Software-Defined Hardware: Enabling Performance and Agility with the BIG-IP iSeries Architecture

White Paper

Introduction

A perfect storm of market trends is shifting the application and IT landscape. Traffic growth continues unabated due to mobile, streaming media, the Internet of Things (IoT), and digital transformation affecting every industry. Cloud adoption is accelerating to respond to enterprise needs for agility and faster deployment as the number of applications increases. Along with the growth in applications and everything connected via the Internet have come security threats, including data breaches and the theft of critical business and personal data. One outcome has been encryption of all data in transit, with SSL traffic doubling in the past year—from 29 percent to 64 percent of all traffic. Lastly, new application architectures and agile development models, including microservices and DevOps, are emerging.

This rapidly shifting IT landscape poses significant challenges for enterprises, including how to cost effectively scale applications and the associated services, handle the complexity of deploying private clouds, and protect apps and data across a hybrid environment. Each product added to the architecture solves part of the problem but results in product sprawl, which creates higher TCO and new attack surfaces.

These challenges necessitate a next-generation, cloud-ready Application Delivery Controller (ADC) platform that provides DevOps-like agility with the scale, security depth, and investment protection needed for both established and emerging apps. The F5® BIG-IP® iSeries is a family of next-generation ADCs specifically designed to meet these challenges. This high performance and programmable ADC can:

- Deliver the best performance and price/performance ratio across a range of application delivery and security metrics to service the explosive increase in client and server traffic.
- Outperform software-based products running on commodity hardware.
- Overcome the tremendous load on SSL/TLS infrastructures caused by encrypting all traffic. It accomplishes this through offloading and accelerating cipher and key exchange processing, including elliptical curve cryptography (ECC) and forward secrecy (FS).
- Provide cost efficient, shared-edge services in two-tier architectures for private clouds.
- Future-proof the application infrastructure through hardware that can be repurposed via user selectable performance optimizations and software updates. This ability to repurpose hardware helps to deliver agility in an infrastructure that can meet evolving business needs.

BIG-IP iSeries appliances are hugely scalable, low-latency application proxies that not only provide layer 4-7 application delivery services but serve as the foundation of the F5 architectural vision for application delivery in the future. The family optimizes the power and flexibility of dedicated hardware and software running on the ideal mix of CPU, memory, and co-processing components, while innovative hardware optimizations deliver unmatched performance and scalability.

BIG-IP iSeries Innovations

The BIG-IP iSeries includes several new capabilities, driven through unique hardware optimizations, which are integrated to meet the changing demands on enterprise IT—including the shift toward private cloud architectures, protection of critical data, and consolidation of application services.

Reprogrammable performance profiles

Gordon Moore famously observed that the number of transistors in an integrated circuit appears to double each year for the same costs. (This observation was later adjusted to reflect doubling every two years). Moore's Law, as it came to be known, was so accurate a predictor that it was extended to other aspects of computing, such as CPU speed, memory speed, network speed, and disk drive storage density. While transistor count and CPU speed continued to double, everyone in the industry knew the trend could not continue forever. Even so, relying on regular speed and density increases allowed for critical functionality to be implemented in software.

For functionality where software and CPU was still insufficient to meet performance needs, the technology industry responded with integrated circuits designed for a specific application: the application-specific integrated circuit (ASIC). An ASIC had the advantage of being able to deliver a specific capability at the fastest possible hardware speed. ASICs found use throughout computing where performance was critical, such as network cards, memory controllers, graphics processors, and encryption. While an ASIC could perform tasks at the fastest possible speed, its capabilities were limited to what was designed into the chip. If an application needed capabilities not available from the ASIC, the application had to use a CPU instead and perform the tasks in software.

Today, CPU speed no longer continues to double so quickly. If a CPU can be viewed as flexible but slow, and an ASIC as inflexible but fast, a third technology operates in the middle: the field-programmable gate array (FPGA). An FPGA is slower than an ASIC but much faster than a CPU, and it can be programmed to perform tasks not envisioned by the FPGA designers.

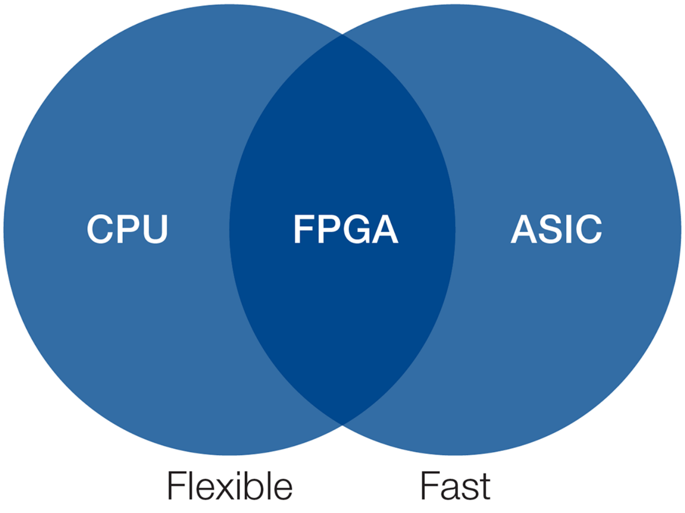

Figure 1: CPUs, ASICs, and FPGAs

FPGA technology in the BIG-IP iSeries enables network and application traffic processing to be offloaded to the FPGA silicon. This frees CPU resources available for other, more complex tasks requiring greater flexibility. The advantages of an FPGA are the abilities to:

- Tailor the offload features to the role of the platform.
- Be reprogrammed, providing software flexibility.

