Firewalls in the Data Center: Main Strategies and Metrics

Joel Snyder, PhD

Senior Partner, Opus One

What You Will Learn

Measuring performance in networks has usually involved looking at one number: throughput. Since the first days of switches and routers, organizations have added up the performance they need, compared it to a total on a manufacturer's data sheet, and used those values to decide whether or not they had the right hardware.

As network managers have added firewalls to improve the security of the data center, the same performance engineering spotlight that shines on routers and switches is being applied to security appliances. When speeds jump above 10 Gbps as these firewalls move closer to the core of the data center, reliable performance metrics are critical.

Unfortunately for security and network practitioners, the same basic metric of throughput cannot be used to evaluate firewall performance. Because a security appliance actively participates in connections from Layer 2 up to Layer 7, you cannot simply look at bits-per-second throughput to predict how a firewall will behave in the data center.

In this document, you will learn key metrics you should use to evaluate firewall performance in the data center and why raw throughput is almost never the most important performance metric to use in your planning. Selecting a firewall does not mean simply picking the fastest firewall, but the one that is designed to handle the rapidly evolving, network-intensive application environment of the data center. You will also learn why today's firewalls must be built from the start to support today's network-based applications, and how to confidently use firewalls to increase security in data centers.

Traditional Firewall Performance

Because organizations always start with "feeds and speeds" (how many ports and how fast do they go) when evaluating switches and routers, it is tempting to apply these same metrics to firewalls: how many bits per second (bps) or packets per second (pps) can the device handle? If the firewall will go in the network core, it seems logical to use the same performance metrics for a security device that you use for a network device.

> Cisco MultiScale™ performance is a combination of breadth and depth. It provides rapid connections per second, an abundance of concurrent sessions, and accelerated throughput. It also enables multiple security services and spans physical, switch, and virtual platforms for exceptional flexibility.
>
> *– Fred Kost, Cisco Systems*

For example, stateless UDP traffic (such as you would see in a Network File System (NFS)) and long-lived TCP connections (such as you would see in a Microsoft Windows file system, an iSCSI Storage Area Network (SAN), or a backup application) are common in many data center networks. These types of applications present continued and heavy load to the network.

When you send file system traffic through a data center firewall, bits-per-second performance measurements are your starting point. But even in these simple cases, other performance metrics are equally important. For example, latency is a critical concern, because if the firewall introduces delays, applications will be affected. Because of the nature of TCP and file system protocols, even a small increase in latency can cause dramatic application slowdowns.

Unfortunately, performance engineering gets more complex as your application mix goes beyond pure file system protocols. After all, the network services that organizations are most interested in securing are application-layer protocols. This is where the problem occurs, because when a firewall is securing complex application traffic, you cannot measure performance in just bits-per-second and milliseconds of latency.

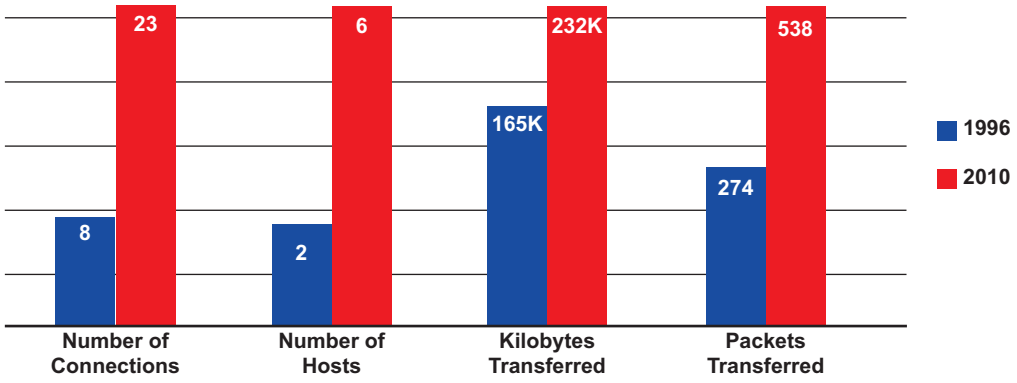
Switches and routers are active participants in the traffic that flows through them, but only up to Layer 3, the IP layer. Firewalls, however, are aware of each TCP connection and UDP session that passes through them. They participate at Layer 4, the session layer, at a minimum. As application connections come and go, the firewall must also create and tear down its own internal data structures to maintain state information for every session. This state information is checked and updated for every packet that passes through the firewall to provide the highest level of protection against sophisticated attacks.

