INTRODUCTION TO CLOUD SYSTEMS

Lecture 8 – Data processing acceleration – QAT, CBDMA/DSA, Kubernetes Operators.

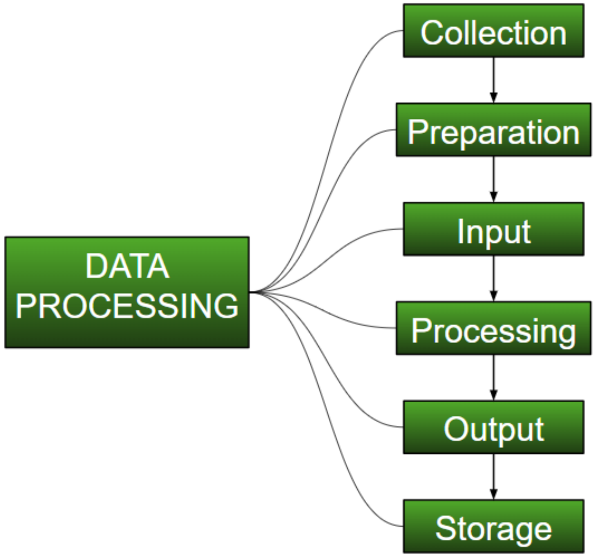
Data processing

Data Processing - Making Information Ready to Use

Data processing

What Is Data Processing?

Data processing occurs after the data collection stage of the data pipeline. In the processing stage, data is prepared for use, then stored in a system that can be accessed by applications and users.



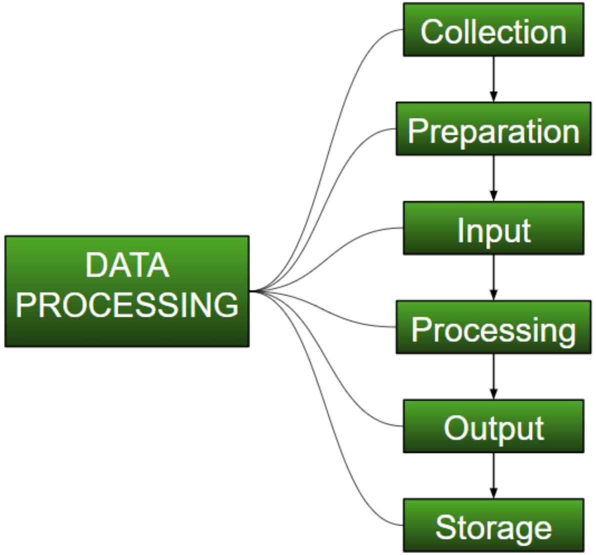
https://www.intel.com/content/www/us/en/analytics/data-processing.html

Data processing

What Is Data Processing?

In order to be analyzed, data must first be processed to ensure that it is clean and high quality. Processing can verify and format data, making it easier to access, query, and store

To achieve insights without delay, organizations must maximize data processing performance and throughput while staying cost effective. Intel® hardware and software technologies work together to accelerate data processing from edge to cloud.



Data processing

Types of Data Processing

There is no one-size-fits-all method for processing data. Different types of workloads and applications require different approaches to make processing performant and cost-effective.

Data processing

Types of Data Processing

Methods for processing data may include:

- **Batch processing**: Batch processing consists of dividing data into groups, or batches, that can be processed as resources become available. During batch processing, batches of data are processed serially, one after another. While batch processing can efficiently process large volumes of data, it typically is best for data that does not require immediate use.

Data processing

Types of Data Processing

Methods for processing data may include:

- **Stream processing**: Stream processing occurs when data is processed continuously as it enters the data pipeline. This type of processing yields faster analysis of smaller amounts of data than batch processing. It typically is used to process data that must be acted upon quickly.

Data processing

Types of Data Processing

Methods for processing data may include:

- **Distributed data processing**: As network technologies have evolved, data processing tasks no longer need to be completed on the same node. With distributed data processing, multiple nodes running in the same cluster work in parallel to process data workloads across a network. Using distributed data processing allows advanced analytics workloads to be processed using lower-cost, lower-power-consumption hardware.

Data processing

Data Processing Technology

As one of the most resource-intensive stages of the data pipeline, data processing efficiency can be significantly impacted by hardware and software optimization.

Today, many leading software vendors optimize their products for Intel® hardware. The Intel® ecosystem of solution and technology partners ensures that many software solutions run best on Intel® hardware and helps customers get the best return on their technology investments.

