INCREASING *the* OBSERVABILITY *of* INTERNET BEHAVIOR

Improvements are needed in network infrastructure and protocols for continued growth of the Internet.

THOMAS M. CHEN

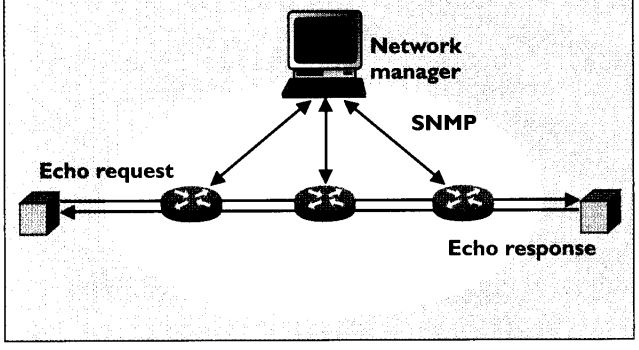
Historically, the Internet has been difficult to monitor and manage due to its heterogeneity, geographic size, and distributed administration. In addition, the Internet protocol was purposely designed with lowest common denominator requirements from each network, including minimal protocol functions, to facilitate performance measurements. For example, IP does not include any timestamps to measure packet delays or packet sequence numbers to detect packet loss. As another example, the widely used *traceroute* utility is not instrumented within IP but exploits the Internet control message protocol (ICMP) "time exceeded" message in an unintended way. Considering the unbounded success of the Internet, it might be natural to expect these shortcomings have been resolved by now, but actually the means for measurements are still quite limited. Most organizations have been more occupied with deploying and maintaining their networks than developing advanced mechanisms for performance monitoring. Also, the Internet is not an ordinary network; it is actually a vast collection of interconnected networks that are owned and operated by separate, competitive organizations. Consequently, most networks do not fully cooperate with externally initiated measurements of performance. The ICMP and IP header options, which would be useful for certain types of measurements, are not universally supported.

While the limited observability has been adequate for best-effort data service, the Internet is evolving toward a more advanced services architecture, such as integrated and differentiated services, with higher expectations on network performance and quality of

service (QOS). Improved observability of the network is a prerequisite to verifying and optimizing network performance for demanding applications. Also, better means for monitoring are needed to overcome the increasing complexity of the Internet caused by its tremendous growth in size, traffic diversity, transmission speeds, and traffic volume.

The traditional cornerstone for Internet monitoring is the venerable *ping*, a query-response tool that sends an ICMP "echo request" to a designated host, which returns an ICMP "echo response" [9]. Although ping is mainly for verifying connectivity to another host, it can be used in more general ways.

Figure 1. Ping and SNMP.



Repeated pings can be useful for "black box" testing of the network (observing the delivery of packets but not the routers traversed by a packet). Packet delays and packet loss seen by pings are reflections of the general network performance.

Network management serves the complementary function of monitoring the status of individual nodes in the network, as shown in Figure 1. In the ubiquitous simple network management protocol (SNMP) paradigm, a centralized network manager can poll nodes for their status data defined by variables contained in the management information base (MIB) [3]. The network manager can also be notified of predefined events, if detected, through SNMP trap messages. Since SNMP was originally intended as a simple interim approach, it has limited capabilities for performance and traffic monitoring. For performance, the network manager can collect rudimentary local performance measurements from each router. Although routers are the ideal points for traffic measurement, they are generally not equipped for exhaustive traffic monitoring. The primary function of routers is regarded as packet forwarding, and management functions are considered to be secondary. The SNMP paradigm allows for collection of aggregate traffic statistics, such as the number of packets per interface over a time interval, but not sufficiently detailed measurements for accurate traffic characterization. Traffic analysts usually depend on special-purpose traffic measurement tools such as protocol analyzers.

