WHITE PAPER

Enterprise and Government Cloud

Making Multi-cloud Work

Seven considerations for optimizing your multi-cloud environment

Multi-cloud Defined

Multi-cloud is the combination of the best-of-breed solutions and services from different cloud providers, including private cloud, to create the most suitable solution for a business. By providing interoperability and portability, multi-cloud gives organizations more flexibility with their cloud solution over different price- points, services offerings, capabilities, and geographical locations. A multi- cloud strategy can employ public, private, and/ or hybrid cloud solutions, depending on the needs of individual companies. Done correctly, multi- cloud creates consistency across the company, independent of the services being consumed.

Table of Contents

Executive Summary

Multi-cloud environments are becoming the standard for enterprises, as public cloud services continue to proliferate, and influence the adoption of cloud-based technologies across the data center. Indeed, spending in cloud computing in 2018 was over $200 billion, and is expected to rise 20 percent in 2019[1].

This increase reflects IT's need to drive the digital transformation of the business. IT is using multi-cloud to move from a cost center to a value center enabling better business decisions, new business strategy, and delivering a cost-efficient model.

These imperatives reflect the priorities of business leaders, 98 percent of whom say IT is critical or very important to their business strategy, with their key priorities being to deliver more value with IT architecture, drive more insights and improve flexibility and scale to react to changes[2]. Nevertheless, IT leaders face real challenges – both architectural and conceptual - to deliver these benefits in a multi-cloud world.

Maximizing the benefit of the multi-cloud environment means abstracting the application from the architecture, having a cloud-aware data strategy, and beginning to think like a cloud service provider (CSP) in the provision of services and platforms to developers, business unit (BU) leads, project managers and other internal customers. IT must understand the value of their data, its 'gravity' to particular applications and locations, the skill of their teams, and the needs of their customers (including security and regulation). Finally, they need a coherent multi-cloud workload placement strategy for new and legacy applications.

This document looks at seven key questions IT leaders need to consider to achieve these goals, and where Intel® technologies, optimizations and ecosystem can help.

Seven questions to consider in making the most of multi-cloud environments

These considerations are based on Intel's experience of establishing our own multi-cloud environment, and of working with and alongside our broad ecosystem to help thousands of enterprises do the same.

1. How do I minimize TCO?

However strategic the IT department is, as a cost center, it is essential to always be looking to maximize the cost benefit to the larger organization. Enterprise IT organizations need to understand how to do this in the context of multi-cloud. This means understanding the cost implications of using public or private cloud for a particular application. It means using containerization and virtualization to improve density and utilization in your on-premises deployment. It could also include a combination of containerization and persistence to virtualize established applications. Data can be expensive to store and move between clouds. A total cost of ownership (TCO) strategy should also include an understanding of the cheapest and most effective ways to store, move and process data.

On an application basis, IT departments need to understand the infrastructural underpinnings of the applications and the skills necessary to support them. They then need to assess the portability of the application. Is it a custom application deployed through containers, legacy software tied to specific infrastructure needs, or are their regulatory requirements on how the data is handled?

Simplified operations are also crucial to multi-cloud TCO, as is a single-pane of glass management view across the various clouds, and maximizing utilization. In the private cloud a core vector of maximum utilization is virtual machine (VM) scheduling. Some applications have significant spikes of activity on a monthly or quarterly basis and are relatively quiet – or totally dormant – the rest of the time. For these episodic applications, such as end-of-quarter financial analyses or payroll, it is important to understand the capacity they need when they burst, and what they use the rest of the time. That way, you can ensure that you neither under- nor over- provision. It is worth analyzing the performance trends across various VMs, and allocating them to systems accordingly so that you can maintain a steadily high utilization (many enterprises aim at 80-85 percent) while leaving adequate room for performance spikes.

Software licencing costs can often be the largest piece of the TCO of a particular application. Much of the licensing pricing of substantial enterprise software is on a per-core basis. To get the most out of the licensing investment it is essential to optimize the per-core performance. Here it could be that the performance bottleneck is not at the CPU, but in storage performance or memory capacity. Developing a balanced system in the private cloud, and understanding the performance characteristics of public clouds is an essential aspect of understanding multi-cloud TCO.

Where organizations are using multiple public clouds, it can make sense for IT to offer a centralized cloud brokerage service. Rather than individual business units (BUs) or developer teams engaging a public cloud instance to get a project started, they work through IT who have control of the process.

For example, Intel IT's internal cloud brokers offer consulting, onboarding, integration, financial stewardship, and security services to BUs when the public cloud is the best fit for an application or workload. Our internal cloud brokers enable us to optimize use of the public cloud to serve our core business enterprise market, where reliability and performance are priorities. We can help business units take advantage of public cloud services when appropriate to the use case, to provide elasticity or other business benefits.

