Achieving End-to-end Throughput Guarantee for TCP Flows in a Differentiated Services Network

Xiaoning He and Hao Che
Department of Electrical Engineering
The Pennsylvania State University
University Park, PA 16802, USA

Abstract—A challenge for the differentiated services (DS) architecture is how to simultaneously provide end-to-end assured service (AS) for transport control protocol (TCP) sessions and the best-effort service traffic, based on a single queue management. Our research results show that both intradomain and interdomain best-effort traffic can have adverse impact on the interdomain TCP traffic. This paper proposes a technique to achieve desired end-to-end throughput guarantee for TCP sessions. The proposed technique is composed of a series of measures which includes: (a) a path pinning mechanism for AS allowing aggregated bandwidth reservation for AS at *each intermediate router* in the forwarding path; (b) a packet marking strategy; (c) a dropping policy; (d) an adaptive dropping-threshold calculation method for queue management based on aggregated reserved bandwidth and real-time traffic measurement. The simulation results demonstrate that with this technique, a high end-to-end service assurance can be achieved for the TCP traffic, while a reasonably high throughput for best-effort traffic is maintained.

Keywords—**Differentiated Services, Quality of Services, Assured Service for TCP, Performance measurement, Adaptive algorithm design, End-to-End performance**

I. INTRODUCTION

As the Internet evolves into a global commercial infrastructure, there is a growing need to support quality of service (QoS) to applications. Recently, a radical approach, known as differentiated services (DS) [1], [8], has attracted much attention. The DS model is based on the assumption that resources are *abundant* in the core and bottlenecks occur only at the border nodes between domains. While offering multiple classes of service (CoSs), the DS model ensures scalability by keeping a stateless core and adhering to the IP (i.e. Internetworking Protocol) hop-by-hop forwarding paradigm. However, a key problem for this model is the conflict between maximizing resource utilization and achieving a high service assurance. In order to provide high service assurance, enough resources need to be provisioned to all the possible paths in the direction from a source to a destination.

Recently, Stoica and Zhang proposed a premium service model [3], [5]. The authors developed a technique (see [4]) that can pin a path without keeping per-flow state in the core routers and they designed algorithms to achieve end-to-end deterministic performance guarantee for premium service sessions using explicit paths. We believe that to provide end-to-end performance guarantee while achieving high resource utilization, a connection-oriented approach is necessary. In fact, as we shall see shortly, both intradomain and interdomain best-effort traffic can adversely impact the performance of interdomain TCP traffic.

The focus of this paper is to design algorithms for providing end-to-end service assurance for TCP applications while achieving high resource utilization and maintaining high scalability. The approach taken is to enable connection-oriented AS with aggregated resource reservation. The connection-oriented AS can be enabled by using a route pinning technique proposed in this paper, the one proposed in [4] or by combining DS with a connection-oriented networking architecture such as multiprotocol label switching (MPLS) [6], [7]. The core routers along the path do not keep per-flow state information but only the aggregated bandwidth reservation information for AS sessions as a whole. A key design is to decouple the treatment of the AS traffic from the best-effort traffic. With the assured and best-effort traffic sharing a single output port queue, an *adaptive* queue management algorithm is proposed. The algorithm is based on the aggregated reserved bandwidth as well as the measured level of aggregated conformant AS traffic. Simulation results show that the algorithm successfully suppresses the negative impact of both interdomain and intradomain best-effort traffic on the performance of the interdomain AS traffic, resulting in a high end-to-end service assurance for the AS traffic.

The rest of the paper is organized as follows. Section 2 presents a background introduction on the current status of the AS design for TCP applications. Section 3 describes the proposed scheme including a queue management algorithm based on the aggregated reserved bandwidth. Section 4 gives the experiment results on the proposed scheme. Finally, Section 5 concludes the paper and presents future research directions.