Data processing

Data Processing Technology

Intel brings a wide-ranging portfolio of hardware and software technologies to accelerate today’s data processing workloads, including:

- **Intel® Xeon® processors**: Offering flexibility that can tackle diverse workloads from many sources, Intel® Xeon® processors include features like Intel® Deep Learning Boost that are optimized for tasks like data normalization and noise reduction for AI processing.

Data processing

Data Processing Technology

Intel brings a wide-ranging portfolio of hardware and software technologies to accelerate today’s data processing workloads, including:

- **Intel® Optane™ SSDs**: Designed for longevity and built to optimize storage and data caching performance, Intel® Optane™ SSDs can help to accelerate streaming and real-time data processing while maintaining high system reliability. applications.

Data processing

Data Processing Technology

Intel brings a wide-ranging portfolio of hardware and software technologies to accelerate today's data processing workloads, including:

- **Open source technologies**: Intel offers a range of open source libraries and platforms that accelerate data processing and analysis, including Intel® oneAPI toolkits, Intel® oneAPI Math Kernel Library (Intel® oneMKL), and Intel® oneAPI Data Analytics Library.

Data processing

Data Processing Technology

Intel brings a wide-ranging portfolio of hardware and software technologies to accelerate today’s data processing workloads, including:

- **Security enhancements**: With Intel® QuickAssist Technology (Intel® QAT), data teams can accelerate encryption and decryption performance to enhance security for data processing applications.

QAT - Intel® QuickAssist Technology

Intel QAT (QuickAssist Technology) can provide extended accelerated encryption and compression services by offloading the actual encryption and compression request(s) to the hardware QuickAssist accelerators, which are more efficient in terms of cost and power than general purpose CPUs for those specific compute-intensive workloads.

https://www.intel.com/content/www/us/en/developer/articles/technical/how-intel-quickassist-technology-accelerates-nfv-use-cases.html

QAT - Intel® QuickAssist Technology

Benefits

There are several benefits to using Intel QAT. For example, it can be scaled by choosing accelerators with different performance characteristics or by employing multiple accelerators in a single platform. Another advantage is the reduction in software development efforts by implementing a consistent set of APIs that can be used across products and over multiple development cycles. The APIs also allow for optional supported features that can be queried at run-time, allowing the same software to run unmodified on different deployed platforms.

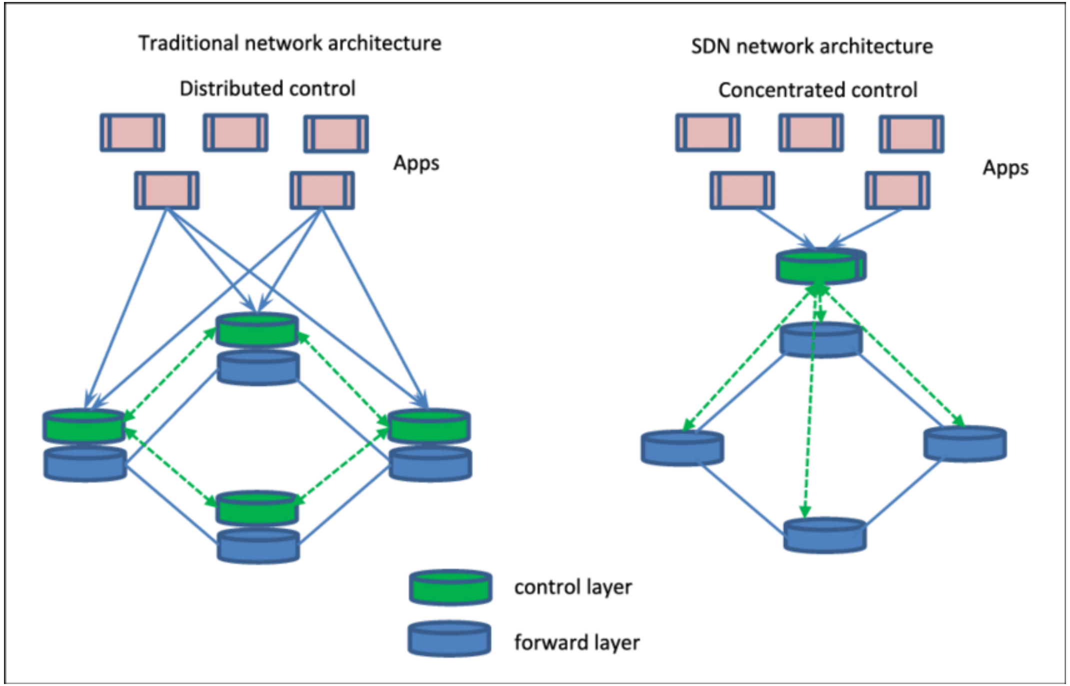
QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions

Hardware-based acceleration services for workloads such as encryption and compression supported by Intel QAT are well suited for use with Software Defined Networking (SDN) and Network Function Virtualization (NFV) implementations on Intel® architecture servers. An accelerator abstraction layer provides a uniform means of communication between applications and accelerators, as well as facilitating management of acceleration resources within the OpenStack* architecture.

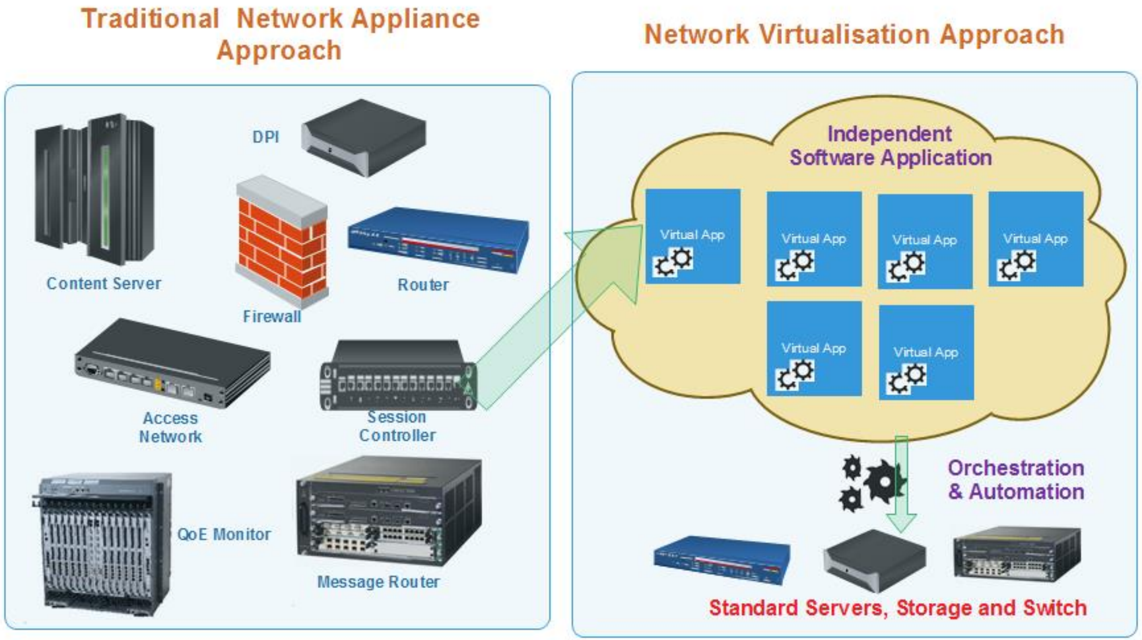
QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions



QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions



QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions

One of the technologies that comes up alongside NFV is SDN or Software-Defined Networking. In many ways, NFV and SDN compliment each other. Let’s take a look at the difference between the two:

- NFV – NFV is used to optimize network services by taking network functions away from hardware. Network functions run at the software level so that provisioning can take place more efficiently.
- SDN – SDN separates the control plane from the forwarding plane and provides a top-down perspective of the network infrastructure. This allows the user to provision network services as they are needed.

QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions

Both of these technologies turn legacy networks on their head in favor of a software-based networking approach. Virtualizing networking services allows resources to be provisioned faster and more efficiently in a way that supports scalability. These two don’t need to be used together but they complement each other in a number of ways.

QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions

For instance, with SDN you can enable network automation to determine where network traffic is sent on. NFV can complement this by allowing you to manage routing controls at the software level. Combining the two allows you to mix automation with software-level routing to create the most efficient service across the network.

