GESTALT IT

AUGUST 23, 2022

DIGITAL INFRASTRUCTURE AT DATACENTER SCALE

IT'S ALL ABOUT THE ECOSYSTEM

A GESTALT IT WHITEPAPER

ATHER BEG
ALASTAIR COOKE
STEPHEN FOSKETT
CRAIG RODGERS

SPONSORED BY

intel®

Table of Contents

Executive Summary

Modern information technology is defined by abstraction, aggregation, and combination of resources according to policy. Rather than building individual vertically-integrated systems, infrastructure components are brought together as a platform to support higher-level services. This model is often treated as a private or hybrid cloud, since it is the same approach taken by public cloud providers. But it is equally suitable for modern virtualized servers and virtual desktop infrastructure. Thus, modern IT architectures are about the whole platform, not any individual component.

Until recently, datacenter infrastructure was a sea of complexity. Technology selections were often made independently per application, leading to bespoke infrastructure maximized for each use case. This may have been generally accepted in previous decades, but the rise of virtualization, containers, and cloud-native platforms has made adaptability more important than extreme performance. Simplification is now the primary theme for on-premises IT infrastructure, with architects focusing on sets of infrastructure components that have been designed to operate together.

For a long time, application performance was driven by having more and faster CPU cores in each server, but cloud-native and software-defined infrastructure is not server-bound. At the same time, work is being off-loaded to specialized accelerators, both inside the CPU and on non-CPU processors. GPUs were among the first such accelerators, but now we have NICs and other add-in cards with their own optimized processors. The other dimension is adding specialized instructions to the core CPU that implement in hardware a feature that would use a lot of CPU time if it were implemented in software. Intel has been adding workload specific CPU extensions since SSE in the Pentium III processor in 1999 and has supported add-in accelerator cards even longer.

A CPU alone is obviously not a complete server or platform; memory, storage, and external connectivity are all essential. Just like selecting an engine, gearbox, and tires for a car, the components of a server are all interdependent; putting a massive engine or mud tires on a car will not transform its performance. Intel's portfolio of server components includes the Xeon Scalable CPU line, Optane technology for SSDs and persistent memory, and Ethernet 800 for fast connectivity. More accelerators are coming in the next generation, along with CXL to help provide more flexible access to resources, including memory expansion. All of these are designed to work together seamlessly and widely supported by popular software.

One must also consider support and availability. Every major server, operating system, and application vendor has a relationship with Intel. This ensures that their products are optimized to work together on

Intel portfolio server. No other company can match Intel’s range of products and volume of production. All of these factors combine to make Intel the preferred platform for the modern datacenter.

Intel’s Original Equipment Manufacturer (OEM) partners have developed server and cloud products to provide a unified platform for modern applications that includes Intel-based processing, I/O, memory, storage, networking, and accelerators with both hardware and software to provide a secure and sustainable solution. Each of these areas has seen radical upgrades over the last decade, with a proliferation of CPU cores accompanied by the advent of persistent memory, FPGA, processor offloads, and advanced I/O and connectivity. Servers will continue to develop in the future, as technologies like UCIe, CXL, and silicon photonics bring greater flexibility and scale.

Because Intel has invested in all of these technologies, they remain a leader in datacenter and cloud servers. The portfolio of Intel server components delivers proven performance in datacenters for enterprise workloads. These components are designed to work together, minimizing the risk of unplanned outages due to component interoperability without requiring extensive integration testing. Combined with Intel’s support for the entire computing ecosystem, from edge to datacenter to cloud, this explains the enduring popularity of Intel solutions.

1. Introduction

The purpose of this whitepaper is to explore trends in IT infrastructure and evaluate how well the Intel product ecosystem matches the needs of modern server architects. It also discusses many technological and business considerations for server infrastructure as well as the risks and benefits of making the decisions today for the infrastructure of tomorrow.

The **Datacenter Trends** section is largely a framework for evaluating different approaches to IT infrastructure and is presented as the context in which vendor solutions are evaluated today. We begin with a look at the ways in which cloud computing is affecting the modern enterprise and the push and pull factors for public cloud. We then look at the emergence of private cloud and the geopolitical factors that make it attractive for global businesses. Finally, we examine the technologies that are contributing to these developments, including accelerators and disaggregated infrastructure.

We will then examine how Intel's portfolio and ecosystem fits into several specific use cases that are familiar to the authors. The **Real-World Use Cases** section illustrates how specific technologies are applied to fulfill technical and business requirements. These use cases include repatriation of cloud workloads, virtual desktop infrastructure, and artificial intelligence, all of which are impacting enterprises and driving investment today.

The final major section examines a hypothetical **Perfect Configuration** for modern server infrastructure which recognizes that optimization is only useful where it is applied to a limiting factor. This "perfect" configuration would balance investment in processors, memory, storage, networking, and accelerators with the benefit each can provide.

2. Datacenter Trends

Although technology enthusiasts tend to focus on CPU core count, performance numbers, network and storage throughput, and similar objective metrics, buyers of enterprise datacenter and cloud computing infrastructure have a more utilitarian approach. A server is much more than a CPU. Computing infrastructure has evolved from server, network, and storage to an integrated cloud computing platform. For this reason, buyers are driven more by the need to build a flexible, integrated, and sustainable solution than by the specific characteristics of individual components.

When evaluating servers, one must begin with the cloud. Today's workloads are defined by the DevOps approach, with software-defined infrastructure from the datacenter to public cloud. The focus is on industry-standard server components rather than customized devices. For this reason, the server platform has become more critical even as individual components are less of a focus. The largest HPC and cloud infrastructure exemplifies this software-centric architecture, being constructed of racks of individual servers that make up a unified platform running distributed applications.

This drive to the cloud provides many benefits, but enterprises are also increasingly aware of the limits of public cloud services. Concerns about data sovereignty and availability of services have driven some to repatriate workloads to on-premises environments that replicate the functionality of the public cloud within a datacenter or local hosting environment. Many organizations are building hybrid clouds that combine datacenter, hosted, and public infrastructure in a unified platform.

This section examines these drivers for modern servers, beginning with a discussion of the modern cloud platform. We then consider the benefits of public cloud as well as the challenges that are driving applications back on-premises. We will also consider the technology that is used in these environments from a holistic perspective, and look for future trends.

2.1 Flexible Cloud Infrastructure

Cloud is a way of operating, rather than a location. There are huge benefits to using public cloud platforms, but to get the most out of those benefits may require changes that are not cost-effective for every application. The application changes required are also different for different combinations of applications and cloud providers. The result is that many enterprise organizations will use multiple public cloud providers as well as on-premises clouds to deliver the myriad of applications and services they require.

2.1.1 Why Cloud?

Public cloud platforms tend to work best for applications that change a lot, either in terms of software or workload. Rapid software changes are a hallmark of DevOps methodologies and digital transformation, where business processes are being developed through software. These environments require highly productive software development and suit cloud-native platforms. Load changes might be a seasonal variation such as consumer retail, or an application that is driven by daily, weekly, or monthly periodic cycles. These applications tend to have highly variable resource utilization. Peak utilization might be ten or twenty times the average utilization.

Both of these use cases involve variability over time and the ability to respond rapidly to these changes. Often these applications are the ones where organizations interact with external customers or users, typically through web or mobile applications.

2.1.2 Why On-Premises?

On-premises platforms tend to favor applications where the business requirement is relatively static or predictable and the users or customers are internal. Consistency and reliability are baked into the processes and often innovation is less important than the value of the application that was built years ago. Often these applications are internal facing, only used by staff to operate the business and fulfill orders. Occasionally there are data sovereignty issues that prevent applications being migrated to the public cloud.

2.1.3 Blurred Lines

Over the last few years, the lines between public cloud and on-premises have blurred. Most public cloud platforms offer some form of VMware vSphere service, allowing on-premises applications to run unmodified in a public cloud datacenter. At the same time, many public cloud platforms are offering an on-premises option to deliver the cloud-native services that suit more dynamic applications. Google Distributed Cloud and AWS Outposts are examples of on-premises delivery of public cloud infrastructure. These services can be very helpful in allowing more options for where applications can reside, delivering the appropriate functionality wherever it is that the non-functional requirements demand each application run.

2.1.4 Non-Factors

A few of the factors which were perceived as deal breakers for public cloud in the past (e.g. security, compliance, and reliability) are simply implemented differently in the public-cloud. Security in public-cloud platforms generally requires the defense-in-depth and least-privilege approaches that are accepted best-practices but often poorly implemented on-premises. Compliance is often more complex in multi-tenant, public-cloud platforms, since responsibility is shared by both the provider and the customer. There is often guidance from the cloud platform on the methodologies that build compliant applications using their services. The perception of the reliability of public-cloud platforms is often driven by the fact that when an outage occurs, its effects are highly visible. On the other hand, outages of on-premises datacenters seldom make the news even though they happen every day. Almost everyone who has worked in a datacenter environment has a story of the day the whole datacenter was taken down by the security guard, the fire suppression system, or a misconfiguration.

Another misconception is the lack of non-CPU hardware capabilities on public clouds. The narrative that "software is eating the world" often also says that generic CPUs are all that is required, and that this is what the public cloud delivers. The reality of public cloud is that cloud instances that expose specialized CPU features are widely available, with most of these being centered around the capabilities of Intel Xeon Scalable processors. CPU-based acceleration, such as AES-NI or security features, make a huge difference in the ability to operate an application and are available both on-premises and in the public cloud. GPUs, AI/ML accelerators, NVMe storage, and specialized PCIe attached accelerators are also available on most public cloud platforms as they are in on-premises server platforms.