Public cloud services can be very cost-effective for certain workloads and data types, and expensive for others. Across the enterprise, a cloud-first strategy is generally being replaced with a right-cloud strategy. Workloads are being both moved to the public cloud and repatriated from it. This is not necessarily a failure of public cloud, but reflects a developing understanding of where and when workloads are best hosted. For example, it may make sense for developers to create and train a machine learning application in the public cloud, but operationalize it with massive data sets in an on-premise cloud.

Lastly, IT needs to understand its goals as a strategic contributor to the business. What services do you want to deliver to provide business value? This information will be important when developing cloud-based services that minimize TCO.

2. How do I manage my infrastructure to provision new and innovative services in the way the business needs?

Focus on efficiency and ease of use, putting the customer first. This means rethinking processes to be as agile and responsive as an external CSP, as that is what your internal customers expect. Services should be easy and quick to instantiate (a web form, not a lengthy meeting) and should mirror the ease and responsiveness of external suppliers. You must be able to go about provisioning a platform for your developers which allows them to create applications independent of specific architectural considerations. By creating an 'application down' rather than 'infrastructure up' application development model you free your developers to be more agile.

Increasingly, as containerization and microservices become part of the enterprise application stack, IT leaders will need to rethink how they support them. This may mean an internal catalog of 'as-a-service' offerings to the business – for example the provision of bare metal to internal teams. Many CSPs offering container-based architectures or containers-as-a-service (CaaS) often base this on bare metal.

This customer focus can also include creating a pre-approved collection of components and services which can be 'purchased' in an on-line store so that developers and business architects can be facilitated in getting services up

and running as quickly and efficiently as possible, without creating the problems and dangers of 'shadow IT' which can happen when IT is considered a roadblock to swift service provision and users go direct to public cloud providers.

Finally, IT departments will need to assess and adapt their roles in service provisioning. As they become responsible for buying, cataloging and aggregating services rather than building and designing them, the skills necessary to support the business will shift.

3. How do I build a consistent and interoperable environment between clouds which supports optimal TCO and service provision?

Increasingly, as applications are abstracted from architecture, workload placement becomes a calculation based on cost, latency, security, data gravity and performance.

Intel IT's Multi-Cloud Strategy: Focused on the Business

Maximize the business value of the cloud. That simple statement underlies a three-year initiative to transform Intel IT's cloud strategy (see figure 1) by:

- Modernizing our application stack by abstracting it from the infrastructure to enable anything-as-a-service (XaaS) capabilities.
- Focusing on business and application needs, not on infrastructure.
- Determining optimal workload placement to balance cost with capability requirements.
- Validating our approach through communication with fellow travelers and industry benchmarks.

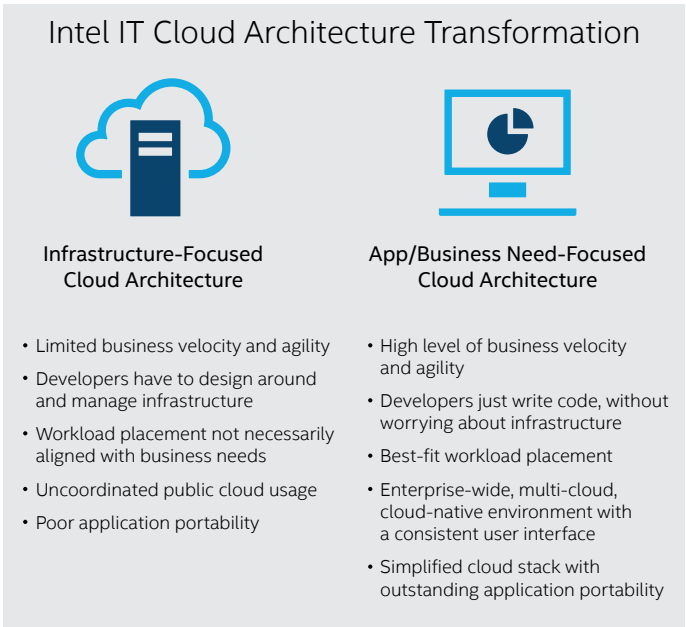


Figure 1. Intel's multi-cloud transition

This has two major implications for the architecture. The first is that it needs to be entirely compatible across clouds. The second is that performance becomes an issue of the whole system – processor, storage, networking – not just per-core performance in the processor.

Compatibility across clouds means that the same (or compatible) technologies, certifications, service level agreements (SLAs) and processes need to be in place in both public and private clouds. All this, though, needs to be in the service of business objectives. Intel itself has done a lot of work in this area, and our own multi-cloud strategy is a good place to look for some directions and lessons.

