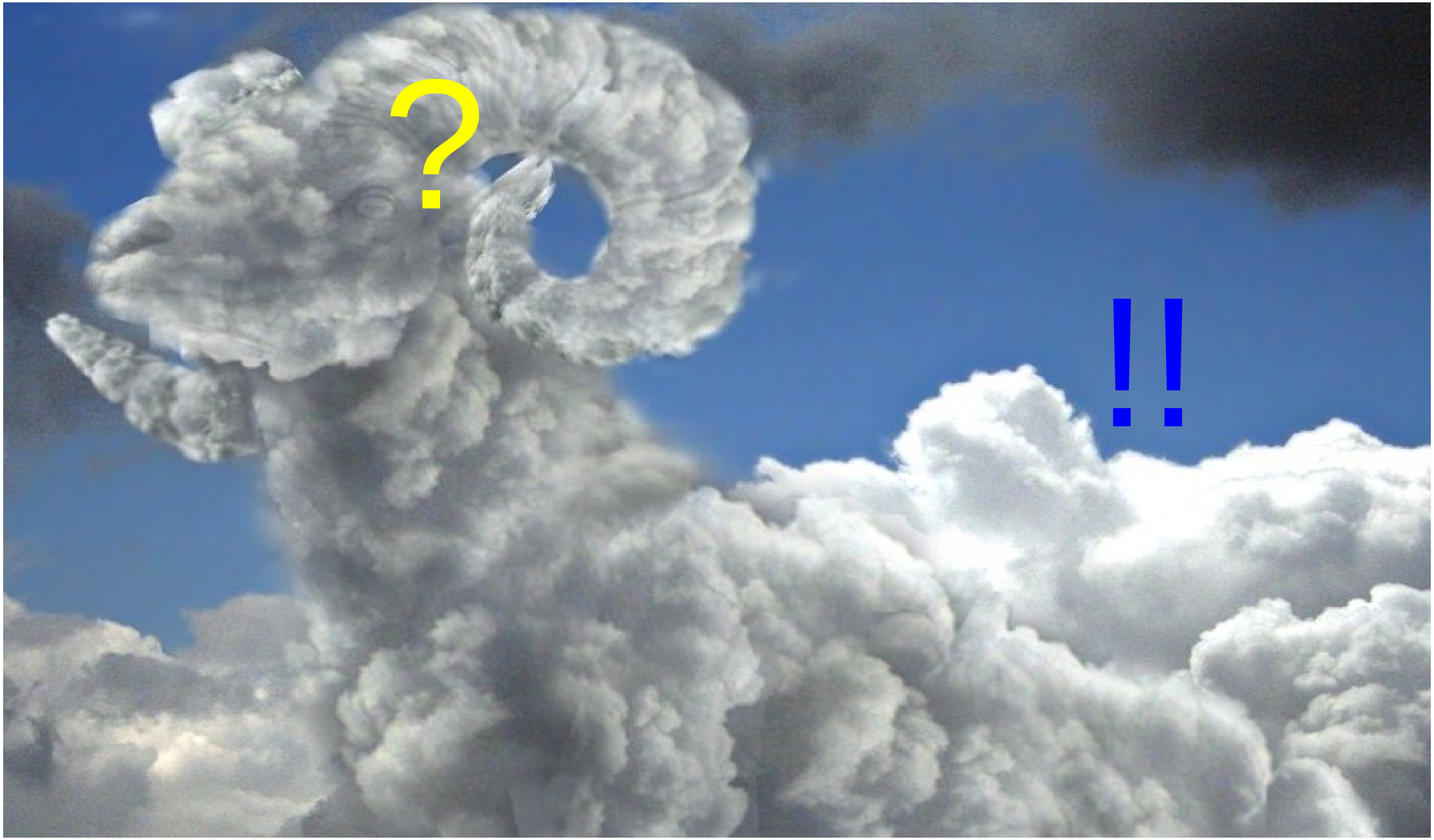
# RAMCloud

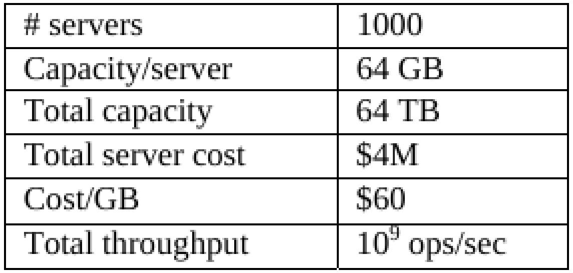


# Overview

- Datacenters split into application and storage servers
- Use RAMCloud for storage
    - All Information is kept in DRAM at all times
    - (not like memcached, data not stored on I/O device)
    - auto scaling, application sees one large storage
    - must be as durable as if stored on disk
- 100x-1000x better performance than current disk-based storage