II. BACKGROUND

The original idea for designing AS for TCP applications was proposed by Clark and Fang [9]. In this model, AS provides better-than-best-effort service. Traffic is policed at every Internet service provider (ISP) domain edge node. At the edge node, conformant packets are marked as in-profile or IN and non-conformant packets are marked as out-profile or OUT. Both IN and OUT packets are injected into the core of the network, and the OUT packets are treated the same way as the best-effort packets. In each core router, a single first-in-first-out (FIFO) queue is used for both AS and best-effort traffic. A 2-level RED (i.e. random early detection) [12] packet dropping algorithm, called RIO (RED in-and-out) [9], is run based on traffic type. The performance of RED has been studied in [13], [14]. IN packets have a lower dropping probability than the best-effort packets and OUT packets. At each and every domain boundary, traffic is policed locally, and packets are subject to remarking before being injected into another domain, based on local congestion situation. However, since there is no end-to-end resource

0-7803-6494-5/00/$10.00 © 2000 IEEE

provisioning, an end-to-end service assurance is not guaranteed. Several other works on the improvement of this model in an attempt to achieve better service assurance and fairness were proposed [10], [11], all based on a connectionless forwarding paradigm.

The above studies did not consider end-to-end performance of the AS TCP sessions, in the presence of possible cross traffic, especially, the cross best-effort traffic with small round-trip time (RTT). The cross traffic could occur within a domain or at a domain boundary. Among other issues, a question is whether local control at each domain boundary can guarantee end-to-end performance for AS flows that cross multiple domains. After all, the ultimate performance measure is end-to-end service assurance, local control at each domain boundary, although has merit in its own right, may not guarantee end-to-end performance. To answer this question, we did simulation tests on the approach proposed in [9], using NS-2 from LBNL (Lawrence Berkerley National Laboratory).

The network setup is shown in Fig. 1. It is composed of three domains, with border routers R1 in domain 1, R2 and R3 in domain 2, and R4 in domain 3. The link bandwidth between routers are 33 $Mbps$ each and the buffer size is 50 packets for each output port of a router. There are ten hosts in domain 1 and each has an AS TCP session with one of the hosts in domain 3. Out of these ten sessions, 5 have target rates of 5 $Mbps$ and the other 5 have target rates of 1 $Mbps$. So the aggregated target rate for AS is 30 $Mbps$, which is lower than the link bandwidth between any two routers.

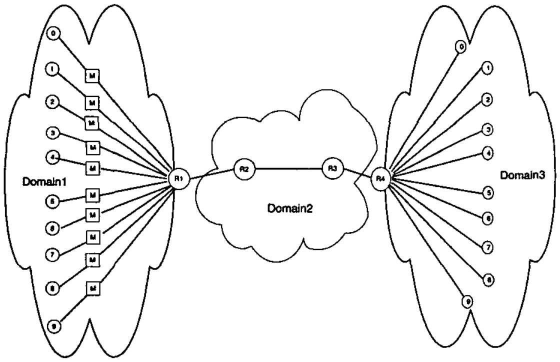

Fig. 1. Network setup for experiment 1

The parameters for RIO are set at $(min_{in}, max_{in}, P_{in}) = (40, 70, 0.02)$ for IN packets and $(min_{out}, max_{out}, P_{out}) = (10, 30, 0.2)$ for OUT packets. For more details on RIO, please refer to [9].

Two experiments were performed. In the first experiment, we assumed that there is no cross traffic. In the second experiment, we added 50 best-effort TCP flows between nodes R2 and R3 with a RTT of 10 ms each, representing the intradomain cross traffic (we can also interpret it as the interdomain cross traffic coming into domain 2 via R2 and going out to other domains via R3).

TABLE I
AGGREGATED THROUGHPUTS/GOODPUTS OF ASSURED SERVICES FOR EXPERIMENTS 1 AND 2

Flow	RTT(ms)	Target	Test1	Test2
0	20	5Mbps	5.2/5.1	3.6/3.5
1	20	1Mbps	3.0/2.8	1.0/0.9
2	40	5Mbps	4.1/3.9	2.8/2.7
3	40	1Mbps	2.0/1.9	1.0/0.9
4	50	5Mbps	4.0/4.0	3.0/3.0
5	50	1Mbps	2.1/1.9	0.8/0.8
6	70	5Mbps	3.7/3.6	2.6/2.6
7	70	1Mbps	1.6/1.5	0.8/0.8
8	100	5Mbps	3.5/3.4	2.6/2.6
9	100	1Mbps	1.2/1.2	0.7/0.6
Total		30Mbps	30.3/29.4	18.9/18.3

