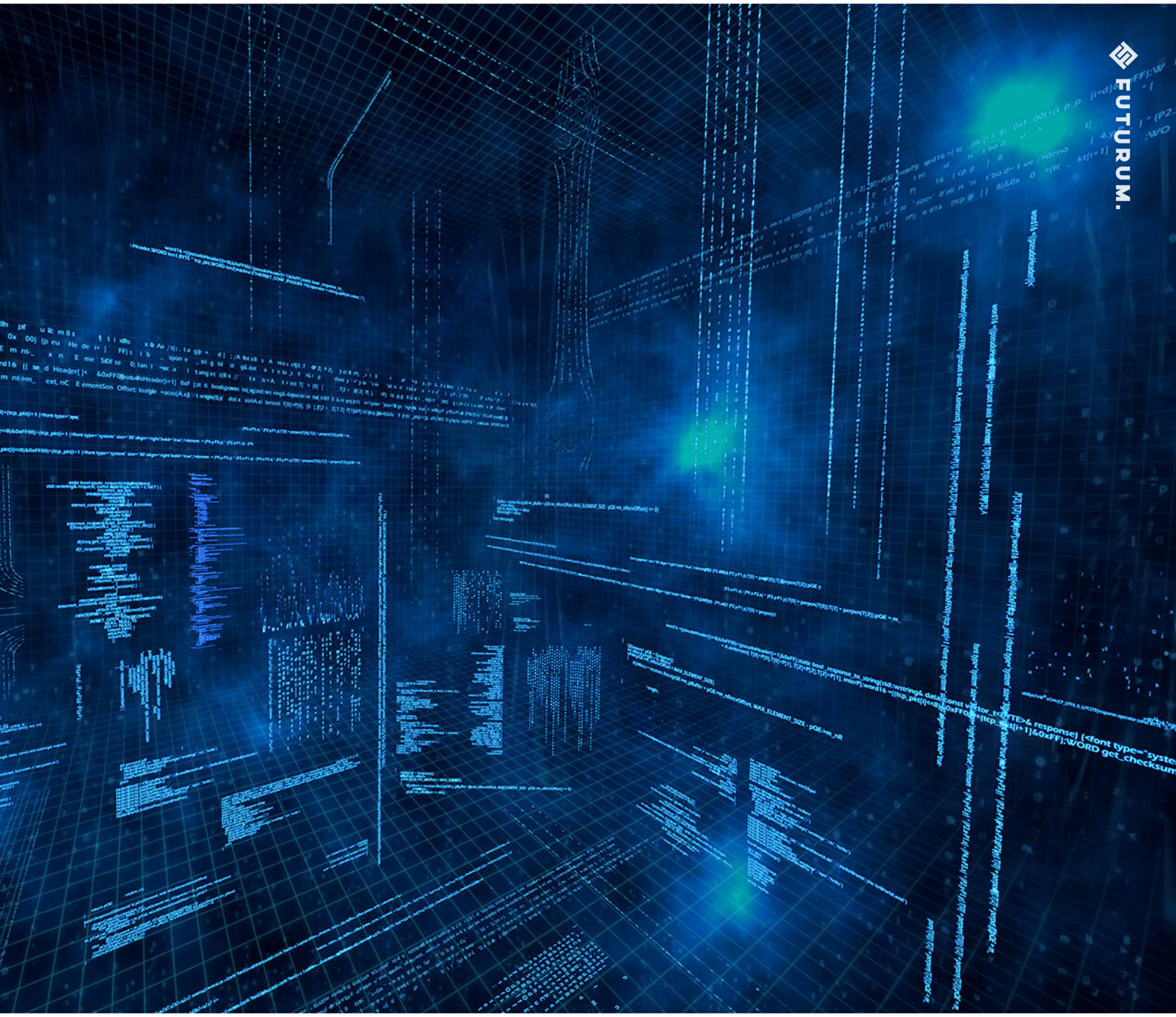


The Rise of Confidential Computing

Trust: The New Battlefield in the Age of Digital Transformation

April 2021

DANIEL NEWMAN
Founding Partner + Principal Analyst

SHELLY KRAMER
Founding Partner + Lead Analyst

Published: April 2021

FUTURUM.

TABLE OF CONTENTS

Part 1. The Commercial Landscape

The insider hack is not a new occurrence. In 1671, Thomas Blood almost managed to steal the Crown Jewels by befriending the keeper of the jewels, Talbot Edwards. After gaining trust, Blood convinced Edwards to let him into the Jewel House. Once inside, the thieves knocked Edwards unconscious and managed to take the jewels. In a more modern context, the Edward Snowden Wikileaks example is more recent and high profile. Snowden had the right credentials and system access to gain access to the data he subsequently made public.

The reasons for highlighting these two cases are that the security in question is nothing to do with perimeter security. In the first case, a more secure front gate or in the second case, a better firewall would not have protected against either of these types of attack.

The IT industry has woken up to this threat vector and a new dynamic is emerging in the common lexicon — Confidential Computing. This research brief will cover:

1. The challenges faced by businesses who rely on the trust of their customers, suppliers, or ecosystem partners. Specific focus will be placed on recent hacks, their impact, and lessons we can learn.

2. A look at Confidential Computing and how vendors are looking to deploy trust-based computing models, specifically in areas such as Trusted Execution Environments, Enclaves, and homomorphic encryption

3. The various approaches key vendors are making to address the business and technical requirements. The focus here will be on underling chip-based approaches from the likes of AMD, Intel and IBM. We will also cover how these approaches are being translated to the public cloud with services from AWS, Azure, and IBM.

The SolarWinds Hack of 2020

Even months after SolarWinds hack, it continues to dominate the narrative in security circles as more details surface. The hack has unprecedented reach with large U.S. federal government agencies, Fortune 500 companies, security and operating system vendors all affected.

The SolarWinds hack has put a harsh spotlight on the fragility of the software development cycle. The key takeaway from this hack: if your Continuous Integration/ Continuous Delivery (CI/CD) pipeline and your code release processes are unsafe, the consequences can be catastrophic. More worryingly for the consensus thinking, all the investment and resources organizations put into vulnerability scanning, automation, and DevOps training is money wasted.

Behind the Headlines – What was the Root Cause of the SolarWinds Hack?

SolarWinds is an Austin, Texas-based information technology firm. One of SolarWinds' products is a software system called Orion that is widely used by companies to manage their IT resources. According to SEC documents, SolarWinds has some 33,000 customers who use Orion. Hackers breached SolarWinds' systems and inserted malicious code into the software build process. The breach of the CI/CD pipeline went undetected for many months and, as a result, numerous product updates were unwittingly shipped by SolarWinds to customers that included the inserted vulnerabilities. The inserted malicious code introduced a backdoor allowing hackers to gain access to the software running on SolarWinds' customers' infrastructures.

The hackers found a way to legitimize the malicious code by injecting it into the build pipeline. The SolarWinds CI/CD build pipeline was producing digitally signed and trusted software for over 18,000 customers worldwide. The real issue for clients is complex. The build pipeline produces builds, and these builds are digitally signed with the SolarWinds certificate trusted by the Certificate authorities in various operating systems and browsers. If clients were to revoke the digital certificate, they would be revoking both the good and the bad software, as it would be impossible to differentiate between good or bad code.