When the firewall is also providing Network Address Translation (NAT) services or running Application Layer Gateways (ALGs) for applications such as voice over IP (VoIP) or video conferencing, the load rises even higher because the firewall has to decode and manage traffic all the way up to Layer 7, the application layer. Moving further and further up the stack makes network engineering more difficult and performance measurement more important.

In the data center, application traffic puts a very different load on the network than file system traffic. Client-server communications between users and servers, and server-server communications between application, database, and directory servers have very different profiles. Application traffic is connection intensive, with connections constantly being set up and torn down. This connection intensity adds another dimension to performance engineering. You must look beyond pure IP throughput and latency and include connection rates and connection capacities.

When a user logs in and connects to a file server, a TCP connection is created that can stay alive for hours. A half-second of delay in connection establishment will not even be noticed. But if the same user runs an application that opens two dozen connections for every page displayed, and each connection takes 500 extra milliseconds to set up because a firewall is adding delay, productivity will be affected.

Figure 1. The size of the Cisco.com web page increased by 30 percent over 15 years, but the number of network connections required tripled



The evolution of modern applications is tipping the balance from pure performance measures to metrics such as connection rate and concurrent-connection capacity. For example, look at the Cisco® website. In the past 15 years, while the size of the homepage has increased by 30 percent, the number of end-user connections required to download the page has tripled. Further, this value refers just to the page that appears to greet visitors. Logged-in users present an even heavier load as cookies are checked and the page is dynamically updated (Figure 1).

These statistics show only the front end of the application. Consider that for every page displayed, the front-end web server may also be querying for directory information, including subscription and entitlement information, checking security permissions, updating back-end state information and setting cookies, tracking user preferences

and history, and consulting content-distribution networks to modify the page in real time. A single click on a web application can cascade into a plethora of transactions across an entire data center.

The growth of the Internet represents one type of surge in use. A similar surge is also occurring inside the firewall, brought about by the increase in connected personal devices. The employee who previously had only a standard desktop computer may now also have a smartphone and a netbook or tablet device, all wireless, silently connecting to applications and services even when inside a pocket. One employee may now be putting two or three times the load on the network than that employee did before mobility became pervasive.

As applications and use patterns change, firewalls must change. The issue, however, is not just raw performance: today's firewall should not be just a faster version of yesterday's firewall. The firewall has to be reengineered to secure a different type of traffic with different performance demands.

Next-Generation Firewall Performance

As the network threat landscape has evolved, attacks have moved up the stack and focused on applications, usually web-based ones. Old threats with cute names like "Smurf" and "Ping of Death" are not significant anymore, as attackers have gone after the weakest link: the application. The capability to counter these advanced attacks has come to define the "next generation" of firewalls: security devices that are application aware and provide advanced intrusion-prevention capabilities (Table 1).

Table 1. Application-Aware firewalls must apply a complex set of security controls without affecting end-user perception of performance

Application-Aware Next-Generation Firewalls Must…	
Handle…	The surge of connections when a user first opens the application webpage
Maintain…	Connections as long as the user is within the application, even if no traffic is being transferred
Analyze…	The content of each application object, looking for prohibited content and potential network threats
Support…	IP throughput and latency requirements of the application to maintain user productivity and comply with internal service-level agreements (SLAs)

Measuring the performance of next-generation firewalls is difficult because the performance of these firewalls is now data dependent. Traditional firewalls required a close analysis of connection rates and connection capacities. Next-generation firewalls applying next-generation protections, such as application awareness and intrusion prevention, will offer different levels of performance depending on the data flowing through them. Unfortunately, as firewalls add increasingly sophisticated threat protections, the data dependence performance problem gets worse, not better.

> Moving up the stack to secure the application layer makes performance measurement difficult… and critical.

Thus, the measure of firewall performance has to go beyond Layer 4 metrics, such as connections per second or maximum number of simultaneous connections, and include Layer 7 measures, such as transactions per second. For example, if the firewall is HTTP aware, as most firewalls are, then HTTP transaction performance can be an important bottleneck. Because each transaction may contain different data and may activate different application protections, each transaction presents a different processing load to the firewall. In other words, two HTTP transactions of identical size may have different performance characteristics due to the data within the transaction.