Table 1. summarizes the results for the two experiments. Let's first focus on the case without cross best-effort traffic. As one can see, the sessions with the lower target rate tend to achieve throughputs/goodputs higher than the target rate, whereas the sessions with the higher target rate tend to receive throughputs/goodputs lower than the target rate. This phenomenon was also observed in the previous papers [9], [2]. The related fairness issues were addressed in [2]. Here we focus on the aggregated throughput and goodput. As one can see, RIO achieves rather high aggregated service assurance with the aggregated throughput slightly higher and goodput slightly lower than the aggregated target rate. However, the situation becomes quite different in the presence of the cross best-effort traffic. One can see that the impact of the cross traffic on the end-to-end performance of the AS traffic is tremendous. Most of the AS sessions fall short of their target rates. Even worse, the achieved aggregated throughput/goodput are only about two third of the target value.

In summary, RIO can not guarantee end-to-end service assurance for the AS TCP sessions in the presence of cross best-effort traffic. One major reason behind this quality deterioration is that dropping an OUT packet and a best-effort packet have quite different impact on the performance of the two traffic types. Another major reason is that RIO is a static algorithm in the sense that the control parameters are set at fixed values, regardless of the AS traffic volume. We shall study these issues in more detail later.

III. PROPOSED SCHEME

In this section, we propose a modified DS model called Measurement-based Connection-Oriented AS (MCOAS) to solve the above mentioned issues.

A. Path Pinning

MCOAS departs from the previous approach in that it introduces connection-oriented end-to-end AS for TCP applications. With MCOAS, there is no need to assume the resource abundance in the core of a domain, nor is there a need to distinguish domain border routers from core routers. An end-to-end connection setup and resource reservation (using, e.g., RSVP) is performed for each AS session. In order to keep a stateless core, we propose a path pinning method. With this method, we pin the routed path for the life-time of an AS TCP session, such as the one proposed in [4]. It is worth mentioning that in an MPLS enabled Internet, connection-oriented DS can be easily supported without designing separate path pinning mechanisms. Since MPLS is a connection-oriented service architecture based on label swapping, service classes can be easily supported by embedding multiple CoS-based trunks in a label switched path [6], [7]. In what follows, we propose a route pinning mechanism to keep the core routers stateless.

The route pinning mechanism we proposed is similar to the IP source routing in the sense that the path information is kept in each data packet of a session, not in the routers along the path. However, the difference is that the path information is represented by the output port numbers, not the IP addresses of the routers in the path and that it is the edge node of the source, not the source itself where this information is attached to the packet of a TCP session. The idea is to map the IP address of each router or router interface to a port number with local significance. When a signaling packet reserves resources along the path all the way back to the edge node of the source, it also copies the input port number of each router, or equivalently, the output port number for data flow. The edge node then caches the port-numbers-to-connection binding information in its cache table. When the edge node receives a data packet of an AS session, it attaches the port list in the options field of the IP header for packet forwarding. A router in the path forwards a packet by first checking the IP header to see if the AS bit is set. If it is, the router immediately reads the corresponding output port number from the head of the port list and then deletes the port number from the list or moves the port number to the tail of the list before it forwards the packet to the output port. If we assume that any core router has less than 256 interfaces with any other core routers and edge routers, and the maximum number of hops for any routed path is less than 30. Then the added overhead will be less than 30 bytes which is reasonably small. Notice that this scheme requires no change to the routing table nor the forwarding table. The only change is that a core router needs to associate its interfaces to all the other routers with a locally significant sequence of port numbers.

However, there are some related issues which need to be addressed. First, since the number of hops differs from one session to another and thus the number of port numbers in the port list varies from one session to another, whether to use a variable length or fixed length IP header needs to be decided. A simple solution is to use a fixed length with 30 bytes allocated for a port list. It is very unlikely that a core router will interface with more than 256 other routers. Also an end-to-end route is unlikely to exceed 30 hops. Second, there is an issue of robustness with respect to route changes. In general, path pinning mechanism usually cause transient packet loss when routes change and the information has not propagated back to the source. This issue can be solved by allowing a core router to immediately change an AS packet to a best-effort packet when it sees the output port the packet is destined to is down. Of course, this implies that in addition to the port list, each AS packet also needs to carry the destination IP address.