2.1.5 Multi-Cloud Challenges

Spreading applications across multiple datacenters is a complex process, and even more so when the datacenters are different cloud providers with different capabilities. The greatest value of each cloud platform comes from using the more advanced and platform specific features of the provider. These different features are also drivers for deploying different applications onto different cloud platforms. In the past, Google has been more associated with analytics applications, Azure more with Windows applications, and AWS more for new web and mobile applications. Modern enterprise organizations will require all of these types of applications and will need to decide whether to embrace the complexity of multi-cloud development and deployment, or standardize on a single platform which may not be as well optimized for any one application.

Taking cloud-native storage as an example, AWS provides S3, Azure has Blob storage, and GCP has Cloud Storage. These three services fulfill the same fundamental need for large amounts of unstructured storage which is used by cloud-native applications. Of these, S3 has the most widely-used interface (API), so many third party storage products provide an "S3 compatible" API. But both Azure Blob Store and Google Cloud Storage have their own APIs which make the unique features of those services available.

To address the cross-compatibility of public clouds and on-premises clouds, customers may use an abstraction layer that is available on all of the platforms, most commonly Kubernetes-orchestrated containers. These abstraction layers hide the complexity of the different platforms to enable workload mobility, but also nullify the different capabilities of each cloud. Enterprises must choose how much to embrace the unique value of each platform, and thereby complicate moving from platform to platform. The decision may be different for each application: Some will use the unique features of one platform and thus be tightly coupled to it, while other applications may sacrifice these to enable cross-platform mobility.

Data movement is always a challenge at scale: Moving large data sets between datacenters can be time consuming and expensive. Therefore, whether the datacenters are on-premises or in public clouds, the best data movement is one that is no longer needed because the applications that share data are contained inside the same cloud or datacenter. An application that generates a terabyte of data per day is easy to accommodate, since it requires well under an hour to transfer that volume of data at 10 gigabit speed. An application that generates a petabyte of data is more of a problem, since it would require over 100 gigabits of bandwidth. It is often more practical and cost-effective to process the bulk data close to where it is generated and transfer only the results between sites. We often see this model in machine learning, IoT, and edge computing, where real-time processing is distributed and only processed data is centralized.

2.2 Complexities of Private Cloud Infrastructure

Design and implementation of a private cloud infrastructure, and ensuring its smooth operation going forward, is a significant undertaking. Organizations committing to their own private cloud infrastructure also accept that it will bring a set of challenges that they will need to budget and prepare for; challenges that they can avoid by utilizing a service provider or the public cloud, thereby making them the vendor’s problem.

2.2.1 Whose Datacenters?

One of the most important decisions early in the process is to decide where to host the infrastructure. It is crucial to make the right choice here because getting it wrong could have major cost implications down the line. Organizations that do not have sufficient on-premises space and utilities provision may have the decision made for them. Otherwise, careful consideration is required whether to self-host the infrastructure or place it in a service provider datacenter.

Given such service providers have ample space, connectivity to multiple carriers and resilient utility provisioning, they are well-equipped to take that responsibility away from an organization, enabling it to focus on its business. However, these benefits come at a cost and with service level and access restrictions – none of which are concerns when an organization hosts its infrastructure itself.

Resilience is another major factor to consider. Building an infrastructure to withstand the failure of individual hosts is easy but preventing service failure despite loss of an entire site is not only technically challenging but also much more expensive. For some organizations, resilience provided by multiple datacenters is enough.

Depending on the nature of business, some organizations may need multi region availability instead – making the technical and cost challenge even bigger.

Provisioning of multi-site/multi-region private cloud implementations requires highly resilient network connectivity, which comes with additional complexity and cost. A team of skilled individuals is often required to ensure smooth operation of such an infrastructure which adds to the recurring cost. This is where service providers and public clouds play an important role by filling in those gaps in expertise.

2.2.2 Elasticity

An area where the public cloud has a substantial advantage over the private cloud is the elasticity of resources, even when hosted with a service provider. All major public cloud providers have significant amounts of spare capacity available at each site, allowing customers to consume as much or as little as needed and exactly when they need it. In contrast, it is virtually impossible for an organization to provide such freedom in resource consumption to its users. While it can forecast resource consumption and keep expanding its provision proactively, that cost is often locked in even if the resource requirement is reduced afterwards.

This is one of the reasons that public clouds have gained such momentum in recent years: They provide a flexible consumption-based model to organizations that have fluctuating resource requirements. The pace of innovation in that space also means hundreds of new capabilities are added to them each year and are instantly available to their customers to innovate with and solve their business challenges.

The public cloud resolves many of the limitations and complexities that private cloud suffers from. It is no wonder that most organizations see the public cloud to be an integral part of their cloud infrastructure strategy and have plans to maintain a hybrid cloud infrastructure for flexibility of resources.

However, there is one reason why the trend towards the public cloud is slowly reversing into the opposite direction and back towards the private cloud: Data sovereignty.

2.3 Geo-Political Realities

Due to national security concerns, data sovereignty has risen to the top of the security agenda of most governments. Those concerns were primarily fueled by the widespread publicity of the PRISM program run by the US National Security Agency, as well as similar programs run by other countries around the world. While trust in government varies dramatically around the world, businesses everywhere are concerned about the implication of these national capabilities for data governance.

Much like governments, commercial organizations (financial, healthcare, and public sector) have become increasingly concerned about data privacy issues. They are required to hold personal information of customers which can range from public to extremely sensitive, depending on the sector they belong to. A breach of that personal information could be disastrous to the individuals concerned and could cause irreparable damage to the reputation of the organizations holding that information. Even in the absence of regulatory requirements, public pressure forces these companies to protect data.

2.3.1 Regulated Privacy

In the past decade, most countries have passed data privacy laws and regulations. GDPR (General Data Protection Regulation) is a well-known example of such regulation where all member states within the European Union have moved to protect online privacy rights of its citizens. The cost of non-compliance for an organization can be extremely high: Up to 20 million Euros or 4% of global revenue, whichever one is higher.

But data privacy laws are complex and they are yet to mature. The biggest challenge is that interpretation and jurisdiction of such laws vary between countries. An example of such legislation is the U.S. Clarifying Lawful Overseas Use of Data (CLOUD) Act[1]. It states that any service provider that is subject to U.S. jurisdiction, may be compelled to disclose the data in its possession, regardless of where the data is located. This means privacy of personal or organizational data is at risk if hosted with a US provider, even if the data does not reside in the US.

However, a counter to that is the famous Schrems II[2] ruling, which invalidated the EU-US Privacy Shield Framework. That framework allowed lawful transfer of personal data from EU/EEA businesses to US-based companies that were listed in the Privacy Shield list. The US companies in that list had to go through a review process to become part of the Privacy Shield list, but the ruling essentially deemed the US level of data protection insufficient for data privacy when compared to GDPR. Therefore, it became illegal for EU/EEA companies to transfer an individual's personal data to those US companies under that agreement. Any EU/EEA company that continues to transfer such data might be considered in breach of GDPR rules and risk the penalties in line with GDPR non-compliance.

This means that an entity not only has to comply with local laws when it comes to data privacy, but also to consider whether the laws of other jurisdictions apply to the handling of that data as well. Given that 92% of western data is understood to be residing on US websites and cloud platforms, governments and other organizations in possession of sensitive data are looking to adjust and co-locate data appropriately.

---

[1] Clarifying Lawful Overseas Use of Data (CLOUD) Act (justice.gov)

[2] The CJEU Judgement in the Schrems II Case (europa.eu)

2.3.2 Sovereign Cloud

The need to maintain data sovereignty has given rise to the sovereign cloud and has moved governments and organizations that handle sensitive data, to prioritize their shift towards such cloud deployments[3]. A sovereign cloud is essentially a private cloud designed to comply with national and regional data privacy requirements, both while the data resides on the platform and when it is transferred.

Sovereign clouds can be connected to the Internet for secure sharing of data with trusted organizations, as is done under the GAIA-X initiative[4] in the European Union, or it can be completely isolated, such as in case of sensitive government clouds. It is a private cloud that is highly secure and customized in both cases, tailored exactly according to the needs of the organization or a government entity.

Keeping data secure is important to organizations of all kinds but is paramount for governments. In recent years, geopolitical realities have meant governments are taking additional steps to protect sensitive data from falling into the wrong hands, especially during regional conflicts when the number of attacks on public infrastructure rise exponentially.

The ability to secure and control access to personal data regardless of complex and ever-changing legislation provides governments and other organizations the confidence that it is secure. The need for this level of trust is fueling the return of the private cloud.

2.4. Server Architecture

Regardless of the needs of the organization or application, there is a computer server architecture to match. Component portfolio architectures must consistently deliver new features, performance benefits, reduced power consumption per operation, all while adhering to ever increasingly stringent security standards to remain competitive. Components need to provide these features to not only allow developers to protect customer workloads at the software layer but provide inherent protection at the hardware layer.

Over the decades, workload processing requirements have steadily risen, saturating the additional computational gains presented by Moore's Law. Although the electrical power and cooling required to support a given system as fallen, the overall scale of IT infrastructure has risen even quicker. Fortunately, this rapid growth and evolution of requirements has been something companies have been able to leverage to help meet sustainability goals as a core component of their corporate, environmental, and social responsibilities.

