

Scaling Indexer Clustering

5 Million Unique Buckets and Beyond

Cher-Hung Chang | Principal Software Engineer

Tameem Anwar | Software Engineer

09/26/2017 | 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.

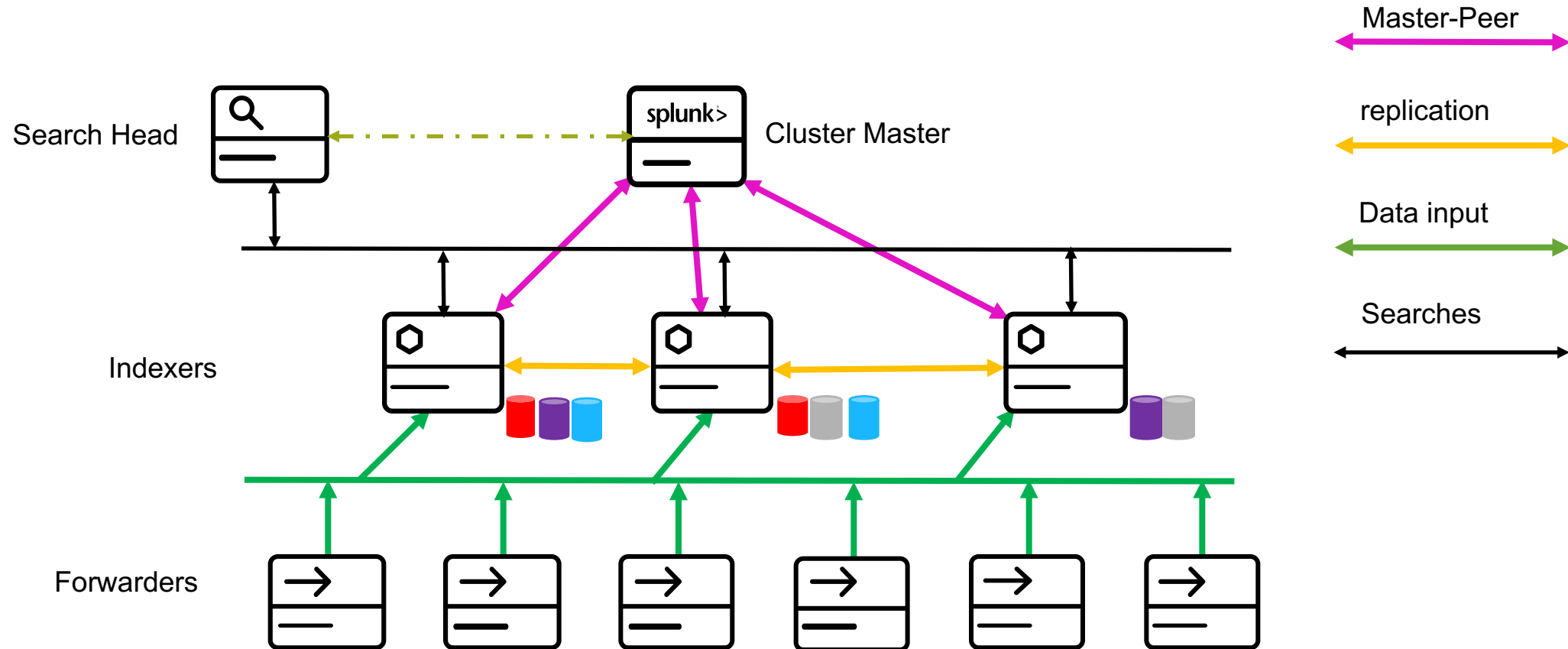
Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2017 Splunk Inc. All rights reserved.

Agenda

1. Introduction to Indexer Clustering
2. Scalability Improvements
3. Performance

Introduction to Indexer Clustering

Indexer Clustering Topology



Splunk
Server

Cluster Nodes

► Cluster Master (CM)

- Stateless
- Maintains in-memory state of all the peers and buckets
- Coordinates the replicating activities of the peer nodes
- Tells the search head where to find data

► Cluster Peer (Indexer)

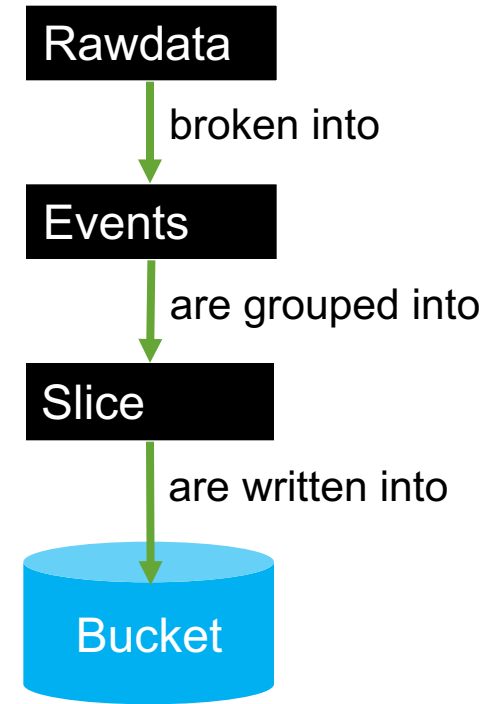
- Indexes, replicates data
- Responds to the incoming search requests
- Report its state and all its buckets to CM