B. Marking Policy

For RIO, both OUT packets of AS flows and the best-effort packets have the same dropping priority. However, dropping an OUT packet and a best-effort packet could impact an AS flow and a best-effort flow in quite different ways. For example, for the case study in Section 2, since the AS TCP sessions have much larger RTTs than the cross best-effort TCP flows, when there is a packet drop, it takes much longer time for an AS TCP session to open up its window than a best-effort flow does. As a consequence, most of the time, the best-effort traffic takes up a large portion of the output buffer no matter how much bandwidth is reserved for the AS traffic.

In order to suppress the impact of the best-effort flows, it is better to distinguish between OUT packets of an AS flow from the best-effort packets. This would allow a certain degree of decoupling of the two types of flows. In this paper, we modify RIO to allow three types of packets. Our packet marking policy can be described as follows: (a) mark the IN-profile packets of RIO as AS packet;(b) mark the OUT-profile packets of RIO as EX packet; (c) mark the best-effort packets as BE. In our approach, a packet is marked only once at the edge node closest to the sender. There is no further remarking at the domain boundaries.

C. Dropping Policy and Adaptive Dropping-Threshold Algorithm

In this subsection, we design a dropping policy to suppress the BE traffic in a proper way so that it will give its way to the AS traffic.

We use the following dropping policy which generalizes RIO:

```
If (Packet Type is AS)
    Process the packet in the same way as RIO processes an IN packet
Else if (packet Type is BE and queue length of the best-effort packets > $K_{be}$)
    Drop the BE packet
Else
    Treat EX and BE packets the same way as OUT in RIO
```

The only difference between this policy and RIO is that in this policy, a threshold K_{be} is imposed to upper bound the number of best-effort packets in the queue. To suppress the best-effort traffic without starving it, a proper design of K_{be} is crucial.

Intuitively, K_{be} should be a function of the total buffer size K, the link bandwidth B, and the aggregated reserved bandwidth B_{as} for the AS sessions at any given time t, i.e., $K_{be} = f(K, B, B_{as})$. Also, $f(K, B, B_{as})$ should be a monotonically decreasing function of B_{as} with boundary conditions $K_{be} = 0$ at $B_{as} = B$ and $K_{be} = K$ at $B_{as} = 0$. However, since the optimal buffer allocation problem that is traffic dependent is a hard one, we take a practical approach by designing an adaptive algorithm based on a given functional form. Note that the algorithm is adaptive in nature simply because B_{as} is a function of time. Also note that the approach only requires that each router keeps and updates a B_{as} value for each output port and thus high scalability is retained. With the path pinning technique proposed in Section 3.1, B_{as} can be easily updated at a router upon each arrival of either a connection setup or teardown packet.

However, there is an issue associated with the above approach. Since an AS session may not always fully use its reserved bandwidth, solely use B_{as} to update K_{be} can cause unfairness issue between the AS traffic and best effort traffic. For example, when the bottleneck link is fully reserved by the AS sessions which means $B_{as} = B$, one has $K_{be} = 0$ based on the above discuss. This means that no BE packets will be allowed to enter the buffer. If no AS sessions are using the link, the utilization of the link will be 0. To solve this problem, we use $min\{B_{as}, B_{ms}\}$ instead of B_{as} to update K_{be}, where B_{ms} is the measured aggregated bit rate for conformant AS traffic averaged in a given time window. In particular, We calculate the K_{be} as follows,

$$K_{be} = K(\frac{B - min\{B_{as}, B_{ms}\}}{B})^n, \quad \text{for } n\text{=1,2,3,...} \qquad (1)$$

Here what n value should be used will be determined by simulation analysis. One can expect that the above formulation allows the best-effort traffic to have a better chance to compete with the AS sessions for sharing the residual bandwidth left over by some other underutilized AS sessions. This formulation also explores the multiplexing gain of the AS traffic.

