Technology Guide

Lowering TCO for Cloud Networking Services

Optimize Cloud Spend with Intel® Xeon® Scalable Processors

Authors

Carlson Bill, Intel

Eric Jones, Intel

Heqing Zhu, Intel

David Lu, Intel

Declan Doherty, Intel

1 Introduction

The type of processor used by cloud compute instances significantly affects the total cost of ownership (TCO) of cloud networking and security services. Improvements in cloud instance processor architecture can lead to enhanced throughput allowing more network traffic to be handled using less resources. Additionally, advanced instruction set capabilities and on-chip accelerators, that manage specific workloads more efficiently than general purpose CPU cores, enable a higher density of applications to run efficiently on fewer instances. Whether measured as reduced up-front costs or as increased capacity, improved TCO can be assumed for each successive generation of cloud instances powered by Intel® Xeon® Scalable processors.

3rd Gen Intel® Xeon® Scalable processors are available in the public cloud. Enterprises can reduce TCO by adopting cloud instances based on the newest Intel processor architecture. The purpose of this technology guide is to review how Intel processors optimize networking and security workloads to lower overall TCO.

Table of Contents

Figures

Tables

Document Revision History

Revision	Date	Description
001	April 2023	Initial release.

1.1 Terminology

Table 1. Terminology

Abbreviation	Description
AVX	Advanced Vector Extension
AWS	Amazon Web Services
CSP	Cloud Service Provider
DL	Deep Learning
GCP	Google Cloud Platform
HCL	HashiCorp Configuration Language
IPsec	Internet Protocol Security
MCNAT	Multi Cloud Network Automation Tool
OneDNN	Intel® oneAPI Deep Neural Network Library
RPM	Red Hat Package Manager
TCO	Total Cost of Ownership

1.2 Reference Documentation

Table 2. Reference Documents

Reference	Source
AWS	https://www.intc.com/news-events/press-releases/detail/1423/intel-xeon-scalable-platform-built-for-most-sensitive
Intel Crypto Acceleration	https://newsroom.intel.com/articles/crypto-acceleration-enabling-path-future-computing
N2 VMs with Intel processor - GCP	https://cloud.google.com/blog/products/compute/compute-engine-n2-vms-now-available-with-intel-ice-lake
3rd Gen Intel® Xeon® Scalable processors	https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html
VMware Carbon Black	https://blogs.vmware.com/security/2022/04/vmware-carbon-black-reduces-cloud-security-costs-up-to-35-with-the-latest-intel-powered-ec2-instances.html
Intel® Deep Learning Boost	https://networkbuilders.intel.com/solutionslibrary/intel-deep-learning-boost-boost-network-security-ai-inference-performance-in-google-cloud-platform-gcp-technology-guide
Amazon EC2 M6i Instances	https://aws.amazon.com/blogs/aws/new-amazon-ec2-m6i-instances-powered-by-the-latest-generation-intel-xeon-scalable-processors/
Amazon EC2 C6i Instances	https://aws.amazon.com/blogs/aws/new-amazon-ec2-c6i-instances-powered-by-the-latest-generation-intel-xeon-scalable-processors/
Specifying a minimum CPU platform for VM instances	https://cloud.google.com/compute/docs/instances/specify-min-cpu-platform

2 Cloud Networking and TCO

Security and networking applications are deployed throughout enterprise cloud infrastructure. Cloud access points, transit gateways, next-generation firewalls, and many other such applications are used throughout cloud deployments to secure corporate data and to protect the networks those applications run on. The instance type and thus the processor chosen to run these critical workloads directly impacts costs.

AWS introduced c6in instance type based on 3rd Gen Intel® Xeon® Scalable processors (Ice Lake) which is ideal for advanced networking applications such as 5G RAN and Core, Multi-Cloud Networking (MCN), and a wide variety of network security applications. These networking instances feature up to 25Gbps of peak networking bandwidth at smaller instance sizes and up to 200Gb/s with the larger instances. The c6in instances are network optimized instances featuring an all-core Turbo frequency of up to 3.5GHz and up to 80Gb/s of EBS (Elastic Block Storage). The instance type is optimal for data plane and high network throughput workloads such as IPsec and SSL. Furthermore, 3rd Gen Intel Xeon Scalable processors integrate acceleration capabilities to improve the performance of networking workloads such as cryptography, compression, network packet analysis,

and AI/ML, resulting in much reduced CPU utilization. These features lower costs by providing similar performance at smaller instance sizes compared to previous generations.

