COST MANAGEMENT

FUNDAMENTALS OF COST OPTIMIZATION

Turn off the light.

- When you stop the instances, you stop paying for them.

Be elastic.

- Support workloads with the right amount of horsepower to get the job done.

Continually optimize.

- Drive recurring and improving savings through cost-aware architectures.

CONTROLLING UNDER UTILIZED RESOURCES

Do you use everything you pay for?

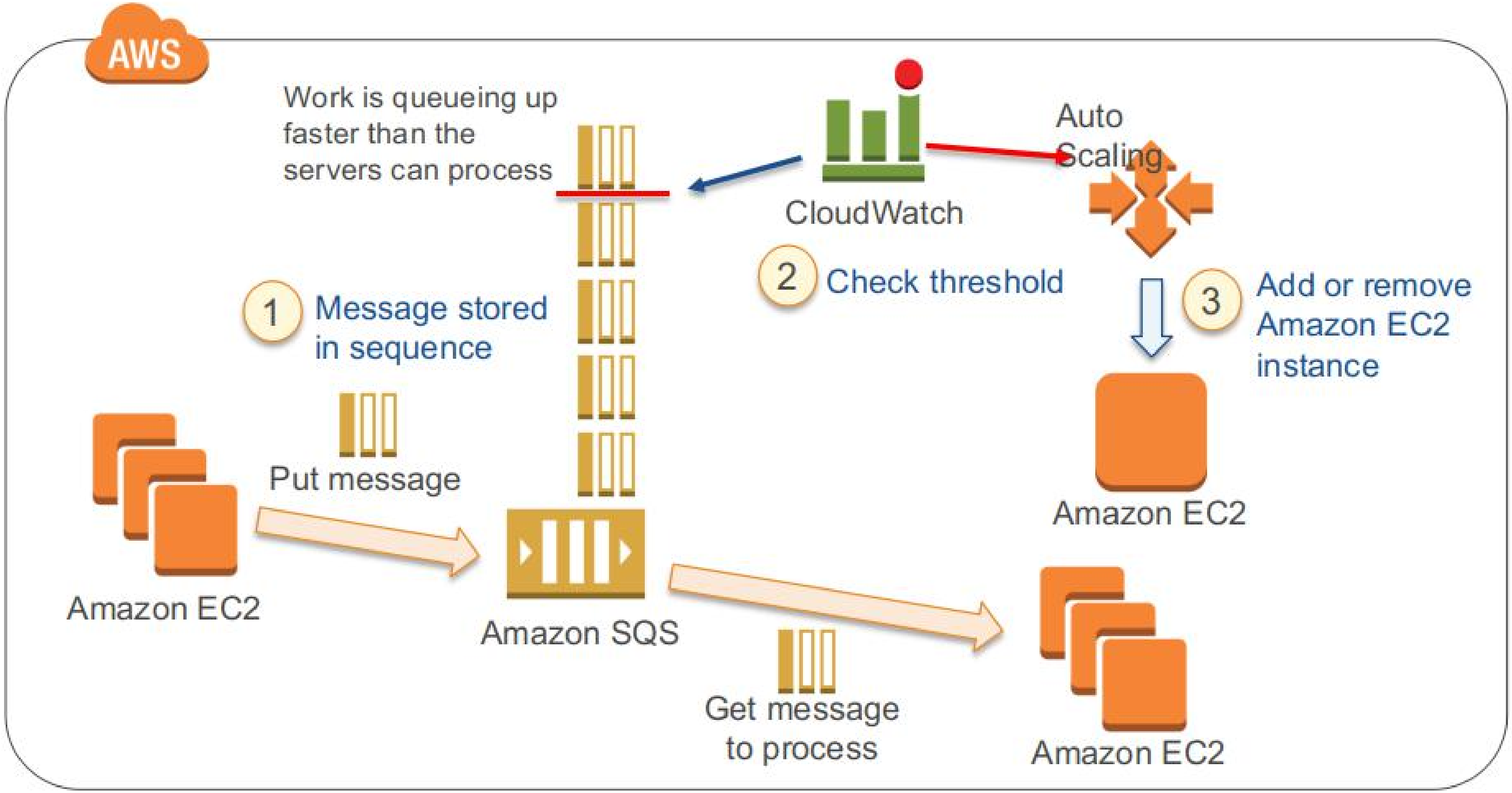
- Cloud is designed to be leveraged for on-demand capacity.
- De-provision unused resources.

Consider a batch processing system.

- Is your batch processing work completed and done? Stop and terminate the unused batch servers.

How do you monitor and remove unused AWS resources?

JOB OBSERVER PATTERN


AWS
Work is queueing up faster than the servers can process
CloudWatch
Auto Scaling
1 Message stored in sequence
2 Check threshold
3 Add or remove Amazon EC2 instance
Put message
Amazon EC2
Amazon SQS
Amazon EC2
Get message to process
Amazon EC2

BEST PRACTICE FOR EVERY CLOUD

The job observer pattern lets you coordinate the number of Amazon EC2 instances based on the number of jobs that need to be processed. Because this pattern automatically scales up or down based on the computational demand, you won’t have to over-pay or hit a bottleneck, and this improves cost-effectiveness. By scaling up as necessary, the overall time for executing jobs can be reduced by processing the jobs in parallel.

Another benefit of this pattern is that even if a batch server fails, the Amazon SQS messages would remain, enabling processing to be continued immediately upon recovery of the Amazon EC2 instance and producing a system that is robust to failure.

This pattern follows the Cloud Architecture Best Practices that we discussed in earlier modules:

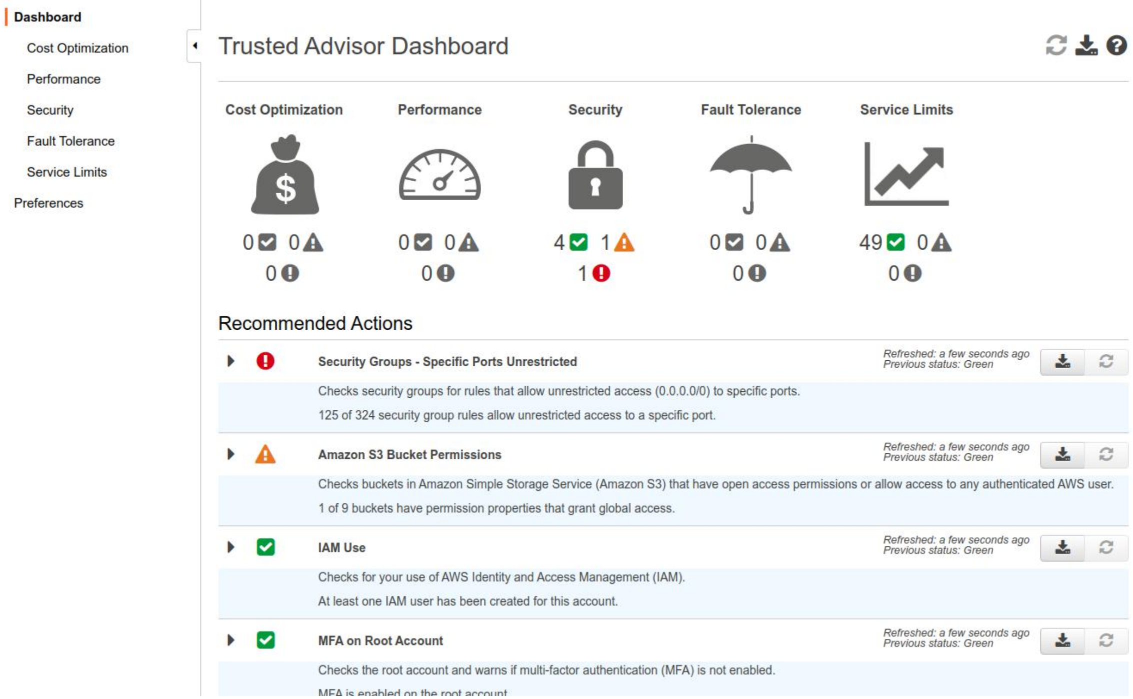
- Think parallel
- Loose coupling
- Do not fear the constraints

TRUSTED ADVISOR

Cost Optimization with Trusted Advisor

AWS Trusted Advisor is a web-based application that inspects your AWS environment and makes recommendations based on best practices.

- Opportunity to save money.
- Improve system performance.
- Close security gaps.
- Improve system fault tolerance



Recommended Actions

❗ **Security Groups - Specific Ports Unrestricted**
Refreshed: a few seconds ago
Previous status: Green
Checks security groups for rules that allow unrestricted access (0.0.0.0/0) to specific ports.
125 of 324 security group rules allow unrestricted access to a specific port.

⚠ **Amazon S3 Bucket Permissions**
Refreshed: a few seconds ago
Previous status: Green
Checks buckets in Amazon Simple Storage Service (Amazon S3) that have open access permissions or allow access to any authenticated AWS user.
1 of 9 buckets have permission properties that grant global access.

✓ **IAM Use**
Refreshed: a few seconds ago
Previous status: Green
Checks for your use of AWS Identity and Access Management (IAM).
At least one IAM user has been created for this account.

✓ **MFA on Root Account**
Refreshed: a few seconds ago
Previous status: Green
Checks the root account and warns if multi-factor authentication (MFA) is not enabled.
MFA is enabled on the root account

FLEET MANAGEMENT

Trade off between fault tolerance and high utilization.

