Gujarati Dharmi

CWID - 20018001

CS 524-A Introduction to Cloud computing – Homework #2

1. **Complete reading Chapter 3 of the textbook and the lecture materials.** Please note the errata: The references to [19] on p. 56 of the book should be replaced with references to [20]! Please also read [20] (available free) at https://www.kernel.org/doc/ols/2007/ols2007v2-pages-87-96.pdf.

2. **Explain the advantage that para-virtualization provides for handling timers in virtual machines.**

➢ Advantage of Para-virtualization (handling timers):
- The para-virtualization is the best to obtain higher speed and efficiency than full virtualization.
- It enables guest operating system to connect with virtual machine monitor (VMM).
- With para-virtualization, a typical modification is to change the idle code to request the VMM to notify itself in a specified time period. Then, time is re-calculated and restored in the guest. The ability to handle timers in virtual machines more effectively is one of the benefits of para-virtualization.
- A virtualized timer provided by the hypervisor serves as the guest operating system's timekeeper in a conventional virtualization context. As a result, timekeeping may be done more precisely and costs can be reduced.

3. **Explain how para-virtualization helps in minimizing access to APIC.**

➢ The way that para-virtualization helps in minimizing access to APIC:
- SMP guest handling is another example. In x86 or x86-64, local Advanced Programmable Interrupt Controller (APIC) is required to support SMP especially because the operating systems need to send IPI (Inter-Processor Interrupt).
- The code needs to access the APIC registers a couple of times. Each access to the APIC registers needs to be intercepted for virtualization, causing over-head (often a transition to the VMM).
- A virtualized access to the underlying hardware resources is provided by the virtualization approach known as para-virtualization, allowing various operating systems to run on the same physical hardware.
- The guest operating system interacts with the hypervisor to access the physical resources when para-virtualization is used since it is aware that it is functioning in a virtualized environment.
- Para-virtualization has the ability to reduce access to the host system's Advanced Programmable Interrupt Controller (APIC), which is one of its benefits. If numerous virtual

machines are contacting the APIC at once, which is a hardware component that handles interrupts in the system, it may become a bottleneck.
- By minimizing access to the APIC, para-virtualization can improve system performance.
- In a para-virtualized setting, the guests operating systems are conscious of the hypervisor's presence and to manage speak with it directly.

4. Find out if Linux (like Unix) has both the user-mode and system-mode stacks for each process it runs.

- Yes, Linux has user-mode and system-mode stacks for every process it runs, much like Unix.
  - In Linux, a user-mode stack and a kernel-mode stack are each given to newly formed processes. While a process is running in user mode, data and instructions are stored in the user-mode stack; when it is running in kernel mode, data and instructions are stored in the kernel-mode stack.
  - A user process runs in the user mode and therefore its code cannot contain privileged instructions, but the operating system must execute these.
  - The system mode stack store the function call frames for the kernel-level routines that are performed to handle the system call or interrupt when the process makes a system call or receives an interrupt.
  - After discussing these points, we can say that Linux (like Unix) has both the user-mode and system-mode stacks for each process it runs.

5. Find out what "unscrambled" means in the description of the Intel LSL instruction (you can, for example, use the Intel manual referenced in the lecture).

- Intel LSL(Load Segment Limit) instructions:
  - LSL Loads the unscrambled segment limit from the segment descriptor specified with the second operand (source operand) into the first operand (destination operand) and sets the ZF flag in the EFLAGS register.
  - In the instructions, "***unscrambled***" is referred to as "Unscrambled Limit" as the limit scaled according to the setting of the G flag in the segment descriptor.
  - It is contrast of the LSL Scrambling which is a technique that was used in early x86 processors to compress the limit value of a segment register from 20 bits to 16 bits in order to save space.
  - The LSL Scrambled instruction was used to load a scrambled limit value into a segment register, and the processor would unscramble the value as it was loaded. However, this technique is no longer used in modern x86 processors, and the LSL instruction always loads an uncompressed limit value into the segment register.
  - Hence, "unscrambled" merely refers to the LSL instruction's loading of an uncompressed limit value into a segment register without further scrambling or unscrambling of the value.