Lessons We Can Learn from SolarWinds Hack

Since the SolarWinds hack, the industry has mobilized to put forward methodologies and approaches to prevent this kind of breach from happening in the future. Of course and inevitably, this unfortunate occurrence, experienced by one of the IT industry's most trusted brands, has led to vendors jumping on the bandwagon and using the hack in their marketing collateral to spread fear, uncertainty, and doubt with their clients. One takeaway however, is clear: there is a burgeoning need for a comprehensive, holistic Development Security Operations (DevSecOps) strategy that aims to provide full observability of the CI/CD build pipeline and its various stages. Simply put, the focus needs to shift toward a concept of a trusted CI/CD pipeline — and quickly.

Other Significant Recent Data Breaches — and the Costs Associated With Those Breaches

Data breaches and security incidents are becoming increasingly costly and an almost daily occurrence. In December of 2019, Norwegian aluminum manufacturer Norsk Hydro, a global manufacturing giant, experienced a breach when malicious code tore through its network, forcing the company to shift to manual mode and causing tens of millions in damages. Norsk Hydro admitted the bill for a crippling cyberattack could top $75 million. Discovered in June 2019 (by a police department and not the company itself), financial services provider Desjardins Group experienced a data breach that involved some 9.7 million active and inactive files of individuals with accounts at the company's credit union branches located primarily in Quebec and Ontario. Desjardins' own security weaknesses were determined to be the cause of the breach, including a series of gaps in administrative and technological safeguards. The company estimated the cost of recovering from the breach is likely to be $108 million, far more than it originally estimated.

British Airways was fined £20 million by the Information Commissioner's Office (ICO) in a data breach scam that compromised the information of more than 400,000 customers. In the Marriott data breach, hundreds of millions of people had their passport and credit card numbers stolen, made possible by a security failing on the part of Marriott. While the $28 million in expenses incurred by the company by the spring of 2019 may seem insignificant, the ICO levied a fine of £99m (more than $120 million US) on the company for violating British citizens' privacy rights under GDPR.

We would be remiss to leave out the 2017 Equifax data breach, which exposed the personal information of 147 million people and resulted in the company agreeing to a global settlement with the Federal Trade Commission, the Consumer Financial Protection Bureau, and 50 U.S. states and territories of some $575 million dollars.

Also note that the DOJ has been investigating these breaches, along with the April 2015 discovery of the years-long breach of the United States Office of Personnel Management (OPM). OPM is the agency that manages the government's civilian workforce, and their personnel files containing extremely sensitive data had been exfiltrated.

Coincidentally, in February of 2020, the U.S. Department of Justice formally charged four members of the Chinese military with that attack, and also noted that the Equifax attack was explicitly linked to the Marriott and OPM breaches.

They key thread connecting all these breaches — poor security practices and smart, determined, patient threat actors.

The financial impact of a data breach is what keeps CISOs and their CEOs awake at night. A new report from IBM and the Ponemon Institute estimates the average cost of a data breach in 2020 is $3.86 million.That sum seems paltry when compared with the millions of dollars in fines in the above-referenced high-profile data breaches, but for many organizations, a hit like that would be significant.

"The overall headline number stayed very similar to what we saw last year," says Charles Debeck, senior threat analyst at IBM X-Force IRIS, "but if you dig deeper into the data, what we saw was an increasing divergence between organizations that took effective cybersecurity precautions versus organizations that didn't. This divergence has been increasing year over year; the organizations that are engaging in effective cybersecurity practices are seeing significantly reduced costs, the organizations that aren't engaging in these same practices are facing significantly higher costs."

The Cost of Being Unprepared for a Data Breach

According to the Ponemon's 2020 Cost of Insider Threats Global Report, organizations face the highest costs with the average being $8.19 million per breach. When you break these costs down to the per record level the average cost of each lost record has gone down slightly to $146 from $150 in 2019. The most expensive type of record to lose was customer PII records, which were involved in around 80% of breaches in the study.

The costs to remediate a data breach and notify clients, however, is not the total picture. Ponemon's report indicates that nearly 40% of the cost of a data breach is driven by lost business. These costs manifest themselves as:

- Increased churn
- Lost revenue driven by system downtime
- Reputational impact
- The cost of customer acquisition

The impact on industry by industry varies. "Healthcare has been the number-one industry in terms of average cost of a data breach for ten years in a row now," says IBM's Debeck. "It's a highly regulated industry and faces a lot of regulatory burdens when it comes to remediation of a breach, and there are a lot of additional costs with medical records compared to other types of record." In late December 2020, IBM discovered that hackers targeted the COVID-19 vaccine supply chain by way of global phishing campaigns launched in Germany, Italy, South Korea, the Czech Republic, greater Europe, and Taiwan and in late February 2021, hackers broke into Oxford University's biochemical systems and COVID-19 research and offered the information online for sale. Oxford University develops the Oxford COVID-19 vaccine in partnership with AstraZeneca.

The takeaway is that if you are in a highly regulated industry such as Healthcare or Banking that not only makes the organization a prime target for threat actors, and the costs and impacts can be significantly higher.

The study looked at the long tail of data breaches and found that organizations are paying the price of a data breach for many months after the initial impact. Around 61% of the cost of a breach comes in the first year, around 24% comes in the next 12 to 24 months, and the final 15% comes more than two years later. The Equifax breach mentioned earlier has the dubious honor of being one of the most expensive data breaches of all time, with a $575 million dollar settlement that could potentially end up at $700 million when it's all said and done.

In addition to material costs, publicly traded organizations will see their stock value affected by data breaches as well. An in-depth article on Comparitech analyzes the stock value of companies on the New York Stock Exchange in the aftermath of 33 data breaches of at least 1 million records.

Share prices of breached companies hit their lowest — around 7.3% down — around 14 market days following a breach and underperform the wider NASDAQ by -4%. While companies are likely to see their share price rebound and perhaps even rise ahead of the market average in the first six months after a breach, they are still likely to underperform on the NASDAQ by -6.5% 12 months later. As a recent example, in November 2019 Macy's stock had dropped 11% in a single day after it disclosed a breach and suffered a "highly sophisticated and targeted data security incident…that affected a small number of customers during a one-week period in October." However, by the end of December that year the company's share price had recovered.

"Companies that leak highly sensitive data like credit card numbers, including Macy's, typically see a steeper drop in share price than companies that leak less sensitive data," says Paul Bischoff, privacy advocate at Comparitech. "Our research shows companies see an initial drop in share price for

about three weeks following a data breach, after which it recovers. Six months post-breach, most companies have fully recovered and even outperform the prior six months in terms of share price. Our analysis also shows more recent breaches have less negative impact on share price than older ones, a sign of breach fatigue among consumers who have grown accustomed to their data being stolen."

The Career Impact of a Data Breach

The consensus opinion based on multiple sources is that 1 in 4 of companies reported executive firings related to application attacks. U.S. companies were more likely to say execs were let go after an incident, as were companies in the technology or financial services sectors.