- Fault tolerance requires less resource use to be able to successfully fail over.

Trade off between instance size and value for money.

- Higher utilization gives better compute value for money.
- Scaling granularity.

AMAZON EC2 PRICING OPTIONS

	On-Demand Instances	Reserved Instances (RIs)	Spot Instances
Term	None; Pay as you go	1 year or 3 years	Bid on unused capacity
Benefit	Low cost and flexibility	Predictability ensures compute capacity is available when needed	Large scale, dynamic workload
Cost	Pay for only what you use; no up-front commitment or long-term contracts	Pay low or no up-front fee; receive significant hourly discount	Spot price based on supply and demand – determined automatically
Use case	Applications with short term, spiky, or unpredictable workloads Application development or testing Billed hour forward	Applications with steady state or predictable usage Applications that require reserved capacity, including disaster recovery Users able to make up-front payments to reduce total computing costs even further	Applications with flexible start and end times Applications only feasible at very low compute prices Users with urgent computing needs for large amounts of additional capacity

AMAZON EC2 RESERVED INSTANCE TYPES

No Upfront

- Access a Reserved Instance without an upfront payment.
- Discounted effective hourly rate for every hour within the term, regardless of usage.
- 1-year reservation available.

Partial Upfront

- Part of the Reserved Instance must be paid at the start of the term.
- Discounted effective hourly rate for the remainder of the term, regardless of usage.
- 1-year or 3-year reservations available.

All Upfront

- Full payment made at the start of the term.
- No other costs incurred for the remainder of the term, regardless of usage.
- 1-year or 3-year reservations available.

RESERVED INSTANCE MARKETPLACE

Flexibility

- Sell your unused Amazon EC2 Reserved Instances
- Buy Amazon EC2 Reserved Instances from other AWS customers
- As your needs change, change your Reserved Instances

Diverse term and pricing options

- Shorter terms
- Opportunity to save on upfront pricing

Identical capacity reservations

AMAZON EC2 SPOT INSTANCES

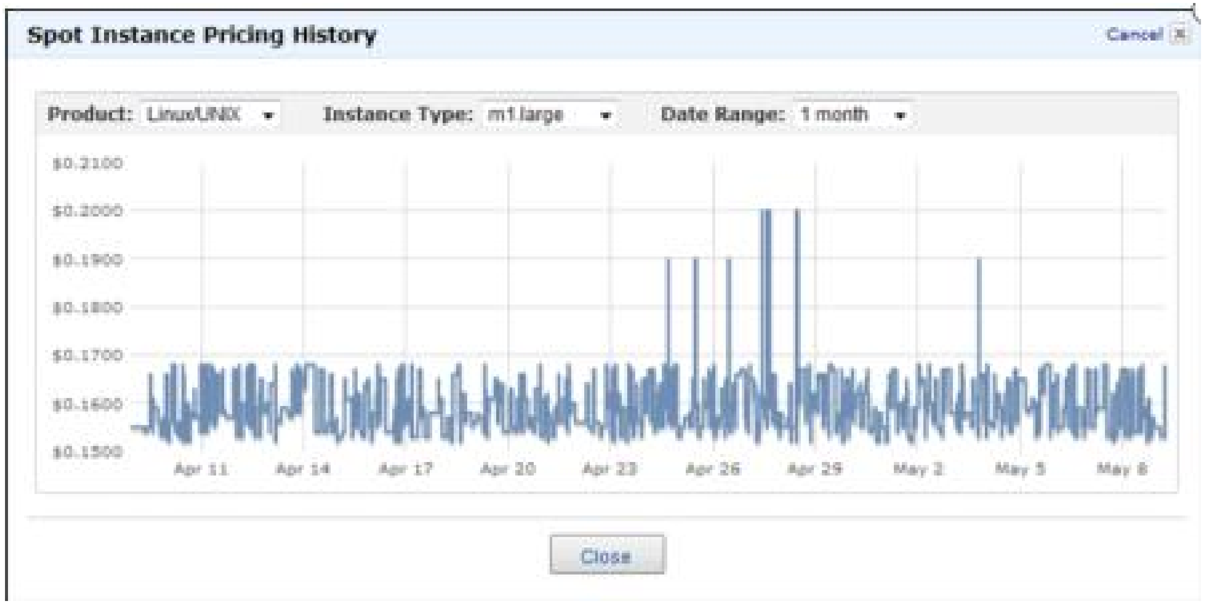
"Preemptible" on Google, "Spot" on Azure

Bid for unused AWS capacity.

Prices controlled by AWS based on supply and demand

Termination Notice provided 2 minutes prior to termination, stored in metadata

Best approach to temporary requests for large numbers of servers.



Spot Instance Pricing History

Product: Linux/UNIX
Instance Type: m1.large
Date Range: 1 month

SPOT USE CASES

Use Case	Types of Applications
Batch processing	Generic background processing (scale out computing)
Web/data crawling	Analyze data
Financial	Hedge fund analytics, energy trading, etc.
Elastic Map Reduce	Hadoop (large data processing)
Grid computing	Scientific trials/simulations in chemistry, physics, and biology
Transcoding	Transform videos into specific formats
Gaming	Back-end servers for Facebook games
Testing	Scale to large server pool to test software, websites, etc.

VIMEO’S SPOT MARKET CONSIDERATIONS (1 OF 2)

Never bid more than threshold (80% of on-demand price).

No more than 10 open spot requests at any time.

Bid 10% more than the average price over last hour.

Use spots for low-priority and less time-critical jobs.

Have more retries for jobs running on spots.

VIMEO’S SPOT MARKET CONSIDERATIONS (2 OF 2)

Watch out for open spot requests (add expiry to your requests).

Billed hour forward unless terminated by AWS.

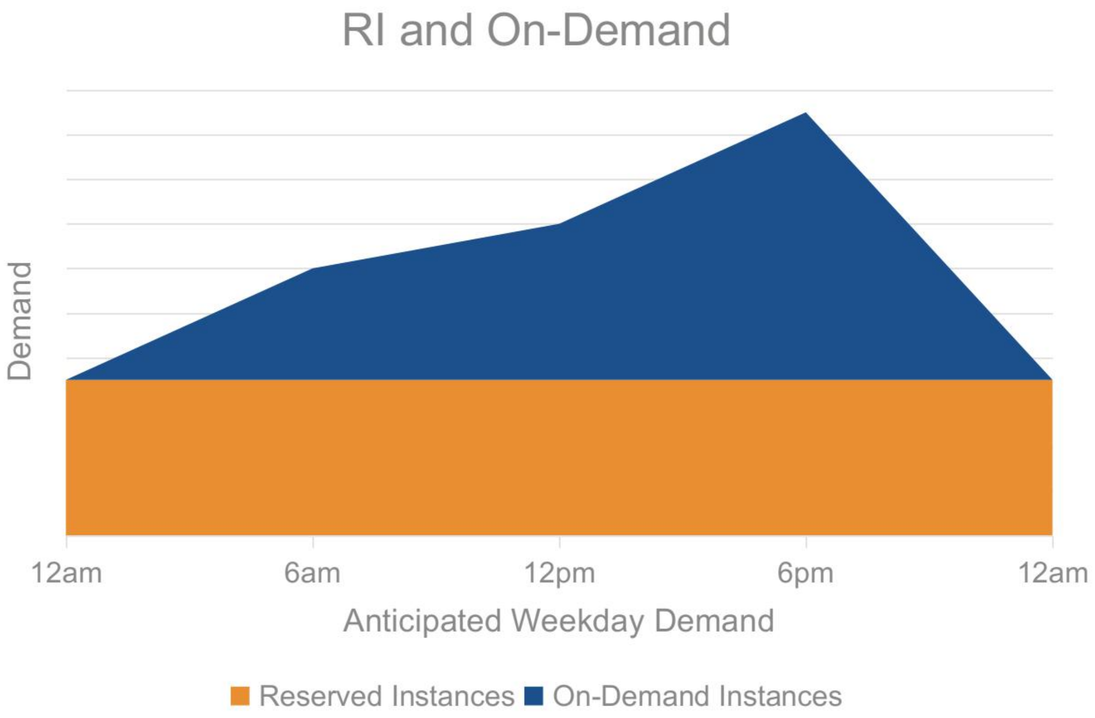
For long-running jobs, either bid higher on spot or use on-demand instances.

Fail over to on-demand when spot market is saturated.

Note Not every rule applies in every situation. For example, the "fail over to on-demand" may not agree with your organization's budget.

- What is your opinion and your situation?

LEVERAGING EC2 PRICING MODELS TOGETHER



BLENDED APPROACH

Choose instance type that matches requirements.

- Start with memory requirements and architecture type (32-bit or 64-bit).
- Then choose the closest number of virtual cores required.

Scale across Availability Zones.

- Smaller sizes give more granularity for deploying to multiple AZs.

Start with on-demand and then assess utilization for RIs.

COSTS FOR DATABASES

Multiple instance types to choose from

- Use small-sized database for data ingestion.