In the quest to write the most interesting and engaging applications, developers have paid little attention to network performance and focused instead on usability. The result for firewall managers is an enormous amount of

uncertainty, as apparently simple applications begin to crush security appliances with the heavy load they put on the network. Old rules of thumb, such as the assumption that web pages will open and close connections quickly, just do not apply any more.

Consider one popular Internet application: Facebook. Although Facebook itself is not a corporate application, it represents a class of web-based applications that are already inspiring enterprise application developers. Data center managers hoping to secure application servers should consider Facebook as a "things to come" example of what they will soon be dealing with inside the corporate network.

> Today's firewall should not be just a faster version of yesterday's firewall; it must be reengineered for today's application performance demands.

Each time a Facebook user decides to check their homepage, the web browser opens multiple connections to multiple hosts; most of those connections download multiple objects, each of which must be individually analyzed by a next-generation firewall. One test found 19 connections opened to seven different hosts, downloading more than 50 HTTP objects, in less than five seconds. While a traditional firewall would be stressed by so many connections, a next-generation firewall is stressed even further by the large number of objects on every single page.

New applications with presence or IM capabilities, such as Facebook, pose an additional challenge by creating persistent connections as well. With Facebook, tests showed that four connections will be long lived, using firewall resources as long as the user is "on" Facebook. This behavior is counter to traditional firewall benchmark thinking, which assumes that all connections last only a second or two, so that concurrent capacity is not especially important. In other words, a firewall with a connection capacity of 8000 sessions will top out at 2000 Facebook users, even if all the users are away from their devices doing nothing.

The problems that Facebook raise for the scaling of firewalls do not map directly to a typical enterprise—or do they? Consider the move to unified communications, a synergistic combination of voice and video, synchronous (instant messaging) and asynchronous (email) communications, and presence information. When unified communications are fully implemented in an enterprise, the advanced applications do not look all that different from Facebook and Skype, wrapped into a single high-bandwidth package.

Going Beyond the Transport Layer

In the buzzword-laden environment of firewall marketing, it can be difficult to understand what differentiates generations of firewalls. This paper uses a simple definition: today's enterprise data center firewalls must go beyond simple access controls at the transport (TCP and UDP) layer to provide additional security. Whether it is called "next generation," or "unified threat management (UTM)," or "deep packet inspection," the point is that the firewall does not stop after it permits or denies access based on the IP address and port number, but provides additional security controls.

These controls can include:

- Ethernet-layer tagging (such as in the Cisco TrustSec™ solution): As policy-based access controls and identity-aware networking are deployed in enterprise and carrier networks, firewalls need to be aware of these security tags, transferred across trusted networks at Layer 2, and they must be able to include them in their access control decisions.
- Application-layer controls: With attacks focused on the application, firewalls need to be able to understand and control applications through technologies such as content filtering, web URL filtering, and malware detection.
- Intrusion prevention: Signature- and anomaly-based intrusion detection, when brought to a very high layer of fidelity, can easily identify common attacks and help protect unpatched systems or misconfigured applications.

This discussion of performance measurement of complex applications raises the question of standardized testing: is there a way for network and security managers to compare products without performing their own testing?

In firewall performance reporting, the only constant seems to be that no two vendors and no two tests use the same methodology. In fact, the same test lab will use different methodologies on nearly identical products a few weeks apart. One place to start, though, is with Internet RFC 3511, Benchmarking Methodology for Firewall Performance. Although RFC 3511 is slightly outdated (April 2003), it does lay out a set of metrics that are a good starting point for anyone installing data center firewalls. Table 2, based on the work of RFC 3511, summarizes the most important performance metrics and their importance to network managers evaluating data center firewalls.

