WHITEPAPER

The Era Of General Acceleration Has Begun

THENEXTPLATFORM | SUPERMICRO intel XEON PLATINUM

Executive Summary

The era of general purpose computing began in April 1964 with the launch of the aptly named System/360 mainframe by IBM. Computing was simpler then: Batch processing of integer data and, eventually, online transaction processing.

But the idea was that one family of systems could support a whole stack of systems software, including operating systems, databases, transaction monitors, programming tools, and such.

Thanks to the System/360's enormous success, soon thereafter dozens of different minicomputer makers stormed the datacenter with their own unique processors and software stacks tuned for them, aimed at specific workloads – some of them back office work like the System/360 mainframes did, some of them focused on new eras such as simulation and modeling.

The RISC/Unix revolution vanquished most of these proprietary systems, and the advent of the Pentium Pro and Xeon server processors from Intel coincided with the rise of the commercial Internet to foster the first nearly universal computing substrate that was affordable and that could run hyperscale workloads that the world had been imagining for years.

The System/360 in the 1960s and the 1970s, the Unix systems of the 1980s and the 1990s, and the X86 platform of the 1990s and 2000s represent three different epochs in the era of general purpose computing. But today, something different is happening. Driven by the need to highly tune workloads to systems, and to meet very tight power, thermal, and monetary budgets and to deliver the most optimal performance and value for the dollar, we are entering what we might call the era of general acceleration.

In this new era, compute engines are going to proliferate

In this new era, compute engines are going to proliferate, offering varying types and capacities of compute and memory to match a set of workloads that is diversifying faster than the SKU stacks of the major CPU vendors. It has been decades since just being able to manipulate integers was sufficient in a compute architecture, which saw the rise of outboard math coprocessors and then integrated vector engines and now, with the advent of artificial intelligence, the rise of matrix math units to accelerate AI inference and in some cases AI training, too. But all kinds of other functions that might otherwise be done by CPU cores are being done by accelerators, often integrated on cores, sometimes integrated in the CPU package and other times in outboard accelerators such as GPUs, FPGAs, or DPUs.

And importantly, for each workload, the right kind and capacities of compute have to be brought to bear in a system that meets the physical and economic requirements of where that workload is running. The era of general purpose computing is dead, and so is the era of the 1U and 2U pizza box server. Both the compute engines and the systems that wrap around them have to be co-designed and optimized along with the workload software stack because the end of Moore's Law demands this.

Acceleration Is Core To The Future Of Compute

The ability to shrink transistors over the past five decades made modern computing possible. But system architecture is not as simple as putting an entire motherboard on a single piece of silicon that fits into a single server socket. The fact that those transistors are getting more expensive as they keep shrinking now means that some tough architectural choices need to be made, and every possible function which might be needed cannot be reasonably crammed into that core anymore.

Over the past decade, Moore's Law roughly equated to transistors getting cheaper by a factor of 2X every 18 to 24 months meaning manufacturers could afford to put more of them on a device. That has begun to slow however while Dennard scaling – as manufacturing process shrinks, you can boost the clock speed to add more performance at the same thermals – has long since stopped. All of which means processor and system architects have been struggling with where to put what kind of compute.

Because variety is the spice of life, there is a progressive offloading of different compute functions from the core to the die or socket to the external data processing unit (DPU), graphics processing unit (GPU) and other accelerators outside of the socket. But not every CPU maker draws the lines in the same way while all CPU makers let their customers, who have very different applications and therefore very different system needs, use different architectures to do this progressive offloading. The end result is that systems tend to be tailored as much as is possible in a semiconductor industry that requires a certain level of volume economics for anything to be built in the first place.

It is a difficult and delicate dance, figuring out where to compute what in a market that has seen a Cambrian explosion in compute engines - there's an increasing diversity of compute engine makers but an increasing diversity of instructions within a socket too.

It is a difficult and delicate dance, figuring out where to compute what in a market that has seen a Cambrian explosion in compute engines.

The reason is simple: efficiency. Sometimes, when a function becomes ubiquitous enough, it makes sense to implement it in an ASIC to drive performance and price/performance rather than implement it in software, where its performance will be bound by the number cores you can allocate to it and the clock speed of those cores. The latter approach uses up valuable CPU cores and burns a lot more electricity per unit of work.

All things being equal, it would have been easier on everyone if everything could have been implemented in high level software and run on a more general purpose X86 processor. This certainly happened during the Dot Com Boom, thanks in large part to shrinking transistors and scaling clock frequencies on CPUs.

