

Indexer Clustering – Internals & Performance

Da Xu

Software Engineer, Splunk

Chloe Yeung

Software Engineer, Splunk

.conf2016

splunk >

Disclaimer

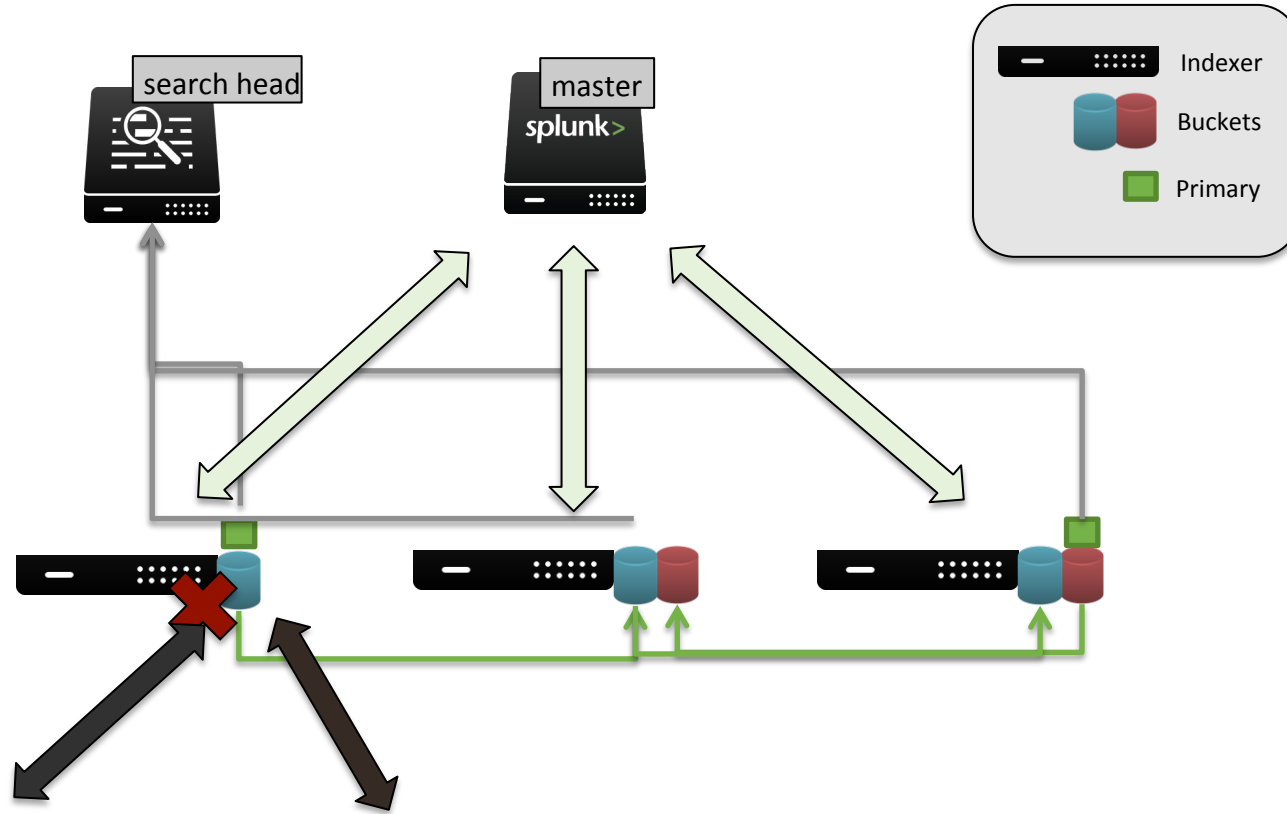
During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Indexer Clustering Overview



.conf2016

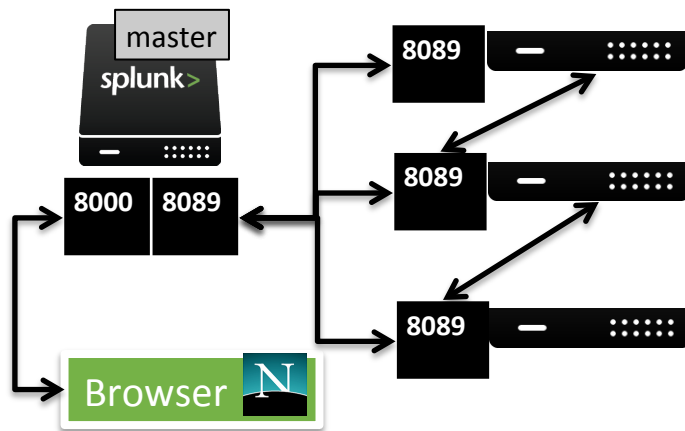
Cluster!



Communication Through Endpoints

The cluster master and peers communicate amongst themselves through the clustering endpoints on the management ports. Some examples:

- Peers->Master:
 - /services/cluster/master/peers
 - ▶ Add Peer to cluster
 - ▶ Heartbeat to master
 - /services/cluster/master/buckets
 - ▶ Alert master there is a new bucket
 - ▶ Alert master a bucket changes (hot -> warm, warm -> frozen)
- Master->Peers
 - /services/cluster/slave/buckets
 - ▶ Change primaries
 - ▶ Become searchable / unsearchable