While the CISO is not always let go — Kaspersky reports that senior non-IT employees are laid off at 27% of enterprises (those with over 1,000 employees) that suffer a breach — their positions can often be at risk if there were clear security failures. Nominet partnered with Osterman Research for a research study of over 400 CISOs in the U.S. and the UK. The CISO Stress Report — Life Inside the Perimeter: One Year On, on the working life of a CISO showed that CISOs are stressed and they most definitely worry about data breaches and the career impact of a breach. The report found that some 6.8% of CISOs in the U.S. and 10% in UK believed that in the event of a breach they would lose their jobs, and just under 30% of survey respondents believed they would get an official warning.

Here are five major security incidents that cost security leaders their jobs in recent years:

Capital One
In July 2019, Capital One announced an attacker had gained access to the personal information of over 100 million card customers and applicants. The bank didn't learn about the breach until 127 days later, when it was tipped off by an outside researcher. Capital One has said it expects the incident to cost between $100 million and $150 million. In November 2019, shortly after the breach, the WSJ reported that Capital One had replaced Michael Johnson, the chief information security officer during the massive breach, with Mike Eason, the chief information officer of Capital One's commercial bank. In December of 2020, Capital One hired Goldman Sachs' CISO Andy Ozment as head of technology risk to help with the breach response.

Equifax
In 2017 Equifax was compromised via an unpatched consumer complaint web portal. This led to some 143 million customer records — including names, addresses, dates of birth, Social Security numbers and driver license numbers — being stolen.

The U.S. House of Representatives Committee on Oversight and Government Reform called the incident "entirely preventable," while U.S. Senate Permanent Subcommittee on Investigations accused the company of a "neglect of cybersecurity." The cost of the incident is estimated to be $1.35 billion. The company paid $575 million (potentially rising to $700 million) with the Federal Trade Commission and others. Many customers impacted by the breach people opted for money rather than free credit monitoring which has quickly become a standard offering. Both CSO Susan Mauldin and CIO David Webb left the company in the weeks after the breach.

Uber

In late 2017, ride-hailing company Uber revealed the data of 57 million Uber customers and the details of 600,000 Uber drivers had been stolen, including names, email addresses, phone numbers and driver's license numbers. Attackers reportedly accessed Uber's private GitHub code repository to gain access to the data. Uber's chief security officer Joe Sullivan, who had previously served as Facebook's CSO for five years, was fired from as a result. In December of 2020, the U.S. Department of Justice charged Sullivan with lying to management about the security breach and paying hush money to the hackers. At a press conference announcing the charges, U.S. Attorney for the Northern District of California David Anderson said that by hiding the Uber hack from authorities and management, Sullivan indirectly helped the hackers breach other companies. If convicted on both charges, Sullivan could face up to eight years in prison. He currently serves as the chief information security officer at internet company Cloudflare.

Target

The 2014 attack on US retailer Target is still spoken about today because it was one of the most notable cases of a successful supply chain attack — hackers exploited poor security in an HVAC vendor to compromise Target's payment systems and steal the payment details of some 40 million customers attack over the Christmas period in 2013. The company later said that hackers also stole personal information, including names, phone numbers, and email and mailing addresses from as many as 70 million customers.

CIO Beth Jacob resigned in the months following the attack as the company overhauled its security posture and appointed its first CISO, former GE CISO Brad Maiorino, shortly afterwards.

JP Morgan

2015 saw both JPMorgan Chase's CSO Jim Cummings and CISO Greg Rattray reassigned to new positions within the bank in the wake of its 2014 breach of over 83 million accounts in the U.S., including names, email and postal addresses and phone numbers.

The consensus opinion based on multiple sources is that 1 in 4 of companies reported executive firings related to application attacks. U.S. companies were more likely to say execs were let go after an incident, as were companies in the technology or financial services sectors.

Going Deeper On the Need for Trust

Now that we have established the burgeoning need for an improved security posture based on the significance, impact and frequency of security breaches we need to understand what the new market dynamic is. Security has been a foundational layer within IT provision for decades so what has changed and why is a new paradigm emerging?

Operational Trust vs. Technical Trust

Before we dive further into the topic of Confidential Computing, for contextual purposes, we'll first delineate between Operational and Technical Trust. The prevalent mindset in the enterprise world as a whole is that security is a people and process problem. The gospel that is preached is that through better and ongoing training of staff, and by applying strict rules around access, and ultimately to compliance and certification that the required level of enterprise-wide security posture can be achieved. This mindset is known as achieving an ultimate state of Operational Trust.

We believe that while Operational Trust is an important element of an enterprise security mindset and culture as a whole, it does not, in and of itself, provide sufficient protections. People will always be people, accidents can and will happen, mistakes will undoubtedly occur, and hackers will hack.

This is where the domain of Technical Trust comes into play. Technical Trust is, as you might imagine, the focus on removing people from the security equation altogether through the deployment of technological solutions to address the fundamental security concern. This approach focuses on technology solutions rather than training, process improvements, compliance, and certification.

Simply put, in order to function at the highest level of secure operations, what the industry needs is a way to make it possible to run applications on an individual's computer, but in such a manner that the owner of that computer can neither influence nor observe what's happening. Moreover, this must be able to be accomplished through the deployment of technology, with no reliance on human intervention.

The States of Data

Let's talk for a moment about data and the states of data. Data exists in one of three states:

- At rest on a storage device
- In transit between two locations across a network
- When it is in use as it's being processed by applications

Credit card data, medical records, personnel records, corporate data, intellectual property, and customer data, from firewall configurations to our geolocation data, protecting sensitive data in all of its states is more critical than ever.

Now, let's talk about protecting data. You can protect data at rest by encrypting it. You can protect data in transit. It's a little trickier, but you can encrypt that, too. What about protecting data while you're using it? You need to unencrypt the data to use it, right? It would be hard to read your email if it's encrypted while you're trying to look at it. So yes, under normal circumstances, you would need to unencrypt data in order to use it. And therein lies the problem. When you unencrypt data in order to use it, that data is then in memory, which can be dumped, potentially providing threat actors access to your unencrypted data.

Cryptography is now commonly deployed to provide both data confidentiality, by stopping unauthorized viewing, and data integrity, which prevents or detects unauthorized changes. While techniques to protect data in transit and at rest are now commonly deployed, the third threat vector — when data is in use as it's being processed by applications — is more of a challenge. That's where Confidential Computing comes in. How to manage and, more accurately, secure data in use is the purview of Confidential Computing. More on that in a bit.

When we want to work with data, we have to decrypt it to allow an application or user to use it. This issue is further complicated in the mode where the administrators of our data repositories are no longer employees we trust or perhaps they are no longer employees. When the date resides on a public cloud service, there's the challenge of ensuring the container or virtual machine operators are operating in a trusted way.

One commonly adopted approach is processes and compliance statements. The ISO/IEC 27001 is the most commonly cited certification in this space. Originally published in 2005, this international standard provides a robust framework for providers and end users alike to focus on as they look to improve the security posture of their operations. However, full deployment and certification of ISO/IEC 27001 would not stop the Edward Snowden style hack where the admin has all the right credentials to access the system.

How to manage and, more accurately, secure data in use, is the purview of Confidential Computing.

