PRIVATE CLOUD COMPUTING – THE DSTA JOURNEY

TOH Thiam Huat Adrian, LIM Victor

ABSTRACT

Cloud computing provides a business model for the delivery of IT services with greater agility and efficiency as compared to traditional IT. Cloud technologies such as server virtualisation and automation are widely implemented by organisations to enable new business capabilities and optimise data centre and engineering resources. This article delves into concepts of cloud computing and describes DSTA's journey into private cloud computing, including some of the key lessons learned.

Keywords: cloud, converged infrastructure, IaaS, virtualisation, fabric

INTRODUCTION

The rapid pace of application development and proliferation of IT services places a significant pressure on existing infrastructure and engineering resources. The optimisation of data centre resources is also essential to support the business needs of DSTA.

As such, DSTA adopted cloud computing in 2012 to improve the agility and efficiency of its IT services. Its cloud computing journey had centered around the optimisation of infrastructure and engineering expertise with proven technologies and security mitigations. Automation of IT workloads was achieved through the provision of virtualisation technology to consolidate computing resources and the implementation of compliance (desired configuration) management to improve IT management efficiency.

CURRENT LANDSCAPE

The National Institute of Standards and Technology (2011) defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

Figure 1 gives a simplified view of the different types of cloud computing environment. Public cloud providers are typically in control of how customer data is managed. It is however the customers' responsibility to manage data protection and IT security. Data can be replicated to other countries for better resiliency and cost efficiency, but this method raises the issue of data sovereignty. The private cloud model was thus adopted by DSTA as it allow full autonomy of the use of IT resources to safeguard its core business processes.

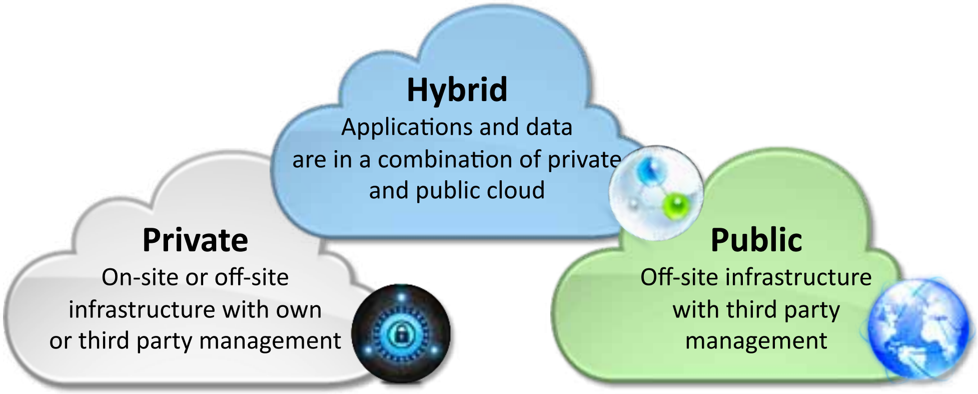
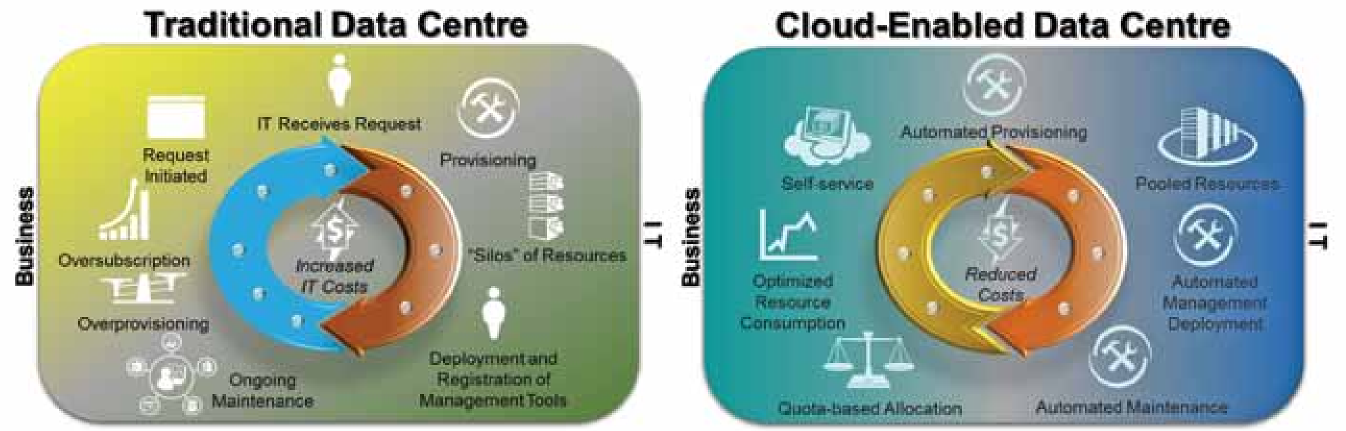

