Cloud Computing

CHAPTER 4

CAN THE CLOUD HELP OPERATIONS?

Dr. JAMIL S. ALAGHA

Can the Cloud Help Operations?

Organizations rely on IT infrastructure to ensure their daily operations achieve organizational goals.

In traditional IT, educated guesses and expensive capital investment purchases were normal

Cloud computing eases these pressures by introducing a ‘pay-as-you-go’ approach

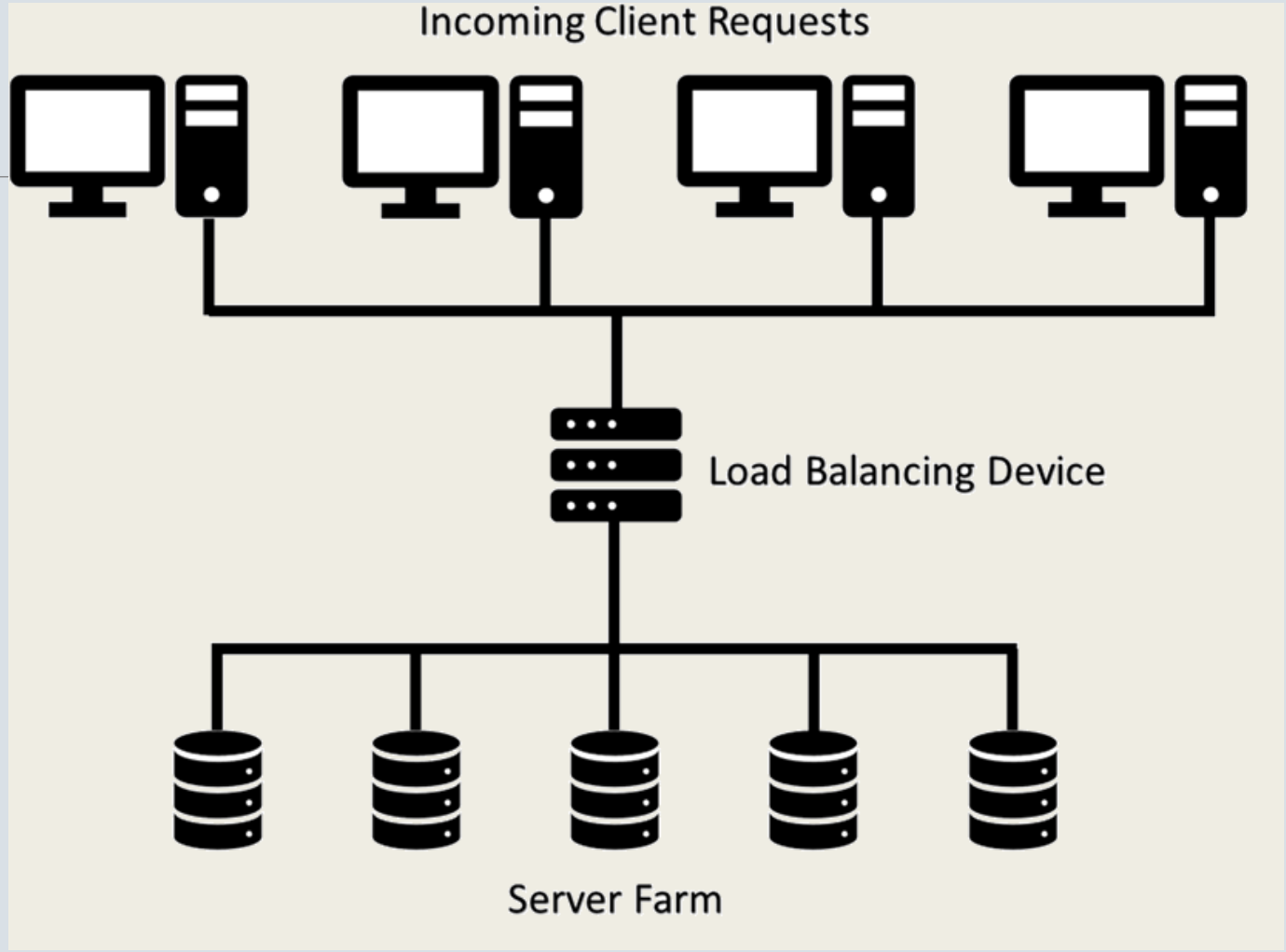
Two tools that help are load balancing and scalability/elasticity

Load Balancing

Process of efficiently distributing incoming network traffic among servers

Multiple network servers used to prevent system from being overwhelmed with request high volumes

Balancing algorithms can be static or dynamic, and centralized or distributed



Load Balancing

Systems expecting constant, uniform requests work well with centralized, static load balancers

Unpredictable, varying loads are better served with using a decentralized, dynamic algorithm that uses information collected at run time

Decentralized approaches require more system communication which slows processing overall

General Load Balancer Tasks

Ensures client and application resource requests (as well as network traffic load) are efficiently allocated across multiple servers

Manages resources to ensure high levels of uptime by monitoring availability and sending requests to online servers within the pool.