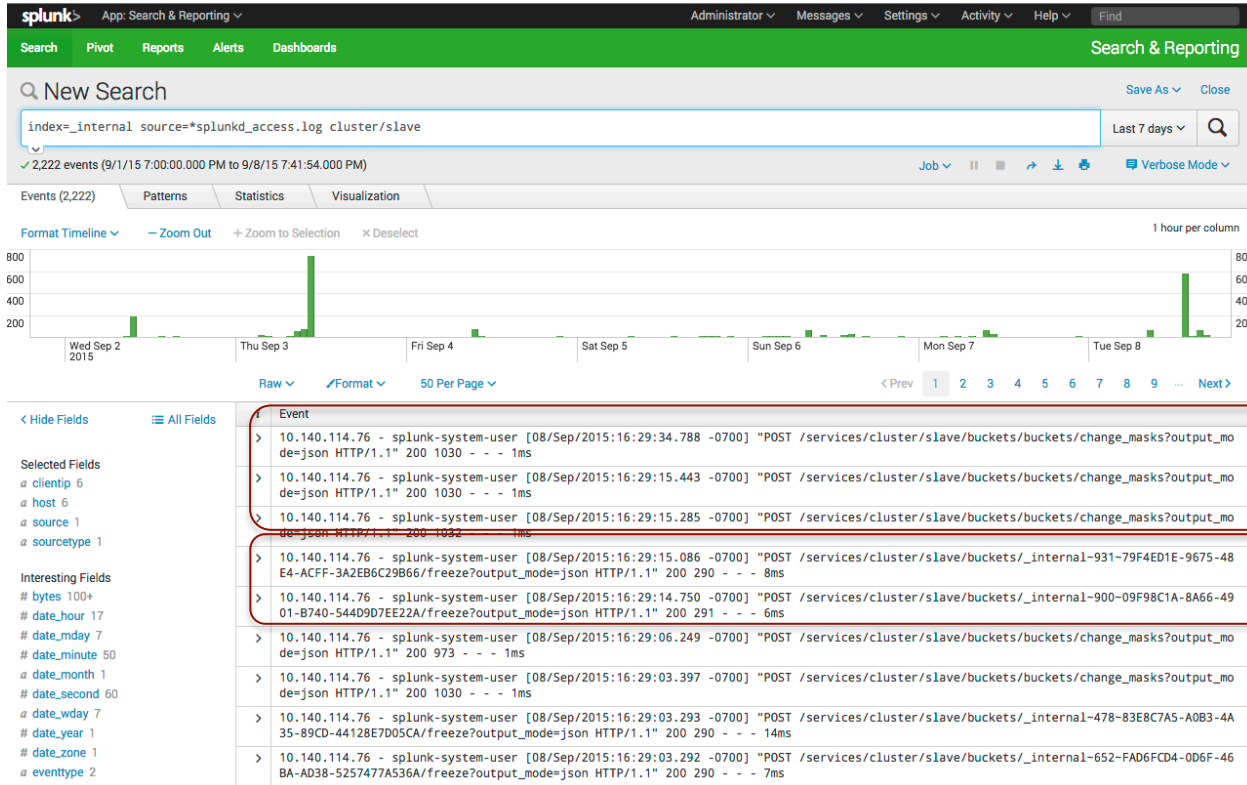


What's My Cluster Doing?



.conf2016

Endpoints Are Logged!



Bucket primary changes!

Buckets being frozen!

Metrics.log

```
> 09-08-2015 22:59:15.184 -0700 INFO Metrics - group=subtask_seconds, name=cmmaster_service, to_fix_streaming=0.000, to_fix_data_safety=0.016, to_fix_gen=0.000, to_fix_rep_factor=0.036, to_fix_search_factor=0.032, to_fix_sync=0.000, service=0.085
> 09-08-2015 22:59:15.184 -0700 INFO Metrics - group=subtask_seconds, name=cmmaster_endpoints, clustermastergeneration_edit=0.018000, clustermasterinfo_list=0.018000, clustermasterpeers_edit=0.185000
> 09-08-2015 22:59:15.184 -0700 INFO Metrics - group=subtask_counts, name=cmmaster_service, to_fix_streaming=0, to_fix_data_safety=97, to_fix_gen=0, to_fix_rep_factor=235, to_fix_search_factor=235, to_fix_sync=0, to_fix_added=0, to_fix_removed=0, to_fix_total=235, count=15
> 09-08-2015 22:59:15.184 -0700 INFO Metrics - group=subtask_counts, name=cmmaster_endpoints, clustermastergeneration_edit=18, clustermasterinfo_list=18, clustermasterpeers_edit=185
> 09-08-2015 22:59:15.184 -0700 INFO Metrics - group=executor, name=cmmaster_executor, jobs_added=0, jobs_finished=0, current_size=0, smallest_size=0, largest_size=0, max_size=0
> 09-08-2015 22:59:15.184 -0700 INFO Metrics - group=cmmaster_servicejobs, serviced=0.000000, current_size=0.000000
> 09-08-2015 22:58:44.184 -0700 INFO Metrics - group=subtask_seconds, name=cmmaster_service, to_fix_streaming=0.000, to_fix_data_safety=0.016, to_fix_gen=0.000, to_fix_rep_factor=0.036, to_fix_search_factor=0.031, to_fix_sync=0.000, service=0.084
> 09-08-2015 22:58:44.184 -0700 INFO Metrics - group=subtask_seconds, name=cmmaster_endpoints, clustermastergeneration_edit=0.019000, clustermasterinfo_list=0.019000, clustermasterpeers_edit=0.181000
> 09-08-2015 22:58:44.184 -0700 INFO Metrics - group=subtask_counts, name=cmmaster_service, to_fix_streaming=0, to_fix_data_safety=97, to_fix_gen=0, to_fix_rep_factor=235, to_fix_search_factor=235, to_fix_sync=0, to_fix_added=0, to_fix_removed=0, to_fix_total=235, count=16
> 09-08-2015 22:58:44.184 -0700 INFO Metrics - group=subtask_counts, name=cmmaster_endpoints, clustermastergeneration_edit=19, clustermasterinfo_list=19, clustermasterpeers_edit=181
```

- Cluster master/slave activity can be found under `cmmaster*` or `cmslave*` groupings/names
- Metrics about cluster endpoints
 - How many times each endpoint was hit
 - How long we spent in those endpoints
- Metrics about jobs (rep fixup jobs, searchable fixup jobs, freeze jobs, etc)
 - How many jobs remain?
- How many # of buckets do we still need to fix?

Clustering Logs/Activity

splunkd_access.log	metrics.log
<ul style="list-style-type: none">• Each individual endpoint access<ul style="list-style-type: none">• (master-side) services/cluster/master/...• (indexer-side) services/cluster/slave/...• How long we've spend at the endpoint (ms)<ul style="list-style-type: none">• Higher times indicate the CM/Indexer is swamped with work (>50ms? >100ms?)• The response (200 = success, non 200 = failure)	<p>Metric information with regards to Clustering Activity, recorded every 30 seconds.</p> <ul style="list-style-type: none">• name=cmmaster_endpoints<ul style="list-style-type: none">• group=subtask_count total number of accesses• group=subtask_seconds time Splunk spent responding to these endpoints• name=cmmaster_executor<ul style="list-style-type: none">• "Jobs" the CM has scheduled, finished, and current size of jobs to complete<ul style="list-style-type: none">• Jobs are responsible for hitting the endpoints and performing the action (move-primary, freeze, etc)• group=jobs, name=cmmaster<ul style="list-style-type: none">• Actual counts of the jobs and their jobnames <p>Indexers have their own corresponding jobs (cmslave)</p>