Hybrid
Applications and data are in a combination of private and public cloud
Private
On-site or off-site infrastructure with own or third party management
Public
Off-site infrastructure with third party management

Figure 1. Different types of cloud computing environment

The diversity of IT workloads in a largely heterogenous cloud computing landscape dictates the use of different hardware platforms and Operating Systems (OS). This diversity creates challenges in terms of manageability and interoperability, such as the need for highly specialised but separate teams to support and manage IT resources. Using proprietary technologies also compounds the issue of infrastructure complexity and supportability.

Harnessing IT to maximise workforce productivity is therefore cardinal. A cloud-enabled data centre, through the standardisation of IT and data centre technologies and processes, will be able to achieve high system reliability, maintainability and supportability (see Figure 2).



(a)

(b)

Figure 2. Key elements of a (a) traditional and (b) cloud-enabled data centre

CLOUD INFRASTRUCTURE

The common cloud service models are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). IaaS was determined to be the most suitable cloud service model for DSTA's operating environment where all IT resources are hosted on site (see Figure 3).

A cloud infrastructure comprises the data centre, mechanical and electrical systems, as well as the corresponding IT resources such as network, storage and compute. A more efficient and effective infrastructure can be created through the consolidation of resources for optimal utilisation.

Fabric Networking

To achieve a unified high-performance computing system, the use of high speed 10G network with fabric network and high density blade computing greatly reduces the amount of network equipment needed. Additionally, it converges the local area network and storage array network data traffic over a common network infrastructure. This optimisation is essential to maximise data centre performance and utilisation.

Alternative Storage

Storage array network used to be the de-facto enterprise storage with proprietary hardware algorithm that makes it costly and complex to migrate. The JBOD[1] (Just a Bunch Of Discs) solution is becoming a good alternative due to its low cost and tier storage capability.

Processor Architecture

Technological advances have enabled chipset manufacturers to make significant progress in the development of applications to be deployed on previously unsuitable hardware platforms as well as enabling portability for applications between different hardware platforms.

Operating System

Specialised competency and effort are required to support an OS. Conscious effort is thus needed to streamline the number of OS and facilitate supportability.