[3] Deploying the Right Data to the Right Cloud in Regulated Industries (vmware.com)

[4] Project GAIA-X – A Federated Data Infrastructure as the Cradle of a Vibrant European Ecosystem (data-infrastructure.eu)

2.4.1 Internal and External Accelerators

As memory, network, and storage vendors strived to deliver ever more performance, the Intel x86 CPU had to deliver more performance to match. The Intel 8086 processor gave rise to x86 architecture in 1976, but workload demands had already increased enough to warrant the release of the first x86 accelerator, the Intel 8087 co-processor, just four years later. This was an add-on chip focused solely on improving floating point calculation performance, and was the first workload accelerator in the open systems era. Computational workload demands have continued to evolve at a steady rate, forcing vendors to create more and more task-specific accelerators. These requirements grew into a portfolio of platform components, driving the industry to adopt new CPU, RAM, storage, and peripheral interfaces. Direct Memory Access (DMA) allowed these components more effectively to offload not just computation from the CPU but also input and output tasks.

Graphic processing benefited greatly with the advent of Graphics Processing Units (GPUs), which offload the special computations associated with computer graphics. From the beginning, GPUs have relied on a huge number of specialized cores working in parallel. In recent decades, GPU applications have evolved to allow for computational workloads to run on these massively-parallel GPUs. This in turn has enabled new applications like cryptography and artificial intelligence (AI), which have forced GPU architectures to evolve beyond graphics processing.

Networking was the next area optimized by special-purpose add-in cards. Once the performance of Ethernet surpassed gigabit speeds, companies began working to accelerate UDP and TCP processing. After 1999, TCP Offload Engine (TOE) cards offered greater performance and reduced CPU impact, especially for protocols like iSCSI. In recent years, network adapters have added Field-Programmable Gate Array (FPGA) and Application-Specific Integrated Circuit (ASIC) processors to these cards. Intel's Infrastructure Processing Units (IPU) enable programmable network devices to intelligently manage system level resources by securely accelerating networking and storage infrastructure functions in a datacenter.

Storage has undergone a radical evolution over this same period of time, with RAID, solid-state, tiering, and caching providing faster access, greater functionality, and ever-increasing bandwidth. Security considerations have been met through the advent of self encrypting drives adding intrinsic protection for companies without impacting the performance of the drive itself. Intel's Optane technology combines attributes traditionally assigned to memory and storage, offering greater capacity at lower cost than DRAM and greater performance than NAND flash. And modern storage devices include dedicated processors to offload specialized storage operations.

After MMX, SSE, SIMD, and AVX instruction sets that greatly benefitted multimedia and vector based workloads, enterprises turned to encryption offload. Intel gave its processors an accelerator purely aimed at increasing throughput for workloads where AES or SHA was essential: Advanced Encryption Standard New Instructions (AES-NI) was introduced in 2008. Virtualization focused accelerators such as VT-x paved the way to allow even greater densities and performance within the virtualization space. Internal and external accelerators have underwritten true industry level innovation.

No two workloads are the same. Enterprises must decide on a performance baseline for their workload requirements. Workloads built factoring internal CPU accelerators can complement external workload accelerators. Intel's portfolio enables enterprises to get the performance they need where they need it. CPU benchmarks only tell part of the story when looking at a graphics intensive workload. High throughput low latency networks leveraging Remote Direct Memory Access (RDMA) is one of a plethora of workloads that will not benefit from additional CPU performance. The impact of workload accelerators cannot be overstated.

Collectively, these external accelerators are becoming known as "X Processing Units" (XPUs) meaning something is being accelerated either on die or externally. The future looks bright given observations within the industry driving innovation bringing components closer together, often sharing space and or resources on die. We expect to see an aggregation of control and monitoring solutions highlighting operational gains realized from these modular components,

3rd Gen Intel® Xeon® Scalable Platform

Feature	2nd Gen Intel® Xeon® Scalable Processor (Cascade Lake)	3rd Gen Intel® Xeon® Scalable Processors (Ice Lake)	Notes
Cores per Socket	4-28	8-40	New Sunny Cove architecture
L1/L2/L3 cache per core	32KB/1MB/1.375MB	48KB/1.25MB/1.5MB	Larger caches to enable fast access to data
Memory Channels and DIMM Speed	6 Up to 2933	8 Up to 3200	Huge boost in memory bandwidth & support for Intel® Optane™ PMem 200
Processor Interconnect: UPI links, speed	2 or 3, 10.4 GT/s	2 or 3, 11.2 GT/s	Improved bandwidth between processors
PCIe lanes per socket	PCIe 3.0, 48 Lanes (x16, x8, x4)	PCIe 4.0, 64 lanes (x16, x8,x4)	2x bandwidth and more PCIe lanes to support new Gen 4 SSD, Ethernet and other adjacencies
Workload Acceleration Instructions	AVX-512 VNNI DDIO	AVX-512, VNNI, DDIO vAES, vPCLMULQDQ, VPMADD52, VBMI, PFR, Crypto, SHA extensions, TME, SGX	Enable new capabilities and speedup performance
Platform Adjacencies		Intel® Optane™ PMem 200 series, Intel® Optane™ P5800X SSD, Intel DC P5510 SSD, Intel E810-C ethernet adapter	

leading to a new disaggregated architecture, as described in section 2.4.4 below.

2.4.2 Power, Management and Security

As of today, changes in operational efficiency through increased core counts and new instruction sets coupled with advancements in lithography are allowing for CPUs with greater power efficiency. 3rd Generation Intel Xeon Scalable processors have contributed numerous advancements to these operational efficiency gains. Cryptographic instruction additions such as AES and SHA acceleration performed via hardware again, within the CPU have greatly enhanced the performance of numerous workloads requiring data encryption, an ever-increasing percentage of workloads in modern times. This is further reinforced when we consider how many more layers of security enterprises are running to protect systems and data.

Additional gains in platform power efficiency can be attributed to storage, shifting the landscape from motors spinning generating larger volumes of heat to more power efficient NVMe devices. Further gains can be attributed to the wider use of SFP and fiber optic network and storage connectors. Leveraging newer processors with newer and more power efficient connections not only increases performance, but also helps organizations contribute to company sustainability objectives by providing greater power efficiency and reduced cooling requirements.

Intel's 2021 Product Security Report[5] provides some insight to known CVE security vulnerabilities indicating that Intel themselves internally detected 50% of the known CVE entries for Intel Products and that number increases to 93% when we factor in their bug bounty program. Not all CPU vendors operate a bug bounty program and Intel has products in many categories within the server architecture landscape that other vendors do not compete in. However, when we look purely at CPU related CVE counts, AMD had 31 from May 2021 to the end of the year and Intel had 16 in 2021 for the year.

2.4.3 The Computational Conundrum

Data Center Infrastructure Management (DCIM) adoption has continued to grow which allows more efficient planning and management of many of the physical aspects of providing datacenter services. The ability to integrate hardware systems into a given DCIM solution is paramount to model and manage datacenter needs effectively. The platform chosen from the portfolio available will have lasting implications. Enterprises must factor in all components and ensure their interoperability before committing to a deployment, something vendors often assure across their portfolio of products. Securing components that are known to work with each

[5] Intel 2021 Product Security Report (intel.com)

other not only helps enterprises maintain SLA commitments, but is less painful to own and manage effectively long term.

2.4.4 Disaggregated Architecture

As discussed above, software-defined cloud infrastructure is inherently designed to scale out beyond the bounds of a server, storage array, or network switch. The evolution of multi-core, multi-processor, and accelerated computing points the way to a future in which separate compute resources act together when bound by software. This history also teaches us much about managing pooled non-uniform resources, scheduling, and dynamic provisioning.

One of the primary challenges to scaling infrastructure has been the boundary of the server itself. I/O that crosses outside a server has traditionally been encapsulated in high-level network protocols to protect it in transit. This has created a logical divide between internal and external I/O that limits what can be placed outside the chassis. But PCI Express (PCIe), the leading I/O technology, is beginning to bridge this divide. Intel's pioneering Thunderbolt technology showed that native, full-speed PCIe links could extend all the way to external peripherals, and NVMe storage has shown that PCIe could connect devices in a dynamic way.

The next step along this path is the migration of directly-accessed memory, storage, and processing outside a traditional server architecture using block and cache-coherent interconnects known as Compute Express Link (CXL). This technology, developed initially by Intel, has been embraced as a standard by nearly every company in the IT industry and is already being implemented. CXL 1.1 products allow memory expansion beyond traditional memory channels and are becoming available in 2023.

The next wave of CXL products will allow nearly any resource to be moved outside the server chassis, including dynamic memory (RAM), persistent memory, storage, general-purpose accelerators, and I/O devices. The development of CXL-capable peripheral and switch chips is proceeding rapidly, with many companies expected to bring products to market in 2023. And the major CPU and server vendors, including Intel, are widely expected to deliver CXL 2 support in their next-generation products. This enables pooling of resources, which is enabled by software to manage access and sharing.

CXL will ultimately enable truly disaggregated systems, with CPU, memory, storage, and I/O resources at multi-chassis, rack, or datacenter scale. Disaggregated servers will enable software-defined infrastructure concepts to reach their full potential, eliminating the traditional scaling constraints of CPU socket, memory channel, and so on. When this disaggregated future arrives, the ultimate differentiator for server components will be broad compatibility not core count or absolute performance.

At the same time, a new technology is being designed to allow more flexible combinations of processor components. Known as Universal

Chiplet Interconnect Express (UCIe), it would enable so-called chiplets, manufactured at different times, on different process nodes, and perhaps from multiple vendors, to be combined on a single processor package. This modularity will spur the creation of special-purpose processors for various use cases. Intel is driving the development of the UCIe specification and enabling the creation of chiplet-based processors, the first of which are already being delivered in 2022.

Another key technology places optical interconnects directly on silicon chips. Silicon photonics products are now appearing, with Intel leading the development in this area as well. It is expected that this technology will be delivered on chiplets, and many expect to see silicon photonics-based PCIe and CXL connections in the future.