Cluster Activity

🔍 New Search

Save As ▾ Close

index=_internal source=*metrics.log host=marsha* group=subtask_counts name=cmmaster_endpoints | timechart max(clustermasterbuckets*)

Last 30 days ▾



✓ 100,058 events (8/8/15 12:00:00.000 AM to 9/7/15 10:09:21.000 PM)

Job ▾



Verbose Mode ▾

Events (100,058)

Patterns

Statistics (31)

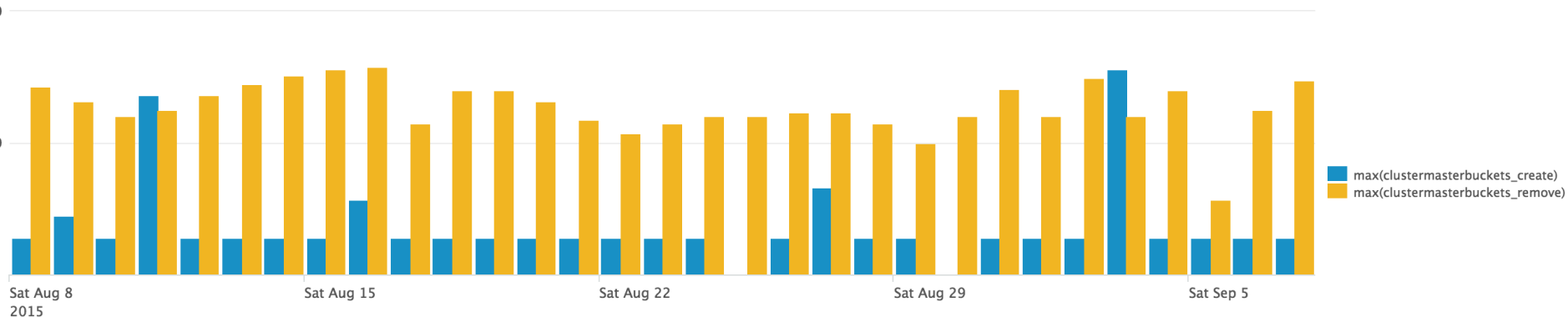
Visualization

Column ▾

Format ▾

100

10



Cluster Activity

🔍 New Search

Save As ▾ Close

```
index=_internal source=*metrics.log name=cmmaster_service | timechart avg(to_fix_gen) avg(to_fix_total) avg(to_fix_rep_factor) avg(to_fix_search_factor) span=30s
```

Date time range ▾ 🔍

✓ 58 events (9/9/15 12:14:00.000 AM to 9/9/15 12:29:18.000 AM)

Job ▾ || ■ ↶ ↷ 📄 📌 Smart Mode ▾

Events Patterns Statistics (31) Visualization

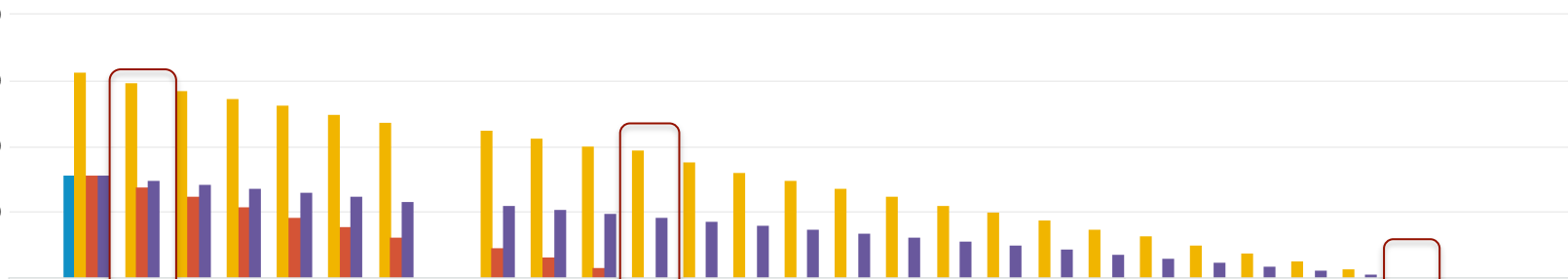
Column ▾ Format ▾

1,000

750

500

250



avg(to_fix_gen)
avg(to_fix_total)
avg(to_fix_rep_factor)
avg(to_fix_search_factor)

