White Paper

Infrastructure and Information Security

November 2020

IT@Intel: "Instant" VPN Scaling and Continuity During Crisis

When the pandemic struck, Intel® architecture-based servers, virtual gateways, a hybrid cloud plan, and fast action from Intel IT rapidly scaled VPN security and kept Intel's global workforce productive

Intel IT Authors

Husni Bahra
Principal Engineer

Kevin Bleckmann
Cloud Architect

Chandra Chitneni
Principal Engineer

Sridhar Mahankali
Principal Engineer, Information Security

Sanjay Rungta
Senior Principal Engineer

Pavel Terekhov
Network Security Engineer, Information Security

Tony Valverde
Principal Engineer

Table of Contents

Executive Summary

When the COVID-19 pandemic struck, it propelled Intel's global workforce of over 100,000 employees, along with contingent workers and ecosystem partners, to begin working remotely practically overnight. Virtual private network (VPN) connections would be needed for all of these workers, but Intel's global VPN capacity was a fraction of what would be needed.

Depending on the geography, Intel IT needed to scale VPN capacity either through a public cloud service provider (CSP) or through on-premises local cloud servers. Either way, as the world cascaded into quarantine, there was no way to procure the necessary VPN gateway appliances. We adapted to the situation and created the needed capacity with virtualized VPN gateways.

In the scant days following Intel's work-from-home order, we not only met Intel's VPN capacity needs but exceeded them. We accomplished this by fine-tuning off-the-shelf solutions based on Intel® architecture-based servers and Intel® Ethernet network adapters enabled with SR-IOV functionality. Based on our internal team's deployment, we were able to scale our virtual VPN gateways by 100 to 200 percent of the performance level seen in some dedicated hardware VPN gateway appliances. The solution spotlights the exceptional capabilities of virtualized solutions, SR-IOV, and multicloud strategies for overcoming considerable challenges in a timely manner.

Business Background

In late January 2020, as the Chinese workforce returned from its new year celebrations, awareness of the COVID-19 pandemic was rippling across the nation. Wuhan and other cities went into lockdown on January 23rd. Intel IT realized that Intel's China-based workers would immediately transition to working from home. From Friday to Sunday, we rushed to work with our Internet service providers and augment ISP capacity by 50 percent, landing a new ISP circuit within 40 hours (a process that typically takes several days or even weeks). Starting Monday morning, with lockdown in effect, virtual private network (VPN) loads rocketed to 2.5x their regular daily levels. We had to devise strategies for traffic prioritization using quality of service policies and manage congestion through improved load balancing based on custom route policies. Ultimately, Intel's workforce in China faced a few weeks of adjusting to new procedures and network traffic priorities, but the system worked well from the first day, and Intel smoothly continued its operations.

Even though Intel managed its pandemic response in China for more than a month, advance notice never passed to other geographies, as the consensus was that the problem had been contained—and so, for Intel IT at least, history repeated itself as the virus exploded globally. Even though we learned several key lessons about managing and scaling VPN capacity from the Chinese pandemic response, other regions presented an entirely different magnitude of challenges. Intel's workforce in China can operate with a 50 percent increase in VPN gateway bandwidth capacity. In contrast, key Intel locations, both in the U.S. and at sites such as India and Malaysia, required many times higher bandwidth levels.

Of course, the rest of the world faced its own COVID-19 challenges in the ensuing weeks. Friday, March 13th marked something of a dividing line in much of the U.S., and this was the day Intel notified its U.S. employees that they would be working from home until further notice. We went on high alert. At that point, Intel's global VPN capacity could handle about 50,000 users, with half of those being in the U.S. That nearly instantaneous 2.5x jump in VPN use seen in China suddenly loomed over us on a global scale. In the U.S. due to our multicloud strategy preparation, we already had resilient, high-bandwidth connectivity between Intel and a CSP. We used that connectivity to burst VPN capacity via the CSP, although we had never before tried running virtual VPN gateways on that scale. Orders for physical VPN gateways were suddenly stalled or canceled as demand raced beyond supply. Our VPN vendors had to prioritize similar orders from healthcare and other pandemic response groups focused on critical institutions. Wait times grew to two months or more.

We needed to scale into the public cloud in the U.S. and into private clouds around the world. This private cloud push depended in part on the addition of new virtualization infrastructure to ensure performance optimization and tuning for VPN capacity. However, physical VPN appliance availability for those local infrastructures had suddenly vanished. Work started on Thursday, March 12. By Friday, teams were deep into tuning settings and standardizing configurations, but that was still just the beginning. If we didn't have solutions in place for Intel's more than 100,000 employees (plus contract laborers and ecosystem partners) who would need to work remotely on Monday morning, the company would be crippled.