All of these technologies show that a truly disaggregated infrastructure future is coming to the datacenter and cloud.

2.5 Datacenter Trends: Conclusion

Things have changed massively since the early days of datacenter computing. Many infrastructure trends have come and gone, with the focus oscillating between private and shared compute environments. When it comes to deciding on the datacenter of the future, one must recognize that balance and flexibility are paramount.

Through all of the changes over the last five decades, Intel has remained a leader in enterprise computing. Its vast portfolio of products are used in every IT environment, including the traditional datacenter, the cloud, and the new world of embedded and edge computing. Given Intel’s continued investment, this trend is likely to continue.

3. Real-World Use Cases

Now that we have considered the primary trends in server infrastructure decisions, we turn to a consideration of the workloads that are being hosted here. As discussed above, public cloud has become increasingly popular with enterprises, but some are also building their own infrastructure to support such workloads on-premises. Another critical workload, especially driven by the global pandemic, is virtual desktops and similar centralized environments for remote access. We must also consider the demands of nascent AI and ML workloads, as these are increasingly common. These workloads are representative of the demands on modern infrastructure, and there are numerous technologies that can be employed to meet them.

3.1 Repatriating Cloud Workloads

Many enterprises are evaluating or currently in the process of repatriating workloads. Long standing "cloud first" approaches have created issues with costs, control, data gravity, and provider failures. Coupled with the advent of global requirements mandating where data can be stored, many enterprises are taking workloads out of public clouds to address compliance or performance requirements. This allows organizations to meet data residency requirements, especially true for workloads with access to sensitive data. This can often catch organizations by surprise having to react to new legislation as it is written and can also prove costly. Prudent enterprises ensure exit agreements with cloud providers should the need arise.

Since the beginning of 2020 companies were also forced to react to the global COVID pandemic pivoting, how they operated, simply to continue to function strong arming cloud adoption further. Architecting a solution with a lead time on hardware deliveries was simply not a viable option. Enterprises have been steadily re-evaluating workload placement ever since.

Edge computing also contributes to the repatriation of workloads as organizations look to make better use of data available outside of public cloud providers. Data locality at the edge can mean it makes sense to process data closer to its source depending on the size of datasets. Workload mobility has also been a huge driver with enterprises aiming to provide highly available, disaster tolerant services across multiple public and private clouds. This is the proverbial holy grail of cloud computing in terms of solution robustness and availability.

While many of the largest global organizations have the in-house staff and technical ability to manually perform this function, it is extremely difficult to achieve. Smaller companies typically engage Managed Service Providers (MSP) to provide Disaster Recovery (DR) and High Availability (HA) services to meet their Business

Continuity (BC) objectives on a more modest scale. But all businesses, large or small, want to avoid putting all of their proverbial eggs in one basket.

Cloud repatriation does not spell the end for hyperscalers; it simply means that customers want options and appropriate solutions for their workloads. Regardless of service level agreements, any infrastructure provider can experience a costly outage, so enterprises are simply working to mitigate that risk. Multi-day outages have been observed among the hyperscalers geographically and within essential areas of functionality, such as routing and DNS, so the prudent approach is to remove sole reliance on any single provider. AFCOM recently reported that 58% of organizations had repatriated workloads in 2021.[6] Although these respondents claimed security, performance, and cost as their primary drivers, many of these could be attributed to the challenge posed by modern DevOps and hybrid cloud approaches.

3.1.1 Hybrid Cloud

Hybrid cloud is a general term that covers a diverse range of solutions, but the term usually suggests an approach that combines both on-premises and cloud. Many have simply embraced the option to run cloud-native applications on-premises or in a cloud provider, while others have built an overlay that allows seamless transition between them. One popular approach is to run test and development in the cloud and later deploy on-premises to minimize long-term costs. But hybrid cloud, while seemingly simple and economical, presents numerous challenges to effectively secure, maintain, and scale.

Vendors such as VMware have developed a hybrid cloud approach for more conventional applications with offerings such as VMware Cloud Foundation (VCF). Real-world experience with this product shows the benefits of a Software Defined Data Center (SDDC) approach. SDDC provides a single control plane, allowing enterprises to manage workloads in multiple locations using a common set of tools and APIs. The availability of Intel Xeon Scalable processors both on-premises and in public cloud SDDC offerings from AWS, Azure, Google, and Oracle means that a consumer can seamlessly and non-disruptively migrate workloads between any of these locations. While deploying this type of SDDC configuration requires effort, it will pay dividends in the long run by greatly simplifying other tasks, such as backups, HA and DR, as well as providing insight and performance metrics. As enterprises and providers strive to add additional nines to their agreed SLAs, when minutes and seconds matter, this level of functionality is invaluable.

3.1.2 Hybrid Cloud Kubernetes

Containerization of microservices-based workloads has been around for a while. Since the “chroot” command in Unix arrived in 1979, administrators have had the ability to isolate and

[6] The 2021 State of the Data Center Report (datacenterworld.com)

segregate file access for processes. Throughout the 2000s, advances from FreeBSD, Linux, Solaris, and others added greater ability to isolate workloads. This led to Google developing process containers in 2006, which was subsequently renamed to Control Groups (cgroups) and merged into the Linux kernel in 2.6.24. In 2013, when Docker emerged, container-based approaches became popular, leading to Kubernetes in 2014. Kubernetes addresses many of the hurdles that came with managing and scaling applications and is becoming popular in enterprise environments. Most modern enterprises applications are designed for a cloud-native Kubernetes-based deployment that can be run on any modern infrastructure.

Other vendors have invested to allow their products and services a deeper level of integration with VMware Cloud Foundation to take advantage of the scale, insight, and APIs available to further bolster their offerings. This presents a further value add for SDDC consumers and deeper integrations with existing solutions within an enterprises technology portfolio. This is the other primary hybrid cloud approach seen today.

The benefit of stateless SDDC workloads is obvious. Kubernetes orchestrates millions of web servers and microservices providers, and once-challenging stateful workloads like database services are now available in cloud-native platforms. The only challenge posed to containerized applications is the need for compatible CPU architecture and accelerators, and Intel's server architecture is widely available in both on-premises datacenters and public cloud offerings.

3.1.3 Hybrid Cloud vSphere

VMware's SDDC approach also allows organizations to silo off workloads that they see fit using a mechanism to define workload domains. Enterprises are free to create workload domains spanning multiple clusters and providers if deemed necessary, have carte blanche on the definition of what work they perform. Obvious choices have been management, general servers, VDI, and database-specific. Modern workload domain choices focus on current business challenges surrounding artificial intelligence, machine learning, and other enterprise-specific use cases that contribute to successful business outcomes. This granular approach to workload architecture allows enterprises to focus external accelerator placement where it's needed.

Multitenancy is achievable by leveraging VMware Cloud Director (VCD), which solves many of the challenges enterprises face. It effectively isolates internal and external customer workloads beyond the VLAN layer deeper into the TCP/IP stack using Generic Network Virtualization Encapsulation protocol (Geneve) network segments. Intel 800 series network cards can accelerate this process and further allow enterprises to define policies on the traffic using

Application Device Queues (ADQ) providing predictable latency for workloads.[7]

Intel accelerators enhance the performance and security across most aspects of a modern hybrid cloud software defined datacenter. This contributes to a lower TCO when compared to raw x86 or x64 performance.[8]

3.2 Virtual Desktop Infrastructure

Desktop and laptop computers make up the largest proportion of the computer power in most enterprise organizations. VDI places all that compute load inside a datacenter and is one of the most challenging workloads. Intel's Xeon Scalable processors provide a large number of fast cores in modest servers, delivering a cost-effective computing platform for VDI. Coupled with Intel's accelerators and Optane technology, Xeon CPUs can help to address some of the density and scaling issues that affect VDI deployments. Today, Optane technology helps address both storage performance and RAM scalability challenges that impact the cost of a large VDI deployment, and we suspect that future servers will offer additional memory channels and CXL-based memory expansion to take this even further.

3.2.1 VMware Licensing is a Constraint

When we look at the cost of a VDI deployment, there are per-user costs and there are per-physical server costs. For cost optimization, we look at optimizing the number of users that can be accommodated per physical server. One element of the per-server cost is the hypervisor, usually VMware ESXi, where each socket license allows a maximum of 32 cores. CPU sockets with 32 cores, or a multiple of 32, will give the most CPU power for the fixed socket license. From this fixed point, we build a server hardware configuration to utilize these cores efficiently. As a rule of thumb, 8 vCPUs per core and 4 GB of VM RAM per vCPU leads us to 1 TB of RAM per 32-core socket9, possibly a little less since ESXi has memory efficiencies. This leads us to a dual-socket server with 32 cores per socket and a total of 1-2 TB available memory as likely to provide the most cost-effective VDI server. The basic rules of thumb work well for CPU and RAM, but the hard part is always the storage design.

3.2.2 Cost Per Desktop

The CPU and RAM rules of thumb are proven in an Evaluator Group study commissioned by

---

[7] Building a Multicloud Analytics Solution with VMware Cloud Foundation (vmware.com)

[8] Scaling Data Center Modernization for Hybrid Cloud (intel.com)

[9] 32 cores x 8 vCPUs x 4 GB = 1 TB

Intel[10]. Using LoginVSI, the de-facto industry standard VDI benchmarking tool, the study showed very high user density using 3rd Generation Intel Xeon Scalable processors, leading to lower per-desktop cost.