12:15 AM
Wed Sep 9
2015

Searchable

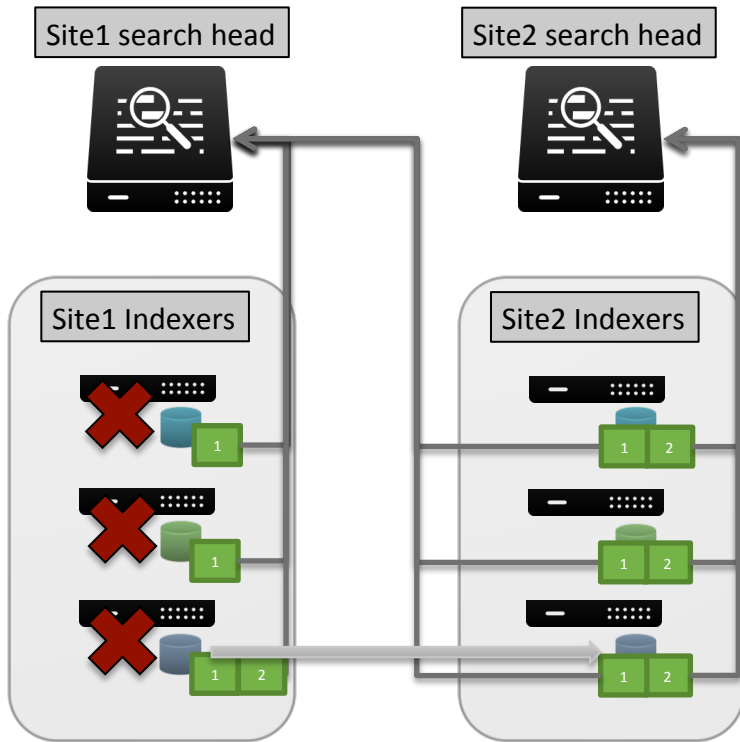
12:20 AM

RF Met

12:25 AM

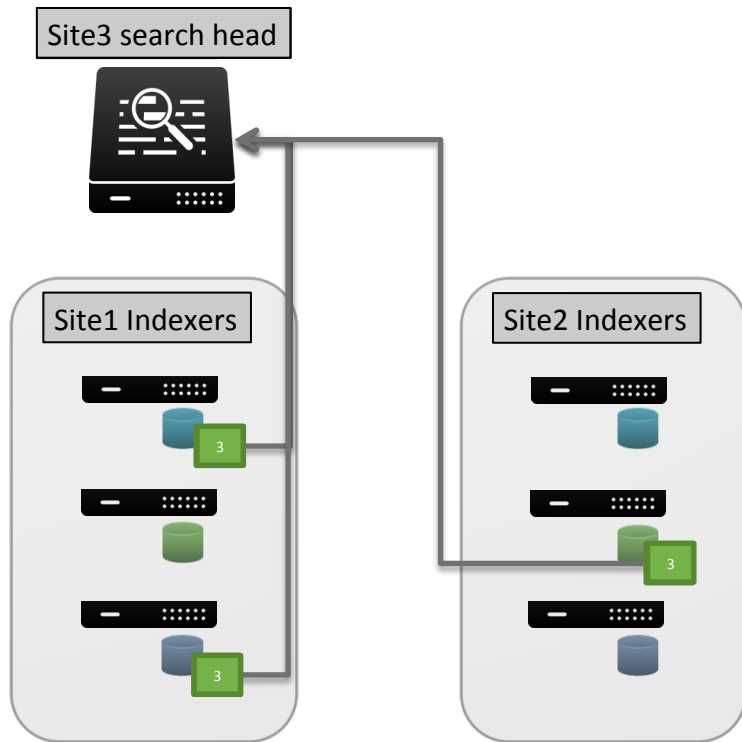
SF Met

Multisite Search Affinity



- When a searchable copy becomes available on a site, splunk will move the primary for that site to its local copy
- Buckets on a site will return events to a searchhead with the same site.
- If a peer goes down, the master will move the primaries that peer had to another copy
- If the entire site goes down, the other site(s) will become primaries

Multisite ~~Search Affinity~~



- Splunk 6.3 – site0
 - Primaries behave just like non-multisite, without any regards to site!
- Pre 6.3
 - Workaround!
 - Add another site to `available_sites`
 - Set SH (no indexers) to new site
 - Make sure to call “`splunk set indexing-ready`” on every CM restart
 - (wont work if your excess ‘total’ sites is greater than the # of non-specified sites... ie `origin:1 total:3` in our illustration will not work, because then the CM will try to put the 2 non-origin copies into a site each, and there are no indexers in site3!)

Buckets



.conf2016

More Buckets More Problems

splunk Administrator 30 Messages Settings Activity Help

Indexer Clustering: Master Node

✓ All Data is Searchable ⚠ Search Factor is Not Met ⚠ Replication Factor is Not Met

150 searchable Peers 0 not searchable Peers 32 searchable Indexes 0 not searchable Indexes

Peers (150) Indexes (32) Search Heads (1)

filter 100 per page Bucket Status

Index Name	Fully Searchable	Searchable Data Copies	Replicated Data Copies	Buckets	Cumulative Raw Data Size
index10	✓ Yes	2	3	34680	65.56 GB
index17	✓ Yes	2	3	34519	66.54 GB
index01	✓ Yes	2	3	33968	65.09 GB
index16	✓ Yes	2	3	33948	64.70 GB
index20	✓ Yes	2	3	33876	66.09 GB
index03	✓ Yes	2	3	33767	63.43 GB
index15	✓ Yes	2	3	33640	66.33 GB
index25	✓ Yes	2	3	33564	60.89 GB
index07	✓ Yes	2	3	33554	70.02 GB
index13	✓ Yes	2	3	33545	64.44 GB
index18	✓ Yes	2	3	33522	63.62 GB
index11	✓ Yes	2	3	33396	64.23 GB
index12	✓ Yes	2	3	33369	65.71 GB
index08	✓ Yes	2	3	33253	62.88 GB
index29	✓ Yes	2	3	33194	63.73 GB
index02	✓ Yes	2	3	33054	64.54 GB
index19	✓ Yes	2	3	33042	63.57 GB
index04	✓ Yes	2	3	32961	61.66 GB
index28	✓ Yes	2	3	32792	60.72 GB
index30	✓ Yes	2	3	32722	62.16 GB
index26	✓ Yes	2	3	32717	61.21 GB
index05	✓ Yes	2	3	32697	64.35 GB
index24	✓ Yes	2	3	32637	62.12 GB
index14	✓ Yes	2	3	32615	64.45 GB
index21	✓ Yes	2	3	32443	62.57 GB
index09	✓ Yes	2	3	32339	62.48 GB
index23	✓ Yes	2	3	31975	60.81 GB
index22	✓ Yes	2	3	31789	61.27 GB
index06	✓ Yes	2	3	31711	62.87 GB
index77	✓ Yes	2	3	31400	67.84 GB

More Buckets More Problems

The screenshot shows the Splunk Clustering: Master Node interface. At the top, there are navigation menus for 'Apps', 'Administrator', 'Messages', 'Settings', 'Activity', and 'Help'. Below the title 'Clustering: Master Node', there are three warning icons with red triangles and exclamation marks, each followed by a text label: 'Some Data is Not Searchable', 'Search Factor is Not Met', and 'Replication Factor is Not Met'. Below these labels, there are two summary statistics. The first one shows '29 searchable' and '1 not searchable' under the heading 'PEERS'. The second one shows '0 searchable' and '13 not searchable' under the heading 'INDEXES'. At the bottom of the interface, there are three tabs: 'Peers (30)', 'Indexes (13)', and 'Search Heads (2)'.

