Technology Guide

Intel® QuickAssist Technology (Intel® QAT) - Accelerating HAProxy* Performance on 4th Gen Intel® Xeon® Scalable Processors

Authors

Divya Pendyala

Joel Schuetze

Stanislas Odinot

Intel Corporation

Willy Tarreau

Emeric Brun

HAProxy Technologies

1 Introduction

Businesses today are facing an ever-increasing demand for network and server performance driven by the increasing number of mobile users and the applications migrating to cloud data centers and edge.

Industry trends are driving the demand for infrastructure scale, increased performance and operational efficiency, ease of management and enhanced security to handle sensitive and mission-critical workloads and data.

As more traffic moves to the cloud and edge, the complexity of modern enterprise networking and security calls for advanced and high-performance load balancers to provide a flexible, secure, and reliable network to deliver business results.

Load balancers handle various form of layer 4 and layer 7 encryption. SSL/TLS processing is an example of compute resource intensive workload and is a driver for enterprises consuming more CPU resources. Performance scalability per additional compute resources provided is never linear due to contention everywhere in the system, resulting in the required number of cores growing faster than the expected performance level. In addition, many users try as much as possible to have load balancer hosted on one server system and prefer not to use more than one active machine for their load balancers.

HAProxy* is an open-source software load balancer and reliable reverse proxy, continuously optimized to run on Intel® Xeon® processors including 4th Gen Intel® Xeon® Scalable processors with built-in Intel® QuickAssist Technology (Intel® QAT) accelerators. This processor delivers a highly optimized and cost-effective solution with real-time data and robust network security addressing the above-mentioned load balancer systems challenges.

This paper explains the integration of Intel QAT on 4th Gen Intel Xeon Scalable processors that provides a powerful solution for HAProxy so that it can address the growing needs of secure load balancing performance, efficiencies, and effectiveness.

This document is part of the Network & Edge Platform.

Table of Contents

Figures

Tables

Document Revision History

Revision	Date	Description
001	June 2023	Initial release.

1.1 Terminology

Table 1. Terminology

Abbreviation	Description
CDN	Content Delivery Network
CPS	Connection per second
CPU	Central Processing Unit
DoS	Denial-of-service
FMA	Fused-multiply add
SSL	Secure Socket Layer
TCO	Total Cost of Ownership
TLS	Transport Layer Security

1.2 Reference Documentation

Table 2. Reference Documents

Reference	Source
HAProxy	HAProxy.com
4th Gen Intel® Xeon® Scalable Processors	https://www.intel.com/content/www/us/en/products/details/processors/xeon/scalable.html
Intel® QuickAssist Technology (Intel® QAT)	https://www.intel.com/content/www/us/en/architecture-and-technology/intel-quick-assist-technology-overview.html
Intel® Quick Assist Technology Engine for OpenSSL* (Intel® QAT Engine for OpenSSL*)	https://www.intel.com/content/www/us/en/developer/articles/guide/building-software-acceleration-features-in-the-intel-qat-engine-for-openssl.html
Intel® Ethernet 800 Series Network Adapters	https://www.intel.com/content/www/us/en/products/details/ethernet/800-network-adapters.html
Intel® Network Builders	https://networkbuilders.intel.com/
GitHub for Intel QuickAssist Technology Engine	https://github.com/intel/QAT_Engine
GitHub for Intel® Integrated Performance Primitives Cryptography (Intel® IPP Cryptography)	https://github.com/intel/ipp-crypto
GitHub for Intel® Multi-Buffer Crypto for IPSec Library	https://github.com/intel/intel-ipsec-mb

2 Overview

2.1 Industry Leading Load Balancer by HAProxy Technologies

HAProxy is a widely used open-source software load balancer, whose development is primarily backed by HAProxy Technologies. It is particularly suited for very high traffic web sites and powers a significant portion of the world's most visited sites. It offers high availability, load balancing, traffic offloading, and a broad spectrum of proxy-based features aiming at optimizing application delivery and keeping applications responsive under heavy loads.