Table 2. Firewall performance evaluation must move far beyond basic metric such as Throughput and Latency to meet today's application requirements

	Metric Name	Description	Why It Is Important
Basic	IP Throughput	• Raw capability of the firewall to pass bits per second or packets per second from interface to interface; similar to router and switch metrics • Typically shown at the IP layer	• If basic throughput is not in place, the firewall obviously cannot handle the traffic • Basic throughput is a good starting point to be sure that the product is close to what is needed
	Latency	• Time traffic is delayed in the firewall, added to the total end-to-end delay of the network • Should be measured in milliseconds at an offered load near the firewall's limit	• Latency affects perceived response time • Most firewalls have low latency when lightly loaded, but latency must be measured and reported when the firewall is at its operating load to understand whether undesirable delays will occur
Traditional Enterprise Firewall	Connection Establishment Rate	• Speed at which firewalls can set up connections and the full three-way handshake to establish a TCP/IP session • dozens of connections to be set up across an enterprise firewall	• Every click on a web-based application may cause • If this burst of connections overwhelms the firewall, user perception of application speed will be adversely affected
	Concurrent Connection Capability	• Total number of open connections through the firewall at any given moment	• To offer better performance, applications are keeping more connections open for longer periods of time, thus avoiding the overhead of connection establishment • These connection counts can increase quickly with modern applications
	Connection Teardown Rate	• Speed at which firewalls can tear down connections and free resources to be used for other traffic	• Since every connection that is set up must also be torn down, the firewall has to keep up in both directions
Next-Generation Firewall	Application Transaction Rate	• Capability of the firewall to secure discrete application-layer transactions (such as HTTP GET operations) contained in an open connection • May include application-layer gateways, intrusion prevention, or deep-inspection technology • Application transaction rate can be highly data dependent	• As attacks move from the network to the application layer, firewalls must peer deep into the application to secure the traffic • Simply providing basic access controls no longer meets the requirements for enterprise security of applications; the next-generation firewall must secure as much of the application stream as possible

Unfortunately, the higher up you move toward applications, the more difficult it is to define testing methodology for metrics such as application transaction rate. The data-dependent nature of application-aware firewalls and the vastly different characteristics of different applications make a single test plan for the firewall industry an impossibility. It falls to data center managers to work closely with their security suppliers to adequately test devices before committing to a final configuration to help ensure that performance will meet requirements.

The conclusion is easy: measuring performance is hard. But without a complete picture of behavior at the application layer, it is impossible to predict how firewalls in data center networks will behave. Fortunately, firewall test and measurement vendors, such as Spirent, Ixia, and Breaking Point, have developed the tools to measure firewall Application Transaction rates. With sound application-specific performance data, designing data center networks with embedded security is possible.

Firewalls in the Data Center without Fear

Perimeter network security, the traditional domain of firewalls, is a clear requirement for all networks: the Internet is a constant threat to anything connected to it. Focusing outward, however, ignores internal threats to the data center, home of the most valuable assets of the enterprise. Yet data center firewall deployments have lagged far behind perimeter installations. Why have network managers been afraid to put firewalls in their data centers? For two reasons:

1. **Performance:** Most firewalls are still designed and built for the traffic patterns of 15 years ago and for the security protections needed during that same era. When a firewall cannot handle the load, the last place you want to put it is in the middle of a data center.

2. **Benchmarks:** Firewall vendors have continued to promote performance benchmarks that tell you almost nothing useful about how the firewall will perform in a data center.

The result is unfortunate: network and security managers have avoided firewalls in their high-performance networks. The lack of access controls has resulted in lower security in enterprise networks. Even when firewalls are installed, network teams often do everything they can to make exceptions and avoid sending traffic through security control points. They resort to building a mess of complex routing and switching elements, suboptimal management capabilities, and reduced access controls. Deep inside the network, especially within the data center, they have the least access controls. In other words, where organizations need security the most, they have avoided it.

Data center deployments have other requirements besides high performance. For example, central management is critical in data centers where firewalls of many types—physical, embedded, and virtual—will be deployed in many locations. Full visibility and control of security in the data center is also needed. Monitoring the firewalls and intrusion-prevention technology together gives the security manager insight and provides the capability to use correlation between devices for early detection of emerging threats.

> As attacks have become more sophisticated and widespread, organizations need to go beyond perimeter security and focus on the data center. Organizations cannot be afraid to secure their own networks.

As attacks and attackers have become more sophisticated, network and security teams are starting to realize that they need to go beyond perimeter security and focus on the data center. Before they can do that, though, firewall vendors need to start designing firewalls for today's loads and application-layer protections. When that happens, organizations will be able to move security to where they need it, deep within the data center, without fear or compromise.

OPUS
1404 East Lind Road
Tucson, AZ 85719
520.324.0494
Fax 520.324.0495
http://www.opus1.com