A Double Solution

We faced two problems that needed different yet related solutions. Eventually, we found that the common denominator between them was Single Root I/O Virtualization (SR-IOV).

The PCI Special Interest Group (PCI-SIG) introduced SR-IOV as an extension to PCIe in 2008. The technology serves as a way for virtualized systems "to address sharing of I/O devices in a standard way."[1] To illustrate within this paper's scope, in a virtualization environment, the hypervisor typically mediates all communications between the VM and network interface controller (NIC). SR-IOV enables network performance enhancement by enabling the VM to get direct access to the NIC, thus minimizing or eliminating the impact of processing through the software switch layer of the intermediate hypervisor stack. This represents a performance gain for network traffic.

Additionally, as shown in Figure 1, SR-IOV allows a PCIe device to present multiple virtualized instances of itself to an operating system or hypervisor, so one SR-IOV-compliant NIC in a system can present up to 256 virtual NICs,[2] each of which could operate within a virtual VPN gateway. Thus, SR-IOV enables a relatively limited amount of off-the-shelf server hardware to provide an extensive amount of VPN gateway capacity—which is exactly what we needed in the face of rapid VPN scaling demands.

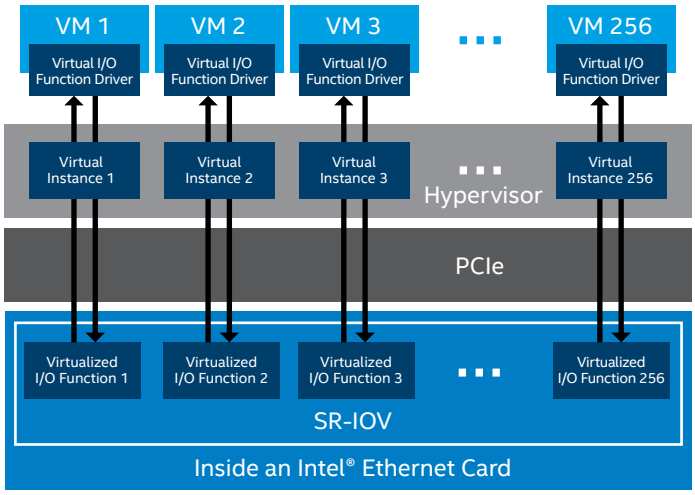

Figure 1. SR-IOV provides a way for the hardware capabilities of a PCI Express device (a network adapter in this case) to be presented to the operating system or hypervisor as multiple virtualized NICs without some of the hypervisor's overhead and processing latency.

In the U.S., we had high-speed WAN connectivity to a top-tier CSP to handle projected VPN bandwidth needs. The question was whether the virtual VPN gateways at the CSP could process the expected influx of VPN traffic. Our vendor's VPN gateway solution was available in the CSP's marketplace and was already optimized for SR-IOV. However, the vendor's VPN gateways hadn't been used in production at this scale.

Over three days, our engineers worked around the clock with counterpart network architects from the CSP to understand the deep nuances of CSP networking. This already difficult task was made even harder by the fact that we had never worked with virtual VPN gateways before. Nevertheless, working in the CSP's environment Thursday through Saturday, we emerged with the ability to support VPN gateways' unique characteristics, such as associating a large pool of IP addresses (for assigning to users as they connect to the VPN gateway) and ensuring that we could route network traffic end-to-end between Intel's network and the VPN gateways at the CSP. We also worked with our VPN vendor to appropriately size and tune the virtual VPN gateways. Soon, we had multiple configured VPN gateways up and running. On Sunday, that solution was accepting test traffic from a relative handful of IT users. Production load hit the CSP's VPN gateways as Monday morning arrived in the U.S.

Meanwhile, other geographies without high-speed connectivity to CSPs needed an alternative plan to scale up their VPN capacity. There was no time to source purpose-built VPN gateway appliances, even if any had been available to buy. Thus, teams around the world repurposed suitable servers using Intel® Xeon® Scalable processors, targeting the core scaling and memory capacities needed to maximize performance while benefitting from the processors' integrated encryption acceleration engine, which enables VPN functionality possible at this scale. We had never deployed virtual VPN appliances, making this a risky proposition. Nevertheless, we scavenged what Intel Xeon Scalable processor-based systems we could, often pulling out existing network adapters and replacing them with Intel 500 and 700 Series Ethernet Adapters that supported SR-IOV. Our engineers configured them over the weekend by working with partner teams around the world, then closely monitored their performance Monday morning to assess how well the newly provisioned virtual VPN servers could handle the load.