# Configuration of a RAMCloud

- Table 1 = currently cost-effective
- With additional servers as large as 500TB possible
- Within 5-10 years depending on DRAM technology up to 1-10 PB at < 5$/GB



| | |
|---|---|
| # servers | 1000 |
| Capacity/server | 64 GB |
| Total capacity | 64 TB |
| Total server cost | $4M |
| Cost/GB | $60 |
| Total throughput | $10^9$ ops/sec |

**Table 1.** An example RAMCloud configuration using currently available commodity server technology. Total server cost is based on list prices and does not include networking infrastructure or racks.

# Motivation

- Databases do not scale well:

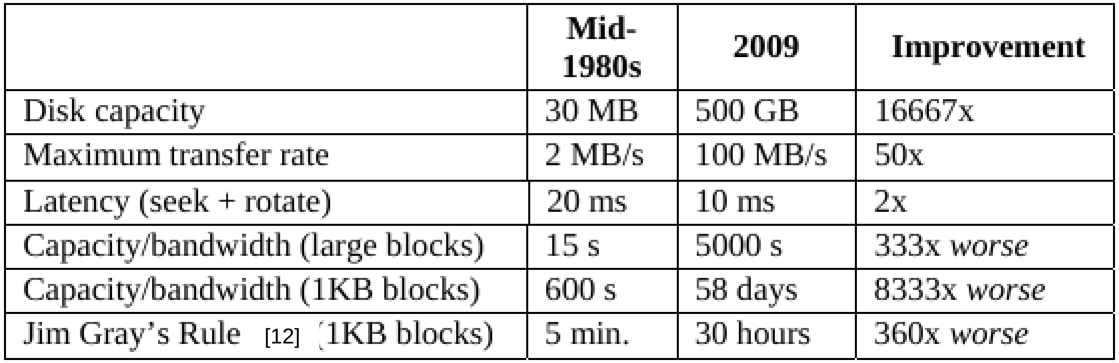
  “virtually every popular Web application […] found [RDBs] cannot meet its throughput requirements”

  require special purpose techniques

- Facebook: 4000 MySQL Servers, still do not meet throughput demand → 2000 mcached servers
- new storage systems (Bigtable, Dynamo) to address scalability issues, but only for specialized scenarios
- give up some ACID properties

# Technology Trends

- Files need to be larger today to achieve 90% maximum transfer rates



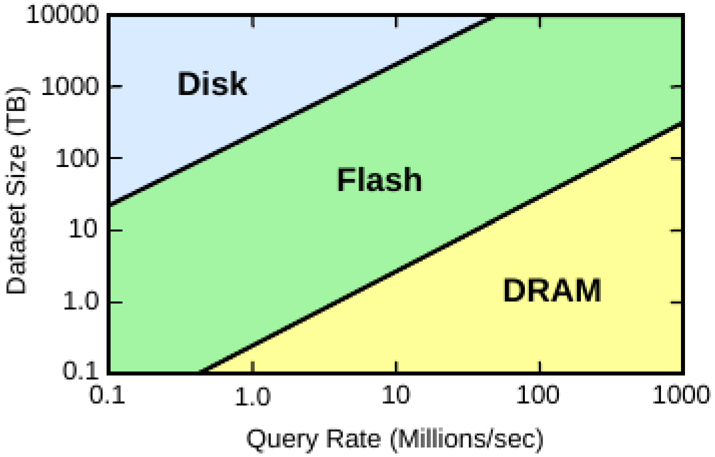
| | Mid-1980s | 2009 | Improvement |
|---|---|---|---|
| Disk capacity | 30 MB | 500 GB | 16667x |
| Maximum transfer rate | 2 MB/s | 100 MB/s | 50x |
| Latency (seek + rotate) | 20 ms | 10 ms | 2x |
| Capacity/bandwidth (large blocks) | 15 s | 5000 s | 333x *worse* |
| Capacity/bandwidth (1KB blocks) | 600 s | 58 days | 8333x *worse* |
| Jim Gray’s Rule [12] (1KB blocks) | 5 min. | 30 hours | 360x *worse* |

# Caching

- Facebook keeps 25% of data in main memory on memcached servers, 96.5% Hitrate
- Incl. database caches, 75% in memory
- RAMCloud would only need 25% more main memory
    - “RAMClouds may cost slightly more than caching systems, but they will provide guaranteed performance independent of access patterns or locality.”