Current limitations motivate a need to explore new methods and possibly network infrastructure changes, which is being addressed by several research projects such as the Cooperative Association for Internet Data Analysis (CAIDA) [4]. Generally, it might be useful to consider current and new methods according to various characteristics. First, methods can be passive, introducing no additional traffic into the network, or active. Traffic monitoring typically consists of passive observation of link packets, whereas performance measurements commonly involve the exchange of test traffic (pings). Performance measurements can also be passive, such as counting packets dropped at a router. For active methods, the amount of additional traffic and its effect on network behavior are always issues. Active methods are often intrusive, possibly affecting active services, whereas passive methods are nonintrusive. Several additional characteristics may be interrelated:

- In-service methods are bound to a specific flow of data packets, while out-of-service methods are not;
- In-service methods can be further classified as in-band (additional fields in the packet header) or out-of-band (measurement packets are separate from data packets);
- Measurements may be performed continuously or on-demand;
- Measurements may be direct or indirect, for example, ping measures roundtrip delay directly but general packet loss must be inferred indirectly from the loss seen by repeated pings;
- Measurements may be one-way or bidirectional;
- One-point methods involve measurements at a single point, two-point methods require two reference points (packet delay), or N-point methods require multiple points (multicast).

The design of methods must consider trade-offs between these various choices. For example, on-demand methods may be preferable to continuous measurements to save bandwidth. For another example, timing information may be carried in every packet header (in-band) or separate packets (out-of-band) to measure packet delays. Timestamps in the packet header would allow accurate delay measurements for each packet, at the cost of significant processing for every packet. In comparison, timestamps in separate packets incur processing only

as needed but measurements cannot be made for specific packets.

IP-Layer Performance Measurements

Wide interest in measuring Internet performance is evidenced by increasing activities in standardization [5, 7] and research [4, 6, 8]. Users depending more on their network services are interested in performance guarantees (for example, service-level agreements) and means to verify and compare service quality. Increased competition is motivating service providers to give more detailed performance guarantees and reports (examples include continuously updated weather maps and reports on the Web). Moreover, service providers need methods to measure performance not only in their own networks but also other service providers' networks, because Internet traffic often traverses more than one domain. Detailed measurements are helpful to fine-tune resource management within a network, and diagnose the location and cause of service problems across domains.

Network performance is generally considered in terms of speed, accuracy, dependability and availability aspects of point-to-point IP packet delivery. Speed parameters are usually the maximum, average, and variation of packet delay. End-to-end packet delays are the accumulation of transmission, propagation, queueing, and processing delays at each router. Maximum packet delay is obviously important for real-time applications, while delay variation can affect buffer underflow/overflow for streaming applications. Different metrics for delay variation are possible, such as the max-min range, upper quantile, or deviation from an expected reference packet pattern. Accuracy is measured in errored packets, that is, delivered packets with bit errors in the header or payload (the header checksum can detect most errors in the header, resulting in a discarded packet). Accuracy is usually not a parameter of great concern because bit errors in the packet payload are assumed to be correctable by the higher-layer protocols if data integrity is important to the application.

Dependability refers to the packet loss ratio or fraction of packets not delivered. IP packets may be lost due to various reasons: router buffer overflows, bit errors in the packet header, expiration of the time-to-live (TTL) field, unrecognized or unreachable destination address, invalid header fields, or inability to fragment if needed. In the current best-effort IP service, the packet loss ratio can be unbounded but is obviously important to certain real-time applications that cannot retransmit lost data. For applications that can recover lost data through retransmissions, a high loss ratio could result in inefficient multiple retransmissions. If the packet loss ratio is very high, the destination will be considered unavailable.

These performance metrics are generally statistical definitions in reference to a long or infinite time horizon. Also, they implicitly assume no fragmentation but need more careful definition in the case of fragmentation (for example, packet delay when the original packet is divided into multiple independent IP datagrams can mean the maximum delay of any fragment).

As mentioned previously, the traditional means for IP-layer performance measurements is the ping tool (and its many variations), which is a pair of ICMP echo request and echo response messages to verify connectivity to a host. The ICMP was designed mainly to report trouble conditions from routers (for example, destination-unreachable causes) instead of a means to carry out performance measurements. Although ping is mainly for verifying connectivity to another host, it can be used in more general ways. Repeated pings are an active, on-demand method to collect an empirical distribution function of roundtrip delay. Packet delays and delay jitter are directly proportional to congestion, so repeated pings can provide an indication of congestion level. Also, since pings are themselves IP packets, they can provide a sample of the general packet loss ratio.

