splunk> .conf2017

Indexer Clustering Internals & Performance

Da Xu | Chloe Yeung

September 28, 2017 | Washington, DC

Forward-Looking Statements

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 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.

splunk> .conf2017

Agenda

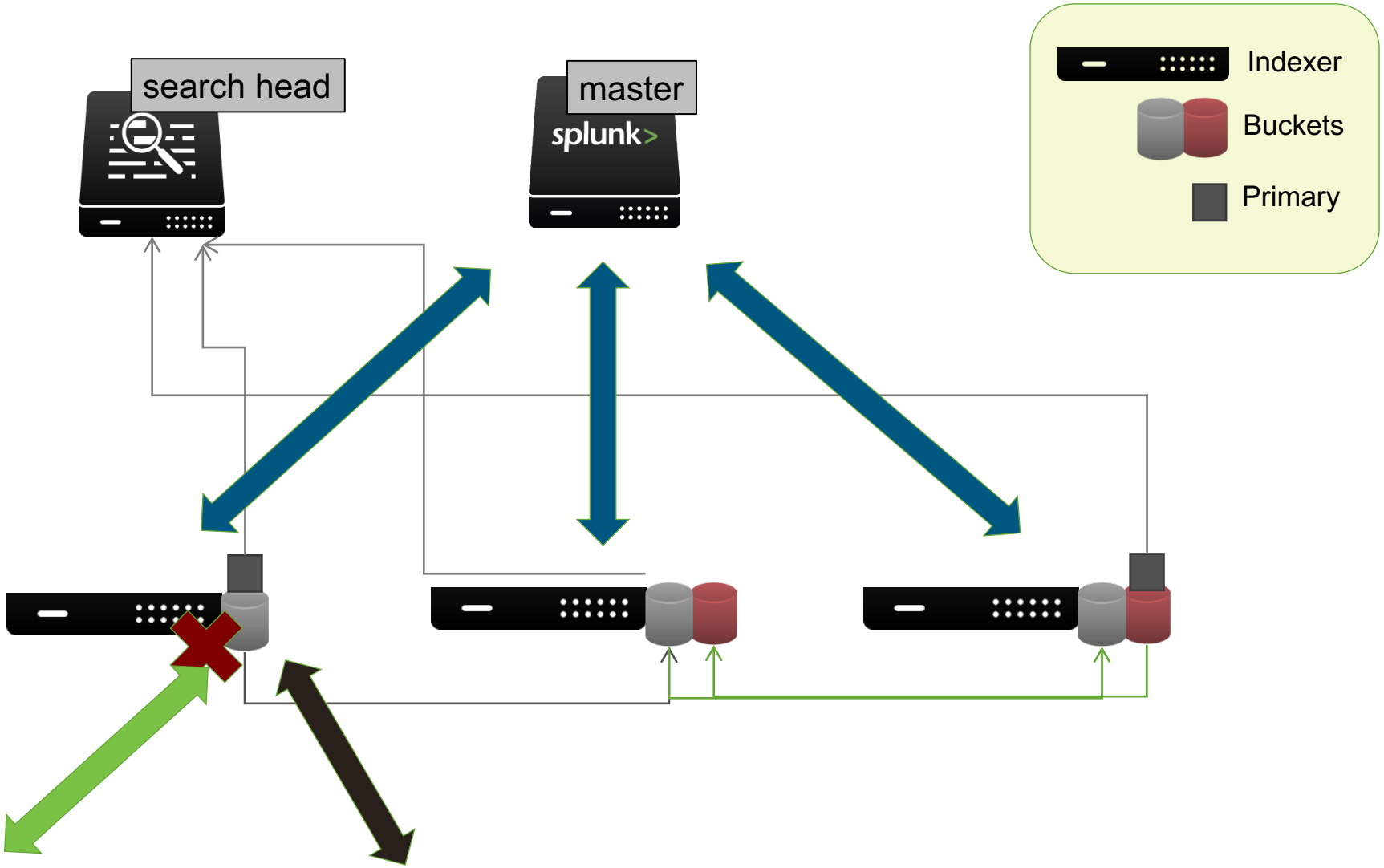
Indexer Clustering Overview	▸ How Clustering works ▸ Communication Through Endpoints ▸ metrics.log, splunkd_access.log ▸ Cluster Activity
Peers and Buckets!	▸ Inspecting Buckets ▸ More Buckets More Problems ▸ Settings for Large Clusters
Performance Testing	▸ Testing Setups ▸ Multisite Testing Results ▸ Bundle Pushing Results ▸ 5M+ Buckets Results
	▸ Q&A

splunk> .conf2017

Indexer Clustering Overview

splunk> .conf2017

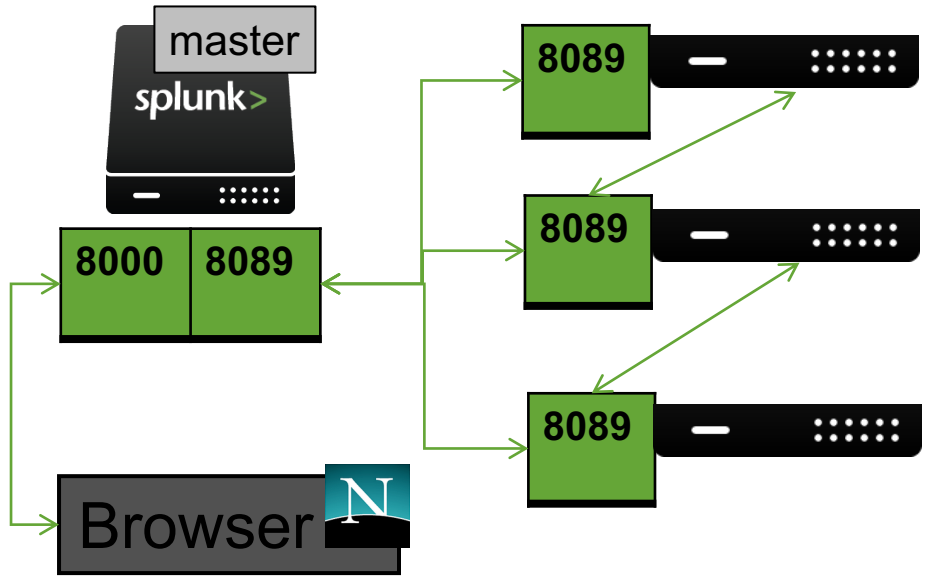
Cluster!