QAT - Intel® QuickAssist Technology

Integration with SDN and NFV Solutions

Intel QAT Adapters are available as PCI Express* Gen 3-compliant cards that support functionality such as the following:

- 4G LTE and 5G encryption algorithm offload for mobile gateways and infrastructure.
- VPN traffic acceleration, with up to 50 Gbps crypto throughput and support for IPsec and SSL acceleration.
- Compression/decompression up to 24 Gbps throughput.
- I/O virtualization using PCI-SIG Single-Root I/O Virtualization (SR-IOV).For enabling Intel® Virtualization Technology including SR-IOV with Intel QAT

QAT - Intel® QuickAssist Technology

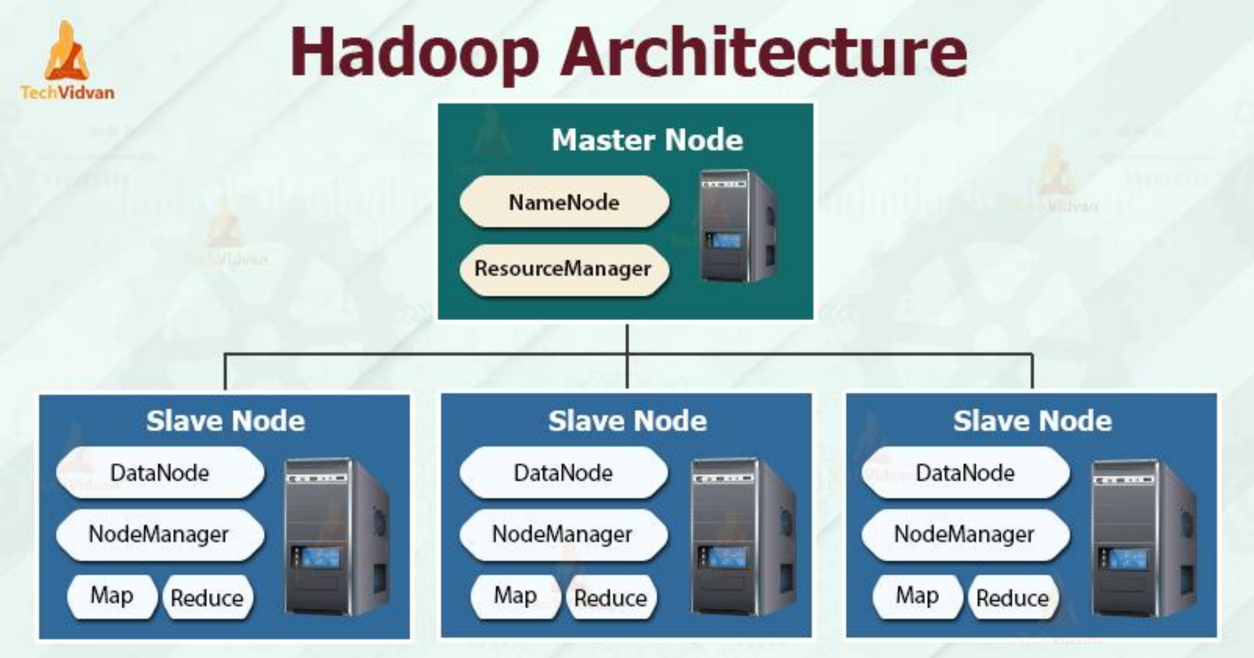
Accelerating Hadoop

Hadoop is a scalable, distributed data storage and analytics engine that can store and analyze massive amounts of unstructured and semi-structured data. The distributed architecture of Apache

Hadoop is comprised of multiple nodes (sometimes thousands), which are individual servers that run an off-the-shelf operating system and the Apache Hadoop software. Input file sizes can be on the order of one terabyte (TB), so moving this data both within and round the server cluster can require a lot of disk and network bandwidth, leading to significant delays.