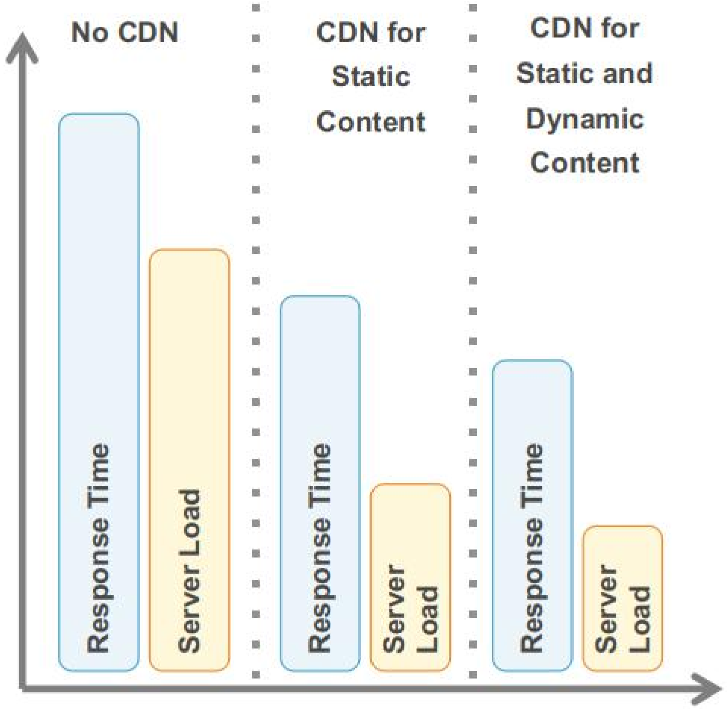
Amazon RDS

- If you have I/O intensive workloads, you may save money with Provisioned IOPS.
- Launch larger database from snapshot for reporting.
  - Enables information sharing without affecting the performance of production database.

OFFLOAD YOUR ARCHITECTURE

The more you can offload, the less infrastructure you need to maintain, scale, and pay for.

- Offload popular traffic to Amazon CloudFront and Amazon S3.
- Introduce caching.



DATA STORAGE AND TRANSFER COSTS – S3

Amazon S3 costs vary by region.

Priced by storage, request, and transfer.

- Storage cost is per GB-month.
- Per-request cost varies, based on type of request.
  - For instance, price per 1,000 PUT requests.

Transfer out has cost per GB-month (except in same region or to Amazon CloudFront), transfer in is free.

Pricing

DATA STORAGE AND TRANSFER COSTS – AMAZON

Reducing outbound costs

- Retrieve only required output.
- Enable Amazon EMR output compression.

Reduced Redundancy Storage (RRS) for Amazon S3

- Reduces replication of Amazon S3 objects.
- Reduces storage costs but drops durability of Amazon S3 objects.
- RRS can be enabled during or after upload.

YOU MAY USE CONSOLIDATED BILLING

Receive a single bill for all charges incurred across all linked accounts.

- Share reserved instances.
- Combine tiering benefits.

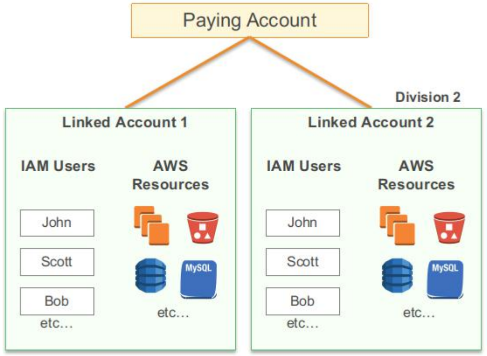
View and manage linked accounts.

Add additional accounts.

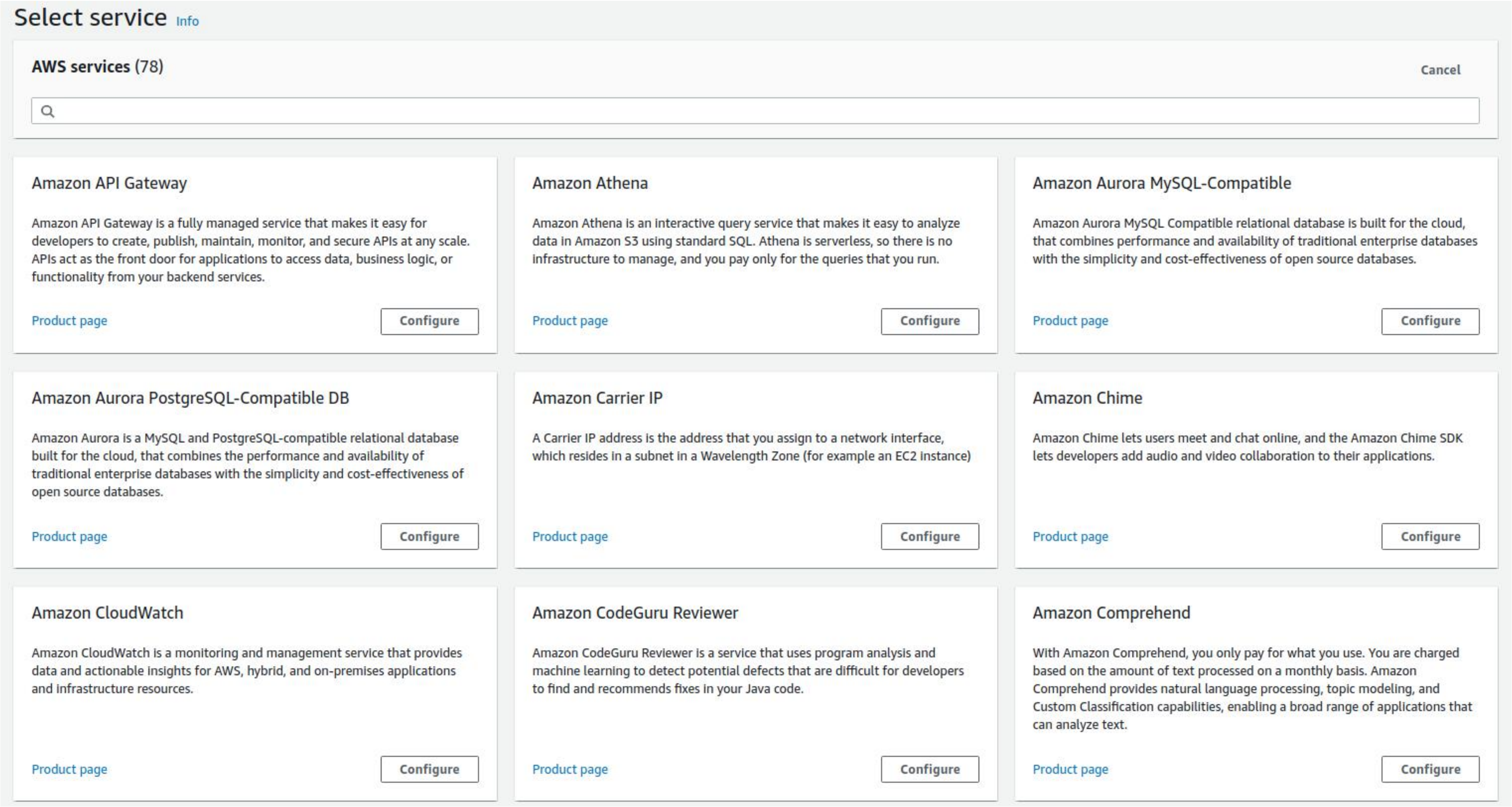
- Consolidated billing only supports one level depth.

However, this may not fit every need

- Many schools, for example, are using reseller to bill individual accounts
- As of re-invent 2020, this is changing and may not be a universal fit-all practices



AWS PRICING CALCULATOR



Select service Info

AWS services (78)

Cancel

Amazon API Gateway

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. APIs act as the front door for applications to access data, business logic, or functionality from your backend services.

Product page | Configure

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

Product page | Configure

Amazon Aurora MySQL-Compatible

Amazon Aurora MySQL Compatible relational database is built for the cloud, that combines performance and availability of traditional enterprise databases with the simplicity and cost-effectiveness of open source databases.

Product page | Configure

Amazon Aurora PostgreSQL-Compatible DB

Amazon Aurora is a MySQL and PostgreSQL-compatible relational database built for the cloud, that combines the performance and availability of traditional enterprise databases with the simplicity and cost-effectiveness of open source databases.

Product page | Configure

Amazon Carrier IP

A Carrier IP address is the address that you assign to a network interface, which resides in a subnet in a Wavelength Zone (for example an EC2 instance)

Product page | Configure

Amazon Chime

Amazon Chime lets users meet and chat online, and the Amazon Chime SDK lets developers add audio and video collaboration to their applications.

Product page | Configure

Amazon CloudWatch

Amazon CloudWatch is a monitoring and management service that provides data and actionable insights for AWS, hybrid, and on-premises applications and infrastructure resources.

Product page | Configure

Amazon CodeGuru Reviewer

Amazon CodeGuru Reviewer is a service that uses program analysis and machine learning to detect potential defects that are difficult for developers to find and recommends fixes in your Java code.