► Search head

► Forwarder

Buckets

- ▶ Unit of data the cluster is aware of
 - ▶ Created on the indexer
 - ▶ Indexer notifies CM upon every state transition of its bucket
 - ▶ Configurable size
 - ▶ Bucket manipulation is centric of CM's responsibility
 - ▶ More data, more buckets
- ▶ Flow of bucket creation
-
- ```
graph TD; Rawdata[Rawdata] -- "broken into" --> Events[Events]; Events -- "are grouped into" --> Slice[Slice];
```



# Heartbeats

- ▶ One of the mechanisms CM uses to communicate with peers
- ▶ Status synchronization
- ▶ Once peer registers to master, it starts to heartbeat to master every `heartbeat\_period` seconds (defaults to 1)
- ▶ CM utilizes `heartbeat\_timeout` to consider if peer is offline and perform fixup if necessary

# Scalability Improvements

---

# Quick Glance of Improvements since 6.6

- ▶ 3x more cluster-wide buckets (vs 6.5)
  - 15M cluster-wide buckets, or 150 PB data (10GB/bucket)
- ▶ Faster cluster management with scale
  - rolling restart: ~2x faster
  - peer failure recovery: ~2x faster
  - Improved responsiveness
    - Better average turn-around time: 2x faster and more
  - and more...
- ▶ Less memory (vs Splunk 6.6)
  - ~15% reduction in CM

# Challenges to scale up

## ► More data, more buckets

- more complexity
- more fixes to reach healthy state
- more memory consumed, less cache hit-rate, higher cache miss penalty

## ► Extreme conditions are difficult to fulfill in time

- Long-running transactions
- When CM's CPU is burned out, requests could be possibly timeout and result in failure recovery
  - Add into cluster, huge amount of fix jobs
  - More fix ups, higher demand of CPU
- False positive scenarios due to above reasons
  - unnecessary fixes

## ► Network latency

# Improvements

- ▶ Peer adding with configurable amount of buckets
- ▶ Limit job processing time
- ▶ Lockless heartbeat
- ▶ Enhanced bucket management

130.60.4 - - [07/Jan 18:10:57:153] "GET /category.screen?category\_id=GIFTS&SESSIONID=5D1SLAFF10ADFF10 HTTP 1.1" 404 720 "http://buttercup-shopping.com/cart.do?action=view&itemId=EST-6&product\_id=F1-SW-03"  
128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D3SL7FF6ADFF0 HTTP 1.1" 404 3322 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=F1-SW-03"  
317 27.160.0.0 - - [07/Jan 18:10:56:156] "GET /oldlink?item\_id=EST-26&SESSIONID=5D5SL9FF1ADFF3 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D18SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=F1-SW-03"  
100.60.4 - - [07/Jan 18:10:57:153] "GET /category.screen?category\_id=GIFTS&SESSIONID=5D1SLAFF10ADFF10 HTTP 1.1" 404 720 "http://buttercup-shopping.com/cart.do?action=view&itemId=EST-6&product\_id=F1-SW-03"  
128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D3SL7FF6ADFF0 HTTP 1.1" 404 3322 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=F1-SW-03"  
317 27.160.0.0 - - [07/Jan 18:10:56:156] "GET /oldlink?item\_id=EST-26&SESSIONID=5D5SL9FF1ADFF3 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D18SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=F1-SW-03"  
100.60.4 - - [07/Jan 18:10:57:153] "GET /category.screen?category\_id=GIFTS&SESSIONID=5D1SLAFF10ADFF10 HTTP 1.1" 404 720 "http://buttercup-shopping.com/cart.do?action=view&itemId=EST-6&product\_id=F1-SW-03"  
128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D3SL7FF6ADFF0 HTTP 1.1" 404 3322 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=F1-SW-03"  
317 27.160.0.0 - - [07/Jan 18:10:56:156] "GET /oldlink?item\_id=EST-26&SESSIONID=5D5SL9FF1ADFF3 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D18SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=F1-SW-03"

► Splunk 6.5 and prior

- ## ► Splunk 6.6+

- Configurable amount of buckets to replace long-running add-peer
- Reduce the requirement of CM computing power
- Reduce the variation caused by network latency
- Better turn-around time
- More responsive CM

# Peer adding - configurable amount of buckets

**Some Data is Not Searchable**

**Search Factor is Not Met**

**Replication Factor is Not Met**

5 searchable

5 not searchable  
Peers

0 searchable

96 not searchable  
Indexes

Peers (10)