HAProxy has a robust design and has evolved significantly to meet the changing needs of modern applications while taking advantage of the evolution of operating systems and processors. Being an application layer reverse-proxy deployed at the edge, its role is to make routing decisions based on traffic analysis, which requires it to take care of data security and integrity. This represents heavy processing concentrated inside a single component. It is made possible due to a highly optimized, extremely low-latency event-driven architecture that takes advantage of systems with multi-core processors by distributing non-blocking operations to the least busy CPU core in a system and sharing as little as possible between them to keep caches hot. These careful design choices have resulted in the best possible combination of both low latency and high throughput, with an extreme robustness.

For reliability, the software features high observability of data processing, and has been developed with protections against software malfunction conditions such as impossible conditions, endless loops, etc. The software checks for these conditions and

will gracefully crash providing information on the problem without corrupting data or leading to lengthy outages, which makes HAProxy an excellent solution for various mission-critical network security workloads.

HAProxy Technologies is a Titanium member of Intel® Network Builders program delivering high-performance load balancing solutions on Intel® platforms.

2.2 Load Balancing Use Cases that Benefit from 4th Gen Intel® Xeon® Scalable Processors and Intel® QuickAssist Technology

The following use cases highlight how HAProxy delivers load balancing performance with the latest Intel® Xeon® Scalable processors:

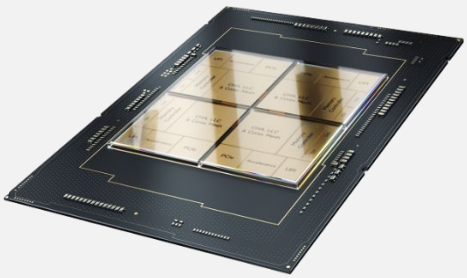
- Content Delivery Network/Web Server: Minimize latency and accelerate asset delivery speed for content delivery networks (CDNs) and web servers. HAProxy on 4th Gen Intel Xeon Scalable processors with Intel QAT improves cryptography and data compression performance allowing each core to serve more clients so that CDNs can deliver content as quickly as possible. In addition, by significantly reducing the SSL processing overhead, the infrastructure better handles fast traffic variations resulting from world-wide events or denial-of-service (DoS) attacks.
- Web Applications: Applications running on 4th Gen Intel Xeon Scalable processors will handle the same load using fewer CPU cores thanks to the built-in Intel QAT accelerators that speed up cryptography, data compression, and data movement. This results in the need for less and/or smaller containers or virtual machines for the same task, lowering operation costs.
- Network Security Appliance: Network security appliances must keep ahead of current trends so that customers can protect their investment. The high performance of 4th Gen Intel Xeon Scalable processors with its new instructions, faster DDR5 memory, PCIe 5.0, built-in accelerators speeding up AI, encryption, and compression is a desirable solution for network security appliance designs that will be extensible in field for several years.

2.3 Performance Provided by 4th Gen Intel Xeon Scalable processor

The 4th Gen Intel Xeon Scalable processor is designed to accelerate performance across the most demanding workloads. The new processor has integrated the most built-in accelerators[1] of any CPU on the market to help maximize performance efficiency for emerging workloads, including those powered by AI. Advanced data security technologies help protect data in an ever-changing landscape of threats while unlocking new opportunities for business collaboration and insights. Together with its ecosystem of partners, Intel makes it easier for enterprises to stay competitive, offering the most choices to scale infrastructure and achieve business value quickly.

Network-optimized 4th Gen Intel Xeon Scalable processors are the next step in accelerating load balancing workloads while increasing energy efficiency, with a high-throughput, low-latency platform engineered for data centers, network core, and scalable to the edge for on-prem or cloud deployments.

Purpose Built for Accelerated Network Workloads



4th Gen Intel Xeon Scalable processors[2] offer:

- **Advanced execution resources** in a range of core counts and feature sets, with improved per-core performance compared to the previous generation, enhanced by the most built-in accelerators in the industry.
- **Enhanced memory subsystem** with up to eight DDR5 channels operating at up to 4800 MT/s, a 1.5x improvement in memory bandwidth and speed compared to the predecessor platform.
- **Faster, higher capacity I/O** based on up to 80 lanes of PCIe 5.0 per socket, compared to 64 lanes of PCIe 4.0 per socket in the prior generation.

