MANAGING END-TO-END QOS

Chang Qian Song Junde Song Mei
Beijing University of Posts and Telecommunications
qian.chang@263.net

Abstract

Efficient network management can provide improved flexibility, scalability and robust by distributing monitoring, analysis and control functionality throughout the networks. End-to-end (wired and wireless) QoS management faces big new challenges. The paper discusses both the online monitoring and the offline inter-network analysis. Monitoring methods in the core networks and the wireless access networks are analyzed respectively. From the analysis above, we infer that the different services and networks need to be measured by using different methods. The effective architecture to implement the end-to-end QoS management is proposed. The components of the integrated inter-network analysis process are discussed in order to provide end-to-end Quality of Service (QoS) guarantee. Simulation results show that these methods have good performance.

Keywords: *Probe; call detail record; inter-network analysis; time correlation; network model based reasoning*

1. Introduction

The paper presents how to manage and measure end-to-end QoS. Currently, the research focus is put on the performance measurement and monitoring of Internet. The total solution set of the end-to-end measurement (for both wired and wireless networks) is studied rarely and it will be the research hotspot.

To answer this question, two research needs must be addressed:

First, research of integrated wired/wireless network measurement is required. Most related research is focused on measuring the wired and wireless networks in isolation; the interactions between the two are not well understood.

Second, measurement metrics that can reflect the performance of the wired/wireless networks, as well as the correlation between those metrics, must be researched.

Probes are used to diagnose faults and predict performance in the IP core networks. They can work in both static way and dynamic way. It is robust and flexible enough to satisfy the requirements of real-time processing and prediction with high accuracy. The working way and the classification of the probes are analyzed in order to reduce the overall measurement throughput and increase the accuracy of the monitoring when the probes are applied.

As to the wireless access networks, probe is not often be used because of the terminal's mobility. Call Detail Records (CDRs), which can provide the statistic performance of the networks, are trend to be used. However, the CDRs are not enough for the QoS guarantee because of its limitation that make the real time analysis difficult, although in real time they are used to send QoS alarm events to fault manager. A new scheme that injects movable test modules in the system is proposed to evaluate the performance in a specific scope. The mobile probe in the mobile terminal is mentioned in the paper.

On the base of analyzing the measurement methodology, the effective methods used to support the inter-network analysis and management are discussed in detail.

The following sections are organized as follows. Section 2 discusses the measurement methodology, including probing and CDR mechanism, as well as the measurement metrics, etc. Section 3 describes architecture of the proposed methods. Section 4 is a simulation result that is to illustrate how to implement the end-to-end QoS management. Finally, Section 4 gives the conclusion of the paper.

2. Measurement Methodology

2.1 Probing Technology

Probe is a very popular technology in the wired network measurement. In the section, various probing techniques are analyzed. The probing techniques differ in the packets constituting a probe, the size, and the path

Proceedings of the 2002 IEEE Canadian Conference
on Electrical & Computer Engineering
0-7803-7514-9/02/$17.00 © 2002 IEEE

traversed by each probe packet, etc. They also differ in the host collecting the probing response and the function used by this host to perform the required estimation. All these will be analyzed in detail in the following sub sections.

2.2.1. Multicast probe and unicast probe. Multicast probe and unicast probe is different depending on the model used to transmit probe traffic. Multicast refers to one-to-many packet transmission. The constituent packets of the multi-destination (multicast) probe do not all target the same destination IP address. Otherwise, the packets of the unicast probe have only one destination.

The notional multicast packets many be of the same end-to-end inferences that can be made by multi-destination unicast probes.

2.2.2. Passive and active network measurement. Active measurement tools inject test packets into the network and observe their behavior. For example, the simple ping tool measure round-trip-time (RTT) of IMCP probe packets.

Active measurement tools can be used to measure bulk throughput, path characteristics, packet delay, packet loss, and bandwidth characteristics between hosts. For example, the probe injects UDP packets into the network according to Poisson process in [1]. The motivation of this method is that sampling at intervals determined by a Poisson process will result in observations that match the time-averaged state of the system.

In contract, passive measurements observe actual traffic without perturbing the network, which is its advantage. Passive techniques typically consist of gathering packet level data at some tap point (either from routers and switches, or from stand-alone traffic meters) in the network. A great deal of information can be extracted from passive monitoring, even at only a single point. Passive monitors must process the full load on the link, which can be problematic on high-speed links. This is its disadvantage.