- More buckets (and more peers) means the CM has to do more work
 - Iterates through each bucket, checking whether it needs to queue up any fixup jobs
 - ▶ Replication Jobs (to meet RF)
 - ▶ Search Jobs (to meet SF)
 - ▶ Primary Jobs (all buckets need to have a primary copy per site)
 - ▶ Other jobs (freezing, checksum, rolling, etc)
- As the number of buckets grows, CM responsiveness goes down

More Buckets More Settings

server.conf	
service_interval (CM)	<p>Specifies how often the CM should look through the buckets, scheduling jobs as necessary. Default = 1.</p> <ul style="list-style-type: none">• Adjust to 1 sec for every 50k buckets.
heartbeat_period (Indexer)	<p>Specifies how often the Indexers contact the CM. Defaults to every 1 second.</p> <ul style="list-style-type: none">• For lots of peers (>50) or lots of buckets (>100k), we can increase this value to 5-30
heartbeat_timeout (CM)	<p>Specifies how long before an Indexer is considered 'Down' when no heartbeats comes in</p> <ul style="list-style-type: none">• Multiple of heartbeat_period, anywhere from 20x – 60x
cxn_timeout (CM+Indexer) rcv_timeout (CM+Indexer) send_timeout (CM+Indexer)	<p>Specifies how long before an intra-cluster connection will terminate. Default = 60</p> <ul style="list-style-type: none">• If a cluster indexer times out, it will re-add itself to the CM, which itself is a busy operation (it needs to resync the state of all its buckets), which can lead to negative feedback loops...• These can be bumped up for busier clusters (300s)
indexes.conf	
rotatePeriodInSecs (Indexer)	<p>Specifies how often to check through all the buckets – rolling them from hot->warm->cold as necessary. Default = 60</p> <ul style="list-style-type: none">• 10min=600

Inspecting Buckets

The screenshot shows a Splunk Atom Feed for the bucket 'cluster/master/buckets'. The feed includes a title, update information, and a list of bucket properties. The properties are organized into sections: 'app', 'eai:acl', and 'peers'. The 'app' section shows 'bucket_size' as 416918 and 'constrain_to_origin_site' as 0. The 'eai:acl' section shows 'owner' as 'system', 'removable' as 0, and 'sharing' as 'system'. The 'peers' section shows 'bucket_flags' as 0x0, 'checksum' as 79F4ED1E-9675-48E4-ACFF-3A2EB6C29B66, 'checksum_state' as StableChecksum, 'search_state' as Unsearchable, 'server_name' as Cindy_Peer, and 'status' as Complete.

```
Splunk Atom Feed: clustermasterbuckets
Updated: 2015-09-09T00:58:27-07:00 Splunk build: 6.3.0
Feed links: create - _acl -
_audit-58-56805911-7851-49A5-8FC5-8B7FC49B0938
bucket_size 416918
constrain_to_origin_site 0
app
can_list 1
can_write 1
modifiable 0
owner system
eai:acl
read
  1. ad_admin
  2. admin
  3. everything
  4. splunk-system-role
perms
  write
    1. ad_admin
    2. admin
    3. everything
    4. splunk-system-role
removable 0
sharing system
force_roll 0
frozen 0
index _audit
origin_site site3
peers
  bucket_flags 0x0
  checksum 79F4ED1E-9675-48E4-ACFF-3A2EB6C29B66
  checksum_state StableChecksum
  search_state Unsearchable
  server_name Cindy_Peer
  status Complete
```

services/cluster/master/buckets

- Which peers does the bucket exist on?
- Which peers is the bucket primary?
- Is the bucket searchable/unsearchable/pending-searchable?

		bucket_flags	0x4
		checksum	
	09F98C1A-8A66-4901-B740-544D9D7EE22A	checksum_state	StableCksum
		search_state	Searchable
		server_name	Bobby_Peer
		status	Complete
		bucket_flags	0x3
		checksum	
	83E8C7A5-A0B3-4A35-89CD-44128E7D05CA	checksum_state	StableCksum
		search_state	Searchable
		server_name	Marsha_Peer
		status	Complete
		bucket_flags	0x0
		checksum	
	FAD6FCD4-0D6F-46BA-AD38-5257477A536A	checksum_state	StableCksum
		search_state	Searchable
		server_name	Jan_Peer
		status	Complete
		site0	83E8C7A5-A0B3-4A35-89CD-44128E7D05CA
primaries_by_site		site1	83E8C7A5-A0B3-4A35-89CD-44128E7D05CA
		site2	09F98C1A-8A66-4901-B740-544D9D7EE22A
		site1	2
rep_count_by_site		site2	1
		site1	2
search_count_by_site		site2	1

Inspecting Buckets

The screenshot shows a Splunk Atom Feed for the bucket 'clustermasterbuckets'. The feed includes a title, update information, and a list of bucket properties. The properties are organized into sections: 'app', 'permissions', 'removable', 'sharing', 'force_roll', 'frozen', 'index', 'origin_site', 'checksum', and 'status'. The 'permissions' section is expanded to show 'read' and 'write' permissions for 'eai:iacl'.

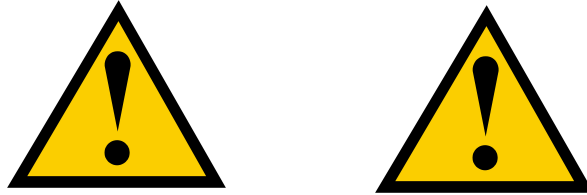
```
Updated: 2015-09-09T00:58:27-07:00 Splunk build: 6.3.0
Feed links: create - _acl -
_audit-58--56605911-7851-49A5-8FC5-8B7FC49B0938
bucket_size 416918
constrain_to_origin_site 0
app
  can_list 1
  can_write 1
  modifiable 0
  owner system
eai:iacl
  read
    1. ad_admin
    2. admin
    3. everything
    4. splunk-system-role
  write
    1. ad_admin
    2. admin
    3. everything
    4. splunk-system-role
removable 0
sharing system
force_roll 0
frozen 0
index _audit
origin_site site3
checksum
  bucket_flags 0x0
  checksum 79F4ED1E-9675-48E4-ACFF-3A2EB6C29B66
  checksum_state StableCksum
  search_state Unsearchable
  server_name Cindy_Peer
  status Complete
peers
```

There's so many buckets! How do I find one that I care about?
Why would I care?

Filters! `services/cluster/master/buckets?filter=`

- Which buckets do not have primaries?
 - `buckets?filter=has_primary=false`
- Which buckets do not meet my RF=3?
 - `buckets?filter=replication_count<3`
- Which buckets are frozen?
 - `buckets?filter=frozen=true`
- Standalone?
 - `buckets?filter=standalone=true`
- Standalone and frozen?
 - `buckets?filter=standalone=true&filter=frozen=true`
 - (don't think this is a thing)
- Don't meet RF=3 and index=main?
 - `buckets?filter=replication_count>3&filter=index=main`

Modifying Buckets



Endpoints!

