White Paper

Hardware Acceleration

Database backups at Enterprise scale

New Intel® QuickAssist Technology Performance on 4th Gen Intel® Xeon® Processors: SQL Server 2022 Backup Compression

4th Gen Intel® Xeon® Processors come equipped with improved Intel® QuickAssist Technology support onboard! We put its performance to test with SQL Server 2022's backup compression feature.

Authors

Manoj Babu
Cloud Software Engineer

Mahmut Aktasoglu
Sr. Cloud Software Engineer

Hamesh Patel
Principal Engineer

Garrett Drysdale
Principal Engineer

Database backups are business-critical operations to support core functionality and availability of many usage models. As such, their performance is important to satisfy business- or user-defined SLAs.

In SQL Server 2022, Microsoft is introducing a ground-breaking optimization feature for backup operations using Intel® QuickAssist Technology (Intel® QAT), addressing the previously mentioned challenges. The feature is designed to work seamlessly with Intel QAT hardware.

In this paper, we take a closer look at SQL Server 2022 support for Intel QAT hardware acceleration and compare its performance to the default XPRESS-based software backup compression implementation. In various experiments performed, we focus on the benefits of hardware-accelerated compression and demonstrate the value of offloading through improved backup times, better compression ratios and reduced CPU utilization, using the new generation of the Intel® QAT accelerators on 4th Gen Intel® Xeon® processors (i.e., HW v2.0).

- Setting up and using Intel QAT hardware acceleration in SQL Server 2022 is extremely straightforward
- The latest Intel QAT devices provide up to 3.27x better compression performance in SQL Server 2022 backup workloads respectively compared to software compression baselines[1]
- Intel QAT hardware acceleration reduces the impact of compression on the CPU by up to 68% compared to software compression, freeing it up to do other more important tasks[1]
- The DEFLATE algorithm used in Intel QAT can provide up to 13% better compression than the default software XPRESS algorithm in the OLTP datasets we tested[1]
- Offload to hardware accelerator also benefitted OLTP workload throughput by up to 3% while performing a backup operation[1]

4th Gen Intel® Xeon® Processors

4th Gen Intel Xeon processors are the latest Intel server class offerings based on next generation architecture. The key feature improvement is the presence of uncore accelerator hardware on die in the CPU. Specialized hardware for common operations being in such proximity to the core provides significant benefits to performance and offload latencies. We look at one such hardware accelerator, Intel Quick Assist Technology, in this white paper.

Intel® QuickAssist Technology (Intel® QAT)

As the complexity of applications continues to grow, systems need more and more computational resources for workloads, including cryptography and data compression.

The 4th Gen Intel Xeon processor family gives customers a scalable, flexible and extendable way to offer Intel QAT cryptography acceleration and compression capabilities to their existing product lines. Intel QAT provides hardware acceleration to assist with performance demands of applications such as 5G UPF, IPsec or TLS networking, or compression/decompression for storage, cloud, enterprise, database or machine learning, while reducing storage footprint and reserving processor cycles for application and control processing.

Server processors come equipped with up to 4 Intel QAT endpoints per processor on the die itself. This significantly reduces latency of access/offload compared to the previous 2nd generation Intel QAT, which allows for improved performance and throughput in the latest generation.

For more information go to: www.intel.com/quickassist

SQL Server 2022

SQL Server 2022 is the most Azure-enabled release of SQL Server yet, with continued innovation across performance, security, and availability. It is part of the Microsoft Intelligent Data Platform, which unifies operational databases, analytics, and data governance.

SQL Server 2022 integrates with Azure Synapse Link and Microsoft Purview to enable customers to drive deeper insights, predictions, and governance from their data at scale. Cloud integration is enhanced with managed disaster recovery (DR) to Azure SQL Managed Instance, along with near real-time analytics, allows database administrators to manage their data estates with greater flexibility and minimal impact to the end-user. Performance and scalability are automatically enhanced via built-in query intelligence. Adding to the list of rich features, SQL Server 2022 adds support for Intel QAT for its backup compression. This feature will help you accelerate your compressed backups with better compression ratios, freeing up both storage and processor usage, allowing for storage and CPU utilization savings.

For more information on SQL Server 2022, go to:

https://www.microsoft.com/en-us/sql-server/sql-server-2022

SQL Server Backup

Backup of a database is a business critical operation that typically occurs periodically in real scenarios for a variety of purposes. Backup operations can be done in two ways. They can be an exact copy of the database or a compressed version of the database. An exact copy occupies a lot of spaces on disk, but can be quite fast as it is a lightweight operation. Compressing the database can provide significant benefits (almost 3x less space than without compression). Performing the compression operation in software comes at the expense of CPU Utilization, i.e. CPU cores that perform the compression operation, which could potentially be used for other purposes.

An example backup command in SQL Server looks like:

```
BACKUP DATABASE { database_name | @database_name_var }
  TO <backup_device> [ ,...n ] [WITH COMPRESSION]
```

The software compression in SQL Server uses the XPRESS9 compression algorithm. It is the default when compression is specified during a backup command.

For more information, see: https://learn.microsoft.com/en-us/sql/relational-databases/backup-restore/backup-compression-sql-server?view=sql-server-ver16 and https://learn.microsoft.com/en-us/sql/t-sql/statements/backup-transact-sql?view=sql-server-ver16

SQL Server 2022 backup compression acceleration using Intel QAT

In this section, we describe the architecture of Intel QAT technology, Intel QAT software stack and how it interacts with SQL Server 2022.