Servers built with 4th Gen Intel Xeon Scalable processors further enhance the scalability and performance of software-defined infrastructure with an enhanced instruction set architecture (ISA). Intel® Ethernet 800 Series Network Adapters complement Intel® architecture processors for load balancer deployments, with enhanced timing synchronization that helps prevent service disruptions. For cloud deployments, optimized acceleration software packages enhance the performance seamlessly.

[1] https://www.intel.com/content/www/us/en/products/details/processors/xeon/scalable.html

[2] All 4th Gen Intel Xeon Scalable processor benchmarks are here: https://edc.intel.com/content/www/us/en/products/performance/benchmarks/overview/-SPR

2.4 Built-in Accelerators Improve Performance

Integration of accelerators into the processor redefines CPU architecture. Using accelerators provides a more efficient way to achieve higher performance than relying solely on increasing the CPU core count for workload processing.

With all-new accelerated matrix multiply operations, 4th Gen Intel Xeon Scalable processors have exceptional AI training and inference performance. Other seamlessly integrated accelerators speed up data movement and compression for faster networking, boost query throughput for more responsive analytics, and offload scheduling and queue management to dynamically balance loads across multiple cores. To enable new built-in accelerator features, Intel supports the ecosystem with OS-level software, libraries, and APIs. This architecture allows users to run cloud and networking workloads using fewer cores with faster cryptography.

2.5 Technology Description

2.5.1 Intel® QuickAssist Technology (Intel® QAT)

Intel QAT (see Figure 1), is one of the built-in accelerators and is designed for speeding up cryptography including private key protection, and data decompression. It performs fixed-function acceleration for asymmetric and symmetric encryption, hashing, lossless decompression and key management. It also enables fusing of multiple operations to improve latency and memory bandwidth.

Intel QAT is designed to speed up these common application and infrastructure functions, while also freeing up a significant number of CPU cores. Because it offloads cryptography functions from the processor, the accelerator helps systems serve a larger number of clients.

Unlike previous generation Intel Xeon processors, Intel QAT acceleration comes as a built-in feature in specific SKUs of the 4th Gen Intel Xeon Scalable processor. This enables customers to use the acceleration technology for cryptographic workloads with no additional physical components.

2.5.2 Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

Intel® Advanced Vector Extensions 512 (Intel® AVX-512) helps accelerate the performance of scientific simulations, financial analytics, AI/deep learning, 3D modeling and analysis, image and audio/video processing, cryptography, data compression, and other intensive workloads. Intel AVX-512 is the latest Intel architecture processor vector instruction set, with up to two fused-multiply add (FMA) units and other optimizations to help accelerate the performance of demanding computational tasks.

These vectorized instructions used in software acceleration libraries, namely Intel® IPP Cryptography and Intel® Multi-Buffer Crypto for IPsec with Intel QAT Engine for OpenSSL, provide the computational power for HAProxy.

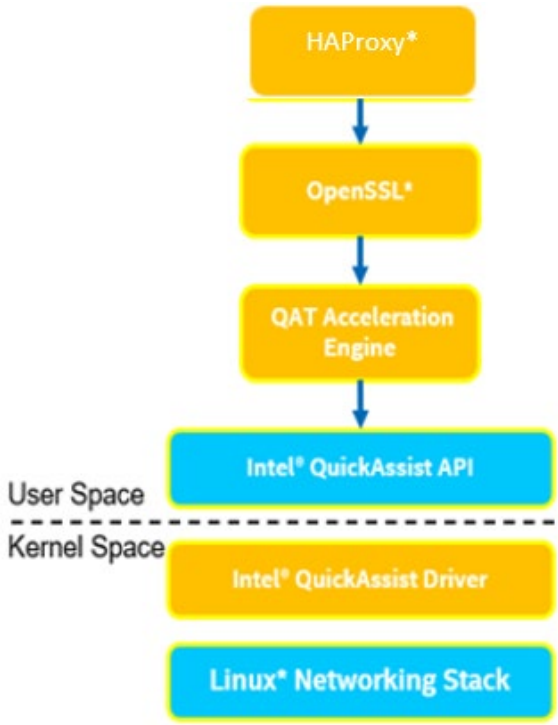
2.5.3 Intel® QuickAssist Technology Engine for OpenSSL* (Intel® QAT Engine for OpenSSL*)