D. Establishment Procedure

In our approach, we treat the domain border routers the same way as the core routers and do not distinguish domain border routers from the core routers. An end-to-end connection setup and teardown for resource reservation and resource release, respectively, are required for an AS session. The intermediate routers keep and update the aggregated bandwidth reservation for AS at each output port by processing connection setup and teardown packets. Each intermediate router is also responsible for assigning the output port number to each connection setup packet. In addition to the roles an edge node plays in the traditional approach, it is also responsible for initiating the connection setup and teardown, and caching and assigning the port number list to each AS session coming into the network. In our approach, since the resource reservation is end-to-end, there is no need to do any packet remarkings at the border routers. The border routers play exactly the same role as the core routers. They run only the packet dropping algorithm proposed in the previous subsection.

IV. Simulation Results and Analysis

Two simulation experiments were performed to test the performance of MCOAS in comparison with RIO. Note that RIO here specifically refers to the RIO packet dropping algorithm, not the AS architecture proposed in [9]. A packet forwarding path is fixed using the path pinning mechanism proposed in Section 3.1.

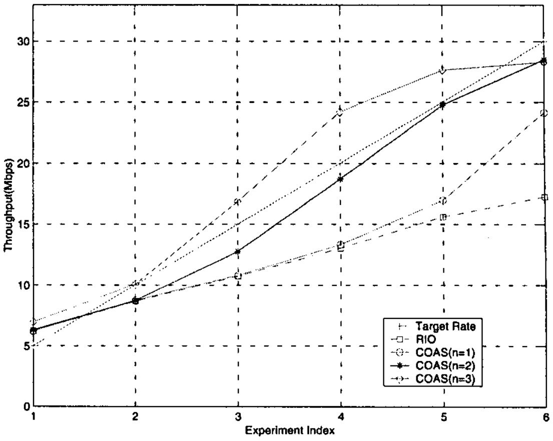

Fig. 2. Aggregated throughputs/goodputs for MCOAS and RIO

TABLE II
Aggregated Achieved Throughputs of Assured/Best-Effort Traffic for MCOAS and RIO

Target	RIO	n=1	n=2	n=3
5 Mbps	6.2/23.1	6.2/22.6	6.3/22.4	7.0/21.9
10 Mbps	8.7/20.1	8.7/20.1	8.7/20.3	10.1/19.0
15 Mbps	10.7/18.3	10.8/18.3	12.8/16.6	16.9/12.9
20 Mbps	13.0/16.2	13.3/16.0	18.7/11.1	24.2/5.8
25 Mbps	15.6/13.9	17.0/12.9	24.8/4.7	27.6/2.9
30 Mbps	17.3/12.5	24.2/6.3	28.5/1.5	28.3/1.4

The first experiment is based on the network setup in Fig. 1. We consider the case where there are 50 cross best-effort TCP flows between R1 and R2. With the other parameter settings identical to the ones in Section 2, we assume the 10 AS sessions have the same target rate and six different aggregated target rates are considered, i.e., 5, 10, 15, 20, 25, and 30 $Mbps$. The experiment is performed for MCOAS with $n = 1, 2, 3$ in (1) and also for RIO. The results for aggregated AS goodputs are shown in Fig. 2 and Table. II. Also listed after slashes in Table. II are the goodputs for the aggregated cross best-effort traffic. From Fig. 2, one can see that MCOAS outperforms RIO at all three n values, with the best performance achieved at $n = 2$. One can also see that RIO offers acceptable performance only when the aggregated target rate for AS is small. From Table II, one can clearly see that the cross best-effort traffic greatly degrades the performance of the AS traffic as its aggregated target rate increases. On the contrary, at $n = 2$, MCOAS successfully suppresses the cross traffic and keep the achieved goodput close to its target rate even with high bandwidth reservation. We observe that even at target rate $B_{as} = 30\ Mbps$, the cross best-effort traffic can still achieve 1.47 $Mbps$ goodput, which suggests that MCOAS not only provide high service assurance for the AS traffic but also a reasonable throughput performance for the best-effort traffic. Note that for all the experiments throughout the paper, a simple discrete control scheme of K_{be} is used, i.e., the largest integer which is no greater than the calculated K_{be} is used as the threshold for dropping best- effort packets. However, to avoid $K_{be} = 0$ due to the discrete control of K_{be}, one best-effort packet is allowed to be buffered with probability equal to K_{be} when the calculated K_{be} is smaller than 1.