To start using Intel QAT devices, you need to install the latest Intel QAT Windows driver for your devices. For installation steps and details on QAT drivers, please refer to Intel QAT technical guide. The driver installer package will copy all necessary files for SQL Server 2022 compression offload feature to work. You will need at least v2.0 of the QAT driver package from the link below.

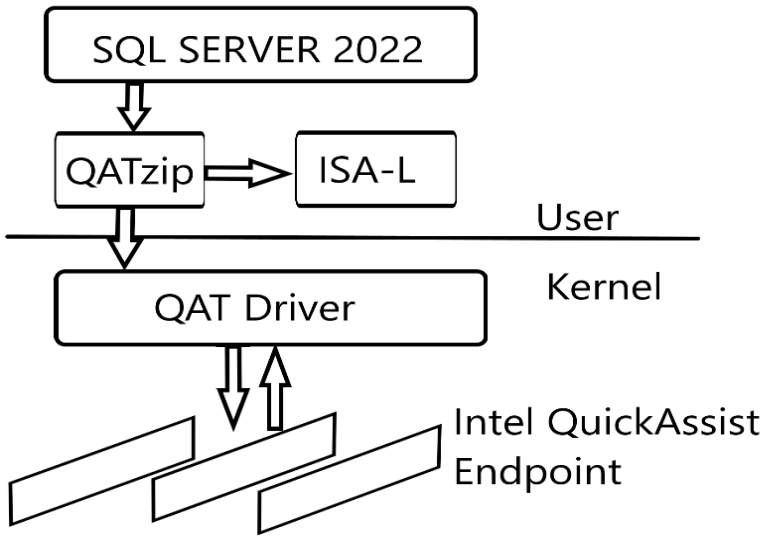

Figure 1. SQL Server 2022 Intel® QuickAssist Technology offload architecture

SQL Server 2022 introduces new DMVs to ensure you can troubleshoot basic installation and usage issues. Details can be found in the links at the bottom of this section.

The Intel QAT offload architecture is illustrated in Figure 1. The main components are:

- QATzip library, which is a user-level library that provides APIs to offload compression jobs to Intel QAT endpoints.

- Intel® Intelligent Storage Acceleration Library (Intel® ISA-L), which provides a highly optimized software implementation of QAT DEFLATE algorithm. It utilizes Intel vector instructions if they are available on the platform and delivers great software compression/decompression performance. QATzip automatically switches to ISA-L APIs in case of device failures or unavailability, which will be referred to as software fallback.

- A kernel mode driver that is responsible for managing various endpoints and work queues for the devices on your system.

Depending on availability of Intel QAT hardware on the platform, SQL Server can perform backup compression acceleration in two modes. During SQL Server start up, if there are Intel QAT endpoints available on the platform, SQL Server will utilize the hardware. The hardware mode is supported by software fallback, which compresses/decompresses the data in the same format. In case of device failures or other offloading issues, SQL Server will seamlessly switch between hardware offloading to software fallback mode when you select QAT_DEFLATE compression algorithm for your backups. If you don't have the hardware, you can still install the Intel QAT compression software components, which will allow you to perform DEFLATE-based compression/decompression on your database backups, and switch to a platform with Intel QAT hardware later seamlessly. In other words, you can decompress your backup files that used Intel QAT hardware to compress on a platform that doesn't have the hardware, and vice versa. This was a key design decision in implementing the feature in SQL Server 2022.

Using Intel QAT compression backups in SQL Server is simple and easy to use. First, we install the QAT drivers to the host machine. Then, we need to do a one-time configuration of SQL Server to load the QAT libraries during startup. Once this is done, Intel QAT is enabled and can be used for any backup at any time. If a QAT accelerator is not present at any time, the same commands in SQL Server will run the format compatible compression, transparently, using software mode (Intel ISA-L) in the QAT software stack.

```
sp_configure 'show advanced options', 1;
GO
RECONFIGURE
GO

sp_configure 'hardware offload enabled', 1;
GO
RECONFIGURE
GO
```

< Stop and restart the SQL Server service.>

```
ALTER SERVER CONFIGURATION
SET HARDWARE_OFFLOAD = ON (ACCELERATOR = QAT);\
```

< Stop and restart the SQL Server service.>

The backup commands are very similar to the default case and only need to specify an additional `algorithm = QAT_DEFLATE` parameter to the original command. This tells the server to use the new Intel QAT offload capability.

```
BACKUP DATABASE { database_name | @database_name_var }
  TO <backup_device> [ ,...n ] WITH COMPRESSION(ALGORITHM = QAT_DEFLATE)
```

While we demonstrate all experiments on bare metal setups, Intel QAT devices can be attached to virtual machines using SR-IOV and provide benefits across the same vectors described in this paper. Check the Technical guide for more information on this process.

