Lee Brunovsky

CSCI 6910 Cloud Computing and Security

CH1 Introduction to Cloud Computing, CH2 Cloud Concepts & Technologies

HW1

Due: Jun 17, 2021

Submitted: Jun 13, 2021

Chapter 1:

Q4: Bahga & Madisetti described the principle of multi-tenancy as a means to facilitate sharing across multiple users via either higher level virtual machines (VMs) that allow applications to mount and run concurrent operations on the same computing and storage resources, or organically by utilizing more highly tailored applications specifically tuned to leverage an architecture of pooled bare metal assets such as load balancers and various types of servers, in addition to sharing an OS (2014).

Q5: In cloud infrastructure, horizontal scaling refers to the provisioning of more servers, where in vertical scaling the number of servers remains static, but the computing capacity of the hardware is increased. More specifically, Bahga & Madisetti elaborated that scaling up concerns improving the capabilities of resources such as memory, storage, or network resources on a single machine, and therefore is better suited for monolithic applications: Conversely, scaling out is more sensible for loosely coupled microservice architecture, where identical services and their dependent resources can be easily replicated or removed to more cost effectively handle dynamic load periods (2014).

Q6: Virtualization provides a means to concurrently run numerous OS instances via VMs running on disparate partitions that mount the same underlying physical memory, network, processor and storage host resources (Bahga & Madisetti, 2014). The major difference between full, para- and hardware-assisted virtualization architecture lies in whether the hypervisor is type-1 (native) or type-2 (hosted), and if the guest OS kernel requires modification. More specifically, Bahga & Madisetti expounded that the guest OS is unmodified in full virtualization, since requests are executed through a binary translation, where as the guest kernel is tailored to more efficiently pass instructions to the hypervisor by means of hypercalls in Para-virtualization architectures; furthermore, the authors explained that there is no need for the para-virtualization approach or binary translation with hardware virtualization due to the on-chip features of current major processor designs from AMD and Intel that support virtualization calls by automatically trapping to the hypervisor (2014).

Q7: If my current company Honeywell wanted to launch another e-commerce website, it would likely require broad network access, high performance via elasticity, a high level or reliability and multi-tenancy, which calls for enterprise grade IaaS service model architecture and a hybrid cloud approach where sensitive data and back-ups may be held privately on-site in edge systems with the bulk of the architecture existing in the public cloud. Additionally, Honeywell outsources to Salesforce already for CRM purposes. Accordingly, AWS EC2 and C2 instances would allow for Honeywell to leverage its very large existing staff of developers to upload their home grown microservice web applications that could easily be horizontally scaled depending on cyclical service peaks, which we often see during the US time zone from high net worth Business Aviation customers requiring urgent support and a high up-time SLA. These customers often use various devices such as iPads, smartphones, and PC's to load aircraft data or interact with our various services such as SATCOM provided internet. All of these devices have customer unique Honeywell user logins where we track MAC addresses for customized content delivery and access to aircraft or equipment loadable software options that can easily cost over $1M USD. Therefore, the dynamic scaling, disaster recovery, and management & monitoring tools that accompany IaaS would be a great fit, while AWS simultaneously extends these enterprise grade features to support both use cases that allow for up scaling when resource dependent applications are needed and to keep high up-time during dynamic loads (Bahga & Madisetti, 2014).

Chapter 2:

Q2: According to Bahga & Madisetti, full and para-virtualization differ in that the guest OS is decoupled in the former, requiring no modification and is facilitated via binary translation where as the OS is modified in the latter to improve efficiency via hypercall instructions (2014).

Q3: Load balancing maximizes throughput, availability and reliability while minimizing response time by intelligently distributing dynamic loads across multiple servers and application instances, which directly facilitates scalability and elasticity in cloud services (Bahga & Madisetti, 2014).

Q4: Sticky sessions are session management policy aimed at simplicity where all of a users requests are routed to the same specific server; however, this methodology is not redundant since there are no failover provisions inherently included (Bahga & Madisetti, 2014).

Q5: Bahga & Madisetti mentioned that traditional scaling methods don't adequately respond to volatile swings in demand since they are provisioned in advance based largely on forecasts comprised of fixed time intervals, which often result in over-provisioning on account of efforts to meet peak-demands and result in heavy initial infrastructure investment; conversely, on demand scaling provides the ability to tailor the cloud architecture configurations to more effectively pair with application demands (CPU, I/O, memory, network) via a continuous refinement of monitoring metrics throughout the deployment life cycle (2014).

References

Bahga, A., & Madisetti, V. (2014). *Cloud Computing: A Hands-On Approach*. Self published, Arshdeep Bahga & Vijay Madisetti.