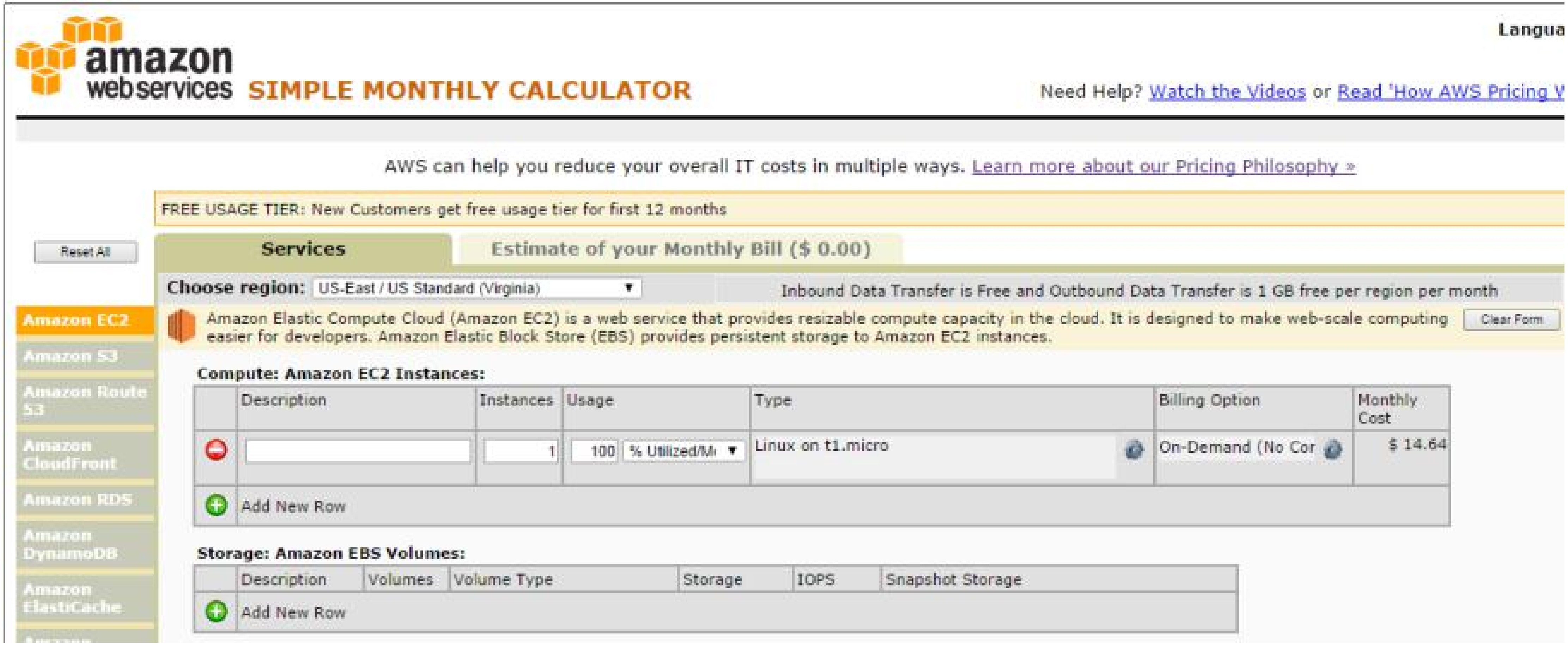
Product page | Configure

Amazon Comprehend

With Amazon Comprehend, you only pay for what you use. You are charged based on the amount of text processed on a monthly basis. Amazon Comprehend provides natural language processing, topic modeling, and Custom Classification capabilities, enabling a broad range of applications that can analyze text.

Product page | Configure

AWS PRICING CALCULATOR



SIMPLE MONTHLY CALCULATOR

Need Help? Watch the Videos or Read 'How AWS Pricing V

AWS can help you reduce your overall IT costs in multiple ways. Learn more about our Pricing Philosophy »

FREE USAGE TIER: New Customers get free usage tier for first 12 months

Choose region: US-East / US Standard (Virginia) ▼

Inbound Data Transfer is Free and Outbound Data Transfer is 1 GB free per region per month

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers. Amazon Elastic Block Store (EBS) provides persistent storage to Amazon EC2 instances.

Compute: Amazon EC2 Instances:

	Description	Instances	Usage	Type	Billing Option	Monthly Cost
⊖		1	100 % Utilized/M ▼	Linux on t1.micro	On-Demand (No Cor	$ 14.64
⊕	Add New Row					

Storage: Amazon EBS Volumes:

	Description	Volumes	Volume Type	Storage	IOPS	Snapshot Storage
⊕	Add New Row					

COMPARING TCO IS NOT EASY

Start by understanding your use cases and the applications that support them.

Take all the fixed costs into consideration.

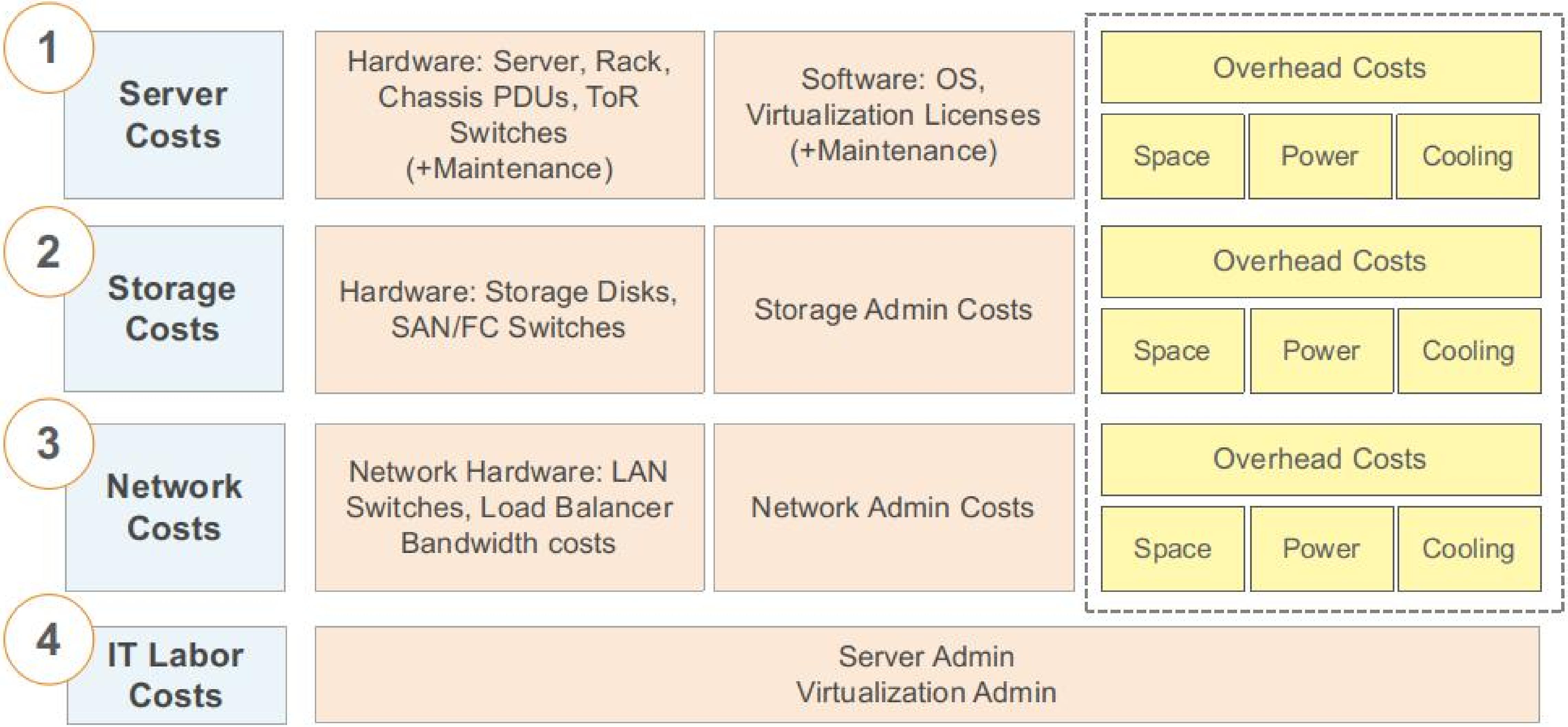
Use updated pricing (compute, storage, and net).

Leverage RI pricing vs. On-Demand pricing vs. spot instances.

Intangible cost savings: take a closer look at what you get as part of AWS.

If you are a higher ed institution you may get a data egress waiver the cloud provider. It is useful as the size of dataset increases.