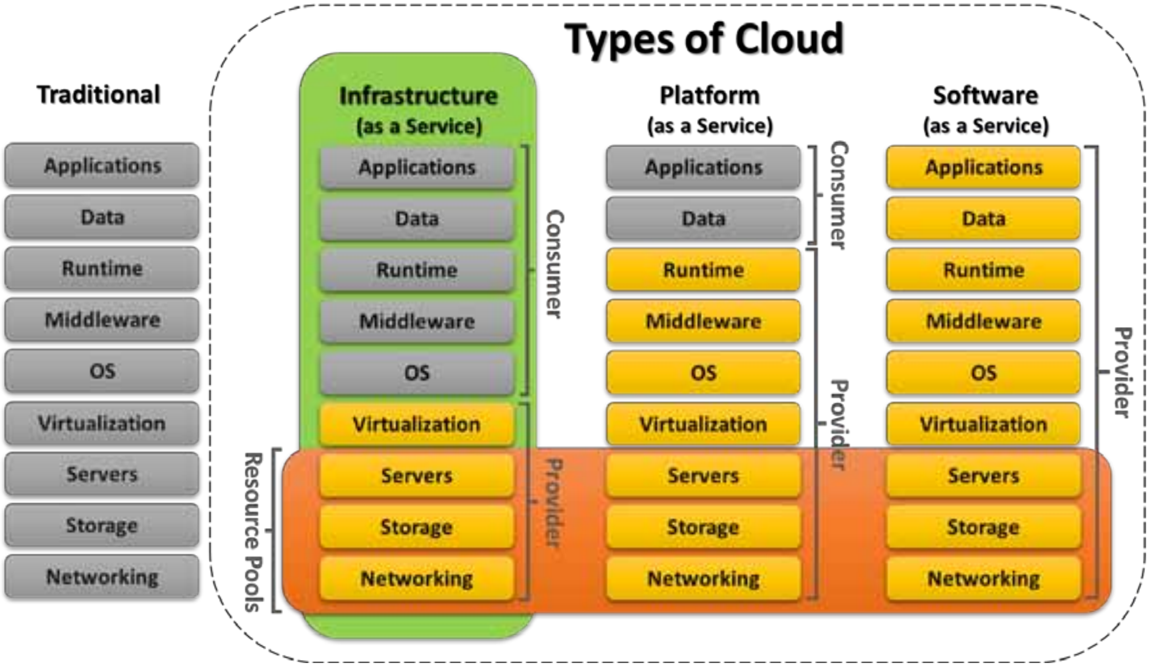


Figure 3. Individual components of cloud service models (Microsoft, 2013)

Consolidation

Consolidation of IT resources is essential for cloud adoption (see Figure 4). By virtualising physical workloads, obsolete resources and under-utilised physical systems are freed up. This addresses the inefficiency of server sprawl[2] and optimises IT resources to facilitate IaaS, PaaS and SaaS.

Virtualisation

This is the abstract layer between the physical entity and the virtual resources that forms the basis of cloud computing. Virtualisation allows workloads to be mobile and removes the dependency on the hardware layer. (See Figure 5).

Software Defined Infrastructure

Using software-centric management to control each of the IT resources creates a unified view of the interdependencies among IT resources and improves fault resolution time. While traditional IT operations require significant manual effort during configuration, this can be streamlined and automated using policy driven provisioning to standardise configuration. The Cisco road map as shown in Figure 6 outlines a shift of the fabric layer, from a traditional distributed architecture to an application driven infrastructure which provides performance and availability assurance for critical business applications.