But because Dennard scaling is dead and Moore's Law is limping along, it means we have to rely on an expanding hierarchy of compute to wring efficiencies out of systems. And therefore, our old habits of counting cores and clock speeds and measuring the relative increase in the integer performance of the cores on a processor, and then picking a pizza box server after the fact is no longer a good way to build a system.

Now, we have to start with the workload, pick the appropriate form factor for the system, including its power, thermal, and physical constraints (the machine may not be in an idealized datacenter environment, but somewhere out on the edge), and then optimize the compute, networking, and storage

within that system along performance, price/performance, or performance/watt vectors. The process is a co-design of hardware and software within a physical and budgetary environment. And you can't optimize for all of these constraints at the same time. You have to pick, and it always involves making tradeoffs.

Let's walk through three different workloads running on various Supermicro systems that employ Intel's latest "Sapphire Rapids" 4th Generation Xeon SP processors to see how this works. Then, we will dive into the feeds and speeds of all of the Intel Acceleration Engines in the Sapphire Rapids CPUs that can be brought to bear to boost the performance of all kinds of workloads.

AI Inference

Artificial intelligence is taking the datacenter by storm, and every company is looking at the best way – and the cheapest way – to weave inference into their workloads.

Artificial intelligence is taking the datacenter by storm, and every company is looking at the best way – and the cheapest way – to weave inference into their workloads. They can add custom ASIC or GPU accelerators from various sources to their application servers, or buy banks of them and offload inference from their application servers entirely. But for latency, security, and cost reasons, these are no longer necessarily the best choices.

As CPUs get matrix math units that can accelerate AI inference, it makes sense to consider keeping inference not only within the same servers that are running the applications, but on the CPUs themselves. Intel is first out of the gate to accelerate inference (and also small-scale AI training) with the Advanced Matrix Extension (AMX) matrix math units in the Sapphire Rapids CPUs.

For application serving including AI inference, the Supermicro X13 Hyper SuperServer (SYS-121H-TNR) is a good platform on which to build.

SUPERMICRO X13 HYPER

The Supermicro X13 Hyper SuperServer system (SYS-121H-TNR)

This machine comes in a 2U, rack-mounted form factor that has been the standard for application serving in the enterprise datacenter for two decades now, and it comes with a pair of Sapphire Rapids processors, 32 DDR5 main memory slots, room for a pair of internal NVM-Express M.2 flash or SATA3 disk drives for the operating system, slots for eight 2.5-inch NVM-Express, SAS, or SATA storage devices, and eight internal fans to keep it all cool. The X13 Hyper SuperServer also comes with two Advanced I/O Module (AIOM) networking ports with a variety of Ethernet speeds, a pair of 1,200 watt Titanium-level energy efficient power supplies, and a dedicated baseboard management controller with its own network interface; it also has a bezel-free configuration if you do not want to pay for unnecessary plastic.

The X13 Hyper SuperServer is able to handle Sapphire Rapids CPUs rated at up to 350 watts, which means you can put in low, middle, or top bin parts depending on the application needs. The performance of the AMX units in the processors will scale more or less linearly with the number of sockets and cores because there is one AMX unit per core. Core for core against the other current X86 server chip in the market, if you do image recognition using the ResNet50 model, natural

language processing using the BERT Large model, or the DLRM PyTorch recommendation model, you will get a factor of 3X improvement in performance on AI inference work over that alternative. And the cost is effectively zero compared to using outboard accelerators because the AMX units are included in the Sapphire Rapids cores essentially for free.

Database Acceleration

Now, let's build a fast database server. Here, we will start with the Supermicro X13 Multi-Processor server, (SYS-241H-TNRTTP), which scales up to four-way configurations and is designed for heavy database workloads where lots of compute, storage, and memory are brought to bear to support a single instance of a database management system.

SUPERMICRO X13 MULTI-PROCESSOR

The Supermicro X13 Multi-Processor system (SYS-241H-TNRTTP)

The X13 Multi-Processor system comes in a 2U rack-mounted form factor that can have up to four of the Sapphire Rapids CPUs. It has 64 DDR5 memory slots, with a maximum capacity of 16TB using 256GB DIMMs. The system has a pair of M.2 NVM-Express or SATA3 devices for the operating system and eight 2.5-inch bays that can hold flash or disk drives, and has a pair of AIOM network interfaces and a pair of 2,700 watt Titanium-level power supplies to juice it all up.