Communication Through Endpoints

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

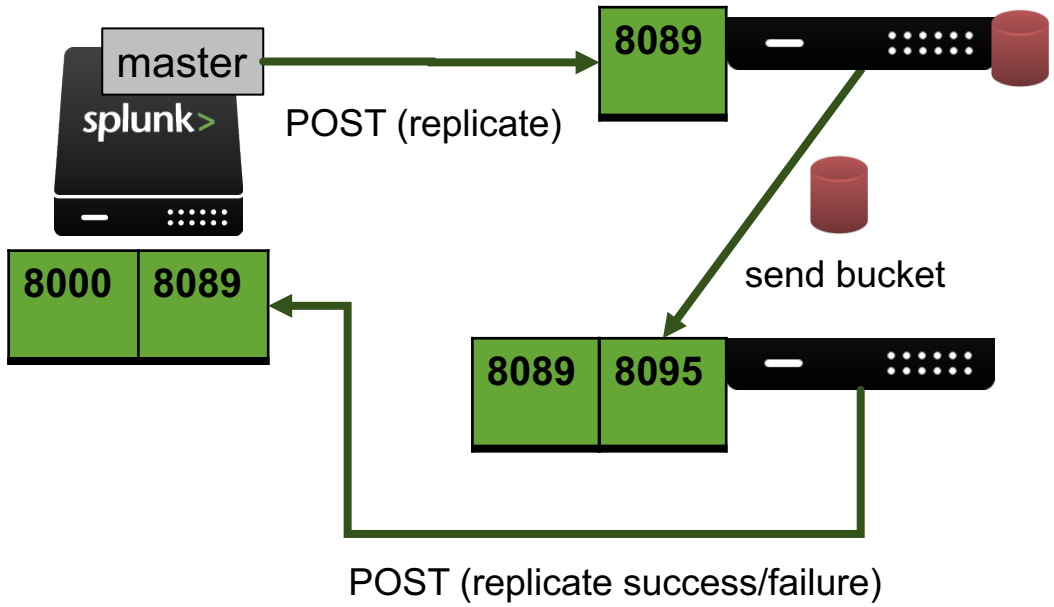
- Peers->Master:
  - /services/cluster/master/peers
    - Add Peer to cluster
    - Heartbeat to master
  - /services/cluster/master/buckets
    - Alert master there is a new bucket
    - Alert master a bucket changes (hot -> warm, warm -> frozen)
- Master->Peers
  - /services/cluster/slave/buckets
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



Communication Through Endpoints

Bucket replication example

1. CM sends a replicate command to the indexer
   - POST slave/buckets/BID1/replicate
2. Indexer receives the command, and starts replication!
   - Sends the bucket through the replication port
- On success / failure, the indexers will report the result to the CM



Inspecting The Cluster

Endpoints, Logs, and Metrics

splunk> .conf2017

Cluster/Master Endpoints

The services/cluster/master/ endpoints contain lots of information about the cluster.

- Cluster/master/generation
  - Peer overview
  - Current generation information (Searchable? RF met? SF met?)
- Cluster/master/peers
  - Bucket counts and states for every peer
  - Bucket primary counts
- Cluster/master/buckets
  - Lists all the buckets in the cluster
  - Individual bucket states and copies

Cluster/Master/Peers

49A0D210-5662-4AC0-AC3B-72777066C271

active_bundle_id	6EFA36E8BE12EB09D0F4D4C97223522A	
apply_bundle_status	**invalid_bundle**	**bundle_validation_errors**
		invalid_bundle_id
	reasons_for_restart	
	restart_required_for_apply_bundle	0
	status	None
base_generation_id	1	
bucket_count	1638	
bucket_count_by_index	**_audit**	79
	_internal	13
	_telemetry	27
	main	1519
buckets_rf_by_origin_site	**default**	1638
buckets_sf_by_origin_site	**default**	1638

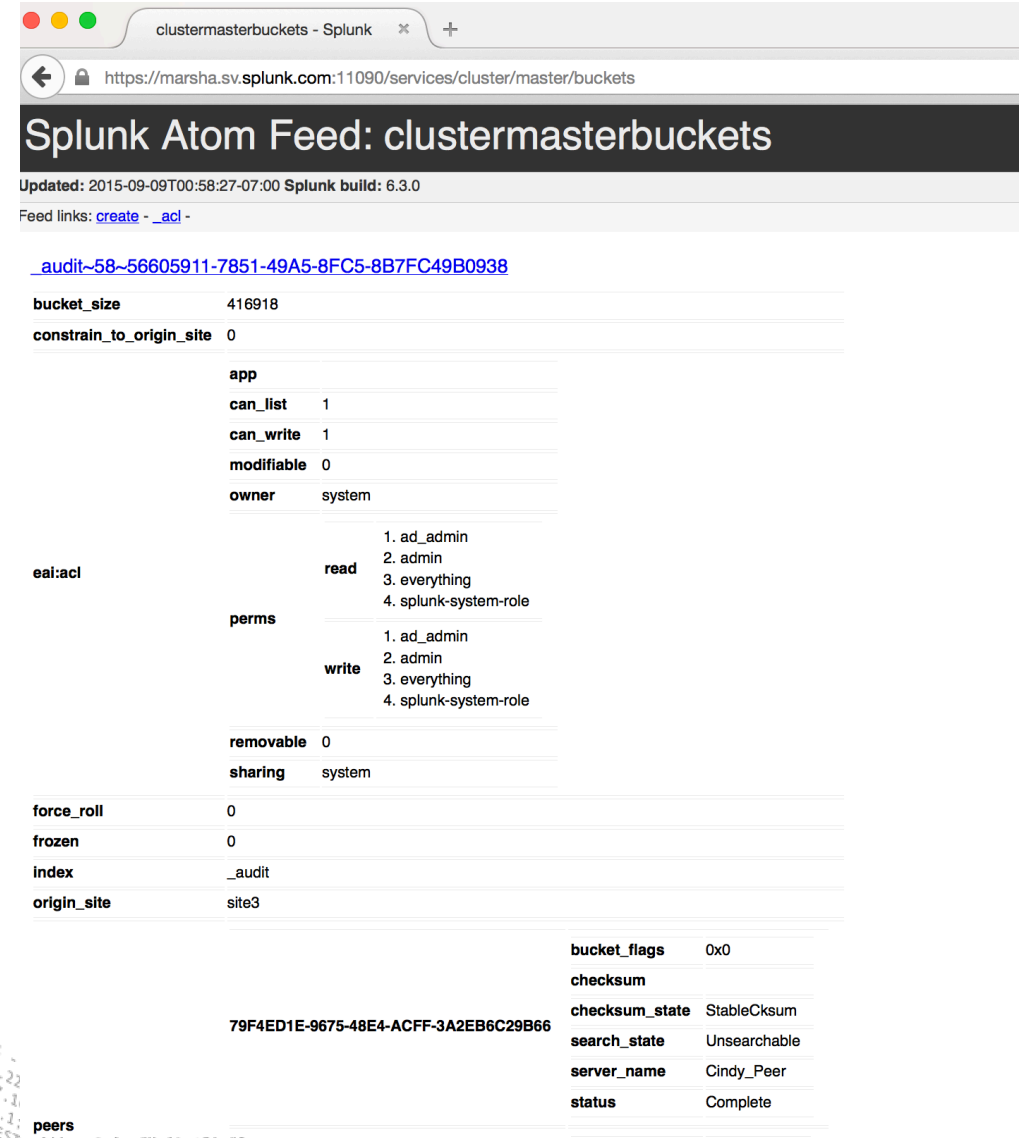
Cluster/Master/Peers

search_state_counter	**Searchable**	1638
	SearchablePendingMask	0
site	default	
status	Up	
status_counter	**Complete**	1631
	StreamingSource	3
	StreamingTarget	4