We use application platforms to abstract underlying infrastructure and deployment details. We also provide internally developed, private database-as-a-service (DBaaS) capabilities, which allow developers to focus on writing the best applications possible, enable enterprise-wide development standards, and permit our applications to modernize. We are completing an application rationalization process to help determine which apps have reached the end of their useful life and which can be migrated to a more appropriate cloud environment.

A cloud strategy that focuses on business and application needs provides the following benefits:

- We gain a high level of business velocity and agility, with built-in redundancy and resiliency in the XaaS capabilities.
- Developers can write code without worrying about infrastructure, allowing IT to make the "best-fit" workload placement.
- A cloud-native environment furnishes a consistent multi-cloud user experience across the enterprise.
- A simplified cloud stack provides outstanding application portability.
- Our transformed cloud strategy supports fast, agile application development.

Read in detail how Intel built its multi-cloud strategy in the paper Intel IT's Multi-Cloud Strategy is Focused on the Business.

4. How do I build a process to best assess my workload requirements?

Deciding which cloud should host which workload, and whether to move them between and across clouds and why, is a critical task in getting the most out of multi-cloud. Some applications can be ported with ease to the public cloud. However, it can be costly to move some applications and data between clouds. Another key vector of workload placement in a multi-cloud environment is to avoid financially disadvantageous vendor lock-in.

Intel, based on research provided by CloudPhysics* and CloudGenera* encompassing over 500+ Enterprise customers, developed a model which rationalizes the approach to workload placement (see figure 2).

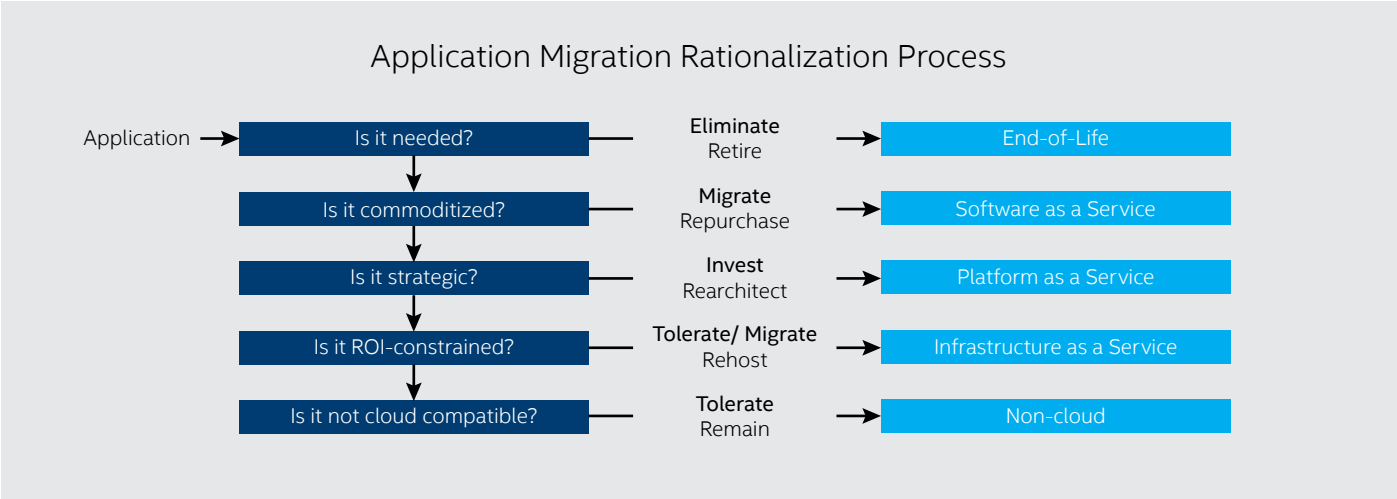


Figure 2. Workload placement rationalization model

It considers whether applications are commoditized, strategic, return on investment (ROI)-constrained or even necessary; whether data or application requirements dictate a particular approach, or whether some applications are cloud-compatible at all.

With a cloud-mature application stack that is abstracted from the infrastructure and that has the ability to systematically identify whether an application is providing business value, we will be well-positioned to take advantage of a multi-cloud environment. Intel has 15,000 software engineers, working on optimizations for thousands of applications to ensure application compatibility across previous, current and future generations of Intel® technology. At Intel IT, we anticipate that we will use our private enterprise cloud for some applications, such as those with strict security requirements or those that are used only internally. We will engage with different public service providers according to the provider's strengths and the application's needs. For example, one provider may have excellent identity management and security features, while another may excel at offering function-as-a-service (FaaS) and CaaS capabilities. As cloud providers innovate, we have the option of moving applications from our private cloud to a public cloud, or from one public provider to another, if doing so better meets our business needs. The flexibility of being able to choose between providers, as well as easily move an application from one hosting environment to another, will enable us to maximize the value of cloud across the enterprise.