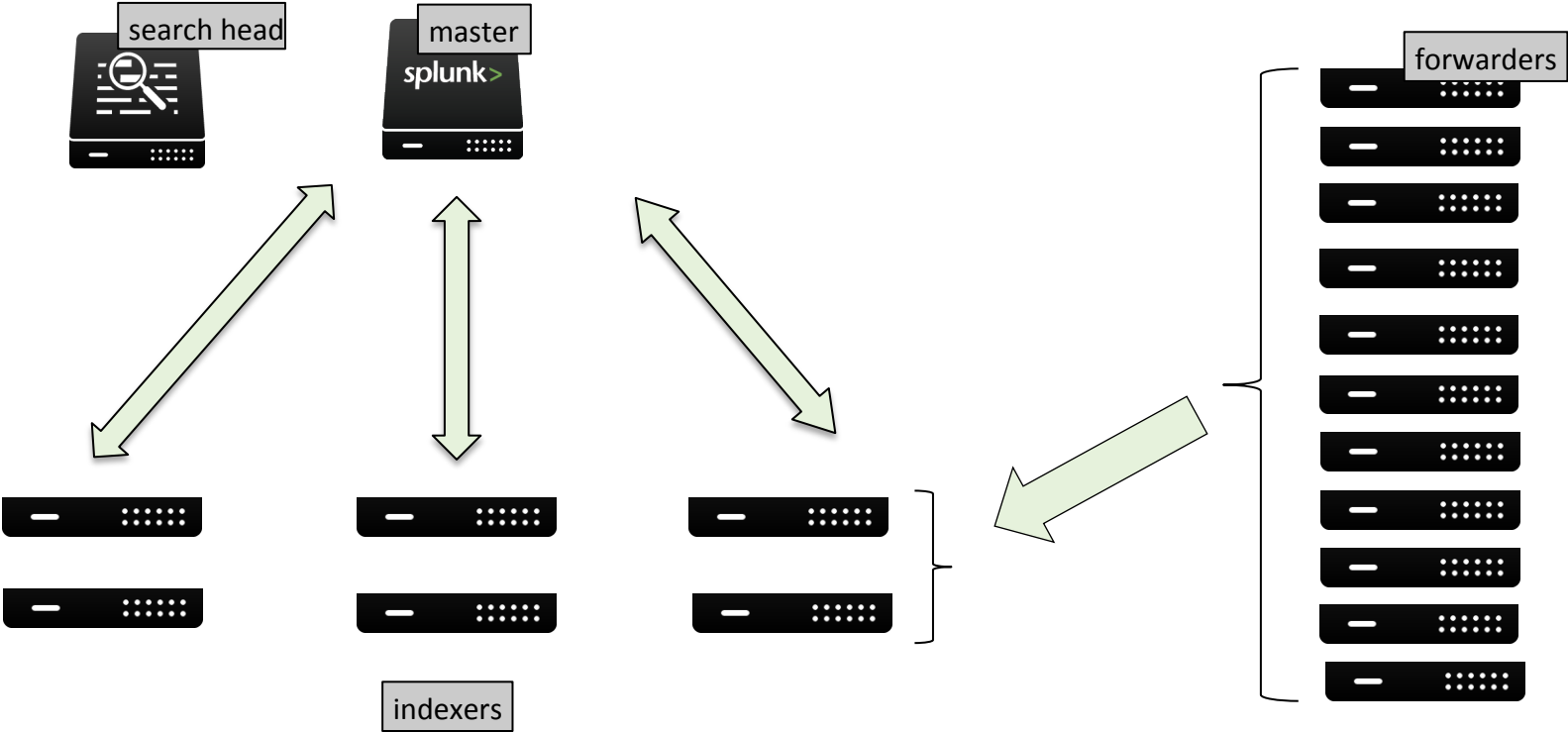
- Freeze a bucket:
 - `curl -k -u admin:changeme https://{indexer}:{mgmt}/services/data/indexes/{INDEX}/freeze-buckets -d bucket_ids=46_11115C7A-E2F0-4225-A740-4ED6BD2D9CE5 -X POST`
- Remove a copy of a bucket:
 - `curl -k -u admin:changeme "https://{master}:{mgmt}/services/cluster/master/buckets/main~1490~D4A07A5D-3C3C-4D36-BD70-D610B432466F/remove_from_peer" -d peer={PEER_GUID}`
- Remove all copies of a bucket:
 - `curl -k -u admin:changeme "https://{master}:{mgmt}/services/cluster/master/buckets/main~1490~D4A07A5D-3C3C-4D36-BD70-D610B432466F/remove_all" -d peer={PEER_GUID}`

Performance Testing

.conf2016

splunk >

Clustering Performance Tests



Scale Testing

Cluster Variable	Tested up to
Indexers	1000
Indexes	1000
Unique buckets	1.5 million
Forwarders	100000

1000 Node Performance Test

Amazon EC2

- 1 hs1.8xlarge master
- 1 c1.xlarge searchhead
- 1000 c1.xlarge indexers
- 10000 forwarders
 - 100 m1.small (10 forwarders / box)

Replication factor: origin:2, total:3

Search factor: origin:1, total:2

Data ingestion rate: 2 MB/s per indexer

Ingested 1 million unique buckets

11 TB of syslog data

Machine type	vCPUz	Memory (GB)	Instance Store (GB)
hs1.8xlarge	16	117	24 x 2000
c1.xlarge	8	7	4 x 420
m1.small	1	1.7	1 x 160

Large Cluster Configurations

Master

Server.conf

```
[sslConfig]
```

```
allowSslCompression = false
```

```
[clustering]
```

```
heartbeat_timeout = 600
```

```
service_interval = 10
```

Indexer

Server.conf

```
[general]
```

```
useHTTPClientCompression =  
true
```

```
[clustering]
```

```
heartbeat_period = 10
```

```
cxn_timeout = 1200
```

```
send_timeout = 1200
```

```
rcv_timeout = 1200
```

Test Case: Peer Failure

1. Take down a peer
./splunk stop -f
2. Wait for:
 - Searchable
 - Replication factor met
 - Search factor met

Splunk Release	Number of buckets	Time to be searchable	Time to meet rf/sf
6.3	30000	160 s	85 s
6.4	51000	36 s	47 s
6.5	935000	2 s	32 s

The screenshot shows the Splunk Admin Console interface for 'Indexer Clustering: Master Node'. The top navigation bar includes 'splunk>', 'Apps', 'Administrator', '923 Messages', 'Settings', 'Activity', 'Help', and a 'Find' search bar. Below the navigation, there are three status indicators, each with a green checkmark: 'All Data is Searchable', 'Search Factor is Met', and 'Replication Factor is Met'. Under 'All Data is Searchable', it shows '1000 searchable' and '0 not searchable' Peers. Under 'Search Factor is Met', it shows '12 searchable' and '0 not searchable' Indexes. There are also buttons for 'Edit', 'More Info', and 'Documentation'.