A balanced approach exploits the advantages of both ASICs and FPGAs.

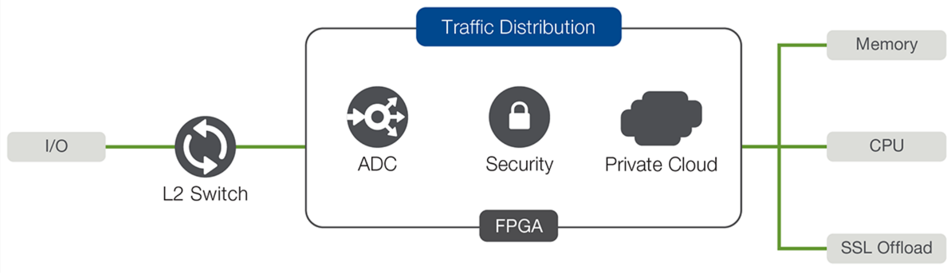

Figure 2: How the BIG-IP iSeries system architecture uses FPGA technology

The FPGA sits at the heart of the traffic passing through the appliance, inspecting traffic, making critical traffic management decisions, and offloading protocol and security processing. It provides:

- Traffic aggregation and disaggregation, steering traffic to the appropriate CPU. (Read more about this capability.)
- Traffic direction to the appropriate cryptographic hardware for SSL/TLS processing.
- Assurance that the correct traffic is sent to compression hardware.
- DDoS mitigation that can absorb larger attack sizes (up to 10 times compared to software only).
- Client white, gray, or blacklisting (also known as shunning).
- Private cloud tunneling protocol performance improvements (up to 50 percent more traffic processing).
- Hardware enhanced L4 load-balancing for:
  - UDP and TCP.
  - IPv4 and IPv6.
- Network overlay encapsulation/de-encapsulation.

In addition, an FPGA can respond to traffic with bounded latencies, ensuring that traffic can be handled appropriately and with a uniform performance level, even under load—unlike a CPU, where software performance can vary as other software vies for the same CPU resources. Specifically, when software is used to direct traffic and the CPU is under load, such as during a DDoS or SSL negotiation attack, the CPU and software responsiveness decreases, reducing the ability of the ADC to manage the attack. Conversely, when an FPGA directs traffic, it will respond predictably, regardless of the load on the CPU.

All of these traffic management tasks can be performed before the traffic is further processed by downstream components. The FPGA enables inspection and dispatch to the appropriate components, facilitating intelligent traffic management at hardware speeds.

TurboFlex functionality in the BIG-IP iSeries

FPGA technology has been incorporated into previous generations of ADCs, but the BIG-IP iSeries significantly enhances the number and type of performance optimizations and how they are selected through the use of F5 TurboFlex™ performance profiles.

In other systems, designers have sought to create a balanced set of features and then have specifically allocated FPGA resources to these features. FPGAs cannot dynamically allocate additional resources to different tasks, since sections of the FPGA are statically assigned to particular tasks. This has the benefit of true parallelism—but at the expense of some flexibility.

The BIG-IP iSeries introduces multiple, customer-selectable FPGA performance profiles. By enabling a specific TurboFlex performance profile through the BIG-IP GUI, a command line, or an API call, customers can allocate resources of the FPGA differently or offload additional tasks to meet requirements.

If the platform's primary use case is as a security-focused edge device, for instance, functions such as DDoS mitigation, TCP processing, and whitelisting might be favored. For a device providing advanced application delivery services in an OpenStack powered cloud, network overlay processing, such as VXLAN or GRE encapsulation, would be a priority. Choosing the most appropriate performance profile will allocate the FPGA resources in the best way possible for the workloads.

Moving high volume but relatively simple tasks to the FPGA boosts overall system performance, both by performing tasks at wire speed and by reducing CPU load. Different TurboFlex profiles optimize performance for different use cases while still providing full capabilities.

It is certain that the coming years will bring further change. Infrastructure will need to support new protocols, deal with new threats, and provide new services. "Static" and "unchangeable" are not good descriptors for a future-looking infrastructure.

With the BIG-IP iSeries, additional task offloads and features (in the form of new TurboFlex performance profiles) will be added to the FPGA technology through BIG-IP software updates. Those additions will provide new capabilities without the need to procure new platforms. In fact, BIG-IP iSeries platforms will evolve with your architecture, adding to its capabilities rather than restricting them.

New hardware SSL capabilities

As attacks on SSL/TLS continue to advance, key lengths must continue to increase to stay ahead of attackers. Longer keys necessarily require more computing capability, and as noted above, CPU speed is no longer progressing at the pace required to match. The performance problem of managing SSL/TLS connections is compounded by an increase both in general Internet use and in the proportion of secure connections. General traffic is increasing, and more of that traffic is being encrypted, due to HTTP/2 and strong adoption by hyperscale web companies such as Google, Apple, Facebook, and Microsoft.