6. **Read the following two papers:**
   a. **Carl Waldspurger and Rosenblum, M. (2012) I/O Virtualization. Communications of the ACM, vol. 55, No 1. January 2012. Pages 66-72; and**
   b. **Muli Ben-Yehuda; Xenidis, J.; Ostrowski, M.; Rister, K.; Bruemmer, A.; Van Doorn, L. (2007). The Price of Safety: Evaluating IOMMU Performance. Proceedings of the Linux Symposium on June 27th–30th, 2007. Ottawa, Ontario. Pages 225-230.**
   - **Explain the advantages and disadvantages of using I/O MMU by citing the appropriate text from the paper.**

- I/O Memory Management Unit (IOMMU) Advantages from following two papers:
  - **Memory access** - It translates I/O-virtual memory addresses to corresponding physical memory addresses, making direct memory access by devices safe and efficient.
  - **Flexibility in mapping** - The IOMMU mapping needs to be updated as well so that the IO to machine mapping will again correspond exactly to the pseudo-physical to machine mapping.
  - **Allow multiplexing** - It enables time and space-multiplexing of I/O devices, allowing multiple logical devices to be implemented by a smaller number of physical devices.
  - **Secure Data** - It represents a logically isolated private network, where the isolation is provided using cryptographic methods to secure data and provide some measure of performance isolation or quality-of-service controls to reflect the relative importance or absolute requirements of diverse VM workloads.
  - **Pre-allocation of resources** - In a hypervisor scenario, pre-allocation is only viable if the set of machine frames owned by the guest.

- I/O Memory Management Unit (IOMMU) Disadvantages from following two papers:
  - **Cost** - Such transparency usually comes at the cost of emulating a fairly complex virtual device interface that was not designed to support virtualization efficiently.
  - **Accessibility** - Direct memory access (DMA) illustrates additional safety and performance issues. It enables an I/O device to read and write host RAM directly without involving the CPU, which is critical for achieving high-performance I/O rates.
  - **Complexity** - IOMMU adds another level of complexity to the system, which needs to be overcome in order to find the optimal caching strategy.
  - **Performance penalty** - IOMMUs can impose a performance penalty due to the extra memory accesses required to perform DMA operations.

- Benefits of adopting an I/O MMU include memory access, flexibility, multiplexing, security and Resource pre-allocation. However, it also has certain disadvantages, including cost, direct accessibility, complexity, and performance penalty. The system's requirements and its criteria's should be taken into account when deciding whether to use an I/O MMU.

7. **Find out what hypervisors Amazon is using in EC2, and describe their major characteristics.**

- Amazon Elastic Compute Cloud (EC2) uses two different types of hypervisors: Nitro and Xen.
  - **Nitro Hypervisor**

    It is a combination of dedicated hardware and lightweight hypervisor enabling faster innovation and enhanced security. The AWS Nitro System is the underlying platform for our next generation of EC2 instances that enables AWS to innovate faster, further reduce cost for our customers, and deliver added benefits like increased security and new instance types.

    It has completely re-imagined our virtualization infrastructure. Hypervisors protect the physical hardware and bios, virtualizes the CPU, storage, networking, and provide a rich set of management capabilities.

    With the Nitro System, we are able to break apart those functions, offload them to dedicated hardware and software, and reduce costs by delivering practically all of the resources of a server to your instances.

    The Nitro Cards are a family of cards that offloads and accelerates IO for functions, ultimately increasing overall system performance.

    The Nitro Security Chip Enables a secure boot process for the overall system based on a hardware root of trust, the ability to offer bare metal instances, as well as defense in depth that offers protection to the server from unauthorized modification of system firmware.