Cluster/Master/Buckets

frozen	0		
index	main		
origin_site	default		
peers	**49A0D210-5662-4AC0-AC3B-72777066C271**	**bucket_flags**	0xffffffffffffffff
		checksum	
		checksum_state	StableCksum
		search_state	Searchable
		server_name	p1
		status	Complete
		summaries_data_models	
		summaries_report_accs	
	99B08D78-F5CD-454A-A72B-11564161E269	**bucket_flags**	0x0
		checksum	
		checksum_state	StableCksum
		search_state	Searchable
		server_name	p2
		status	Complete
		summaries_data_models	
		summaries_report_accs	
rep_count_by_site	**default**	2	
search_count_by_site	**default**	2	

Cluster/Master/Buckets



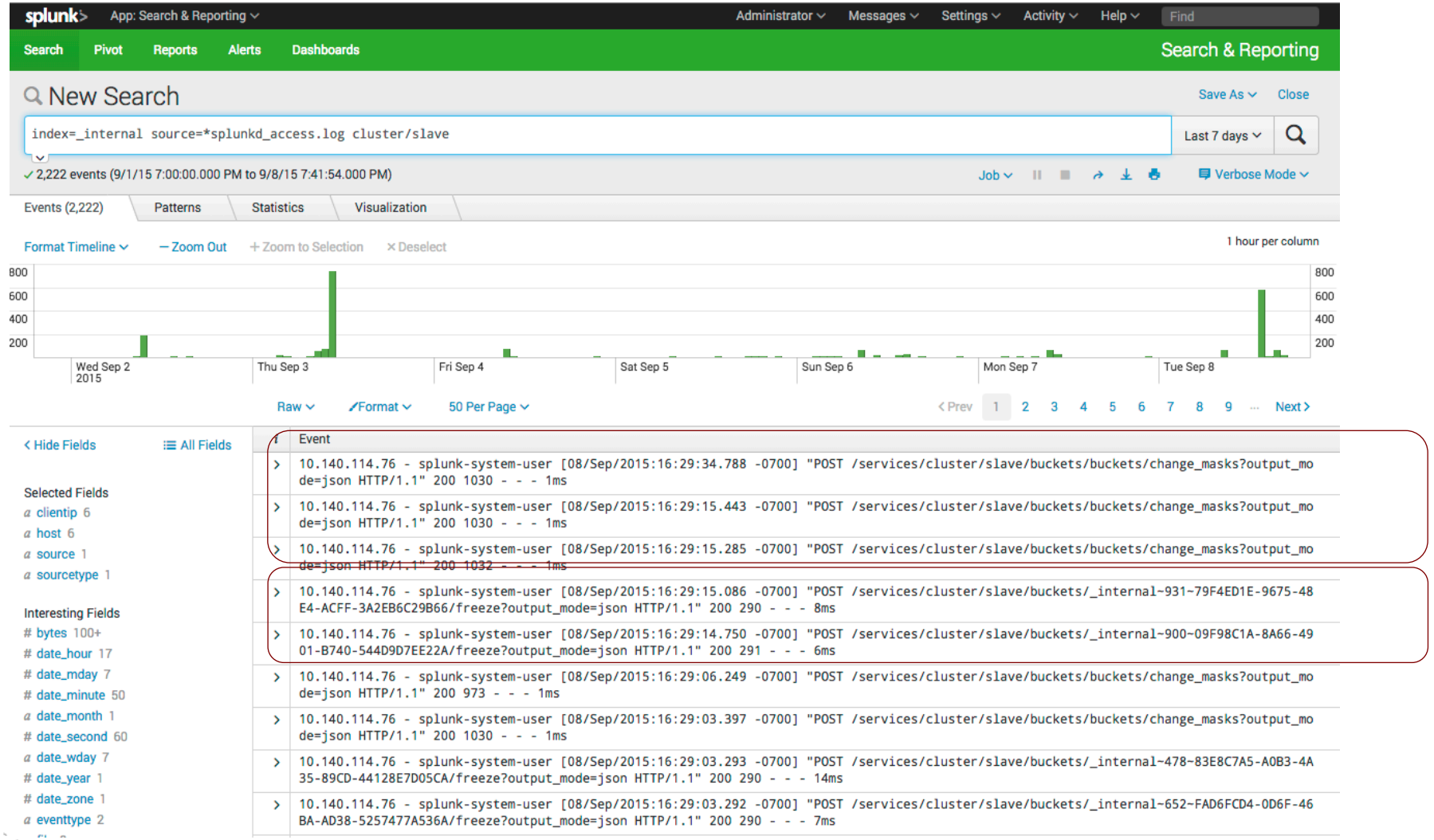
There's so many buckets! How do I find one that I care about? Why would I care?

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

splunk> .conf2017

Endpoints Are Logged!



Bucket primary changes!
Buckets being frozen!

splunk> .conf2017

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_fac
tor=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_fi
x_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_fac
tor=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_fi
x_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 # of buckets do we still need to fix?

splunk> .conf2017

Clustering Logs/Activity

splunkd_access.log

- Each individual endpoint access
  - (master-side) services/cluster/master/…
  - (indexer-side) services/cluster/slave/…
- How long we've spend at the endpoint (ms)
  - Higher times indicate the CM/Indexer is swamped with work (>50ms? >100ms?)
- The response (200 = success, non 200 = failure)

metrics.log

- Metric information with regards to Clustering Activity, recorded every 30 seconds
- name=cmmaster_endpoints
  - group=subtask_count total number of accesses
  - group=subtask_seconds time Splunk spent responding to these endpoints
- name=cmmaster_executor
  - "Jobs" the CM has scheduled, finished, and current size of jobs to complete
    - Jobs are responsible for hitting the endpoints and performing the action (move-primary, freeze, etc)
- group=jobs, name=cmmaster
  - Actual counts of the jobs and their jobnames
- Indexers have their own corresponding jobs (cmslave)

splunk> .conf2017

Jobs Metrics (metrics.log)

splunk> .conf2017

Jobs By Time (splunkd.log)