Part 2. Confidential Computing: What it is, Why it's a Game-Changer

As we mentioned earlier, Confidential Computing aims to address computational trust and security for data in use, enabling encrypted data to be processed in memory without exposing it to the rest of the system by way of the utilization of Trusted Execution Environments (TEEs). It also aims to reduce exposure to sensitive data and provide greater transparency for users. Although in its nascent stages, we are very excited about Confidential Computing and the wide range of benefits it will afford organizations who very much need the data protection it can provide.

Here are some of the areas we believe Confidential Computing will ultimately be able to deliver significant value:

- Confidential Computing is focused on enabling end-to-end security encryption.
- It will put into place technology-focused approaches, rather than compliance and policy-based approaches. This will ensure the protection of data while in the state of being utilized or processed.
- Confidential Computing will provide a model where cloud customers have higher authority over their data and their processing at all points in the management of data.
- It will allow the development of technology deployment options to ensure protection against nefarious use by insiders and system administrators.
- The management of encryption keys is also a discipline within Confidential Computing that should not be overlooked.
- Confidential Computing should make it easier for organizations to move data between different environments without exposing any sensitive data.

The Community Approach – Confidential Computing Consortium

One of the most exciting things about Confidential Computing is that although in early stages, some of the biggest names in technology are already working in the space. Even better, they are partnering and working to use their powers for good.

Established in 2019, the Confidential Computing Consortium (CCC) is a Linux Foundation project and community dedicated to refining and accelerating the adoption of Confidential Computing. The intent to form the organization was announced at Open Source San Diego in 2019, and brings together cloud providers, hardware vendors, open source experts, developers, and academics to accelerate the Confidential Computing market.

This Consortium was among the first industry-wide initiatives to address data in use as more companies move more of their workloads to span multiple environments, from on premises to public cloud and to the edge.

The current members of the CCC include:

- Alibaba
- Arm
- Baidu
- ByteDance
- Fortanix
- Google Cloud
- Huawei
- IBM
- Intel
- Microsoft
- Red Hat
- Swisscom
- Tencent
- VMware

Whether on the public cloud, on-premises servers, or the edge, the CCC is working on making it easier to run and move quickly between various environments. This initiative is also working on:

- Supporting Confidential Computing by hosting technical open-source projects and open specifications.
- Bringing hardware vendors, cloud providers, and developers together to grow its market value.
- Setting up the regulatory standards.
- Building an open-source tools environment for TEE development by building proper open-source tools.

The Vital Role of Hardware in Confidential Computing

When we look at a full stack view, the overall security of the stack is only as strong as the layers below it. If any layer lower down the stack is compromised, every layer above it in the stack is inherently compromised.

As a result, when we are looking to implement the ultimate layer of security, we must focus on the lowest elements of the stack. In this case, that means a focus on the silicon components of the hardware. By focusing on security at the lowest layers of hardware, it is therefore possible to remove the role of vendors in the following layers of the stack:

- Operating system
- Device driver
- Platform
- Peripheral vendors

Then, we can go even further into the realm of who administers and operates the systems, be they on-prem or in the cloud and remove them from the equation as well. In this case, we are specifically talking about service providers and their admins. This means we can remove these admins from the list of required trusted parties and, as a result, reduce exposure to potential compromise at any point in the system lifecycle.

To further emphasize the crucial role that hardware plays in providing Confidential Computing and on delivering the goal of decreasing the reliance on proprietary software for Confidential Computing environments, the Confidential Computing Consortium has excluded from its scope Trusted Execution Environments (TEEs) that have only software roots of trust and focused on hardware-based security guarantees for Confidential Computing environments.

What is a Trusted Execution Environment?

According to Trustonic, a Trusted Execution Environment (TEE) is an environment for executing code, in which those executing the code can have high levels of trust in the asset management of that surrounding environment because it can ignore threats from the "unknown" rest of the device.

The definition of a TEE as defined by the Linux Foundation's Confidential Computing Consortium is an environment that provides a level of assurance of the following three properties:

1. Data confidentiality: Unauthorized entities cannot view data while it is in use within the TEE.
2. Data integrity: Unauthorized entities cannot add, remove, or alter data while it is in use within the TEE.
3. Code integrity: Unauthorized entities cannot add, remove, or alter code executing in the TEE.

In the context of Confidential Computing, unauthorized entities could include other applications on the host, the host operating system and hypervisor, system administrators, service providers, and the infrastructure owner — or anyone else with physical access to the hardware.

Together, these attributes provide not only an assurance that the data is kept confidential, but also that the computations performed are actually the correct computations, allowing one to trust the results of the computation as well.

A hardware-based TEE uses hardware-backed techniques to provide increased security guarantees for the execution of code and protection of data within that environment. This assurance is often missing in approaches that do not use a hardware-based TEE.

The Goal of Confidential Computing

The goal of Confidential Computing aims to reduce the ability for the system administrator of a platform to access data and code inside TEEs sufficiently such that this path is neither an economically or logically viable attack during execution.

Use of the phrase "economically or logically viable" makes assumptions, of course, about the types of attackers considered. There are some attackers where such considerations may be weighted differently from others, including nation state actors and some academic institutions. There is an associated recognition that there is no absolute security, but TEEs can raise the bar significantly over other techniques available for protecting data in use by various measures beyond confidentiality and

integrity protection, including usability and cost. This improvement allows designers, implementers, and operators of systems managing sensitive data and algorithms to concentrate on other aspects of the system.

The Scope of Confidential Computing

The following threat vectors are considered to be in-scope for Confidential Computing:

Software attacks. Software attacks include attacks on the operating system, hypervisor, BIOS, other software and stacks.

Protocol attacks. Protocol attacks include side attacks on protocols associated with attestation as well as workload and data transport.

Cryptographic attacks. Cryptography is an evolving discipline, with vulnerabilities being found over time in ciphers and algorithms, including mathematical breakthroughs, availability of computing power, and new computing approaches such as quantum computing. In some cases, defense-in-depth may be appropriate, for instance employing quantum- resistant cryptography within TEE instances whose implementation is not itself quantum-resistant.

Basic physical attacks. While long-term intrusive attacks on the CPU are considered out-of-scope, other attacks are considered in-scope, including cold DRAM extraction, bus and cache monitoring and plugging of attack devices into an existing port, e.g., PCIe, Firewire, USB-C.

The Confidential Computing Consortium also believes that there exist opportunities to provide guidance to those designing, implementing and operating workloads around which types of applications which may be more vulnerable to attacks than others, as well as issues around lifecycle management to help mitigate attacks.

What Can Be Considered Out of Scope for Confidential Computing

Threat vectors which are generally considered to be out-of-scope for Confidential Computing include:

Sophisticated physical attacks. Sophisticated physical attacks that typically require long-term and/or invasive access to hardware, including chip scraping techniques and electron microscope probes.

Upstream hardware supply-chain attacks. These exclude attacks on components of a host system that is not directly providing TEE capabilities, but does include attacks on, for instance, a CPU. Examples include attacks at chip manufacturing time and attacks at key injection/generation time.

What is Fully Homomorphic Encryption (FHE)?