As mentioned above, a good example is the IPsec workload. IPsec (Internet Protocol Security) is a protocol used to secure network traffic by providing encryption and authentication for IP traffic. Both CPU count and network throughput are important resources to consider for IPsec. **Error! Reference source not found.** shows the hourly instance pricing and the benchmark configuration for a user space (VPP) IPsec workload on Amazon EC2 instances. Table 4. compares the gen-to-gen price-performance ratios of the first generation Intel® Xeon® Scalable processor (Skylake) based C5n.4xLarge and C5n.9xLarge instances to the 3rd Gen Intel Xeon Scalable processor based C6in.4xLarge instance. The gen-to-gen performance improvement for the 4xLarge instances, at a packet size of 1420, is 60% resulting in a gen-to-gen costs savings of over 50%.

Table 3. c5n vs c6in IPsec Price & Workload Configuration[1]

Instance	Price/hr (Demand)	Theoretical Max. Throughput	Core/Thread Count	Rx Queue Count Per Port	IPsec Tunnel Count	IPsec Mode/Cipher Suite
c5n.4xlarge	$0.864	25Gbps	6/6	6	256	Tunnel/ESP AES-GCM-128
c5n.9xlarge	$1.944	50Gbps	14/14	14	256	Tunnel/ESP AES-GCM-128
c6in.4xlarge	$0.907	50Gbps	6/6	6	256	Tunnel/ESP AES-GCM-128

Table 4. c5n vs c6in IPsec gen-to-gen Price-Performance Ratios

	c5n.4xlarge		c5n.9xlarge		c6in.4xlarge		c5n.4xlarge-to-c6in.4xlarge		c5n.9xlarge-to-c6in.4xlarge	
Packet Size	Throughput (Gbps)	Rate (Mpps)	Throughput (Gbps)	Rate (Mpps)	Throughput (Gbps)	Rate (Mpps)	Perf. Imp.	Price Perf. (Gbps/$) Imp.	Perf. Imp.	Price Perf. (Gbps/$) Imp.
128	2.349	2.294	2.15	2.100	2.736	2.67	1.16x	11.38%	1.27x	172.75%
512	8.778	2.143	8.308	2.028	10.07	2.41	1.15x	9.28%	1.21x	159.79%
1024	14.651	1.789	16.751	2.045	21.03	2.38	1.44x	36.73%	1.26x	169.08%
1420	17.834	1.572	22.598	1.989	28.834	2.538	1.62x	54.01%	1.28x	173.48%

While the number of vCPUs is held constant across the c5n and c6in 4xLarge instances, it is reasonable to ask what the gen-to-gen price performance ratios might be when comparing instances with similar maximum network throughput. Comparing the c5n.9xLarge and c6in.4xLarge instance types reveals the gen-to-gen costs savings at 50Gbps maximum throughput. While the larger sized c5n.9xLarge instance has better performance at larger packet sizes when compared to the smaller c5n.4xLarge instance, the TCO increases significantly due to the more expensive hourly usage rate of the larger instance. However, for a packet size of 1420 bytes, the c6in.4xLarge instance not only outperforms the larger c5n.9xLarge instance with up to a 28% throughput advantage but also achieves a significant cost savings with an estimated 173% price-performance improvement.

[1] Instance pricing collected from https://aws.amazon.com/ec2/pricing/on-demand/, all prices reflect the us-west-1 region as of July 26th, 2023.

