network, service differentiation is still needed to support applications' utilization of the network. Applications with similar traffic characteristics and performance requirements are mapped into different diffserv service classes based on end-to-end behavior requirements of the applications. However, some network segments may be configured in such a way that a single forwarding treatment satisfy the traffic characteristics and performance requirements of two or more service classes. For such cases, it may be desirable to aggregate two or more service classes into a forwarding treatment. This document provides guidelines for aggregation of service classes into forwarding treatments.

Table of Contents

1. Introduction

In the core of a high capacity network, it is common for the network to be engineered in such a way that a major link, switch, or router can fail and the result be a routed network that still meets ambient SLAs. The implication of this is that there is sufficient capacity on any given link that all SLAs sold can be simultaneously supported at their respective maximum rates, and this remains true after re-routing (either IP re-routing or MPLS protection-mode switching) has occurred.

It is frequently argued that such over provisioning meets the requirements of all traffic without further QoS treatment, and from a certain perspective that is true. However, as the process of network convergence continues, certain services still have issues. While delay and jitter is perfectly acceptable for elastic applications, real time applications are negatively affected, and in extreme cases (such as some reported around the September 2001 attacks on the US East Coast, or under extreme DOS load) such surges could disrupt routing.

The treatment aggregates recommended herein are designed to aggregate the service classes in Diffserv Service Classes [5] in such a manner as to protect real-time traffic and routing, on the assumption that real-time sessions are protected from each other by admission at the edge, and provide a staged response to stress.

The document Diffserv Service Classes [5] provides the basic diffserv classes from the points of view of the application requiring specific end-to-end behaviors from the network. At some network segments of the end-to-end path, the number of levels of network treatment differentiation may be less than the number of service classes that the network segment needs to support. In such situation, that network segment needs to use the same treatment to support more than one service class. In this document we provide guidelines of how multiple service classes may be aggregated into a forwarding treatment aggregate. Notice in a given domain, we recommend the supported service classes be aggregated into forwarding treatment aggregates, this does not mean all service classes needs to be supported and hence not all forwarding treatment aggregates needs to be supported. Which service classes and which forwarding treatement aggregates is supported by a domain is up to the domain administration and may be influenced by business reasons. We've also provided some terminology and requirement for performing this aggregation.

1.1. Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [3].

2. Terminology

We try to use existing definition of terms from current RFCs. We have also added new definition of terms here when necessary.

- Treatment Aggregate. This term is used here to indicate the aggregate of DiffServ service classes. This is different from Behavior Aggregate and Traffic Aggregate because Treatment Aggregate is only concerned with the treatment of the aggregated traffic. It does not concern with how the aggregated traffic is marked, and hence does not put a restriction on the aggregated traffic having a single codepoint that have a single PHB.

3. Overview of Service Class Aggregation

In some deployments, especially in the middle of the network where network capacity is higher, traffic treatment differentiation may be less granular. In these deployments, aggregation of the different service classes may be more practical.

These aggregations have the following requirements:

1. The end-to-end network performance characteristic required by the application MUST be supported. This performance characteristic is represented by the use of Diffserv Service Classes [5].

2. The treatment aggregate MUST exhibit the strictest requirement of its member service classes.

3. The treatment aggregate SHOULD only contain member service classes with similar traffic characteristic and performance requirements.

4. The notion of the individual end-to-end service classes MUST NOT be destroyed when aggregation is performed. Each domain along the end-to-end path may perform aggregation differently, based on the original end-to-end service classes. We RECOMMEND an easy way to accomplish this by NOT altering the DSCP used to indicate the end-to-end service class. But some administrative domains may require the use of their own marking, when this is needed,

the original end-to-end service class indication MUST be restored upon exit of such administrative domains.

5. Each treatment aggregate have limited resource, hence traffic conditioning and/or admission control MUST be performed for each service class aggregating into the treatment aggregate.

4. Service Classes to Treatment Aggregate Mapping

The service class and DSCP selection in Diffserv Service Classes [5] has been defined to allow in many instances mapping of two or possibly more service classes into a single treatment aggregate. Noticing there is a physical-space/time relationship between link speed, queue depth, delay, and jitter. The degree of aggregation, hence the number of treatment aggregates, will depend on if the domain implementing the aggregation will have link speed high enough to minimize the affects of mixing traffic with different packet size, different transmit rates on buffering/queue depth, and finally its impact on loss, delay, and jitter. With the general rule of thumb being higher link speeds allows higher degree of aggregation/smaller number of treatment aggregates. But all requires some forms of traffic conditioning and/or admission control.