Test Case: Site Failure

1. Take down 1 site (333 peers)

`./splunk stop -f`

2. Wait for:

– Searchable

Splunk Release	Number of buckets	Time to be searchable
6.3	30000	791 s
6.4	51000	328 s
6.5	935000	451 s

The screenshot shows the Splunk Admin interface for 'Indexer Clustering: Master Node'. The top navigation bar includes 'splunk>', 'Apps', 'Administrator', '923 Messages', 'Settings', 'Activity', 'Help', and 'Find'. Below the title, there are three status indicators, each with a green checkmark: 'All Data is Searchable', 'Search Factor is Met', and 'Replication Factor is Met'. Under 'All Data is Searchable', it shows '1000 searchable Peers' and '0 not searchable Peers'. Under 'Replication Factor is Met', it shows '12 searchable Indexes' and '0 not searchable Indexes'. On the right side, there are buttons for 'Edit', 'More Info', and 'Documentation'.

Test Case: Master Restart

1. Force restart of master

`./splunk restart -f`

2. Wait for:

- Master to restart
- Searchable
- Replication factor met
- Search factor met

Splunk Release	Number of buckets	Time to be searchable	Time to meet rf/sf
6.3	30,000	361 s	1007 s
6.4	51,000	286 s	488 s
6.5	935,000	720 s	723 s

The screenshot shows the Splunk Admin Console interface for 'Indexer Clustering: Master Node'. The top navigation bar includes 'splunk>' and 'Apps' on the left, and 'Administrator', '923 Messages', 'Settings', 'Activity', 'Help', and 'Find' on the right. Below the navigation, the page title is 'Indexer Clustering: Master Node' with 'Edit', 'More Info', and 'Documentation' links. The main content area displays three green checkmarks indicating successful status: 'All Data is Searchable', 'Search Factor is Met', and 'Replication Factor is Met'. Below these, statistics are shown: '1000 searchable Peers' and '0 not searchable Peers' for the first two, and '12 searchable Indexes' and '0 not searchable Indexes' for the third.

Test Case: Rolling Restart

1. Perform rolling restart (on cm)

`./splunk rolling-restart cluster-peers`

2. Wait for:

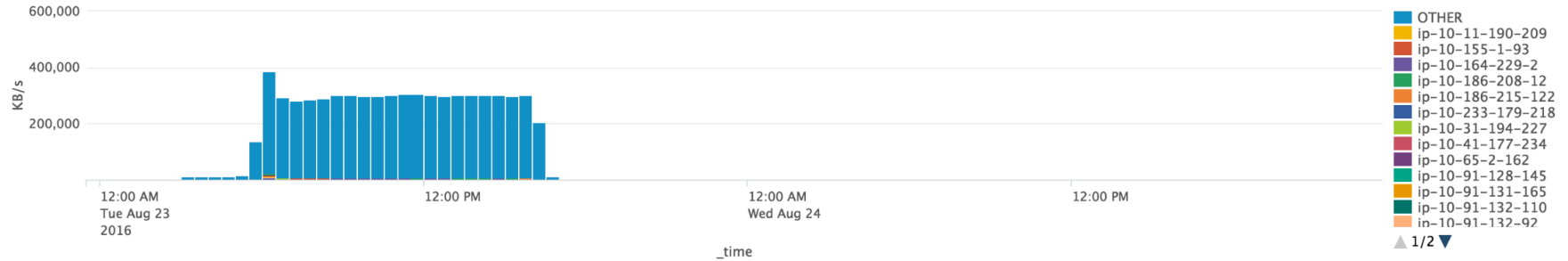
- All peers to restart
- Searchable
- Replication factor met
- Search factor met

Splunk Release	Number of buckets	Time to be searchable	Time to meet rf/sf
6.4	51000	746 s	1127 s
6.5	935000	328 s	470 s

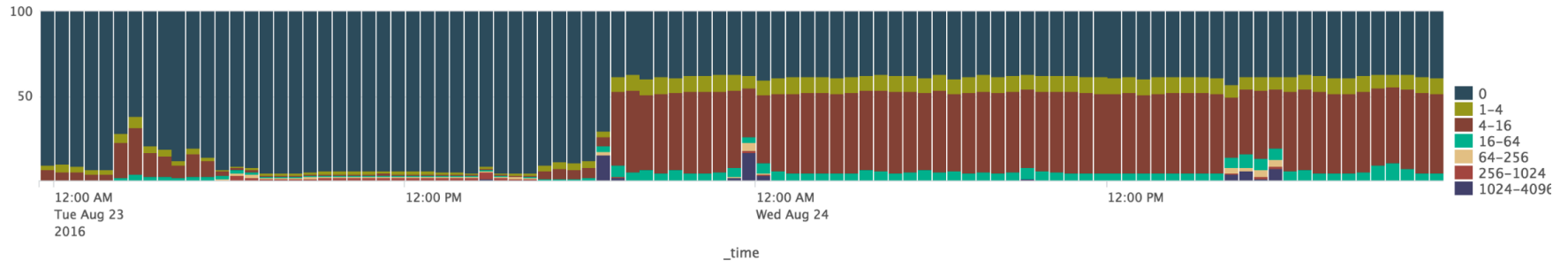
The screenshot shows the Splunk Admin Console interface for 'Indexer Clustering: Master Node'. The top navigation bar includes 'splunk>' and 'Apps' on the left, and 'Administrator', '923 Messages', 'Settings', 'Activity', 'Help', and 'Find' on the right. Below the navigation, the page title is 'Indexer Clustering: Master Node' with 'Edit', 'More Info', and 'Documentation' links. The main content area displays three green checkmarks indicating successful status: 'All Data is Searchable', 'Search Factor is Met', and 'Replication Factor is Met'. Below these, there are two summary statistics: '1000 searchable 0 not searchable Peers' and '12 searchable 0 not searchable Indexes'.

