C* Long Term Road Test

Kavika Vollmar & Chris Romary

Introductions

- Christopher Romary
- David “Kavika” Vollmar
- Gnip / Twitter
- Sponsor: DataStax

Outline

- Current C* Usage / Size / Configurations
- Issues
- Best practices
- Questions

Background

- 4 C* clusters across at least 2 environments (Production, staging, optional review)
- Prod Configurations
  - Tango 16 nodes, 7 of 40 TB
  - Uniform 20 nodes, 6 of 20 TB
  - Yankee 20 nodes, 18 of 33 TB
  - Whiskey 8 nodes, 0.3 of 10

Different node and drive types

- EC2 and bare metal machines
- Using ephemeral and EBS volumes in EC2
- Dedicated SSDs on bare metal

Versions

Cassandra 1.2

Astyanax 1.56

CentOS 6.4 (and now 7)

Access Patterns

- Very different access patterns
  - Rows written once versus continuously updated
  - Fairly stable size (e.g. 30 days of data) vs continuously growing
  - Read once versus reading same rows over and over again

Customer Compliance Use Case

- Ever Expanding Data Set
- Lots of small rows
  - 1 row per tweet, 2-5 columns of longs per row
- Heavy Write Load
  - 10.000 writes / second
- Heavy Read Load
  - 3.000 reads / second
  - Random read distribution

Picking an Instance Size

- Ephemeral (with spinning disks) did not provide required IOPS
- Moved to provisioned IOPS
- Picked most inexpensive host type with presumed sufficient IOPs and cores

VNodes

- VNodes
  - Easy to add/remove nodes
  - Lose any 2 nodes and your cluster is down (depending on R/F)
- Token Rings (Single Token)
  - We don’t recommend it
  - Hard to add more capacity

Memory Usage

- Bloom Filters
- Row Cache
- Key Cache
- Mem Tables
- Indexes

Bloom Filters

- Live Off Heap
- Grow proportional to data size
- % Chance to avoid hitting disk
- Can Cause OOM

Memory Settings

Small amount of RAM used by Heap

Let OS manage paging/caching

Tune off heap memory to prevent OOM

Seed Configuration

- Point of contact for the cluster
- Only important when joining the cluster
- Test behavior with seeds across Data Centers

IOPs Throughput issues

- Did not achieve expected throughput
- Usually performed up to expectations
- New nodes (sometimes) had issues

More on IO

- iostat
  - best case showed expected 3000 r/sec
  - worst case showed < 1000 r/sec
- block size confusion
  - Did one os “read” correspond to 1 EBS iop?

Pre-warming volumes

- Read each sector to make sure it’s allocated
- Recommended By Amazon
- Wasn’t needed in most cases
- Takes several hours so do it in advance

Migration to new hardware

- Newer Instance Types
  - Newer Hardware
  - Less maintenance time
- Bigger Instance Types
  - 2x RAM
  - Ephemeral SSDs

Actually Migrating

Several ways to do it

- Multiple data centers
- Add/remove nodes

Not Always Smooth

- Relationship to consistency settings
- Test Config Settings
- Load on Prod System

Data Per Node

More Data per Node:

- Costs Less

But…

- Repairs Take Longer
- Adding/Removing nodes takes longer
- Everything takes longer

General Best Practices (1)

- Generation of config files
- Disk space alerts at 40%
- Script Everything
  - Spin up
  - Repairs
- Lock down prod with different credentials
  - Securing EC2 in general is a much larger topic

General Best Practices (2)

- Verify client settings and configurations
- Always use the same KPI’s
- Performance testing with separate applications

General Best Practices (3)

Have a “staging” system that mirrors prod as close as possible

Test all changes in staging first

Questions?

The Future...

How far will this scale? As...

- The size of the data grows
- The number of nodes in the cluster grows

What will the next AWS issue be?