Intel® QAT Engine for OpenSSL* is a software package that supports acceleration for both hardware and optimized software based on vectorized instructions. The advancement in cryptographic acceleration provides users more options to accelerate their workloads. The Intel QAT Engine for OpenSSL supports the ability to accelerate the standard OpenSSL using basic Intel instruction set to either the hardware acceleration path (via the Intel QAT hardware (qat_hw) path) or via the optimized software path (qat_sw lib).

The Intel QuickAssist Technology accelerator is accessed through a device driver in kernel space and a library in user space. Cryptographic services are provided to OpenSSL through the standard engine (provider in OpenSSL-3.0) framework. This engine (Figure 1) builds on top of the user space library, interfacing with the Intel QAT API, which allows it to be used across Intel QAT generations without modification. This layering and integration into the OpenSSL framework allow for seamless usage by applications.

HAProxy makes use of the Intel QAT Engine for OpenSSL. Applications such as HAProxy interface to OpenSSL, a toolkit for TLS/SSL protocols that includes a modular system to plugin device-specific engines. With this implementation, HAProxy can be deployed on-prem or cloud for network security workloads using fewer cores with faster cryptography allowing each core to serve more clients.

HAProxy, using Intel QAT crypto engine, reduces the number of cores needed for the software depending on the implementation used Intel QAT hardware versus Intel QAT software optimizations. For users, this could mean faster crypto processing and more cores available to run more applications.



HAProxy: https://github.com/haproxy/haproxy/

OpenSSL: https://github.com/openssl/openssl

QAT Acceleration Engine: https://github.com/intel/QAT_Engine

QuickAssist Driver : https://www.intel.com/content/www/us/en/developer/topic-technology/open/quick-assist-technology/overview.html

Figure 1. Intel QuickAssist Technology Stack Diagram

3 Benchmarks and Key Performance Indicators

3.1 Connections per Second (CPS)

Clients send HTTPS connection requests without requesting data. This will utilize Key Exchange + Certificate Authentication exercising the TLS-1.2 handshake only with no data transfer.

3.2 Test Setup

For this gen-gen test Intel® Xeon® Platinum 8470N, 4th Gen Intel Xeon Scalable processor and Intel® Xeon® Gold 6338N, 3rd Gen Intel® Xeon® Scalable Processor are used. The server with Intel Xeon Platinum 8470N is interconnected to a Cisco Nexus C9336C switch with a total possible aggregated link bandwidth of 400 Gbps (4x100 GbE links) and Intel Xeon Gold 6338N is connected to Arista DCS-7060CX2 Switch with a total link bandwidth of 100 Gbps (4x100 GbE links).

Table 3. Hardware Configuration

Component	Config 1 – 4th Gen Intel Xeon Scalable Processor	Config 2 – 3rd Gen Intel Xeon Scalable Processor
CPU Model	Intel Xeon Platinum 8470N	Intel Xeon Gold 6338N
Memory	8*16 GB @4800 MT/s DDR5	8*32 GB @2666 MT/s DDR4
Hard Drive	960GB- INTEL SSDSC2KG96	500GB- KINGSTON SA400M8
Ethernet Adapter	2x Intel Ethernet Controller E810-XXV for SFP	2x Intel Ethernet Network Adapter E810-CQDA2 for QSFP
Intel QAT Hardware	Intel QAT Gen 4	Intel QAT Gen 2
Intel Turbo Boost Technology	Enabled	Enabled
Base Frequency	1.7 GHz	2.2 GHz
All-core Maximum Frequency	2.7 GHz	3.5 GHz
Maximum Turbo Frequency	3.6 GHz	3.5 GHz
Microcode	0x2b000181	0xd000375
Test by Intel as of	03/21/2023	03/21/2023