New Search

Save As Close

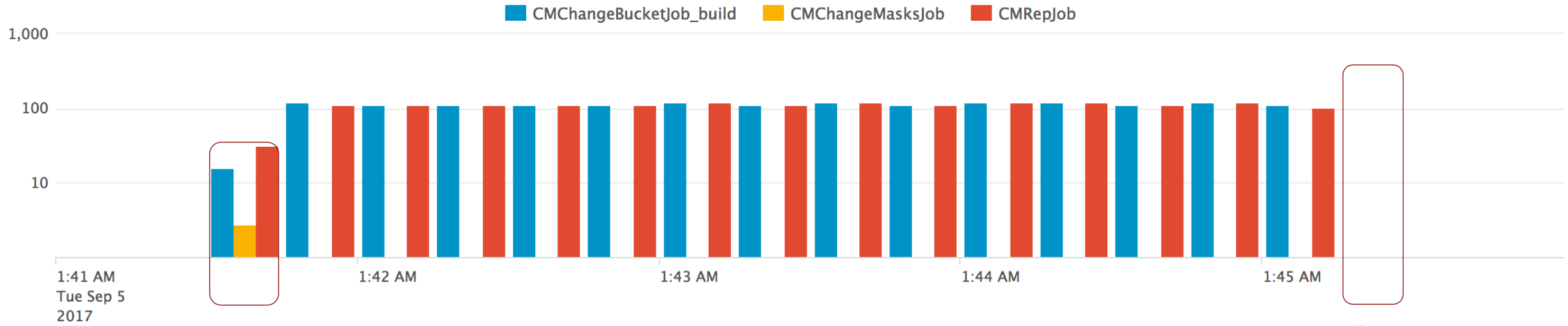
```
index=_internal source=*master*splunkd.log CMRepJob | timechart span=15s count by job
```

Date time range

✓ 3,264 events (9/5/17 1:41:00.000 AM to 9/5/17 1:46:00.000 AM) No Event Sampling Job Smart Mode