2.2.3. Sender-based probe and receiver-based probe. They are different depending on where inferences are made.

As to the sender-based probe, the probes and their replies are recorded only at the location of the probe sender. Sender-based measurement has the enormous advantage of not requiring access to the remote site in order to instrument the probe arrivals. On the other hand, sender-based measurement has its limitation that from it one can say little about how traffic behaves along the path's two different directions. For example [2], suppose a measurement consists of sending a flight of 10 "ping" packets from A to B, and timing at A the arrival of their echoes. If only 6 echoes return, we have no way of knowing whether B never sent the 4 others, because their corresponding ping never arrived at B; or if B did send them, but they were lost on their journey from B back to A; or if some combination of loss from A to B and loss from B to A occurred. Consequently, it is difficult to say anything concrete about the nature of the loss event.

This consideration becomes more subtle, but equally important, when applied to analyzing packet delay. A sender-based scheme can only observe round-trip time (RTT) delays. These are perforce the sum of the one-way transit time (OTT) delays in the two directions, plus the (unobserved) delay of the receiver generating its reply. If the goal of the timing measurement is to estimate capacity along the forward path, then any delay variations incurred on the return path are pure noise, and at best dilute the precision with which the sender can estimate the path capacity.

All the disadvantages of the sender-based probe are the advantages of the receiver-based probe.

End-to-end measurement is often done using either "sender-based only" or "receiver-based only" measurement. If we can trace the transfers at both the sender and the receiver, we can get the measurement about the forward path, the reverse path, and the processing delays at both the sender and the receiver. As a result, the mechanisms of coordinating measurement between sender and receiver are considered currently.

2.2.4. Monitor placement within the network. A clear understanding of network topology, particularly link and router location, is a prerequisite to monitor placement. Upon discovering the network topology, one can identify IP net-blocks (i.e. ranges of IP address) that appear on each link, and can target traffic flows of interest. While it is tempting to measure the traffic between every pair of sites, the cost does not scale with the benefit. Instead, one might identify which links carry the most traffic, and locate monitors there. Alternatively, one could begin by monitoring traffic at the border routers of one's infrastructure.

2.2 Call Detail Record

Unlike the wired network, the mobile terminal in the wireless network has the mobility that makes it difficult to be measured by using probes. Currently, the call detail record could be used to measure the performance of the wireless network.

Call detail record is used to provide rating and billing information. Billing plays a major role in communication services provided over a network. However, the rating and billing information be can also used for fraud control if available in real-time.

Like passive monitors mentioned above, CDR must

process the full load on the link, which makes its real time analysis difficult.

Fig. 1 shows a model that can be used both on-line and off-line.

The inference engine is the heart of the architecture. The engine processes the incoming stream of call detail records. It calculate rates and discounts and call detail record according to customized rating and discounting functions that depend on the list of features and services subscribed to by the customer, computes summary fields associated with the customer, generates and compresses a processed call detail record that includes the rating and discounting information, and stores an in-memory image of summary fields and customer information essential for rating, discounting, and real-time queries.

Call detail database: The Call detail database stores processed call detail records. The processed call detail records contain rating and discounting information, and any information relevant information.

By filtering the large amount of data by some rules, the inference engine can filter the important information and stores them in an in-memory image. The in-memory has a real-time interface for real-time queries. As a result, he architecture allows for a connection between the billing data stream and network management data, quick creation of new features, and automatic detection of interactions between features. The management system can either take on-line query through the real time interface or analyze the large amount of data in the call detail database.

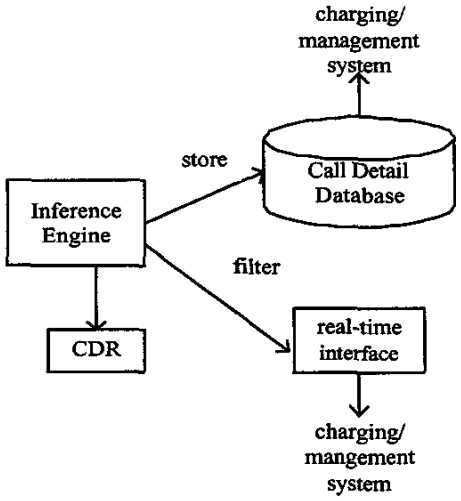

Fig. 1. CDR Processing Architecture

2.3 Metric Selection

The most commonly used network metrics are latency, packet loss percentage, link utilization, and availability. Although these metrics are commonly used, they are not always clearly defined.

As to the IP field, [3] has developed a framework for performance metrics, and is producing standards for metrics such as connectivity, one-way delay, one-way packet loss, round-trip delay, delay variation, bulk transfer capacity, etc. These newer metrics will become more important as providers move to implement different QoS in their networks. All the above metrics measure the behavior of packets on a link; they only provider an indirect view of a network's performance as experienced by its users. Attempts to measure used-perceived network performance [4] require research to determine valid metrics.

As to the wireless field, QoS classification and related metrics are defined in [5].

2.4 Data collection and archiving

Traffic analysis requires collection of monitor data into one or more archive locations. One common approach involves building a trace file repository enabling users to request a report on specific sites, metrics, or time periods. Alternately, a meter can filter and process data in real-time to reduce data storage requirements.