The QAT driver can be downloaded from:

https://www.intel.com/content/www/us/en/search.html?ws=idsa-default#q=quickassist&sort=relevancy&f:@tabfilter=[Downloads]

For more information on how to enable Intel QAT on SQL Server, go to:

https://learn.microsoft.com/en-us/sql/relational-databases/integrated-acceleration/overview?view=sql-server-ver16

https://learn.microsoft.com/en-us/sql/relational-databases/integrated-acceleration/use-integrated-acceleration-and-offloading?view=sql-server-ver16

SQL Server 2022 Backup Compression Performance Study

Next up, we present our performance results in the following sections. We start with platform and configuration details, followed by workloads used and main performance metrics we track. Following that, we highlight 2 common backup scenarios and measure metrics and draw conclusions.

PLATFORM SPECIFICATIONS

To test the new Intel QAT hardware-accelerated compression feature in SQL Server 2022, we used 4th Gen Intel Xeon processors with multiple configurations and datasets. In this section, we detail each component used in our system-under-test (SUT).

CPUs tested

Your choice of CPU matters when it comes to SQL Server performance. Our Intel QAT performance measurements are made against SQL Server's default software compression algorithm. In our experiments, we used the Intel® Xeon® Platinum 8490H in a 2-socket configuration. This SKU has 4 Intel QAT endpoints per socket.

BIOS settings

After deciding on CPU and platform, the next step is to decide on the BIOS settings. We decided to mimic a configuration tuned for heavy OLTP workloads, and for that reason, we picked the best-known BIOS settings to get optimal throughput on such workloads. We find such BIOS settings through a series of experiments, where we test different configuration options of relevant BIOS options on the same workload and isolate the settings that provide optimal performance.

In conclusion, we used the BIOS settings illustrated in Table 1, which we refer to as best-known configuration (BKC) BIOS settings for the rest of the paper. We will highlight any changes to BKC BIOS settings, if applicable in each scenario's configuration table.

Another point worth noting: the configuration names are as presented on our platforms and might differ on other OEM platforms. Please refer to your system technical documentation on how to change the settings for your platform. Most BIOS settings are mostly out-of-the-box options, and we indicate the configurations changed from the default value by an asterisk(*)

Table 1. BIOS configuration for various settings

Processor Configuration	
Virtualization	Disabled
MLC Streamer	Disabled
MLC Spatial Prefetcher	Disabled
DCU Data Prefetcher	Enabled
DCU Instruction Prefetcher	Enabled
LLC Prefetcher	Disabled *
Power and Performance	
Uncore Power Management	
CPU Power and Performance Policy	Balanced Perf
Workload Config	IO Sensitive
Perf P-Limit	Enabled
Uncore Freq Scaling	Enabled
Uncore Freq RAPL	Enabled
CPU P-state	
SpeedStep	Enabled
Turbo Boost	Enabled
Energy Efficient Turbo	Disabled *
HW P-State	
Hardware P-States	Disabled *
CPU C-State	
Package C-State	C0/C1 state *
C1E	Disabled *
Processor C6	Disabled *

Operating Systems

In all our experiments, we use Windows Server 2022 Datacenter Edition (10.0.20348.587).

Storage Subsystem

Storage system setup is an important contributor to overall backup performance, not only because it is an IO-heavy operation, but also storage and file layout of your backups and databases have an impact on degree of parallelism. SQL Server determines the number of threads for backup based on the number of mountpoints for the database being backed up and the number of backup files/stripes being written into. The compression operation happens in the writer threads and hence, by scaling how many files we specify for the backup, we can modulate the number of compression threads we use.

In our experiments, we aim to avoid IO becoming a bottleneck so that we can do a fair CPU vs. accelerator performance comparison. For that reason, we use high-performance Intel® P4608 NVMe storage units that provide low IO latencies and ample IO bandwidth. On top of that, we tuned disk layout and striping parameters to achieve optimal backup performance. Our experiments are configured to use 8 writer threads and 4 reader threads for each backup. Figure 2 represents the storage layout of our experiments.

Figure 2. Diagram of storage layout where each blue blob represents a disk mountpoint. Each black line is a source file, i.e. the original database file being read from. Each yellow line is a backup stripe being written into after compression.

The hardware configuration is summarized in the Table 2 below

Table 2 Hardware Configuration used for experiments on 4th Gen Intel® Xeon® Processors and Intel® QAT

	Config 1 (4th Gen Intel® Xeon® Processor Xpress Software Compression)	Config 2 (4th Gen Intel® Xeon® Processor with Intel® QAT Hardware Accelerated Compression)
CPU Model	8490H	8490H
Sockets	2	2
Cores per Socket	60	60
Hyperthreading	Enabled	Enabled
CPUs	240	240
Intel Turbo Boost	Enabled	Enabled
Base Frequency	1.9GHz	1.9GHz
All-core Maximum Frequency	2.9GHz	2.9GHz
Frequency during Run	2.9GHz	2.9GHz
NUMA Nodes	4	4
Prefetchers	DCU HW	DCU HW
Accelerators (4th Gen Only)		**Intel® QAT Gen4 (x4 per socket)**
Installed Memory	1024GB,16*64GB,4800MHz,DDR5	1024GB,16*64GB,4800MHz,DDR5
Disk	3x P4608 NVME Drives	3x P4608 NVME Drives
BIOS	EGSDCRB.SYS.OR.64.2023.09.2.02.0429.1	
Microcode	0x2b0001b000000000	0x2b0001b000000000
OS	Windows Server 2022	Windows Server 2022
Kernel	10.0.20348.587	10.0.20348.587
TDP	350W	350W
Frequency Governor	Performance	Performance
Power & Perf Policy	Performance	Performance
Max C-state	C0/C1	C0/C1

WORKLOADS

In this paper, what we refer to workload depends on the context. Our main focus is on backup performance, so any backup operation we perform is in and of itself a workload. What matters in this scenario is the database size, data type/composition of the database that is being backed up – which impacts how compressible data is, and how many concurrent backups are happing on the SUT.

