Openstack Swift at Scale

OPENSTACK SUMMIT 2013
HONG KONG

rackspace®
the open cloud company

In the Beginning

- Cloud Files 2.0
- 5 developers
- 4 ops
- 9 months
- 10K lines of code
- Openstack!

Openstack Swift at Rackspace

- 6 Datacenters
- More than 85 Petabytes of raw disk
- Over half a trillion requests since release
- 60 Gb sustained peaks of throughput to a single cluster
  - Over 200 Gb for backend services

The Original Goal

- 100 petabytes of data
- 100 billion objects
- 100 gigabits per second throughput
- 100 thousand requests per second

Use Cases

- Internal and external
- Backup
- Media
- CDN
- Logs

Swift as a Complete System

- Openstack Swift Software
- Hardware
- Network
- Monitoring

Hardware

- Then: 24 1TB drives per box, 1G network
- Now: 90 3TB drives per box, 10G network
- SSD drives for Account/Containers
- Commodity SATA drives
- Test, test, test

Network

- Then: 1GB to host
- Now: 10GB to host
- Network aggregation
- Haproxy for load balancing

Monitoring

- The usual suspects
- Error log lines
- Replication times
- Dispersion report
- Async pendings

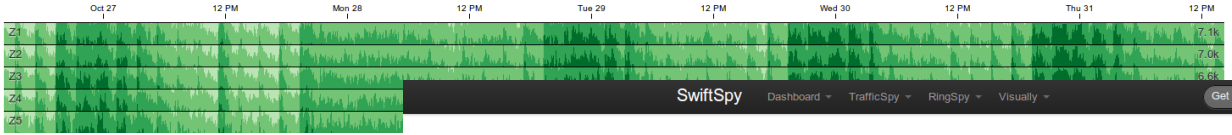
Extra Tools

- swift ring master
  (https://github.com/pandemicsyn/swift-ring-master)
- swift stalker
  (https://github.com/pandemicsyn/stalker)
- graphite/statsd/swift-informant
  (https://github.com/pandemicsyn/swift-informant)
- swift spy

SwiftSpy Dashboard TrafficSpy RingSpy Visually Get Nodes Login

Per zone total writes.

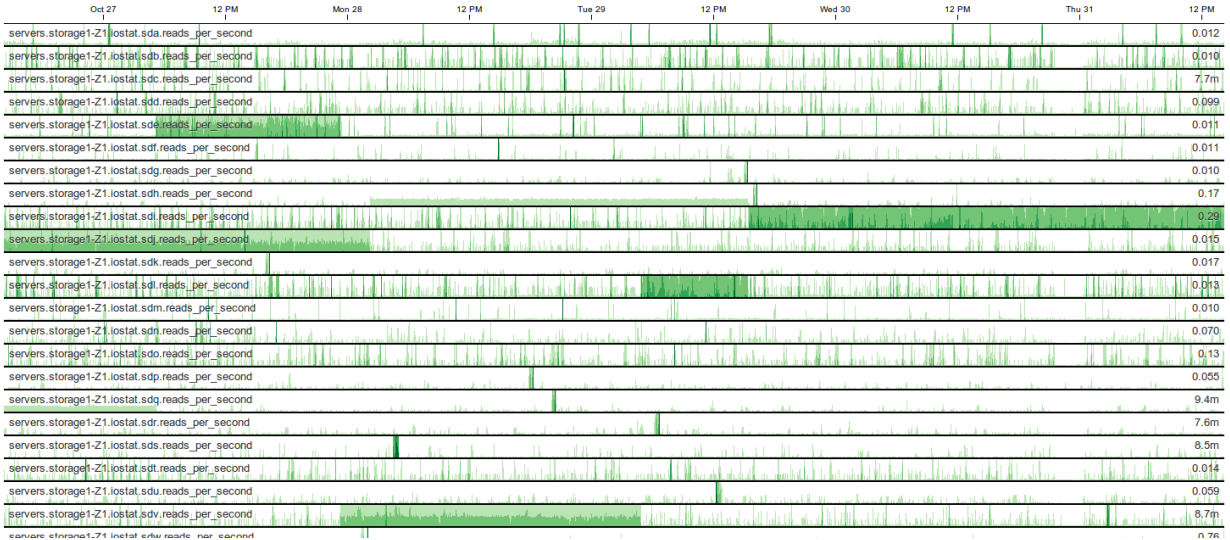
Available graphs: loadavg iops iowait reads writes txbit rxbit gigabyte_used gigabyte_avail service_time util_percentage average_queue_length await