  - **Xen Hypervisor**

    It provides services that allow multiple computer operating systems to execute on the same computer hardware concurrently. It runs in a more privileged CPU state than any other software on the machine, except for Firmware.

    Prior to switching to the Nitro hypervisor, Amazon started out using the Xen hypervisor for its EC2 instances.

    Administrators can live migrate Xen virtual machines between physical hosts across a LAN without loss of availability. It uses para-virtualization technique, which host operating system into the host domain.

    Most operating systems which can run as a Xen HVM guest like Linux, FreeBSD, OpenBSD, NetBSD, MINIX, etc.

    Third-party developers have built a number of tools (known as Xen Management Consoles) to facilitate the common tasks of administering a Xen host, such as configuring, starting, monitoring and stopping of Xen guests.

Xen is an open-source hypervisor that was originally developed at the University of Cambridge. Amazon initially used the Xen hypervisor for its EC2 instances before transitioning to the Nitro hypervisor.

Some of the major characteristics of the Xen hypervisor include:

a. **Para-virtualization**: The Xen hypervisor uses a para-virtualization technique, which allows virtual machines to communicate directly with the hypervisor and other virtual machines, resulting in better performance than full virtualization.
b. **Resource allocation**: The Xen hypervisor provides granular control over resources such as CPU, memory, and network bandwidth. This allows users to allocate resources to virtual machines in a fine-grained manner.
c. **Support for multiple operating systems**: The Xen hypervisor supports a wide range of operating systems, including Linux, Windows, FreeBSD, and Solaris.

8. **Find out the URL to the source code of the Nitro hypervisor.**

- AWS launched the Nitro hypervisor. For its AWS services, which include EC2, Amazon built this own hypervisor.

  The aim of Nitro is to provide performance that is indistinguishable from metal.

  The Nitro hypervisor's source code is not accessible to the general public. As a result, it is not obtainable by anyone except AWS.

9. **Examine the Amazon EC2 VM offer capabilities and particularly the Amazon Machine Image (AMI) ((https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AMIs.html) and answer the following questions:**
   a. **How (i.e., in what units) does EC2 measure the CPU power of a virtual machine and how is the unit in question translated into the power of the physical processors?**
   b. **What kinds of machine instances are there as characterized by the power of their respective CPUs, platform (i.e., 32-bit or 64-bit), memory, storage, etc.? Please list all the instances in the nomenclature along with their respective characteristics.**
   c. **Which operating systems are available on the above systems?**
   d. **What is an AMI and what is its relationship to an instance?**
   e. **What are the components of an AMI?**

- **Amazon Elastic Compute Cloud (EC2):**
  - It is a cloud computing web service that provides resizable computation power. Customers may use EC2 to easily create and grow virtual servers to meet their unique needs without making huge upfront hardware costs. EC2 offers a number of virtual machine configurations suited for certain use cases, including as general-purpose computing, high-performance computing, and more.

- **Amazon Machine Image (AMI):**
  - It is a supported and maintained image provided by AWS that provides the information required to launch an instance. You must specify an AMI when you launch an instance. You can launch multiple instances from a single AMI when you require multiple instances with the same configuration. You can use different AMIs to launch instances when you require instances with different configurations.

  - **AMI Benefits:**

    A stable, secure, and high-performance execution environment for applications running on Amazon EC2.

    Provided at no additional charge to Amazon EC2 users.

    Repository access to multiple versions of MySQL, PostgreSQL, Python, Ruby, Tomcat, and many more common packages.

    Updated on a regular basis to include the latest components and these updates are also made available in the yum repositories for installation on running instances.

    Includes packages that enable easy integration with AWS services, such as the AWS CLI, Amazon EC2 API and AMI tools, the Boto library for Python, and the Elastic Load Balancing tools.