Indexes (96)

Search Heads (1)

10 per page ▾

| i | Peer Name ▾            | Site ▾ | Fully Searchable ▾ | Status ▾    | Buckets ▾ ? |
|---|------------------------|--------|--------------------|-------------|-------------|
| > | perf086                | site3  | No                 | BatchAdding | 1564000     |
| > | idx_08_204.107.141.240 | site2  | ✓ Yes              | Up          | 1625673     |
| > | perf089                | site3  | No                 | BatchAdding | 1483000     |
| > | perf084                | site1  | ✓ Yes              | Up          | 1342597     |

# Peer adding - configurable amount of buckets

- ▶ `buckets_per_addpeer` = <non-negative integer> in `server.conf`
- ▶ Defaults to 1000
- ▶ The more buckets in add-peer, the quicker it adds to cluster
  - Needs larger REST receive timeout `rcv_timeout` and heartbeat timeout `heartbeat_timeout`
  - Less responsive CM
  - Less up-to-date cluster
- ▶ Needs tweaking along with
  - Computing power of CM
  - Related timeout settings
  - Number of buckets
  - Number of indexers

# Limit job processing time

► Splunk 6.5 and prior

- CM used to process all jobs within a single service duration, which could be many seconds or even minutes
  - Replication (to meet RF)
  - Search (to meet SF)
  - Primality (all buckets need to have a primary copy per site)
  - Others (freezing, checksum, rolling, etc)

► Splunk 6.6+

- An option to limit how long CM spends per fixup
- CM is forced to context switch to serve other requests
- More responsive CM, however fixups take longer

# Limit job processing time

- ▶ `max_fixup_time_ms` = <zero or positive integer>
- ▶ Defaults to 0 (unused)
- ▶ Limits how long each fixup level runs
- ▶ Useful on larger clusters having massive amount of buckets
- ▶ Enable it when there are massive amount of buckets, and cluster failures result in extra busy cluster master fixup activity (where service duration takes more than 10 seconds)

► Splunk 6.5 and prior

- ## ► Splunk 6.6+

- 

# Enhanced bucket management

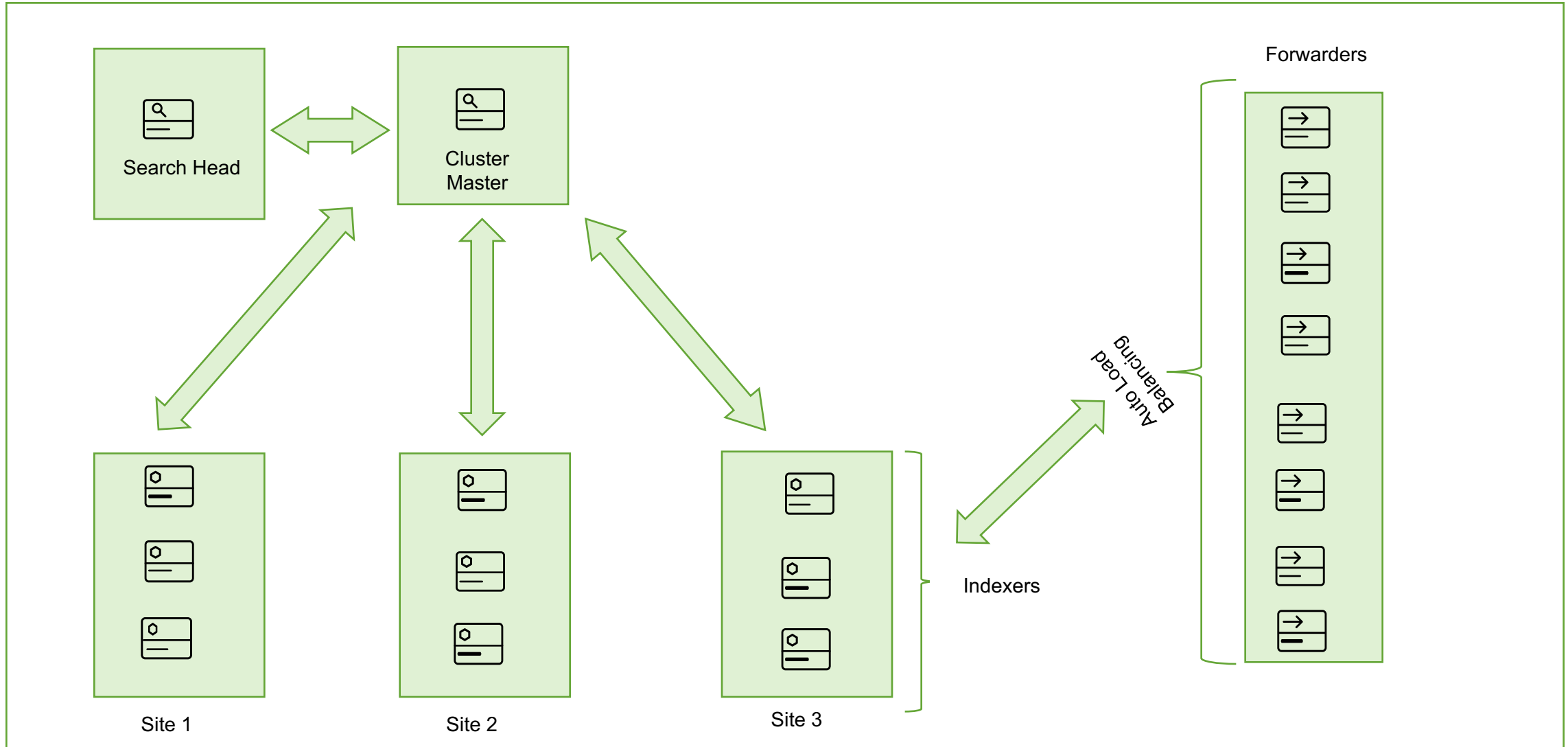
- ▶ Bucket is centric of CM's responsibility
  - Bucket manipulation is required by mostly all processing in CM
  - Improving it will generally improve all critical paths
- ▶ Available since Splunk 7.0
- ▶ Bonus: ~15% memory saving in master node