TCO ESTIMATES FOR ON-PREMISES DEPLOYMENTS



1 Server Costs

Hardware: Server, Rack, Chassis PDUs, ToR Switches (+Maintenance)

Software: OS, Virtualization Licenses (+Maintenance)

Overhead Costs

Space

Power

Cooling

2 Storage Costs

Hardware: Storage Disks, SAN/FC Switches

Storage Admin Costs

Overhead Costs

Space

Power

Cooling

3 Network Costs

Network Hardware: LAN Switches, Load Balancer Bandwidth costs

Network Admin Costs

Overhead Costs

Space

Power

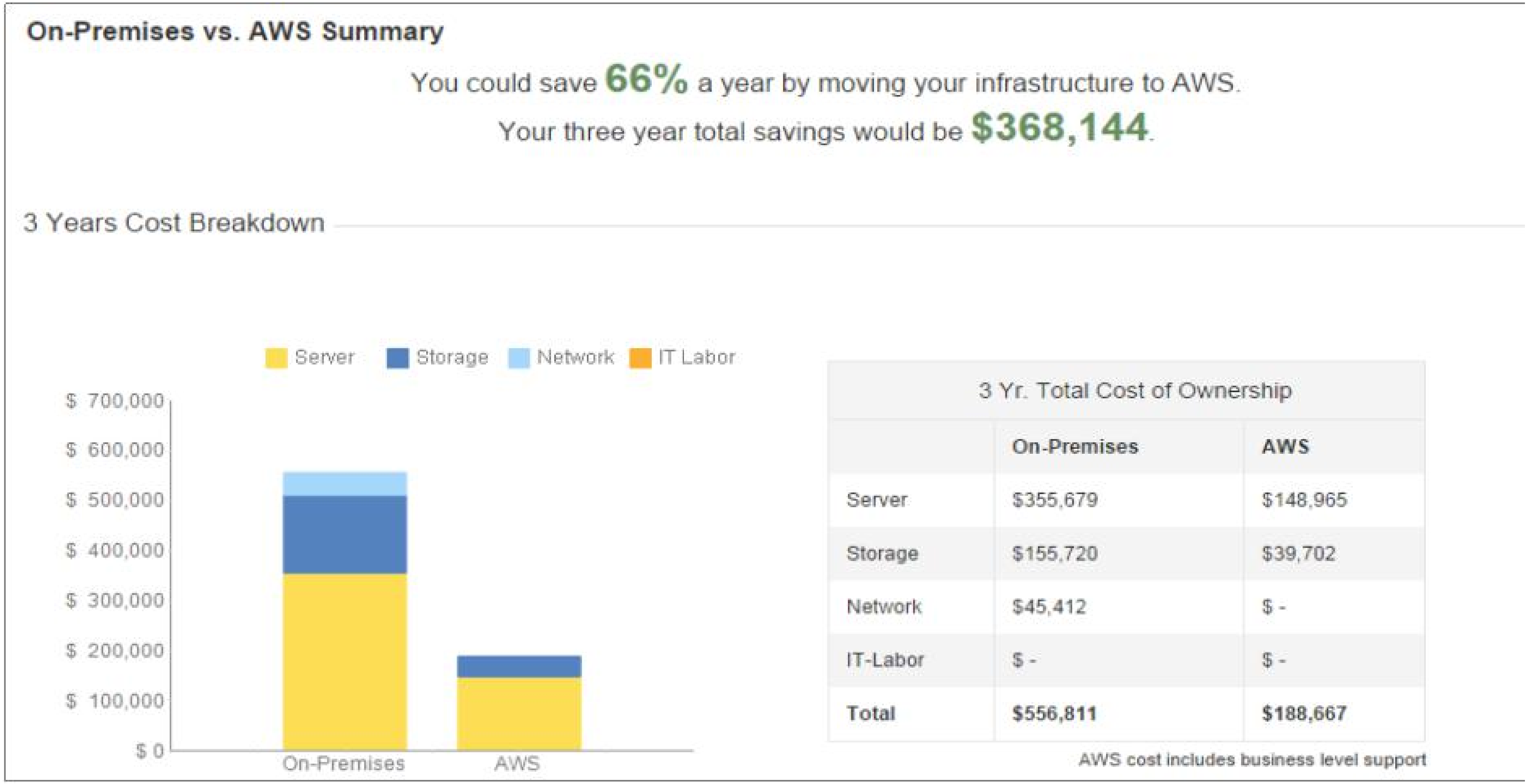
Cooling

4 IT Labor Costs

Server Admin
Virtualization Admin

AWS ONLINE TCO CALCULATOR

https://calculator.aws/#/



On-Premises vs. AWS Summary

You could save **66%** a year by moving your infrastructure to AWS.

Your three year total savings would be **$368,144**.

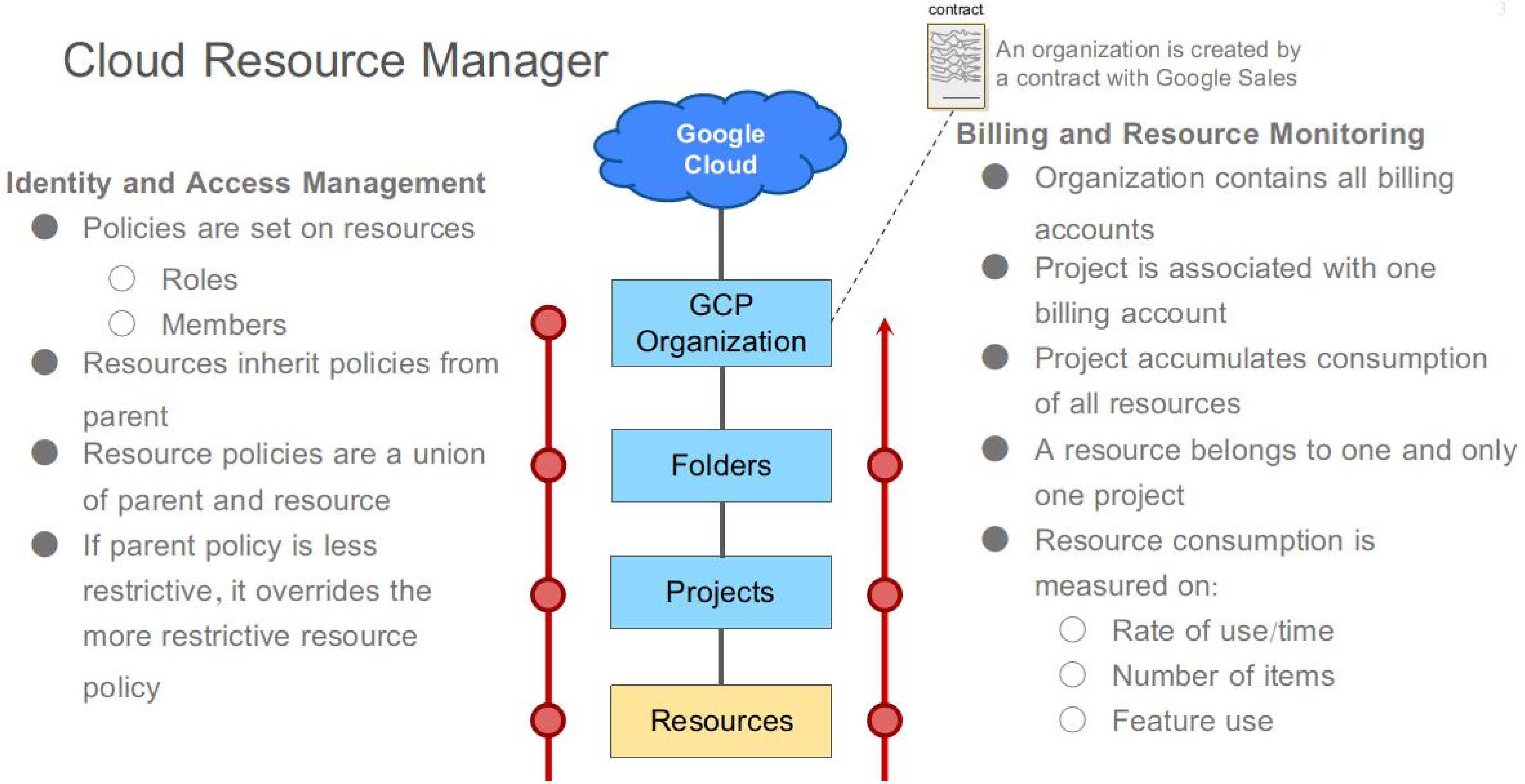
3 Years Cost Breakdown

3 Yr. Total Cost of Ownership

	On-Premises	AWS
Server	$355,679	$148,965
Storage	$155,720	$39,702
Network	$45,412	$ -
IT-Labor	$ -	$ -
Total	**$556,811**	**$188,667**

AWS cost includes business level support

COST PLANNING ON GCP



Cloud Resource Manager

Identity and Access Management

- Policies are set on resources
  - Roles
  - Members
- Resources inherit policies from parent
- Resource policies are a union of parent and resource
- If parent policy is less restrictive, it overrides the more restrictive resource policy

An organization is created by a contract with Google Sales

Billing and Resource Monitoring

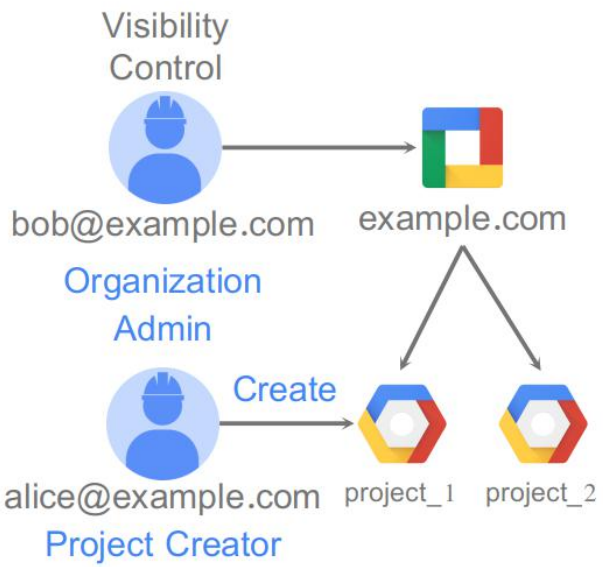
- Organization contains all billing accounts
- Project is associated with one billing account
- Project accumulates consumption of all resources
- A resource belongs to one and only one project
- Resource consumption is measured on:
  - Rate of use/time
  - Number of items
  - Feature use

