CS524

Homework 02

Muhammad Owais Imran
CWID: 20025554

Question01:

The strategy without paravirtualization is that hypervisor would need to schedule interrupt timers back-to-back at certain interval for non-idle machine when the guest operating system is scheduled back for execution by the scheduler, which could not be a scalable choice for virtualization in hardware processes. With paravirtualization, the virtualization code is altered in a way to request notification at specified time which makes real-time processing easier, in this way, instead of hypervisor setting up interrupts, the guest machine would itself request interrupt at certain intervals.

Question02:

The way x86 based processor architecture uses Advanced Programmable Interrupt Controller (APIC) for interrupt redirection to support Symmetric Multiprocessing (SMP). Accessing APIC in virtual mode is a costly operation because of the transition into and outside the hypervisor for each access request/operation. With paravirtualization having the complete view of virtual machine code, all calls to APIC are replaced with a single hyper call.

Question03:

In Linux and Unix, there are separate modes for each process it runs.

In CPUs, there is at least one dedicated register for the user (running user applications) and system (operating system application and processes) is present. However, the switch from user to system is made based on some configured interrupts on system level. Once these interrupts occur, the system saves the current execution state onto the stack and executes the interrupt handler, once it executes the interrupt handler, the user state is restored from the stack.

Question04:

In Intel manual, section 5.10.3, it states that “Unscrambled” in Intel Load Segment Limit (LSL) instruction mean the limit scaled as per the value set in the G flag in the segment descriptor. The unscrambled limit is loaded when the privilege level and type check pass into destination and set a ZF flag in the EFLAGS register.

Upon access of any segment by the processor, a limit check is performed to ensure the segment size is in accordance with the limit specified and the limit is highly dependent on the value of the G flag set in the LSL instruction.

Question05:

- **Advantages of I/O MMU**
  - The large memory region of can be allocated without it being a sequential block. The I/O MMU maps the contiguous virtual memory address to the underlying physical memory.
  - The existing capabilities of the infrastructure can be used to address the problems related to availability and load balance across different I/O channels.
  - Decoupling enables popular VM related features including suspension/resumption of VM as well as live migration of VMs from one physical machine to another.
- **Disadvantages of I/O MMU**
  - Degradation In performances of translation
  - Management Overhead
  - Consumption of physical memory to store I/O page translations.

Question06:

AWS EC2 uses Xen based hypervisors. Its major characteristics are:

- Support of Live VM Migration and Storage from one host to another, much needed feature to support upscaling and downscaling of resources.
- High availability by supporting automatic restart in case of host machine failure.
- By utilizing the features of embedded hardware, it downsizes the data center cost by effectively utilizing only the desired resources.

Question07:

There is no open-source code for Nitro Hypervisor, as it is a tool totally custom to the AWS and is purposely built for AWS based resource.

Question08:

a. Amazon EC2 measures computing power by means of EC2 Compute Units (ECU). 1 ECU ~ 1.0-1.2GHz of 2007 Opteron or 2007 Xeon Processor.
b. Following are the types of Compute Instances in AWS
    a. General Purpose: (M7g, M7i, M7i-flex, M7a, Mac, M6g, M6i, M6in, M6a, M5, M5n, M5zn, M5a, M4, T4g, T3, T3a, T2)
        i. Features the latest DDR5 memory
        ii. 20% enhanced networking bandwidth
        iii. EBS-optimized by default
    b. Compute Optimized: (C7g, C7gn, C7i, C7a, C6g, C6gn, C6i, C6in, C6a, C5, C5n, C5a, C4)
        i. Support for Clustering
        ii. SSD Backed instance storage
    c. Memory Optimized (R8g, R7g, R7i, R7iz, R7a, R6g, R6i, R6in, R6a, R5, R5n, R5b, R5a, R4, X2gd, X2idn, X2iedn, X2iezn, X1, X1e, High Memory, z1d)
        i. Designed for memory extensive operations. Such as databases, in-memory caches.
    d. Accelerated Computing (P5, P4, P3, P2, G5g, G5, G4dn, G4ad, G3, Trn1, Inf2, Inf1, DL1, DL2q, F1, VT1)
        i. 3rd gen AMD EPYC processor
        ii. Up to 8 NVIDIA h100 tensor core GPUs
    e. Storage Optimized (I4g, Im4gn, Is4gen, I4i, I3, I3en, D2, D3, D3en, H1)
        i. Up to 38 Gbps of Network Bandwidth
        ii. Featuring up to 15TB of NVMe SSD
    f. HPC Optimized (Hpc7g, Hpc7a, Hpc6id, Hpc6
        i. Up to 64 Cores of Gravitron3E processor.
c. Amazon EC2 supports a variety of OS including Windows, Amazon Linux, CentOS, Debian, Oracle Linux, Ubuntu Server, etc.
d. AMI stands for Amazon Machine Image, it is a pre-built template which users can use to preconfigure their AWS EC2 instance, each AMI comes with a specific set of configurations like OS, Networking Configuration, OS Boot scripts.
e. Following are the typical components of an AMI:
    i. Root Volume
    ii. Operating System
    iii. Boot Loader
    iv. File System
    v. Installed Packages and Software
    vi. Configuration Settings
    vii. Metadata
    viii. Manifest file
    ix. Permission and Ownership
    x. Customization Script (Optional)

Question09:

Amazon EC2 is free to try, with on-demand, reserved, spot and savings plan instance.

In on-demand pricing, you pay for what you use with no-upfront payments.

In reserved pricing, you reserve specified computing power for a specific timeframe, it is good if your workload is predictable, this includes up-front payment.

In a spot instance, you bid for unutilized and available AWS ec2 instance, you can use the system if your bid price is above the specified threshold.

In a savings plan, you end up paying the full amount as an upfront for securing a specific number of machines, you can provision them as per your needs.

Question10:

a. SLA is a contract b/w the service provider and an end user defining the scope, responsibilities, and roles of each side. To maintain free services of AWS EC2, one needs to signup in an AWS free tier, and the usage of all the associated services should be less than or equal to the threshold specified for free tier.
b. The steps for creating a free instance are:
    a. Create an AWS EC2 instance with the required OS, networking configuration, security policies, IAM roles, authentication keys, and required software and services (ensuring all these services are within the free-tier limit)
    b. Configure auto-scaling (if needed)
c. It is possible, you need to create an EC2 instance on AWS, and replicate all the hardware configurations as per your PC hardware. After that connect your PC with the cloud instance and transfer all the files and system image to the cloud EC2 instance.