Chapter 05

Cloud Networking

Episode 5.01

Compute Introduction

Cloud Compute Defined

- Computer calculations in the cloud
  - Operating systems
  - Services
  - Applications
  - Functions

Cloud Compute Benefits

- Dynamic performance improvement
- Use only when required
- Test new hardware capabilities
- Evaluate new software
- Implement specialized processing

Cloud Compute Challenges

- Latency of results
  - Traffic has to cross the Internet and back
- Learning new methodologies
- Understanding cloud architectures
  - Cloud structure and networking
- Understanding service provider options

Quick Review

- Cloud compute is computer-based calculations and processing in the cloud
- Dynamic performance improvements can be implemented easily in the cloud through automatic scaling
- You can easily evaluate new software in the cloud without significant expense
- It is important to consider latency introduced in cloud computing

Episode 5.02

CPU Capabilities

Central Processing Units (CPUs)

- The core compute engine
- Cloud providers offer varying technologies
- Plan for proposed resources
  - What is required at peak
- Implement available resources
  - What is required at all times
  - Use elasticity to achieve peak requirements

CPU Technologies

- Hyperthreading
  - Multiple threads of concurrent operation
  - Results in multiple virtual CPUs
  - For example, a 4-core hyperthreaded CPU = 8 virtual CPUs

CPU Technologies

- VT-x
  - Virtualization technology in the CPU
  - Intel's solution
  - AMD implements AMD-V

Overcommitment Ratios

- Utilize real resources for virtual machines well
- Scenario:
  - 2 CPUs
  - Each is quad core
  - Each is hyperthreaded
  - Total of 16 virtual CPUs
  - Run 4 virtual machines, each with 8 CPUs
  - Result is a 2:1 overcommitment ratio
    - 2 virtual processors for each of the CPUs (including hyperthreading)
- Overcommitment is the primary factor in private clouds
- Cloud service providers hide this from you and perform it themselves

DEMO

- Show Openstack “Overcommitting CPU and RAM”

https://docs.openstack.org/arch-design/design-compute/design-compute-overcommit.html

Quick Review

- The CPU is the core compute engine and cloud providers offer varying levels of capability
- Hyperthreading is available in many CPUs and provides multiple virtual processors in each core
- Overcommitment can be used in virtualization so that more virtual machine CPUs can exist than actual CPUs in the host system

Episode 5.03

Memory Requirements

Memory Requirement Factors

- Operating system
- Services
- Applications
- Processes

Memory Ballooning

- A feature of virtualization platforms
  - Unused, allocated memory for one guest can be used by another
  - Allows for overcommitment of memory
- Mostly used in private clouds from a configuration perspective
  - Service providers may used it, but you won't configure it
- Bursting
  - The action of ballooning

DEMO

- Discuss
- https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/memory-optimized-instances.html
- AWS Memory Optimized Instances

Quick Review

- Consider everything running on a virtual machine when determining memory requirements
- Memory ballooning allows unused memory from one VM to be used by another
- Cloud providers offer special compute instances that are memory optimized

Episode 5.04

Performance Considerations

Performance Factors

- CPU
- Memory
- Disks
- Network

DEMO

- Optimizing CPU options in AWS

  https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-optimize-cpu.html

- Storage Optimized Instances in AWS

  https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/storage-optimized-instances.html

- Load testing in GCP

  https://cloud.google.com/community/tutorials/load-testing-iot-using-gcp-and-locust

Quick Review

- CPU and memory are the primary performance factors specifically for compute
- Disks are important for local data access
- Network through is important for transfer of data (both latency and speed of transfer)

Episode 5.05

Cost Considerations Lab

Hands-On

- Azure Pricing Calculator

  https://azure.microsoft.com/en-us/pricing/calculator

- AWS Pricing Calculator

  https://calculator.aws/#/

- GCP Pricing Calculator

  https://cloud.google.com/products/calculator/

Quick Review

- Costing in the cloud is a combination of desired resources and performance requirements
- Cloud vendors provide costing calculators you can use before implementation
- Cloud vendors provide monitoring and budgeting solutions to constrain spending

Episode 5.06

Energy Savings

Public/Community Cloud Energy Savings

- Shared resources = energy savings
  - Only use what you need on a shared system

Traditional Private Deployments

- Departmental servers
- Multiple data centers
- Localized servers in distributed companies
- High-powered desktops in many cases

Private Cloud Deployment Energy Savings

- Virtualization changed everything
- Private cloud is basically automated virtualization
  - With some extra bells and whistles
- The new deployment that saves energy
  - Multiple virtual servers on a single box
  - Services accessed across the Internet in the private cloud
  - Virtual desktops with high computing power
    - Possible shared among multiple resources

Quick Review

- Shared resources in the public and community clouds results in overall energy savings
- A factor of 10-20 times energy savings is not uncommon
- Private cloud offers energy savings for the same reason; however, the savings are the same for any virtualization implementation