Table 4. Software Configuration

Component	Config 1 – 4th Gen Intel Xeon Scalable Processor	Config 2 – 3rd Gen Intel Xeon Scalable Processor
Operating System	Ubuntu 22.04.1 LTS	Ubuntu 22.04.1 LTS
Kernel	5.15.0-52-generic	5.15.0-52-generic
HAProxy	2.7.0	2.7.0
OpenSSL	1.1.1k	1.1.1k
Intel® QAT driver	QAT20.L.1.0.10-00005.tar.gz	QAT.L.4.19.0-00005
Intel® QAT Engine for OpenSSL	v0.6.19	v0.6.19
Intel® Multi-Buffer Crypto for IPsec Library	V1.3	V1.3
Intel® Integrated Performance Primitives Cryptography (Intel® IPP Cryptography)	IPP Crypto 2021.6	IPP Crypto 2021.6

For each dual-node client, two sets of 2000 outstanding connection requests, each task set to the cores of each node are created with h1load, HAProxy's own implementation of load generator, on the client. Additionally, tested with ApacheBench The requests run both continuously and simultaneously for 300 seconds. Two Client machines are used to ensure that there are no limitations from the Client side. Each Client process establishes a secure connection, exits gracefully, and sends a new request to establish a secure connection. At the end of the 300- Second run, the Connections per Second from each process is summed up and reported on the screen. One haproxy process is allocated for every hyper-thread (nbthread) in the HAProxy configuration file (haproxy.conf). Example: 2C4T (2 core, 4 thread config) has 4 haproxy threads. Each test is mapped to CPU cores and hyperthreads for consistency of results. Example – 2C4T: nbthread 4 and cpu-map 1/all 0-1,104-105 are defined in haproxy4t.cfg, HAProxy configuration file.

For more details on Haproxy configuration, refer to: https://www.haproxy.com/blog/the-four-essential-sections-of-an-haproxy-configuration/.