QAT - Intel® QuickAssist Technology

Accelerating Hadoop

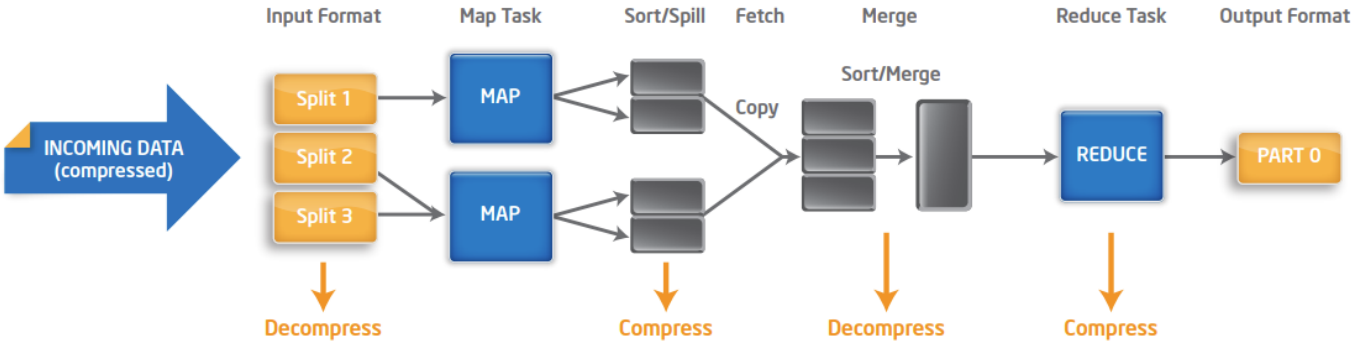


https://techvidvan.com/tutorials/hadoop-architecture/

QAT - Intel® QuickAssist Technology

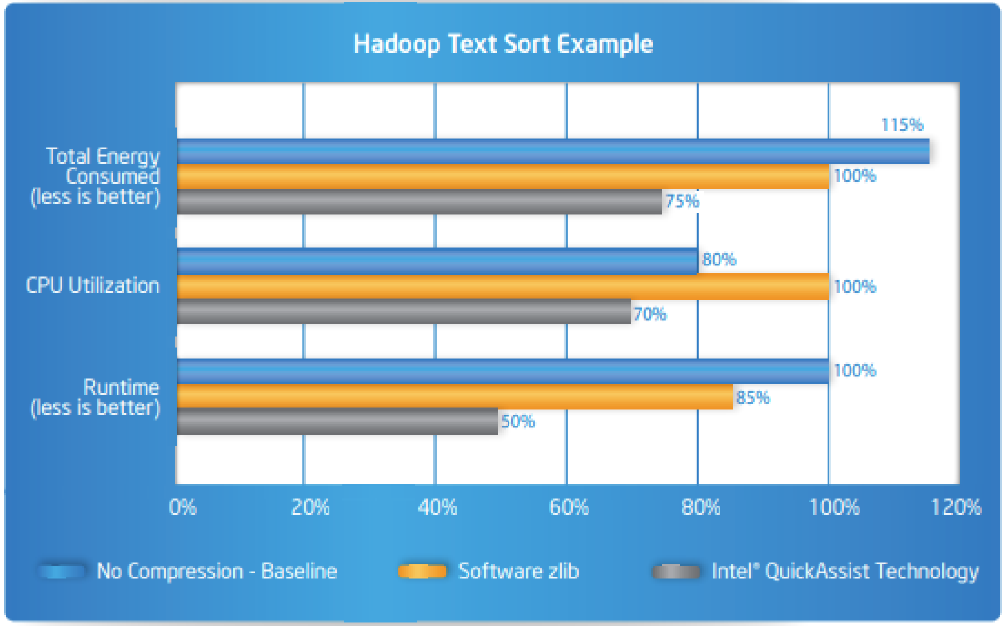
Accelerating Hadoop

One way to reduce the impact of transferring large data volumes is to compress them before they are sent onto the next Hadoop phase. The phases of a Hadoop job are shown in figure below, along with potential points where compression and decompression processing could be inserted to improve performance, and to reduce the load on the network and server I/O.