| Test Case            | Splunk 6.6 | Splunk 7.0 |
|----------------------|------------|------------|
| 1M bucket Insertions | 60ms       | 2.42ms     |
| 1M bucket Lookups    | 60ms       | 2.26ms     |

# Indexer Clustering Performance

Scale tests with 5 Million Unique Buckets

# Deployment

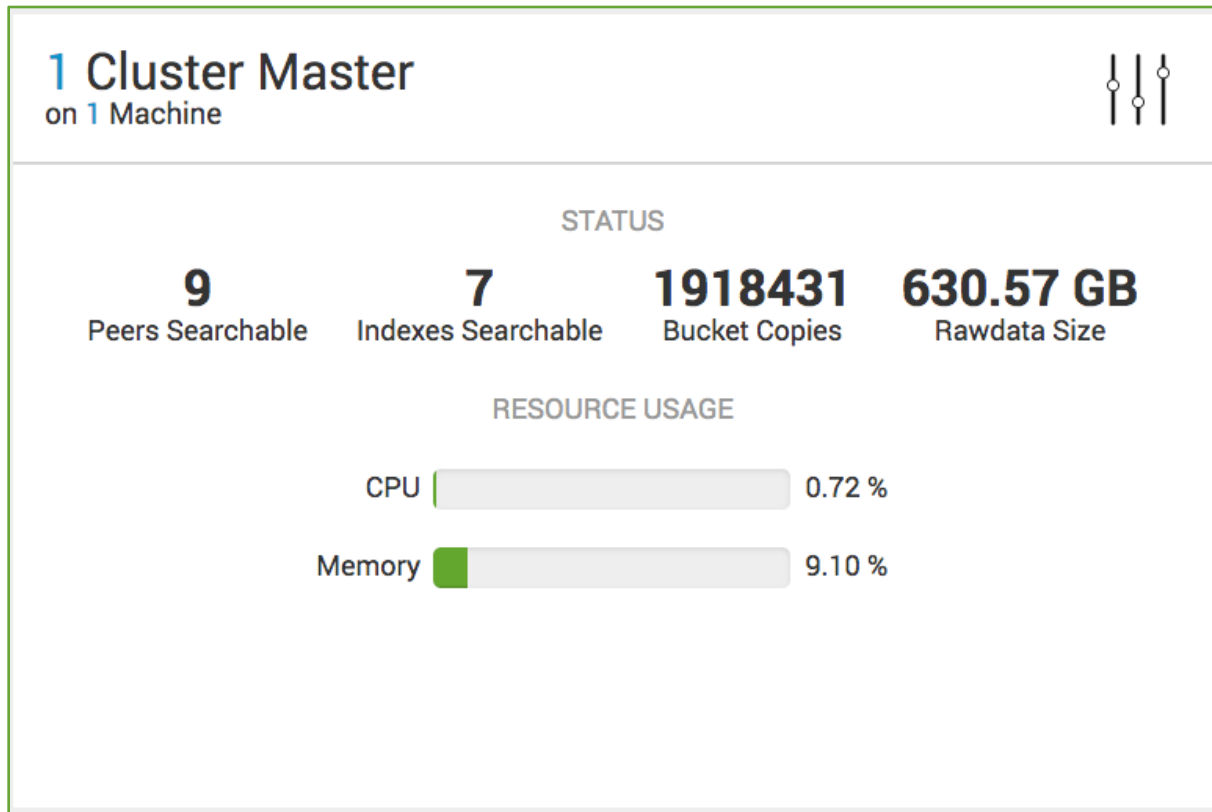


# Test Configuration

- Machine Specs
  - 2x12 Xeon 2.30 GHz
  - 24 cores (48 w/HT)
  - 128 GB RAM
  - 8 x 300GB 15k RPM disks in RAID-0
  - 1 Gb Ethernet NIC
  - CentOS 7.6
- Multi-site Cluster Configuration
  - 3 sites
  - Replication factor – origin:2, total:3
  - Search factor – origin:1, total:2
- No other load on the box