With the Sapphire Rapids CPUs, Intel is announcing a new set of circuits called the In-Memory Analytics Accelerator, or IAA for short, that boosts the performance of analytics primitives and CRC calculations. It also works with another set of algorithms called Quick Assist Technology, or QAT, to decrease the memory capacity and memory bandwidth requirements for analytics and database workloads.

The jump in performance per socket for the Sapphire Rapids chips compared to the prior generation "Ice Lake" 3rd Gen Xeon SPs is significant. With the Clickhouse DB database, performance is 59 percent higher on a 60-core Xeon SP-8490H than on a 40-core Ice Lake Xeon SP-8380 machine, and with the RocksDB database, performance is up to 3X higher compared to the Ice Lake equivalent. By using the accelerators, that means a database server like the Supermicro X13 Multi-Processor system could do that much more work, or companies could cut back on the cores they deploy to get the same performance and possibly radically lower their database licensing costs.

In this case, customers could optimize for both performance and price, which is a rare thing indeed in the datacenter.

5G Edge Workloads

As our smartphones become the most important personal computing device in our lives and as AI and other workloads move out closer and closer to them to provide the lowest possible latency for those workloads, systems have to be deployed in somewhat hostile environments out at the edge. For such workloads, the Supermicro X13 SuperEdge system (SYS-211SE-31D) is the platform on which to start building.

SUPERMICRO X13 SUPEREDGE

The Supermicro X13 SuperEdge (SYS-211SE-31D)

The X13 SuperEdge server is a half-depth, 2U rack-mounted server designed explicitly for telcos and other service providers. It has three hot pluggable, single-socket server nodes that can slide into its enclosure to provide compute and storage for telecom DRAN, CRAN, Flex-RAN, and Open vRAN and all kinds of edge applications. The server nodes support a single Sapphire Rapids CPU rated at between 85 watts and 300 watts, up to 2 TB of DDR5 memory across eight DIMMs, a pair NVM-Express M.2 flash drives for the operating system and local data, and three PCI-Express slots for network and other peripheral expansion. The chassis has a pair of 2,000 watt power supplies.

With the "Golden Cove" cores used in Sapphire Rapids CPUs, the AVX-512 unit has been tuned to accelerate the vRAN software stack that does the conversion of analog radio signals to digital forms. That means they can be transmitted over the backhaul in the fiber optic networks that link base stations to datacenters and their service, for example. This is done with Intel's FlexRAN and Data Plane Development Kit (DPDK) software, which runs certain functions on the AVX-512 unit if it sees that it is a Sapphire Rapids CPU. With this Open vRAN acceleration, this part of the software stack used in 5G mobile networks can be boosted by up to 2X. This can potentially mean putting fewer systems out at the edge for 5G networks.

Drilling Down Into The Accelerator Engines

The core is no longer the universal unit of performance in the datacenter that it has been over the past two decades.

The core is no longer the universal unit of performance in the datacenter that it has been over the past two decades. More and more functions that were previously outside of the socket have been added to the core – things like PCI controllers, memory controllers, and floating point vector math units to accelerate HPC. And now, we see CPUs getting matrix math units to accelerate AI inference and sometimes AI retraining workloads too.

These present fixed ratios of functions relative to integer performance in the cores. But for other kinds of acceleration – data encryption, data compression, network and storage acceleration, and so forth– putting all of this onto each core would be overkill and would actually limit the number of cores that could be put on a chip to do real work.

Having integrated accelerator engines does not necessarily mean customers won't implement certain functions outside of the CPU package or in software. The necessity of precisely tuning hardware to the needs of a specific application is going to drive complexity across the entire server base, and there is no way out of it as long as the market demands flexibility. Whilst this seems unavoidable, with a slight shift in thinking, the complexity that enables that flexibility is also desirable. So, expect certain functions to be embedded in the CPU cores, on the CPU packages, in the IPUs and SmartNICs, and in other peripheral cards attached over the PCI-Express bus, and in many cases using memory coherence over the CXL protocol.

When you look at how the applications are actually running, some functions should be outside of the core and shared by a mesh of cores, but still be placed on the die to ensure low latency and high performance. It really comes down to how important a function that was implemented in higher level software becomes to datacenter hosting operations.

If the function implemented in software consumes a lot of CPU cycles, it will be used when necessary but sparingly. This was the case for data encryption, for more than a decade. Basically until enough transistors were in the processor

Suddenly it became technically and economically feasible to start encrypting all data at rest and in motion, without a performance penalty.

budget and there was enough customer demand – particularly from the hyperscalers and cloud builders which were most definitely not encrypting any data internally within their datacenters, but at the same time absolutely knew they needed to do so because they were building shared compute utilities.