3.2.3 Storage Performance

VDI is a notoriously fickle storage workload:

- When desktop VMs start up or are re-imaged there is a large and sustained sequential read workload
- When users log on to their desktops there are a large number of random writes
- During the working day, there is a moderate write load, unless desktop VM RAM is short, then there will be a mix of reads and writes

All these different demands can be applied to one piece of storage over a single day. NVMe SSDs with optimized NAND flash or Intel's Optane technology provide high-performance storage capacity with low latency and reasonable write endurance. The NVMe interface is ideal for VDI as it features multiple IO queues, which are required to support multiple different VMs reading and writing at the same time. The use of local SSDs with Optane technology in each VDI host can deliver cost-effective and high-performance storage for stateless desktops. Using Optane technology as the performance tier of a VSAN cluster will also deliver improved performance for persistent desktops, provided it is coupled with high-performance networking, such as Intel Ethernet 800.

3.2.4 More Desktops with More RAM

Another way to use Optane technology is as Persistent Memory (PMem) on the RAM bus. For VDI, PMem is most often used to expand the VDI host RAM using memory mode. Like many virtualization workloads, VDI is often limited by the amount of RAM that is installed on the server. In memory mode, the Optane DIMMs provide the memory capacity for a virtualization host and a smaller amount of DRAM provides a cache. The amount of DRAM cache is typically a quarter of the amount of PMem. Optane technology delivers a lower cost per GB than DRAM, even accounting for the requirement to use DRAM as cache.

The performance studies that Intel published using both VMware Horizon and Citrix Virtual Desktops showed almost double the number of desktops on a 2TB PMem system than on a 1TB DRAM-equipped server. The 2TB PMem system delivered approximately a 20% cost saving over three years compared to the 1TB DRAM configuration[11].

Although Intel is not continuing the development of Optane technology beyond 2022, it is a cost-

---

10 "Intel Select Solutions for Virtual Desktop Infrastructure with VMware Horizon"

11 "Intel Select Solutions for VDI"

effective way to increase system memory for VDI workloads with today's Xeon Scalable processors. Intel is expected to add wider and deeper memory channels for next-generation Xeon Scalable processors, allowing them to support more DRAM. In the future, CXL is emerging to dramatically increase the availability and flexibility of RAM deployment as well.

3.3 AI and ML

Artificial intelligence (AI) technology is being deployed more frequently in enterprise and cloud applications, and this deserves special attention. Modern AI typically leverages machine learning techniques, which build an algorithm based on sample input data ("training") which can then be deployed in production ("inferencing"). Most such models employ deep learning techniques to build an artificial neural network that can make decisions based on rich and varied input data.

Most deep learning systems use specialized processor instructions or massively parallel processors to accelerate the floating point math used for both training and inferencing. Training systems require fast access to large amounts of data storage as well as flexible and fast connectivity between processing and storage nodes. This makes AI training one of the most challenging applications for modern data processing systems. But even AI inferencing often requires specialized processor accelerators, including special instruction sets and dedicated processor hardware.

3.3.1 ML Training

Machine learning training typically requires extremes of computational, networking, and storage infrastructure. Most training infrastructures are engineered "stacks" which deliver these resources using multiple servers with CPU, GPU, and memory, integrated network with offload processors, and high-performance scale-out storage. Every producer of these components is working to deliver an AI training solution, but the real-world performance of these varies widely. The most successful offer parallel and distributed connectivity between host and storage, allowing the processors to access a massive volume of training data efficiently.

If a customer requires higher ML processing performance, Intel's Habana Labs offers a specialized ML training processor called Gaudi which offers high-performance matrix multiplication with a heterogeneous processor and integrated memory. It leverages multiple 100 Gb RDMA over Converged Ethernet links to connect to a host[12]. Habana's SynapseAI software suite integrates with TensorFlow and PyTorch, allowing developers to train deep learning models on Gaudi with minimal software changes.

Intel also recently introduced a high-performance distributed parallel file storage solution known as Distributed Asynchronous Object Storage (DAOS). This solution leverages Intel's Optane technology for persistent memory and NAND

[12] "Affordable Deep Learning Training with Intel Habana Gaudi," Sree Ganesan, AI Field Day 3 (techfieldday.com)

flash SSDs to overcome the bottlenecks typically encountered in big data environments[13]. DAOS is one of the leading high-performance scalable storage solutions and is found in many high-performance computing (HPC) environments. Although it is useful for many applications, DAOS is particularly valuable for AI workloads.

3.3.2 ML Inferencing

Although machine learning inference does not require as much computational or storage resources as training, it is still computationally intensive. For this reason, most ML inferencing solutions rely on specialty processors such as GPUs to parallelize processing and accelerate performance. But this can be expensive, both in terms of operational (power and cooling) and capital (hardware) costs. Therefore, most organizations that have deployed machine learning in production are looking for a simpler platform.

Intel has recently been positioning the Xeon Scalable processor line as an alternative to expensive GPUs for ML inferencing. Although a CPU is quite different architecturally from a GPU, Intel has long equipped their CPUs with special accelerated hardware and instruction sets for such computationally-intensive tasks as matrix math.

Many organizations have discovered that the Intel Xeon Scalable processors with deep learning boost (DL Boost) capability can perform ML inferencing satisfactorily with dramatically-reduced investment compared to GPUs. For example, Twitter was able to achieve satisfactory performance in natural language processing using a single thread on a Xeon CPU at much lower cost than a specialized GPU instance.[14] Twitter leveraged the Vector Neural Network Instructions (VNNI) in 2nd Generation Intel Xeon Scalable processors with DL Boost in Google Cloud, but the Bfloat16 capabilities in the 3rd Generation Intel Xeon Scalable processors would likely be even more efficient.

Leveraging internal CPU accelerators like DL Boost allow companies to get greater return on investment from existing hardware or to simplify their infrastructure. Less expensive instances can be used in a cloud platform, and these can be more easily re-used or shared with non-ML applications in a private cloud. A simplified infrastructure can reduce environmental impact as well.

3.4 Implementation Details

Our research into modern infrastructure revealed a few key implementation details of note. One of the key enablers of high performance and

[13] "Intel DAOS 2.0: Storage Software Stack and Ecosystem Update," Kelsey Prantis, Storage Field Day 22 (techfieldday.com)

[14] Speeding up Transformer CPU inference in Google Cloud (twitter.com)

efficiency is direct access to data, but RAM is extremely expensive. This is why we were impressed by Intel's Optane technology, since it increased the capacity of system memory at a dramatically lower cost. We experimented with Intel Optane technology, and will report our findings here. We also implemented a hyper-converged infrastructure (HCI) solution with Intel Xeon Scalable processors and Optane technology components and have included our findings related to this experience.

Although Intel announced in 2022 that it would cease further development of Optane technology, the current Optane and Xeon products were designed to work together and provide many benefits for current workloads. In fact, our research shows that Optane PMem provides unmatched levels of performance by expanding the available system memory at far lower cost than traditional DIMMs. Intel is also committed to supporting current Optane products and they remain widely available in the channel. For this reason, we remain confident in our recommendation of this technology.

3.4.1 Persistent Memory

One of the operating modes for Optane PMem DIMMs is as high-capacity memory for a system, often a hypervisor. In this mode, the Optane DIMMs provide the installed system memory for the server. Optane DIMMs have a lower cost per GB than DRAM, making a 2TB PMem system less expensive than one with 2TB of DRAM. Unfortunately, Optane DIMMs have higher latency than DRAM, so the system will also need some DRAM to act as cache to deliver lower latency for frequently accessed memory. In this mode, a PMem DIMM is paired with a much smaller DRAM DIMM to provide cost-effective capacity and performance.

In our testing, the DIMM pairing was very evident on an HPE ProLiant system. A single memory channel could only have one PMEM DIMM and a single DRAM DIMM. The HPE BIOS also required that the PMEM DIMM be either four or eight times as large as the DRAM DIMM. This 4:1 or 8:1 ratio is in line with Intel's best practice, but we were surprised that the HPE BIOS was very strict about enforcing the configuration and ratio.

When designing a server system to use Optane PMEM as system memory, one must be mindful of the number of memory channels and the number of DIMM slots per channel. More DIMM slots without more memory channels may not provide the benefits expected. When deploying the system, it will report the total PMEM size as its installed memory. The DRAM DIMMs will not contribute to memory capacity as they provide a cache for the data in PMEM, not an additional tier of memory.

3.4.2 HCI Details

The initial value proposition with hyperconverged infrastructure (HCI) was the elimination of the storage array and its dedicated storage network from the virtualization platform. The storage network was replaced with the same ethernet network that VMs communicate, and the array was replaced with a storage cluster provided by the hypervisor hosts. The real value of HCI showed in the ability to hide as much complexity as possible, allowing customers to focus on managing the virtual machines that run their business applications. HCI simplifies the entire lifecycle of a virtualization platform, from purchasing through operation and updates until retirement.

One of the challenges for HCI vendors is guaranteeing the performance of the storage cluster which shares resources with the compute cluster and doing so without excessive resources allocated to storage. The Intel Xeon CPUs include the capability to bind storage processes to specific CPU cores, and the VM processes to different cores, thus preventing resource contention without having to allocate entire CPU sockets to storage processes.[15]

Matching an HCI platform to the workload is important, every vendor offers different server configurations with different resources. HCI is often used as the default virtualization platform for workloads that do not have large resource demands. HCI is particularly well suited to Virtual Desktop Infrastructure (VDI) as both scale-out, HCI with servers that deliver consistent resources and desktops that require large numbers of relatively small VMs. Some HCI workloads are more storage sensitive, often leading to the use of SSDs with Optane technology as a cache in the HCI storage.

HCI platforms often differentiate themselves on how simple it is to deploy updates to an HCI cluster. The simplicity is possible because the vendor only offers a specific set of hardware choices and is able to validate the updates and update process before handing the updates to customers. This is where Intel's HCI vendor partners play a crucial role.