# 1.9 Million Bucket Test

Splunk 6.5 vs Splunk 7.0



- ▶ 640,000 unique buckets
- ▶ 200,000 buckets/peer
- ▶ Default cluster timeout configurations
- ▶ No Search or Indexing load

# Test Results

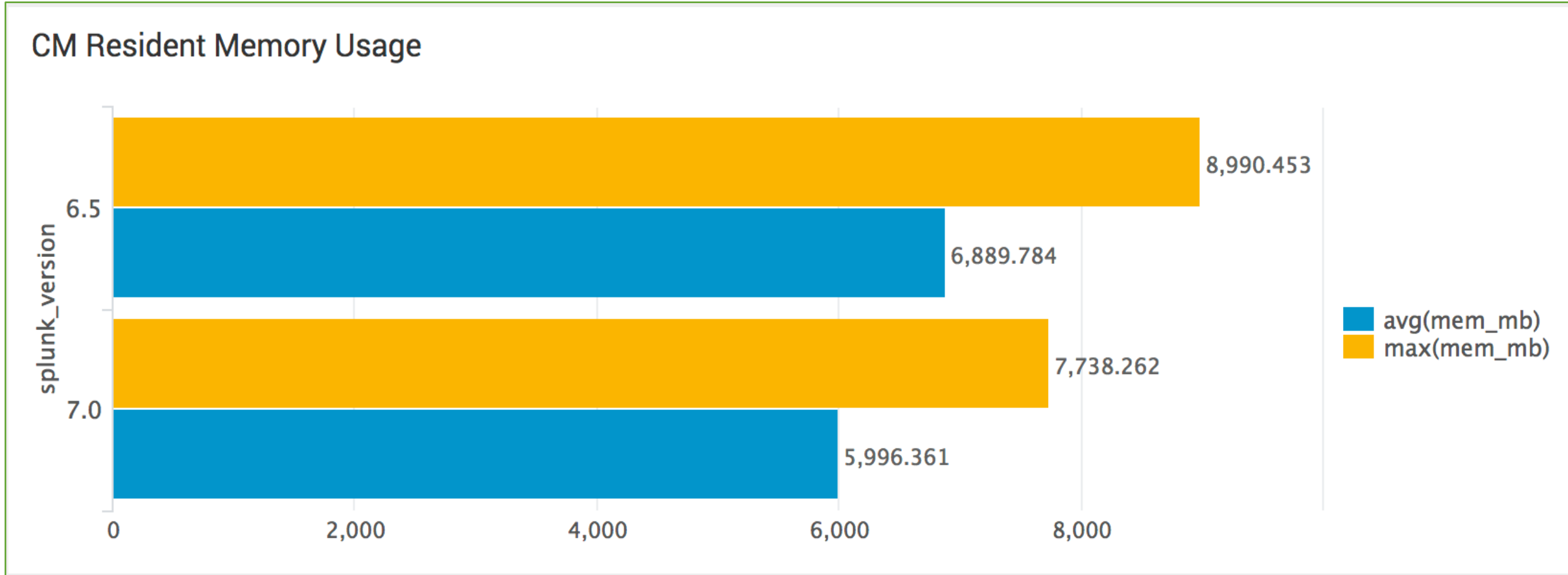
| Test Case       | Success Condition  | Time for Completion |               |
|-----------------|--------------------|---------------------|---------------|
|                 |                    | Splunk 6.5          | Splunk 7.0    |
| CM Restart      | Cluster Complete   | 43.08 secs          | 36.7 secs     |
| Peer Failure    | Cluster Searchable | 60 + 73.8 secs      | 60 + 8.6 secs |
| Rolling Restart | Cluster Complete   | 19.1 mins           | 10.1 mins     |
| Bundle Push     | 100 MB Pushed      | 49.2 secs           | 25.8 secs     |