In another context, we also use an OLTP workload alongside the backup operation on the same database that is being backed up.

There are a few considerations on how to pick an appropriate workload in this context.

- Transaction types
- Transaction composition
- Workload characteristics (i.e., steady-state, high-CPU utilization, etc.)
- Amount of interference to backups

To demonstrate Intel QAT benefits, we picked our workloads that exhibit a steady state, high-CPU utilization profile with differing level of interaction with on-going backup operation (I.e., low and high interaction). One could characterize such OLTP workloads as the worst-case scenario to run backups within SLAs without impacting operational SLAs.

In our experiments, we use a Microsoft proprietary stock trading application emulator (STAE), a light-weight cloud OLTP workload, as low interaction workload, and HammerDB as the high interaction workload (TPROC-C like).

The STAE workload consists of mostly short, CPU-intensive read-only queries and update heavy profile in terms of non-read only. Read-only queries account for roughly 70% of the transactions in flight at any given time, whereas UPDATEs constitute 23% and rest are INSERTs and DELETEs. This database is ~500GB in size and has 4 DB files and 1 log file.

HammerDB is the leading benchmarking and load testing software for the world's most popular database engines. We use HammerDB to create a TPROC-C test schema, load it with data and simulate the workload of multiple virtual users against the database for both transactional and analytic scenarios. This workload can then be used to derive meaningful information about your environment such as hardware performance comparisons and software configurations. The database used is 12000 WareHouse and is ~470GB in size and comprises 4 DB files and 1 log file.

The databases used in our experiments are summarized in Table 3 here:

Table 3 Databases used in backup experiment

	TPROC-C like	STAE
Database	**TPROC-C likedatabase (HammerDB)**	**Stock trading application emulator from Microsoft**
ScaleFactor/Warehouse	5000 WH	20,000 SF
Size on Disk	472 GB	516 GB
DB Size (Used Space)	393 GB	516 GB
Number of Backup stripes	8 files	
Number of mountpoints	4 disks for source data and 2 disks for destination backup	

METRICS

Backup time reduction

Backup time is the time taken from the start of the backup to the completion of the backup. The lower this metric the better it is for user experience. The Intel QAT accelerator can improve the time by optimizing the compression pathway of the backup operation.

Compression ratio improvement

The compression ratio is simply calculated as (uncompressed size)/(compressed size). It can also be depicted using space savings as 1-(compressed size)/(uncompressed size) . A higher compression ratio helps reduce space needed to store data, which can reduce time and cost of storing or moving the data. Intel QAT uses the DEFLATE algorithm that provides better compression than the current SQL Server standard of Xpress9 algorithm.

CPU utilization reduction

Intel QAT is a dedicated HW accelerator that performs compression operation. In the case of SQL Server, it shifts the compression path from the CPU to the accelerator. This has multiple benefits – frees up the CPU to do other tasks, improves backup time and improve compression ratio.

Transaction Throughput improvement

In cases where we run an OLTP workload while backing up the database, we also measure transaction throughput of queries when SQL Server 2022 is using HW offloading vs. default XPRESS software compression.

PERFORMANCE IMPROVEMENT WITH INTEL QAT HW v2.0 over SQL Server 2022 Software-compressed backups

Scenario 1: Backing up a database with compression on an idle system

In this scenario, we are looking into various configurations and measure performance in the case where the database backup is the only operation on the platform. This scenario lets SQL Server utilize available platform resources for backup operations. In other words, there is no other significant disturbance to thread scheduler and backup reader/writer threads will face no contention.

We ran backup experiments on a bare metal 4th Gen Intel Xeon Scalable system and observed up to 3.27x improvement in backup time and up to 69% reduction in CPU utilization to backup up a database using SQL Server. Significant CPU cycles saved while providing faster throughput is the hallmark of good accelerators like QAT. Figure 3 and Figure 4 summarizes the performance gains and CPU savings provided by offloading the backup to the QAT accelerator.