4.1. Mapping Service Classes into Four Treatment Aggregates

For most of today's high speed links, the use of one network control traffic treatment aggregate and three user traffic treatment aggregates is sufficient to handle the requirement of all the service classes indicated in Diffserv Service Classes [5]. We use the performance requirement (tolerance to loss, delay, and jitter) from the application/end user as the guidance on how to map the service classes into treatment aggregates. We have also used Section 3.1 of RFC 1633 [6] to provide us with guidance on the definition of Real Time and Elastic application requirements. An overview of the mapping between service classes and four treatment aggregates is provided by Figure 1, with the mapping based on performance requirement.

Notice we recommended certain service classes be mapped into specific treatment aggregates. But this does not mean that all the service classes recommended for that treatment aggregate needs to be supported. Hence for a domain, a treatment aggregate may contain a subset of the service classes recommended in this document, they being the service classes supported by that domain. A domain's treatment of none-supported service classes is that domain's local policy. This local policy may be influenced by its agreement with its customers. Such treatment may use the Elastic Treatment

Aggregate, dropping the packets, or some other arrangements.

Treatment Aggregate	Tolerance to			Service Class	Tolerance to		
	Loss	Delay	Jitter		Loss	Delay	Jitter
Network Control	Low	Low	Yes	Network Control	Low	Low	Yes
Real Time	Very Low	Very Low	Very Low	Telephony	VLow	VLow	VLow
				Signaling	Low	Low	Yes
				Multimedia Conferencing	Low - Medium	Very Low	Low
				Real-time Interactive	Low	Very Low	Low
				Broadcast Video	Very Low	Medium	Low
Assured Elastic	Low	Low - Medium	Yes	Multimedia Streaming	Low - Medium	Medium	Yes
				Low Latency Data	Low	Low - Medium	Yes
				OAM	Low	Medium	Yes
				High Throughput Data	Low	Medium - High	Yes
Elastic	Not Specified			Standard	Not Specified		
				Low Priority Data	High	High	Yes

Figure 1: Treatment Aggregate and Service Class Performance Requirements

4.1.1. Network Control Treatment Aggregate

The Network Control Treatment Aggregate aggregates all service classes that is functionally necessary for the survival of a network during a DOS or other high traffic load interval. The theory is that whatever else is true, the network must protect itself. This

includes the traffic that Diffserv Service Classes [5] characterizes as in the Network Control Service Class.

The DSCPs of the original service class remain an important consideration and should be preserved during aggregation. Traffic bearing these DSCPs is carried in a common queue or class with a PHB as described in RFC 2309 [9] and RFC 2474 [4] for CS6. And for a lower probability of packet loss, bearing a relatively deep target mean queue depth (min-threshold if RED is being used).

4.1.2. Real Time Treatment Aggregate

The Real Time Treatment Aggregate aggregates all real time (inelastic) service classes. The theory is that real-time traffic is admitted under some model and controlled by an SLA managed at the edge of the network prior to aggregation. As such, there is a predictable and enforceable upper bound on the traffic that can enter such a queue, and to provide predictable variation in delay it must be protected from bursts of elastic traffic.

This treatment aggregate may include the following service classes from Diffserv Service Classes [5], in addition to other locally defined classes: Telephony, Signaling, Multimedia Conferencing, Real-time Interactive, Broadcast Video.

Traffic in each service class that is going to be aggregated into the treatment aggregate should be conditioned prior to aggregating. It is recommended that per service class admission control procedure be used followed with per service class policing so that any individual service class does not generate more than what it is allowed. Further, additional admission control and policing may be used on the sum of all service classes aggregated.

The DSCPs of the original service classes remain an important consideration and should be preserved during aggregation. Traffic bearing these DSCPs is carried in a common queue or class with a PHB as described in RFC 3246 [11] and RFC 3247 [12].

4.1.3. Assured Elastic Treatment Aggregate

The Assured Elastic Treatment Aggregate aggregates all elastic traffic that uses the Assured Forwarding model as described in RFC 2597 [10]. The premise of such service is that an SLA is negotiated that includes a "committed rate" and the ability to exceed that rate (and perhaps a second "excess rate") in exchange for a higher probability of loss using AQM [9] or ECN flagging [13] for the portion of traffic deemed to be in excess.

This treatment aggregate may include the following service classes from Diffserv Service Classes [5], in addition to other locally defined classes: Multimedia Streaming, Low Latency Data, OAM, High Throughput Data.