The white paper Optimal Workload Placement for Public, Hybrid, and Private Clouds goes into more detail about workload placement considerations.

5. How do I architect for data gravity and affinity in a multi-cloud environment?

It has become a commonplace to say that data volumes are exploding. Nevertheless, more data has been created in the past two years than in the entire previous history of the human race. Over the next 10 years, the world's data will grow 10-fold[3].

Although data is proliferating, most enterprises are able to expose less than 1 percent of their data to analytics[4]. Various factors are contributing to this dilemma:

- Disconnected silos of data across the company that are difficult to access
- Data that has been archived and is not retrievable
- Older data infrastructures not geared for ingesting and/or blending data from multiple sources
- Disparate governance rules and inconsistent metadata and formatting
- The ever-increasing expense of storing data
- Most importantly, the difficulty of figuring out what to retain and what to analyze

It is no surprise that 71 percent of enterprise CIOs[5] indicate that legacy infrastructure is a barrier to their ability to innovate. Most companies must solve these challenges in their data layer before they can truly embark on an analytics strategy.

Nevertheless, as CloudGenera points out[6], the solution is not simply a public cloud-first approach – aggregating all the data in a public environment. This can incur massive costs and can reduce rather than increase flexibility.

In a multi-cloud world, all these issues need to be resolved in the context of application and data placement. Some applications need to be close to the data they service for latency reasons (perhaps for real-time or streaming analytics or where they contain massive data sets). Sometimes it is more practical to move the compute to the data than the data to the compute (for example, edge applications, where massive amounts of data need to be processed and sorted before being sent back to the centralized data center).

For many applications and workloads, data affinity becomes a significant issue when the application is migrated. Accessing massive data sets then porting or replicating the data can be impractical. When using a public cloud, further elements of economic complexity need to be factored in.

Additionally, data governance regulations like the General Data Protection Regulation (GDPR) must be understood and applied as data moves between sites and countries.

Yet many applications need to run close to their data. For some, it may be practical to port or to synchronize the data along with the VM, while for others the data may need to sit in a low-latency repository accessible by all the sites running the VM. Public CSPs offer the ability to create profiles for different types of workload requirements and manage data access for you. When architecting private or hybrid cloud solutions, the same items must be addressed. Intel® architecture implements telemetry which feeds into modern hybrid cloud models to help prioritize utilization and balance workloads in the cloud.

Finally, there are economic factors. It can be very cheap to move data into the public cloud, and very expensive to move it back out again. Given that application and data repatriation is a significant trend, thinking through the likely storage costs and requirements is critical. Likewise, when keeping data on-premise it is important to understand the usage and cost profile of the data and store it as part of a balanced system on the right media. The data storage platform needs to be adaptable, elastic and flexible.

6. How do I build an optimized container and VM migration strategy?

We have talked about the ability to move and migrate applications and workloads between clouds as a benefit of multi-cloud and a key vector in maximizing the benefits of flexibility and cost control which a multi-cloud architecture can provide. Making the most of this flexibility requires thought and planning.

Do you intend to shut down workloads and spin them up somewhere else as a permanent move, effectively replicating the workload on a different cloud? Or do you intend to migrate a 'live' workload, for example using a 'follow the sun' model to maximize efficiency and customer experience? Cold migration of some workloads – for example offloading them to the public cloud – can have economic and practical benefits. Effectively, VMs are moved and data replicated, and the initial instances shut down. Applications where this can be useful include server consolidation, legacy application preservation or simplifying the management of many workloads into one virtualized support model. Cold migration can also be used to evacuate older legacy system platforms into a newer virtualized model whether it is on-premise in private clouds, or off-premise in the public cloud. This helps to reduce the impact of data center legacy debt by removing older, obsolete hardware and moving workloads into modern infrastructure.

'Live migration' is where the VM is moved between sites, either with its attendant data or with dual access to a low-latency network or data architecture. The VM and its workloads continue to run.

Google Cloud's Anthos*, optimized on Intel® technology

Google Cloud's Anthos* is a Google-managed software stack powered by industry-leading application modernization technologies and optimized for Intel® architecture. It transforms IT operations on-premise and in the cloud, letting enterprises build and manage modern hybrid applications across environments using Kubernetes* optimized for Intel architecture. Enterprises can more securely deploy new and existing applications with containers, microservices architecture, and a service mesh delivered and managed by Google and built on Intel® platforms, enabling fast time to market, low administrative overhead, and increased innovation.

This delivers unified cloud resource management, seamless application portability, data mobility, and compelling total cost of owndership (TCO) benefits across cloud environments on-premise and from public cloud providers.