3.5 Use Cases: Conclusion

Cloud repatriation, virtual desktop infrastructure, and machine learning are three key workloads driving server architecture selection today. These workloads require a homogenous infrastructure made up of server, storage, and network components that can be combined into a unified platform. But they also require extreme levels of performance, memory, and capacity, driving businesses to adopt technologies like internal and external accelerators and persistent memory. Achieving an appropriate level of performance and consistency is one reason Intel's server portfolio is so attractive: It offers not just CPU performance but an ecosystem of accelerators from both inside and outside the company, and is widely supported by major vendors.

---

[15] "Boosting Hadoop Performance and Cost Efficiency with Caching, Fast SSDs, and More Compute"

4. The Perfect Configuration

Server virtualization changed the computing landscape forever, and it can be credited with influencing most infrastructure innovations in recent decades. For example, network virtualization, and application containerization both followed widespread adoption of server virtualization in enterprise IT. Virtualization also started the focus on balanced infrastructure rather than optimum performance, since flexibility trumps a highly-tuned design when workloads change.

Recognizing the wide variety of implementations, manufacturers are producing a wide variety of hardware to meet nearly any need. This wide range of choices presents a conundrum: Should one go for the application-centric optimum configuration or take a hardware-centric approach and aim for a balanced configuration that is suitable for all workloads?

This section starts with a discussion of the challenges virtualization brought in addition to all of its goodness. It moves into how to opt for a particular configuration before attempting to build a perfect configuration based on publicly available architecture and performance whitepapers produced by Intel and VMware after extensive testing. Finally, it discusses why is it important to choose vendors with a diverse ecosystem of products when building robust systems fit for the datacenters of the future.

4.1 The Virtualization Challenge

Virtualization brought the ability for organizations to utilize their hardware resources efficiently by consolidation of workloads. It is considered normal now to fit as many workloads as possible on a smaller number of hosts, apart from some specialized use cases or where raw performance of an application is of paramount importance.

When consolidating virtual workloads, memory is typically the limiting factor. Lack of it typically has a significant negative effect on an application, especially in cases where the working set is large. For that reason, an allocation ratio of 1:1 physical to virtual RAM is common in almost all cases.

This has historically been an expensive problem to solve as DRAM costs rise sharply with size, which ultimately limits the number of virtual workloads that can be safely consolidated on a host. Additionally, growth of DRAM capacities has not kept pace with enhancements in processor capabilities, effectively capping the amount of memory that a host with a certain number of processors can support.

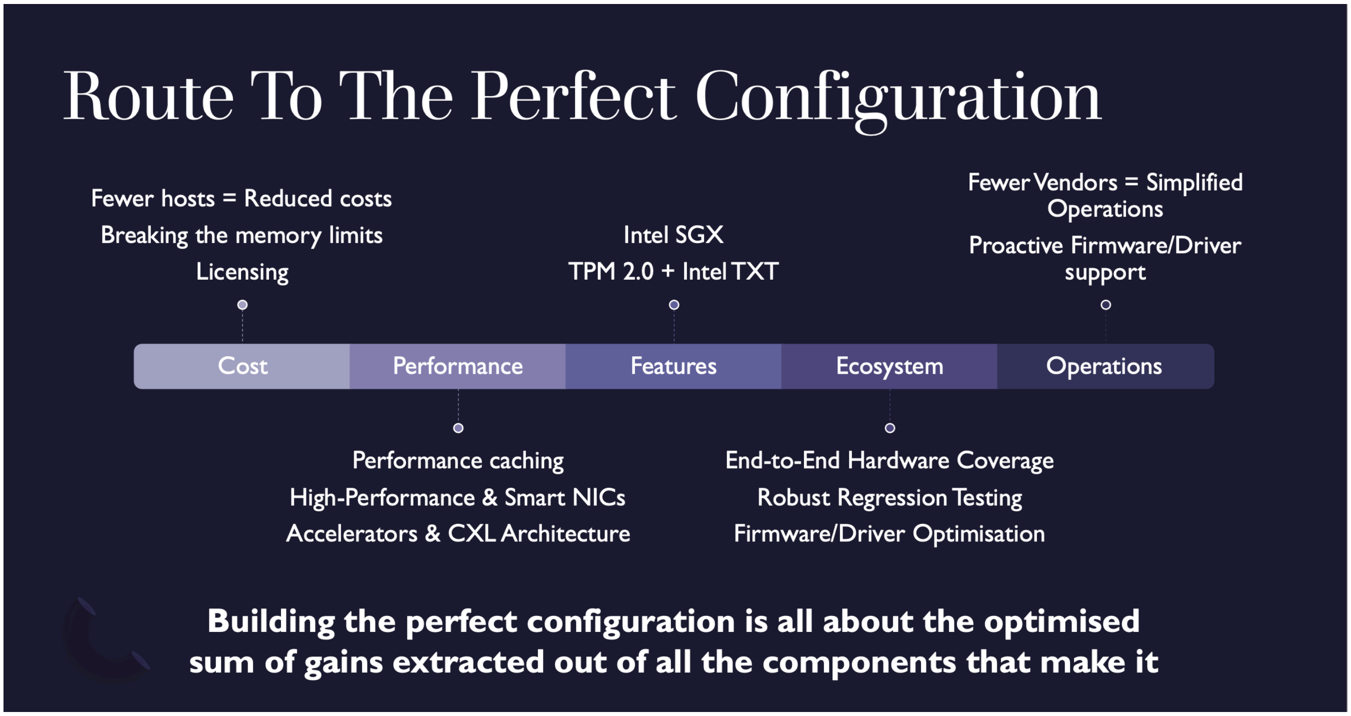
A tiered memory solution based on Intel's Optane Persistent Memory is an elegant and cost-effective solution to that challenge. When paired with the latest versions of VMware vSphere, the supported configurations break the boundaries that DRAM only configurations are constrained by. This allows consolidation of significantly more virtual workloads on the same number of hosts. Best of all, transparency of "memory mode" ensures that no changes are required to the hypervisor, operating systems, or the applications.

Having said that, building the perfect configuration for a datacenter requires more than enhancements in memory capacity alone. Such a configuration needs efficiencies at all levels, to provide gains in all aspects of the configuration. They provide the best possible combination of cost and performance when combined.

4.2 The Optimal vs. The Perfect Configuration

Careful observation and study of the performance characteristics of a specific application over time is required to determine the most optimal configuration for that application. This approach is best suited for specific business-critical applications, and the investment of time and effort in developing that configuration is well-justified.

However, most organizations or service providers, typically running multi-tenant environments, want to strike the right balance between cost and performance. That balance is extremely difficult to achieve given that workloads running on that platform are heterogenous in nature. For such cases, taking a hardware-centric approach provides the "perfect" configuration in the shortest amount of time.



More importantly, such a configuration provides reasonable confidence that it will be able to deliver the required performance to all virtual workloads in the most cost-effective manner.

4.3 Building The "Perfect" Configuration

While looking for that "perfect" configuration, making hardware selections based on best practice recommendations provides the best chance for a balanced configuration. Intel and VMware have done extensive testing with various configurations and made their results public, which makes the task of building such a configuration possible.

In this section, the factors that have the most impact are listed, with appropriate rationale. While those factors assume usage of VMware vSphere as the hypervisor layer, most of these principles translate well to other hypervisors too. For fairness, only the publicly available results and statistics have been considered.

Disclaimer: The following are important factors to consider when making hardware buying decisions, however, they should not be seen as purchase advice.

4.3.1 Processor Family

When creating a datacenter architecture for the future, it makes sense to start with the latest generation of processors. With that in mind, the 3rd Generation Intel Xeon Scalable processor family is the obvious choice, as it supports 8 memory channels per socket, operating at 3200 MHz, thus providing up to 32 percent more bandwidth compared to the previous generation[16].

This family of processors is designed to support the latest Intel Optane Persistent Memory 200 series, which we also recommend for current configurations, as well as all of the accelerators mentioned. This makes the 3rd Generation Intel Xeon Scalable processor family the basis for all of the following recommendations.

4.3.2 Processor

While considering a good balance between cost and performance, it appears that a dual-socket 32-core processor configuration provides the optimum value for a heterogeneous workload scenario. Having 32 cores per processor is important because starting from vSphere 7.0, one CPU license covers the first 32 cores of the processor[17]. Maximizing the number of cores per vSphere license optimizes consolidation of workloads per host. 32-core processors are also

[16] Technology Update on 3rd Gen Intel Xeon Scalable Processors (techfieldday.com)

[17] Licensing for ESXi Hosts (vmware.com)

less expensive on a per-core basis than higher core count alternatives.

Therefore, 32 core 63xx and 83xx series processors should be considered. For example, the Intel Xeon Gold 6338 Processor[18] is the most cost-effective choice for most workloads. At a higher cost, a range of 32 core 83xx series processors are available for workloads requiring a higher level of performance.

4.3.3 TPM and Intel TXT

Software attacks are becoming increasingly complex. Software solutions alone are not enough to protect an organization from a security breach. For that reason, hardware-based solutions are required to provide a secure "Root of Trust," which can securely store encryption keys, certificates, and other security artifacts that can be used to authenticate the platform in use.

A Trusted Platform Module (TPM) chip provides that functionality and is now commonly available on most hardware platforms. TPM 2.0 is the current standard and supported by the latest Windows, Linux operating systems and hypervisors. vSphere has supported TPM 2.0 since version 6.7[19] and it is highly recommended to buy hardware that contains a TPM 2.0 chip and have it enabled in the UEFI firmware.

