Quality of Services for Remote Control in the High Energy Physics Experiments: a Case Study

T. Ferrari, A.Ghiselli, C.Vistoli

Italian National Institute for Nuclear Physics, CNAF

Abstract

The development of new advanced applications and the evolution in networking are two related processes which greatly benefit from two-way exchanges and from progress in both fields. In this study we show how mission-oriented applications can act as effective support to research when they can take benefit of the recent enhancements in network engineering, in particular of the support of traffic differentiation in packet networks.

In this article we focus on a specific class of network-based applications: remote instrumentation control and remote display of analysis data applied to the high energy physics experiments. They require a reliable transmission channel, in particular they are one-way delay sensitive and they need guaranteed bandwidth. This can be achieved through the support of Quality of Service (QoS), i.e. through the differentiated treatment of packets on the end-to-end data path.

Several technologies and protocols for QoS support in packet networks have been devised during the last years by the research community. In this study we focus on the Differentiated Services approach, an architecture characterized by high scalability, flexibility and interoperability.

We identify the application requirements and we quantitatively define the corresponding services needed to fulfill them. The network is designed according to the differentiated services network model by defining the distribution of the diffserv functional blocks: policing, classification, marking and scheduling. For each of them the configuration best suited to remote control support is defined.

Keywords: Quality of Service, Differentiated Services, remote control, CDF

1 Introduction

Remote control is a functionality with a lot of application cases in the research field. In this paper we address the problem of remote instrumentation control applied to the high energy physics, in particular to CDF experiment: *the Collider Detector at Fermilab* [1].
In CDF remote control means two functionalities: the control of the configuration of the trigger hardware from a remote size and the monitoring of the analysis through the remote display of results in multiple client sites. The *Rpc and Object Broker INterface* (ROBIN) [2] is the software platform which provides support for the former task, while ROOT, the *Object Oriente Data Analysis Framework* [3] is the tool deployed for the latter activity.

2 QoS: the differentiated services architecture

Classification consists in the differentiation of traffic classes through the association of an identifier to a set of one or more data flows. In the differentiated services architecture [4] two packet classification approaches can be deployed.

With *Multifield Classification* packets are identified through the content of a combination of IP header fields like the source or destination IP address, the port number and the protocol type.

On the other hand, *DSCP-based Classification* distinguishes packets according to the Diffserentiated Services Code Point in the IP header. The DSCP can be set by the application and/or by the edge router of the first diffserv domain encountered on the data path to the destination. The edge router should perform admission control to guarantee that non-zero DSCP values are correctly deployed, i.e. that only applications which are entitled to do so actually mark packets by placing the right DSCP [5] into the IP header.

Marking consists in the placing of a code point into the IP header of a packet for the identification of the class that packet belongs to. Marking is normally performed at the edge of a diffserv domain: In this way only edge routers need to keep the information for the mapping of flows or microflows into diffserv classes and to perform policy control. For scalability sake core diffserv nodes trust the marking performed at the edge: Flows and/or microflows are replaced by the *Behavior Aggregate* (BA) concept and core nodes treat packets on an aggregated basis by only taking into consideration the class they belong to. Re-marking can occur at the boundary between two distinct diffserv regions when a given packet treatment is identified differently in contiguous domains.

3 Application characterization

Whatever technology is deployed for QoS support, the application needs to be characterizd to identify the most suitable service, which can be defined either qualitatively or quantitatively. The sensitiveness to delay and/or jitter, the tolerance to packet loss, the traffic volume issued by the application, the burstiness degree of the sources, the presence of potential traffic collision points in the networks, are some of the factors which need to be taken into consideration. Data packets issued by our application need to be identified and classified.

3.1 Hardware remote control

This task is performed through the *bidirectional* exchange of IP datagrams between server an client. A negligible traffic volume is produced. Hardware control is scarcely interactive and only small information items (commands) are set out over the network. The frame payload size is less than 100 bytes and burstiness is negligible.

IP frames are transported through the TCP protocol and both clients and servers are identified by a known IP address. In addition, the server port number is fixed: 50038, while the client port number can vary. As a consequence, packet marking can be easily implemented through multifield classification based on the content of the source/destination address, on the protocol type TCP and on the source/destination port number.

Traffic is delay bound since the effect of commands on the trigger hardware should be as immediate as possible. In addition, packet loss under TCP implies the retransmission of datagrams, as a consequence this application requires high reliability during transmission.

3.2 Monitoring of the analysis

Also in this case data exchange is bidirectional, but the traffic volume is asymmetric as most of the data is transmitted in the server-to-client direction. Data is exchanged so that limited amounts of data (ROOT objects) are exhanged for the display of analysis results in remote sites and burstiness is negligible. The server and remote clients are identified by well-known IP addresses and traffic deploys the TCP transport protocol.

Given the off-line nature of monitoring, this application is packet loss tolerant, however since it is also interactive, in case of congestion a minimum amount of bandwidth should be guaranteed to protect. However, monitoring is not one-way of jitter sensitive.

The overall traffic sent out by the server should be provided with a maximum bandwidth guarantee equal to threshold max, while exceeding traffic could be either treated as best-effort or dropped. In addition, traffic to a given client should be protected from mis-behaving users by guaranteeing a minimum per-user bandwidth bw_i. However, each client should be given the possibility to deploy up to the maximum overall capacity max (when not allocated to other monitoring agents).

4 Service level specifications

According to the previous application characterizations, sections 4.1 and 4.2 instroduce service level specifications (SLS) for both hardware remote control and monitoring.