Events Patterns Statistics (20) Visualization

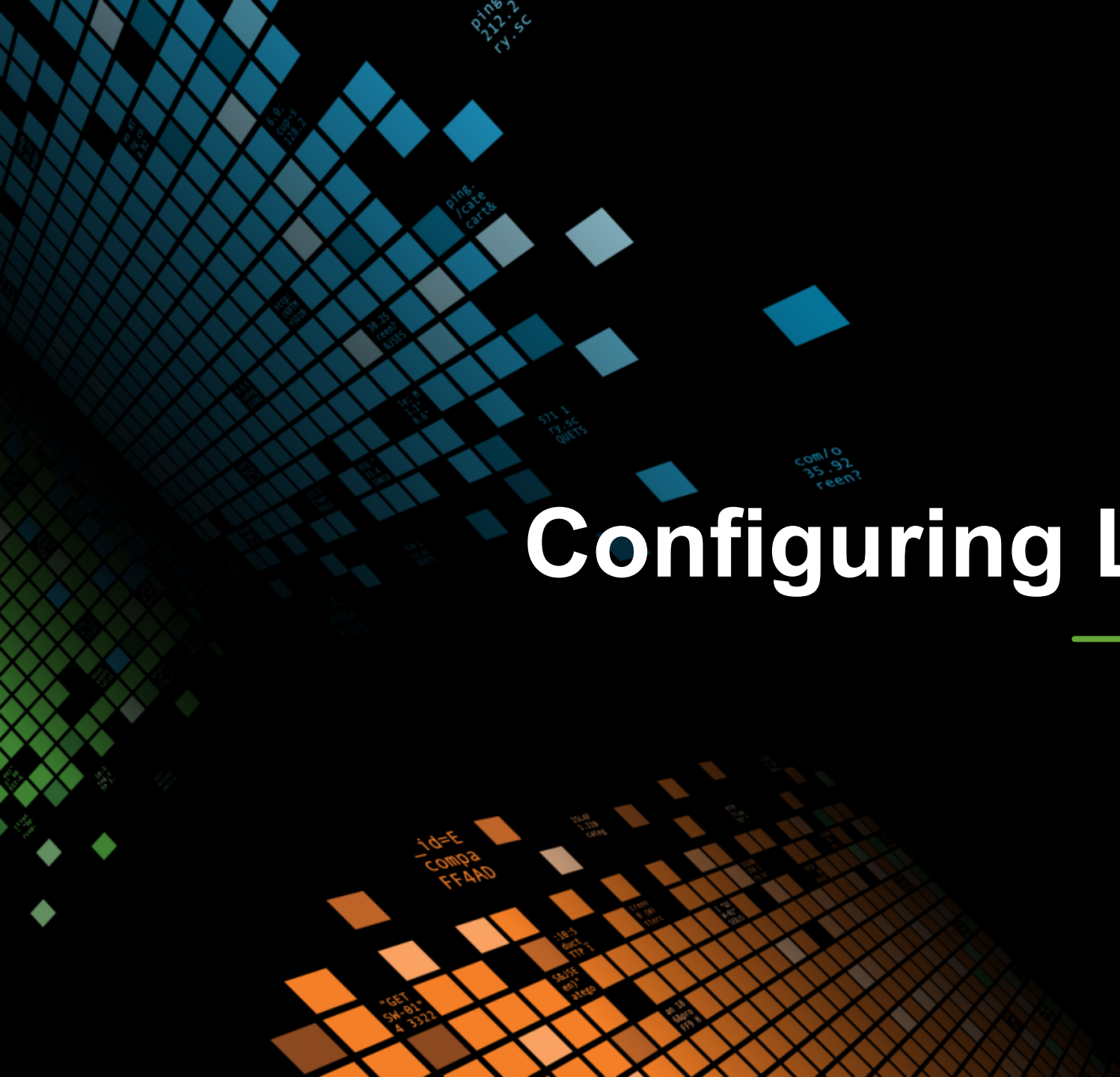
Column Chart Format



searchable

RF+SF Met

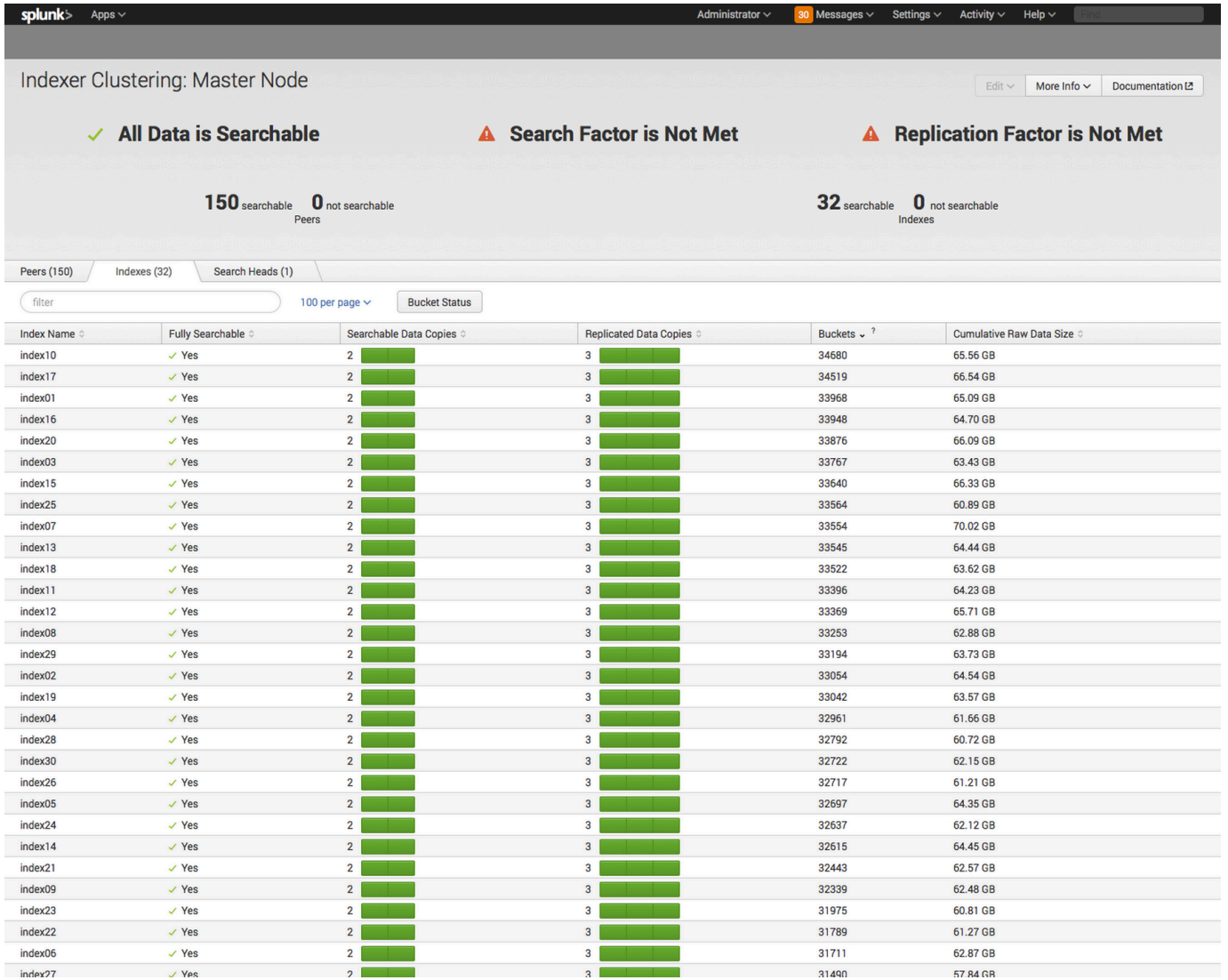
splunk> .conf2017



Configuring Large Clusters

splunk> .conf2017

More Buckets More Problems



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	62.15 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
index27	Yes	2	3	31490	57.84 GB

More Buckets More Settings

server.conf	
cxn_timeout rcv_timeout send_timeout (CM+Indexer)	Specifies how long before an intra-cluster connection will terminate. Default = 60. • 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). • These can be bumped up for busier and larger clusters (300s).
indexes.conf	
rotatePeriodInSecs (Indexer)	Specifies how often to check through all the buckets – rolling them from hot->warm->cold as necessary. Default = 60 • 10min=600

splunk> .conf2017

Splunk 6.6 Scaling Improvements

Since Splunk 6.6, we scale well with 5 million+ buckets. Lots of negative feedback loops have been broken. Busy clusters no longer have their indexers continuously re-add (remove all buckets, add all buckets).

See the talk “**Scaling Indexer Clustering – 5 Million Unique Buckets and Beyond**”

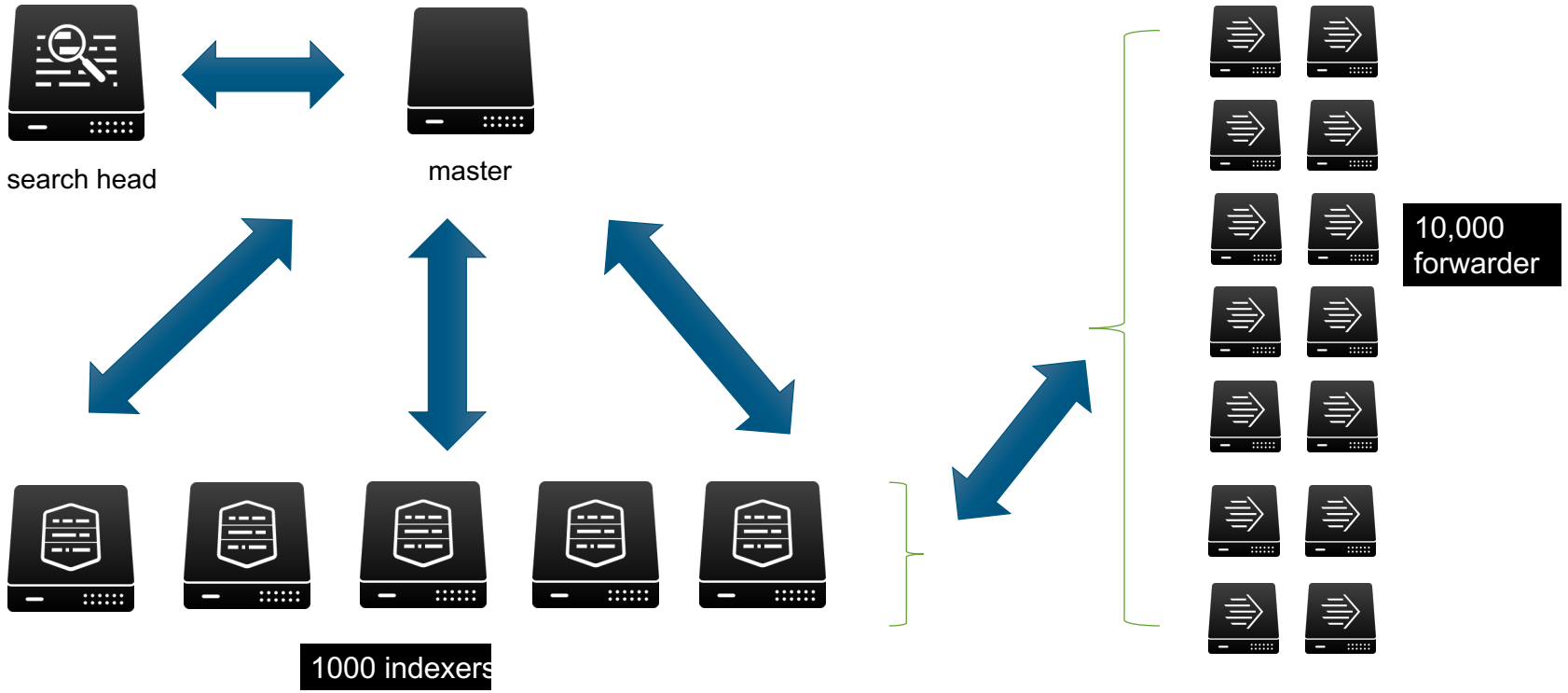
Performance Testing

splunk> .conf2017

Scale Testing

Component	#
Indexers	1,000
Indexes	1,000
Unique buckets	5,000,000
Forwarders	100,000