And so, encryption acceleration came to server CPUs and suddenly it became technically and economically feasible to start encrypting all data at rest and in motion, without a performance penalty. Putting data encryption accelerators in the server CPU package transformed it from a generic software cost that burned a lot of cores or that required an external accelerator hanging off the PCI-Express bus to being included in the server CPU architecture. And with a minor incremental cost or none at all compared to prior CPU generations and freed up the cores that were running encryption algorithms.

It is hard to beat inexpensive acceleration embedded on the CPU chiplet in terms of latency, and you can't beat 'included for free' in terms of price/performance. And there is a multiplicative effect as software functions are taken off the cores, freeing them to do work, and functions are implemented in a relatively small number of transistors that perform those functions much faster.

There are more than a dozen Accelerator Engines in the most recent 4th Gen Intel® Xeon® Scalable processors, and the number has been gradually increased in the past four processor generations:

SUPERMICRO WP ACCELERATOR ENGINE GENERATION TABLE

	Intel® Xeon® Scalable processors (Sky Lake)	2nd Gen Intel® Xeon® Scalable processors (Cascade Lake)	3rd Gen Intel® Xeon® Scalable processors (Ice Lake)	4th Gen Intel® Xeon® Scalable processors (Sapphire Rapids)
Intel® Advanced Vector Extensions 512 (Intel® AVX-512)	X	X	X	X
Intel® Crypto Acceleration			X	X
VNNI, BF16 (Intel® Deep Learning Boost)		X	X	X
Intel® Advanced Vector Extensions (Intel® AVX) for vRAN				X
Intel® Advanced Matrix Extensions (Intel® AMX)				X
Intel® Volume Management Device		X	X	X
Intel® Control-Flow Enforcement Technology (Intel® CET)				X
Intel® Software Guard Extensions (Intel® SGX)			X	X
Intel® Trust Domain Extensions (Intel® TDX)				Limited
Intel® Speed Select Technology (Intel® SST)		X	X	X
Intel® Data Direct I/O Technology (Intel® DDIO)	X	X	X	X
Intel® Dynamic Load Balancer (Intel® DLB)				X
Intel® QuickAssist Technology (Intel® QAT) (integrated)				X
Intel® Data Streaming Accelerator (Intel® DSA)				X
Intel® In-Memory Analytics Accelerator (Intel® IAA)				X

These Accelerator Engines are part of the same coherent, shared memory space as the "Golden Cove" cores on the Sapphire Rapids package, which means the latency between the accelerators and the cores is no worse than between the cores on the die.

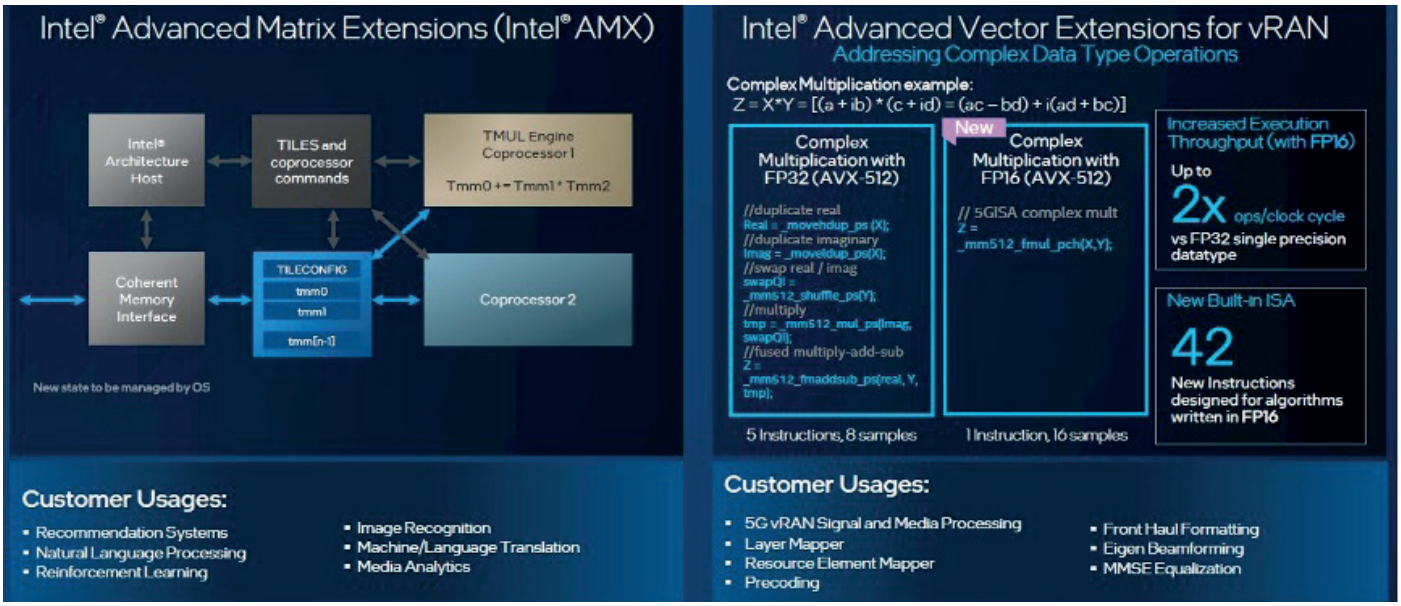
The six most important of the Accelerator Engines, according to Intel, that are on the Sapphire Rapids processors, are the following:

