

Couchbase

MEMBASE: CLUSTERED BY ERLANG

ERLANG FACTORY SF 2011
SEAN LYNCH
MATT INGENTHRON



Agenda

Membase at a high level

Let's create a cluster

What we needed when building

What we built

Lessons learned

WHAT IS MEMBASE?



Membase is a distributed database

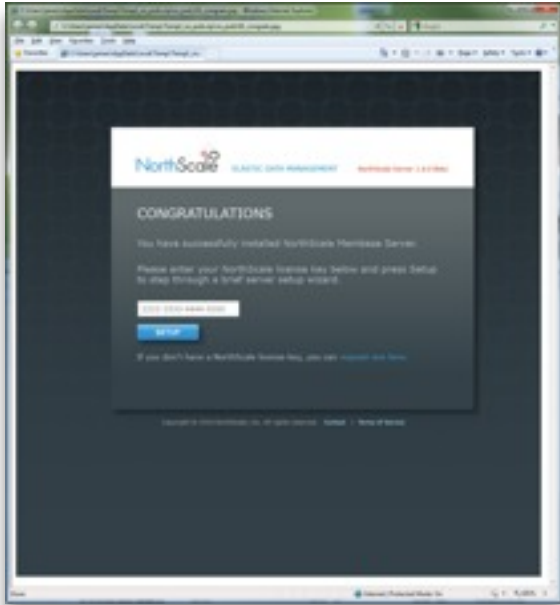


In the data center



On the administrator console

Membase is Simple, Fast, Elastic



- ✧ Five minutes or less to a working cluster
 - Downloads for Linux and Windows
 - Start with a single node
 - One button press joins nodes to a cluster
- ✧ Easy to develop against
 - Just SET and GET – no schema required
 - Drop it in. 10,000+ existing applications already “speak membase” (via memcached)
 - Practically every language and application framework is supported, out of the box
- ✧ Easy to manage
 - One-click failover and cluster rebalancing
 - Graphical and programmatic interfaces
 - Configurable alerting

Membase is Simple, Fast, Elastic



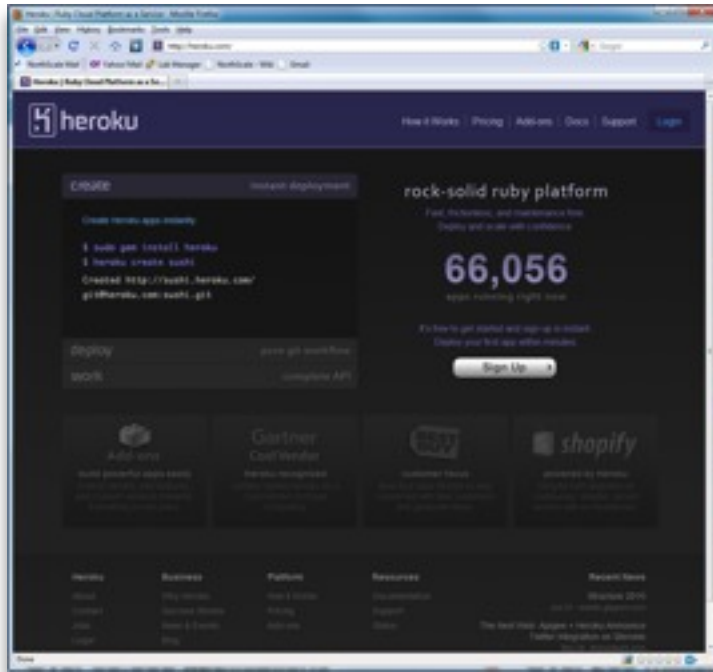
- ✧ Predictable
 - “Never keep an application waiting”
 - Quasi-deterministic latency and throughput
- ✧ Low latency
 - Built-in Memcached technology
- ✧ High throughput
 - Multi-threaded
 - Low lock contention
 - Asynchronous wherever possible
 - Automatic write de-duplication

Membase is Simple, Fast, Elastic



- ❖ Zero-downtime elasticity
 - Spread I/O and data across commodity servers (or VMs)
 - Consistent performance with linear cost
 - Dynamic rebalancing of a live cluster
- ❖ All nodes are created equal
 - No special case nodes
 - Any node can replace any other node, online
 - Clone to grow
- ❖ Extensible
 - Filtered TAP interface provides hook points for external systems (e.g. full-text search, backup, warehouse)
 - Data bucket – engine API for specialized container types