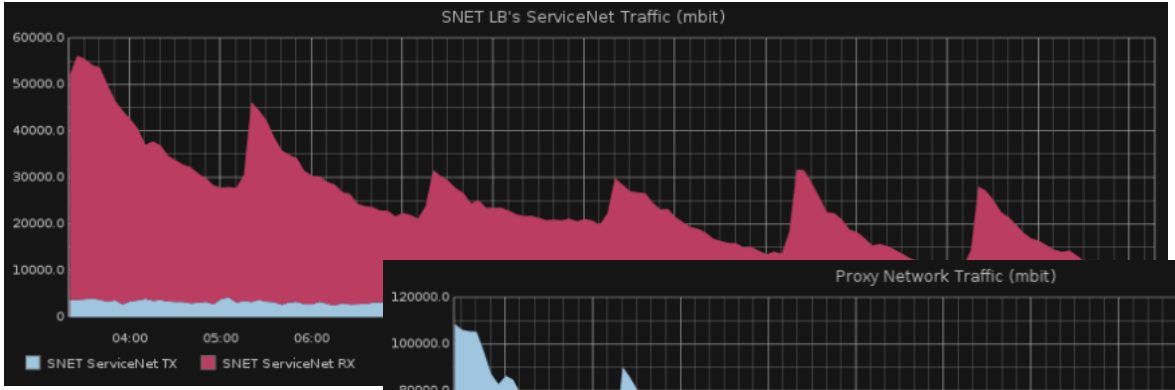


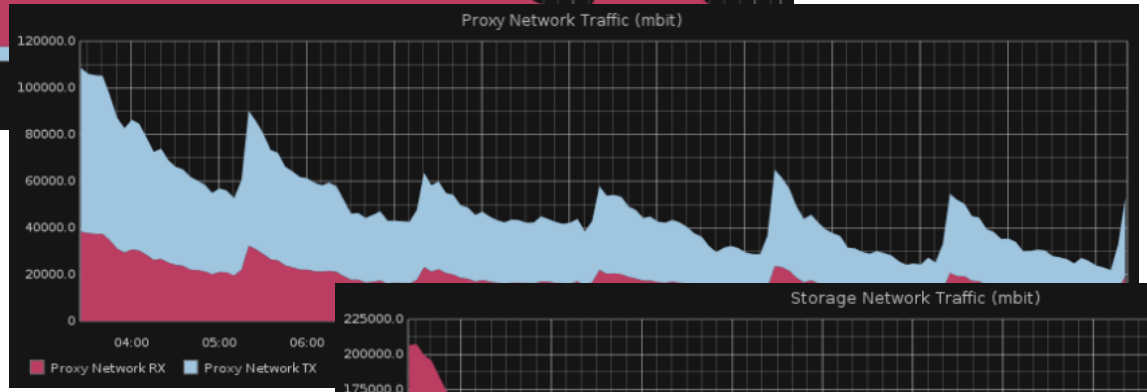
SwiftSpy Dashboard TrafficSpy RingSpy Visually Get Nodes Login

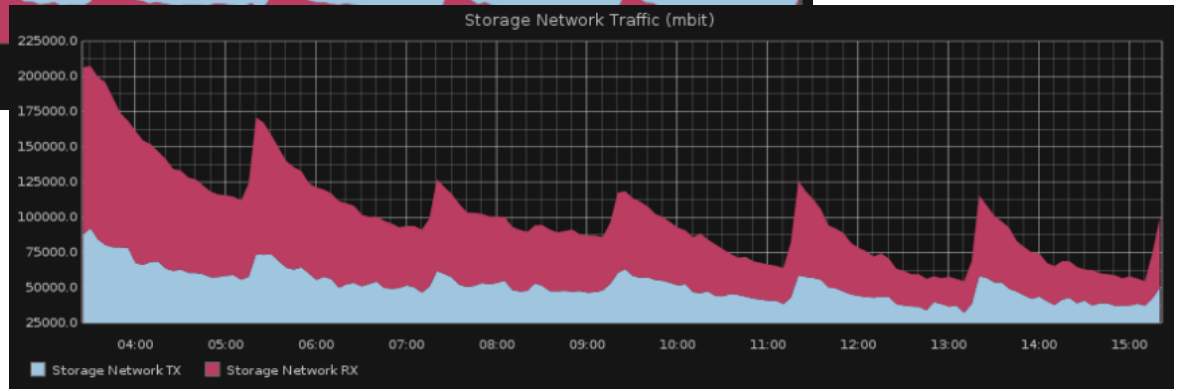
reads for storage1-Z1

Available graphs: loadavg iops iowait reads writes txbit rxbit gigabyte_used gigabyte_avail service_time util_percentage average_queue_length await




SNET LB's ServiceNet Traffic (mbit)
60000.0
50000.0
40000.0
30000.0
20000.0
10000.0
0
04:00
05:00
06:00
SNET ServiceNet TX
SNET ServiceNet RX


Proxy Network Traffic (mbit)
120000.0
100000.0
80000.0
60000.0
40000.0
20000.0
0
04:00
05:00
06:00
Proxy Network RX
Proxy Network TX


Storage Network Traffic (mbit)
225000.0
200000.0
175000.0
150000.0
125000.0
100000.0
75000.0
50000.0
25000.0
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
Storage Network TX
Storage Network RX

The Road Ahead

- Better replication
- Better handling of full disks
- Better error handling/limiting
- Container sync

Thank You!

Chuck Thier

Principal Engineer, Rackspace

chuck.thier@rackspace.com

@creiht