The AVX-512 vector and AMX matrix math units are on each core in the Xeon architecture.

SUPERMICRO ACCELERATOR ENGINE SIX KEY

For AMX, Intel is working with the upstream AI frameworks, toolkits, and libraries – starting with TensorFlow and PyTorch – as well as adding support for AMX in its own oneDNN neural network. In time, we believe that AMX could be used to support certain kinds of low-precision HPC calculations, we also believe that certain portions of HPC solvers will be based on AI models rather than brute force calculations to further

SUPERMICRO ACCELERATOR ENGINE SIX KEY



accelerate HPC performance. These are works in progress at the moment.

Quick Assist Technology, or QAT, is new to the Xeon SP processor complex with Sapphire Rapids. What's interesting is that Intel has been moving hardware based QATzip data compression and SSL encryption algorithms upstream towards the CPU package over the last fifteen years. QAT started out on Ethernet network interface cards way back in 2007, and then moved onto certain server chipsets, then into chipsets co-packaged with Xeon D chips, and now is on the chiplets (but not the Golden Cove cores) in the Sapphire Rapids package.

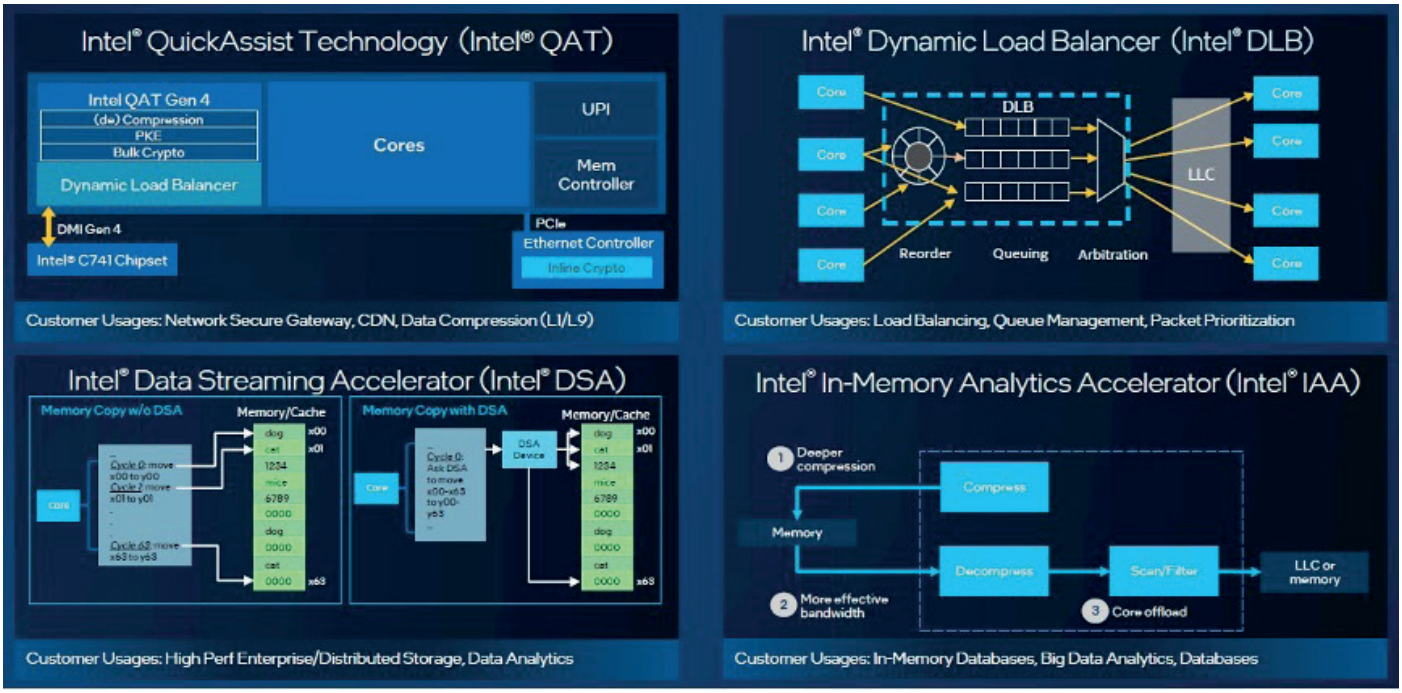
QAT is useful for speeding up and securing distributed storage systems, file systems, databases (with support for the open source RocksDB being highlighted), Hadoop and Spark data analytics, NGINX Web application serving, and such. Customers who create their own middleware and database software can access QAT acceleration using instructions, just like any other feature of the Xeon SP CPU. But Intel has worked with third party software providers and open source projects to make sure QAT is supported wherever it is appropriate so customers don't have to think of it.

