RAMCloud: Scalable High-Performance Storage Entirely in DRAM

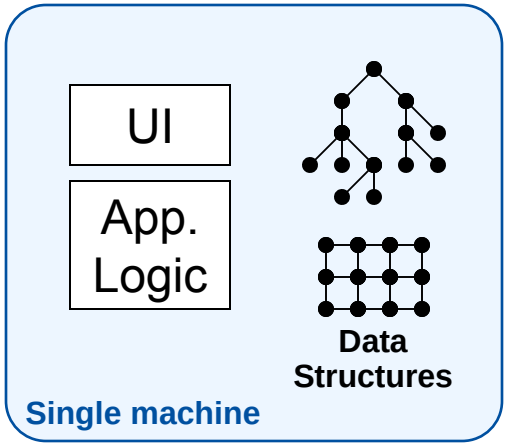
John Ousterhout, David Mazières, and Mendel Rosenblum

Stanford University

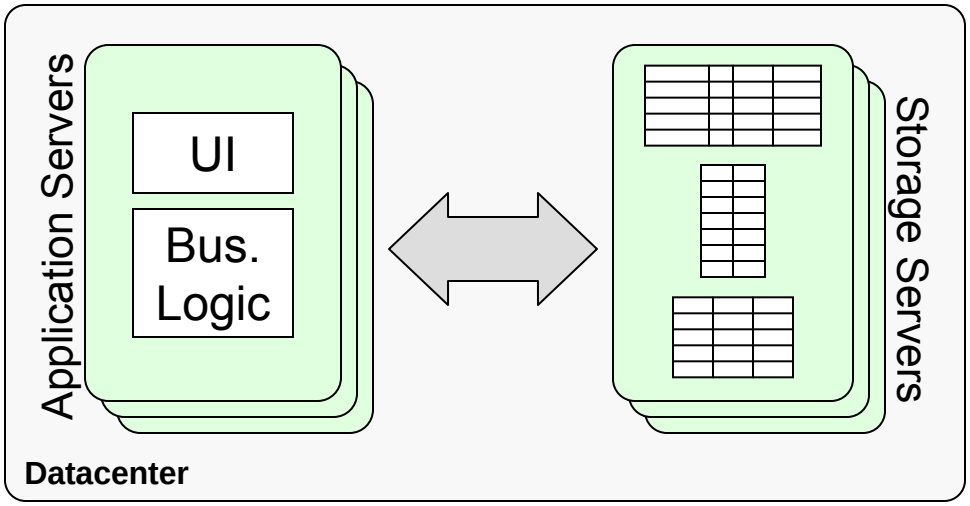
http://www.stanford.edu/~ouster/cgi-bin/papers/ramcloud.pdf

Motivation: Latency at Scale

Traditional Application



Web Application



<< 1µs latency

0.5-10ms latency

- **Large-scale apps struggle with high latency**
- **RAMCloud goal: low latency and large scale**
- **Enable a new breed of information-intensive applications**

RAMCloud Concept

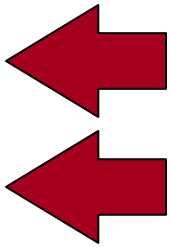
- **Lowest possible latency?**
  All data always in DRAM
- **5-10µs RPC within datacenter**
- **Scale: aggregate 10-10000 commodity servers**
- **High throughput: 1M ops/sec/server**
- **Durable and available**

	Today	5-10 years
# servers	1000	1000
GB/server	64GB	1024GB
Total capacity	64TB	1PB
Total server cost	$4M	$4M
$/GB	$60	$4

RAMClouds are practical today

Research Issues

- Achieving 5-10 µs RPC
- Durability at low latency
- Data model
- Concurrency/consistency model
- Data distribution, scaling
- Automated management
- Multi-tenancy
- Node architecture



Conclusion

- **100TB - 1PB @ 5-10µs for 1000-10000 clients**
- **For more on motivation & research issues:**
  - "The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM"
  - To appear in *Operating Systems Review*
  - http://www.stanford.edu/~ouster/cgi-bin/papers/ramcloud.pdf
  - Or, google "RAMCloud"
- **Questions/Comments?**