130.60.4 - - [07/Jun 18:10:57:153] "GET /category.screen?category\_id=GIFTS&SESSIONID=5D1SLAFF10ADFF10 HTTP 1.1" 404 720 "http://buttercup-shopping.com/cart.do?action=view&itemId=EST-6&product\_id=FI-SW-03" "Opera/9.80 (Win  
128.241.220.82 - - [07/Jun 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D3SL7FF6ADFF9 HTTP 1.1" 404 3322 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=KQ-CW-01" "Mozilla/4.0 (Compaq i486)  
ows NT 5.1; SV1; .NET CLR 1.1.4322" 468 125.17 14.0 "http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D1SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=remove&itemId=EST-3&product\_id=KQ-CW-01" "Mozilla/4.0 (Compaq i486)  
itemId=EST-16&product\_id=RP-LI-02" 468 125.17 14.0 "http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D1SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=remove&itemId=EST-3&product\_id=KQ-CW-01" "Mozilla/4.0 (Compaq i486)  
ofaction=purchase&itemId=EST-26&product\_id=KQ-CW-01" 468 125.17 14.0 "http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D1SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=remove&itemId=EST-3&product\_id=KQ-CW-01" "Mozilla/4.0 (Compaq i486)  
http://buttercup-shopping.com/cart.do?action=changequantity&itemId=EST-18&product\_id=AV-CB-01&SESSIONID=5D1SL8FF3ADFF9 HTTP 1.1" 200 2423 "http://buttercup-shopping.com/cart.do?action=remove&itemId=EST-3&product\_id=KQ-CW-01" "Mozilla/4.0 (Compaq i486)

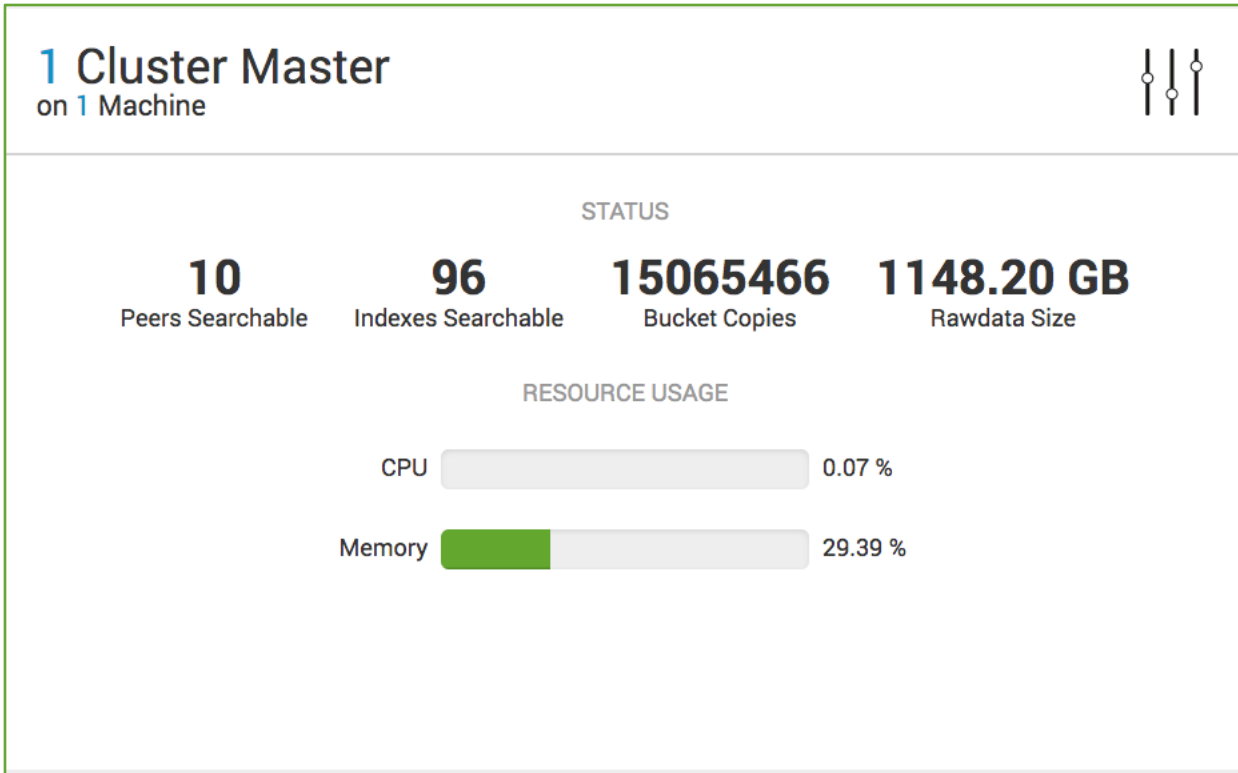
# Improved CM responsiveness

| Endpoints                           | Average Time (milliseconds) |            |
|-------------------------------------|-----------------------------|------------|
|                                     | Splunk 6.5                  | Splunk 7.0 |
| /services/cluster/master/generation | 778 ms                      | 262 ms     |
| /services/cluster/master/peers      | 1972 ms                     | 119 ms     |
| /services/cluster/master/buckets    | 818 ms                      | 440 ms     |
| /services/cluster/master/info       | 4293 ms                     | 1796 ms    |
| /services/cluster/master/indexes    | 1874 ms                     | 598 ms     |