- **Answer: a**
  - A virtual CPU is a statistic used by EC2 to quantify the CPU power of a virtual machine. A virtual CPU is a metric that represents a part of a real virtual core. One virtual machine represents a single hardware thread of a real CPU core.
  - An EC2 instance type, for example, may have 5 vCPUs, which implies it has access to four hardware threads of a physical CPU core. If the actual CPU core has a clock speed of 3.5 GHz, the instance's total CPU power will be around 10 GHz (5 vCPUs x 3.5 GHz per vCPU).
  - The conversion of vCPUs to actual CPU power is determined by the hardware utilized by EC2. EC2 makes use of a range of physical CPU types, including Intel Xeon, AMD EPYC, and Graviton2 CPUs, each with a unique performance profile. Moreover, various EC2 instance types may employ different CPU generations or settings, affecting performance.
  - Amazon provides a statistic called an EC2 Compute Unit to assist consumers compare the CPU power of different EC2 instances. One ECU is defined as the CPU capability of a 2007 Opteron or Xeon processor running at 1.0-1.2 GHz, depending on the instance type.

- **Answer: b**

  Amazon provides a variety of EC2 instance types with different capabilities:

- **General Purpose Instances** (A1, M5, T3a): These are mostly utilized in services related to web servers, mobile or gaming development environments or apps, or enterprise-level applications like ERP or CRM.
- **Compute Optimized Instances** (C5n, C6g): These instances deliver high performance at a cost-effective price and are typically used in applications like web servers and scientific modeling.
- **Memory Optimized Instances** (R5n, R6g, x1e): These instances are used for memory-intensive workloads that are required to process large datasets at a fast speed.
- **Storage Optimized Instances** (D2, H1, I3): These instances are suitable for cloud-running applications that run high transaction and low latency workloads in use cases such as in-memory databases, data warehousing, and data analytics.
- **Accelerated Computing Instances** (P3, G4, F1): These instances use additional hardware accelerators like Graphics Processing Units (or GPUs) and Field Programmable Gate Arrays (or FPGAs) that enable higher throughput in compute-intensive applications with more parallelism.

Depending on workload and application requirements, we can choose from the different Amazon EC2 instance types that we have explained in this guide.

- **Answer: c**

Amazon Web Services (AWS) EC2 instances:

- t2.micro, c5.18xlarge and m5.large with vCPU, EBS storage and 64-bit platform.
- Operating systems available include Amazon Linux, Ubuntu, CentOS, Windows Server, and others.
- Google Cloud Platform (GCP) Compute Engine instances:
  Operating systems available include Debian, CentOS, Ubuntu, Windows Server, and others.

- **Answer: d**

An Amazon Machine Image (AMI) is a template that contains a software configuration (for example, an operating system, an application server, and applications). From an AMI, we launch an instance, which is a copy of the AMI running as a virtual server in the cloud. We can launch multiple instances of an AMI.

- An instance is a virtual server in the cloud. Its configuration at launch is a copy of the AMI that we specified when we launched the instance.
- An AMI and an instance have the same connection in that an instance is produced from an AMI. An AMI serves as the foundation for generating an EC2 instance, and the instance inherits all of the AMI's configuration and software settings.

- **Answer: e**

  An Amazon Machine Image (AMI) is composed of several components that are used to create a virtual machine instance in Amazon Elastic Compute Cloud (EC2).

  These are the components of an AMI:

  - **Root volume**: This is the operating system base disk, which contains the root file system, installed apps, and any other data that was there when the AMI was built.
  - **Launching**: These are the permissions that govern who may use the AMI to launch an EC2 instance. We can select specific AWS accounts or groups, or can make the AMI available to anybody.
  - **Block device mapping**: This defines the storage volumes that will be associated to the instance when it is started from the AMI. Amazon Elastic Block Store (EBS) volumes or instance store volumes can be included in the block device mapping. Metadata: This includes information about the AMI, such as its name, description, version number, and architecture.
  - **Tags**: These are labels that can assign to the AMI to organize and manage resources.