To further examine the performance of MCOAS, we consider a network setup with one more domain in the data path as shown in Fig. 3. In this experiment, There are 10 AS TCP sessions between the hosts in domain 1 and domain 4, where domains

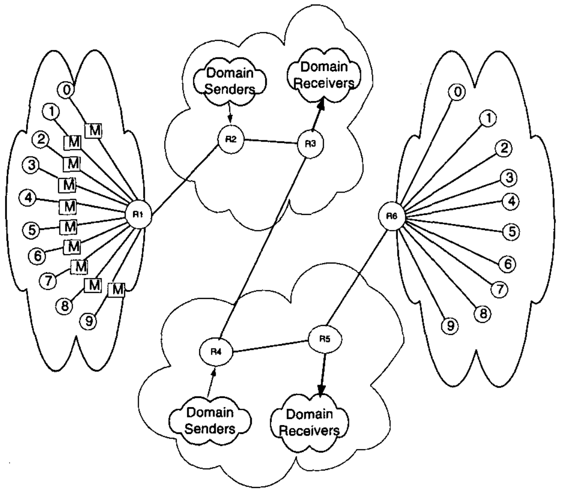

Fig. 3. Aggregated throughputs/goodputs for MCOAS and RIO

are numbered increasingly from the left to the right. Their target rates and RTTs are listed in Table III. The aggregated target rate is 30 $Mbps$. The link capacities between R1 and R2, R2 and R3, R3 and R4, and R5 and R6 are all 33 $Mbps$. The link capacity between R4 and R5 is 50 $Mbps$. There are 30 cross best-effort flows from R2 to R3 in domain 2 and 30 cross best-effort flows from R5 to R6 in domain 3. The simulation is performed for MCOAS with $n = 2$ and RIO. Both throughput and goodput are measured. This time we want to test the performance of each session and the results are listed in Table III, with a slash separating the throughput from the goodput.

TABLE III
INDIVIDUAL THROUGHPUT/GOODPUT FOR MCOAS AND RIO

Flow	RTT(ms)	Target	RIO	MCOAS
0	20	5	3.4/3.3	4.7/4.5
1	20	1	0.9/0.8	2.7/2.6
2	40	5	3.2/3.1	4.4/4.3
3	40	1	0.6/0.6	2.0/1.9
4	50	5	2.9/2.8	4.0/4.0
5	50	1	0.8/0.7	2.0/1.8
6	70	5	2.4/2.4	3.8/3.7
7	70	1	0.8/0.7	1.5/1.4
8	100	5	2.8/2.8	3.0/3.0
9	100	1	0.5/0.5	1.3/1.2
Total		30	18.1/17.6	29.2/28.4

As one can see, MCOAS outperforms RIO for all the AS sessions and again, it offers superior performance to RIO in terms of aggregated throughput guarantee. However, without isolation among AS sessions themselves, the fairness issue still exists when MCOAS is used.

To see the performance of the best-effort traffic, Table IV lists the aggregated throughput/goodput for the cross best-effort traffic in both domains. As expected, RIO fails to suppress the cross best-effort traffic and leads to a bandwidth over utilization by the best-effort traffic. On the contrary, MCOAS gracefully suppresses the best-effort traffic in both domain 2 and domain 3, offering rather high goodputs at about 1.5 $Mbps$ and 17.5 $Mbps$ in the respective domains. Hence, MCOAS can locally suppress cross traffic, resulting in a near-optimal global resource utilization. In fact, for both experiments, every bottleneck link achieves a link utilization as high as 95 %.