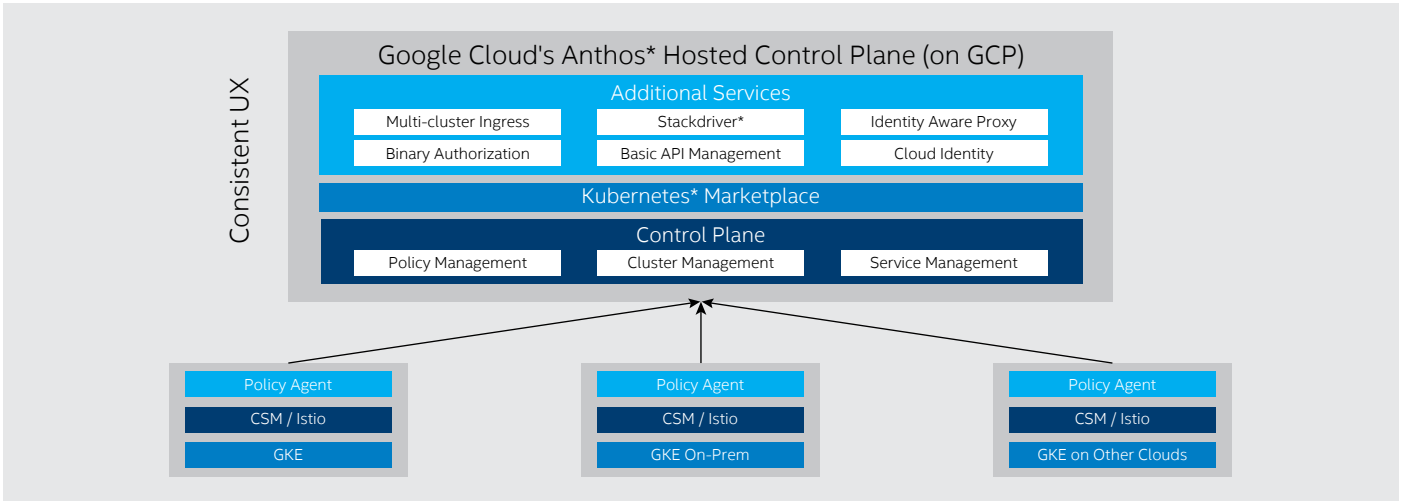


Figure 3. Google Cloud's Anthos* architecture

This is useful in the 'follow the sun' model, or where workloads are being moved between public and private clouds as capacity and priorities require.

The scenario includes such applications as deep learning, inference, modeling and other compute-intensive applications. A key benefit to live, or hot migration is having a well-developed application architecture that allows for your data to be highly accessible with minimal downtime.

Working through a rigorous workload placement methodology can help to clarify decisions and optimize utilization and TCO. Do you have an economic and practical model to establish which workloads should be placed where on your multi-cloud architecture, and those which it makes sense to migrate? You can decide this by considering things like latency, security and integration requirements of the applications and workloads you need to run.

Also take into account any plans to move towards a hyper-converged infrastructure (HCI). You should plan how you will oversee the management and automate the provision of virtualized services. Consider using a 'single pane of glass' management infrastructure so that you can see the performance of all your servers and systems.

To read more about these and other key VM/container migration strategy elements, read the white paper: Some Like it Hot – VM and Container Migration in Hybrid Cloud Environments.

7. How do I go about maximizing security?

In a multi-cloud environment it is necessary to secure data across the public and private clouds, and as information moves between and out of them. This requires platform-wide, hardware-level security and encryption, security-optimized workload placement, and a consistent model to match the security needs of applications with the characteristics of the architecture they are on.

Each cloud instance must have clearly defined boundaries of protection for data coming in and out of it, as well as its applications and workloads. Silicon-level security in the underlying hardware layer can improve the security performance for every VM, container and service that is deployed into modern private cloud solutions.

The north-south traffic must take into account that data will be traveling over the internet (or other network) that is outside of the local data center's control. Using Intel® Xeon® Scalable processors will improve the encryption of data in transit, at rest, and in use. Intel® storage and memory technologies can also encrypt data when in use.

For internal east-west traffic, Intel® Ethernet 700 Series Network Adapters provide the required bandwidth, with up to 40GbE between systems inside the data center. These network adapters move data faster to ensure the Intel Xeon Scalable processors can focus on workloads.

Ensuring there is adequate networking bandwidth to enable a multi-cloud scenario is imperative to good performance. Modern hardware-based security reduces the overhead of cloud security management so each virtual networking interface and firewall rule can be centrally managed via cloud-capable software tools.

Intel® technology enabling multi-cloud success

A successful multi-cloud strategy requires a fully tuned platform – not just processor but storage, networking and software optimizations working together to provide the performance required. Intel provides an optimal technology foundation, protected by hardware-level security, which is architected for real-time business and allows you to scale and extend workloads.