The same is true of the other Accelerator Engines in the Sapphire Rapids processors. The idea is to make this as seamless and invisible as possible.

The other three major new accelerators going clockwise – Dynamic Load Balancer, In-Memory Analytics Accelerator, and Data Streaming Accelerator – are brand new with Sapphire Rapids CPUs. Let's walk through these.

- **Dynamic Load Balancer (DLB):** DLB is addressed through the Intel Data Mover Library and it is used to dynamically redistribute the flow of data across a core complex and its network interfaces in place of an ingress gateway that runs in software. DLB can underpin IPSec security gateways, VPP routers, virtual switching, streaming data processing, and the handling of elephant flows in network function virtualization (NFV) workloads. DLB reduced the latency of load balancing across a microservices workload by 96 percent at the same throughput compared to using the Istio ingress gateway.
- **In-Memory Analytics Accelerator (IAA):** IAA accelerates analytics primitives and CRC calculations

SUPERMICRO ACCELERATOR ENGINE FOUR NEW

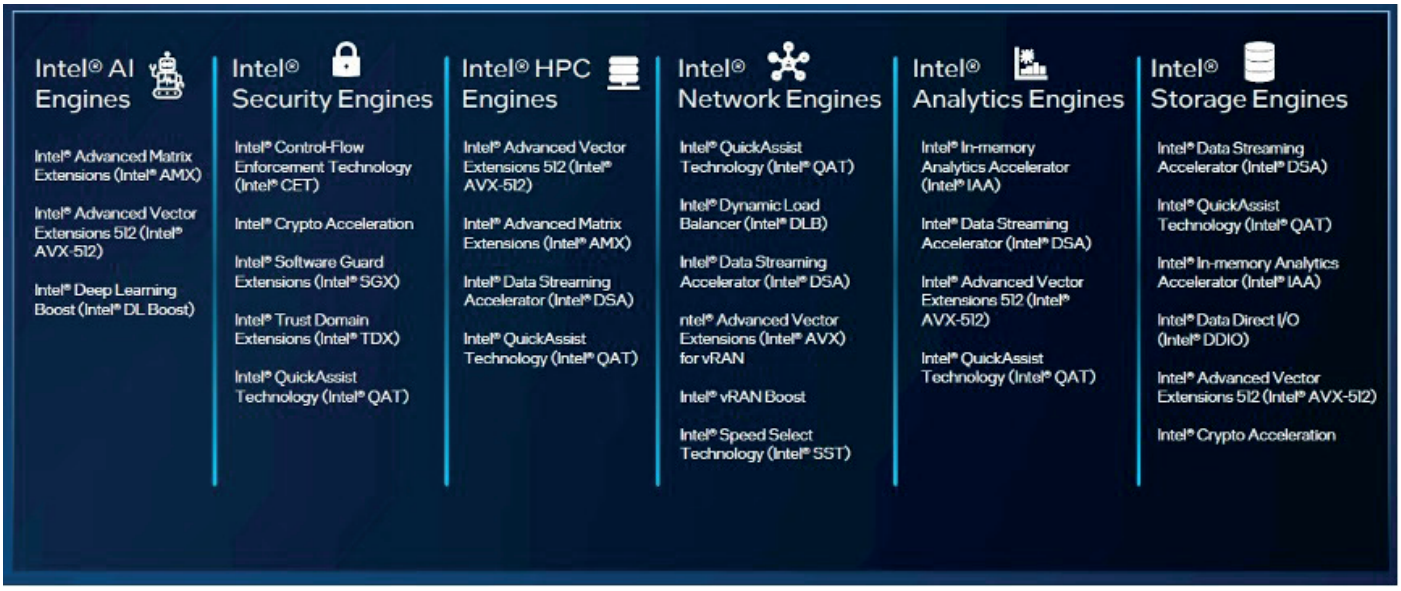


and works with QAT to decrease the memory capacity and memory bandwidth requirements for analytics and database workloads – particularly in-memory databases, but also relational databases and datastores – and helps to boost the throughput of these system programs. IAA is enabled through the combination of Intel's Query Processing Library and its Data Mover Library, and is certified to accelerate RocksDB, Redis, Cassandra, MySQL, and MongoDB as well as several commercial databases. On RocksDB benchmarks, turning on IAA boosted performance by 2.1X compared to not having it on and using Zstd software, which was created by Facebook and which is the reference standard in C, to do compression.

