Wintersession 2024

3C – Controlling Cloud Costs

Check-in
For the first time in a decade, managing cloud spend (82%) surpasses security (79%) to become the number one challenge.
Cloud Waste Trends

Public cloud spend was over budget by an average of 18%, up from 13% the previous year.

Flexera State of the Cloud Report
Key Cloud Deployment Components

Cloud Computing

- Mobile
- Applications
- Database
- Compute
- Storage
- Private Cloud
- Public Cloud
- Hybrid Cloud

Compute

Storage

Applications

Database

Mobile
Cost Factors Overview

**Type:** Compute, Memory, General

**Size:** CPU/GPU and memory – i.e., large, x-large

**Purchasing Method:** On Demand, Reserved, Spot

**Region:** where instance will be launched

**Database // Serveless DB**

**Type:** Based on Database Transaction Units

**Size:** Based on the number of virtual cores

**License:** Open Source or Commercial

**Purchase Method:** (On Demand, Reserved, Spot)

**Capacity:** Provisioned read/write

**Storage:** Including indices

**Services:** Such as Data transfer or streams

**Throughput:** Reserved provisioned throughput

**Type of Storage:** Amount of redundancy and accessibility speed

**Data Transfer:** In/out Frequency and number of requests

**Quantity:** Number of files storage units requested

**Region:** Where data storage is requested

**Hidden Costs**
- Data transfer
- API requests
- Data redundancy
- Logging
- Availability

**Compute // Serverless Compute**

**Requests:** Number of executions your function will perform

**Memory:** The amount of allocated memory per function

**Time:** Average execution time

**Region:** Where function will be launched

**File Storage**

**Time:** Average execution time

**Region:** Where function will be launched
Compute Cost Factors

**TYPE**
Optimized for a particular workload - Compute, Memory, General, etc.

**SIZE**
Instance size in terms of CPU/GPU and memory i.e. large, xlarge, etc.

**REGION**
Region where instance is launched i.e. US East (Virginia)

**PURCHASING METHOD**
On Demand Instances
Reserved Instances
Spot Instances

\[ \text{Compute Cost Factors} \times \text{Quantity} \]
## Compute Purchasing Methods

<table>
<thead>
<tr>
<th>On Demand Instance</th>
<th>Full Price</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reserved Instance</strong></td>
<td>Up to 75% discount</td>
</tr>
<tr>
<td>(Convertible instance option)</td>
<td>Commitment</td>
</tr>
<tr>
<td>Payment Options (No upfront, partial upfront, all upfront)</td>
<td>1-12 months or 12-36 months</td>
</tr>
<tr>
<td><strong>Spot Instance</strong></td>
<td>Up to 90% discount</td>
</tr>
<tr>
<td>Allocation Strategy (lowest price, maximum price, or bidding)</td>
<td>Three Types:</td>
</tr>
<tr>
<td></td>
<td>1. Request</td>
</tr>
<tr>
<td></td>
<td>2. Request &amp; Maintain</td>
</tr>
<tr>
<td></td>
<td>3. Reserve for Duration</td>
</tr>
<tr>
<td>Instance Type</td>
<td>GCP</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>General Purpose</td>
<td>e2-standard-4</td>
</tr>
<tr>
<td></td>
<td>$0.134</td>
</tr>
<tr>
<td>Compute Optimized</td>
<td>c2-standard-4</td>
</tr>
<tr>
<td></td>
<td>$0.208</td>
</tr>
<tr>
<td>Memory Optimized</td>
<td>m1-ultramem-40</td>
</tr>
<tr>
<td></td>
<td>$6.293</td>
</tr>
<tr>
<td>Accelerated Computing</td>
<td>a2-highcpu-1g</td>
</tr>
<tr>
<td></td>
<td>$3.678</td>
</tr>
</tbody>
</table>
# Price – 1-year commitment (no prepayment)