4.1 Service 1: remote hardware control

- Bandwidth guarantee: only traffic up to an upper threshold bw, e.g. 512 Kbps, is marked as high-priority, while excess traffic is transmitted as best-effort[1]. The value of parameter bw can be estimated according to formula: $bw = x * 64Kbps$, where x is the number of servers which can be accessed through router $R1$ or of clients which a given edge router $R2$ connects to the diffserv domain.
- Burst tolerance: any packet belonging to a burst whose size is smaller than the upper threshold $B = 64Kbytes$ is marked with a high-priority label, otherwise the packet is treated as best-effort [2].
- Delay bound: one-way delay is upper-bounded. One-way delay corresponds to parameter $Type - P - One - way - Delay$ as defined in RFC 2679 [6]: *`Type-P-One-way-Delay` from Src to Dst at T is dT means that Src sent the first bit of a Type-P packet to Dst at wire-time T and that Dst received the last bit of that packet at wire-time T+dT.*
  The upper delay bound D expressed in msec can be estimated through the formula: $D = \frac{1}{2}RTT + x * 10\,msec$, where RTT is the round trip time of a packet of 100 bytes, x the number of routers on the data path and 10 msec is an approximation of the maximum nodal delay introduced by a diffserv router deploying priority queuing as service policy when the average best-effort packet size is of 1028 bytes[3].

The service specification defined above can be deployed in both directions, i.e. from server to client and vice versa. The value of paremeters defining tolerances, like the bandwidth guarantee, burst tolerance and the delay bound can be tuned according to the traffic volume in each direction.

4.2 Service 2: monitoring of the analysis

While in the previous case a unique service can satisfy the application requirements in both traffic directions, for this application two different services have to be defined given the asymmetry of the two data streams.

[1] 512 Kbps is a reference value, a different and more appropriate bound can be chosen depending on the number of local clients connected to the edge router.

[2] The maximum buffer size B can be tuned as needed. The optimum value can depend on the instantanous traffic volume, i.e. on the number of local servers or clients.

[3] By picking 1028 bytes we get a worst-case estimation of the nodal delay, since in production networks the average datagram size is in the range [300, 400] bytes.

4.2.1 Client → Server SLS

- Bandwidth: each client is guaranteed with a minimum rate of 64 Kbps to the server [4], however it is allowed to generate a traffic volume up to 256 Kbps (or an equivalent value higher than 64 kbps) in case of resource availability.
  Packets for which the instantaneous traffic rate exceeds the upper rate threshold are dropped [5]. Packets exceeding the lower threshold (64 Kbps) are dropped first.
- Burst tolerance: 16 Kbytes (or an equivalent value defined through tuning).

4.2.2 Server → Client SLS

- Bandwidth: for each client the server can deploy 64 Kbps of guaranteed bandwidth. This means that the overall amount of guaranteed bandwidth is equal to $64Kbps * m$ where m is the number of clients (if for each client the amount of guaranteed bandwidth $bw(i)$ is the same). 64 Kbps is a reference value which can be appropriately tuned.
  If plenty of bandwidth is available, the overall rate of traffic generated by the server can be up to 5 Mbps, this means that if some clients are not active at a given time, a given client can deploy from 64 Kbps to 5 Mbps. Traffic exceeding 5 Mbps is dropped: in case of congestion packets exceeding the minimum per-client guaranteed bandwidth (64 Kbps) are dropped first.
- Burst tolerance: traffic bursts up to 128 Kbytes are tolerated. This reference value can be tuned and modified appropriately, if needed.

5 Conclusions

Remote control applied to the high energy physics is a representative example of mission-critical research applications requiring the support of new and enhanced types of data transmission in order to be a reliable tool for researchers. In this paper we propose a quality of service architecture for the implementation of Quality of Service in packet networks by detailing first the application requirements and the corresponding service and by engineering the differentiated services network model in terms of placement and configuration of functional blocks like packet classification, marking, policing and scheduling. The network design here prented requires testing in a network testbed for the achievement of the ultimate goal: the deployment of advanced software tools in a production QoS-capable networking infrastructure.

References

1 *The the Collider Detector at Fermilab:* `http://www-cdf.fnal.gov/`
2 *The Rpc and Object Broker INterface* `http://www-b0.fnal.gov:8000/ROBIN/`
3 *An Object Oriented Data Analysis Framework*, `http://root.cern.ch/`
4 *RFC 2475: An Architecture for Differentiated Services*; S.Blake et al., Dec 1998.
5 *RFC 2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*; K. Nichols et al., Dec 1998.

[4] 64 Kbps is a reference value which can be tuned appropriately.

[5] Alternatively excess packets could be treated with a higher dropped priority or they could be provided with a best-effort service. In the second approach the drowback of the is that in this way packets belonging to the same microflow but marked as best-effort are placed into a different queue. As a consequence, in this way the scheduler changes the packet order within each microflow. However, reordering should always be avoided to prevent packet retransmissions and additional overhead on the receiving hand.

6 *RFC 2679: A One-way Delay Metric for IPPM*; G. Almes, S. Kalidindi, M. Zekauskas, Sep 1999.
7 *RFC 2598: An Expedited Forwarding PHB*; V. Jacobson et al., June 1999.
8 *RFC 2597: Assured Forwarding PHB Group*; J. Heinanen et al., June 1999.
9 *RFC 2698: A Two Rate Three Color Marker*; J.Heinanen, R.Guerin, Sep 1999.