04 Pricing and Billing -Q&A - KirkYagami

1. Sustained Use Discounts vs Committed Use Discounts

Q: Explain the difference between sustained use discounts and committed use discounts in GCP. When would you recommend using each?

A:

- Sustained Use Discounts:
  - Automatically applied discounts for running specific Compute Engine resources for a significant portion of the billing month.
  - No upfront commitment required.
  - Discount increases with usage, up to a 30% net discount for instances running the entire month.
  - Recommended for workloads with consistent but unpredictable usage patterns.
- Committed Use Discounts:
  - Require a 1-year or 3-year commitment to use a specific amount of resources.
  - Offer larger discounts, up to 57% for most machine types.
  - Recommended for workloads with predictable, long-term resource needs.

2. Per-Second Billing

Q: How does GCP's per-second billing work, and for which services does it apply? Are there any exceptions?

A:

- Per-second billing charges for compute resources by the second, with a minimum of 1 minute.
- Applies to:
  - Compute Engine
  - Kubernetes Engine
  - Cloud Run
  - App Engine flexible environment VMs
- Exceptions:
  - Premium OS licenses are billed by the minute
  - Sustained Use Discounts are calculated hourly

- Benefits include more accurate billing and cost savings for short-lived resources.

3. Resource-Based Pricing Model

Q: Describe the purpose and functioning of GCP's resource-based pricing model. Give an example of how this applies to a specific GCP service.

A:

- Purpose: To provide flexible and granular pricing based on actual resource consumption.
- Functioning: Charges are calculated based on specific resources used (e.g., CPU, memory, storage) rather than predefined instance types.
- Example with Cloud Run:
  - Billed precisely for:
    - CPU time used to execute requests
    - Memory allocated while requests are running
    - Networking for handling requests and responses
  - Only pay for resources consumed during request processing, not for idle time.

4. Custom Machine Types

Q: What is a custom machine type in GCP, and how might it help optimize costs compared to predefined machine types?

A:

- Custom machine types allow you to create VM instances with specific amounts of vCPUs and memory.
- Cost optimization:
  - Avoid paying for unused resources by tailoring the VM to exact needs.
  - Can be cheaper than the next larger predefined machine type if you need just slightly more resources.
  - Allows for fine-tuning performance and cost for specific workloads.
- Example: Instead of choosing between an n1-standard-4 (4 vCPUs, 15 GB memory) and n1-standard-8 (8 vCPUs, 30 GB memory), you could create a custom instance with 6 vCPUs and 23 GB memory if that better fits your needs.

5. Egress Charges

Q: Explain the concept of "egress charges" in GCP. How can you minimize these charges when designing a multi-region application?

A:

- Egress charges are fees for data transferred out of GCP services or between GCP regions.
- Strategies to minimize charges:
  - Use Cloud CDN to cache content closer to users.
  - Optimize data transfer between regions (e.g., use regional endpoints for Cloud Storage).
  - Use the same region for interdependent services where possible.
  - Consider using Network Service Tiers (Standard tier is cheaper for some egress traffic).
  - Compress data before transfer.
  - Use GCP's private network for transfer between GCP services when possible.

6. GCP Pricing Calculator

Q: What is the purpose of GCP's pricing calculator? Walk through the process of estimating costs for a simple three-tier web application using this tool.

A:

- Purpose: To estimate costs for GCP resources and services before deployment.
- Process for a three-tier web application:
  1. Open the GCP Pricing Calculator
  2. Add Compute Engine instances for the web tier:
     - Select instance type, number of instances, and usage time
     - Add load balancer
  3. Add Cloud SQL for the database tier:
     - Select database type, instance size, and storage
  4. Add App Engine or Kubernetes Engine for the application tier:
     - Estimate instance hours and memory usage
  5. Include networking costs:
     - Estimate egress traffic
  6. Add any additional services (e.g., Cloud Storage, Cloud CDN)
  7. Review and adjust the estimated costs
  8. Save or share the estimate

7. Billable vs Non-Billable Projects

Q: Describe the difference between billable and non-billable projects in GCP. How does this impact organization-level billing?

A:

- Billable projects:
  - Can use and be charged for GCP resources
  - Must be associated with an active billing account
- Non-billable projects:
  - Cannot use paid GCP resources
  - Can only use free services or quotas
  - Typically used for testing or learning purposes
- Impact on organization-level billing:
  - Only billable projects contribute to the organization's overall costs
  - Non-billable projects help separate test environments from production
  - Allows for better cost allocation and budget management across different teams or departments

8. GCP Free Tier

Q: What are GCP's "free tier" offerings, and what are the limitations? How should these be considered when planning a new project?

A:

- GCP Free Tier includes:
  1. Always Free: Certain amounts of many common resources are always free to use
  2. Free Trial: $300 credit for new customers to use for 90 days
- Limitations:
  - Always Free resources have specific usage limits
  - Some services are excluded from the free tier
  - Free trial requires credit card for verification
- Considerations for new projects:
  - Use free tier resources for initial development and testing
  - Plan to scale beyond free tier limits for production workloads
  - Be aware of which services are not covered by the free tier

- Set up billing alerts to avoid unexpected charges when transitioning from free tier

9. Budgets and Alerts

Q: Explain how GCP's budgets and alerts work. How would you set up a budget alert system for a project with variable monthly spending?

A:

- GCP Budgets and Alerts:
  - Allow setting spending thresholds for projects or billing accounts
  - Can trigger notifications when spending reaches defined percentages of the budget
- Setting up for variable spending:
  1. Analyze historical spending patterns to establish a baseline
  2. Set up a budget slightly higher than the average monthly spend
  3. Create alerts at multiple thresholds:
     - 50% of budget: Early warning
     - 80% of budget: Review and potential action required
     - 100% of budget: Immediate attention needed
  4. Use rolling budgets to account for month-to-month variations
  5. Set up email and SMS notifications for key stakeholders
  6. Consider automating responses (e.g., disabling new resource creation) at critical thresholds

10. Labels for Cost Allocation

Q: What is the purpose of labels in GCP, and how can they be used effectively for cost allocation and tracking? Provide an example of a labeling strategy for a medium-sized enterprise.

A:

- Purpose of labels:
  - Organize and categorize resources for better management and billing analysis
  - Attach metadata to resources for filtering and reporting
- Effective use for cost allocation:
  - Apply consistent labels across all resources

    - Use labels to identify project, environment, department, cost center, etc.
    - Create detailed billing reports based on label combinations
- Example labeling strategy for a medium-sized enterprise:

    1. environment: prod, dev, test, staging
    2. department: marketing, sales, engineering, finance
    3. project: website, mobile-app, data-warehouse
    4. cost-center: cc-1234, cc-5678
    5. owner: email@company.com
    6. confidentiality: public, private, sensitive

    This strategy allows for detailed cost breakdowns and helps identify areas for optimization across different business units and projects.