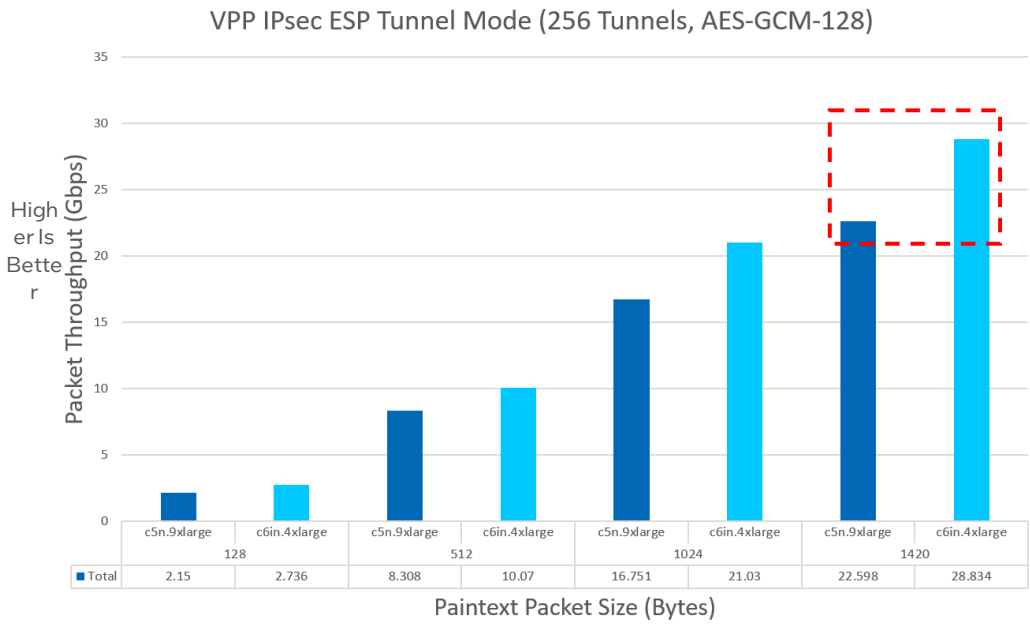

Figure 1. IPsec Perf. c5n.9xLarge vs c6in.4xLarge

3 Cloud Networking Performance Characteristics

As customer demand increases, service vendors who are currently running on instances based on older generation processors may be tempted to simply scale their applications to a larger sized instance of the same type. However, as demonstrated by the above IPsec example, this may not always be the most cost-effective approach. Instead, choosing the optimal instance type that matches a workload's resource requirement can cut costs by preventing overprovisioning or underutilization. Yet, determining the appropriate instance type and size for a given workload can prove to be a difficult and time-consuming task — especially in a multi-cloud environment.

Multi Cloud Network Automation Tool (MCNAT) developed by Intel helps to address this challenge by automating the cloud benchmarking and characterization process. The tool utilizes open-source software like Terraform, Packer, Ansible, and Python to deploy sophisticated workloads and benchmarking tools in an automated fashion from start to end. The tool can help developers and architects to easily compare performance data on cloud instances to predict how well applications perform amongst the various instances of different cloud service providers (CSPs). The tool can also be used for compute resources in on-prem data centers and in edge locations.

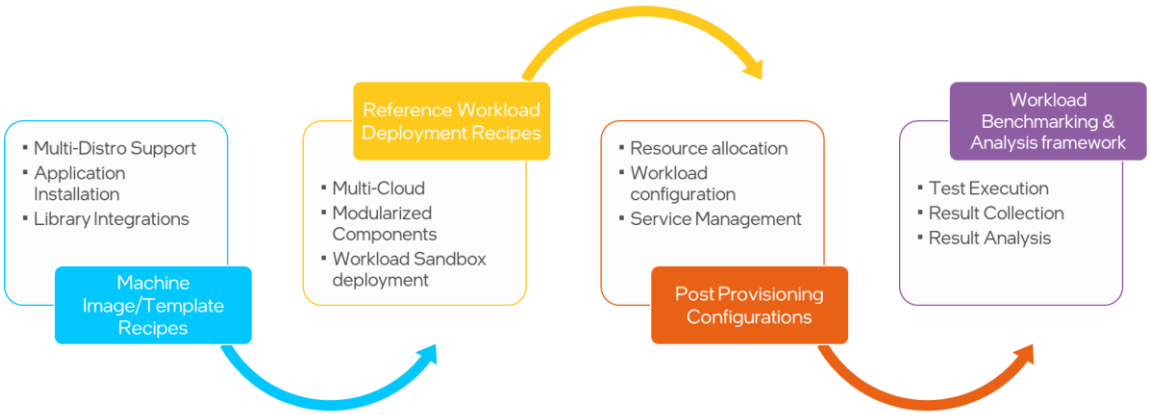

Figure 2. MCNAT Artifact Pipeline

Much of the data collected for this technology guide was gathered with MCNAT. Please contact the author for more information on MCNAT.