Episode 5.07

Dedicated vs. Shared Compute

DEMO

- AWS Dedicated Hosts and Instances
  - https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/dedicated-hosts-overview.html
- Azure Dedicated Host
  - https://azure.microsoft.com/en-us/services/virtual-machines/dedicated-host/

Quick Review

- A dedicated host is a physical server that is used only for your instance(s)
- A dedicated instance is an instance specified to run on a definite host (definition may vary by provider)
- Dedicated hosts can result in better performance as you control what runs on the actual machine

Episode 5.08

High Availability and Disaster Recovery for Compute

HA/DR Effect for Compute

- High availability and disaster recover (HA/DR)
- Server or service must be there
  - Virtual servers demand the full virtual server be available
  - Serverless compute requires the function to be available

HA/DR Effect for Compute

- Availability functions for compute
  - Clustering
    - Multiple instances with a primary and failover
  - Load balancing
    - Multiple instances with workload rotating between them
  - Serverless limits
    - Imposed by service provider

DEMO

- AWS Lambda Limits
  - https://docs.aws.amazon.com/lambda/latest/dg/limits.html

Quick Review

- Clustering allows for the use of multiple virtual machines in providing the same application with failover options
- Load balancing allows for the use of multiple virtual machines in providing the same application with rotation among accessed machines
- Serverless compute handles high availability automatically

Episode 5.09

Monitoring Lab

Monitoring Terminology

- Target object
  - Baselines
  - Anomalies
- Alerts
- Events
  - System can log
  - Event collection for analysis

Event Correlation

- Event timestamps are used to correlate
- Ex:
  - Event A happened at 10:17:32 and Event B happened at 10:17:33
  - Event B and A are related
  - Maybe Event B was caused by Event A
- Correlation benefits
  - Determination of cause
  - Locating attack points
  - Identifying errant code

Hands-On

- Monitoring in AWS

Quick Review

- An anomaly is an event outside of the ordinary expectations
- Event correlation is the linking of two different events based on timestamps
- Cloud providers offer monitoring solutions like CloudWatch in AWS

Episode 5.10

Forecasting

Forecasting Required Resources

- Forecasting is looking into the future to determine needs
  - Look at today to predict tomorrow
- Baseline
  - Standard normal performance today
  - Current average utilization
  - Recent peak utilization
  - How often?

Forecasting Required Resources

- Upsize/increase or downsize/decrease resources to meet future demands
  - CPU
  - Memory
  - Storage

Quick Review

- Forecasting is about looking at yesterday and today to predict tomorrow
- A baseline is required to detect changes in use
- Current average utilization and recent peak utilization are both useful metrics

Episode 5.11

Policies

Policies and Monitoring

- Monitoring may reveal sensitive data
  - Can monitor down to the process level
  - Identifying processes can give you insight into points of attack

Policies and Monitoring

- Policies should be in place
- Policies in support of event collection
  - What can be monitored?
  - When should it be monitored?
  - What can be correlated?
- Policies to communicate alerts appropriately
  - How should it be reported?
  - Who should be notified?

Quick Review

- Monitoring can reveal sensitive data
- Policies should define what can be monitored when
- Policies should also define correlation allowances

Episode 5.12

vGPUs and Cloud Compute Performance

vGPUs

- vGPUs provide physical GPUs to one or more virtual machines
- Some implementations may provide direct access to GPU hardware
- Others implement it through software layers

vGPUs Use Cases

- Virtual Desktop Insfrastructure (VDI)
- High-Performance Computing (HPC)
- 3D Rendering and Design
- Machine Learning (ML) and AI
- Video Encoding and Streaming

Compute Performance

- Perceived compute performance is real compute performance to the user
- One thing is not the thing that results in perceived performance outcomes
  - CPU
  - Network
  - Memory
  - vGPU

Episode 5.13

Containers and Orchestration

Containers Defined

- Definition
- Container formats include Docker, Open Container Initiative (OCI), App Container (appc), and LXD image format
- Container benefits:
  - Isolation
  - Portability
  - Scalability
  - Resource Efficiency

Containers Applications

- Microservices implementation
- Continuous Integration/Continuous Deployment (CI/CD)
- Dynamic solutions (scalability)
- Multi-cloud deployments

Container Orchestration

- Management and automation of containers
- Orchestration tasks:
  - Deployment
  - Scaling
  - Load Balancing
  - Health Monitoring
  - Rolling Updates

Orchestration Solutions

- Private Cloud
  - Examples: Kubernetes, Docker Swarm, Apache Mesos
- CSPs
  - Google Kubernetes Engine
  - Amazon Elastic Kubernetes Service (EKS)
  - Azure Kubernetes Service (AKS)

Episode 5.14

Additional Compute Considerations

Subscription Services

- May reference a software licensing model
  - Month-to-month
  - Year-to-year
- May reference the simple access to the CSP