Fully Homomorphic Encryption (FHE) is a class of encryption methods envisioned by Rivest, Adleman, and Dertouzos in the late 1970s, and first constructed by Craig Gentry a former IBMer in 2009.

Prior to Gentry's breakthrough work at IBM in 2009, cryptographic schemes that allowed processing on encrypted data were limited to partial homomorphic schemes. These approaches meant that when two parties want to transmit data securely, the following process occurred:

- The sender would encrypt the data with a public key
- The recipient would decrypt the data using the key
- The recipient would perform a computation on the data, re-encrypt the data and send it back to the original sender

This approach means that data is scrambled for transmission and, therefore, if a third party intercepts the data, it could not be stolen in its clear form. However, utilizing this approach, senders must trust recipients with their data, as recipients have to decrypt the data to perform the computation.

Homomorphic encryption differs from typical encryption in that it allows computational tasks to be performed directly on encrypted data without requiring access to a secret key. The result of such a computation remains in encrypted form and can at a later point be revealed by the owner of the secret key.

How Can FHE Be Applied?

The prevalence of cloud computing and storage have fundamentally changed how businesses and individuals use and manage data at scale. As cloud has grown, so has the deployment of encryption methods, such as Advanced Encryption Standard (AES), as they are fast, and allow data to be stored conveniently in encrypted form. However, to perform even simple analytics on the encrypted data, the cloud compute layer must have access to the secret key. This access is the problem, from a potential security standpoint, anyway.

Access naturally leads to security concerns. Absent granting access, the owner of the data must download, decrypt, and operate on the data locally, which can create operational challenges.

That's where FHE is particularly valuable. FHE can be deployed and quickly simplify the operational constraints, as FHE methods enable the cloud compute layer to directly operate on the encrypted data. As a result of a FHE deployment, only the encrypted result is returned to the owner of the data.

Benefits of an FHE Deployment

The benefits of FHE have not yet been fully realized in commercial deployments at scale as the technology is still in early stages of commercial roll-out. However, the benefits fall into three main domains:

- **Radical improvements in privacy.** With an FHE deployment, data can be processed by third parties without divulging the data itself or any insights from processing the data as intermediate and final results are also kept encrypted.

- **Stricter Regulatory compliance.** As penalties for data breaches increase, FHE can help organizations process workloads involving encrypted data without ever exposing unencrypted and sensitive information.

- **Enhanced Cloud Security.** FHE provides a means to keep data encrypted in a third party or untrusted domain, such as a public cloud deployment. At the same time, FHE allows for use and computation on that data.

Despite Gentry having left the company, IBM continues to be a leader in the FHE space, however the technology is still nascent and practical applications are still not mainstream. IBM has made noises about this technology coming to its mainframe line of mission critical servers and this would make sense, given the company's role in corporate datacenters as a 'System of Record.'

Part 3. The Vendor Landscape

Part 1 of this paper covered the commercial landscape and drivers behind why security is dominating the collective consciousness. Part 2 of this paper then went on to unpack the fundamentals of Confidential Computing, its technical merits and foundational benefits, including the role of Trusted Execution Environments and Fully Homomorphic Encryption. In part 3 we will cover what the major vendors are doing to address this space. We will split the analysis in this section into hardware-based approaches at the silicon layer and then cover how these manifest themselves in public cloud deployment approaches.

Processor Vendors

Intel

Intel's work in Confidential Computing has centered around its Software Guard Extension (SGX) technology. Intel SGX improves the security posture by adding another layer of defense by assisting through the reduction of the attack surface. Intel SGX enables the protection of code and data from attack by malicious software and privileged escalations while data is being processed. SGX enables the development community to create Trusted Execution Environments (TEEs) at a silicon level directly within the processor/memory domain.

SGX is the latest iteration of trusted computing designs, that aims to solve the secure remote computation problem by leveraging trusted hardware in the remote computer. The trusted hardware establishes a secure container, and the remote computation uploads the desired computation and data into the secure container. The trusted hardware protects the data's confidentiality and integrity while the computation is being performed on it.

SGX creates isolated environments at the memory level called enclaves. SGX uses strong encryption and hardware-level isolation to ensure the confidentiality of data and code and to prevent them from being tampered with. Intel designed SGX to protect apps and code even when the operating system, hypervisor, or BIOS firmware is compromised.

While Intel is making progress with SGX, it is not without challenges. ArsTechnica reported in late 2020 that researches had disclosed new flaws in the approach. In addition, given the size of the enclaves created, developers face challenges around deployment at scale. Further, given these size constraints, organizations are faced with the challenge of what to put in the enclave spaces. This choice does not permit a pervasive approach to security and therefore leaves other workloads or data unsecured. Intel SGX can be considered a useful tool for very specific workloads, such as securing encryption keys, reviewing SSL and TLS connections, and signing certificate requests for a certificate authority. However, the technology comes with a significant choice and set of challenges associated with a need to redesign applications for Intel SGX enclaves, and the underlying performance impact is considered significant in certain deployment scenarios.

AMD

Introduced in 2019 on AMD's second-generation Extreme Performance Yield Computing (EPYC) processors, SEV-ES is part of AMD Infinity Guard, which uses an additional security processor located on the processor to deliver what the chipmaker refers to as "hardware root of trust."

This security process made its first appearance on AMD's original EPYC processors but had the limitation of 15 encryption keys. EPYC 2 extends that to more than 500 encryption keys.

AMD's Secure Encrypted Virtualization (SEV) is a technology designed to protect virtual machines by transparently encrypting the memory of each VM with a unique key. SEV can also calculate a signature of the memory contents, which can be sent to the VM's owner as an attestation that the memory was encrypted correctly by the firmware. SEV is especially relevant to cloud computing deployments. In this approach, VMs are hosted on remote servers which are not under the control of the VMs' owners, since it can reduce the amount of trust VMs need to place in the hypervisor and administrator of their host system.

AMD SEV is a hardware accelerated memory encryption for data-in-use protection. The approach takes advantage of new security components available in AMD EPYC processors. These AMD processors have two fundamental characteristics:

- **AES-128 encryption engine embedded in the memory controller.** This encryptions engine encrypts and decrypts data in main memory when the appropriate key is provided.

- **AMD Secure Processor.** Provides cryptographic functionality for secure key generation and key management.

In a virtualized environment such as VMware, it is critical to have protection of data not only from other virtual machines, but from the hypervisor itself. The additional layer of security means that customers can now encrypt data throughout their environment without needing to make changes to their applications.

AMD's SEV-ES also has advantages beyond ease of implementation. The technology provides options for implementation and can be enabled for certain workloads, where it is left disabled for others, and these workloads can also coexist operationally. For clients, this means that flexibility is an option that should not be overlooked and further, there is the convenience that deploying this technology can be done at a client's own pace.

Key components of the AMD approach include:

AMD Secure Memory Encryption (SME). Uses a single key to encrypt system memory. The key is generated by the AMD Secure Processor at boot. SME requires enablement in the system BIOS or operating system. When enabled in the BIOS, memory encryption is transparent.