ORGANIZATION NODE

Organization node is root node for Google Cloud resources

2 organization roles:

- Organization Admin: Control over all cloud resources
- Project Creator: Controls project creation



PROJECTS

Track resource and quota usage

- Enable billing
- Manage permissions and credentials
- Enable services and APIs

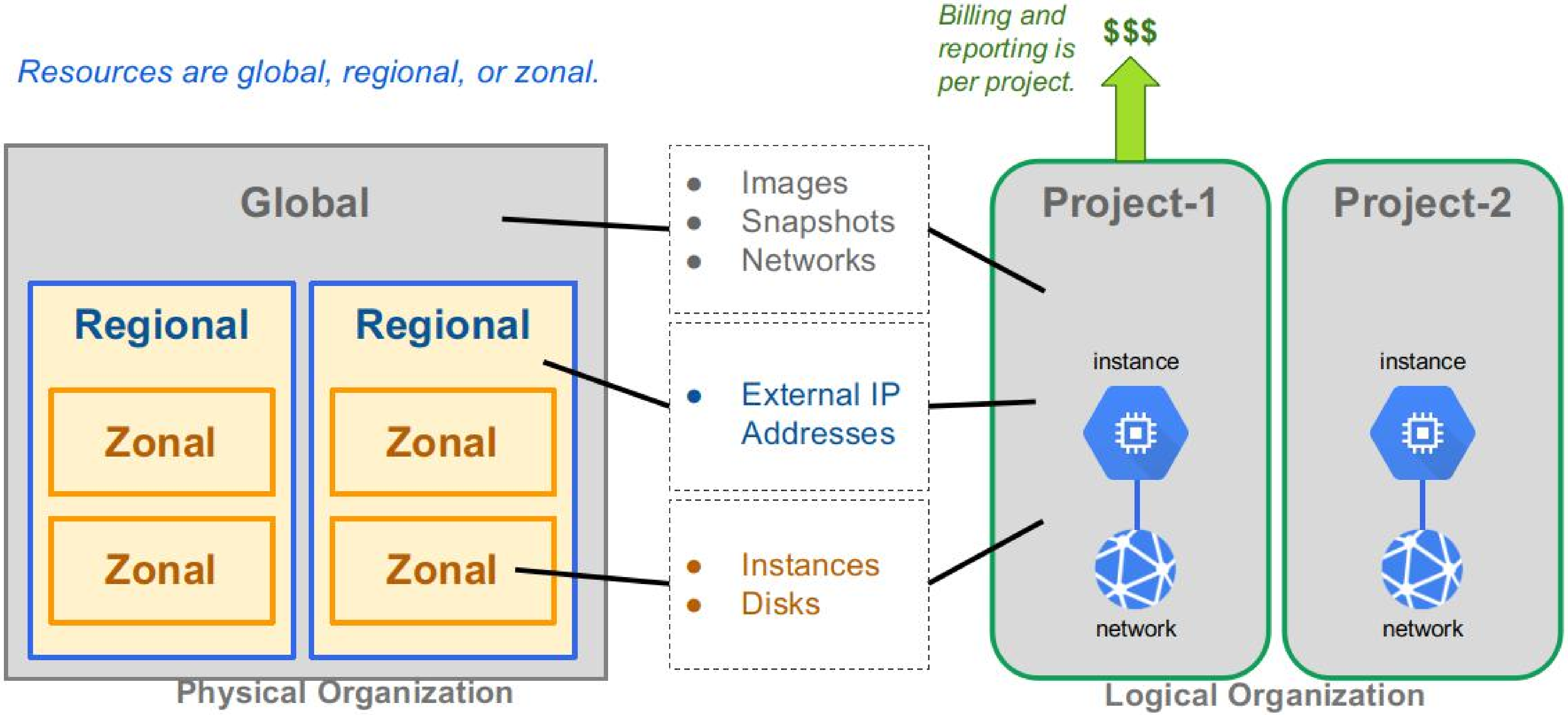
Projects use three identifying attributes:

- Project Name
- Project Number
- Project ID, also known as Application ID

Google Cloud Platform Console or the Cloud Resource Manager API

RESOURCE HIERARCHY

Resources are global, regional, or zonal.



PROJECT QUOTAS

All resources are subject to project quotas or limits.

- Typically fall into one of three categories:
    - How many resources you can create per project
    - How quickly you can make API requests in a project—rate limits
    - Some quotas are per region
- Quota examples:
    - 5 networks per project
    - 24 CPUs region/project
- Most quotas can be increased through self-service form or a support ticket
    - IAM & admin -> Quotas

WHY USE PROJECT QUOTAS?

Prevent runaway consumption in case of an error or malicious attack

Prevent billing spikes or surprises

Forces sizing consideration and periodic review

LABELS

A utility for organizing Cloud Platform resources

- Attached to resources: VM, disk, snapshot, image
- Console, gcloud or API

Example uses of labels:

- Search and list all resources (inventory)
- Filter resources (ex: separate production from test) Labels used in scripts
  - Help analyze costs
  - Run bulk operations

https://cloud.google.com/resource-manager/docs/using-labels

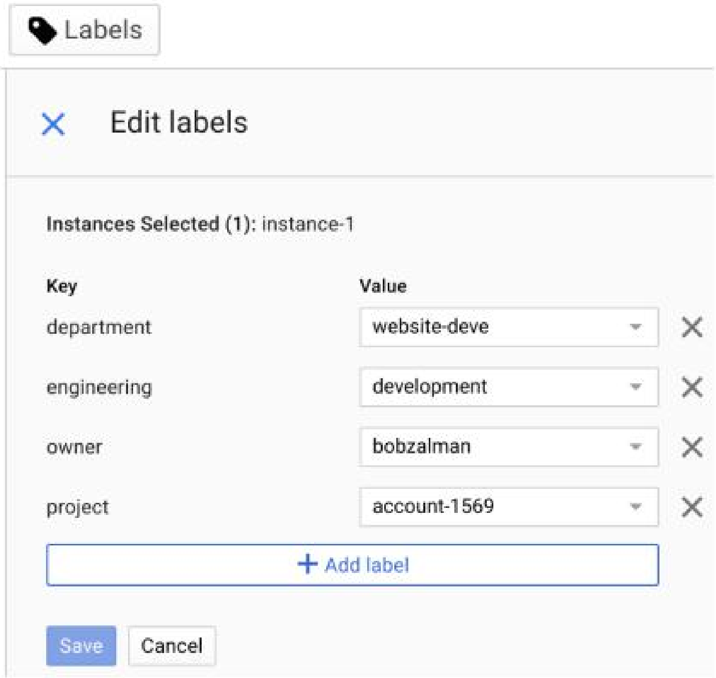
LABEL SPECIFICATION

A label is a key-value pair.

Label keys and non-empty label values can contain lowercase letters, digits, and hyphens, must start with a letter, and must end with a letter or digit. The regular expression is: a-z

The maximum length of label keys and values is 63 characters.

There can be a maximum of 64 labels per resource.



Instances Selected (1): instance-1

Key	Value
department	website-deve
engineering	development
owner	bobzalman
project	account-1569

LABEL PRACTICES

Team or Cost Center

- Distinguish projects owned by different teams.
- Useful in cost accounting or budgeting.
- Examples: `team:marketing`, `team:research`

Components

- Examples: `component:redis`, `component:frontend`

Environment or stage

- Examples: `environment:prod`, `environment:test`

Owner or contact

- Person responsible for resource or primary contact for the resource
  - Examples: `owner:gaurav`, `contact:opm`
- State
  - Examples: `state:inuse`, `state:readyfordeletion`

COMPARING LABELS AND TAGS

Labels are a way to organize resources across GCP

- disks, image, snapshots...