Resource Metrics

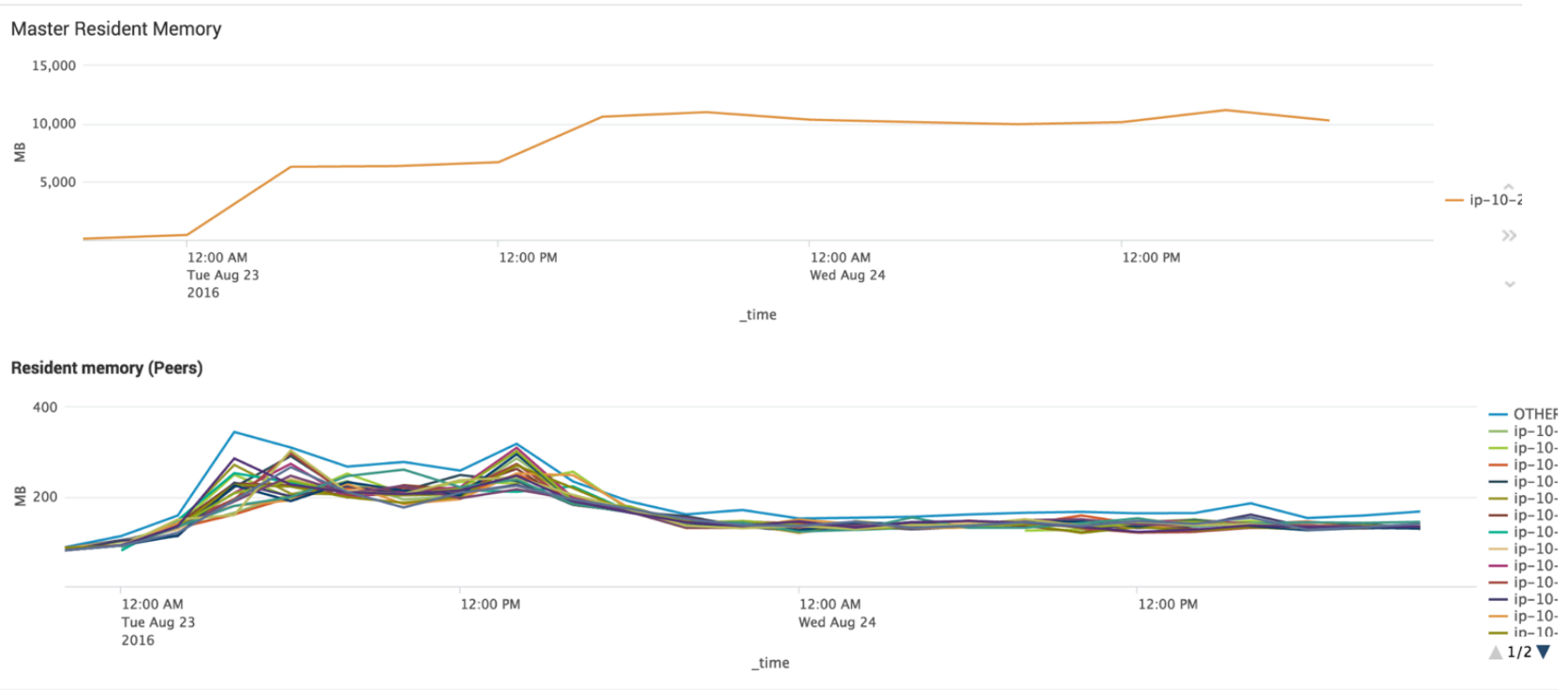
Indexing Throughput



Index Queue Health

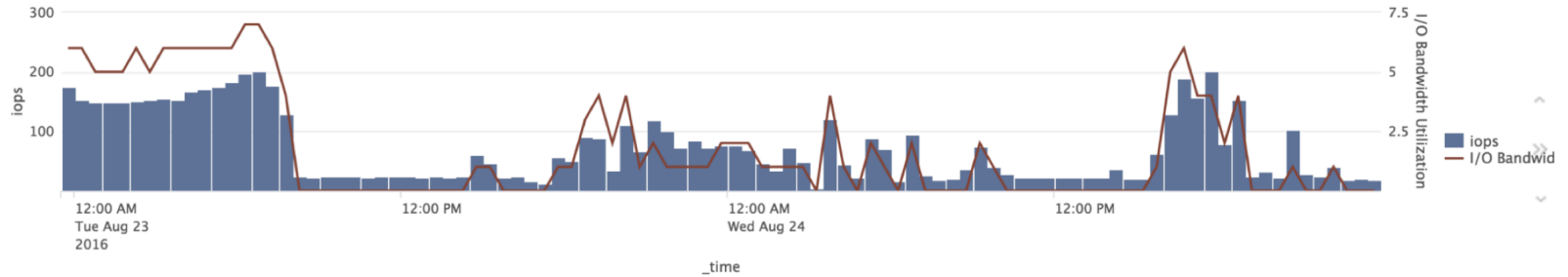


Resource Metrics

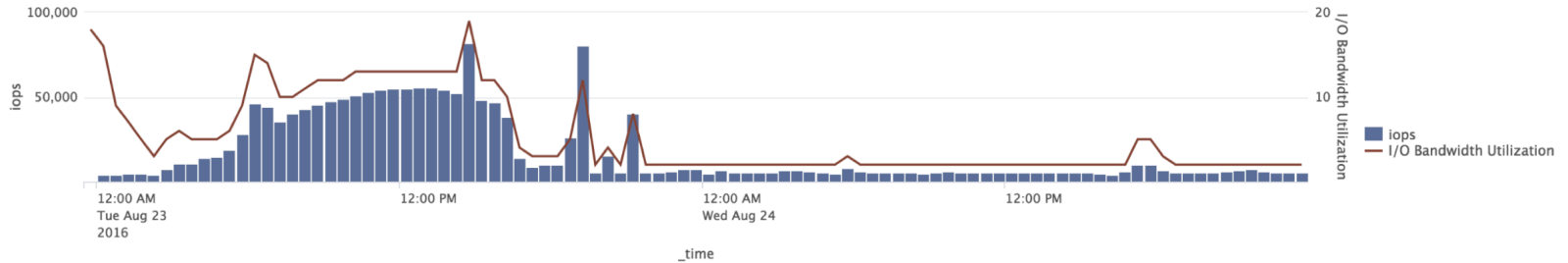


Resource Metrics

Average I/O Usage and Bandwidth Utilization (cm)



Average I/O Usage and Bandwidth Utilization (Peers)



Performance Tests Every Release

- Local in-house 10 indexer lab cluster
- 1000 indexer multi-site cluster
- 500k unique buckets, 150 indexer cluster

Regression tests we run:

- Peer and site failure
- Master restart
- Rolling restart
- Bundle push
- Indexing thruput

And of course, test newest features!

Miscellaneous

Q&A

THANK YOU

.conf2016