RAMCloud Update

SEDCL Retreat
June, 2012

John Ousterhout

Stanford University

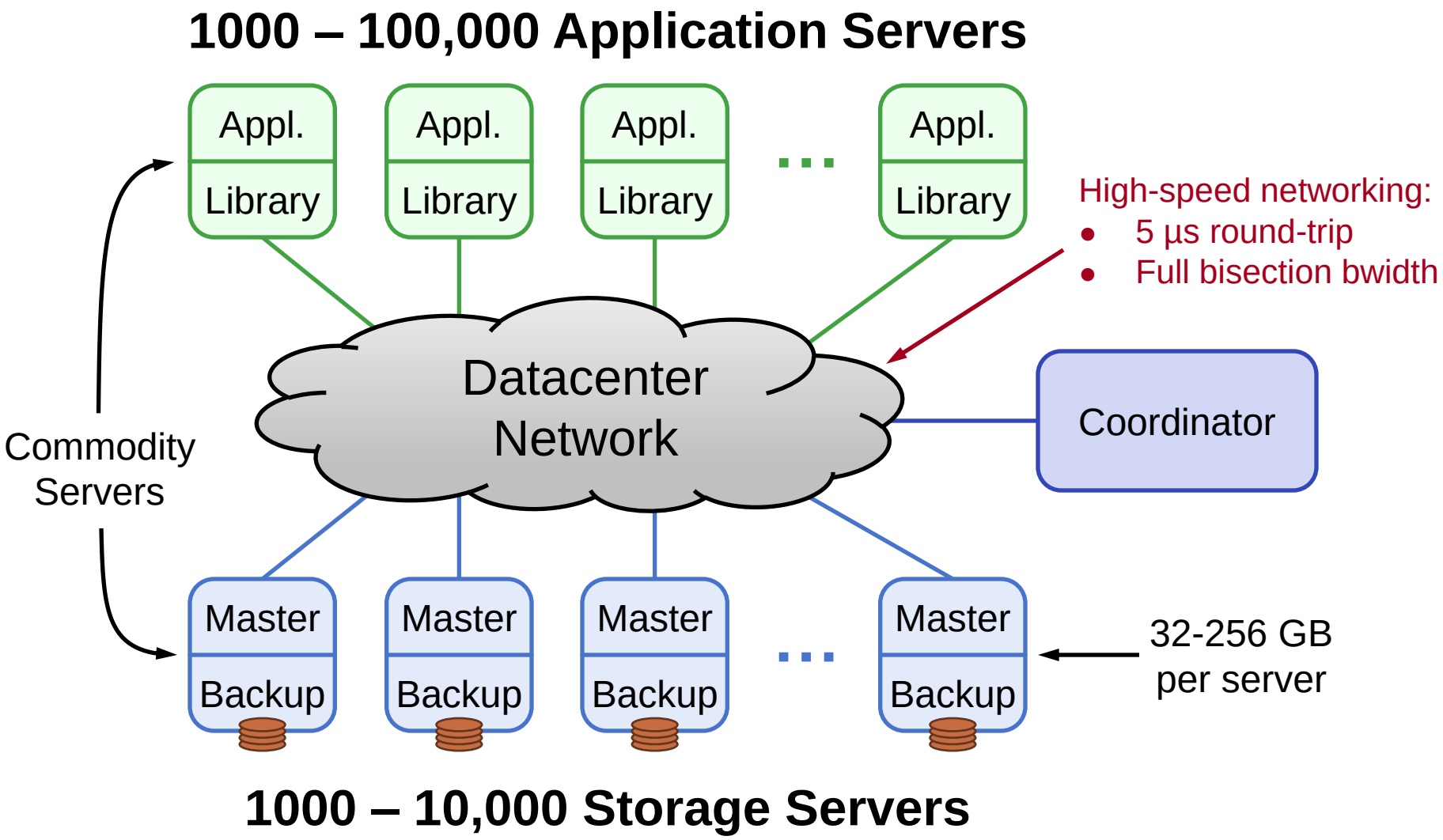
What is RAMCloud?

General-purpose storage system for large-scale applications:

- **All data is stored in DRAM at all times**
- **Large scale: 1000+ servers, 100+ TB**
- **Low latency: 5-10 µs remote access time**
- **As durable and available as disk**
- **Simple key-value data model (for now)**

Project goal: enable a new class of data-intensive applications

RAMCloud Architecture



Status at May 2011 Retreat

- **Skeletal prototype running:**
  - Read/write operations working
    Read 100 bytes in 5.2 µs
  - Recovery implemented for masters
    Recover 6 GB from crashed server in 1.1 seconds with 33 nodes
- **System not yet usable, many features missing:**
  - Recovery incomplete:
    - No recovery from backup crashes
    - RPC system can't handle crashes
    - No cluster membership management
    - No cold start
  - Cleaner not yet usable
  - Coordinator still in skeletal form, not robust
  - Single-threaded

Progress Since May 2011

- **Overall goal: push towards a 1.0 release:**
  - "Least usable system" for real applications
- **Cluster upgrade:**
  - Increase from 40 -> 80 nodes
  - Flash memory for backups
  - Now recovering 35 GB in 1.6 seconds with 60 nodes
- **Major revision of log cleaner:**
  - First usable version
  - Operates in parallel with reads and writes
  - Two-level approach: higher memory utilization without overloading disks
  - Steve will present performance measurements
- **Added multi-threading in servers**

RAMCloud Progress, cont'd

- **Improved crash recovery:**
  - Cluster membership management
    - Detect server failures
    - Disseminate information about server entries/exits
  - New architecture for replica management
    - Handle backup failures
    - Use new threading facilities, cluster membership
  - RPC transports report timeouts gracefully (but no retry yet)
- **Switched to variable-length keys**
- **Performance tools:**
  - 2 benchmarking frameworks (standalone, cluster)
  - Web site for collecting metrics (Dumpstr)
- **Publicity: SOSP paper, LinkedIn talk, articles**

What's Left for RAMCloud 1.0?

- **Fault-tolerant coordinator (complete rewrite underway)**
- **LogCabin (leader election, configuration storage)**
- **A few more bits for recovery**
  - Simultaneous failures
  - Cold start
  - RPC retry
- **Table enumeration**
- **Synchronous backup writes**

1.0 is "constant 3-6 months away"

Acceptance Test for 1.0

- **Run an 80-node cluster for a few weeks:**
  - Synthetic workload, capable of detecting data corruption
  - Force servers to crash at random times
  - Multiple simultaneous failures
  - Coordinator failures
  - Complete cluster crashes
- **Survive with no loss of data**