QAT - Intel® QuickAssist Technology

Accelerating Hadoop



QAT - Intel® QuickAssist Technology

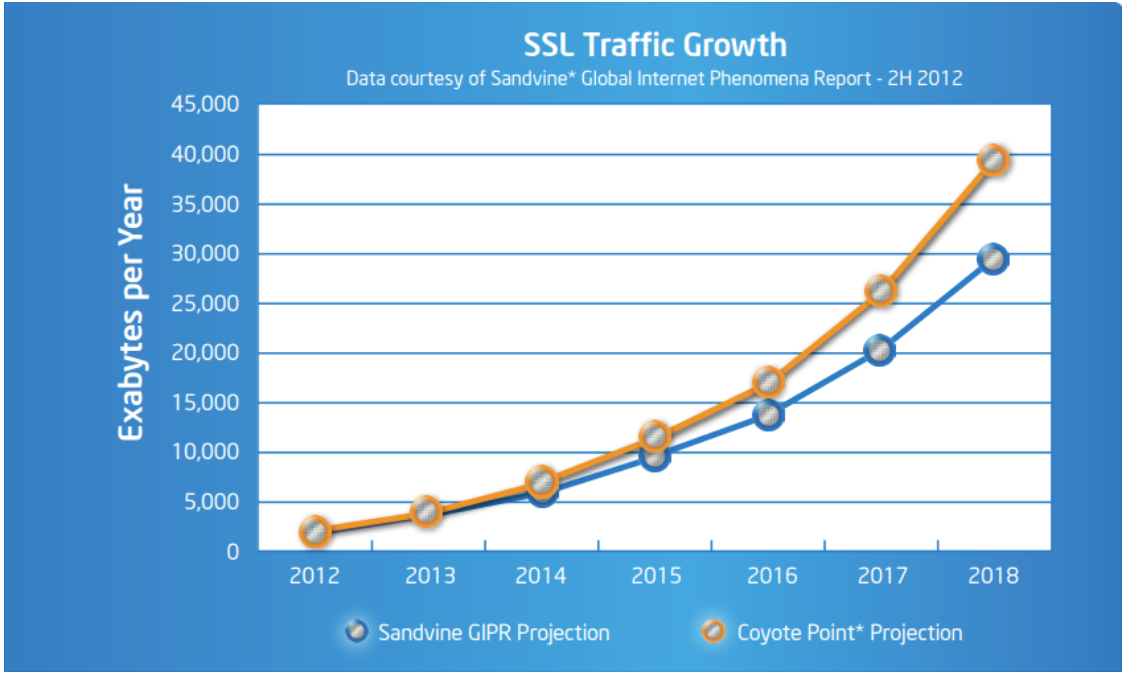
Accelerating OpenSSL

Historically, the demand for secure data transmissions over the Internet was driven primarily by institutions conducting e-commerce and banking transactions. Today, the volume of secured communications is skyrocketing, as personal information of all sorts is being encrypted by applications like Gmail, Twitter, and Facebook using the HTTPS protocol. As a result, servers in data centers, telecom networks, and enterprises are expected to handle increasing amounts of traffic using the Secure Sockets Layer (SSL) protocol, increasing compute requirements. Industry forecasts, shown in figure below, suggest SSL traffic could increase by as much as 16-fold from 2012 to 2018

https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/accelerating-openssl-brief.pdf

QAT - Intel® QuickAssist Technology

Accelerating OpenSSL



https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/accelerating-openssl-brief.pdf

QAT - Intel® QuickAssist Technology

Accelerating OpenSSL

Open Source Optimization - The OpenSSL project provides an open source implementation of the SSL/TLS3 protocols and is a commonly deployed library for SSL/TLS world-wide. The SSL/TLS protocols consist of two phases: an initial session-initiation/handshake and a bulk data transfer. Over time, the implementation has been modified to increase performance, including contributions by Intel that speed up the cryptographic workloads in both phases. OpenSSL performance can be further improved with Intel QuickAssist Technology, which is supported by select Intel architecture platforms. Equipment manufacturers who utilize the current version of OpenSSL will benefit from the optimizations already integrated in the release, as well as the straightforward installation process

QAT - Intel® QuickAssist Technology

Accelerating OpenSSL

Not surprisingly, contributors to OpenSSL continue to work to make the implementation more efficient, and one of these efforts is referred to as Asynchronous OpenSSL, which is expected to be delivered as a patch in a development branch (i.e., parallel development project) with the potential of many-fold performance increases

QAT - Intel® QuickAssist Technology

Accelerating OpenSSL

Asynchronous OpenSSL The standard release of OpenSSL is serial in nature, meaning it handles one connection within one context. From the point of view of cryptographic operations, the release is based on a synchronous/ blocking programming model. A major limitation is throughput can be scaled higher only by adding more threads (i.e., processes) to take advantage of core parallelization, but this will also increase context management overhead. Asynchronous OpenSSL is a non-blocking approach that supports a parallel-processing model at the cryptographic level for SSL/TLS protocols, which in turn allows for other types of optimizations. Although the concept of a non-blocking mode of operation is not new within OpenSSL, it previously had not been applied to cryptographic transforms.