While modern CPUs handle TLS key exchange well, offloading these tasks still makes sense in the context of a hardware ADC—especially when the available CPU is required for higher-level tasks, such as web application firewall services, that require more complex processing. In addition, using specialized hardware is more cost effective per handshake compared to a CPU.

This is especially true when considering the demand for new cipher suites to protect traffic. Until recently, the dominant cipher for key exchange has been RSA. RSA has a significant flaw, however. While the ciphers themselves are secure, if the private key is compromised, a third party can use it to replay recorded traffic encrypted with that private key.

This has a significant implication. Anyone who records SSL traffic encrypted with RSA can replay it at any time in the future once the key is compromised. Countless private keys have been compromised through human error or software failures such as the Heartbleed bug. Any traffic encrypted with any of those keys is readable by anyone holding those keys.

This concern alone is driving organizations toward a concept known as forward secrecy (FS), which renders immune a replay using an exposed private key. Diffie-Hellman ciphers, the first widespread ciphers to provide forward secrecy, have demonstrated implementation defects in most software, but the more advanced Diffie-Hellman elliptical curve cryptography (ECDHE) ciphers do not have known vulnerabilities. The move to forward secrecy has triggered sites to support and prefer ECDHE cipher suites.

Older-generation, hardware SSL acceleration components do not support ECDHE, meaning that software and CPU resources are consumed to process connections, placing additional load on the system. As ECDHE adoption continues to rise, managing these connections in hardware becomes ever more important.

By including the latest generation of cryptographic acceleration hardware and carefully integrating it into the BIG-IP system, F5 has designed the BIG-IP iSeries to offer hardware offload of ECDHE across all platforms. This enables the rapid adoption of elliptical curve cryptography (ECC) and ECDHE cipher suites by providing hardware acceleration for ECDHE as well as by supporting existing ciphers, even in high-load TLS environments.

Increased memory for more objects and more connections

Internet traffic is growing. In fact, in a 2016 report, Cisco predicted a threefold increase in traffic over the next five years[1]. Some of this traffic growth will be organic, due to more consumer connectivity and richer web experiences. Some will come from the predicted explosion in connected devices—the Internet of Things. Gartner predicts over 20 billion connected devices by 2020. Along with this growth in devices comes an increase in the number of connections infrastructures must support. There may be a parallel mitigating effect caused by the adoption of HTTP/2, which uses fewer connections per client, but it's sure that front end connections numbers will rise significantly.

At the same time, application architectures are moving from a monolithic architecture to a system of multiple, loosely coupled services. Traffic management between services (east-west traffic) will require similar management and security services to client application (north-south) traffic. This too will result in significant growth in the number of connections to manage.

The BIG-IP iSeries has been engineered to support significantly more connections than their predecessors—up to double—due to increased memory provisioning and more efficient connection handling. This will provide the power and scalability to manage and protect traffic from a new generation of connected devices and applications.

Part of a wider vision

The BIG-IP Series is a central part of the F5 architectural vision for the future. While the new hardware features of the iSeries make it an ideal platform for many use cases and situations, its integration with the rest of the F5 ecosystem makes deploying, integrating, and managing the BIG-IP appliances both agile and operationally efficient.

Conclusions

The BIG-IP iSeries is the next generation of ADCs, with key innovations that enable customers to meet the challenges of rapidly shifting landscapes. Beyond a significant increase in capacity, the BIG-IP iSeries mixes the power of dedicated hardware with the flexibility of a programmable, updatable platform. By carefully integrating software on the CPU with the latest ASIC and FPGA technology, BIG-IP iSeries devices provide maximum performance, investment protection, and the flexibility to consolidate multiple app and security services onto a unified platform.

Specifically, placing the FPGA early in the traffic path delivers predictable latency for inspecting and dispatching traffic to appropriate components, all without using the CPU. Reprogrammable TurboFlex performance profiles permit the FPGA to be tuned to specific uses, such as DDoS mitigation and TCP whitelisting for security-focused edge use, or network overlay processing for cloud use.

The programmability of both the software and TurboFlex performance profiles ensures that the BIG-IP iSeries can meet today's needs and the unforeseen needs of tomorrow. Support for hardware acceleration of cipher suites, including ECDHE, and capacity for the explosion of connectivity in the coming years place the BIG-IP iSeries at the center of next-generation data center and private cloud architectures. Integration into a wider F5 architecture enables the speed of service deployment to match the speed of traffic processing.

The BIG-IP iSeries marks a dramatic jump forward in the evolution of Application Delivery Controllers.

[1] Cisco Visual Networking Index: Forecast and Methodology, 2015–2020. Updated June 1, 2016.

F5 Networks, Inc.
401 Elliott Avenue West, Seattle, WA 98119
888-882-4447 f5.com

Americas
info@f5.com

Asia-Pacific
apacinfo@f5.com

Europe/Middle-East/Africa
emeainfo@f5.com

Japan
f5j-info@f5.com