Deployments Leading Membase



- Leading cloud service (PAAS) provider
- Over 65,000 hosted applications
- **Membase Server** serving over 1,200 Heroku customers (as of June 10, 2010)



- Social game leader – FarmVille, Mafia Wars, Café World
- Over 230 million monthly users
- **Membase Server** is the 500,000 ops-per-second database behind FarmVille and Café World

DEMO: BUILDING A CLUSTER



ERLANG OTP: RAW MATERIAL FOR BUILDING A CLUSTER

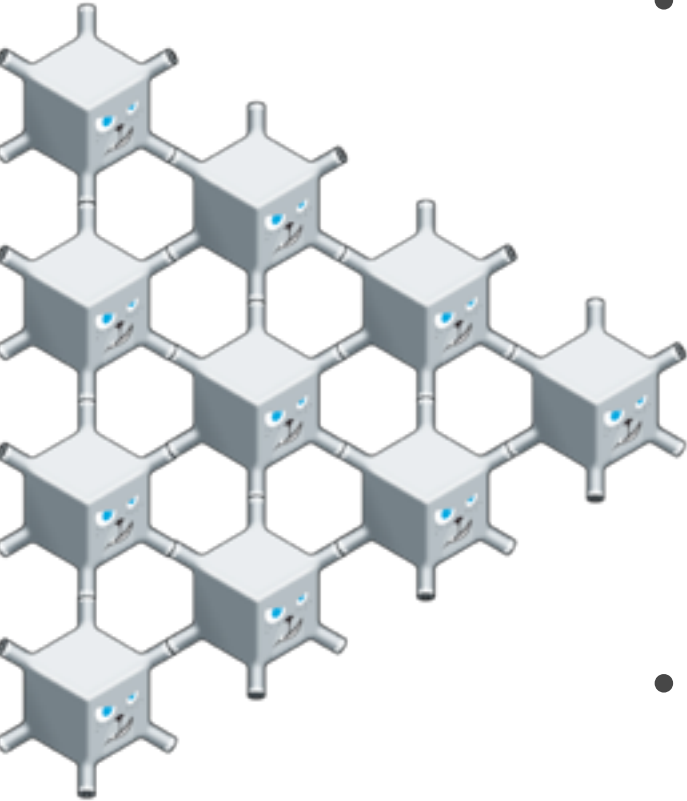


Building a Cluster: Supervisors and Heartbeat

- Our own Supervisors and hierarchy
 - Minorly modified C processes
 - Monitor OS processes as Erlang processes with `ns_port_server`
 - Supervisor cushion
 - Slow down fast startup failures while keeping normal exit/crash fast
- Custom 'heartbeat'
 - Determine failure and gather system resource basics, versions