Usually, pings are not repeated frequently enough to significantly affect network performance. However, the accuracy of delay measurements is a more signficant issue. First, if the network behavior is periodic, periodically repeated pings may not observe the behavior accurately. Second, periodic sampling may affect network behavior by some unpredictable synchronization effect. Poisson sampling (with exponentially distributed random times between samples) as a way of random sampling has been recommended [7]. Third, one-way delay is more important for certain applications than roundtrip delay (and one-way delay is a more typical QOS parameter). One-way delay might be measured if the ICMP echo request message contains a timestamp, but the accuracy of this method has two requirements: clocks (time-of-day) at the source and destination hosts must be synchronized (today, synchronization within 100μsec is affordable by means of global positioning systems); and hosts must be able to write timestamps immediately prior to packet transmission and read timestamps immediately after packet receipt (possibly with special hardware).

Ping is an example of black box testing in which the network is used only for packet delivery. This has

the advantage of simple implementation (no special packet processing within the network), but no information is collected about intermediate routers. Observability of intermediate routers is enabled by another widely used program, traceroute, which exploits the ICMP time exceeded message normally generated to signal a packet loss due to expiration of the TTL field. Traceroute deliberately sends a sequence of IP packets with limited TTL fields to discover routers within a proximity. Because IP is connectionless, however, traceroute does not guarantee the discovered route was the actual route followed by a packet.

For a connectionless protocol, the only way to make performance measurements associated with a specific packet's route is through the packet header. Actually, IP allows a packet's route to be traced using the IP header record route option. This option follows the IP (version 4) mandatory header, leaving empty space for several IP addresses to be filled by routers visited by the packet. Similarly, the IP header timestamp option leaves empty space after the mandatory IP header for routers to fill in their IP addresses and timestamps. In fact, since ICMP messages are carried within IP datagrams, ping can also use the record route or timestamp options to obtain route information as well as test connectivity. Unfortunately, these options have limitations in the current implementation. First, the space is currently limited to nine addresses or timestamps. The 4-bit header length field in the IP header imposes a maximum header of 60 bytes including a 20-byte mandatory part, leaving 40 bytes for header options. In this space, 3 bytes is taken to specify the option and each IP address or timestamp requires 4 bytes, which leaves space for only 9 addresses/timestamps. Various solutions are possible, for example, since a 20-byte header is mandatory, the header length field can be reinterpreted as the length of the options. This would increase the space to 60 bytes for options, or 14 addresses/timestamps. The header length field can be lengthened to allow more addresses or timestamps, but will always be limited to a certain number. A second issue is that IP header options are not universally supported by all routers. If these two issues could be resolved, header options might become a powerful mechanism for on-demand performance monitoring. For example, new options could be defined to collect all types of performance data directly for any packet, for example, utilization per hop, queueing delay per hop, packets lost since the last header option, and so forth.

An open question is the relation between IP-layer performance metrics and the performance observed at higher protocol layers. For example, the relation between IP performance and TCP behavior is not straightforward. TCP throughput is extremely dynamic and sensitive to both packet loss and delay. TCP tracks samples of roundtrip delay and adapts its retransmission timer. Its complex congestion avoidance algorithm is very sensitive to packet loss and retransmission time-outs, which cause the sender to back off into slow start. Hence TCP throughput depends not only on the packet loss ratio but also on the loss pattern. More research is needed to relate IP-layer performance and application-layer performance. Alternatively, more tools are needed to measure performance at the transport layer (for example, TReno measures TCP bandwidth) and application layer (for example, Keynote measures Web server response times).

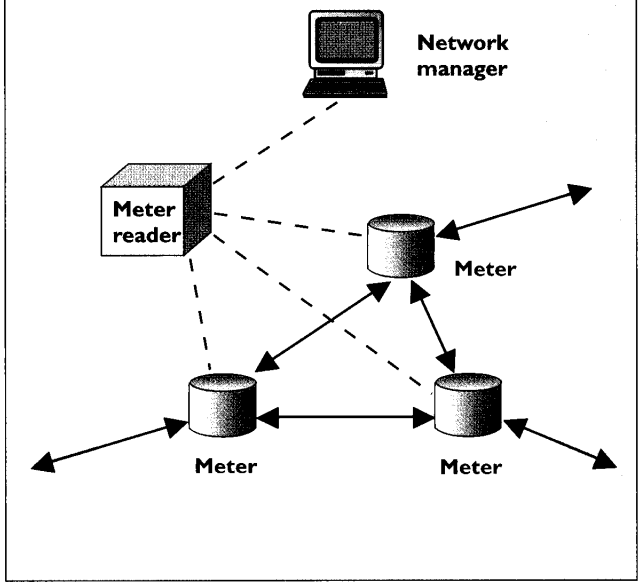
ALTHOUGH ROUTERS ARE IDEAL POINTS FOR TRAFFIC MEASUREMENTS, THEY ARE GENERALLY NOT EQUIPPED TODAY FOR EXHAUSTIVE TRAFFIC MONITORING.