Once a measurement data repository is in place, it is important to provide a clear, easy-to-use web interface to its data. There is no point in collecting data if users can't access it easily so as to make effective use of it.

3. Proposed Solution

3.1 Architecture

Since the service-oriented networks have several important differences from traditional networks, both the high quality TMN resource management and the best-effort IP network management are inadequate. The new requirements for the end-to-end QoS management emerged to enable efficient service management.

First, provision of new services will have to use a broader set of resources including computation, storage, and services in different networks (wired and wireless). The management information has to flow across management domain boundaries to provide end-to-end view. As a result, presenting a unified encompassing view on networks, systems, services and applications is required to provide integrated services.

Secondly, defining standard mechanisms to share selective measurement information among the various networks is a key item to enable interactions between different management domains for provision of

service-oriented management.

Finally, efficient mechanisms are needed to ensure that service level agreement (SLA) provided by the service provider to its customers. The value-added services are likely to have their own specific notions of QoS. Since the resources allocated to the service providers will often be not reserved for exclusive use, new technologies are evoked for trade-offs between all the network resources.

To support these new requirements, we propose a feasible architecture for end-to-end QoS management. The architecture is showed in Fig. 2.

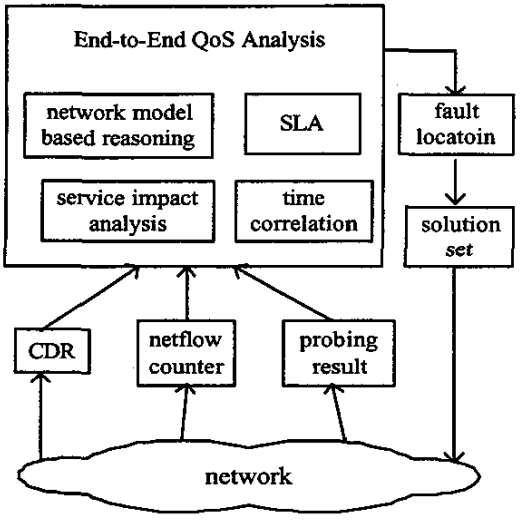

Fig. 2. End-to-End QoS Management Model

CDR, probing result and the net flow counter are the measurement data collected from the network. They include the user profile, network statistic data and the real-time data. CDR is statistic measurement for both wireless and wired networks, as well as on-line analysis and off-line analysis (See section 2.2). Probing is mostly used in wired networks (See section 2.1). However, we can also add the probe to the mobile terminal. In the case, the real-time measurement of the wireless network will be got. The net flow counters are obtained from the network by using several network management interface protocols, such as SNMP/CORBA for IP networks and CMIP/CORBA for telecommunication networks. At this interface, a protocol independent model – Integrated Reference Point (IRP) [5] is recommended.

The module End-to-End QoS Analysis (EQA) module is the heart of the architecture. This module has four sub modules.

SLA module stores the information related to SLA and SLS (service level specification). The SLA is between the users and the service providers to specify which level of QoS should be guaranteed. This module can be dynamic and changes locally to reflect the changes in the client or system state.

Network model based reasoning module is to analyze the data to infer the network performance with the understanding of the network characteristics.

Service impact analysis is to find how the abnormality of a network element affects other network elements and services. By doing so, a potential fault can be detected advanced.

Through time correlation, we can examine the time-dependent properties of the measurement from the different networks and determine the fault location. With some intelligent engine, the related solution sets are determined. It consequently can be applied to the network to improve the service performance.

4. Simulation

To illustrate how to manage end-to-end QoS, we use the Openet Modeler to build a simple network consisting of both wireless network and wired networks (See Fig. 3).

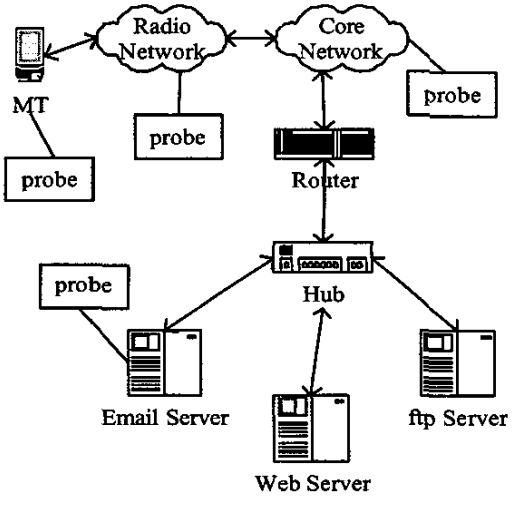

Fig. 3. Network Architecture for Test

In order to simplify the problem, we use probes as the main measurement. We put probes in the networks to complement the measurement task. As shown in Fig. 3, probes can also be put in the mobile terminal to measurement the net flow in the wireless environments.