Ensures flexibility of resource pool size

Provides tools for elasticity and scalability according to demand levels

Centralized Balancing Algorithms

Rely on a control management node used to gather information from nodes eligible to process loads

Central node makes distribution decisions

Hardware or software load balancer sits between the clients and server farm performing a “traffic cop” role

Requests are allocated across all available servers in an intelligent way.

Servers are utilized to maximize speed and enhance utilization.

Servers used uniformly to extend their lifespan

Decentralized Balancing Algorithms

No central manager exists

Nodes eligible for processing exchange their system state information with every other node

Inter-node communication can slow processing

Static Load Balancing

Determined upfront by IT specialists and then put into place for use

Best practice approaches that may not use specific network characteristics to determine or adjust load allocation

Non-preemptive = once a request has been allocated to a specific server, it will not be transferred to another

Seeks reduction of execution time

Less server monitoring and communication takes place

Primary downside is that current system state not considered when allocations are made so if a server is slowed, it may still receive additional workload

Requestors may experience noticeable differences in response times from visit to visit

Static Load Balancing Algorithm Examples

Random: Randomly selecting the next server to receive a request is one method used to distribute workload

Round Robin Algorithm: Best suited when servers are homogeneous, and requests are approximately the same magnitude. Server requests are put into an ordered sequence

IP Hash: Like random method. Uses IP address of client to select host to service request

Dynamic Load Balancing

Use available information when trying to decide what server should receive next assignment

Monitor each servers' status and related changes to distribute work

Centralized dynamic algorithms rely on a central node for communication and load assignment

Decentralized dynamic algorithms move the decisions to nodes

Dynamic Load Balancing Algorithm Examples

Central: Centralized and dynamic approach. Requests stored in FIFO queue. Oldest node request for work receives next processing task

Transfer: Looks at on-going server loads and determines if node is overworked. If so, central manager can move task from one server to another

Least Connection Algorithm: Looks at load distribution based on number of active connections between a server and clients. Distribution of new loads (or re-distribution of tasks) seeks to keep connections approximately even

Cloud Load Balancing Algorithms

Cloud environments that are highly virtualized often have specialized load balancing algorithms

Virtual server nodes can be created as demand increases

Interesting implementation algorithms include variants of the central manager and threshold approaches

Sophisticated and imaginative approaches exist

Example Cloud Load Balancing Algorithms

Central Manager: Central server selects node for load transfer. Generally, server having least current load chosen to receive newest request. New processing nodes created as load increases, or nodes removed if load reduces

Threshold Algorithm: As demand increases, new processing nodes are created. Loads assigned immediately to new nodes.

Hybrid: Uses multiple stage approaches to determine best node for processing task. Two-stage model may: (1) perform static balancing algorithm; and (2) enhance node selection with dynamic analysis.

Ant colony: Intelligent agents move forward and backward to track overloaded and underloaded nodes. Ants update 'pheromones' to track node resources. Foraging pheromones help find overloaded nodes and trailing pheromones help find underloaded nodes.

Bee's life: Inspired by natural world and honeybee's process of searching for food

Genetic algorithm: Employs a 'survival-of-the-fittest' paradigm to ensure algorithms with higher efficiency produce 'offspring' that inherit best features

Client-side load balancing in the cloud

A few of the balancing algorithms covered in the prior section alluded to clients being able to ask that work be off-loaded. Another approach, called client-side load balancing, does not refer to the end-user as the client, but rather middleware acting as a client for back-end services. Just as a refresher, remember that middleware is software that acts as a bridge between backend services like a database, and users' applications. Front-end services interface with the user. Client-side Load Balancers (CLB) use an elastic cloud storage service to choose back-end database or web servers based on pre-determined criteria like current load, queue length, average processing time or other factors. In this case, the client finds the best service instead of a central manager or agent doing it.

Hardware versus Software Balancing

Hardware Balancing

- In traditional IT operations, load balancing relied on hardware solutions.
- Hardware resided in on-premise data center maintained by in-house IT specialists
- Balancing used special appliance from a vendor like Cisco, Citrix, or Barracuda.
- Specialized devices sat in server racks and distributed traffic to directly connected physical servers

Software Balancing

- Most cloud providers use software solutions running on virtual machines
- Part of the server suites included by cloud providers
- Affordable even for small companies

NGINX Plus

This load-balancing software, used by many leading cloud-based organizations, helps with processing requests. High-volume sites like Netflix and Dropbox have used this cloud load balancing solution to ensure content delivery in secure, reliable fashion.