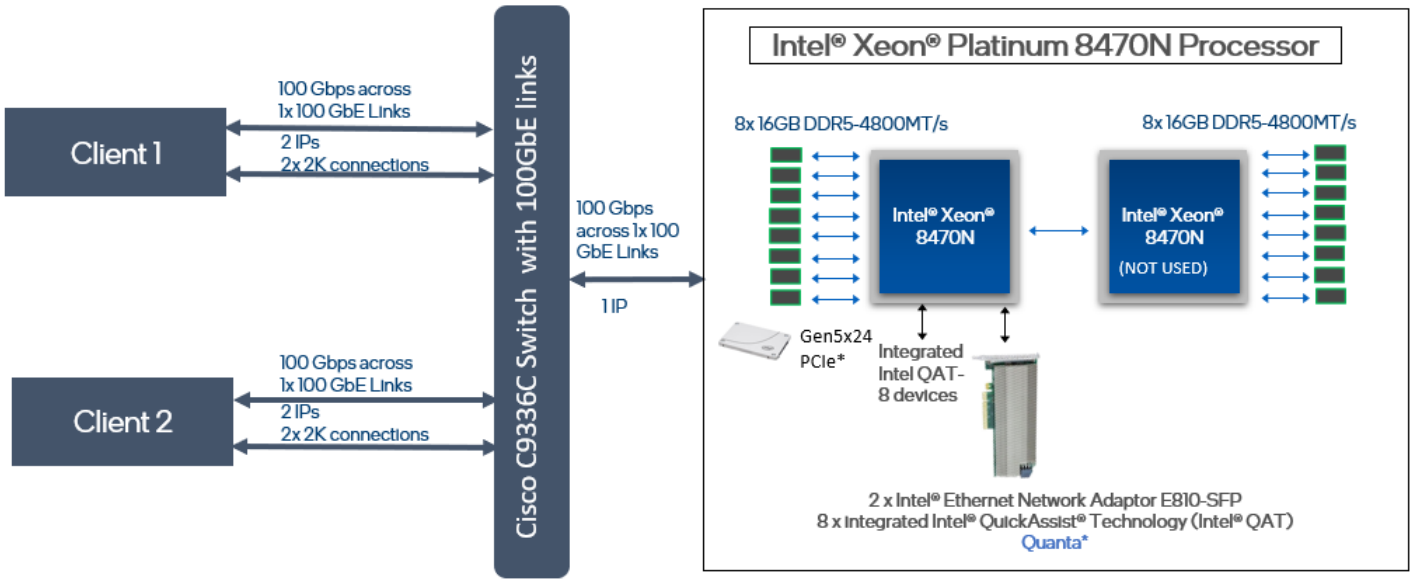


Figure 2. Hardware configuration – 4th Gen Intel Xeon Scalable processor

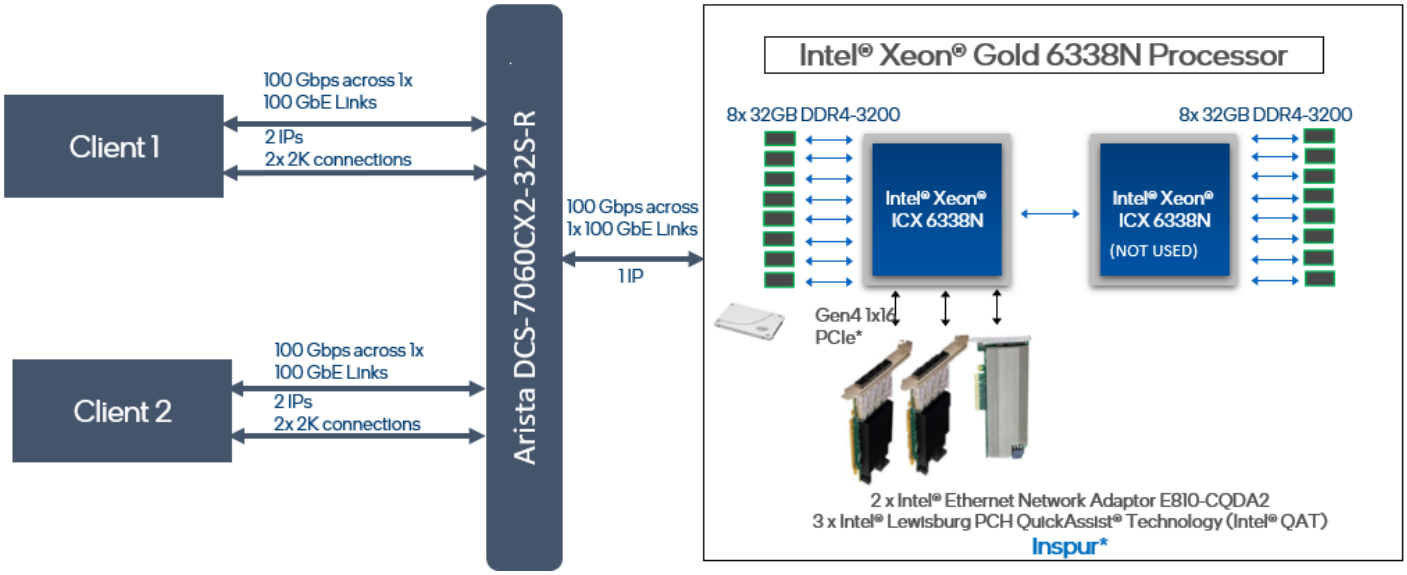


Figure 3. Hardware configuration – 3rd Gen Intel Xeon Scalable processor

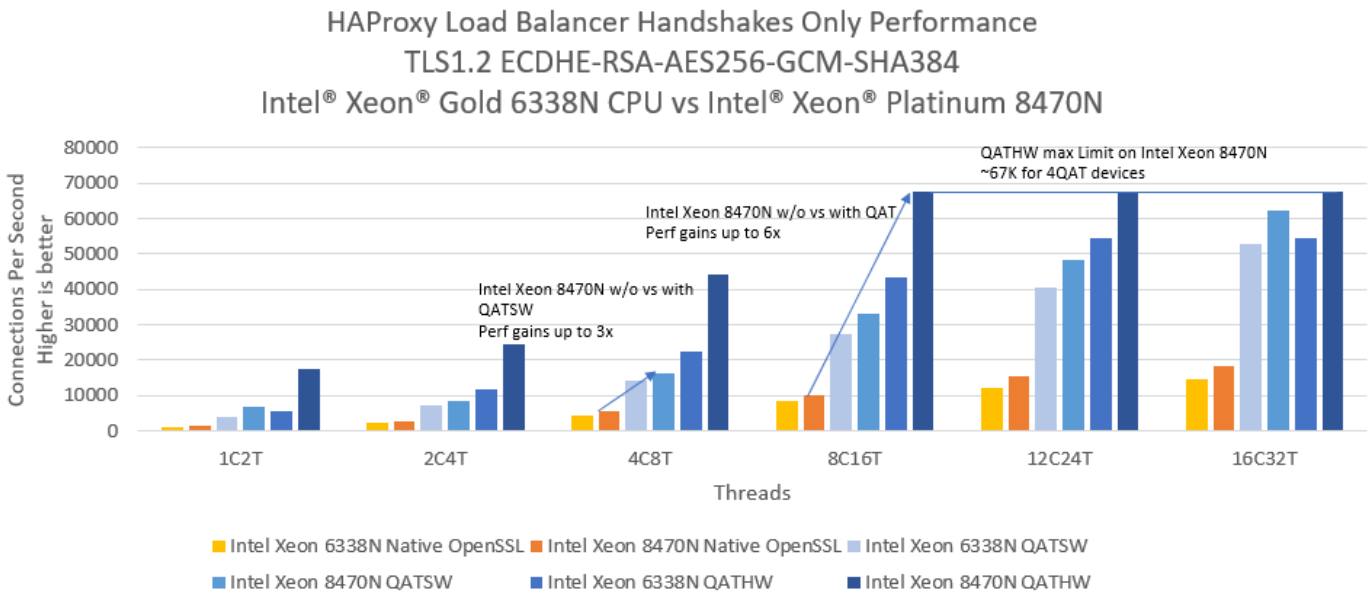


Figure 4. HAProxy Load Balancer Handshake Performance

As Figure 4 demonstrates, offloading ECDHE-X509-RSA2K to Intel QAT provides the highest performance per core. That is, Intel QAT provides up to 6x higher performance than default software stack, native OpenSSL and up to 3x higher using the optimized crypto software stack. It must be noted that Intel QAT has a peak performance, and it will be hitting this point for ECDHE-X509-RSA2K by 8 cores. Once Intel QAT reaches its maximum performance per device, its performance will plateau.

As shown in Figure 4 ,the gen-gen performance benchmarking at 4 cores and 8 threads, the 4th Gen Intel Xeon Scalable processor performs approximately 20 percent better than the 3rd Gen Intel Xeon Scalable processor with native OpenSSL, ~20% better using Optimized crypto software and 2x times better with Intel QAT.