AMD Secure Encrypted Virtualization (SEV). Uses one key per virtual machine to isolate guests and the hypervisor from one another. The keys are managed by the AMD Secure Processor. SEV requires enablement in the guest operating system and hypervisor. The guest changes allow the

VM to indicate which pages in memory should be encrypted. The hypervisor changes use hardware virtualization instructions and communication with the AMD Secure processor to manage the appropriate keys in the memory controller.

AMD Secure Encrypted Virtualization-Encrypted State (SEV-ES). Encrypts all CPU register contents when a VM ceases running. The leakage of information in CPU registers to components like the hypervisor is prevented, but further detection of malicious modifications to a CPU register state is also possible.

IBM

IBM was early to the deployment of Trusted Execution Environments and secure enclaves with its deployment of Secure Service Container (although the use of the word container is misleading in this context) as part of its Z and LinuxONE systems. IBM announced its initial foray into Confidential Computing capabilities in March 2018 at Think, its annual user event, with the launch of Hyper Protect Services which were based on z14 generation mainframe servers.

In September 2019, IBM announced the z15 next generation mainframe and LinuxONE III that are Linux only systems with the capability to deliver up to 16TB of secured memory that can support Confidential Computing workloads. Additionally, IBM's Pervasive Encryption feature set supports the processing of encrypted data in memory while having negligible impact on system performance due to execution at the silicon level.

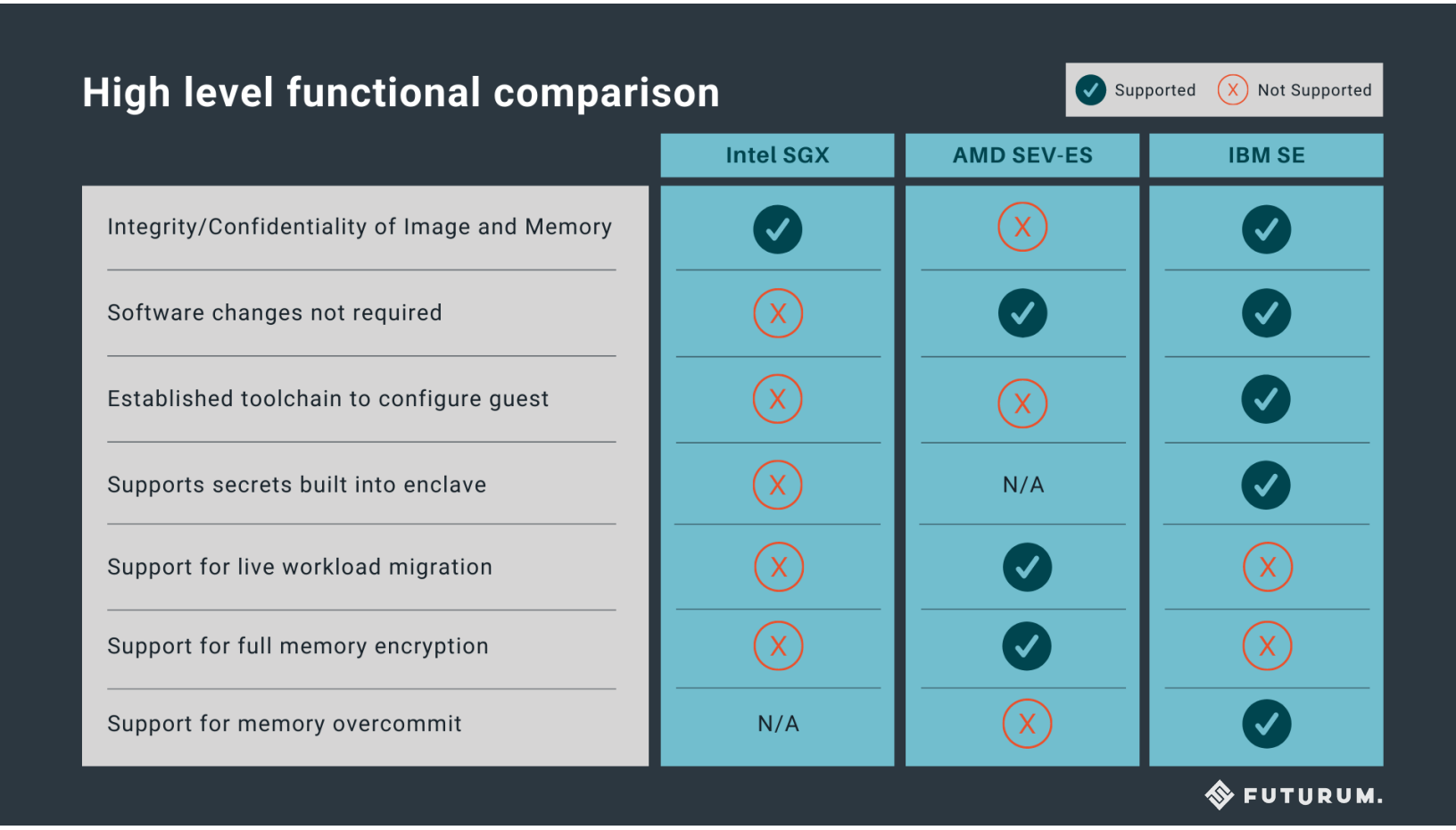
Announced in April 2020, IBM Secure Execution for Linux enables clients to isolate large deployments of workloads with granularity and at scale. The Secure Execution capability help to deliver protection for workloads from internal and external threats and protect from the threat vector of the system or VM administrator with elevated access credentials.

In June 2020, IBM announced new toolkits that allow MacOS and iOS developers to experiment with Fully Homomorphic Encryption (FHE) to keep data protected and processed simultaneously. In the summer of 2020, IBM also debuted a new FHE toolkit for Linux, bringing FHE to multiple Linux distributions for IBM Z and x86 architectures.

IBM also has a FIPS 140-2 Level 4 Hardware Security Module within its system that provides the highest level of certification for tamperproof security of encryption keys and for certain workloads and industries this is becoming a key selection criteria that should not be overlooked in vendor analysis or selection exercises.

Functional Comparison of TEE Approaches

In this section, we will cover a high-level functional comparison of the various TEEs in the market today from AMD, Intel and IBM. Every chip vendor has their own take on the technology as no standard lexicon for this space currently exists. The 3 technologies in market today include the main two vendors Intel and AMD however the IBM approach is worthy of consideration and should be evaluated in Futurum's opinion.



AMD SEV Challenges

AMD is on the 2nd generation of its technology and challenges still exist. One of the largest and most commonly seen weakness is that there are limited parameters or guardrails for overcommitting workloads. In a production deployment when clients overcommit on AMD more than one existing guest may disappear, meaning that every guest must fit in physical memory.

The current AMD SEV ES technology in market has multiple known vulnerabilities, including most worryingly no integrity protection. This vulnerability has a huge downside as lack of integrity protection can be exploited to compromise confidentiality of data. Another key weakness is the immaturity of AMD software toolchain, as little or no development Software Development Kit for programing within the TEE exists.

Intel SGX Challenges

Concerns are prevalent whether Intel approach to a TEE with SGX are enterprise ready. The very small enclave size of 128 MB is intended for entry level server solutions or encrypting keys securely or for personal computers and concerns are not without merit on the deployment models in enterprise scale applications.