Serverless Applications

- Applications are built and run without management of the infrastructure
- Servers are still involved
- You don’t have to manage the servers
- Typically, event-driven, scalable, pay-per-use, and reduced overhead
- Solutions include AWS Lambda, Google Cloud Functions, and Azure Functions

APIs

- Application Programming Interfaces (APIs) allow access to consistent functions without coding
- CSP APIs provide access to the cloud system
- Application APIs provide access to functions
- API selection factors:
  - Vendor support
  - Version and features
  - Type: REST, WebSocket, custom, OS-dependent

Deploying VMs and Images

- Private cloud
  - Through cloud management layers
  - Through virtualization engine software
- Public cloud
  - Through machine images
  - Through templates
- Custom images may be created and deployed in either model

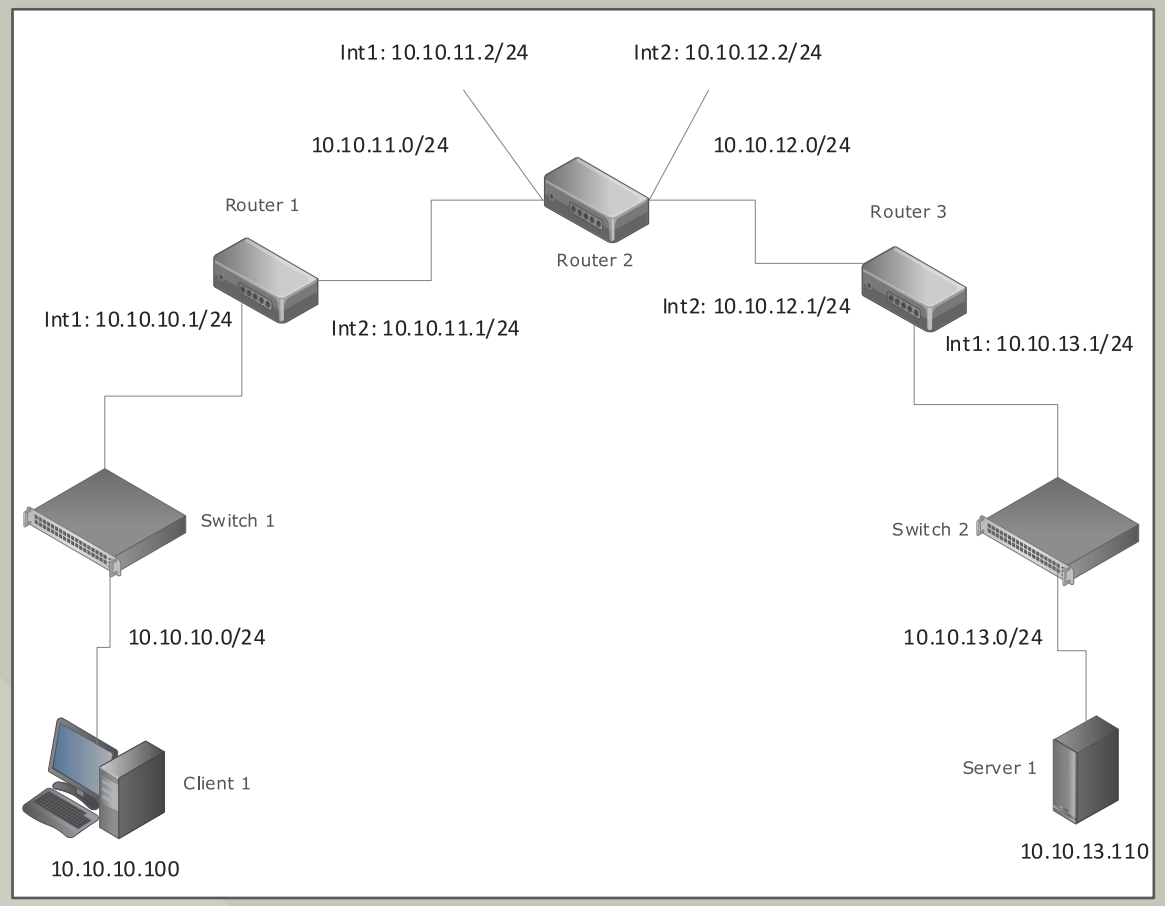
Templates and Validation

- Templates are used to consistently deploy solutions
- Templates reduce validation efforts
- Validation is still always important

Episode 5.15

Troubleshooting Network Problems Lab

Scenario



In the network configuration, Client 1 desires to access Server 1, which provides an interface into the organization's private cloud. However, when the client attempts to reach the server, it is unable to connect. The application used to access the server is a web browser and a DNS name is used to access the server with no numbers or digits in the URL. In addition to the information shown in the diagram, ACLs exist on Router 2 that disallow communications through the router on TCP ports 80 and 21. All other communications are allowed. ACLs exist on Router 3 that allow all communications on all ports. Finally, ACLs exist on router 1 that disallow TCP port 21 and allow all communications on all other ports. What is the cause of this communication problem?