In the following few days, we learned that our strategy was working exactly as hoped. The virtual VPN gateways met the surge with complete transparency for end users. Note that we had not configured SR-IOV in our private cloud resources during opening deployment days. In several days of observing results with our CSP, we learned that SR-IOV could deliver highly beneficial performance improvements. This led us to pursue similar functionality and results in our private clouds in the two to three weeks following initial rollout as we worked to optimize our solutions. We expected (and observed) that SR-IOV would allow us to increase the performance and capacity of our private cloud virtual VPN gateways well beyond their initial non-SR-IOV baseline.

Not least of all, we understood that while we were able to burst VPN traffic into the public cloud, we might eventually want to bring some or all of that traffic back to private infrastructure. Between VPN traffic being network-intensive and public cloud network charges being based on usage, public cloud traffic could easily cost thousands of dollars per day. Even as locations using CSPs for VPN gateway services were being configured and deployed, Intel IT teams worked in parallel over roughly five weeks spanning March and April to prepare on-premises systems to handle that region's VPN traffic.

Results

On March 11, Intel's global VPN capacity remained at its then-usual ceiling of 75,000 users. As shown in Figure 2, on March 16, actual demand surpassed this level. Fortunately, Intel's VPN capacity managed to scale in advance of the need. By the time actual demand peaked at over 80,000 global users, we had reached capacity for 125,000 and went on to achieve readiness for more than 200,000 by mid-April. We nearly tripled Intel's VPN capacity in under three weeks.

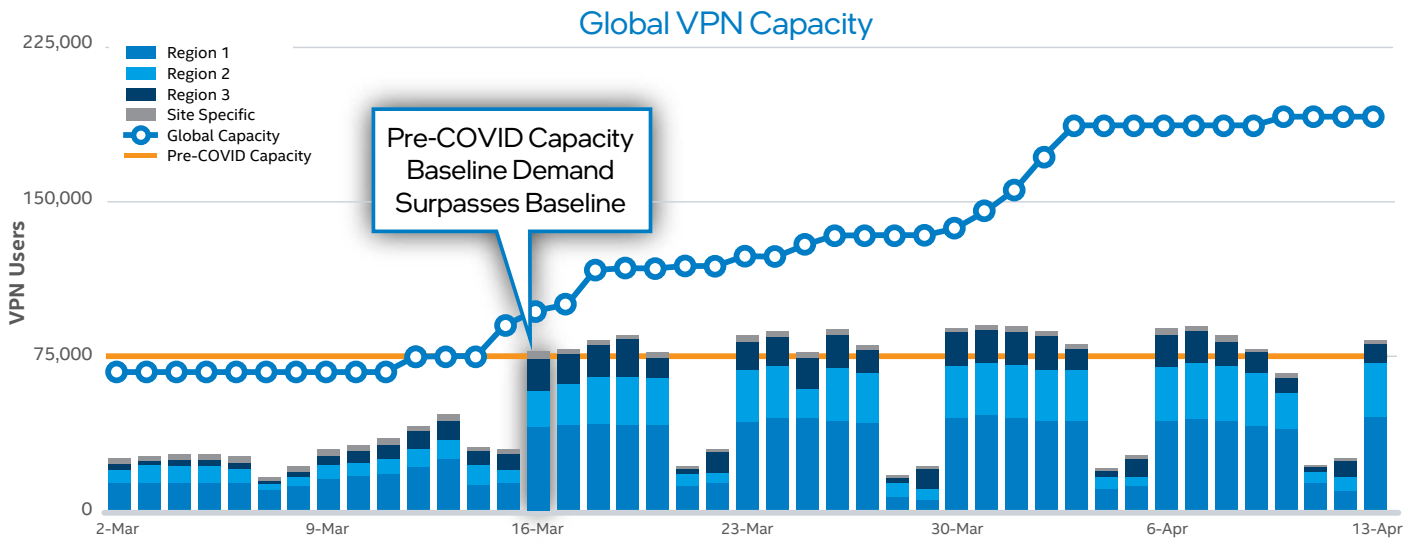

Figure 2. Intel's multicloud virtual VPN gateway strategy enabled rapid scaling and offers a compelling example for other challenged companies to follow.

We divided our solution strategy into four stages.