Another drawback for SGX is that the approach requires major software and code changes, meaning that developers must rebuild applications to deploy fully and reap the benefits.

Another setback for Intel is that SGX does not support built-in secrets. Since SGX enclaves cannot have built-in secrets, an SGX enclave has to perform additional tasks at enclave bootup. From an operational perspective, this approach seems unnecessarily complex and can result in longer startup times than are experienced in other deployment alternatives

IBM Challenges

IBM's approach does not support full memory encryption, however, IBM would likely argue that this is not required since protection is provided by allowing access only to the KVM guest. However, we see the main challenge of the IBM approach is the Secure Execution technology being largely based on the s390x chip architecture and the underlying dependencies on an Ubuntu stack. We'll note that RHEL and SUSE also support Secure Execution, and also that a majority of apps are Java or Node.js based, so this may not be a huge issue, but we felt it worth mentioning. That said, enclaves are not about data encryption, but rather data protection. IBM's Z enclaves appear to provide that focus.

While the IBM approach has technical merits and is probably the most complete deployment of a TEE and Enclaves, the needs for the client to internalize the benefits of the big-endian nature of IBM's Z processor and its limited software support may mean that clients only evaluate this approach for their most secure workloads where these barriers are outweighed by the completeness of the technical deployment IBM has achieved. Based on our analysis of IBM's social media and press release activity, it appears as though one such area where IBM is having initial market traction is in the burgeoning Cryptocurrency and digital asset custodian space.

Public Cloud Vendors

Microsoft Azure Attestation Service

Microsoft's Azure Attestation service is part of Microsoft's Confidential Computing efforts, which also includes virtual machines delivered on Intel SGX chip architectures.

Based on review of available information, the still relatively new attestation service claims to provide technical protections and assurances on the security of using of cloud-based services. The focus of the Azure Attestation service is the security of operations that get processed in memory on virtual machines. Microsoft claims security assurances on the processing of data to be the last piece in the cloud security puzzle in that services provided on Azure already have protections for data when they are in transit and at rest.

Microsoft's cloud services leverage TEE components, or "enclaves," which can be either hardware- or software-based. For the hardware-based approach, Microsoft takes advantage of Intel's SGX, while Microsoft's Hyper-V software-based TEE solution is called Virtualization-Based Security. Microsoft asserts that the Azure Attestation service will provide "a unified solution for attesting" which will get used across multiple Azure services, including "Confidential Containers, Confidential VMs, IOT edge devices and more."

Microsoft is encouraging clients to use confidential containers, write enclave-aware applications with the Open Enclave SDK, utilize a third-party solution to run workloads, or deploy the latest virtual machine from Azure with an approach based on underlying Intel SGX-enabled hardware.

Another key component of Microsoft's approach to Confidential Computing requirements is Azure Kubernetes Service (AKS), which supports adding DCsv2 Confidential Computing nodes again powered by Intel SGX. These nodes allow clients to run sensitive workloads within a hardware-based TEE. Intel's SGX execution model also removes the intermediate layers of Guest OS, Host OS, and Hypervisor, thus reducing the attack surface area. The hardware based per container isolated execution model in a node allows applications to directly execute with the CPU, while keeping the special block of memory encrypted per container.

AKS Confidential Nodes Features include:

- Hardware-based and process level container isolation leveraging the features of Intel SGX
- Heterogenous node pool clusters (a mix of confidential and non-confidential node pools)
- Encrypted Page Cache (EPC) memory-based pod scheduling (requires add-on)
- Intel SGX DataCenter Attestation Primitives (DCAP) driver pre-installed
- CPU consumption based horizontal pod autoscaling and cluster autoscaling
- Linux Containers support through Ubuntu 18.04 Gen 2 VM worker nodes

The use cases Microsoft is listing for its Confidential Computing services include: Prevention of fraud and waste, anti-corruption, anti-terrorism, records and evidence management, intelligence analysis, global weapons systems and logistics management, vulnerable population protection (including child exploitation, human trafficking, etc.), anti-money laundering, digital currencies, blockchain, transaction processing, customer analytics, proprietary analytics/algorithm, disease diagnostics, drug development, and contact tracing.

IBM Cloud

According to IBM, its Cloud Data Shield enables users to run containerized applications in a secure enclave on an IBM Cloud Kubernetes Service host, providing data-in-use protection. IBM Cloud Data Shield supports user-level code to allocate private regions of memory, called enclaves, that are protected from processes running at higher privilege levels. It extends Intel Software Guard Extensions (SGX) language support from C and C++ to Python and Javac, while also providing pre-converted SGX applications for MySQL, NGINX and Vault. Powered by the Fortanix Runtime Encryption platform and Intel SGX technology, these tools enable organizations with sensitive data to leverage cloud computing with more confidence.

IBM's Hyper Protect Services is another component of IBM Cloud and the combined value of both Hyper Protect and IBM Data Shield highlights how IBM Cloud is different, and perhaps more mature in its approach to Confidential Computing by providing a diversified set of enclave technologies that are designed to meet developers where they are and provide an integrated set of confidential computing cloud services, spanning everything from micro-service and container-based apps, to virtualized apps, to scalable data and key management services.

According to IBM, its Cloud Data Shield / Hyper Protect Services have been designed to allow customers to have complete authority over their sensitive data, workloads, and encryption keys, and not even IBM Cloud admins have access.

The range of services within the IBM Cloud portfolio include:

- Database-aaS – with PostgreSQL and MongoDB database flavors available
- Virtual Machines
- Crypto – built on the most secure FIPS 140-2 Level 4 certified Hardware Security Module

These services are obviously a result of continued IBM investment into Confidential Computing technologies over many decades, especially given IBM's early entry into the space courtesy of Gentry et al. They are built on IBM LinuxONE secure enclaves, which offer inbuilt protection for data at rest and in flight plus protection of data in use. The LinuxONE chip architecture developed in-house by IBM is now on its 4th generation of secure enclave technology and is also leveraging Fully Homomorphic Encryption (FHE), both of which have been foundational to IBM's range of mission critical servers. According to IBM's marketing materials, the Hyper Protect services make it easy for application developers to build applications that deal with highly sensitive data and help customers meet regulatory compliance requirements.

IBM Cloud Hyper Protect Services are based on secure enclave technology that integrates hardware and software and leverages what the company calls "the industry's first and only FIPS 140-2 Level 4 certified cloud hardware security module (HSM)." The IBM portfolio currently includes three services: IBM Cloud Hyper Protect Crypto Services, Hyper Protect DBaaS and Hyper Protect Virtual Servers. These provide customers complete authority over sensitive data; associated workloads and the cloud encryption keys.

Since that initial release, IBM Cloud has continued to discuss the critical importance of securing customers' sensitive data and workloads and has added new features to Hyper Protect Services. These include advances that meet key compliance requirements for GDPR, ISO 27K, HIPAA Ready, IRAP Protected and SOC 2 Type 1 reports. Those are critical capabilities for global enterprises and companies working in compliance-focused industries.