Architected for real-time business

Run the most compute-intensive workloads – current and future – fast and cost effectively on Intel technology-based cloud instances, including SAP HANA*, high-performance computing (HPC), and artificial intelligence (AI).

- 2nd generation Intel® Xeon® Scalable processors' advanced compute core provides leaps in compute performance, memory capacity and bandwidth, I/O scalability and new core frequencies of up to 3.8 GHz up to 4.4 GHz with Intel® Turbo Boost Technology, which deliver workload-optimized cloud performance for compute- and memory-hungry applications.
- Intel® Deep Learning Boost (Intel® DL Boost) brings new embedded performance acceleration for AI workloads in the 2nd generation Intel Xeon Scalable processor, and up to 30x performance improvement for Inference workloads compared to the previous generation[7].
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) offers up to 2x more flops per cycle than previous generation technologies[8], enabling improvements in workload speed and data application.
- Intel® Optane™ DC persistent memory is uniquely architected for inherent persistence, retaining data even during power cycles without requiring external power back-up components.
- Up to 36 percent increased VM density per node on Microsoft Windows Server* using Intel Optane DC persistent memory[9].
- Modernize your infrastructure with 2nd generation Intel Xeon Scalable processors and Intel Optane DC persistent memory to optimize core business applications and workloads:
  - Up to 50 percent node reduction at up to 39 percent lower TCO on Microsoft SQL Server[10]
  - Up to 8x higher performance on Spark SQL[10]
  - 100 percent more capacity at 34 percent less cost per TB on SAP HANA[10]

- Optimize your cloud performance with 2nd generation Intel Xeon Scalable processors, Intel® SSD Data Center (Intel® SSD DC) and Intel Ethernet 700 series products.

Optimal technology foundation

Harness seamless application portability, data mobility, and compelling TCO benefits across the broadest range of workloads and services, whether compute, networking, or storage.

- Intel® architecture is supported by the largest independent software vendor (ISV) ecosystem, providing ultimate application migration and portability to the public cloud.
- Intel architecture-optimized offerings, through CSP marketplaces, provide turnkey solutions to jumpstart adoption of cloud services or migrate existing infrastructure to the public cloud.
- Intel architecture provides a foundation for application and workload mobility/portability across clouds, allowing you to use current skills and accelerate time to market.
- Intel® hardware-assisted technologies, Intel® FPGAs, Intel® QuickAssist Technology (Intel® QAT), and Intel Optane DC persistent memory, provide flexibility and scalability in architecting and adopting new workloads.
- 2nd generation Intel Xeon Scalable processors provide 2x better performance[11] and 30x improvement in inference[7] by running general compute workloads and scaling to meet most compute- intensive workloads without changing your hardware or compute environment.
- A growing suite of developer-focused frameworks, training, open source libraries and tools created for Intel Optane DC persistent memory including the Persistent Memory Developer Kit (PMDK) and VTune™ Amplifier.

Scale and extend workloads

More securely and reliably scale and extend workloads from enterprise to cloud, without the need for reconfiguration, application changes, or testing, thanks to broad deployment of Intel® Xeon® processors.

- Intel architecture delivers hardware-enabled security capabilities directly on the silicon to help protect every layer of the compute stack (hardware, firmware, operating systems, applications, networks, and the cloud).
  - 2nd generation Intel Xeon Scalable processors deliver hardware-enhanced threat detection through Intel® Threat Detection Technology (Intel® TDT).
  - Intel® Security Libraries for Data Center (Intel® SecL - DC) are the building blocks of a variety of security usage models and layers that can be rooted in hardware-based capabilities.
- Intel® technology-powered cloud instances deliver the ability to scale globally from two sockets to eight sockets and beyond to meet business needs (bare metal instances).
- Seamlessly scale using Intel® Virtualization Technology (Intel® VT) for live application migration between clouds, deploy new applications and scale up or down as needed to meet changing requirements.
- Widest depth and breadth of data center applications optimized, developed, and tested on Intel architecture, providing workload scale from private to public cloud.
- Dynamic crypto and compression processing with full Intel® Key Protection Technology (Intel® KPT) via integrated Intel QAT.

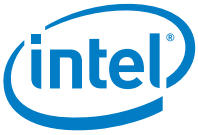
Conclusion

Multi-cloud technologies offer huge benefits to the business and IT teams alike. With careful thought and planning to select the right combination of technologies, it's possible to create a cloud-native platform that frees applications to run at their best, independent of the architecture that supports them.