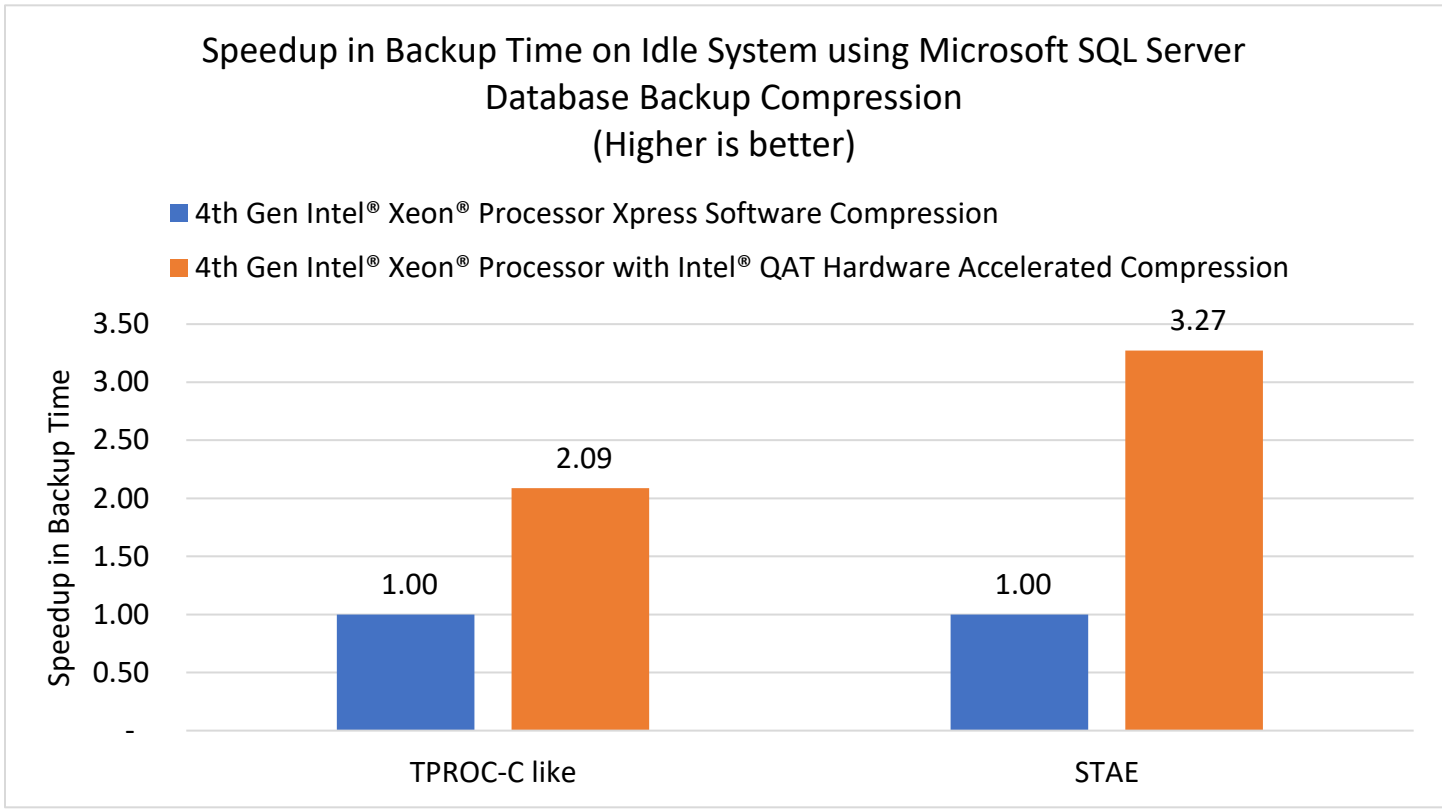

Figure 3 Speedup in Backup Time using default Xpress and Intel® QAT accelerator when platform is otherwise idle