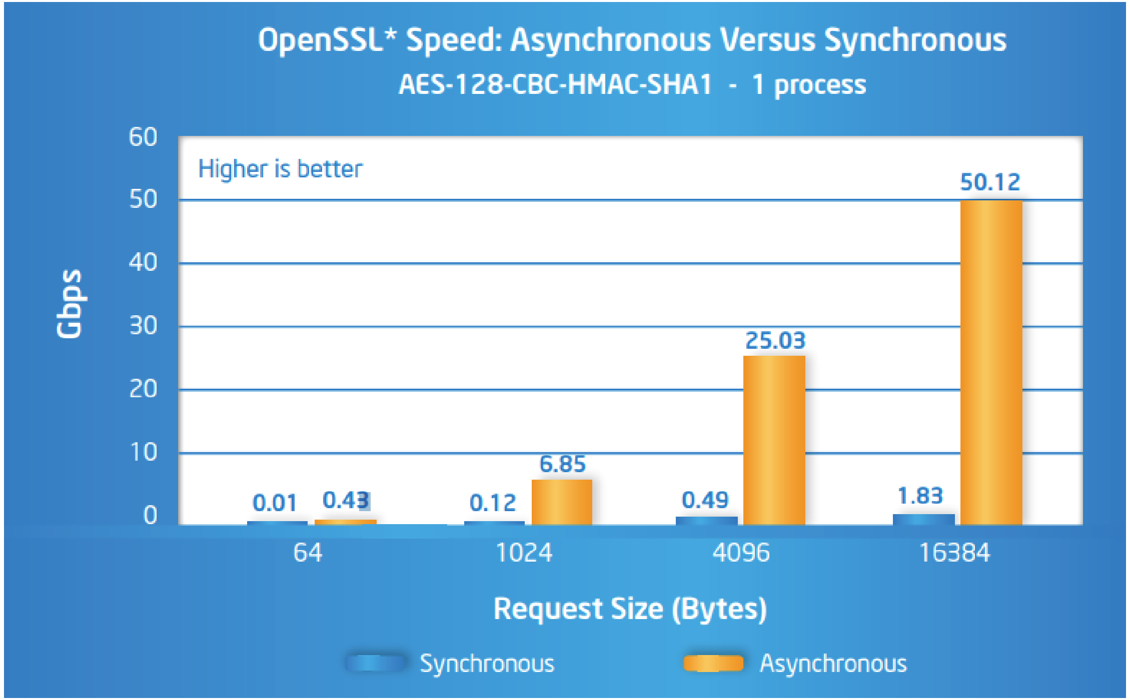
QAT - Intel® QuickAssist Technology

Accelerating OpenSSL

This capability allows cryptographic transforms to be processed on dedicated hardware engines or on separate logical cores, thereby allowing the protocol stack and applications to run other tasks unencumbered. Asynchronous operation also enables single-threaded applications to efficiently handle multiple SSL connections because individual flows are not blocked when using hardware acceleration. Two major benefits of asynchronous OpenSSL are increased single-flow throughput, leading to maximum performance, and fewer contexts, thus reducing context management overhead.