TABLE IV
AGGREGATED THROUGHPUTS/GOODPUTS OF BEST-EFFORT TRAFFIC FOR MCOAS AND RIO, AND THE TOTAL THROUGHPUTS/GOODPUTS FOR MCOAS IN DOMAIN 2 AND DOMAIN 3

	In Dmain 2	In Domain 3
RIO	14.40/12.38	29.25/27.46
MCOAS	1.74 /1.51	19.30/17.48
Total Rate with RIO	32.53/30.00	47.38/45.08
Total Rate with MCOAS	30.95/29.93	48.51/45.90

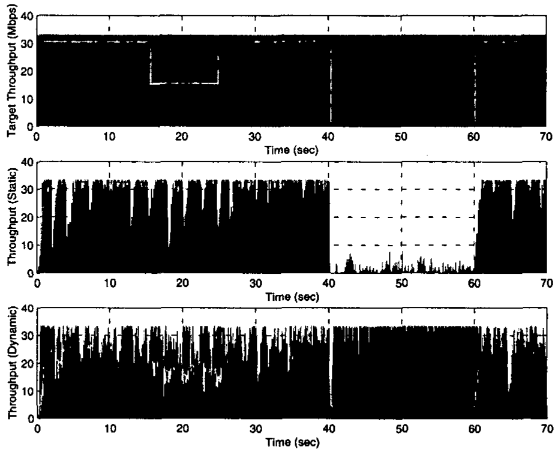

Fig. 4. MCOAS performance with on-off AS sessions

The following experiments are used to show the necessity of updating K_be based on the measured aggregated bandwdith B_{ms} for AS. First we demonstrate the deficiency of MCOAS in the presence of AS bandwidth under utilization if B_{as} instead of $min\{B_{as}, B_{ms}\}$ is used in (1). The network scenario in Fig. 1 is used for the experiment. The other network settings are the same as in experiment 1 except that the 10 AS TCP sessions are random on-off sources. The first plot in Fig. 4 shows the ideal traffic throughput composition when the aggregated conformance AS traffic level varies due to the on-off behaviors of the AS sessions (the black area).The gray area is a mixture of the best-effort traffic and the non-conformant AS traffic. The link bandwidth is fully used for the ideal case. The second plot in Fig. 4 is the actual throughput composition. We notice that the best-effort traffic does not receive a fair share of the residual bandwidth with AS sessions in the time interval between 14 and 25 seconds, where half of the AS sessions are turned off. Note that in this plot, the black area represents the aggregated throughput for AS sessions and the gray area represents the aggregated throughput for the best-effort traffic. Almost all the residual bandwidth has been taken away by the other half of the active AS sessions. Even with all the AS sessions turned off from 40 to 60 seconds, the best-effort traffic cannot take over the available bandwidth, simple because the queue threshold K_{be} is calculated based on the aggregated reserved bandwidth B_{as} for AS sessions, which is unchanged when some or all of the AS sessions are in their off-periods.

Now, we run MCOAS for the same simulation setting but use $min\{B_{as}, B_{ms}\}$ instead of B_{as} for K_{be} updating. The measurement window size for B_{ms} is fixed at 0.1 second and B_{ms} is updated every 0.1 second. The last plot in the Fig. 4 presents the simulation results. From the lower plot and with a reference to the upper plot, we observe that the best-effort traffic receives a rather fair share of the residual bandwidth with the AS sessions from 14 to 25 seconds and it takes over all the available bandwidth from 40 to 60 seconds. Note that the MCOAS performance is not quite sensitive to the selection of measurement window size for B_{ms} in the range between 0.01 seconds and a few seconds. The added complexity in MCOAS is the measurement of the conformant AS traffic level or B_{ms} at each output port and periodic updates of K_{be} at each B_{ms} update interval, in addition to the updates upon each arrival and termination of an AS session.

V. Conclusions and Future Work

In this paper, a measurement based connection-oriented AS for TCP applications was proposed. The goal is to achieve end-to-end service assurance for TCP applications, a high network resource utilization, and high scalability, simultaneously. A path pinning mechanism was proposed to facilitate scalable flow state maintenance, connection setup, and resource reservation for the AS traffic. Packet marking and dropping policies proposed in [9] were also modified to provide better service isolations. Making use of the information about the current aggregated reserved bandwidth for the AS traffic, we were able to design an simple adaptive dropping-threshold algorithm to locally suppress the best-effort traffic from overloading the AS traffic at each intermediate router. We were able to show that MCOAS can guarantee a rather high level of end-to-end service assurance for the aggregated AS TCP traffic, while still retaining a reasonably high throughput for the best-effort traffic. Complex scheduling schemes like weighted round robin (WRR), weighted fair queueing (WFQ) etc, are not required.