4 3rd Gen Intel® Xeon® Scalable Processor Optimizes Cloud Networking

3rd Gen Intel Xeon Scalable processors are based on a balanced, efficient architecture that increases core performance, memory, and I/O bandwidth to accelerate diverse workloads from the data center to the intelligent edge.

The 3rd Gen Intel Xeon Scalable processor provides **1.46x**[2] average gen-on-gen overall platform performance improvement vs. 2nd Gen Intel Xeon processors. This includes **up to 1.60x**[3] higher memory bandwidth, **up to 2.66x**[4] higher memory capacity, and **up to 1.33x**[5] more PCI Express lanes per processor vs. the prior generation Intel Xeon processors. [4]

The 3rd Gen Intel Xeon Scalable processors offer a variety of microarchitecture and Instruction Set Architecture (ISA) enhancements compared to previous generations. They provide customers with built-in acceleration, advanced security capabilities and larger mid-level and last-level caches, which, when combined with higher memory bandwidth, provide significant improvements in performance.

Table 5. M5 vs M6i Network Optimized Instances Pricing and Performance[6]

Instance	Price/hr (Demand)	Core Count	Theoretical Max. Throughput	Memory
M5.8xlarge	$0.9680	32	10Gbps	128GiB
M5.8xlarge	$0.9680	32	10Gbps	128GiB
M6i.8xlarge	$1.5360	32	12.5Gbps	128GiB
M5.16xlarge	$3.0720	64	20Gbps	256GiB
M6i.16xlarge	$3.0720	64	25Gbps	256GiB

Table 6. AWS EC2 M5 vs M6i Instance Performance [7]

Performance Improvement	Cores Count	Memory Bandwidth per vCPU	Network Bandwidth	EBS Bandwidth	Recommended Workloads
15% better	33% more	20% higher	2x	2x+	Network Security Software, SAP, MS SQL Server, PostgreSQL

The 3rd Gen Intel Xeon Scalable processors improve upon Intel® Advanced Vector Extensions 512 (Intel® AVX-512) SIMD instruction set. Intel AVX 512 boosts performance and throughput for the most demanding computational tasks in modern applications including modeling and simulation, data analytics and machine learning, data compression, visualization, and digital content creation. Compared to previous generations, the 3rd Gen Intel Xeon Scalable processor increases memory bandwidth, improves frequency management, and adds two times the FMA (intrinsic for floating point fused multiply-add [FMA] operations), enabling improved performance for specialized workloads.

Figure 3. GCP N2 instance 2nd Gen vs 3rd Gen Intel® Xeon® Scalable processor performance gains

3rd Gen Intel Xeon Scalable processors (Ice Lake) in GCP n2 instance type enables over 30% better price-performance in Google Cloud Environment for variety of workloads compared to n2 machines of the same size running 2nd Gen Intel Xeon Scalable processors (Cascade Lake). n2 instances are ideal for workloads that require high performance per thread or for workloads that specifically take advantage of special instruction sets available in Intel Xeon Scalable processors. To ensure the selection of n2 instances based on 3rd Gen Intel Xeon Scalable processors, it is necessary to specify the desired processor architecture for the n2 machine types as described in [9]. Otherwise, certain instance configurations may default to older generation processors. **Error! Reference source not found.** illustrates compelling performance gains for various workloads running on n2 3rd Gen Intel Xeon Scalable processors vs. 2nd Gen Intel Xeon Scalable processors. [3]

[2] See [125] at www.intel.com/3gen-xeon-config. Results may vary.
[3] 3rd Gen Intel Xeon Platinum 8380 CPU: 8 channels, 3200 MT/s (2 DPC) vs. 2nd Gen Intel Xeon Platinum 8280 CPU: 6 channels, 2666 MT/S (2 DPC).
[4] 3rd Gen Intel Xeon Platinum 8380 CPU: 8 channels, 2 DPC (256GB DDR4) vs. 2nd Gen Intel Xeon Platinum 8280 CPU: 6 channels, 2 DPC (128GB DDR4).
[5] 3rd Gen Intel Xeon Platinum 8380 CPU: 64 lanes of PCI Express 4 per processor vs. 2nd Gen Intel Xeon Platinum 8280 CPU: 48 lanes of PCI Express 3 per processor.
[6] Instance prices change over time and vary by region.