Such security can be further enhanced by procuring hosts with processors with Intel Trusted Execution Technology (Intel TXT) support, such as the 63xx/83xx series processors mentioned in the previous section[20], and enabling it at all levels (i.e., BIOS, OS and hypervisor). Intel TXT works in conjunction with TPM to provide hardware-based trust attestation to the BIOS, OS and the hypervisor. It also works with cloud management solutions to provide policy and compliance enforcement, automation, auditing, workload placement, and geolocation-based migration control etc. Like TPM, Intel TXT is also supported by the latest Windows, Linux operating systems, and vSphere.

These technologies work together to allow an organization to take advantage of "Secure Boot" when booting vSphere. This ensures that the deployed boot image only contains digitally signed code and verifies that it has not been tampered with. That improves the security posture of the hypervisor deployment and is auditable via the security report, available from vCenter.

Once booted, the workloads can optionally be further protected by security policy-based attestation and placement control. The periodic integrity checks can be run on the hypervisors and any compromised hosts can be taken out of service proactively and rebuilt with clean images.

---

[18] Intel® Product Specification Comparison (intel.com)

[19] Securing ESXi Hosts with Trusted Platform Module (vmware.com)

[20] https://ark.intel.com/content/www/us/en/ark/compare.html?productIds=212307,217216,212285,217215,212284,212282

4.3.4 Intel Software Guard Extensions (Intel SGX)

For applications that require enhanced security due to strict isolation requirements, Intel has developed Intel Software Guard Extensions (Intel SGX) which is a set of new instructions that allow developers to partition their applications into “enclaves” (placed in trusted memory), where they can run in an encrypted state and in complete isolation[21].

Application of such levels of security is typically found in the military and highly regulated commercial organizations, where it can provide hardware level protection against memory attacks.

For maximum future compatibility, it makes sense to include Intel SGX in the requirements so that it is available if required. Most Intel 3rd Generation Intel Xeon Scalable processors support these extensions.

However, there are some important considerations that should be borne in mind when using these extensions. For example, when used with 3rd Generation Intel Xeon Scalable processors in vSphere, enabling this feature requires the host to register its SGX cryptographic identity with a public Intel service[22], something a highly secure organization may not allow. A workaround is to use Intel SGX Data Center Attestation Primitives (Intel SGX DCAP)[23] which allows organizations to develop their own local attestation service. This removes the need for an externally provided service.

In addition, due to the nature in which Intel SGX enabled applications run (i.e., within encrypted and trust memory-based enclave), certain well-known vSphere features become incompatible and are restricted. Those features include vMotion, Fault Tolerance, suspending, or taking a snapshot of a VM, to name a few.

It remains an important option available in the security portfolio to secure workloads if the requirement arises, which is an important consideration while designing a datacenter that will serve customers with diverse requirements over its lifetime.

4.3.5 Memory

The combination of 3rd Generation Intel Xeon Scalable processors with Intel Optane Persistent Memory 200 series brings enhanced capabilities compared to other server processors, but also means more memory slots will be available for

---

21 “Confidential Computing with Intel SGX,” presented by by Scott Raynor and Laura Martinez at Cloud Field Day 10 on March 10, 2021 (techfieldday.com)

22 Limitations of Virtual Software Guard Extensions (SGX) in vSphere (71367) (vmware.com)

23 Intel SGX MultiPackage DCAP Support (01.org)

capacity.[24] As discussed above, Intel's 2022 announcement that it is not pursuing further development of Optane technology does not change our view that these products give game-changing benefits to current-generation servers. Since Intel is committed to supporting current Optane products and they are widely available in the channel, we do not hesitate to recommend using them with today's Xeon Scalable server platforms. In fact, it only strengthens the financial advantage of Intel platforms.

Given the common practice to allow a 1:1 physical to virtual memory ratio for workloads, it pays to add as much memory to a host as the architecture allows. To maximize the amount of memory per host while keeping the costs in check, it is recommended to use a combination of traditional RDIMMs and Intel Optane Persistent Memory in "Memory Mode". In practical terms, the host will have 8 channels per socket with 2 DIMM slots per channel, providing 16 DIMM slots per socket. That equates to 32 DIMM slots per host for a dual-socket host.

To achieve the maximum capacity and bandwidth, an "8+8" configuration is recommended. That means for each of the 8 memory channels, the first slot contains a 32 GB 3200 MHz DRAM RDIMM and the second slot has a 128 GB Intel Optane Persistent Memory 200 Series DIMM. Doing so provides a 1:4 ratio of DRAM vs PMem with the total tiered system memory of 2 TB with 512 GB of RAM acting as the cache tier. As the intended workload for our use case is heterogenous, accurate performance data is not likely to be available. For such a case, a ratio of 1:4 provides the most effective performance coverage and is therefore, the recommended choice.

While memory prices can vary, the cost of 2 TB of tiered memory is roughly the same as 1 TB of traditional DRAM that the system would otherwise have. Depending on actual workload characteristics at the time, there might be some performance impact due to this replacement. However, it doubles that amount of memory available for workload consolidation on the host for the same memory cost.

The cost saving is not limited to just the amount of memory. This memory configuration swap allows organizations to go beyond the native memory capacity of that architecture. As the number of hosts required to fit the same amount of workloads is potentially halved as a result, further cost savings are made as a positive consequence.

4.3.6 Storage

Traditional appliance-based external storage has always been a cost-effective option for vSphere environments. However, vSAN remains a popular choice for organizations that do not want to manage such appliances and prefer a hyperconverged and natively integrated storage solution.

---

[24] "A New Vision for the Memory-Storage Hierarchy with Intel," presented by Kristie Mann at Storage Field Day 22 on August 4, 2021 (techfieldday.com)

Due to its architecture, vSAN relies heavily on data objects being written to its cache quickly, before the data is destaged to the capacity disks. Depending on the protection mechanism selected, I/O amplification causes multiple write operations for the same data, which adds to the cache write activity. Read requests are also directed towards the cache in the first instance; the capacity storage only needs to serve the read request on a cache miss. Finally, the destaging process itself requires the cache to serve data.

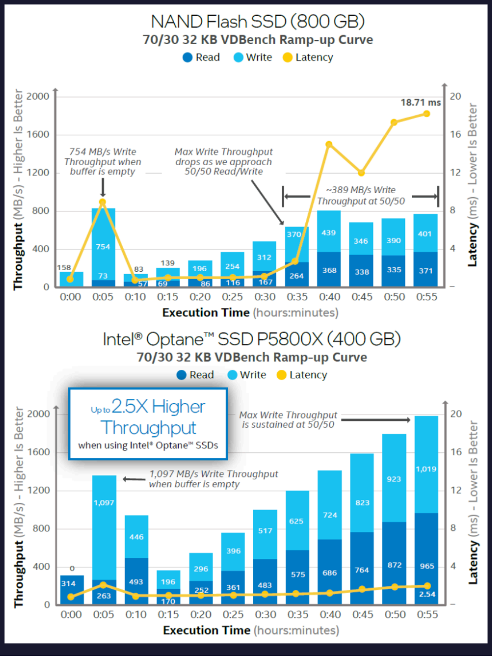
All these I/O intensive operations are constantly occurring which makes the cache the most critical component of the vSAN architecture. Given that a vSAN datastore cannot function without its cache, the performance and endurance of this storage are the two most critical factors to consider when making the selection.

The Intel P5800X SSD, which brings both NVMe and Optane technology, is our recommended choice for cache devices, with each allocated to its own disk group. Intel Optane SSDs write-in-place technology enables it to provide a much higher write throughout than NAND Flash SSDs, which need to go through a program-erase cycle before going back to a programmed state. For that reason, Intel Optane SSDs can sustain the performance required for normal operation while destaging data to the capacity tier without affecting performance.

In addition, Intel Optane SSDs can have up to 16 times more endurance than a NAND Flash SSD, which is a crucial factor due to the dependency of vSAN on high write throughput. Using Intel Optane SSDs ensures the cache will outlive the typical five-year warranty period and will

Physical Cache Disk Throughput Comparison

- 2.5x higher throughput
- Enables smaller cache disk size
- Endurance enough to outlast the host!



reliably keep serving its cache function for the expected lifetime of the vSAN deployment.

In terms of sizing of the cache disk, tests have shown that when using Intel Optane SSD P5800X (that provides 100 DWPD), a disk size of 400 GB each is sufficient (as compared to the typical cache disk size of 600 GB). That is made possible because of the ability of Intel Optane SSDs to sustain a higher read/write bandwidth, allowing them to destage data to the capacity layer as quickly as the I/O activity demands.

4.3.7 Network

If packing as many virtual workloads as possible on a particular host is the aim and no restriction on network consumption is assumed, it is prudent to choose the best networking connectivity that can be afforded for that host.

However, on a typical virtualized environment, the virtual workloads may not be the biggest consumers of network bandwidth. Instead, it is often the storage environment that exerts the most pressure on the network provision. That is generally true when traditional storage appliances are used, but especially so when hyperconverged storage systems like vSAN datastores are providing the host storage. With the availability of vSAN HCI Mesh now, which enables vSphere clusters to consume remote vSAN datastores, having the ability to provide high speed networking connectivity has never been more important.

Keeping this in mind, we recommend using Intel Ethernet 800 Series adapters with PCIe 4.0 interfaces. These adapters support RDMA and RoCE v2 and use of those, along with 100 GbE speeds (25 GbE minimum) is recommended by VMware where possible. Setting the Maximum Transmission Unit (MTU) to 9000 is also highly recommended[25].

4.3.8 Other Factors

In addition to these hardware components, it is crucial to ensure all the related configurations also match appropriately in order to extract maximum performance out of a host.