Test Configuration



Test: Peer Failure

1. Take down a peer

   ./splunk stop –f

2. Peer registers as “Down” on cluster master
3. Wait for:
   - **Searchable**
   - Replication factor met
   - Search factor met

Test: Peer Failure - Results

Splunk Release	Non-clustering/Clustering/ Multi-site	Avg Indexer Throughput	Total Test Time
6.5	Non-clustering	15.07 MB/s	-
6.5	Clustering	11.6 MB/s	261s
6.5	Multi-site	11.7 MB/s	352s
6.6	Non-clustering	15.49 MB/s	-
6.6	Clustering	12.8 MB/s	297.6s
6.6	Multi-site	12.64 MB/s	358.2s

Test: Site Failure

1. Take down 1 site (333 peers)

   ./splunk stop –f

2. All peers register as “Down” on cluster master
3. Wait for:
   - Searchable
     - (we don’t wait for RF/SF since we’ve lost an entire site)

Test: Site Failure - Results

Splunk Release	Non-clustering/Clustering/ Multi-site	Total test time
6.5	Non-clustering	-
6.5	Clustering	-
6.5	Multi-site	7.8s
6.6	Non-clustering	-
6.6	Clustering	-
6.6	Multi-site	8.4s

Test: Master Restart

1. Force restart of master

   ./splunk restart -f

2. Master completes restart
3. Wait for:
   - Searchable
   - Replication factor met
   - Search factor met

Test: Master Restart - Results

Splunk Release	Non-clustering/Clustering/ Multi-site	Total test time
6.5	Non-clustering	-
6.5	Clustering	10.8s
6.5	Multi-site	10.8s
6.6	Non-clustering	-
6.6	Clustering	11.4s
6.6	Multi-site	11.4s

splunk> .conf2017

Test: Rolling Restart

1. Perform rolling restart (on cluster master)

   ./splunk rolling-restart cluster-peers

2. All peers perform restart
3. Wait for:
   - Searchable
   - Replication factor met
   - Search factor met

Test: Rolling Restart- Results

Splunk Release	Non-clustering/Clustering/ Multi-site	Total test time
6.5	Non-clustering	-
6.5	Clustering	370s
6.5	Multi-site	353s
6.6	Non-clustering	-
6.6	Clustering	360.2s
6.6	Multi-site	355.2s

splunk> .conf2017

Resource Utilization

Splunk Release	%CPU (Cluster master)	Memory (Cluster master)	%CPU (indexer)	Memory (indexer)
6.5	1.96%	64.58 MB	232.03%	218.54 MB
6.6	1.97%	67.51 MB	216.85%	203.45 MB

Bundle Push Testing

splunk> .conf2017

Configuration Modifications

Master's server.conf

- Max_peers_to_download_bundle = 0 (default)
- Max_peers_to_download_bundle = 1
- Max_peers_to_download_bundle = 2
- Max_peers_to_download_bundle = 3

Test Results

```
index=main | chart max(total_time) by bundle_size, max_peers_to_download
```