The DSCPs of the original service classes remain an important consideration and should be preserved during aggregation. Traffic bearing these DSCPs is carried in a common queue or class with a PHB as described in RFC 2597 [10]. In effect, appropriate target rate thresholds have been applied at the edge, dividing traffic into AFn1 (committed, for any value of n), AFn2, and AFn3 (excess). The service SHOULD be engineered so that AFn1 marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Since the traffic is elastic and responds dynamically to packet loss, Active Queue Management [9] SHOULD be used primarily to reduce forwarding rate to the minimum assured rate at congestion points. The probability of loss of AFn1 traffic MUST NOT exceed the probability of loss of AFn2 traffic, which in turn MUST NOT exceed the probability of loss of AFn3.

If RED [9] is used as an AQM algorithm, the min-threshold specifies a target queue depth for each of AFn1, AFn2, AFn3, and the max-threshold specifies the queue depth above which all traffic with such a DSCP is dropped or ECN marked. Thus, in this Treatment Aggregate, the following inequality should hold in queue configurations:

- min-threshold AFn3 < max-threshold AFn3

- max-threshold AFn3 <= min-threshold AFn2

- min-threshold AFn2 < max-threshold AFn2

- max-threshold AFn2 <= min-threshold AFn1

- min-threshold AFn1 < max-threshold AFn1

- max-threshold AFn1 <= memory assigned to the queue

Note: This configuration tends to drop AFn3 traffic before AFn2 and AFn2 before AFn1. Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

4.1.4. Elastic Treatment Aggregate

The Elastic Treatment Aggregate aggregates all remaining elastic traffic. The premise of such service is that there is no intrinsic SLA differentiation of traffic, but that AQM [9] or ECN flagging [13] is appropriate for such traffic.

This treatment aggregate may include the following service classes from Diffserv Service Classes [5], in addition to other locally defined classes: Standard, Low Priority Data.

The DSCPs of the original service classes remain an important consideration and should be preserved during aggregation. Traffic bearing these DSCPs is carried in a common queue or class with a PHB as described in RFC 2309 [9]. The AQM thresholds for Elastic traffic MAY be separately set, so that Low Priority Data traffic is dropped before Standard traffic, but this is not a requirement.

4.2. Treatment Aggregate Summary

The behavior for the above mentioned Treatment Aggregates are summarized in the following table:

Treatment Aggregate	Treatment Aggregate Behavior	DSCP name
Network Control	CS (RFC 2474)	CS6
Real Time	EF (RFC 3246)	EF, CS5, AF41, AF42, AF43, CS4, CS3
Assured Elastic	AF (RFC 2597)	CS2, AF31, AF21, AF11
		AF32, AF22, AF12
		AF13, AF23, AF33
Elastic	Default (RFC 2474)	Default, (CS0)
		CS1

Figure 2: Treatment Aggregate Behavior Summary

5. Using MPLS for Treatment Aggregates

RFC 2983 on DiffServ and Tunnels [7] and RFC 3270 on MPLS Support of DiffServ [8] provided very good background on this topic. This document provides an example of using the E-LSP, EXP Inferred PHB Scheduled Class (PSC) Label Switched Path (LSP), notion indicated in MPLS Support of DiffServ [8] for Treatment Aggregates.

When Treatment Aggregates are represented in MPLS using EXP Inferred PSC LSP, we recommend the following usage of MPLS EXP field for Treatment Aggregates.

Treatment Aggregate	MPLS EXP	DSCP name	DSCP value
Network Control	110	CS6	110000
Real Time	100	EF	101110
		CS5	101000
		AF41,AF42	100010,100100
		AF43	100110
		CS4	100000
		CS3	011000
Assured Elastic	010*	CS2	010000
		AF31	011010
		AF21	010010
		AF11	001010
	011*	AF32	011100
		AF22	010100
		AF12	001100
		AF33	011110
		AF23	010110
		AF13	001110
Elastic	000*	Default (CS0)	000000
	001*	CS1	001000

Figure 3: Treatment Aggregate and MPLS EXP Field Usage

Notes *: For Assured Elastic (and Elastic) Treatment Aggregate, the usage of 010 or 011 (000 or 001) depends on the drop probability.

The above table indicates the recommended usage of EXP field for Treatment Aggregates. Because many deployment of MPLS is on a per

domain basis, each domain have total control of its EXP usage, each domain may use a different EXP field allocation for the domain's supported Treatment Aggregates.

5.1. Network Control Treatment Aggregate with E-LSP

The usage of E-LSP for Network Control Treatment Aggregate needs to cohere to the recommendations indicated in section 4.1.1 of this document and section 3.2 of Diffserv Service Classes [5]. Reinforcing these recommendations, there should be no drop precedence associated with the MPLS PSC used for Network Control Treatment Aggregate because dropping of Network Control Treatment Aggregate traffic should be prevented.