It is important to ensure that the system's BIOS supports the latest 3rd Generation Intel Xeon Scalable processors along with the Intel Optane Persistent Memory 200 Series. Systems that support these two technologies also have arbitration profiles to determine the algorithm that the memory controller uses to switch between DRAM and Intel Optane Persistent Memory when carrying out transactions. Tests have shown that when using "Memory Mode," it is best to use the "Balanced Profile" to provide a good balance between bandwidth and latency.

It is also worth confirming the status of the following three settings as by default, they might not be set to provide maximum performance:

- Intel Virtualization Technology (Intel VT): Enabled

[25] Intel Ethernet Overview and 800 Series (youtube.com) presented by Brian Johnson at Networking Field Day 21 on October 1, 2019 (techfieldday.com)

- Intel Virtualization Technology for Directed I/O (Intel VT-d): Enabled
- Hyper-Threading: Enabled

Some applications could possibly be negatively impacted by enabling hyperthreading. However, as the workload is assumed to be heterogeneous, the performance benefit gained by enabling hyperthreading far outweighs the potential performance risk to some applications.

This is not meant to be an exhaustive list of factors to consider. Rather, it is provided as a baseline upon which the thought process can begin. When the aim is to find a configuration with the best balance between cost, flexibility, and performance, this commentary provides a good place to start.

It is also clear that striking that fine balance is not reliant on a particular component, such as a processor with a high number of cores or extreme operating frequency. In fact, such components are often the most expensive option, and they do not provide good value for money.

The optimum way is to select the widest range of capabilities at the most cost-effective price and refine performance through incremental optimization of every component that makes up the system. This strategy also distributes the risk to performance system wide and therefore, the system is not affected materially when one component must be replaced or does not perform as expected.

It can be safely concluded that consistently high performing systems do not rely on just the number of cores in a processor or their clock speed as a processor is just one component out of many that make up a host. Ultimately, it's all about the sum of the performance gains extracted out of all the components in a host, provided through optimizations made at every level due to the ever improving firmware/code, and all those components working together seamlessly.

4.4 The Importance of a Diverse Ecosystem

The value of building systems using components from the same ecosystem cannot be overstated. It is obvious that achieving performance gains through optimization is more likely when most (if not all) components belong to the same ecosystem and where the manufacturer has complete end-to-end visibility of how those components operate. Having that comprehensive view allows the manufacturer to work towards optimizing those components to operate together in the most efficient way.

A fact often missed is that such optimizations and compatibilities are often introduced into the ecosystem components long before the beneficiaries of them come into existence. A good example of that is when support for new instructions precedes the introduction of the hardware itself. As new technologies are developed, the manufacturer can start rolling out code that prepares its customer base for the upcoming innovations. That capability is a major benefit for customers operationally, in terms of proactive compatibility and performance optimization, and one that can only be meaningfully delivered by a manufacturer with a complete range of datacenter products.

To have a stable datacenter, it is also critical to minimize the number of manufacturers where possible. That strategy minimizes supply chain and support issues. It is not uncommon to see finger-pointing to start between vendors when multiple manufacturers are involved in an incompatibility or bug, causing major delays and inconvenience when troubleshooting.

Datacenters are designed to serve well into the future and therefore, it is imperative to choose a platform that is likely to serve an organization's requirements today but also going forward in the face of constant innovation. To do that effectively, one should carefully consider the pedigree but also keep a keen eye on the future and upcoming technologies from the manufacturers being considered.

These considerations are as important as the technology options themselves and should be given due attention when designing the datacenter fit for the future. It is for that reason that the recommendations made in this section are the combined result of these considerations and the performance and capacity options as publicly documented by Intel and VMware.

5. CONCLUSION

Architecting a datacenter or cloud computing infrastructure requires a very different approach from building a standalone computer. A flexible cloud architecture has been adopted both inside the datacenter and outside. This is the platform of choice for modern applications. Many organizations are using public cloud providers, but most have a hybrid approach, with some workloads hosted on-premises or locally due to regulatory or data sovereignty considerations. But the software-defined cloud platform approach is becoming ubiquitous.

With this in mind, we considered the workloads that are driving purchasing today. Exemplary of these are support for cloud applications, hyper-converged and virtual desktops, AI/ML, and other high performance data processing. Our research focused on these workloads. We even took the time to construct an HCI stack in the CTO Advisor Lab to gain up to date hands-on experience. Thus, we feel confident making recommendations for this type of infrastructure.

When purchasing server hardware, one must consider the entire compute stack from CPU core to memory channels and capacity to I/O, networking, and storage. Increasingly, components outside the CPU have a greater impact on performance, scalability, and sustainability as those cores must be supported with data. Most large-scale systems do not “max out” core count per socket and focus instead on balance between all of these elements. Thus, we recommend a server infrastructure leveraging 3rd Generation Intel Xeon Scalable processors with support from Intel Optane Persistent Memory, Ethernet 800, and accelerators.

6. References

6.1 Whitepaper References

- Twitter Tail Latency Caching with Ethernet Series-800 (pelikan.io)
- Lenovo Optimizing PMEM (lenovo.com)
- The Competition (neurips.cc)
- Intel® Labs Presents Research at NeurIPS 2021 (intel.com)
- Technology Update on 3rd Gen Intel Xeon Scalable Processors, Intel Data Center Update 2021 (youtube.com)
- Technical and Performance Overview for 3rd Gen Intel Xeon Scalable Platform, Intel Data Center Update 2021 (youtube.com)

6.2 Gestalt IT and CTO Advisor Research

- The CTO Advisor OCI VMware Report (thectoadvisor.com)
- The CTO Advisor VxRail Research (thectoadvisor.com)
- Storage Field Day 23: Intel Optane Reference (youtube.com)
- More Twitter Real World Podcast (thectoadviser.com)
- Driving Data Tiering Innovation with Intel Optane and CXL, Tech Field Day 25 (techfieldday.com)
- Confidential Computing with Intel SGX, Cloud Field Day 10 (techfieldday.com)
- Intel Ethernet Overview and 800 Series, Networking Field Day 21 (techfieldday.com)
- Optical Networking at Scale with Intel Silicon Photonics, Networking Field Day 25 (techfieldday.com)
- A New Vision for the Memory-Storage Hierarchy with Intel, Storage Field Day 22 (techfieldday.com)
- VMware and Intel Optane – Delivering Support for Today's Demanding Storage Workloads, Storage Field Day 22 (techfieldday.com)

About The Authors

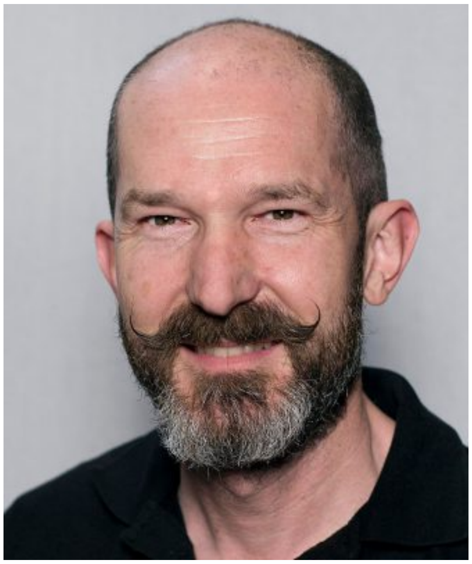
ATHER BEG

Ather is an Enterprise Solutions Architect and works for Rackspace Technology. His focus is on all things related to cloud, technology, storage, virtualization and whatever comes in between.

Being in the industry for nearly 25 years, he feels ancient. If you feel that inclined, he can bore you with stories on how he used to manually park heads on a hard drive or bound protocols to network cards. Seriously though, he has designed, deployed and managed many enterprise environments, involving all the major cloud platforms, VMware Clouds, virtualization and storage technologies – to name a few.

Ather started blogging over 11 years ago so that he can share some of his knowledge with the community. He has been a vExpert for 9 years running and is also vExpert NSX/HCX. He has been an official VMware blogger at VMworld EU and US for the few years too. He is one of the founding members and contributor to Open HomeLab Wiki and co-hosts @OpenTechCast as well.

Ather's natural habitat is tech events like VMworld, Cloud (and other) Field Days, VMUGs etc. and he thrives on meeting like-minded people and having a good old chat about technology. He's friendly and not dangerous at all so please do interact with him whenever you spot him in such surroundings.



ALASTAIR COOKE

Alaistair is a self-employed consultant and writer based in New Zealand. Much of his work is communicating about how IT infrastructure works and what it means for IT professionals. Outside of work he is datacenter and VMware focussed, having created the AutoLab which is a free tool to simplify the creation of a vSphere test or training

STEPHEN FOSKETT

Stephen Foskett is an active participant in the world of enterprise information technology, currently focusing on enterprise storage, server virtualization, networking, and cloud computing. He organizes the popular Tech Field Day event series for Gestalt IT and runs Foskett Services. A long-time voice in the storage industry, Stephen has authored numerous articles for industry publications, and is a popular presenter at industry events. He can be found online
at TechFieldDay.com, blog.FoskettS.net, and on Twitter at @SFoskett.

CRAIG RODGERS

Craig has over 20 years IT experience working with SMB's and Enterprise's providing solutions within the cloud, virtualisation, storage, network and Microsoft stacks. Recent roles have involved hands on architecture and deployments up to enterprise and government sector clients. Craig holds certifications from Microsoft, VMware, Citrix, Cisco, DellEMC, Huawei and Zerto, has previously been awarded VMware vExpert and is a Veeam Vanguard for the past 3 years. Craig is passionate about technology and thrives on a good challenge.