The Shortfall of Network Load Balancing

Applications running across networks encounter a wide range of performance, security, and availability challenges that cost organizations an enormous amount in lost productivity, missed opportunities, and damage to reputations. Faced with requirements for application delivery that have evolved beyond the capabilities of traditional network load balancers, enterprises must migrate to a more advanced application delivery architecture.

White Paper
by Nathan Pearce

Introduction

The way in which applications are consumed has undergone rapid and significant change in recent years—both in the growing list of user device types to support, each with its own unique requirements, and in how those devices are connected, with variations in network characteristics between fixed, mobile, and remote networks. Furthermore, the nature of how applications behave also has changed. Applications are now far more interactive, no longer just presenting web content and waiting for the next request but constantly streaming information, in and out of the data center, during run time.

Enterprises must now look at the delivery of applications and services differently, taking into account these new challenges and addressing the shortcomings of basic, network-oriented load balancing products.

Why Change?

Network load balancers were created as a tactical solution to a scaling requirement: spreading network connections across more than one server. As the technology evolved, certain basic features were added, including scheduled heartbeat checks to see if the servers responded.

But the way in which applications fail often is not binary, not an on-or-off state. Applications can fail to deliver data quickly, accurately, securely, or appropriately for the requesting device. Network load balancing alone delivers no protection against these types of delivery failure. This significant limitation is due to a lack of appreciation of the context in regard to:

- The performance of remote networks beyond the data center.
- Unexpected application behavior.
- The capabilities of the user device.
- Real-time user experience analytics.

Network Performance

The slowest segment of data's journey between server and a remote device is typically the last mile of delivery—in the network connected directly to the user technology. This is increasingly so for mobile devices in particular. Irrespective of the medium, however—wireless or fixed-line—end-to-end performance management is a critical step toward a positive user experience.

The effects of poor performance caused by networks beyond the hosting data center are amplified by the inefficiencies of web applications and their protocols. Web applications are often developed in sterile lab environments lacking the difficulties of real world application access. Consequently, behavioral testing in these labs returns acceptable results that are not mirrored in production environments. In the case of interactive applications, where data is received from and pushed to the data center, the impact of minor latency becomes significantly increased.

Unexpected Application Behavior

Unforeseen errors occur during application run time. For example, uncaught exceptions (as they are referred to in programming terminology) occur when an application reaches a state that was unexpected by the developer. They make their way into applications via a number of means—inexperienced developers, oversights caused by pressure to release functionality quickly, and bugs in application programming interfaces (APIs), to name just a few—and they typically result in an application crash. Unfortunately, asking developers not to make mistakes isn't a viable solution. They're a consequence of the human element in computing.

The resulting experience for users can range from meaningless or confusing application error messages to long delays that end with a session timeout and a web browser error. The cost to the enterprise is loss of business and reputation damage.

Device Capabilities

Optimal application responses for laptops, tablets, and smartphones are not alike. For example, Apple's iOS, the operating system found on iPhones and iPads, cannot consume web applications developed using Adobe Shockwave Flash, nor can they interpret web applications developed in Microsoft Silverlight.

The data center network must concern itself with the unique requirements of a growing list of user device types. Failing to do so leads to poor user experiences— errors, confusion, and disappointment.

Real-Time Analytics

Analytics provide enterprises with operational visibility. However, unless that visibility is derived from real-time behavior and includes detailed application and user device information, then the data is of little use. Only user experience-related analytics can be used to measure successful application delivery. Those analytics must include performance; operating system and web browser versions; and network and application responsiveness, but few components of a traditional network architecture can provide them.

What to Change

Network load balancing was derived as a response to enterprise needs for a tactical solution to a performance and resilience problem in the data center. Increased performance resulted when the workload was spread, albeit unintelligently, across more than one web server. Resilience also improved marginally due to the elimination of the single point of failure inherent in a single-server deployment. This solution was sufficient for enterprises not running mission-critical, revenue-driven business applications.

Unlike network load balancing, application delivery is concerned with making mission-critical applications highly available, fast, and secure. It addresses performance and application behavior, responds to device capabilities, and can provide useful real-time analytics to guide decision-making and development.