Traffic Monitoring and Characterization

In contrast to performance measurement, traffic monitoring is concerned with the flow characteristics of the traffic. Flow characteristics can be viewed in spatial, temporal, and composition aspects. Spatial characteristics refer to the patterns of traffic flow and demand relative to the network topology. Spatial patterns are important for proper network design and planning, such as provisioning for local versus long-distance connections, identification of concentrated demand in hot spots, and avoidance of likely bottleneck links.

Temporal characteristics refer to the stochastic behavior of a traffic flow, usually described in general statistical terms, a stochastic traffic model, or deterministic rate bounds. Temporal characteristics are important for service providers to understand for efficient resource management and traffic control. Statistical characterization usually consists of important parameters such as peak and average rates, as well as burstiness. Burstiness is a measure of unpredictability, which is the cause of queueing and possibly congestion, and can be quantified in various ways. Other important statistical characteristics may include frequency-domain components and long-range dependence or self-similarity. To explain the underlying traffic behavior, many stochastic traffic models have

Figure 2. RTFM architecture.



been proposed for various types of applications [1]; important classes of traffic models include Poisson, Markovian, and self-similar. A major issue is the accuracy of models for Internet traffic, which can be very complex and dynamic. Instead of modeling traffic behavior, deterministic rate bounds seek an algorithmic envelope for worst-case traffic, for example, the leaky bucket algorithm.

Composition of the traffic refers to a breakdown of packets according to content (data, control, acknowledgments), application (Web, email, file transfer, for example), packet length, flow duration, route length, and other attributes. The composition of traffic can provide insights into the applications using the network, which helps to explain its temporal and spatial characteristics. For example, studies of traffic in the vBNS (very high-speed backbone network service) Internet backbone revealed its largest component is Web traffic, which suggests client/server patterns will be dominant and Web server sites may be possible hot spots [12].

Although routers are ideal points for traffic measurements, they are generally not equipped today for exhaustive traffic monitoring. A collection of aggregate traffic statistics, such as the number of packets per interface over a time interval, is allowed in SNMP, but not sufficiently detailed measurements for accurate traffic characterization. Traffic analysts typically depend on special-purpose protocol analyzers, usually high-speed and expensive, or develop their own traffic collection devices such as OC3MON [12]. The OC3MON traffic collection system is an enhanced PC for capturing ATM cells on an optical OC3 (155Mbps) transmission link and collecting a trace of the encapsulated IP packets.

Some routers have advanced but proprietary traffic measurement and reporting capabilities, such as Cisco routers with NetFlow. The real-time flow measurement (RTFM) architecture is an attempt at an open, flexible system for traffic measurement and reporting, based on extending the ubiquitous SNMP network management paradigm [2]. As shown in Figure 2, the RTFM architecture prescribes a set of "traffic meters" located around the network (most likely at routers) capable of observing flows of packets that pass through the meter. A traffic meter should be configurable to selectively observe a specific packet flow and its various attributes defined by a program of rules from a network manager. As packets pass by a traffic meter, they are classified into groups according to the rules. The specified attributes of the flow (number of packets or bytes observed) are recorded in a database similar to an SNMP MIB, which can be retrieved by meter readers. In turn, applications can fetch data from meter readers through regular FTP or SNMP protocols.

Network Infrastructure

The PingER project at the Stanford Linear Accelerator Center is exploring the use of repeated pings around various sites for active performance monitoring [10]. Although ping can make only limited performance measurements (for example, roundtrip delay, packet loss, reachability), the correlation between IP-layer performance and application-layer performance is being studied. It exemplifies the use of current methodologies and existing infrastructure.

As agreements are being sought on new protocols and methods for network monitoring, a natural question is whether improvements are needed and desired in the Internet infrastructure. If needed, the Internet infrastructure may be evolved in two distinct ways: existing routers may be enhanced with new capabilities, or additional equipment may be deployed into the network. As an example of the first approach, routers might be enhanced with capabilities for:

- support of current and new ICMP messages or IP header options;
- clock synchronization, for example, by using global positioning systems;
- accurate writing and reading of timestamps; and
- support of RTFM traffic metering.