There are still many issues to be solved. One of the issues is that the present scheme does not provide service isolations among AS flows themselves and thus it does not solve the fairness issue. Yeom and Narasimha [11] proposed a scheme to solve this issue, however, at the expense of a much increased complexity at edge nodes. Hence, one of our future work is to address this issue in the context of the proposed scheme.

The other issue is that the proposed pinning algorithm requires special processing at each router to look into the options field in the IP header of a packet, which is not part of the standard DS architecture. However, without path pinning, it is hard to guarantee end-to-end TCP service assurance. As we mentioned in the Introduction section, MCOAS would be particularly useful when the DS is combined with MPLS because MPLS provides a standard approach to enable connection-oriented services. So, one of our future work is to study the detailed design issues for the application of MCOAS to a DS enabled MPLS network.

This paper didn't address resource management issues. A potential problems will arise if an established connection is disconnected without properly releasing its reservation. This will cause bandwidth leakage. Routers will not be able to release the reserved bandwidth automatically. A establish and release protocol is being developed and fully addressed in the next paper.

This paper is only concerned with the design of two CoSs, i.e., AS and the best-effort service, based on a single queue management. However, if more than two CoSs are to be supported, interesting issues arise as to how to allocate network resources among difference CoSs and how to do CoS-based routing. A salient feature of the DS architecture is the decoupling of traffic forwarding behavior from the service design. This feature allows service abstractions that enable future extension to incorporate CoS-based routing. The key to enable service abstractions is to logically segregate network resources, such as link bandwidth and buffer, for different CoSs, creating CoS-based virtual networks. Hence, part of our future research is to design dynamical resource allocation/sharing algorithms for different CoSs in DS. On the basis of these algorithms, the next step is to design CoS-based routing algorithms for DS.

References

[1] D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," RFC2475, Dec., 1998, and the references therein.

[2] I. Yeom and A. Reddy, "Realizing throughput guarantees in a differentiated services network," *Proc. of IWQoS*, June 1999.

[3] I. Stoica and H. Zhang, "Providing Guaranteed Services Without Per Flow Management," *ACM SIGCOMM'99*, Boston, MA, Sept 1999.

[4] I. Stoica and H. Zhang, "LIRA: A Model for Service Differentiation in the Internet," *NOSSDAV'98* .

[5] I. Stoica, H. Zhang, S. Shanker, R. Yavatkar, D. Stephens, A. Malis, Y. Bernet, Z. Wang, F. Baker, J. Wroclawski, and S. R. Wilder, "Per Hop Behaviors Based on Dynamic Packet States," Internet Draft, draft-stoica-diffserv-dps-00.txt, Feb. 1999.

[6] T. Li and Y. Rekhter, "Provider Architecture for Differentiated Services and Traffic Engineering (PASTE)," IETF Draft: draft-li-paste-01.txt, September 1998.

[7] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, and J. McManus, "Requirements for Traffic Engineering Over MPLS," IETF RFC 2702, September 1999.

[8] Y. Bernet, J. Binder, S. Blake, M. Carlson, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, and W. Weiss, "A Framework for Differentiated Services," Internet Draft <draft-ietf-diffserv-framework-01.txt>, Oct., 1998, and the references therein.

[9] D. Clark and W. Fang, "Explicit Allocation of Best-Effort Packet Delivery Service," *IEEE/ACM Transactions on Networking*, Vol. 6, No. 4, p. 362, Aug. 1998.

[10] I. Yeom and A. Reddy, "Impact of marking strategy on aggregated flows in a differentiated services network," *Proc. of IWQoS*, June 1999.

[11] I. Yeom and A. Reddy, "Marking of QoS Improvement," http://dropzone.tamu.edu/~ikjun/papers.html.

[12] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," IEEE/ACM Trans. Networking, Vol. 1, No. 4, p. 397, Aug. 1993.

[13] T. J. Ott, T. V. Lakshman, and L. H. Wong, "SRED: stabilized RED ," INFOCOM '99, Vol. 3, p. 1346, Aug. 1999.

[14] D. Lin, and R. Morris, "Dynamics of Random Early Detection ," SIGCOMM '97, 1997.