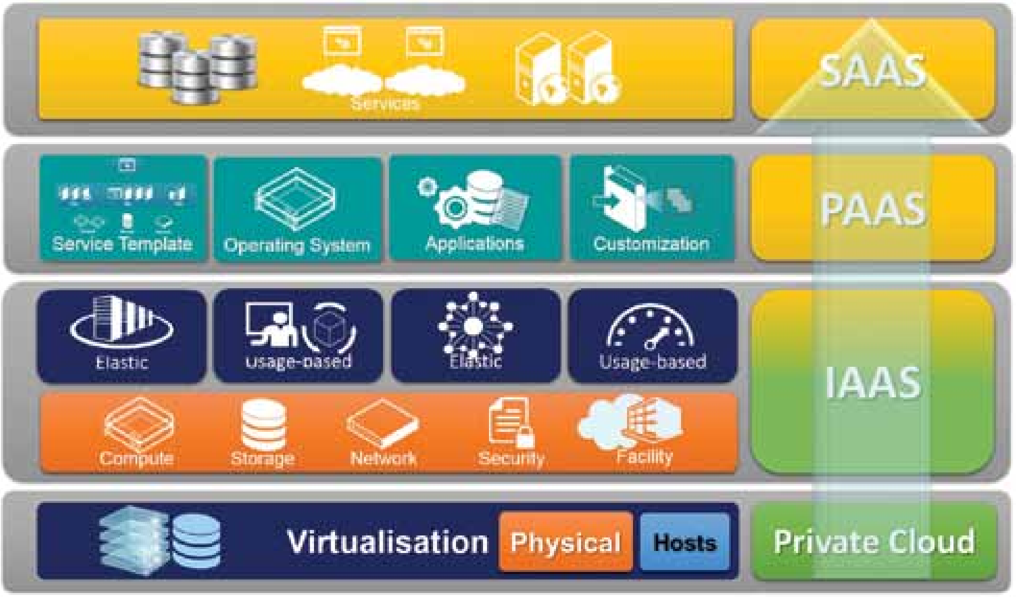


Figure 4. Ongoing cloud adoption

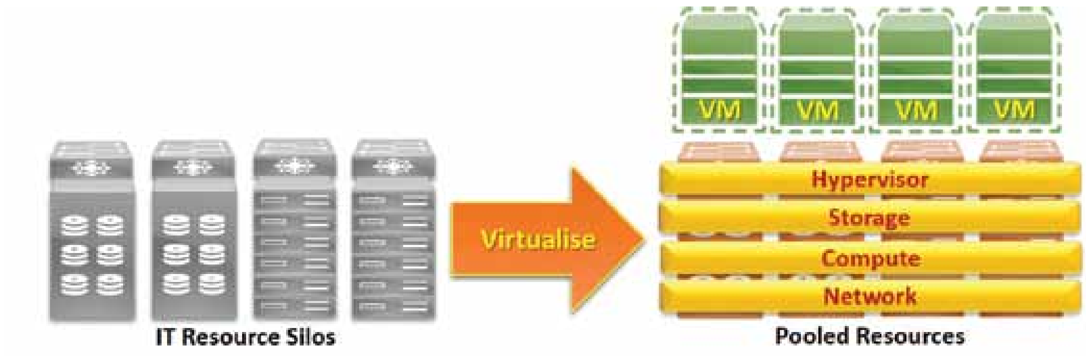


Figure 5. The process of virtualisation

KEY CONSIDERATIONS

The convergence of IT systems has its own set of challenges and risks. Creating strong IT interdependencies places the entire IT infrastructure at higher risk as it can be affected by even a single IT component failure. Moreover, a private cloud is not necessarily less susceptible to exploitation than that of a public cloud. Security controls are mandatory to protect the infrastructure and data. Below are some of the considerations and mitigation measures that should be taken into account in order to address security concerns and system performance.

Security Concerns

Data Classification

One key consideration is the risk of data loss. Different data classification mandates different security measures such as data encryption and rights of access. By putting data of different classification into the same cloud, security measures may end up being overly stringent when applied to less sensitive workloads. Conversely, inadequate security measures will risk the exposure of information in the event of a security breach.

Personnel with physical access to IT resources may access unauthorised classified data. To mitigate this risk, DSTA secured data on the hypervisors[3] using Trusted Platform Module encryption, while the software-based volume encryption, Advanced Encryption Standard-256, was used to secure data on storage resources.

Zonal Segregation

DSTA segregated its network using network zoning. Each zone represents one type of data profile. For example, web applications are categorised under Zone 1 and data warehouse workloads are categorised under Zone 2. Data of the same profile will communicate via the same fabric switches. Different physical fabric switches are used for handling different data profiles. All communication between these switches are conducted via secured gateways such as firewalls.

In a virtualised environment, it is essential to ensure infrastructure workloads and enterprise workloads are separated. Hence, a segregation of network and identity namespace was implemented by DSTA (see Figure 7). Physical infrastructure and management workloads supporting the fabric layer are allocated in the Infrastructure Zone (i.e. Zone 3 in Figure 7) and production workloads supporting corporate requirements are maintained in the Enterprise Zone (i.e. Zone 1 and 2 in Figure 7). Workloads in both zones are maintained in different namespace and distinct networks as if they are in separate environments, with minimal communication between these zones.