Cloud-based Balancing

Best practice to provision load balance server in the same environment as what it balances

For organizations using cloud, cloud-based balancing is preferable

Cloud-based systems must balance goals in addition to loads: one is to ensure traffic and load are routed in efficient and sensible ways; another is to maintain security

Noisy Neighbors Problem

Cloud environments provide cost savings, elasticity, and other benefits due to multi-tenant architecture.

Like an apartment complex, if the walls are thinly insulated or if too much infrastructure is shared, presence of annoying neighbor becomes obvious

In cloud, noisy neighbors monopolize bandwidth, storage access, CPU, or other resources.

System performance may be impacted by what neighbors are doing

‘Cheap' cloud means not enough infrastructure exists to accommodate all the tenants.

Noisy neighbors are avoided by ensuring adequate bandwidth is present and having up-to-date fast storage devices and servers.

Having effective balancing algorithm will ensure requests receive prompt fulfillment

Effective VLAN can help solve issues

Cloud Load Balancing vs. DNS Load Balancing

DNS balancing

- Network optimization technique
- Used to route web domain's incoming traffic flows to appropriate web server
- Concerned with faster access to resources
- Balances load requests for domain

Cloud load balancing

- Roots in DNS balancing
- Instead of transferring requests to a pool of web servers, it routes loads among data centers, server pools, virtual machines, and/or other resource groups
- Cloud load balancing much broader and meant to handle a wider variety of traffic and workload needs

Scalability versus Elasticity

- Elasticity adjusts a system on-the-fly to match workload changes. Can be up in scale or down in scale
- Scalability related to system's capability to accommodate larger loads easily by adding resources. Might mean improving hardware power (scale up) or adding additional nodes to the server farm (scale out).
- Scalability is capability to grow, elasticity is ability to react to changes dynamically

Scaling Up versus Scaling Out

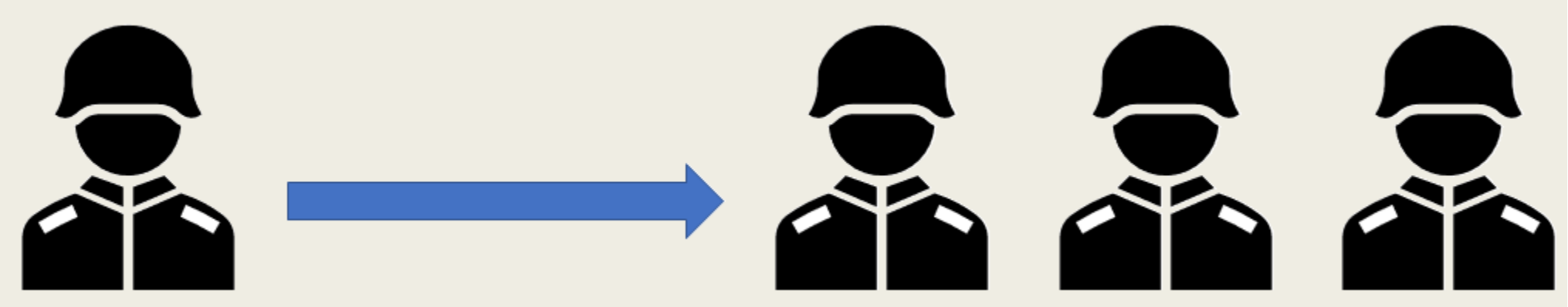
Scaling up increases individual capabilities.