# What about FlashCloud?

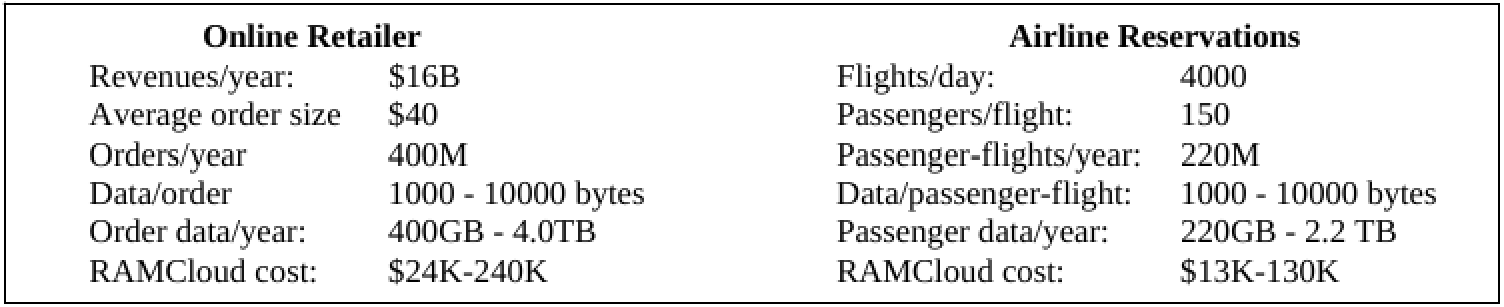
- Might be a good compromise
- but believe that DRAM-based is more attractive because of higher performance
- RAMCloud still 5x-10x better
- Phase-Change memory?
  - might still benefit from techniques developed for RAMClouds



# Applicability

- Facebook @ 260TB (upper limit for RAMCloud)
- DRAM prices today ~=~ disk prices 10 years ago

  → any data that could be stored cost-effectively on disk then can be stored cost-effectively in RAM today
- RAMClouds not good for Images / Video / Audio but mainly for data



| Online Retailer | | Airline Reservations | |
|---|---|---|---|
| Revenues/year: | $16B | Flights/day: | 4000 |
| Average order size | $40 | Passengers/flight: | 150 |
| Orders/year | 400M | Passenger-flights/year: | 220M |
| Data/order | 1000 - 10000 bytes | Data/passenger-flight: | 1000 - 10000 bytes |
| Order data/year: | 400GB - 4.0TB | Passenger data/year: | 220GB - 2.2 TB |
| RAMCloud cost: | $24K-240K | RAMCloud cost: | $13K-130K |

# Research Issues
# - Low latency RPC -

- Ethernet typical: 0.3-0.5ms RTT
- think it is possible to reduce to 5 [illegible]
  - reduce latency in switches (al[illegible]tter with 10GE)
  - reduce software overhead [illegible]
    - → no GP-OS, dedicat[illegible] of network on one core
  - modify TCP proto[illegible]e other reliable UDP based protocol
    - → retran[illegible]neouts too high in TCP, degrade latency
    - → lit[illegible]age in flow-oriented nature of TCP
    - [illegible] protocol can use and otimized ack scheme

Sudden Attack Of Virtualization (SAOV)