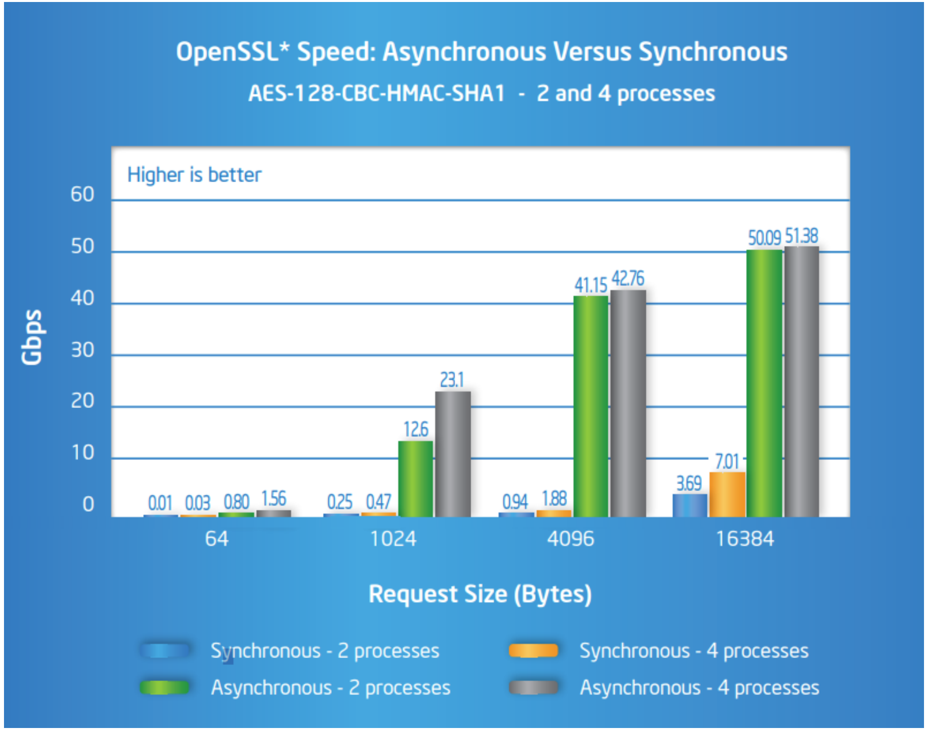
QAT - Intel® QuickAssist Technology

Accelerating OpenSSL



QAT - Intel® QuickAssist Technology

Accelerating OpenSSL



CBDMA/DSA

CBDMA

Crystal Beach DMA (CBDMA) is a DMA engine in the processor, which is extremely efficient in performing memory copy operations.

No CPU intervention during data transfer

Direct memory access (DMA) is a feature of computer systems that allows certain hardware subsystems to access main system memory independently of the central processing unit (CPU).

CBDMA/DSA

CBDMA

CBDMA/DSA

I/O Acceleration Technology (I/OAT)

I/O Acceleration Technology (I/OAT) is a DMA engine (an embedded DMA controller) by Intel bundled with high-end server motherboards, that offloads memory copies from the main processor by performing direct memory accesses (DMA). It is typically used for accelerating network traffic, but supports any kind of copy.

CBDMA/DSA

I/O Acceleration Technology (I/OAT) - DSA

Intel® DSA (Data Streaming Accelerator) is a high-performance data copy and transformation accelerator that will be integrated in future Intel® processors, targeted for optimizing streaming data movement and transformation operations common with applications for high-performance storage, networking, persistent memory, and various data processing applications.

Intel® DSA replaces the Intel® QuickData Technology, which is a part of Intel® I/O Acceleration Technology.

CBDMA/DSA

I/O Acceleration Technology (I/OAT) - DSA

The goal is to provide higher overall system performance for data mover and transformation operations, while freeing up CPU cycles for higher level functions. Intel® DSA enables high performance data mover capability to/from volatile memory, persistent memory, memory-mapped I/O, and through a Non-Transparent Bridge (NTB) device to/from remote volatile and persistent memory on another node in a cluster. Enumeration and configuration is done with a PCI Express compatible programming interface to the Operating System (OS) and can be controlled through a device driver.

CBDMA/DSA

I/O Acceleration Technology (I/OAT) - DSA

Besides the basic data mover operations, Intel® DSA supports a set of transformation operations on memory. For example:

- Generate and test CRC checksum, or Data Integrity Field (DIF) to support storage and networking applications.
- Memory Compare and delta generate/merge to support VM migration, VM Fast check-pointing and software managed memory deduplication usages.

CBDMA/DSA

I/O Acceleration Technology (I/OAT) - DSA

Therefore, it is used to perform the following operations:

- Move data from CPU to RAM and vice versa.
- To access incoherent areas of memory with different memory addressing, address conversion can be performed automatically, so technically we are dealing with an updated DMA.
- It also has access to persistent or non-volatile memory, so it can also access NVMe SSDs, Intel Optane modules, NVDIMMs, etc.
- Through NTB and in the server environment it gives access to another RAM memory or non-volatile memory from another disc in the data center or server.
- It has built-in functions to apply the above points to virtual machines.

Kubernetes Operators

Kubernetes is designed for automation. Out of the box, it has lots of built-in automation from the core of Kubernetes. Kubernetes can be used to automate deploying and running workloads, and to automate how Kubernetes does that.

Kubernetes' operator pattern concept can extend the cluster's behaviour without modifying the code of Kubernetes itself by linking controllers to one or more custom resources.

Kubernetes Operators

Deploying operators

The most common way to deploy an operator is to add the Custom Resource Definition and its associated Controller to your cluster. The Controller will normally run outside of the control plane, much as you would run any containerized application. For example, controller can be run in cluster as a Deployment.

Kubernetes Operators

Using an operator

Once you an operator is deployed, it can be used by adding, modifying or deleting the kind of resource that the operator uses.