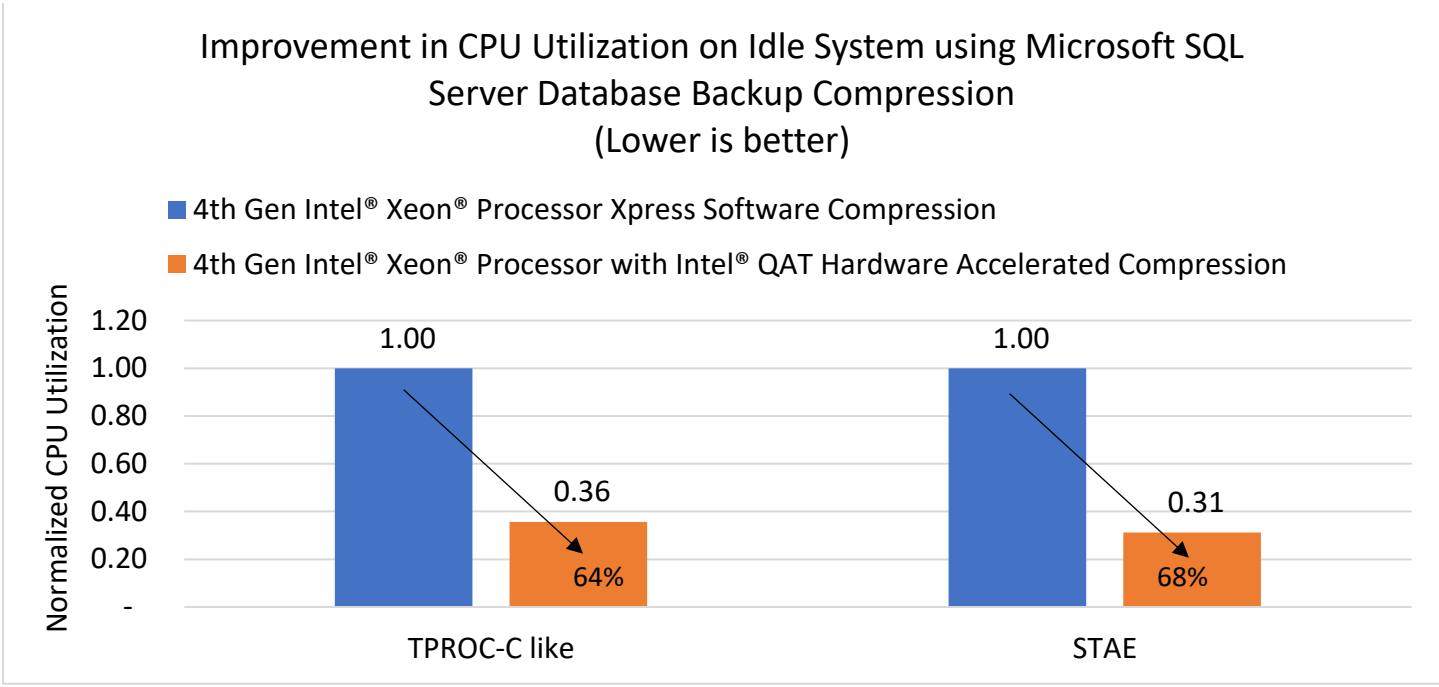

Figure 4 CPU Utilization impact of offload from CPU to accelerator

We note that, STAE database has a higher speedup compared to TPROC-C like database. This highlights the effect of dataset on the benefits of QAT and the compression algorithm used. STAE database is more suited to compression using the DEFLATE algorithm and hence shows better speedup. This can be observed from the compression ratios of the two databases as shown in Figure 5. You can see that the STAE database is more compressible of the two. Additionally, this means that for the STAE database, using QAT will amount to ~13% savings on disk and for TPROC-C like database, the savings are around 2% compared to using Xpress backup method. The compression ratio is calculated as $Compression\ Ratio = \frac{uncompressed\ size}{compressed\ size}$

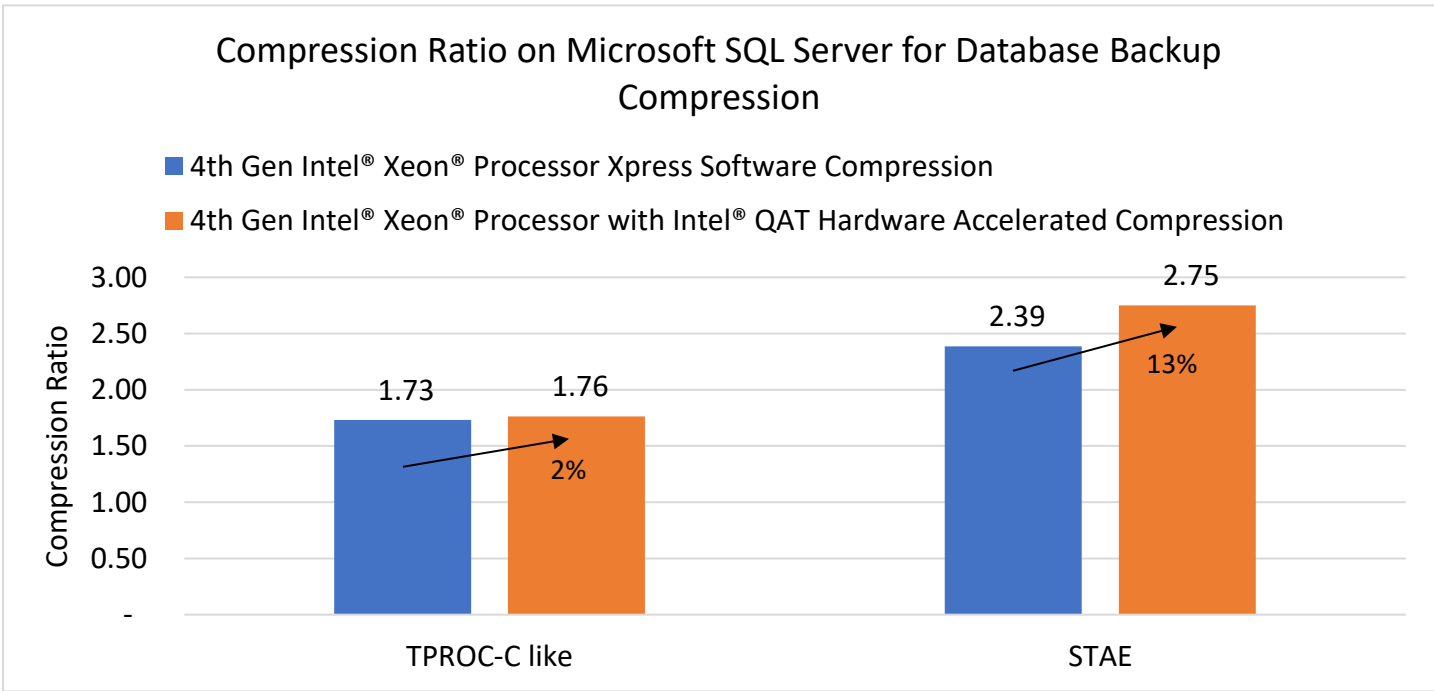

Figure 5 Compression ratio of different mode (algorithms) and different datasets

Scenario 2: Backing up a database with compression under load

For this scenario, we ran an Online Transactional Processing (OLTP) workload on the same database that is being backed up. This is a typical use case for database servers and datacenter platforms in the field. While the transactions are running, we start a backup operation with compression enabled. In the Intel QAT case, the compression is not handled by the CPU, which is now busy with running transactions, the backup is completed much faster.