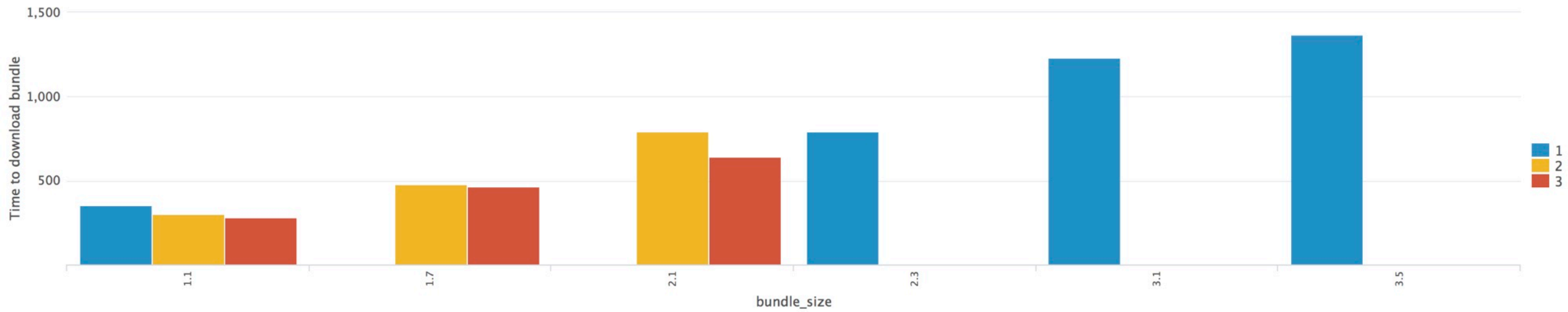
Last 24 hours

✓ 48 events (2/1/17 3:00:00.000 PM to 2/2/17 3:18:08.000 PM) No Event Sampling

Job Smart Mode

Events Patterns Statistics (6) Visualization

Column Chart Format



Validation

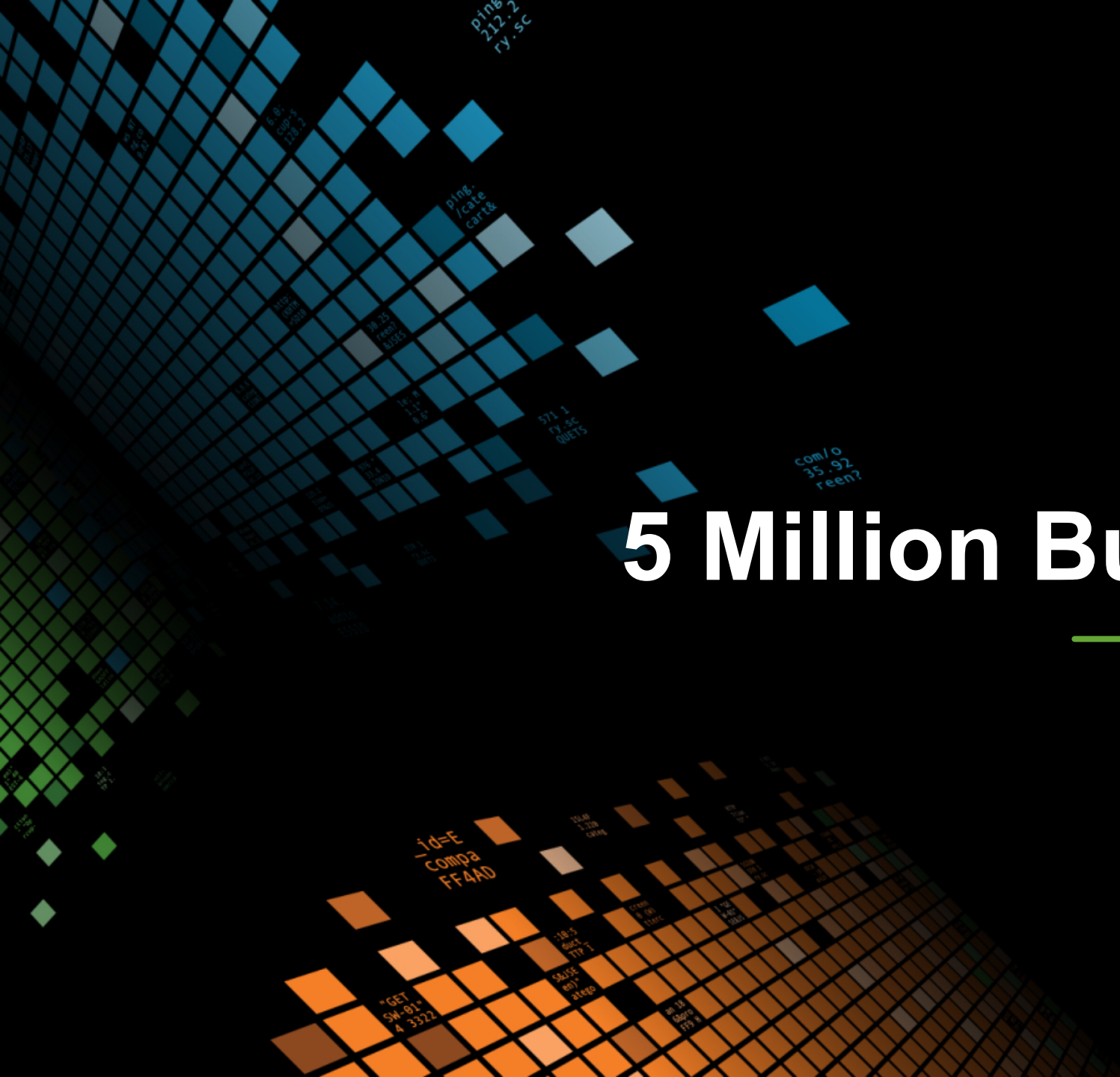
```
master
        active_bundle
               checksum=55BD1789FD910A50C8F245CD9F605F03
               timestamp=1486067118 (in localtime=Thu Feb  2 12:25:18 2017)
        latest_bundle
               checksum=55BD1789FD910A50C8F245CD9F605F03
               timestamp=1486067118 (in localtime=Thu Feb  2 12:25:18 2017)
        last_validated_bundle
               checksum=A92FF82E8AFAD1125783CA3B67D258A3
               last_validation_succeeded=1
               timestamp=1486067118 (in localtime=Thu Feb  2 12:25:18 2017)
        cluster_status=None

 idx_02_204.107.141.240  0A6F2664-4A4F-4DD9-A3DD-BF367A688228    site2
        active_bundle=55BD1789FD910A50C8F245CD9F605F03
        latest_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validated_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validation_succeeded=1
        restart_required_apply_bundle=0
        status=Up

 idx_05_204.107.141.240  1E60D874-687D-40AE-B94C-B63655115849    site2
        active_bundle=55BD1789FD910A50C8F245CD9F605F03
        latest_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validated_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validation_succeeded=1
        restart_required_apply_bundle=0
        status=Up

 idx_09_204.107.141.240  53BE0C69-56EB-4C7D-9107-56A10C5EB934    site3
        active_bundle=55BD1789FD910A50C8F245CD9F605F03
        latest_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validated_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validation_succeeded=1
        restart_required_apply_bundle=0
        status=Up

 idx_08_204.107.141.240  6016D796-91C2-4F36-A546-39D79AB9A774    site2
        active_bundle=55BD1789FD910A50C8F245CD9F605F03
        latest_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validated_bundle=55BD1789FD910A50C8F245CD9F605F03
        last_validation_succeeded=1
        restart_required_apply_bundle=0
        status=Up
```



5 Million Buckets Testing

splunk> .conf2017

5 Million Buckets Cluster

1 Cluster Master
on 1 Machine

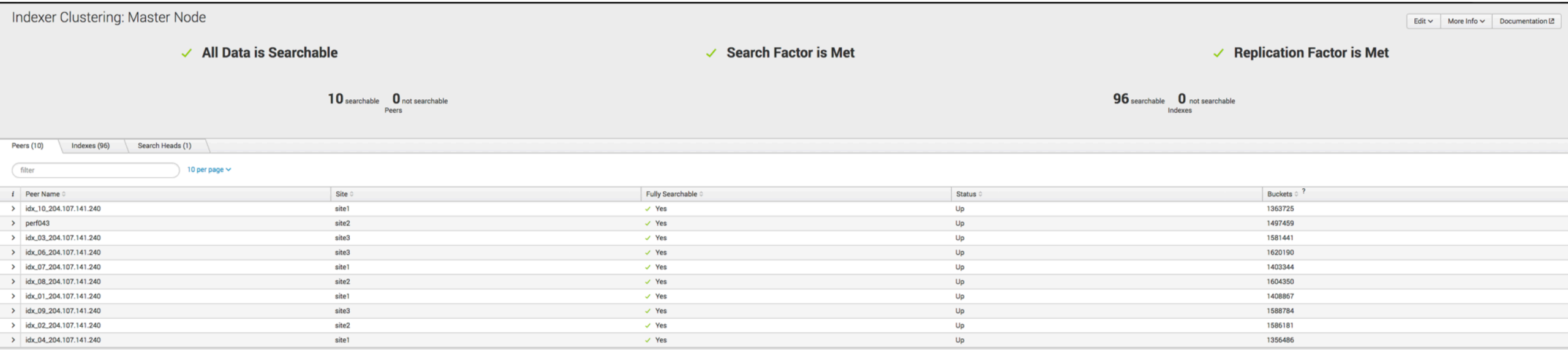
STATUS

10	96	15010827	1236.42 GB
Peers Searchable	Indexes Searchable	Bucket Copies	Rawdata Size