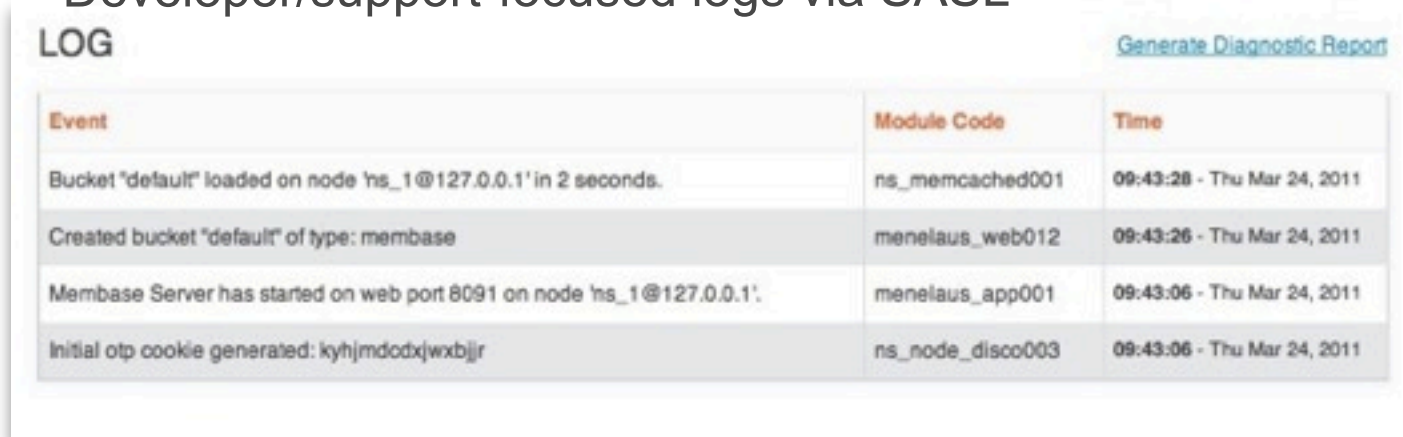
Building a Cluster: Key Modules



- The main cluster module:
`ns_cluster`
 - Joining is more than just ping/pong
 - Leaving, need to leave your cookie and config behind
 - Removing a node is actually done when it rejoins, it sees it's not in the config
- `dist_manager`
 - Change nodenames

Building a Cluster: Key Modules

- User-visible logs for cluster: ns_log
 - Same messages show on all nodes.
 - Currently gossiped/merged among all nodes. In the future, will be replicated among a subset with CouchDB.
 - Developer/support-focused logs via SASL



LOG [Generate Diagnostic Report](#)

Event	Module Code	Time
Bucket "default" loaded on node 'ns_1@127.0.0.1' in 2 seconds.	ns_memcached001	09:43:28 - Thu Mar 24, 2011
Created bucket "default" of type: membase	menelaus_web012	09:43:26 - Thu Mar 24, 2011
Membase Server has started on web port 8091 on node 'ns_1@127.0.0.1'.	menelaus_app001	09:43:06 - Thu Mar 24, 2011
Initial otp cookie generated: kyh mdodx wxbj r	ns_node_disco003	09:43:06 - Thu Mar 24, 2011

- tick_server sending a clock for the cluster
 - Needed a cluster wide centralized time service to synchronize stats from separate nodes
 - Punted on attempting to synchronize node clocks
 - Time can shift if ns_tick server moves

Building a Cluster: Processes once in a cluster

- Singleton Processes
 - Originally called “global singleton”... jokingly... but it stuck
 - Not everything runs everywhere, requirement to run some processes in only one place
 - Used for cluster orchestration
 - Rebalancing, kicking off janitors, etc.
 - Used for tick server

LESSONS LEARNED



Lessons Learned: Networking, other Surprises

- Networks are more fluid
 - Developer laptops
 - Cloud compute environments
- Anyone need some I/O?
 - Look for the +A
 - “+A size: Sets the number of threads in async thread pool, valid range is 0-1024. Default is 0.”
- Sends to remote nodes can block

Lessons Learned: Platform Resources, Maturity

- os_mon, disk_mon
 - Virtual is still virtual
 - Disk info not quite what we needed
- Own log handler due to 64k size maximum log entry size
- Joining a cluster when you are a cluster of one

THE FUTURE



What is Couchbase?



The market's leading **caching** and **clustering** technology.

+



The most reliable and full-featured **document database**.

=



The fastest, most complete and most reliable NoSQL database on the planet.

1+1 really does equal 3

Couchbase

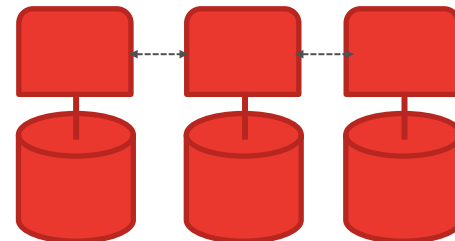
- Feb 08: Merger Announced
- March 10: Developer Preview of Couchbase Mobile - Apache CouchDB for iOS devices
- March 15: Couchbase 1.1 released
 - Includes CouchDB++, GeoCouch, ready for support
- New Membase update in the summer
 - Integrate some features from CouchDB
- Elastic Couchbase later in the year



Mobile Couchbase



Couchbase



Elastic Couchbase*

Q & A



COUCHBase

Data management for interactive web and mobile applications.

SEAN LYNCH
SEAN@COUCHBASE.COM
IRC: FREAKAZOID
@DRPRETTYBAD

MATT INGENTHRON
MATT@COUCHBASE.COM
@INGENTHR