Server scales up by adding more memory or more storage

Scaling out refers to duplication and expansion by adding more

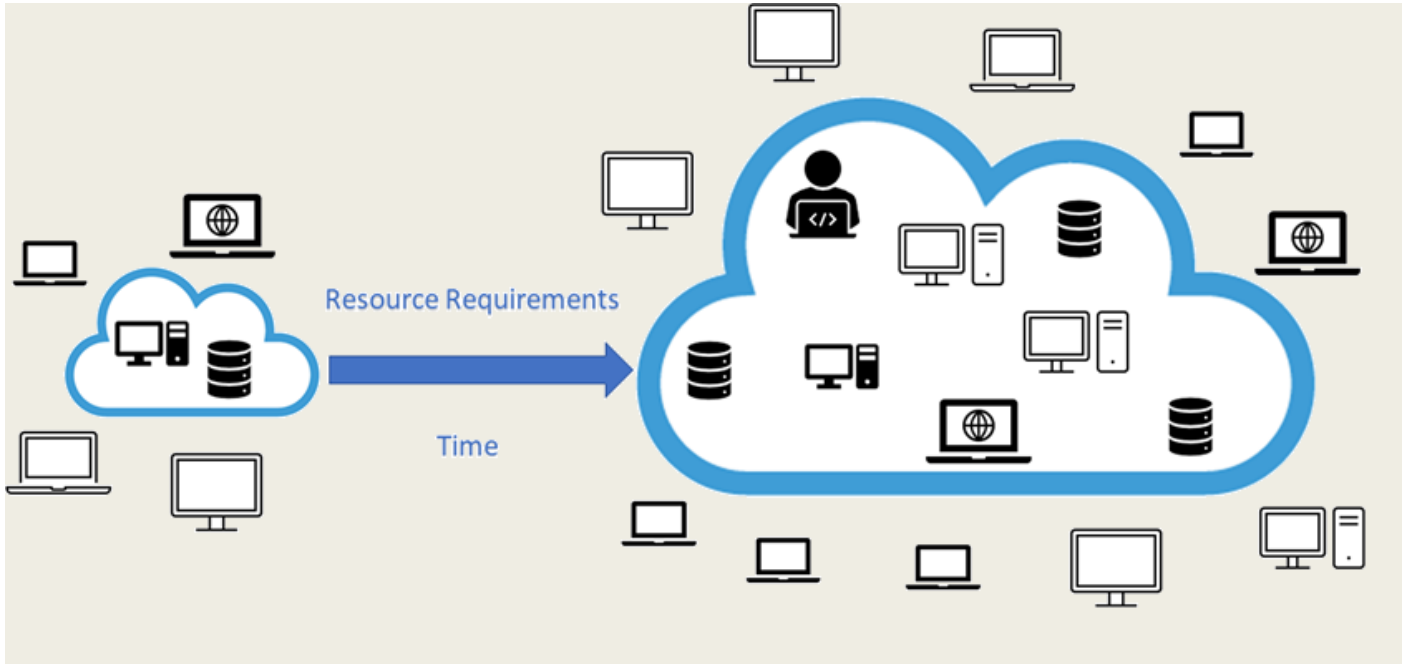
Server scales out by adding more servers to the pool

In a military group, scaling up means an individual soldier receives training so she can perform better and scaling out means more soldiers added to complete a task



Elasticity in Cloud Environments

In cloud infrastructure, hypervisors create new VMs or containers as system demand ramps up. Likewise, as demand decreases, hypervisors can remove extra VMs or containers. Hypervisors work in real time to ensure computing resources meet demand.



Elasticity Challenges

Learning Curve: Requires rethinking the ways resources are allocated, costs are accounted, and systems are licensed/used

Response Time: Implementing elasticity requires response time

Monitoring Elastic Applications: Inherently volatile so tools for monitoring resource use need to be sophisticated

Stakeholder Needs: Requirements of each user need to be gathered and considered during the implementation.

Multiple Levels of Cloud Control: Serves the needs of many users so control structures must be in place.

Security: Resources appear and disappear and access to resources must be managed carefully

Privacy and Compliance: Data privacy and compliance with international laws are inherently complex

Benefits of Elasticity

Ease of Implementation: Elasticity often transparent

Failover and Fault Tolerance: Failing server can be cloned proactively and deployed on a new VM before old one stops

On-Demand Computing: When usage requirements spike, capacity is readily available

Pay only for what you use: Economic benefits of paying only for computing, storage, and networking resources an organization uses is a huge benefit

Standardization of Server Pool: Techniques such as IaC ensure homogeneity of infrastructure

Autoscaling

Cloud computing best practices include several methods for managing the elasticity feature called *autoscaling*. Most often, existing servers are watched or monitored. If an important usage threshold is reached, the system expands and adds a server. Autoscaling can also be based on business rules. For instance, in advance of 'Cyber Monday' (e.g. in the U.S., this is the Monday after Thanksgiving when massive sales become available via Internet retailers) an organization might ramp up its server capacity in anticipation of greater demand on its website. A related approach is to use scheduled scaling to deploy server resources based on regular high demand periods. For instance, a customer service organization may do most of its work from 8AM to 5PM.

Chapter 4 Summary

Concepts at root of cloud computing: load balancing and elasticity/scalability

Cloud load balancing ensures workloads are distributed in efficient and effective manner

Resources balanced in the cloud including processing capability, network interfaces and services, application instances, storage acquisition, and more

Cloud load balancing done using software solutions from cloud vendors

Large-scale public clouds provide balancing features to ensure applications maintain high availability and performance across multiple virtualized application servers

Cloud load balancing relies on the concept of elasticity to manage demands as resources grow and shrink

Elasticity reallocates resources to scale up to meet increasing demands or scale down as resource needs decrease

Scalability is ability to upsize operations from a managerial perspective