- **Data Streaming Accelerator (DSA):** DSA optimizes streaming data movement and data transformation operations that are common in storage, networking, and analytics workloads. It is accessed through the Intel Data Mover Library. On a top-end 60-core Sapphire Rapids part, DSA was able to push 1.7X higher I/O operations per second on a benchmark juggling large packet sequential reads compared to running Intel's own Intelligent Storage Acceleration Library (ISA-L) software on the Sapphire Rapids cores themselves.

Intel has not been precise about how much area these accelerators take up. But collectively and not including the AMX and AVX-512 units on the Golden Cove cores, the Accelerator Engine die area looks to be a little bit bigger than a PCI-Express controller, which is not very big at all.

SUPERMICRO ACCELERATOR ENGINE WORKLOAD



Having the Accelerator Engines turned off when they are not needed is as important as having them turned on when they are.

As you can see from the IAA case above, it is rarely the case that these Accelerator Engines work alone. They often work in collaboration with each other, and this is how Intel organizes them as grouped functions for accelerating specific kinds of workloads:

Like incandescent light bulbs in days gone by, having the Accelerator Engines turned off when they are not needed is as important as having them turned on when they are. Not every

SUPERMICRO INTEL ACCELERATOR ENGINE SERVER TABLE

	AI	HPC	5G	Storage/Networking	Enterprise/Analytics
Intel® CPU Max (HBM)	✓	✓			
Intel® Advanced Matrix Extensions (AMX)	✓	✓			
Intel® Quick Assist Technology (QAT)				✓	
Intel® Data Streaming Accelerator (DSA)				✓	
Intel® Dynamic Load Balancer (DLB)			✓	✓	
Intel® In-memory Analytics Accelerator (IAA)					✓
Intel® vRAN Boost*			✓		
Suggested Supermicro Systems	SuperBlade Hyper Twin Family	Hyper BigTwin	GrandTwin SuperEdge Hyper-E	Petascale Hyper CloudDC	SuperBlade Petascale MP Servers