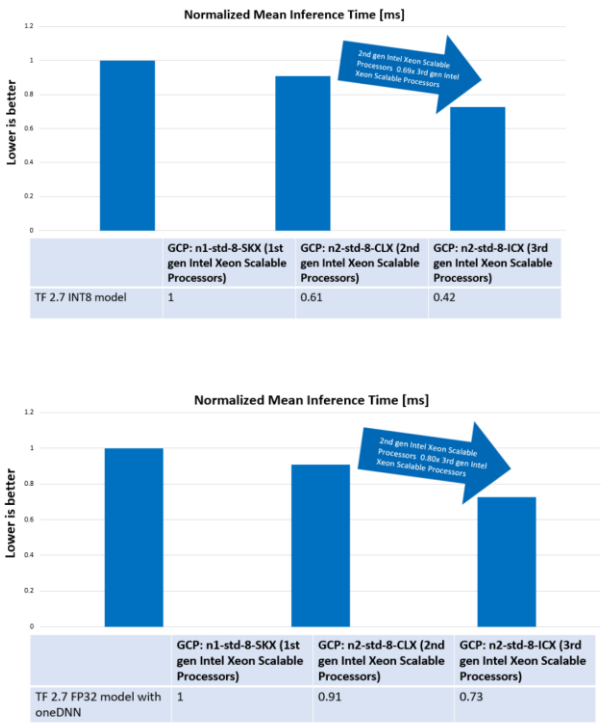

Figure 3. Gen-to-gen comp. of normalized mean inference time of MalConv FP32 frozen model w/ and w/o oneDNN

3rd Gen Intel Xeon Scalable processors incorporate instructions sets and extensions specifically designed for AI/ML inferencing. Intel Advanced Vector Extensions 512 (Intel AVX-512) is a 512-bit instruction set that can accelerate the AI inferencing at the core of applications that prevent network threats and detecting malware.

In addition to Intel AVX-512, Intel® Deep Learning Boost (Intel® DL Boost) is a group of acceleration features introduced in 2nd Gen Intel Xeon Scalable processors that aims to provide significant performance improvements to inference applications built with leading deep learning frameworks such as PyTorch, TensorFlow, MXNet, and Open Neural Network Exchange (ONNX). The foundation of Intel DL Boost is Vector Neural Network Instructions (VNNI), a specialized instruction set that uses a single instruction for DL computations that previously required three separate instructions. We have tested performance of inferencing model using generations of Intel Xeon processors available in GCP and observed continuous performance improvements over generations.

The mean inference time can be dramatically improved under TensorFlow by applying Intel oneDNN and Intel Neural Compressor. With Intel oneDNN and Neural Compressor, the 3rd Gen Intel Xeon Scalable processors-based instance gets performance improvements from an initial 1.34x to 2.39x[7]. Performance tuning with oneDNN and Intel Neural Compressor is both easy and straightforward. Intel DL boost is a standard and universally available feature in 2nd and 3rd Gen Intel Xeon Scalable processors. [6]

Recently, VMware conducted a performance test with VMware Carbon Black Cloud to compare the experience of CPU-intensive applications hosted on different Amazon EC2 instance types. The results conclusively showed that 3rd Gen Intel Xeon Scalable processors with built-in AI accelerators delivered performance boosts in event processing time ranging from 20-35% **compared to** M5 instance types based on previous generation processors**.** Reducing the event processing time provides significant savings of up to 35% cost reduction for VMware Carbon Black Cloud.[5]

5 Summary

3rd Gen Intel® Xeon® Scalable processors provide balanced price/performance across a range of VM shapes and are ideal for a wide variety of workloads that require high network performance or take advantage of the specialized instruction sets and accelerators available in Intel Xeon Scalable processors. This technology guide has described several different instance types across two different CSPs showing how Intel Xeon Scalable processors are able to improve TCO for networking and security workloads. The MCNAT was introduced as a system designed to aid in the identification of the correct Intel architecture-based instance that meets the specific resource and performance needs of customer workloads and to help to avoid the costs associated with overprovisioning or underutilization.

Hardik Tanna, Cooper Trevor made the contribution to this solution paper when they worked for Intel, we want to thank them for their contributions.

[7] For workloads and configuration please visit http://www.intel.com/PerformanceIndex. Results may vary.