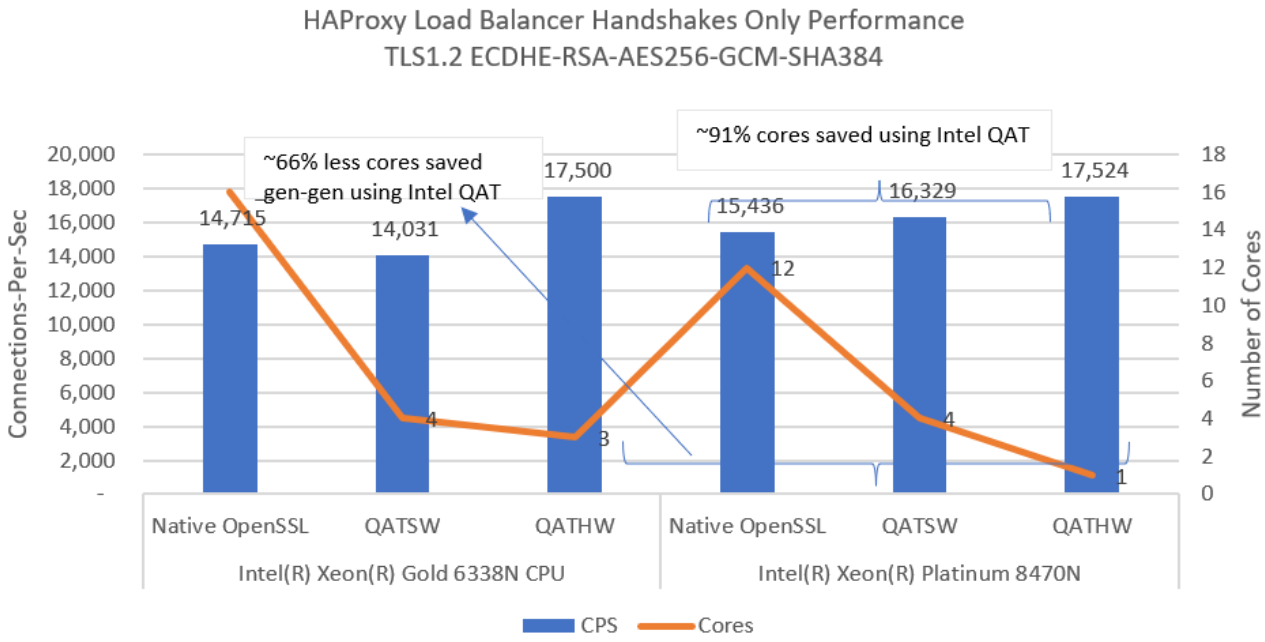


Figure 5. Intel QAT provides core savings for HAProxy handshakes

The 4th Gen Intel Xeon Scalable processor provides significant performance gains per core, and TCO benefits, when utilizing Intel QAT for establishing the TLS connection.

As shown in Figure 5, for HAProxy to achieve ~15k CPS with default OpenSSL, the 4th Gen Intel Xeon Scalable processor requires 25% less cores than that of the 3rd Gen Intel Xeon Scalable processor. By using optimized crypto software, it is ~33% less core utilization. By using integrated Intel QAT, ~66% less cores are used compared to the 3rd Gen Intel Xeon Scalable processor.

Crypto acceleration with integrated Intel QAT in the 4th Gen Intel Xeon Scalable processor, significantly simplifies the hardware configuration unlike when using Intel QAT as an add-in PCIe card on the 3rd Gen Intel Xeon Scalable processor.

4 Benefits

Intel QAT, now built into 4th Gen Intel Xeon Scalable processors, accelerates cryptography and compression. Intel QAT can significantly boost CPU efficiency and application throughput while reducing data footprint and power utilization, thereby enabling organizations to strengthen encryption without sacrificing performance.

Performance is delivered using a mainstream release of OpenSSL and HAProxy for asynchronous processing of TLS handshake operations, plugging in accelerated cryptographic implementations through a standard provider/engine into OpenSSL.

HAProxy load balancer delivers up to 6x better crypto performance using Intel QAT on 4th Gen Intel Xeon Scalable processor compared to the native OpenSSL stack.

By using optimized crypto instructions on this latest platform, the performance improves by up to 3 times than the native OpenSSL stack.

5 Summary

HAProxy has incorporated features and capabilities introduced in 4th Gen Intel Xeon Scalable processors including built-in Intel QAT. The new capabilities deliver better performance using fewer cores for the same workload, reducing contention, and leaving more room for application layer processing and improving TCO. With the built-in accelerators available on 4th Gen Intel Xeon Scalable processors, users can get the benefit of higher performance and efficiency across network security workloads at lower costs.

0623/DN/WIT/PDF 778237-001US