First, we started with ramping up to meet expected VPN demand growth. Next, we optimized VPN gateway host systems for greater performance and efficiency. Part of this involved working with our VPN vendor and Intel's Network Product Group. Even though SR-IOV had been optimized on this vendor's solutions in the CSP marketplace for years, there was a flaw in its NIC implementation that had gone undiscovered until we worked with the VPN vendor to enable SR-IOV in our private cloud with their VPN gateway solution. Our team helped to devise issue fixes, which went immediately into deployment across several geographies. In fact, the VPN vendor confided that Intel IT was two to three days ahead of several other clients who were pursuing similar VPN objectives, and our fixes would imminently enable their own business continuity needs. The lessons we learned quickly rippled across the cloud and server hardware ecosystem. In addition, during this "VPN Optimization" stage, we discovered multiple issues with SR-IOV enablement in our private cloud virtual VPN gateways. We addressed these and worked on tuning the solution for optimal results.

We discovered that a combination of virtual VPN gateway systems based on SR-IOV, the Data Plane Development Kit (for accelerating packet processing workloads), and Intel Xeon Scalable processors (which contain hardware-based acceleration for the demanding encryption that underlies VPN sessions) enabled us to scale our virtual VPN gateways by 100 to 200 percent of the performance level seen in some dedicated hardware VPN gateway appliances. We are still evaluating the maintenance overhead presented by introducing the virtualization layer and will be able to assess it over time. In any case, the benefit of immediate scaling up proved itself and enabled Intel to transition to a company-wide remote work model with extremely short notice.

Digging deeper, we discovered that SR-IOV gave 8x the network I/O bandwidth of the same virtual VPN gateway solution deployed on the same system with non-SR-IOV NICs, as shown in Figure 3.[3] This increase enabled an 8x increase in network bandwidth and twice the number of VPN users. Under highly loaded conditions, we noticed that users were experiencing high network latency on the order of 40 to 50 ms. Thorough investigation revealed that the latency was introduced by high I/O operations. Subsequently, we worked with our partners to get a suitable bugfix for SR-IOV enabled across these servers. This helped us normalize the latency introduced by the VPN gateways. In fact, with the SR-IOV solution, we were able to scale performance beyond 600 Mbps per VPN server. Without SR-IOV, these servers were getting bottlenecked due to virtual appliance CPU load, which in turn limited scaling. The performance improvements enabled by SR-IOV were and remain significant enough to allow considerable reduction in the physical footprint of the systems providing VPN gateway functionality. One general-purpose server equipped with an Intel Xeon Scalable processor and an SR-IOV NIC could replace multiple VPN appliances, thus saving power and reducing rack space needs.

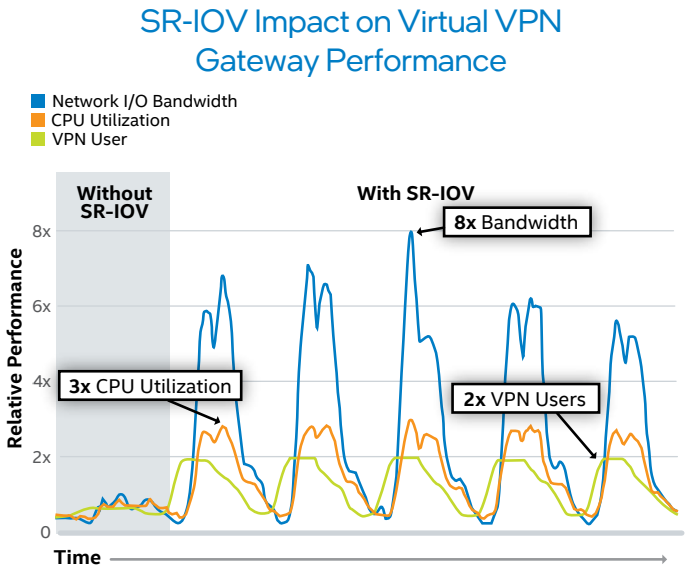

Figure 3. Intel IT testing of on-premises virtual VPN gateway systems show clear benefits from SR-IOV enablement. Note that we capped the gateway at 25 percent CPU utilization for load distribution across multiple gateways.

Achieving higher throughput from fewer boxes helped us in many other ways. We conserved Internet IP addressing and used IP address space from Intel's current allocation. Expanding IP address space requires engagement with Internet service providers, and new IP subnet advertisements could take several days if not weeks in some geographies. Using fewer rack boxes meant conserving data center space and lowering power requirements, as well. Intel Xeon Scalable processors enabled us to have fewer servers with many instances of virtual gateways to meet demand surge in very cost-effective and timely manner.