Further Reading:

- Streamlined multi-cloud solutions from Intel and collaborators
  - Video: Intel's Multi-Cloud Strategy
  - Intel® Select Solutions for Hybrid Cloud
  - Learn about Workload Optimization in a Multi-Cloud Environment
- Explore streamlined multi-cloud and hybrid cloud solutions optimized on Intel from Google* Cloud, Amazon Web Services* (AWS*), and Microsoft Azure*
  - Google Cloud's Anthos*
  - AWS
  - Microsoft Azure

Solution Provided By:



Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information, visit intel.com/benchmarks.

[1] https://resources.cloudgenera.com/wp-content/uploads/2019/03/BlueprintForCloudSuccess-2.pdf drawing on data from Gartner and Wikibon.

[2] IDG Research: "Stakes Rise for IT: The IT Transformation Journey" https://pages.insight.com/rs/366-UKY-221/images/Infographic-IDG_Fall_2017-Final.pdf

[3] Source: https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#7efdd94b17b1

[4] Source: https://hbr.org/2017/05/whats-your-data-strategy

[5] Source: https://www.logicmonitor.com/blog/83-percent-of-enterprise-workloads-will-be-in-the-cloud-by-2020/

[6] Source: https://resources.cloudgenera.com/wp-content/uploads/2019/03/BlueprintForCloudSuccess-2.pdf

[7] 30x inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost: Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/2933 MHz), BIOS: SE5C620.86B.0D. 01.0241.112020180249, Centos 7 Kernel 3.10.0- 957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d9 4195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer synthetic Data: 3x224x224, 56 instance/2 socket, Datatype: INT8 vs. Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb /s, 25nm, MLC). Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=5 6, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690 af267158b82b150b5c.Inference measured with "caffe time -- forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

[8] Up to 2x average generational gains on 2-socket servers with new 2nd Gen Intel® Xeon® Platinum 9200 processor. Geomean of est SPECrate2017_int_base, est SPECrate2017_fp_base, STREAM Triad, Intel® Distribution of LINPACK, server-side Java*. Platinum 92xx vs.. Platinum 8180: 1-node, 2x Intel® Xeon® Platinum 9282 cpu on Walker Pass with 768 GB (24x 32GB 2933) total memory, ucode 0x400000A on RHEL7.6, 3.10.0-957.el7. x86_64, IC19u1, AVX512, HT on all (off Stream, LINPACK), Turbo on all (off Stream, LINPACK), result: est int throughput=635, est fp throughput=526, STREAM-Triad=407, LINPACK=6411, serverside Java=332913, test by Intel on 2/16/2019. vs..1-node, 2x Intel® Xeon® Platinum 8180 cpu on Wolf Pass with 384 GB (12 X 32GB 2666) total memory, ucode 0x200004D on RHEL7.6, 3.10.0-957.el7.x86_64, IC19u1, AVX512, HT on all (off Stream, LINPACK), Turbo on all (off Stream, LINPACK), result: est int throughput=307, est fp throughput=251, STREAM-Triad=204, LINPACK=3238, serverside Java=165724, test by Intel on 1/29/2019.

[9] 36percent more VMS per node for multi-tenant virtualized OLTP databases with Intel® Optane™ DCPMM: 1- node, 2x 26-core 2nd Gen Intel Xeon Scalable Processor, HT on, Turbo on, 768GB, 0(24 slots / 32GB / 2666 DDR)1x Samsung PM963 M.2 960GB, 7 x Samsung PM963 M.2 960GB, 4x Intel SSDs S4600 (1.92 TB), 1xIntel X520 SR2 (10Gb), Windows Server 2019 RS5-17763, OLTP Cloud Benchmark, test by Intel as of 1/31/2019 vs. 1-node, 2x 26-core 2nd Gen Intel Xeon Scalable Processor, HT on, Turbo on, 192GB, 1TB(12 slots / 16 GB / 2666 DDR + 8 slots /128GB / 2666 Intel Optane DCPMM), 1xSamsung PM963 M.2 960GB, 7x Samsung PM963 M.2 960GB, 4x Intel SSDs S4600 (1.92 TB), 1x Intel X520 SR2 (10Gb), Windows Server 2019 RS5-17763, OLTP Cloud Benchmark, test by Intel as of 1/31/2019

[10] SQL server/Hyper-V* - Multi-Tenant Virtualization. Config1-DDR4 (Similar Cost), Tested by Intel 01/31/2019, Platform: Confidential-Refer to M. Strassmaier if a need to know exists, # Nodes: 1, # Sockets: 2, CPU: CLX b0-steping 26c @2.6GHz, Cores/socket, Threads/socket: 26/52, ucode: 0x4000014, HT: On, Turbo: On, BIOS version: C2030.BS.1C03.GN1, BKC version: WW42, AEP FW version: 5253, System DDR Mem Config: slots/cap/run-speed: 24 slots/32GB/2666, Total Memory/Node (DDR,DCPMM): 768GB, 0, Storage boot: 1 x Samsung PM963 M.2 960GB, Storage – application drives: 7 x Samsung PM963 M.2 960GB, 4 x Intel SSDs S4600 (1.92TB), NIC: 1 x Intel X520 SR2 (10GB), PCH: LBG QS/PRQ – T – B2, OS: Windows Server 2019 RS5-17763, Workload & version: OLTP Cloud benchmark.

