Cloud Foundry AutoScaler Service: Features, Design and Roadmap

Julian Friedman
Ying Liu
Pradyut Sarma
Manage Application Capacity with Auto-Scaling

Without Auto-Scaling

• Monitor application
• Change instance manually
  • cf scale MYAPP –i 3
  • cf scale MYAPP –i 1
• Keep monitoring …

With Auto-Scaling

• Create a service instance
• Define your policy
• Bind the service
Dynamic Scaling

Without Auto-Scaling:

• Memory Used
• Memory Utilization
• Throughput
• Response Time
Dynamic Scaling

With Auto-Scaling:

- Memory Used
- Memory Utilization
- Throughput
- Response Time
Scheduled Scaling

- **Recurring Schedule**
  - Day-of-Week , Day-of-Month
  - Start time ~ End time

- **Specific Date Schedule**
  - Start date / time ~ End date / time

- **Dynamic scaling + scheduled scaling**

---

**Instance changes with Scheduled + Dynamic Scaling**

- max instance count
- initial min instance count
- min instance count
Policy definition

```json
{
    "instance_min_count": 1,
    "instance_max_count": 4,
    "scaling_rules": [
        {}
    ],
    "schedules": {
        "timezone": "Asia/Shanghai",
        "recurring_schedule": [
            {}
        ],
        "specific_date": [
            {}
        ]
    }
}
```

<table>
<thead>
<tr>
<th>Metric_type</th>
<th>stat_window_secs</th>
<th>Breach_duration_secs</th>
<th>threshold</th>
<th>operator</th>
<th>adjustment</th>
<th>cool_down_secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput</td>
<td>300</td>
<td>600</td>
<td>8000</td>
<td>&lt;</td>
<td>-1</td>
<td>300</td>
</tr>
<tr>
<td>Throughput</td>
<td>300</td>
<td>600</td>
<td>15000</td>
<td>&gt;</td>
<td>+1</td>
<td>300</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>start_time</th>
<th>end_time</th>
<th>Day_of_week/Day_of_month</th>
<th>instance_min_count</th>
<th>initial_min_instance_count</th>
<th>instance_max_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00</td>
<td>18:00</td>
<td>1,3,5</td>
<td>3</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>18:00</td>
<td>23:00</td>
<td>1,2,3,4,5,6,7</td>
<td>5</td>
<td>8</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>start_date_time</th>
<th>end_date_time</th>
<th>instance_min_count</th>
<th>instance_max_count</th>
<th>initial_min_instance_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-10-11T10:00</td>
<td>2017-10-13T10:00</td>
<td>5</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>2017-12-23T10:00</td>
<td>2017-12-26T10:00</td>
<td>5</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>
# Command Line

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>autoscaling-api, asa</code></td>
<td>Set or view AutoScaler service API endpoint</td>
</tr>
<tr>
<td><code>autoscaling-policy, asp</code></td>
<td>Retrieve the scaling policy of an application</td>
</tr>
<tr>
<td><code>attach-autoscaling-policy, aasp</code></td>
<td>Attach a scaling policy to an application</td>
</tr>
<tr>
<td><code>detach-autoscaling-policy, dasp</code></td>
<td>Detach the scaling policy from an application</td>
</tr>
<tr>
<td><code>autoscaling-metrics, asm</code></td>
<td>Retrieve the metrics of an application</td>
</tr>
<tr>
<td><code>autoscaling-history, ash</code></td>
<td>Retrieve the scaling history of an application</td>
</tr>
</tbody>
</table>

- https://github.com/cloudfoundry-incubator/app-autoscaler/blob/develop/src/cli/readme.MD
Steps to implement Auto-Scaling:

- Understand the traffic and workload type of the application
- Benchmark the application to understand the performance of the application
- Scale application manually to understand how the application behaves when scaling out (how long it is needed to warm up, what is the impact of existing load and sticky session if it is used)
- Identify the performance bottleneck from the benchmarking result, and decide which metric should be used to dynamically adjust the instance number.
- Define the initial scaling policy and enable Auto-Scaling service
- Simulate the workload and test with Auto-Scaling, to adjust detailed settings of the policy, including thresholds, steps of scaling, statistics window, breach duration and cool-down period
- Define scheduled scaling for peak hours
- Simulate the peak hour workload, and adjust the min/max instance number settings in the scheduled policy
- Apply the refined policy and let it go
Deep Dive to AutoScaler
Deploy AutoScaler

• Prerequisite:
  • Install Cloud Foundry
  • Download AutoScaler bosh release:
    git clone https://github.com/cloudfoundry-incubator/app-autoscaler-release

• Deploy with bosh v2 or cf-deployment
  • Bosh v2 manifest
    • bosh update cloud-config ./example/cloud-config.yml
    • ./scripts/generate-bosh-lite-manifest -c <path to cf-release deployment manifest> -p ./example/property-overrides.yml
    • ./scripts/deploy
  • CF-deployment
    • bosh -e YOUR_ENV -d app-autoscaler deploy templates/app-autoscaler-deployment.yml --vars-store=bosh-lite/deployments/vars/autoscaler-deployment-vars.yml -v system_domain=bosh-lite.com -v cf_admin_password=<cf admin password>
Enable AutoScaler Service

- Register AutoScaler service
  - cf create-service-broker autoscaler <brokerUserName> <brokerPassword> <brokerURL>
  - cf enable-service-access autoscaler

- Now, here it is:

<table>
<thead>
<tr>
<th>service</th>
<th>plans</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>autoscaler</td>
<td>autoscaler-free-plan</td>
<td>Automatically increase or decrease the number of application instances based on a policy you define.</td>
</tr>
</tbody>
</table>
Architecture Diagram

create/bind service

Cloud Controller

Service Broker

Scaling Engine

Scheduler

Aggregator/
EventGen

metric collector

Loggreator

Cloud Foundry

Postgres

Raw events

Scaling metrics

Trigger scheduled scaling

Trigger dynamic scaling

• Define policy
• Retrieve metric
• Retrieve history

Autoscaler
Autoscaler on SAP Cloud Platform

- BETA release during SAPPHIRE (July 2017)
  - See strong internal/ramp-up customer adoption for varied use cases

- Multi-Cloud strategy
  - AWS
  - Azure
  - OpenStack
  - GCP (....is next)

- Performance tests

- Preparing to go GA on all platforms
Coming Soon …
Thank you!

- **GIT:**
  - [https://github.com/cloudfoundry-incubator/app-autoscaler](https://github.com/cloudfoundry-incubator/app-autoscaler)
  - [https://github.com/cloudfoundry-incubator/app-autoscaler-release](https://github.com/cloudfoundry-incubator/app-autoscaler-release)

- **Slack:** #autoscaler on cloudfoundry.slack.com
  - [https://cloudfoundry.slack.com/messages/autoscaler/](https://cloudfoundry.slack.com/messages/autoscaler/)

- **Email:** cf-dev@lists.cloudfoundry.org