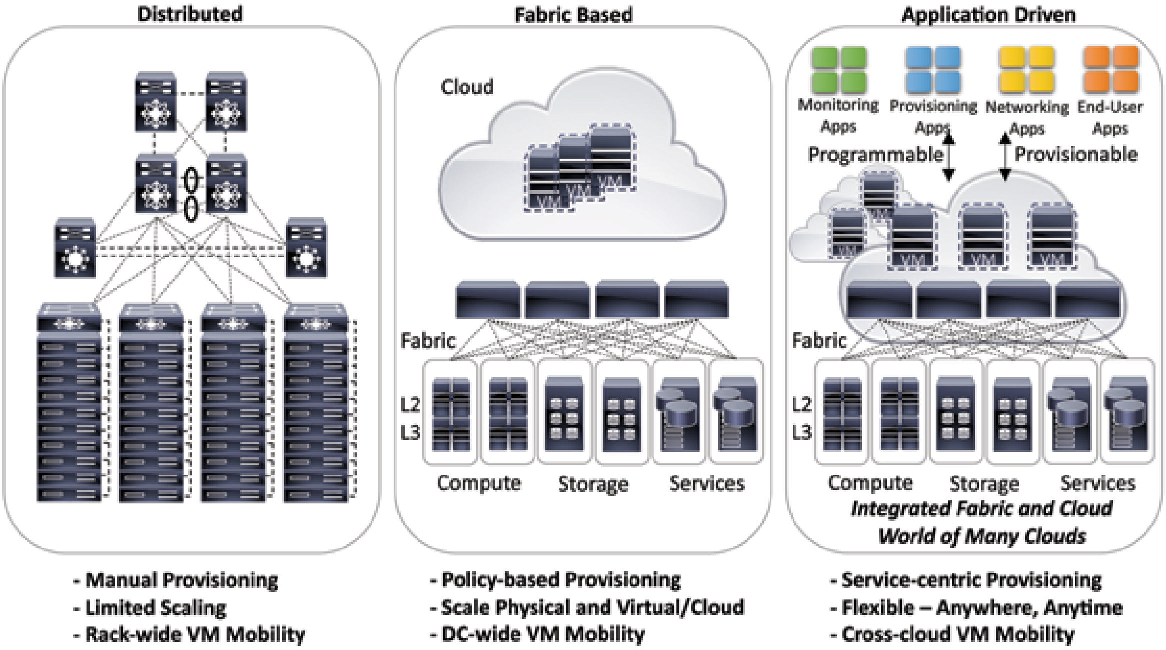


Figure 6. Road map on the fabric layer (Tucker, 2012)

Separate sets of identities are used to access the resources in the respective namespace. This means that personnel with access to the infrastructure workloads should not have access to the enterprise workloads, and vice versa. This separation limits access to classified data, hence strengthening overall security.

Hypervisors Protection

Hypervisors are susceptible to vulnerabilities like any other OS. Bare-metal hypervisors can be used to enhance its protection and unused roles and binaries removed to reduce the attack surface. Regular compliance scans on all hypervisors will ensure that the system undergoes up-to-date patching.

Secure Boot configuration is important in a highly virtualised environment as a compromise on the physical server can impact all the Virtual Machines (VM) on it. It essentially helps to validate the cryptographic signatures that guard against malware attacks on the BIOS before the OS is started up, effectively preventing unauthorised firmware, OS or drivers from being executed.

Imageless Provisioning

In general, service images are used to deploy new workloads. However, a constant effort is needed to maintain these images to mitigate security threats. This issue can be mitigated by integrating patches, hardware drivers and configuration files into the binaries to automate the set-up effectively.

Logging, Auditing and Monitoring

With evolving security threats, proactive monitoring and analytical mechanisms are vital to safeguard systems against anomalies. Key activities such as provisioning, remediating and orchestrating should be logged. These logs should be archived and protected from tampering.

Desired Configuration Management

To safeguard the configurations and integrity of the computing resource, it is essential to enforce change management governance with a centralised policy. Workload categorisation is also vital for effective policy management and remediation against unauthorised change.