Appendix A Platform Configuration

Name	c5n-4xlarge	c5n-9xlarge	c6in-4xlarge
System	Amazon EC2 c5n.4xlarge	Amazon EC2 c5n.4xlarge	Amazon EC2 c6in.9xlarge
Baseboard	Amazon EC2 Not Specified	Amazon EC2 Not Specified	Amazon EC2 Not Specified
Chassis	Amazon EC2 Other	Amazon EC2 Other	Amazon EC2 Other
CPU Model	Intel® Xeon® Platinum 8124M CPU @ 3.00GHz	Intel® Xeon® Platinum 8124M CPU @ 3.00GHz	Intel® Xeon® Platinum 8375C CPU @ 2.90GHz
Microarchitecture	Sky Lake	Sky Lake	Ice Lake
Sockets	1	1	1
Cores per Socket	8		8
Hyperthreading	Enabled	Enabled	Enabled
CPUs	16	36	16
Intel Turbo Boost	Enabled	Enabled	Enabled
Base Frequency	3.0GHz	3.0GHz	2.9GHz
All-core Maximum Frequency	3.4GHz	3.5GHz	3.5GHZ
Maximum Frequency	3.5	3.5	3.5
NUMA Nodes	1	1	1
Prefetchers	L2 HW, L2 Adj., DCU HW, DCU IP	L2 HW, L2 Adj., DCU HW, DCU IP	L2 HW, L2 Adj., DCU HW, DCU IP
Accelerators	QAT:0, DSA:0, IAA:0, DLB:0	QAT:0, DSA:0, IAA:0, DLB:0	QAT:0, DSA:0, IAA:0, DLB:0
Installed Memory	42GB (1x42GB DDR4 2666 MT/s [Unknown])	96GB (1x32GB DDR4 3200 MT/s [Unknown])	32GB (1x32GB DDR4 3200 MT/s [Unknown])
Hugepage Size	1048576 kB	1048576 kB	1048576 kB
Transparent Huge Pages	madvise	madvise	madvise
Automatic NUMA Balancing	Disabled	Disabled	Disabled
NIC	3x Elastic Network Adapter (ENA)	3x Elastic Network Adapter (ENA)	3x Elastic Network Adapter (ENA)
Disk	1x 8G Amazon Elastic Block Store	1x 8G Amazon Elastic Block Store	1x 8G Amazon Elastic Block Store
BIOS	1	1	1
Microcode	0x2006c0a	0xd000331	0xd000331
OS	Ubuntu 20.04.5 LTS	Ubuntu 20.04.5 LTS	Ubuntu 20.04.5 LTS
Kernel	5.15.0-1026-aws	5.15.0-1026-aws	5.15.0-1026-aws
TDP			
Frequency Governor			
Frequency Driver			
Max C-State	9	9	9

Name	n1-std-8	n2-std-8	n2-std-8
Time			
System	Google Google Compute Engine	Google Google Compute Engine	Google Google Compute Engine
Baseboard	Google Google Compute Engine	Google Google Compute Engine	Google Google Compute Engine
Chassis	Google Other	Google Other	Google Other
CPU Model	Intel® Xeon® Platinum 8173M Processor	Intel® Xeon® Gold 6268CL Processor	Intel® Xeon® Platinum 8373C Processor
Microarchitecture	Sky Lake	Cascade Lake	Ice Lake
Sockets	1	1	1
Cores per Socket	8	8	8
Hyperthreading	Enabled	Enabled	Enabled
CPUs	16	16	16
Intel Turbo Boost	Enabled	Enabled	Enabled
Base Frequency	2.0GHz	2.8GHz	2.6GHz
All-core Maximum Frequency	2.7GHz	3.4GHz	3.4GHz
Maximum Frequency	3.5GHz	3.9GHz	3.5GHz
NUMA Nodes	1	1	1
Accelerators	QAT:0, DSA:0, IAA:0, DLB:0	QAT:0, DSA:0, IAA:0, DLB:0	QAT:0, DSA:0, IAA:0, DLB:0
Installed Memory	30GB	32GB	32GB
Hugepagesize			
Transparent Huge Pages	madvise	madvise	madvise
Automatic NUMA Balancing	Disabled	Disabled	Disabled
NIC			
Disk			
BIOS	Google	Google	Google
Microcode			
OS			
Kernel			
TDP			
Power & Perf Policy			
Frequency Governor			
Frequency Driver			
Max C-State	9	9	

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Results have been estimated or simulated.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

1122/DN/WIPRO/PDF 751287-001US