User-defined strings in `key-value` format

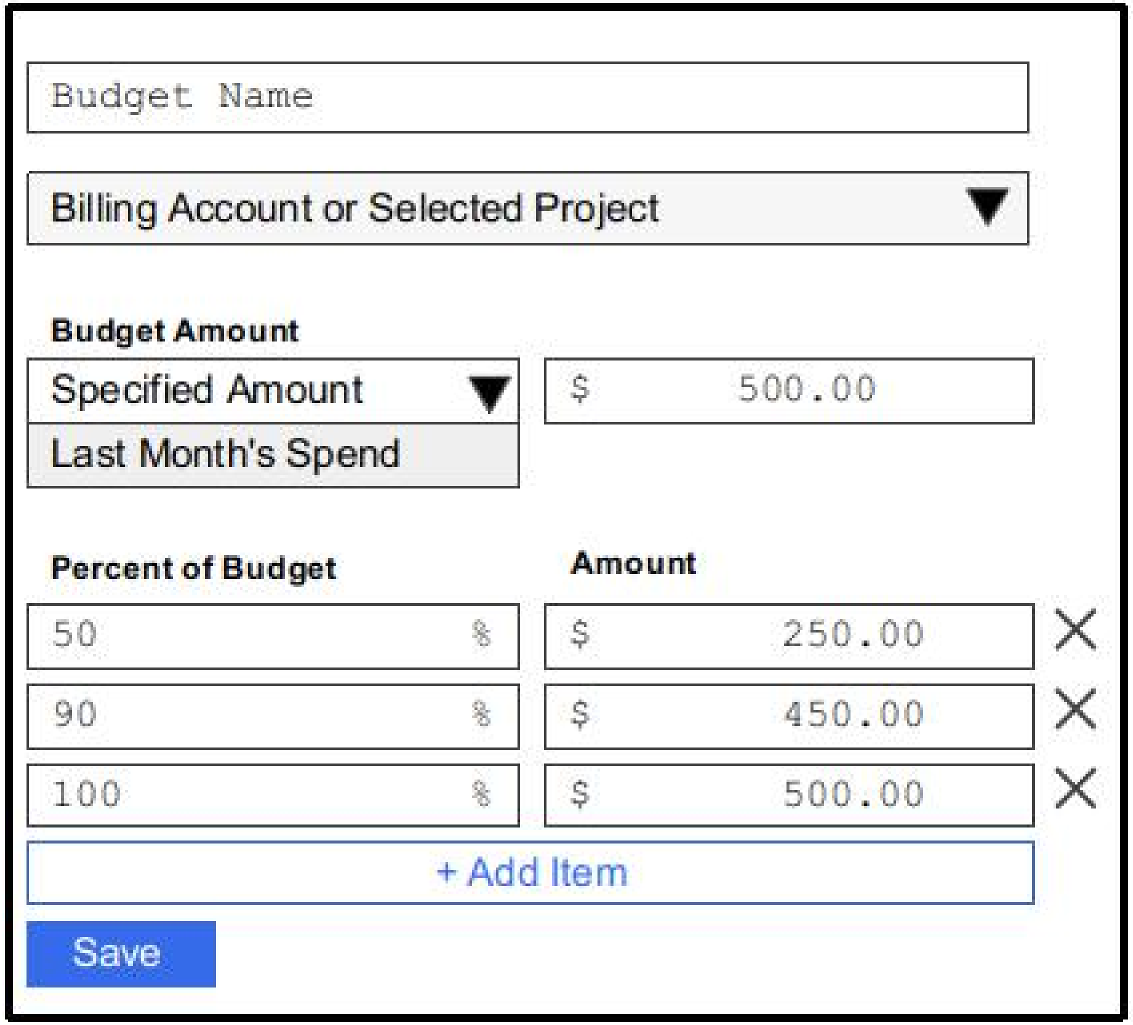
Propagated through billing

Tags are applied to instances only

User-defined strings

Tags are primarily used for networking (applying firewall rules)

BUDGETS AND ALERTS



Budget Name

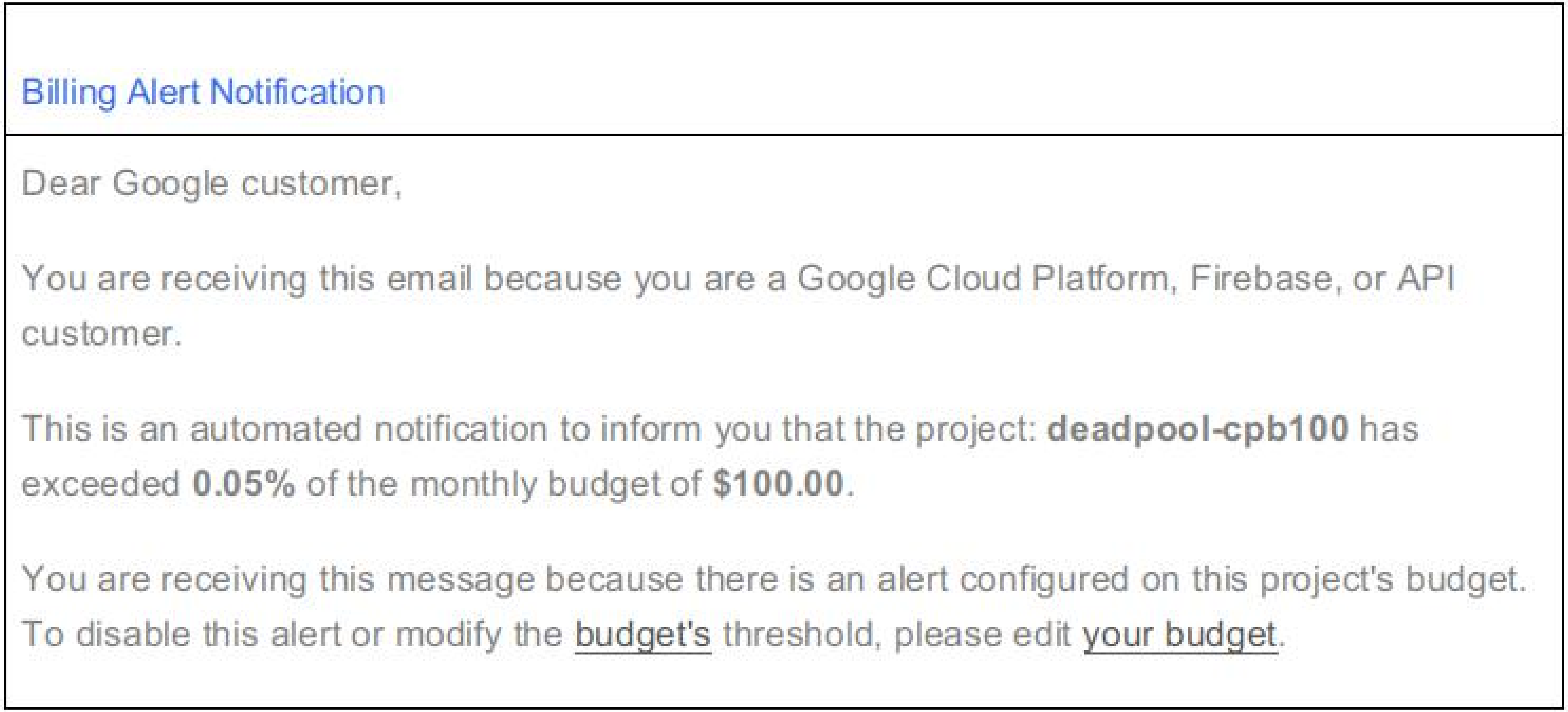
Billing Account or Selected Project

Budget Amount

Specified Amount ▼ $ 500.00

Percent of Budget	Amount
50 %	$ 250.00
90 %	$ 450.00
100 %	$ 500.00

EXAMPLE NOTIFICATION EMAIL



Billing Alert Notification

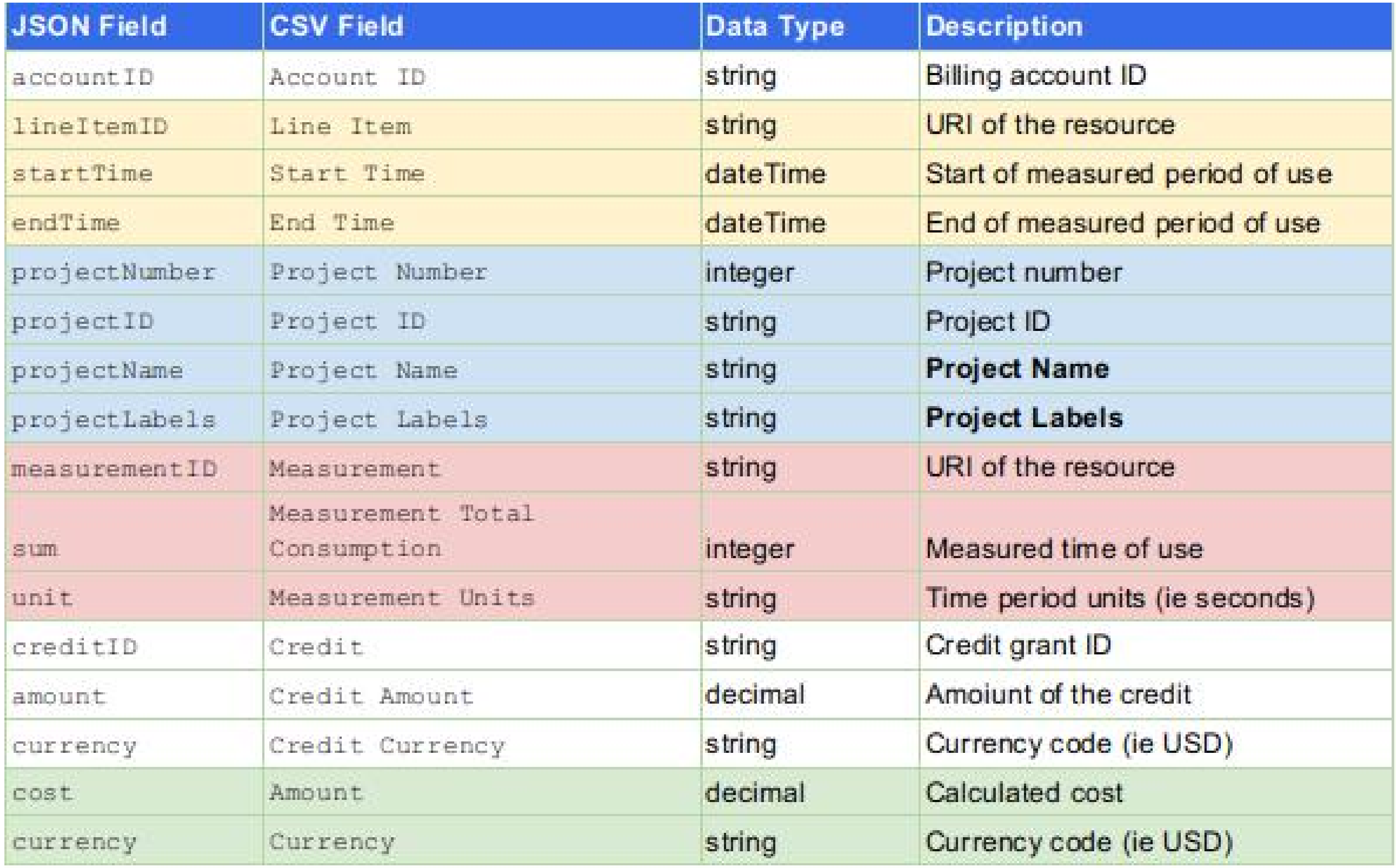
Dear Google customer,

You are receiving this email because you are a Google Cloud Platform, Firebase, or API customer.