Config2-AEP (Similar Cost), Tested by Intel 01/21/2019, Platform: Confidential-Refer to M. Strassmaier if a need to know exists, # Nodes: 1, # Sockets: 2, CPU: CLX b0-steping 26c @2.6GHz, Cores/socket, Threads/socket: 26/52, ucode: 0x4000014, HT: On, Turbo: On, BIOS version: C2030.BS.1C03.GN1, BKC version: WW42, AEP FW version: 5253, System DDR Mem Config: slots/cap/run-speed: 12 slots/16GB/2666, System DCPMM Config: slots/cap/run-speed: 8 Slots/128GB/2666, Total Memory/Node (DDR,DCPMM): 192GB, 1TB, Storage boot: 1 x Samsung PM963 M.2 960GB, Storage – application drives: 7 x Samsung PM963 M.2 960GB, 4 x Intel SSDs S4600 (1.92TB), NIC: 1 x Intel X520 SR2 (10GB), PCH: LBG QS/PRQ – T – B2, OS: Windows Server 2019 RS5-17763, Workload & version: OLTP Cloud benchmark.

	1 - Baseline			2 - Config Description	
# of Systems	1			1	
Memory Sub System Per Socket	DRAM - 384GB (12x32GB)			2GB (4x128GB AEP + 6x16GB DRAM, 2-2-1, Memory Mode	
CPU SKU | # per System	8270 (CLX, Plat, 26core)	2		8270 (CLX, Plat, 26core)	2
Storage Description | Total Storage Cos	# of HDD/SDD's	$7,200		# of HDD/SDD's	$7,200
SW License description | Cost per Syste	SW Cost (per/core or per system)	$0		SW Cost (per/core or per system)	$0
Relevant Value Metric	22.00			30.00	
Type of System	DRAM - Purley			AEP - Memory Mode	
CPU & Platform Match	TRUE			TRUE	
	1 - Baseline			**2 - Config Description**	
	Description	**Total Cost**		**Description**	**Total Cost**
CPU Cost	2 x 8270 (CLX, Plat, 26core)	$14,810		2 x 8270 (CLX, Plat, 26core)	$14,810
Memory Subsystem	Total Cap: 768GB (385GB/Socke	$5,808		Total Cap: 1024GB (512GB/Socket	$5,500
DRAM	24x32GB	$5,808		12x16GB	$2,124
AEP	N/A - DRAM Based System	$0		8x128GB	$2,276
Storage	# of HDD/SDD's	$7,200		# of HDD/SDD's	$7,200
RBOM	Chassis; PSUs; Bootdrive etc.	$1,300		Chassis; PSUs; Bootdrive etc.	$1,300
SW Costs	SW Cost (per/core or per system	$0		SW Cost (per/core or per system)	$0
Total System Cost		$29,118			$28,810
Total Cost	1 Sys x $29,118	$29,118		1 x Sys $28,810	$28,810
System Cost	1			0.98942235	
Indexed Value Metric		1.00			1.36
Indexed Value/S	***Baseline***	1.00			1.38
Pricing Guidance as of May 21, 2019, valid until Jun 29, 2019					

[11]2x average performance improvement compared with Intel® Xeon® Platinum 8180 processor. Geomean of est SPECrate2017_int_base, est SPECrate2017_fp_base, Stream Triad, Intel Distribution of Linpack, server side Java. Platinum 92xx vs Platinum 8180: 1-node, 2x Intel® Xeon® Platinum 9282 cpu on Walker Pass with 768 GB (24x 32GB 2933) total memory ucode 0x400000A on RHEL7.6, 3.10.0-957.el7.x86_65, IC19u1, AVX512, HT on all (off Stream, Linpack), Turbo on all (off Stream, Linpack), result: est int throughput=635, est fp throughput=526, Stream Triad=401, Linpack=6411, server side jaca=332913, test by Intel on 2/16/2019 vs. 1-node, 2x Intel® Xeon® Platinum 8180 cpu on Wolf Pass with 384GB (12x 32GB 2666) total memory, ucode 0x200004D on RHEL7.6, 3.10.0-957.el7.x86_65 C19u1, AVX512, HT on all (off Stream, Linpack), Turbo on all (off Stream, Linpack), result: est int throughput=307, est fp throughput=251, Stream Triad=204, Linpack=3238, server side java=165724, test by Intel on 1/29/2019. https://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/2nd-gen-xeon-scalable-processors-brief.pdf