Performance

Network load balancing is concerned only with forwarding data packets to the next, directly attached device with no appreciation of acceleration, optimization, or workload offloading techniques that, together, vastly improve performance. While some of these techniques can be introduced into application software development, they require specialist programming expertise and ongoing management. On the other hand, many of these performance improving techniques cannot be handled by the application in any case but can easily be provided by the right Application Delivery Controller (ADC).

Application Behavior

Poor application behavior is properly handled with an understanding of the context —that is, by having an intelligent services framework within the network that can differentiate between successful application transactions and service failure. An advanced ADC can handle application failure in two ways. In the event of a short-term failure, the ADC will discard the failed response and re-request the transaction from another server. Should the entire service be unavailable, with no other servers able to fulfill the transaction, instead of resorting to a browser timeout, the ADC will present a corporate branded out-of-service holding page, which is a far more graceful response to a larger problem.

> " BIG-IP ensures maximum availability, enhanced performance, and security for mission-critical Oracle applications. In fact, at Oracle we use approximately 170 BIG-IP devices in front of over 13,500 different applications and services in four data centers worldwide.
> —Leonid Stavnitser Senior Manager Global IT, Oracle USA, Inc.

Device Capability

Network load balancing devices are incapable of differentiating between one client device or another, resulting in a mismatch of what devices can consume which service. Only an advanced ADC has the ability to identify the requesting device and then route it to a service capable of successful rendering, thus, avoiding errors, confusion, and user disappointment.

Real-Time Analytics

Analytics provided by capable ADCs deliver real-time visibility into application performance from the user's perspective, allowing enterprises to assess, isolate, address, and track problem areas before critical outages have occurred. Such analytics, using metrics not visible with legacy solutions, enable efficient troubleshooting, application visibility for capacity planning, and performance tuning and optimization.

Why F5 ADCs?

Choosing the right Application Delivery Networking architecture requires organizations to assess ADCs, and by extension ADC suppliers, for their:

- Application fluency.
- Programmability.
- Collaborative support resources.

Application fluency in an ADC closes the knowledge gap between data packets traversing a network and high performing, reliable, and secure business applications. By understanding both the network and the applications being delivered, a fluent ADC can adjust the network to accommodate changes in the application environment or vice versa, responding to network fluctuations by managing application access and performance. The application fluency of F5® ADCs is derived from F5 Networks' extensive Technology Alliance Partner program, which brings with it co-developed, validated best-practice guides. For an enterprise, this translates to removing or reducing risks to business-critical application availability and with that, increased peace of mind.

The programmability of F5 ADCs, which ensures custom tuning to meet the needs of the organization and its particular network environment and applications, delivers integrated workflows and data center agility. Combined, these benefits enable the rapid deployment of new services and quick reactions to changing demands on existing services. Through programmability, organizations achieve extensibility, turning the network into a pool of reusable resources for resilience, security, and optimization.

Finally, real-world applications are developed and consumed by the people comprising the F5 DevCentral™ community, a collaborative resource of over 110,000 active subscribers and the leading source of application delivery solutions. DevCentral, which enables expert and peer sharing of experiences, turnkey programming tools, proven code, and other resources, is the online authority F5 customers can use to ensure that their applications are always fast, secure, and available to employees and customers.

F5 is the only ADC supplier that is 100 percent focused on advanced Application Delivery Networking. This focus underlies the application fluency and unmatched programmability of F5 products that is key to managing enterprise-grade application user experiences. It has created the longstanding ecosystem of application delivery experience that supports F5 customers through increasingly constant change.

Conclusion

With a broad range of devices to support and varying remote network connectivity problems, IT departments must deliver access from anywhere and on any device securely, fast, and with unbroken availability, 24 hours a day. There is little strategic value in the unintelligent forwarding of packets from device to device as performed by network load balancers. Rather, solving today's application delivery issues requires:

- **Application awareness:** Total insight into how the application is supposed to look on the wire.
- **User awareness:** The ability to see which users are trying to access what application, from which devices.
- **Resource awareness:** The intelligence to tie all the pieces of the application delivery infrastructure together to provide real-time visibility into the entire Application Delivery Network.

Mitigating against the many ways in which applications can fail demands an advanced Application Delivery Controller with all three types of awareness. F5 ADCs in the network provide an intelligent services framework capable of such awareness and thus enabling secure, reliable application delivery.