System Performance

Virtualisation

There are different types of virtualisation platforms. Virtual workloads cannot move between platforms seamlessly without incurring migration efforts and downtimes. Similarly, different platforms entail the use of separate management software as well. Thus, it is crucial to standardise virtualisation efforts on a single platform to improve system agility.

In addition, some applications do not support virtualised workloads. Assessment is needed to determine the suitability of such applications in a cloud environment – those deemed unsuitable should not be virtualised to avoid downstream service disruption and supportability issues.

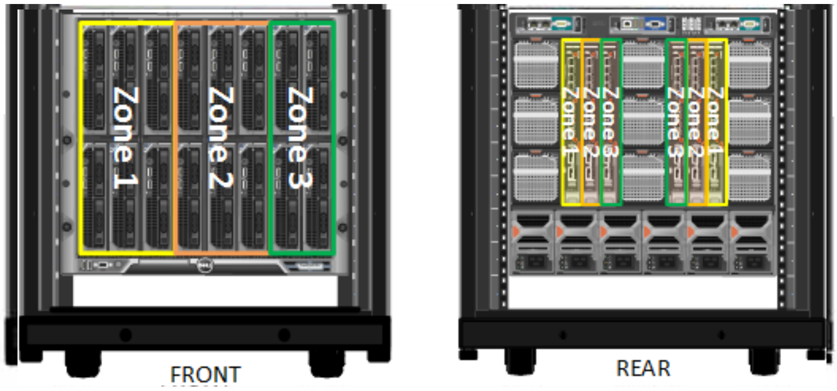


Figure 7. Schematic representation of blade computing zonal segregation

Resilient Workloads

Cloud computing enables virtualised workloads to move across hypervisors seamlessly to improve hardware resiliency. Clustering can be created to achieve high resilient workloads for mission critical services. By adopting an Anti-affinity configuration, the failure of a VM workload will not affect the operation of another that is located in a separate blade enclosure (see Figure 8).

Automated Hypervisor Maintenance

With cloud computing, automation is leveraged to reduce infrastructure management complexity and vulnerability, thus improving system availability. To achieve this, the hypervisor is placed on maintenance mode to shift its workloads to another hypervisor to facilitate the updating of patches and physical drivers without causing any downtime (see Figure 9).

Workload Profiles

Higher processor cores have more computing power to handle multiple workloads concurrently. Hence, it is vital to categorise these workloads based on their computing profiles to ensure reliability and good performance. For instance, a physical machine is able to host a maximum of 12 web application workloads but this is significantly lower for intensive data-warehousing workloads.

Storage Profiles

While virtualised workloads are typically integrated in a common physical storage layer, different workloads have different storage performance demands. Therefore, it is important to categorise these workloads to allocate the right storage profiles for optimal utilisation and performance.

Data that are accessed frequently such as an OS should be thick provisioned while those that are accessed less frequently – such as file server data – should be thin provisioned. Thin provisioning involves the increasing of disk capacity as it is being used while thick provisioning means to pre-allocate disk capacity at the point of workload creation.

Anti-virus software is essential in protecting systems but it also has considerable impact on storage resources if it is not configured properly. Resource-intensive activities such as system backup, signature updates and file scanning need to be coordinated carefully to avoid resource contention. It is thus important to choose an anti-virus solution that is compatible with virtualisation to mitigate against AV storms[4].

Quality of Service

Quality of Service is implemented in the cloud environment to predefine IT resources and ensure that workload performance and security are not compromised by each another. It regulates against volume based attacks such as the execution of distributed denial of service on the VM, hypervisor and fabric layers.

