AliYun and Infrastructure as Code

Jan 2015

Infrastructure as Code

- With Cloud APIs, datacenters become programmable:
  - servers
  - load balancers
  - networks
- The “product” encompasses the application, the platform, and the servers.
- All can be managed as a software project

AliYun Enhancements are Enablers

- New AliYun features are critical to our ability to deliver:
  - Faster
  - More reliably
  - To a wider audience

Here's the top 10...

Availability Zones

- Control which AZ an instance launches in
- Utilize 3 AZ's concurrently for new instance creation
- Huge win in **reliability**

Suggested Action: Open 3 AZs for users to create instances into.

- cn-hangzhou-a
- cn-hangzhou-b
- cn-hangzhou-c

Local Instance Store (SSD)

- Flexible storage options
- High I/O performance at the cost of losing data during reboot
- Allow us to use disk as a temporary work space, not permanent data storage

Requested Action: make the documented “ephemeral-hio” disk type available in all regions and zones.

Tags

- We need tags for grouping our instances into clusters and environments
- We implement tags for ourselves, and most large customers probably do too
- Tags on ECS and SLB instances are the minimum

Requested Action: make an API feature for each model to support tags.

```
Action=CreateInstance&Tag.1.Name=role&Tag.1.Value=web
```

or

```
Action=CreateInstance&Tags='{["role","app"]}'
```

ESS Scaling Alarms

- ESS currently only supports scaling up or down based on CPU or RAM
- SLB latencies, and custom 监控 metrics like application task queue length
- Very limited statistics are available: mean/max/min of 2, 5, or 15 minutes

Requested Action: allow any 监控 metric to be useable as a Scaling Alarm metric. Allow statistic to be calculated over any time span. Allow for sum and count statistics as well.

Noisy Neighbors

- We see evidence of physical hosts SAN access being over-burdened
- We re-create instances we suspect of having greedy neighbors
- This costs a lot of time to track down and replace

Requested Action: Enforce SAN connectivity quotas and caps more strictly

or

Indicate to users when a VM on the same physical host is bursting above the cap

Prepaid Products and Vouchers

- Obtaining, applying, and tracking vouchers is difficult
- Our deliveries have been delayed waiting on AliPay vouchers, activation, and purchasing of prepaid products.
- Minimum 24 hour turnaround from statement of need to product in-use.

Requested Action: Continue disabling requirement of positive account balance. An AliPay account for AliYun which Quixey Beijing can manage and fund in emergencies.

Schedule Maintenance, Not Downtime

- AliYun API maintenance announcements are good!
- Disabling the service during maintenance is not.
- Cloud APIs should be as reliable as the cloud itself.

Requested Action: Do maintenance with elevated risk of downtime, but not planned downtime.

Upgraded Instance Types

- AliYun instance types are limited
- Many low CPU and RAM choices

Requested Action: Add an instance family with 8 CPUs and up to 96GB RAM. Add an instance family with 1Gbps network but only up to 12GB RAM. Allow customers to adjust the resources they need.

Access Control

- Every AliYun API key has full account access
- This is **scary**
- We have no way of restricting to least-privileges for AliYun API access
- RAM looks promising, but requires many new AliYun accounts

Requested Action: Allow API keys within one account to have controlled access to AliYun APIs.

AliYun API Python SDK

- We use boto for AWS automation
- **A lot**
- The Python SDK provided by AliYun does not fit our needs

Requested Action: Create and maintain a higher-level Python SDK for AliYun's APIs.

or

Adopt or help maintain our open-sourced AliYun API Python SDK.