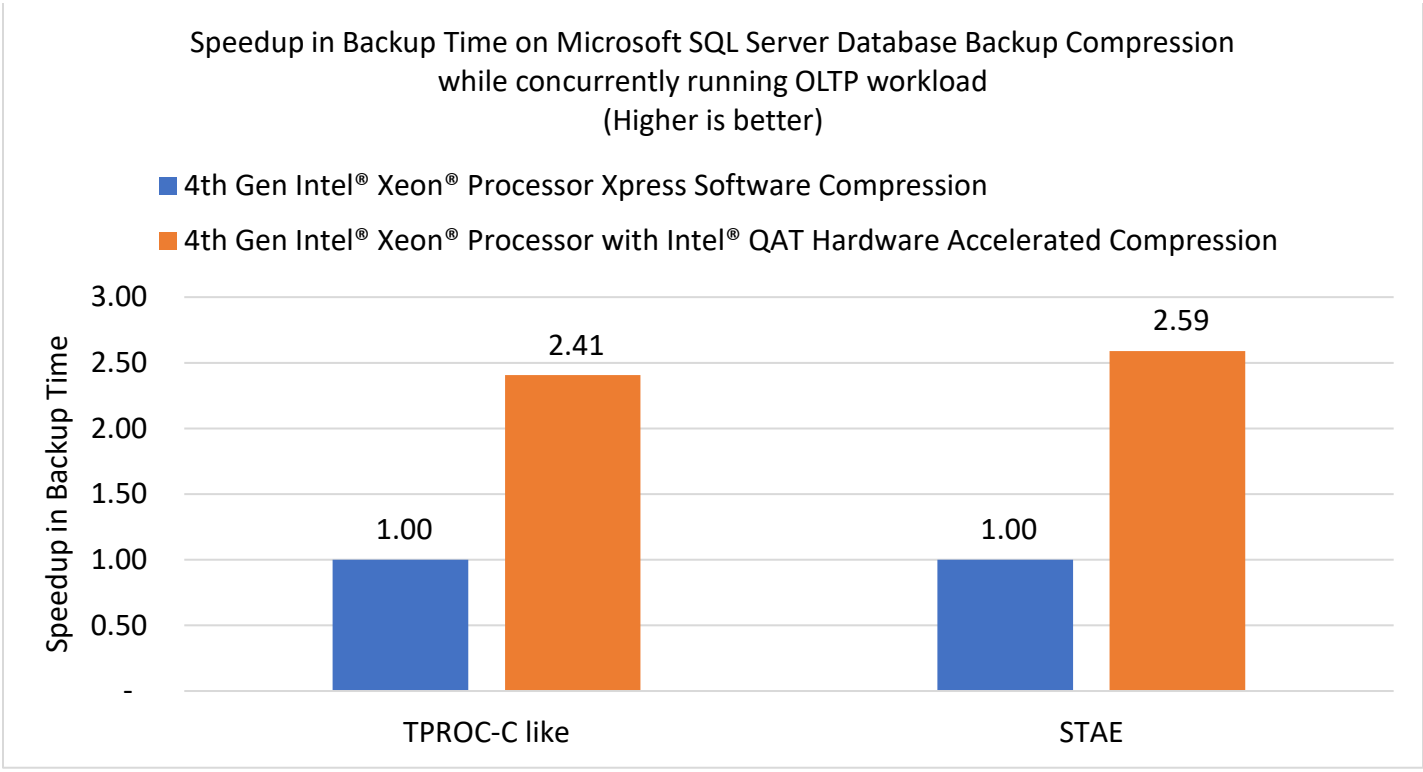

Figure 6 Speedup in Backup Time using default software Xpress and Intel® QAT accelerator when platform is under load with different OLTP workloads

This scenario reveals further benefits of Intel QAT HW acceleration, which is the impact on the OLTP workload performance in terms of throughput. When the compression is offloaded to Intel QAT, the freed-up CPU cycles are used to perform more transactions with reduced impact in comparison to doing the same compressed backup on the same cores. Per our measurements, we observe a ~3% improvement in the throughput of OLTP workloads during the backup phase. Note that this is during the backup phase and with the hardware accelerated Intel QAT, the time when the throughput drops also reduces as per the backup time improvements discussed previously.

