Relative Differentiated Services in the Internet: Issues and Mechanisms

Constantinos Dovrolis
University of Wisconsin, Madison
dovrolis@ece.wisc.edu

Dimitrios Stiliadis
Bell Laboratories
stiliadi@dnrc.bell-labs.com

1 Introduction

In the context of relative differentiated services the network traffic is grouped in N classes of service which differ in their *relative forwarding quality*, i.e., class i is *better* than class $i-1$ in terms of local (per-hop) forwarding measures of the queueing delays and packet losses. Users and applications, on the other hand, select adaptively the class that best meets their requirements and/or cost constraints. In this short paper we present, focusing on queueing delays only, four different service models that offer such a class relative differentiation. Simulation results illustrate that a *predictable and controllable delay differentiation* can be achieved in only one of these models. A more complete treatment of this subject appears in [1].

2 Relative Delay Differentiation

A. Strict Priorities: In this service model, the highest backlogged class is always serviced first. The main problem here is that the network operator has no means to control the delay spacing between classes, and also, lower classes can experience extensive service starvation if there is no restriction on the load of higher classes.

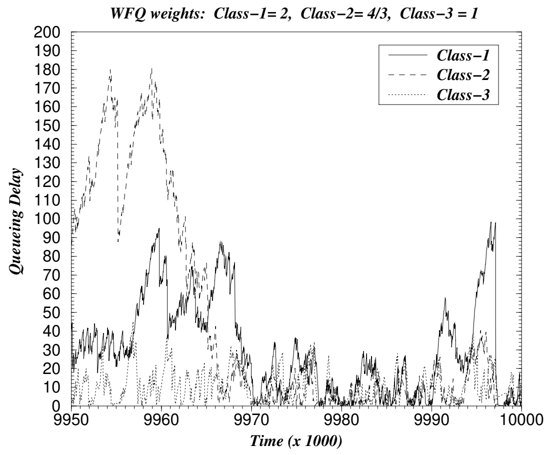
B. Service Rate Differentiation: In this model, each class is allocated a certain service rate depending on its expected arrival rate. Higher classes have a larger service-to-arrival rate ratio, and thus, they are expected to have lower queueing delays. This can be implemented with any Weighted Fair Queueing (WFQ) type of packet scheduler. Although such schedulers provide the network operator with a set of differentiation parameters that serve as 'tuning knobs' for the delay spacing between classes, this spacing is neither predictable, nor controllable. For example, if a high class becomes overloaded in a certain time period, it may have a lower service-to-arrival effective ratio than lower classes, and so it will be worse than those classes in terms of queueing delays. In general, a scheme that is only based on service rate differentiation cannot control the queueing delay differentiation between classes, since the latter depends on the short-term relation between the service rate distribution and the class load distribution.

C. Absolute Delay Differentiation: In this model, the network operator attempts to control the absolute delay spacing between classes. If $\bar{d}_i(I)$ is the average delay of class i measured in a relatively short time interval I, the absolute delay spacing has the form: $\bar{d}_i(I) - \bar{d}_j(I) = D_i - D_j$, with $D_1 > D_2 > \ldots > D_N$ being the *absolute delay differentiation parameters.* These constraints can be approximated in heavy load conditions using an Earliest Deadline First (EDF) scheduler in which a class i packet that arrives at time a_i is assigned a virtual deadline $a_i + D_i$; packets are serviced in the order of increasing virtual deadlines. The problem with the absolute delay differentiation model, however, is that it is effective only in high load conditions, where the delay spacing $D_i - D_j$ is feasible. Additionally, an absolute delay spacing is of limited value because it depends on the (varying) absolute magnitude of the queueing delays. For example, a delay spacing of 5 msec will be important if the queuing delays are in that order of magnitude, while it will be insignificant if the queueing delays are many tens of msec.