Since VPN connectivity was the primary mode of connecting to Intel resources, in the third stage, we focused on cross-site VPN capacity and resiliency so that we could sustain failure of all VPN gateway or ISPs at one site and still have sufficient VPN and ISP capacity at another site within the same geographic region. And finally, in the fourth stage, we analyzed traffic and usage to optimize costs across public and private cloud resources. As expected, after peak demand needs were met and local infrastructure was in place, we brought all public cloud virtualized VPN gateway functionality back on premises. Meanwhile, our efforts in equipping Intel's CSP to provide on-demand VPN gateway services remains in place, ready to accommodate future bursting if needed.

Conclusion

Intel created its Pandemic Leadership Team in 2002 in response to SARS and H1N1. That group was and remains instrumental in guiding Intel's globally coordinated health and business responses, within which VPN capability is a key component. As enterprises continue to struggle with expanding their VPN capabilities in the face of ongoing remote work demands, our recent work with VPN scaling showcases the benefits of Intel Xeon Scalable processors and SR-IOV and their potential for all companies to realize similar benefits and flexibility.

At the same time, this use case illustrates the benefits of a multicloud strategy for enterprises. The ability to quickly flow scaling demand out to the public cloud and then back to private infrastructure as conditions and cost pressures dictate was critical to Intel's weathering the pandemic's spread. And note that our solution continues to evolve. After the initial COVID-19 rush, we made changes to our proxy infrastructure and other optimizations that make having a multicloud strategy even more advantageous and will help ensure smooth transitions between platforms in the future.

Our virtual VPN gateway solutions validate the applicability of virtual appliances over purpose-built appliances in at least one case, and it's likely that others will follow. For example, we are now considering a similar approach to using off-the-shelf servers with SR-IOV to improve on radio access network solutions, which could become a major benefit to providers and enterprises rolling out 5G implementations throughout this decade.

No matter what happens with COVID-19 in the coming months, this virus will not be the world's last pandemic. The need for workforces of all sizes to transition work modes and venues overnight could arise again at any time. Our successful navigation of the COVID-19 transition with Intel Xeon Scalable processors (and their instruction set for encryption processing), SR-IOV, and multicloud point the way toward greater enterprise agility, efficiency, and scalability for whatever challenges await Intel.

Related Content

If you liked this paper, you may also be interested in these related white papers:

- Data Center Strategy Leading Intel's Business Transformation
- Security Architecture Enables Intel's Digital Transformation
- New Network Paradigm for Multi-Cloud Enterprise Paper
- Securing the Cloud for Enterprise Workloads-the Journey Continues

For more information on Intel IT best practices, visit intel.com/IT

IT@Intel

We connect IT professionals with their IT peers inside Intel. Our IT department solves some of today's most demanding and complex technology issues, and we want to share these lessons directly with our fellow IT professionals in an open peer-to-peer forum.

Our goal is simple: improve efficiency throughout the organization and enhance the business value of IT investments.

Follow us and join the conversation:

- Twitter
- #IntelIT
- LinkedIn
- IT Peer Network

Visit us today at intel.com/IT or contact your local Intel representative if you would like to learn more.

Acronyms

CSP	cloud service provider
ISP	Internet service provider
NIC	network interface controller (or card)
SR-IOV	Single Root I/O Virtualization
VPN	virtual private network

[1] Intel, "PCI-SIG SR-IOV Primer An Introduction to SR-IOV Technology," intel.com/content/www/us/en/pci-express/pci-sig-sr-iov-primer-sr-iov-technology-paper.html

[2] Scott's Weblog, "What is SR-IOV?" https://blog.scottlowe.org/2009/12/02/what-is-sr-iov

[3] Information on private cloud virtual VPN gateway servers: custom 2nd generation Intel® Xeon® Scalable Processors (Cascade Lake) with a sustained all-core Turbo frequency of 3.6 GHz and single-core Turbo frequency of up to 3.9 GHz or 1st generation Intel Xeon Platinum 8000 processor series (Skylake-SP) with a sustained all-core Turbo frequency of up to 3.4 GHz and single-core Turbo frequency of up to 3.5 GHz; Intel® AVX, Intel AVX2, Intel AVX-512, Intel® Turbo; EBS Optimized; Enhanced Networking. Also, used a combination of Skylake (mostly Intel Xeon Gold 5120 processor, some 5115 and 6126) systems and Broadwell E5 systems. Due to COVID-19, we had to repurpose systems where we could. We deployed third-party virtual VPN appliances in both cases.

 1120/WWES/KC/PDF 344645-001US