# CM Memory Footprint



# 15 million Bucket Test



- ▶ 5 million unique buckets
- ▶ 3x more buckets cluster-wide compared to Splunk 6.5
- ▶ No Search or Indexing Load

# Buckets per Peer

## Indexer Clustering: Master Node

Edit ▾

More Info ▾

Documentation ↗

✓ All Data is Searchable

✓ Search Factor is Met

✓ Replication Factor is Met

**10** searchable **0** not searchable  
Peers

**96** searchable **0** not searchable  
Indexes

Peers (10)

Indexes (96)

Search Heads (1)

filter

10 per page ▾

| i | Peer Name ▾            | Site ▾ | Fully Searchable ▾ | Status ▾ | Buckets ▾ ? |
|---|------------------------|--------|--------------------|----------|-------------|
| > | perf086                | site3  | ✓ Yes              | Up       | 1767520     |
| > | idx_08_204.107.141.240 | site2  | ✓ Yes              | Up       | 1625686     |
| > | perf089                | site3  | ✓ Yes              | Up       | 1749115     |
| > | perf084                | site1  | ✓ Yes              | Up       | 1348108     |
| > | perf090                | site1  | ✓ Yes              | Up       | 1432264     |
| > | perf082                | site2  | ✓ Yes              | Up       | 1621469     |
| > | perf085                | site2  | ✓ Yes              | Up       | 1560173     |
| > | perf081                | site1  | ✓ Yes              | Up       | 1344318     |
| > | perf087                | site1  | ✓ Yes              | Up       | 1449131     |
| > | perf083                | site3  | ✓ Yes              | Up       | 1167682     |

130.60.4 - - [07/Jan 18:10:57:153] "GET /category.screen?category\_id=GIFTS&SESSIONID=5D5LAF10ADFF10 HTTP 1.1" 404 720 "http://buttercup-shopping.com/cart.do?action=view&itemId=EST-6&product\_id=F1-SW-03" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

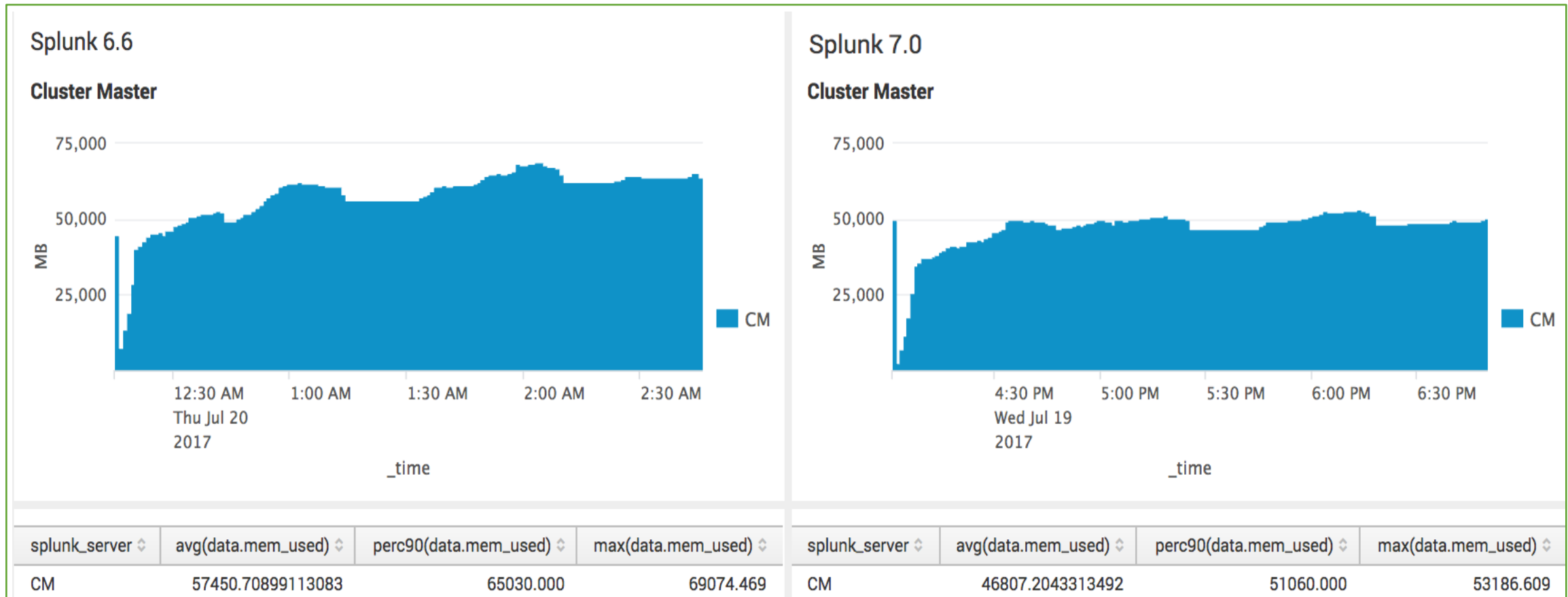
128.241.220.82 - - [07/Jan 18:10:57:123] "GET /product.screen?product\_id=FL-DSH-01&SESSIONID=5D5SL7FF6ADFF9 HTTP 1.1" 200 1318 "http://buttercup-shopping.com/cart.do?action=purchase&itemId=EST-26&product\_id=FL-DSH-01" "Opera/9.80 (Windows NT 5.1; SV1; .NET CLR 1.1.4322)" 468 125.17 14.000