The probes are located in the radio network and core network to measure the performance of core network (CN, See Fig. 4) and radio network control (RNC) in wireless network (See Fig. 5). Fig. 4 shows packet granted/released/queued in the Core Network. The network bottleneck can be estimated approximately. Fig. 5 shows total transmit/received throughput of RNC,

showing the processing capability of the RNC.

The Fig. 6 shows the number of download response time for three different types of flows. To summarize, Email streams are mostly small and short lived, but there are a few large, long-lived streams (t=18ms). FTP streams indicates its relatively stability. Web stream sizes are continuously changed largely.

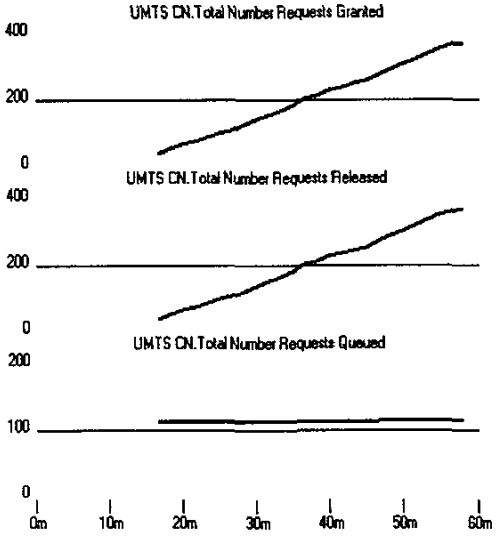

Fig. 4.Core Network Packet Granted/Released/Queued

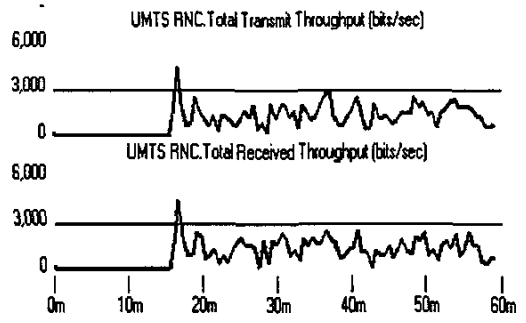

Fig. 5. Total Transmit/Received Throughput of RNC

5. Conclusion

The introduction of 3rd generation networks will quickly enable the Service and Content Providers to offer new and more complex mobile services to the end user. End-to-end QoS management is required but more challenging. Through analyzing the features of the different measurement method and the characteristics of different networks, we infer that the different services and networks need to be measured by using different methods.

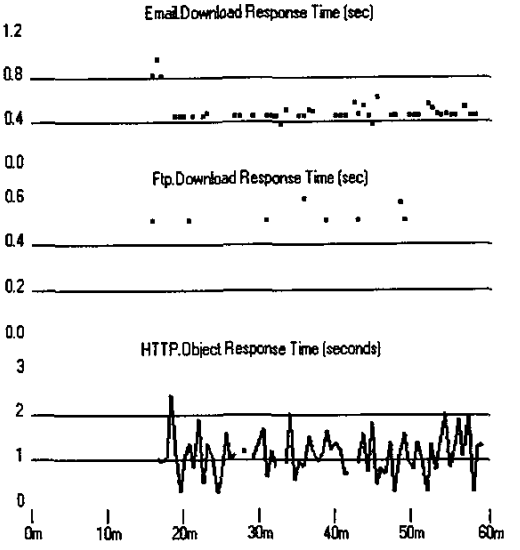

Fig. 6. Download Response Time

On the base of considering the various methods deeply, an effective architecture is proposed. In the architecture, different measurement techniques are applied on demand. The kernel analysis module is able to process and analyze all the data collected from the network, thus the end-to-end performance is inferred. Finally, a simulation is done to show how to perform a simplified end-to-end QoS management.

References

[1] Dan Rubenstein, Jim Kurose, and Don Towsley, " Detecting Shared Congestion of Flows Via End-to-end Measurement," Technical Report 99-66, Department of Computer Science, pp. 1-28, November, 1999.

[2] Vern Paxson, "Measurements and Analysis of End-to-End Internet Dynamics," Ph.D. Thesis, Computer Science Division University of California, Berkeley, April, 1997.

[3] IP Performance Metrics website, http://www.ietf.org/html.charters/ippm-charter.html

[4] Performance Measures for Multimedia Applications, Hughes, W. R., Proceedings 38th IETF meeting, Memphis, April 1997, http://www.ietf.org/proceedings/97apr/ops/rtfm-2/index.htm

[5] 3GPP TS 23.107 QoS Concept and Architecture, 1999

[6] 3GPP TS 32.102. Telecom Management Architecture, 1999.