10. Find out about the pricing of the EC2 platforms and provide a few examples.

- There are multiple ways to pricing of EC2 platforms: On-Demand, Savings Plans, Reserved Instances, and Spot Instances.
  - **On-Demand**: With On-Demand instances, we pay for compute capacity by the hour or the second depending on which instances you run. No longer-term commitments or upfront payments are needed. Data Transfer OUT From Amazon EC2 to Internet price for US East (Verizon) in New York is $0.01 per GB.
  - **Spot instances**: Spot instances allow you to request spare Amazon EC2 computing capacity for up to 90% off the On-Demand price.
  - **Savings Plans**: Savings Plans are a flexible pricing model that offers low prices on EC2 and Fargate usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a one- or three-year term.
  - **Reserved Instances**: It provides a significant discount (up to 72%) compared to On-Demand pricing and provide a capacity reservation when used in a specific Availability Zone.
  - **Dedicated Hosts**: It is a physical EC2 server dedicated for use. Dedicated Hosts can help to reduce costs by allowing using existing server-bound software licenses, including Windows Server, SQL Server, and SUSE Linux Enterprise Server and can also help you meet compliance requirements.

11. **From the above exercise, you will learn that it is possible to create a free machine instance. Please, do the following:**
    a. **Find out and document the essence of the respective Service Level Agreement (SLA) on; in particular write down what one needs to do in order to maintain this service free.**
    b. **Describe the process (i.e., what exactly one needs to do) to create a free machine instance that could be used as a server. (Do not, however, create anything yet!)**
    c. **Can you create a machine instance equivalent to your own PC and then transfer your own PC image there? If so, how would you achieve that?**

- **Answer: a**
    - To create a free machine instance on Amazon Elastic Compute Cloud (EC2), we can use the AWS Free Tier. This tier grants free access to a variety of AWS services, including EC2, for the first 12 months of use. Nevertheless, there are certain restrictions on how these services can be used to ensure that they stay free.
    - The Amazon Free Tier for EC2 gives up to 750 hours of free Linux, RHEL, or SLES t2.micro instances each month. Furthermore, the user can use up to 30 GB of free Amazon Elastic Block Storage (EBS), General Purpose (SSD).
    - To keep this service free, we should keep track of EC2 instances and EBS storage consumption to ensure they stay below the free tier restrictions. Moreover, we should be aware of additional Amazon services that may involve fees, such as data transfer fees and premium support.

- **Answer: b**

  To create a free machine instance on Amazon EC2, one should follow following steps:

    - Log in to an AWS account.
    - From the AWS Services tab, select compute. This will launch the dashboard of E2C.
    - Choose the AWS Region in which we want to provision the EC2 server then click on "Launch Instance".
    - Choose the desired instance type, such as the t2.micro, which is included in the AWS Free Tier.
    - Select the preferred AMI, such as a Linux distribution, that is available for free.
    - Configure the instance details, such as the number of instances, network settings, and storage options.
    - Review the configuration and launch the instance.
    - After launching the instance, one can connect to it using SSH or other remote access protocols and use it as a server for various applications and services.

- **Answer: c**

- It is feasible to establish a machine instance that is similar in our own computer and then move the PC image to an EC2 instance. However, the picture transfer procedure might be complicated and may necessitate the use of specialist software and settings.
- To transfer an EC2 instance, use software such as VMware or Virtual Box to generate a virtual machine image of the PC. After that, the image may be uploaded to Amazon S3, a cloud storage service, and imported into an EC2 instance.
- To verify that the image's compatibility with the EC2 platform, one may need to alter it by installing relevant drivers and adjusting network settings.
- Another option is to utilize a third-party migration application, such as Cloud Endure, to automate the process of transferring a PC image to an EC2 instance.
- These programs can handle image format conversion, picture transfer to S3, and running the EC2 instance with the imported image.