This is an automated notification to inform you that the project: **deadpool-cpb100** has exceeded **0.05%** of the monthly budget of **$100.00**.

You are receiving this message because there is an alert configured on this project's budget. To disable this alert or modify the budget's threshold, please edit your budget.

BILLING EXPORT



JSON Field	CSV Field	Data Type	Description
accountID	Account ID	string	Billing account ID
lineItemID	Line Item	string	URI of the resource
startTime	Start Time	dateTime	Start of measured period of use
endTime	End Time	dateTime	End of measured period of use
projectNumber	Project Number	integer	Project number
projectID	Project ID	string	Project ID
projectName	Project Name	string	**Project Name**
projectLabels	Project Labels	string	**Project Labels**
measurementID	Measurement	string	URI of the resource
sum	Measurement Total Consumption	integer	Measured time of use
unit	Measurement Units	string	Time period units (ie seconds)
creditID	Credit	string	Credit grant ID
amount	Credit Amount	decimal	Amoiunt of the credit
currency	Credit Currency	string	Currency code (ie USD)
cost	Amount	decimal	Calculated cost
currency	Currency	string	Currency code (ie USD)

QUIZ

No resources in GCP can be used without being associated with...

- A. A user
- B. A virtual machine
- C. A bucket
- D. A project

QUIZ

A budget is set at $500 and an alert is set at 100%. What happens when the full amount is used?

- A. Everything in the associated project is suspended because there is not more budget to spend.
- B. A notification email is sent to the Billing Administrator.
- C. You have a 4-hour courtesy period before Google shuts down all resources.
- D. Nothing. There is no point to sending a notification when there is no budget remaining.

QUIZ

How do quotas protect GCP customers?

- A. By preventing resource use in too many zones in a region.
- B. By preventing resource use by unknown users.
- C. By preventing resource use of too many different GCP services.
- D. By preventing uncontrolled consumption of resources.

COST QUESTION #1

How do you make sure your capacity matches but does not substantially exceed what you need?

Anti-pattern

- Over-utilization
- Over-provisioning

Best practice

- Approaches:
  - Demand-based using Auto Scaling
  - Queue-based using Amazon SQS
  - Time-based using scheduling
- Appropriately provisioned

COST QUESTION #2

How are you optimizing your usage of AWS services?

Best practice

- Service-specific optimizations, such as:
  - Minimize I/O for Amazon EBS
  - Avoid uploading too many small files into Amazon S3
  - Use Spot instances extensively for Amazon EMR

COST QUESTION #3

Have you selected the appropriate resources to meet your cost targets?

Best practice

- Match your **instance profile** based on need (compute, memory, storage)
- Determine appropriate instance types using **third-party products** such as CopperEgg or New Relic
- Determine processor load using **Amazon CloudWatch**
- Load custom memory scripts and inspect memory usage using Amazon CloudWatch **custom metrics**
- **Profile your applications** to know which type of Amazon EBS to use (magnetic, general purpose (SSD), provisioned IOPS)

COST QUESTION #4

Have you selected the appropriate pricing model to meet your cost targets?

Best practice

- Use **Spot instances** for select workloads
- Perform regular **analysis of usage** and purchase Reserved Instances accordingly
- Factor in cost when choosing a **region**
- **Automate** turning off unused instances when not needed
- Sell Reserved Instances you no longer need on the **Reserved Instance Marketplace**, and purchase others

COST QUESTION #5

Are there managed services (higher-level services than Amazon EC2, Amazon EBS, Amazon S3) you can use to improve your ROI?

Best practice

- Consider **other application level services**:
  - Amazon Simple Queue Service (SQS)
  - Amazon Simple Notification Service (SNS)
  - Amazon Simple Email Service (SES)
- Achieve the benefits of **standardization** and **cost control** using:
  - AWS CloudFormation templates
  - AWS Elastic Beanstalk
  - AWS OpsWorks
- **Consider appropriate databases:**
  - Amazon RDS (PostgreSQL, MySQL, Microsoft SQL Server, Oracle, MariaDB, Amazon Aurora)
  - Amazon DynamoDB

COST QUESTION #6

What access controls and procedures do you have in place to govern AWS usage?

Best practice

- Establish **groups and roles**
  - Create environment groups and roles such as dev/test/prod
  - Use AWS governance methods such as IAM to control who can spin up instances and resources in each group
- Track, measure, and audit the **life cycle** of projects, teams, and environments

COST QUESTION #7

How are you monitoring usage and spending?

Best practice

- **Tag all resources** to be able to correlate changes in your bill to changes in your infrastructure and usage
- Have a standard process to load and interpret **Detailed Billing Reports**
- Have a plan for both usage and spending in designing a **cost-efficient architecture**
- Use **AWS Cost Explorer**
- **Monitor** usage and spend regularly using Amazon CloudWatch or a third-party provider (Cloudability, CloudCheckr)
- Set up **notifications** to let key members of your team know if your spending moves outside of defined limits.
- Use a **finance-driven charge back method** to allocate instances and resources to cost centers (such as tagging)

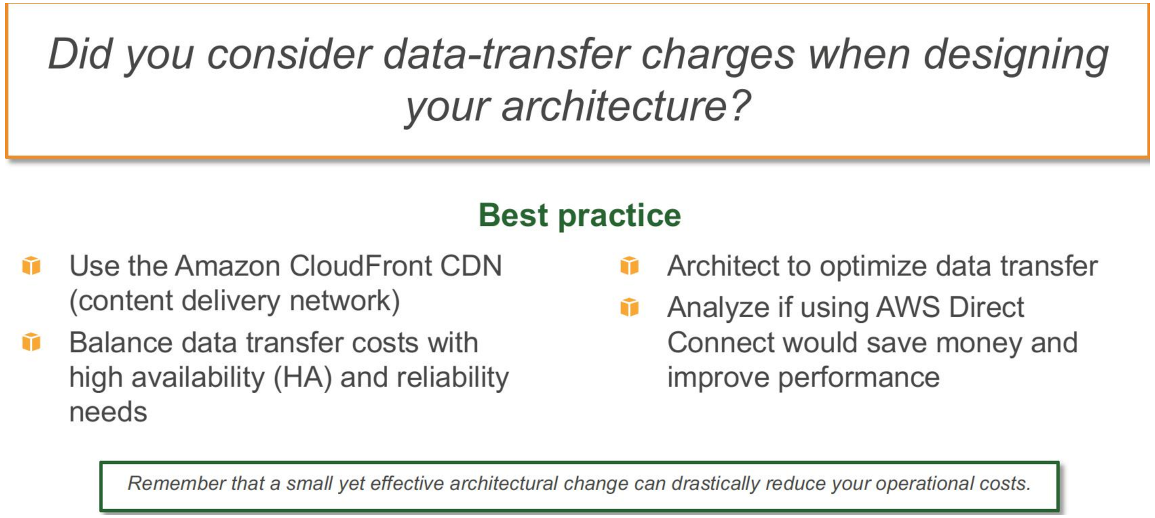
COST QUESTION #8

Do you decommission resources that you no longer need or stop resources that are temporarily not needed?

Best practice

- Design your system to **gracefully handle instance termination** as you identify and decommission non-critical or unrequired instances or resources with low utilization
- Have a process in place to **identify and decommission** orphaned resources
- **Reconcile** decommissioned resources based on either system or process

COST QUESTION #9



Did you consider data-transfer charges when designing your architecture?

Best practice

- Use the Amazon CloudFront CDN (content delivery network)
- Balance data transfer costs with high availability (HA) and reliability needs
- Architect to optimize data transfer
- Analyze if using AWS Direct Connect would save money and improve performance

Remember that a small yet effective architectural change can drastically reduce your operational costs.

Note : CDN is not for everybody. You can achieve significant improvements with S3 alone, see this Sumologic resource for a good summary

COST QUESTION #10

How do you manage and consider the adoption of new services?

- Meet regularly with you solutions architect, consultants, account team
- Consider which new services or features you could adopt to save money

CONGRATS ON COMPLETION