<table>
<thead>
<tr>
<th>Instance Type</th>
<th>GCP</th>
<th>AWS</th>
<th>Azure</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Purpose</td>
<td>e2-standard-4</td>
<td>m6g.xlarge</td>
<td>B4MS</td>
</tr>
<tr>
<td></td>
<td>$0.0137</td>
<td>$0.097</td>
<td>$0.0974</td>
</tr>
<tr>
<td>Compute Optimized</td>
<td>c2-standard-4</td>
<td>c6g.xlarge</td>
<td>F4s v2</td>
</tr>
<tr>
<td></td>
<td>$0.0214</td>
<td>$0.086</td>
<td>$0.10</td>
</tr>
<tr>
<td>Memory Optimized</td>
<td>m1-ultramem-40</td>
<td>r6g.xlarge</td>
<td>E4a v4</td>
</tr>
<tr>
<td></td>
<td>$0.0205</td>
<td>$0.127</td>
<td>$0.1482</td>
</tr>
<tr>
<td>Accelerated Computing</td>
<td>a2-highcpu-1g</td>
<td>p2.xlarge</td>
<td>NC4as T4 v3</td>
</tr>
<tr>
<td></td>
<td>$2.313</td>
<td>$0.614</td>
<td>$0.3093</td>
</tr>
</tbody>
</table>
Serverless Cost Factors

- **REQUESTS**: Number of executions your function will perform
- **TIME**: Average Execution time for the function
- **MEMORY**: The amount of allocated memory for the function
- **REGION**: Region where function is launched i.e. US East (Virginia)

\[ \text{Serverless Cost Factors} \times \text{Quantity} \]
Serverless Compute Cost Example

Requests per Minute: 500,000 (21,600,000,000 Monthly)
  RAM: 256 MB
  Execution Time: <10ms

<table>
<thead>
<tr>
<th></th>
<th>Amazon Lambda</th>
<th>Microsoft Azure</th>
<th>Google Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>$9,000</td>
<td>$8,500</td>
<td>$10,000</td>
</tr>
</tbody>
</table>
Database Cost Factors

- **TYPE**: Based on Database Transaction Units (DTU)
- **LICENSE**: Open Source or Commercial Database license
- **SIZE**: Based on the number of virtual cores (vCores)
- **PURCHASE METHOD**: On Demand Instances, Reserved Instances, Spot Instances

\[ \text{Database Cost Factors} \times \text{Quantity} \]
Serverless Database Cost Factors

**CAPACITY**
Provisioned read/write capacity

**STORAGE**
Storage, including indices

x Quantity

**SERVICES**
Other services, such as data transfer or streams

**THROUGHPUT**
Reserved provisioned throughput
File Storage Cost Factors

**TYPE OF STORAGE**
Amount redundancy and accessibility speed

**DATA TRANSFER**
In/Out frequency and number of requests

**QUANTITY**
Number of File Storage units requested

**REGION**
Region where data storage is requested i.e. US East (Virginia)
Cost Saving Strategies

**RIGHT SIZE**
Right Size the machine for the job

**SPRAWL CONTROL**
Control sprawl of virtual machines

**SHUTDOWN SCHEDULE**
Control running hours for resources

**DISCOUNTS**
If discount available, use it!

**EXPIRATION DATE**
Monitor & delete expired instances

**TAGS**
Keep track of resources with logical tags
Cost Saving Strategies

**ORPHAN CHECK**
Check for orphaned storage and lingering snapshots

**ALERTS**
Alerts and alarms notify you of reaching budget constraints

**LOGGING**
Log user activities for audit and cross reference

**BUDGETS**
Utilize budgets that will ensure you won’t overspend

**LIFECYCLE MGMT**
Automatically move data to a less expensive storage or purge data

**AUTOMATE**
Use Templates to stand-up/bring down stacks
Cost Optimization Process

Cloud Cost Optimization

Right Size

Analyze Usage

Monitor Cost

Config Schedule
## Resources

<table>
<thead>
<tr>
<th>AWS</th>
<th>Azure</th>
<th>Google Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://aws.amazon.com/tco-calculator/">https://aws.amazon.com/tco-calculator/</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Resources

Presenter

ik8@princeton.edu
Irene Kopaliani, PHD, C|CISO
CISA, CISM, CDPSE, CISSP, CCSP

Research Computing

https://researchcomputing.princeton.edu/systems/cloud-computing

OIT Cloud Accounts

https://princeton.service-now.com/service?id=sc_cat_item&table=sc_cat_item=sys_id=06268c7c1bc444d098d1217e6e4bcb4f