Unfortunately, it may be impractical to expect immediate hardware and software changes to all existing routers.

In the alternative approach, additional equipment might be deployed selectively at important points in the network, with no changes in existing routers. The National Internet Measurement Infrastructure (NIMI) project is an example of this approach, envisioning the deployment of enabling "platforms" at selective sites [8]. The NIMI platforms are generic in the sense that they are not tied to any specific measurement method or protocol. Instead, their purpose is to facilitate any methods, for example, traceroute or ping. As another example of this approach, the National Laboratory for Applied Network Research is deploying active measurement program (AMP) probes at various high-performance connection sites to enable active performance and throughput measurements [6]. Performance metrics include roundtrip delay, packet loss, and reachability; AMP probes are complemented by passive OC3MON traffic monitors.

For long-term evolution, routers should be designed with a powerful set of native traffic monitoring capabilities, and offer flexible control to support future management protocols. A recent trend toward programmable routers and so-called active networks is promising [11]. A programmable Internet would provide a separation between software control and hardware packet forwarding, and thus allow new monitoring methods and protocols to be developed without having to depend on changes in the network infrastructure. More long-term, active network technology could allow new monitoring methods to be dynamically installed and executed at selected routers via mobile code.

Conclusion

It is becoming apparent that the Internet cannot continue its successful growth without overcoming the limitations of current network monitoring techniques. Improvements are needed in the Internet infrastructure and protocols to facilitate performance and traffic monitoring. Furthermore, given the necessity for cooperation to make the Internet work, more agreements are needed between service providers for a common measurement infrastructure, protocols, and metrics.

Improving the observability of Internet behavior is only a first step toward the ultimate goal of more accurate monitoring. Simply collecting more raw data would be too overwhelming. An equally important problem is development of advanced tools to process the raw data and provide new insights and guidelines for application design. Tools and algorithms are needed to make use of raw data to optimize resource management and traffic control. Ultimately, the Internet needs to be better observed and understood if we are to make the best use of it. C

References

1. Adas, A. Traffic models in broadband networks. *IEEE Commun. Mag., 35* (July 1997), 82–89.
2. Brownlee, N., Mills, C., Ruth, G. Traffic flow measurement: Architecture. Internet RFC 2063, Jan. 1997.
3. Case, J., et al. Simple network management protocol (SNMP). Internet RFC 1157, May 1990.
4. Claffy, K., Monk, T. What's next for Internet data analysis? Status and challenges facing the community. In *Proceedings of the IEEE, 85* (Oct. 1997), 1563–1571.
5. ITU-T Draft Recommendation I.380. Internet protocol data communication service—IP packet transfer and availability performance parameters. Geneva, June 1998.
6. Mcgregor, T., Braun, H.-W., and Brown, J. The NLANR network analysis infrastructure. *IEEE Communications Magazine 38* (May 2000), 122–129.
7. Paxson, V. et al. Framework for IP performance metrics. Internet RFC 2330, May 1998.
8. Paxson, V. et al. An architecture for large-scale Internet measurement. *IEEE Commun. Mag., 36,* (Aug. 1998), 48–54.
9. Postel, J. Internet control message protocol. Internet RFC 792, Sept. 1981.
10. Stanford Linear Accelerator Center. WAN monitoring; www.slac.stanford.edu/comp/net/wan-mon.html.
11. Tennenhouse, D., et al. A survey of active network research. *IEEE Commun. Mag., 35,* (Jan. 1997), 80–86.
12. Thompson, K., Miller, G., Wilder, R. Wide-area Internet traffic patterns and characteristics. *IEEE Network, 11* (Nov./Dec. 1997), 10–23.

THOMAS M. CHEN (tchen@seas.smu.edu) is an associate professor in the Department of Electrical Engineering at Southern Methodist University in Dallas, TX.

Glossary

AMP: Active Measurement Program
ATM: Asynchronous Transfer Mode
CAIDA: Cooperative Association for Internet Data Analysis
ICMP: Internet Control Message Protocol
IP: Internet Protocol
MIB: Management Information Base
NIMI: National Internet Measurement Infrastructure
NLANR: National Laboratory for Applied Network Research
PING: Packet Internet Groper
QOS: Quality of Service
RTFM: Real-time Flow Measurement
SNMP: Simple Network Management Protocol
vBNS: Very High-speed Backbone Network Service