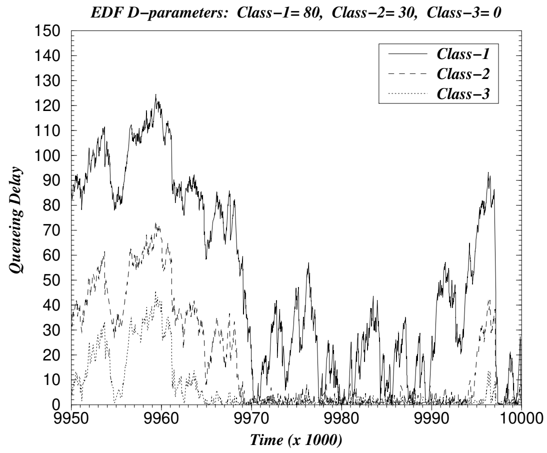
D. Proportional Delay Differentiation: In this model, the network operator controls the relative delay spacing between classes based on a set of proportionality constraints. If $\bar{d}_i(I)$ is the average delay of class i measured in a relatively short time interval I, the proportional delay spacing has the form: $\bar{d}_i(I)/\bar{d}_j(I) = \delta_i/\delta_j$, with $\delta_1 = 1 > \delta_2 > \ldots > \delta_N > 0$ being the *proportional delay differentiation parameters.* These constraints can be approximated in heavy load conditions using the Waiting-Time Priority (WTP) scheduler (originally proposed by L.Kleinrock) according to which the priority of a class i packet at time t, $w_i(t)$ time units after the packet's arrival, is $p_i(t) = w_i(t)/\delta_i$. Packets are serviced in the order of increasing priorities. An important property of the proportional delay differentiation model is that it does not require a priori knowledge of the arrival rate in each class. In fact, it can be implemented with dynamic adjustment of the service rate distribution between classes based on the current load of each class. In the case of the WTP scheduler, the load of a class is reflected in the waiting time of the packet at the head of each queue. Another attractive feature of the proportional delay differentiation model is that it has the same semantics throughout the utilization range, providing the network operator with predictable and controllable relative differentiation between classes independently of the load conditions.

3 Simulation Results

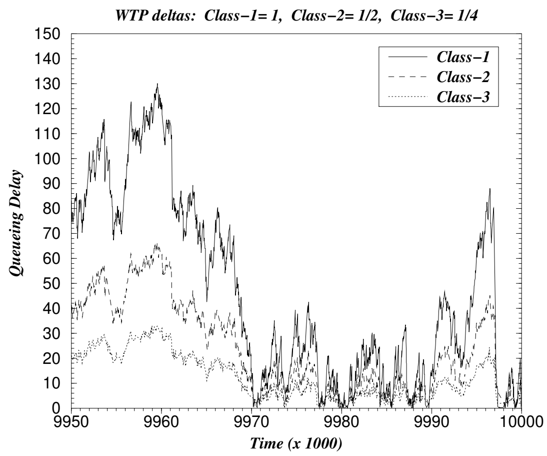
Figure 1 shows typical graphs of the queueing delays of three classes for the last three delay differentiation models. All figures refer to the same time interval (about 4200 packets) and to the same input streams in each class. The simulator models a single link that receives packet streams from three sources, with one source per-class. The average utilization is 95%, the per-source packet interarrivals follow a Pareto distribution with $\alpha = 1.9$, each packet transmission time is 11.2 time units, the reported delays are measured in packet transmission times, and the average load distribution between classes is: Class-1:50%, Class-2:30%, and Class-3:20%. The WFQ weights, the EDF D-parameters, and the WTP δ-parameters have been selected so that, in the long-run, the average-delays of the three classes are spaced as: $\bar{d}_1/\bar{d}_2 = 2$ and $\bar{d}_2/\bar{d}_3 = 2$. Note that the service rate differentiation that WFQ provides is not able to maintain a persistent delay differentiation between classes; in fact, in the left part of that graph Class-2 experiences larger delays than Class-1. EDF is able to provide the specified absolute delay differentiation only in times of high load overloads, while it is unpredictable in lower load conditions. WTP, finally, is able to maintain quite accurately the proportional delay differentiation irrespectively of the aggregate load and the class load distribution.



(a) Service Rate Differentiation



(b) Absolute Delay Differentiation



(c) Proportional Delay Differentiation

Figure 1: The per-class queueing delays in three relative differentiation models.

References

[1] C.Dovrolis, D.Stiliadis, and P.Ramanathan, "Proportional Differentiated Services: Delay Differentiation and Packet Scheduling," Tech. Rep. ECE-99-2, University of Wisconsin, Feb. 1999. Available at http://www.cae.wisc.edu/~dovrolis/TR-ECE-99-2.ps.