5.2. Real Time Treatment Aggregate with E-LSP

In addition to the recommendations provided in section 4.1.2 of this document and in Diffserv Service Classes [5], we want to indicate that Real Time Treatment Aggregate traffic should not be dropped, as some of the traffic carried in the Real Time Treatment Aggregate does not react well to dropped packets. As indicated in section 4.1.2 of this document, admission control should be performed on each Service Class contributing to the Real Time Treatment Aggregate to prevent packet loss due to insufficient resource allocated to Real Time Treatment Aggregate. Further, admission control and policing may also be applied on the sum of all traffic aggregated into this treatment aggregate.

5.3. Assured Elastic Treatment Aggregate with E-LSP

EXP field markings of 010 and 011 are used for Assured Elastic Treatment Aggregate. The two encodings are used to provide two levels of drop precedence indications, with 010 encoded traffic having a lower probability of being drop then 011 encoded traffic. This provides for the mapping of CS2, AF31, AF21, and AF11 into EXP 010; and AF32, AF22, AF12 and AF33, AF23, AF13 into EXP 011.

5.4. Elastic Treatment Aggregate with E-LSP

EXP field markings of 000 and 001 are used for Elastic Treatment Aggregate. The two encodings are used to provide two levels of drop precedence indications, with 000 encoded traffic having a lower probability of being drop then 001 encoded traffic. This provides for the mapping of Default/CS0 into 000; and CS1 into 001. Notice with this mapping, during congestion, CS1 marked traffic may be starved.

5.5. Treatment Aggregates and L-LSP

Because L-LSP (Label Only Inferred PSC LSP) supports a single PSC per LSP, the support of each Treatment Aggregate is on a per LSP basis. This document does not further specify any additional recommendation (beyond what had been indicated in section 4 of this document) for Treatment Aggregate to L-LSP mapping, leaving this to each individual MPLS domain administration.

6. Treatment Aggregates and Inter Provider Relationships

When Treatment Aggregates are used at the provider boundaries, we recommend the Inter Provider Relationship be based on Diffserv Service Classes [5]. This allows the admission control into each Treatment Aggregate of a provider domain be based on the admission control of traffic into the supported Service Classes, as indicated by the discussions in section 4 of this document.

If the Inter Provider Relationship needs to be based on Treatment Aggregates specified by this document, the exact Treatment Aggregate content and representation must be agreed between the peering providers.

7. Security Considerations

This document discusses policy of using Differentiated Services and its service classes. If implemented as described, it should require the network to do nothing that the network has not already allowed. If that is the case, no new security issues should arise from the use of such a policy.

It is possible for the policy to be applied incorrectly, or for a wrong policy to be applied in the network for the defined aggregation. In that case, a policy issue exists that the network must detect, assess, and deal with. This is a known security issue in any network dependent on policy-directed behavior.

A well known flaw appears when bandwidth is reserved or enabled for a service (for example, voice transport) and another service or an attacking traffic stream uses it. This possibility is inherent in DiffServ technology, which depends on appropriate packet markings. When bandwidth reservation or a priority queuing system is used in a vulnerable network, the use of authentication and flow admission is recommended. To the author's knowledge, there is no known technical way to respond to or act upon a data stream that has been admitted for service but that it is not intended for authenticated use.

8. IANA Considerations

To be completed.

9. Acknowledgements

10. Normative References

[1] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[2] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

[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] Babiarz, J., "Configuration Guidelines for DiffServ Service Classes", draft-ietf-tsvwg-diffserv-service-classes-01 (work in progress), July 2005.

[6] Braden, B., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[7] Black, D., "Differentiated Services and Tunnels", RFC 2983, October 2000.

[8] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[9] Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K., Shenker, S., Wroclawski, J., and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.

[10] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[11] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An

Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[12] Charny, A., Bennet, J., Benson, K., Boudec, J., Chiu, A., Courtney, W., Davari, S., Firoiu, V., Kalmanek, C., and K. Ramakrishnan, "Supplemental Information for the New Definition of the EF PHB (Expedited Forwarding Per-Hop Behavior)", RFC 3247, March 2002.

[13] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

Authors' Addresses

Kwok Ho Chan
Nortel Networks
600 Technology Park Drive
Billerica, MA 01821
US

Phone: +1-978-288-8175
Fax: +1-978-288-4690
Email: khchan@nortel.com

Jozef Z. Babiarz
Nortel Networks
3500 Carling Avenue
Ottawa, Ont. K2H 8E9
Canada

Phone: +1-613-763-6098
Fax: +1-613-768-2231
Email: babiarz@nortel.com

Fred Baker
Cisco Systems
1121 Via Del Rey
Santa Barbara, CA 93117
US

Phone: +1-408-526-4257
Fax: +1-413-473-2403
Email: fred@cisco.com

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Statement

Acknowledgment

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