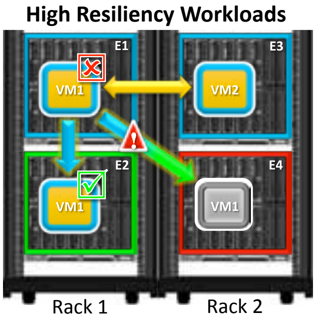


Figure 8. Anti-Affinity configuration for high resilient workloads

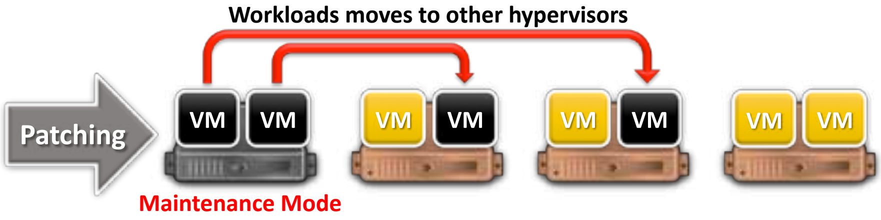


Figure 9. Automated hypervisor maintenance

While cloud computing offers elasticity and economy of scale, it is still susceptible to the same threats common to all computing environments. Such risks can be reduced through careful planning, identifying areas of concerns and mitigating them accordingly.

CONCLUSION

A robust and resilient IT infrastructure architecture is an element of most organisational strategies today. With cloud adoption, DSTA has further strengthened its IT management regime and business IT resilience as well as optimised data centre resources and agility to help business innovate and create new possibilities.

REFERENCES

Chdelay. (2013, July 24). *Getting started with Windows Azure: Part 2, what are cloud services?*. [Blog post]. Retrieved from http://blogs.technet.com/b/xdot509/archive/2013/07/24/getting-started-with-windows-azure-part-ii-what-are-cloud-services.aspx

The National Institute of Standards and Technology (NIST). (2015). *NIST cloud computing program*. Retrieved from http://www.nist.gov/itl/cloud/

Tucker, L. (2012). *The ever changing cloud, CloudExpo 2012*. [Powerpoint slides]. Retrieved from http://www.slideshare.net/lewtucker/the-ever-changing-cloud-cloudexpo-2012

ENDNOTES

[1] JBOD refers to a collection of hard disks that can either operate independently or as a single logical volume. JBOD is typically controlled through software-centric solutions instead of propriety hardware based solutions.

[2] Server sprawl refers to a situation in which applications are hosted on dedicated systems instead of co-existing with other workloads to maximise utilisation. With limited data centre resources, this inefficiency inhibits IT growth while creating unncessary maintenance costs. This situation is usually eliminated through workload consolidation or virtualisation, which inherently reduces the number of physical servers and hence the associated maintenance.

[3] A hypervisor is a program that allows the hosting of virtual machines on a single physical server hardware. Application workloads are subsequently hosted in each of these virtual machines.

[4] An AV storm refers to a situation where anti-virus scans or signature updates are occurring simultaneously on multiple virtual machines. These activities will usually create overheads on the computing resources. While the effects are not prominent on physical servers, this effect is accumulative on a hypervisor. This often leads to a spike in resources, which in turn leads to performance degradation on workloads.

BIOGRAPHY

TOH Thiam Huat Adrian is a Senior Principal Engineer (InfoComm Infrastructure). He leads and directs IT resiliency design, planning and implementation to maintain a secure, stable and robust IT infrastructure for the Ministry of Defence (MINDEF), the Singapore Armed Forces and DSTA. His Corporate IT Email Disaster Recovery project clinched the MINDEF Corporate IT Award in 2009. Adrian has obtained certifications in Project Management and IT Business Continuity Management from the Singapore Computer Society. He is also an EXIN certified Data Centre Specialist and is currently on the Board of Assessors for the Singapore Computer Society Certification in the IT Business Continuity Management Programme. Adrian graduated with a Bachelor of Science (Business IT) degree with Honours from the University of Central England, UK, in 1998.

LIM Victor is a Principal Engineer (InfoComm Infrastructure) who is currently designing cloud computing platforms. He led a team overseeing the development, implementation and maintenance of the Corporate IT Messaging Systems for MINDEF and the Enterprise Collaboration Private Cloud Platform for DSTA. He was also involved in the development, implementation and architecting of mobility related solutions for DSTA and MINDEF. Victor graduated with a Bachelor of Engineering (Mechanical Engineering) degree from the National University of Singapore in 2004.