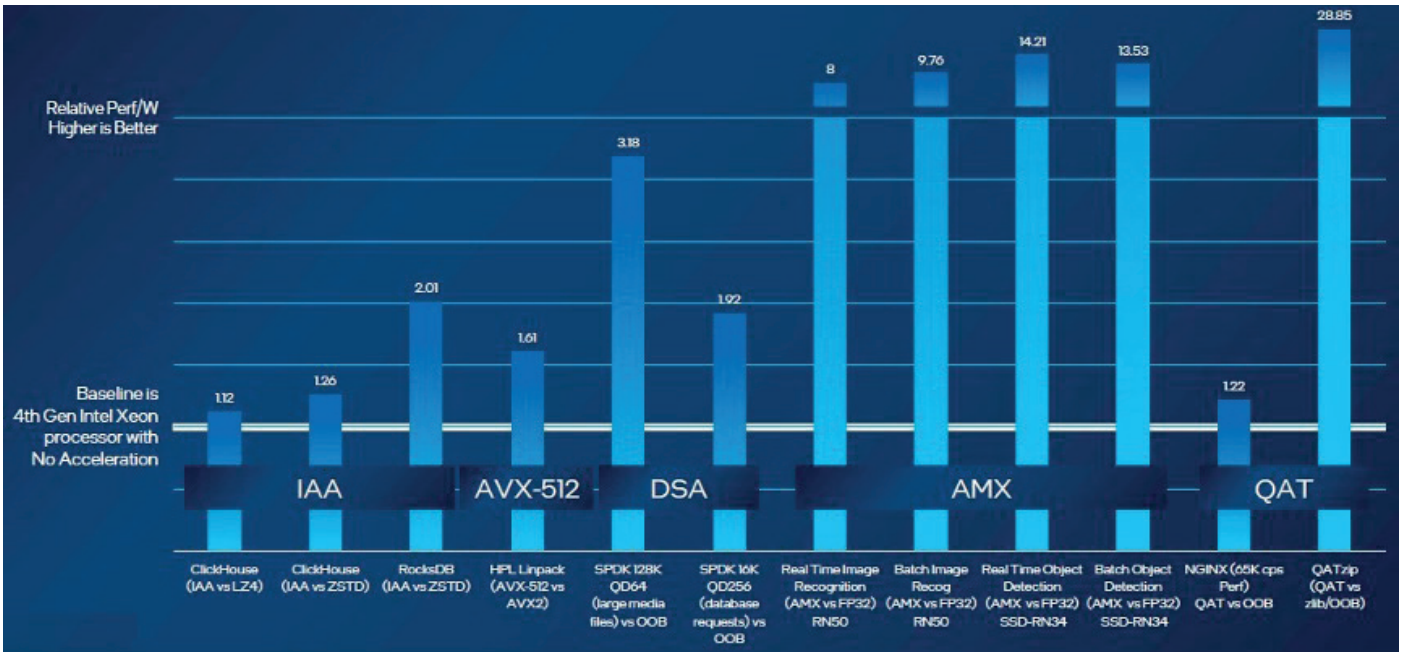
workload will require acceleration, or all kinds of acceleration. In some cases, there should be – and will be – variants of CPUs that don't have some or all accelerated functions included in the architecture, or do not have them activated if they are included in the architecture.

For instance, we fully expect for some CPUs to not have vector math units or matrix math units in their cores for customers whose workloads do not either require these functions, or they want to offload them to beefier accelerators outside of the CPU. With the Max Series CPUs in the Sapphire Rapids family of CPUs for example, which are the ones with HBM2e memory support included, the four Accelerator Engine tiles are actually replaced with HBM memory controllers, so only the AMX, AVX-512, and DSA accelerators that are located in the cores are included.

And in some cases, the Intel Acceleration Engines are available through an On Demand activation mechanism, which means customers only turn them on and pay for them when they actually use them.

Where The Accelerator Rubber Hits The Server Roads

SUPERMICRO ACCELERATOR ENGINE WORKLOAD



Using the Sapphire Rapids in-socket accelerators improves performance and performance per watt, and often by factors of 10X or higher, compared to the prior "Ice Lake" Xeon SPs. Exactly how much depends on the workload:

In the chart above, the baseline is a Sapphire Rapids machine with no acceleration turned on. The performance per watt improvements are shown for those with the appropriate Acceleration Engines turned on. The wattage of the accelerators, with the exception of the AVX-512 and AMX units, is nominal, which means the performance increase is quite significant in many cases.

There is an insatiable appetite to use AI to execute faster and bring new insights and innovations to industries to completely disrupt traditional business models, products and services.

There is an insatiable appetite to use AI to execute faster and bring new insights and innovations to industries to completely disrupt traditional business models, products and services. Intel has chosen to answer customer demand for more compute and greater efficiency by offloading specific workload tasks to power efficient accelerator engines. Doing this frees up the CPU to do general compute operations which in turn increases overall performance. Intel Advanced Matrix Extensions (AMX) is a new built-in accelerator designed to improve the performance of deep-learning training and inference on the CPU for INT8 and BF16 data types. It is ideal for workloads like natural-language processing, image recognition and recommendation systems.

Additionally, AI is an emerging use case on VMware vSAN, and the speed with which AI and deep learning (DL) are evolving means they will soon be built into nearly every enterprise application and analytics tool. Supermicro and Intel have partnered to bring to market an enterprise application solution for VMware vSAN that is optimized by latest generation technologies. X13 SYS-221BT-DNTR systems with 4th Gen Intel® Xeon® Scalable processors feature AMX which gives AI-enabled applications the ability to deliver flexible and efficient performance.

Sponsored by Supermicro.

For more details, visit
www.supermicro.com/en/featured/accelerate-everything

ABOUT SUPERMICRO

Supermicro is a global technology leader committed to delivering first-to-market innovation for Enterprise, Cloud, AI, Metaverse, and 5G Telco/Edge IT Infrastructure. We are a Rack-Scale Total IT Solutions provider that designs and builds an environmentally friendly and energy-saving portfolio of servers, storage systems, switches, software, along with global support services.