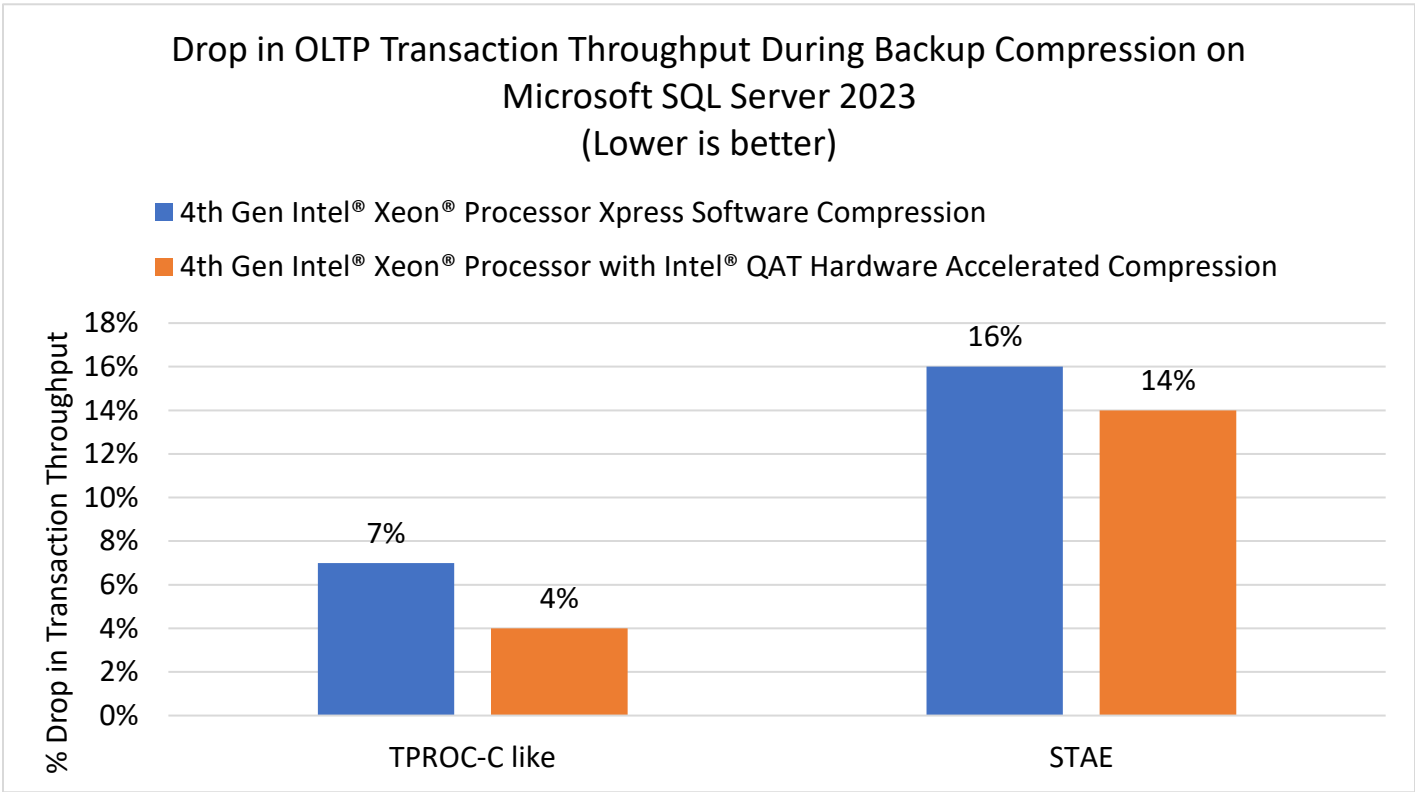

Figure 7. QAT has better sustained throughput during backup operation due to offload of compression tasks

SUMMARY AND CONCLUSIONS

Hardware technology is extremely fluid, with new capabilities every year. Intel QAT is one such technology that provides significant compression/decompression advantages. Microsoft SQL Server has demonstrated their adaptability in their swift adoption of this new Intel QAT technology. SQL Server 2022 has embraced accelerator offload and incorporated database backups using Intel QAT hardware accelerator. This provides database administrators another powerful knob to tune their 4th Gen Intel Xeon servers to extract the most out of their platforms. We have demonstrated results for:

- Setting up and using Intel QAT hardware acceleration in SQL Server 2022 is extremely straightforward
- The latest Intel QAT devices provide up to 3.27x better compression performance in SQL Server 2022 backup workloads respectively compared to software compression baselines[1]
- Intel QAT hardware acceleration reduces the impact of compression on the CPU by up to 68% compared to software compression, freeing it up to do other more important tasks[1]
- The DEFLATE algorithm used in Intel QAT can provide up to 13% better compression than the default software XPRESS algorithm in the OLTP datasets we tested[1]
- Offload to hardware accelerator also benefitted OLTP workload throughput by up to 3% while performing a backup operation[1]

Intel QAT on SQL Server is an exciting new feature that can only improve with more features and optimizations planned in the future.

[1] See [D18] at intel.com/processorclaims: 4th Gen intel® Xeon® Scalable processors. Results may vary. Performance varies by use, configuration and other factors. **SOFTWARE BASELINE:** Test by Intel as of 3/20/2023. 1-node, 2x 4th Gen Intel® Xeon® Scalable Processor, >40cores, HT On, Turbo On, Total Memory 1024 GB (16 slots/ 64GB/ 4800 MHz [run @ 4800MHz]) DDR5 memory, ucode 0x2b0001b000000000, Windows Server 2022, 10.0.20348.587, SQL Server 2022 16.0.1000.6, OLTP database backup with Xpress software compression using 12 total threads. **WITH INTEL® 4th GEN QUICKASSIST TECHNOLOGY:** Test by Intel as of 3/20/2023. 1-node, 2x 4th Gen Intel® Xeon® Scalable Processor, >40cores, HT On, Turbo On, Total Memory 1024 GB (16 slots/ 64GB/ 4800 MHz [run @ 4800MHz]) DDR5 memory, ucode 0x2b0001b000000000, Windows Server 2022, 10.0.20348.587, 4th Gen Intel® QuickAssist Technology, 2.0.10.18 driver version, SQL Server 2022 16.0.1000.6, OLTP database backup with Intel® QAT compression using 12 total threads

0623/EP/PDF 355448-001US