RESOURCE USAGE

CPU		0.18 %
Memory		81.21 %

5 Million Buckets Cluster



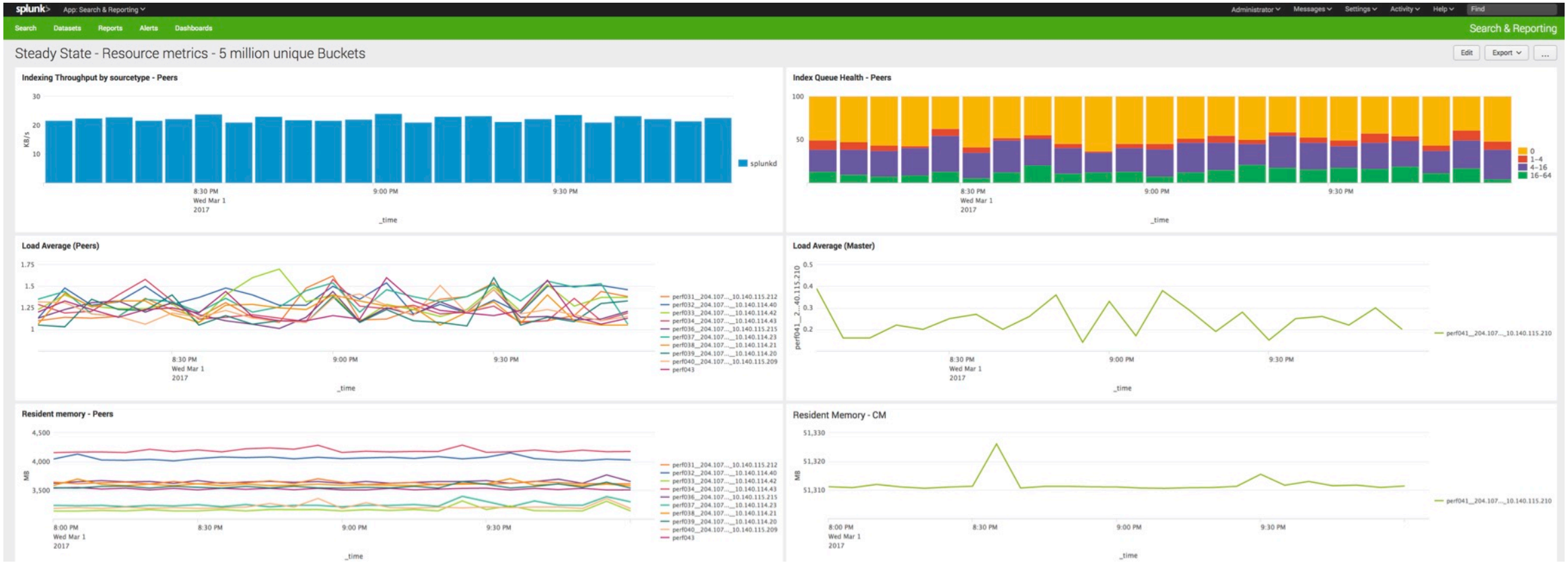
splunk> .conf2017

Regression Test Results

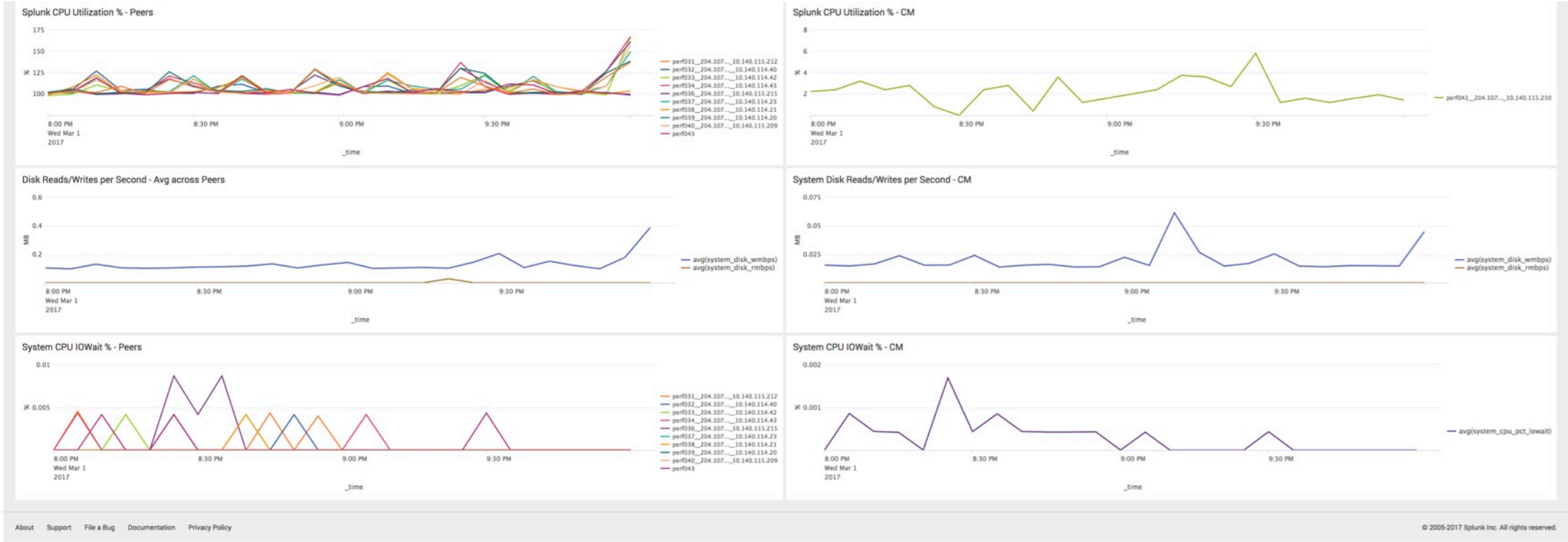
Splunk Release: 6.6

Test	Test time	Time to be searchable	Time to meet rf	Time to meet sf
Peer failure	104.4s	10.3s	-	-
Site failure	196.2s	194.9s	-	-
Master restart	206.4s	159.2s	10.1s	31.5s

Resource Usage



Resource Usage



Thanks - Q&A

Don't forget to rate this session in the .conf2017 mobile app

splunk> .conf2017

APPENDIX Searches

1. index=_internal host=MASTER source=*splunkd.log* CMRepJob running job | timechart count by job
   - Master jobs ran
2. index=_internal source=*metrics.log* name=cmmaster group=jobs | timechart max(CM*)
   - Master jobs metrics
3. index=_internal source=*metrics.log* *fix* host=MASTER | timechart max(to_fix_*)
   - to_fix list sizes
4. index=_internal source=*metrics.log* group=subtask_seconds name=cmmaster | timechart max(service)
   - Master time spent calling service() in between previous log to metrics.log (every 30s)