# Test Results

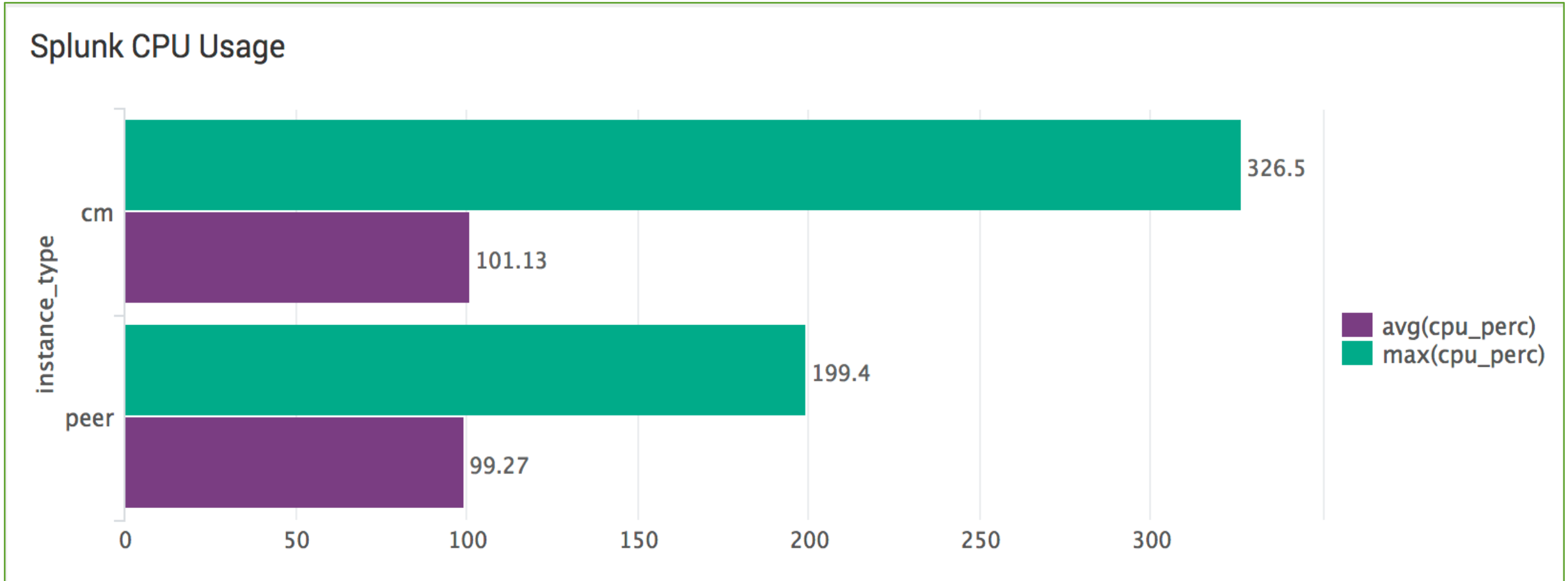
| Test Case       | Success Condition  | Time for Completion |
|-----------------|--------------------|---------------------|
| CM Restart      | Cluster Complete   | 5.73 mins           |
| Peer Failure    | Cluster Searchable | 600 + 140.1 secs    |
| Rolling Restart | Cluster Complete   | 31.91 mins          |
| Bundle Push     | 100 MB Pushed      | 112.2 secs          |

# CM Memory Footprint

- 18.5 % reduction in avg. Resident Memory & 23 % reduction in max Resident Memory



# Splunk CPU Usage



# Cluster Configuration

## Cluster Master - server.conf

### [clustering]

- heartbeat\_timeout = 600
- rcv\_timeout = 600
- send\_timeout = 600
- cxn\_timeout = 600
- max\_fixup\_time\_ms = 5000

## Indexers - server.conf

### [clustering]

- heartbeat\_period = 40
- cxn\_timeout = 600
- send\_timeout = 600
- rcv\_timeout = 600
- buckets\_per\_addpeer = 1000

# Key Takeaways

1. Enhancements - Peer adding with configurable amount of buckets, Lockless heartbeat, Limit job processing time etc.
2. CM scales up to 5+ million unique buckets (3x more compared to Splunk 6.5)
3. Faster Cluster recovery and Improved CM responsiveness.
4. ~15% CM memory reduction.

# Thank You

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

splunk® .conf2017