Currently, IBM's production-ready Confidential Computing solutions are being used by customers, including Daimler and Bank of America. The company also brought this same technology to Apple CareKit via the IBM Hyper Protect Software Development Kit (SDK) for iOS available in the Apple CareKit open-source GitHub community

These services are gaining traction, as they form the basis of IBM's much touted Financial Services Cloud, which is an industry-focused approach to public cloud deployment. While this approach is relatively new, we see merit for highly regulated sectors such as Financial Services and Healthcare.

AWS Nitro

Back in October 2020, AWS announced the general availability of its Nitro Enclaves, "making it easy for customers to create isolated compute environments within Amazon Elastic Compute Cloud (Amazon EC2) instances to further protect their highly sensitive workloads."

According to Amazon, Nitro Enclaves will help customers reduce attack surfaces for their applications by providing a highly isolated and hardened environment for data processing. These EC2-based services provide the capability to make it easier for customers to securely process highly sensitive data and protect it when it must be unencrypted at the point of use by providing an isolated environment for data processing.

AW's move here was making Nitro Enclaves generally available as a new capability of EC2 that consists of each Enclave being a virtual machine with no persistent storage, no administrator or operator access, and no external networking. AWS Nitro Enclaves are an isolated environment running beside the EC2 instance, and this isolation means that applications running in an Enclave remain inaccessible to other users and systems, even to users within the customer's organization. It uses the CPU and memory resources from a customer's EC2 instance, but it is isolated from the instance on the hypervisor level, so that an individual instance cannot access the enclave, even by cloud admins on the OS-level.

An AWS Nitro Enclave owner can start and stop, or assign resources to an Enclave, but even the owner cannot see what's being processed inside of AWS Nitro Enclaves. Users can develop enclave applications using the AWS Nitro Enclaves software development kit set of open-source libraries with AWS Nitro Enclaves SDK integrating with AWS Key Management Service, allowing customers to generate data keys and to decrypt them inside the enclave.

"Customers often tell us that powerful built-in protections like the locked-down security model of the Nitro System are among the primary reasons why they trust AWS with their workloads," David Brown, vice president for Amazon EC2, said in a statement. "Nitro Enclaves builds on those same security and isolation models that have separated AWS for so many customers, delivering a more efficient method for securely processing highly sensitive data. This means customers can build and innovate faster in a way that still meets the highest bar for security."

In addition to the general availability of AWS Nitro Enclaves, AWS also announced the launch of AWS Certificate Manager for Nitro Enclaves, a new Enclave application that makes it easy for customers to protect and manage Secure Sockets Layer/Transport Layer Security certificates for their web servers running on Amazon EC2.

AWS customers in industries as diverse as Financial Services, Defense, Media & Entertainment, and Life Sciences routinely process highly sensitive data are the target clients for these new services.

Google Cloud – Confidential VMs

Google Cloud is trying to convince their clients to move their most sensitive data to the cloud with its new Confidential Virtual Machines product, which relies on a silicon-level security feature within AMD's second-generation EPYC processors. According to available information from Google in mid-2020, confidential VMs makes Google Cloud "the first major cloud provider to offer this level of security and isolation while giving customers a simple, easy-to-use option for newly built as well as 'lift and shift' applications." Considering what else is already happening in this still nascent market, this sounds a bit like marketing bluster to us.

Google Cloud's focus is ease of use, low performance impact, and scalability and the company clearly sees these as paramount to further adoption of Confidential VMs. As a result of this focus, the cloud provider decided to go with AMD SEV over Intel SGX. According to Google, its customers don't have to worry about redesigning or tweaking any of their applications to move them to Confidential VMs.

"The beauty of this idea: the customers don't need to change anything," Nelly Porter, a lead product manager at Google Cloud said. "Every application that they ran before in normal VMs will continue running in Confidential VMs."

Confidential VMs are based on Google Cloud's N2D series instances that run on AMD's second-generation EPYC processors, all of which come with an expanded version of SEV that supports 509 encryption keys that are generated by the processors' Arm-based secure co-processor. With the co-processor's key manager generating the VM encryption keys, neither Google Cloud nor any VMs running on the hypervisor can access them, according to Porter, which is crucial to gaining the trust of organizations that want to move confidential and sensitive data to the cloud.

"This means nobody, not AMD, neither Google have access to those keys," she said.

Tellingly, Porter also stated that Google Cloud didn't go with Intel SGX due to availability of processors supporting the feature, as well as the complexity associated with adapting applications for the technology.

Porter went onto say "From our perspective, for the workloads we're trying to enable, for use of use that we're looking at and performance penalty,' to tell customers [they] have to pay based on those three things, Intel SGX was an interesting idea, and we continue to look and work with Intel on that, but it's not yet applicable to the workloads and scale that we're looking at."

Conclusion

As mentioned at the onset of this research brief, Confidential Computing is still a nascent and emerging technology and has yet to gain widespread adoption, even by hypersclalers such as AWS and Azure.

We see this space gaining more traction in the coming months, especially within heavily regulated industries, such as Banking and Healthcare. We also see niche deployment opportunities in areas such as Fintech, payments rail providers and Cryptocurrency organizations who can provide differentiation for their services based on enhanced trust in their operational models.

As clients look to deploy Confidential Computing technologies, our advice is to go wider than the mainstream choices and conduct a thorough evaluation of all the technical options as the space is still emerging and the winners and losers are yet to be established.

Close attention should be paid to all approaches, however niche, as these may provide the ultimate trust that your use case requires. We recommend casting a broad scope in your evaluation criteria and looking beyond the marketing speak of vendors in the space, and digging deeply into use cases, customer success instances, and thinking as much about future use cases as about current use cases. A technology decision that serves you and your organization well into the future is worth doing the work to get to.

As this domain develops in the coming years, we see standards emerging and a convergence of approaches as open-source collaborative projects such as the Confidential Computing Consortium from the Linux Foundation take hold. These moves toward standards and commonality of deployment architecture are to be applauded, as they will be necessary for widespread adoption within a hybrid multi-cloud architecture, especially if workload portability is a requirement.

Most importantly, we see Confidential Computing as the future of security in the enterprise, as we collectively make the necessary moves beyond traditional security measures that involve people and processes and embrace technology solutions that do what people cannot — ensuring the protection of data while in the state of being utilized or processed.

Important Information About This Research Brief

CONTRIBUTORS:
Daniel Newman
Founding Partner + Principal Analyst
Futurum Research

Shelly Kramer
Founding Partner + Lead Analyst
Futurum Research

PUBLISHERS:
Daniel Newman
Founding Partner + Principal Analyst
Futurum Research

Shelly Kramer
Founding Partner + Lead Analyst
Futurum Research

INQUIRIES:
Contact us if you would like to discuss this report and Futurum Research will respond promptly.

CITATIONS:
This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Futurum Research." Non-press and non-analysts must receive prior written permission by Futurum Research for any citations.

DISCLOSURES:
Futurum Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT FUTURUM RESEARCH
Futurum is an independent research, analysis, andadvisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets. Read our disclaimer statement here.

CONTACT INFORMATION

Futurum Research, LLC | futurumresearch.com | 817-480-3038 | info@futurumresearch.com
Twitter: @FuturumResearch