# Research Issues
# - Durability and Availability -

- RAMCloud should be at least as good as today's disk-based systems

→ at m

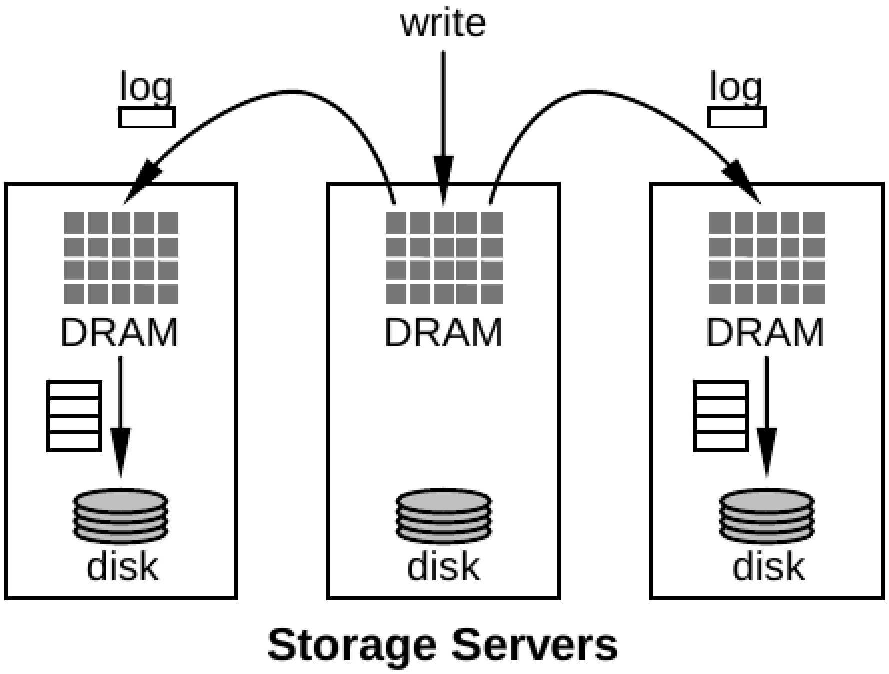
server

(

ng)

→ also

power?)



Buffered Logging

# Research Issues
# - Data model -

- Low latency RPC
  - Ethernet typical: 0.3-0.5ms RTT
  - think it is possible to reduce to 5-10us
    - reduce latency in switches (already better with 10GE)
    - reduce software overhead
      - → no GP-OS, dedicated polling of network on one core
    - modify TCP protocol or use other reliable UDP based protocol
      - → retransmisson timeouts too high in TCP, degrade latency
      - → little advantage in flow-oriented nature of TCP
      - → custom protocol can use and otimized ack scheme

# Research Issues
# - Distribution and Scaling -

- Should scale transparently, software should not be aware of the distributed nature of the storage
- Issue: where to place data?
- No replication needed for performance reasons (b/c low latency / high bandwidth)
- Should enable data migration with applications running

# Research Issues
# - Concurrency, consistency -

- How to handle interactions between simultaneously served requests?
  - ACID scales poorly, many web applications do not need ACID and don't wish to pay for it
  - RAMClouds extremely low latency may enable higher level of consistency than other systems of comparable scale
    - Reason: ACID is only expensive if there are many transactions competing → low latency = less aborts!
- Strong consistency still expensive if replication over data centers needed!

# Research Issues
# - Others -

- Multi-tenancy
  - system must house applications of varying sizes
  - must scale on short notice
  - access control / security mechanisms needed
  - performance isolation?
- Server – client functionality distribution
  - client side library
    - may hide object model
  - migrate functionality (code) to storage servers? security?
- Self - Management

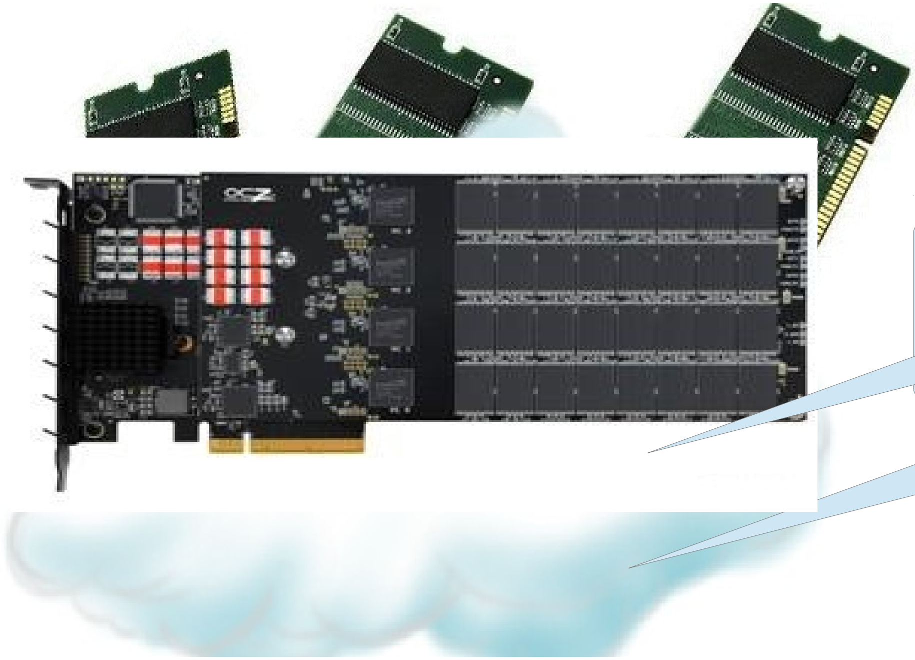
# Disadvantages

- High cost per bit
- High energy usage per bit
- Floor space

  → not effective for large amounts of data
- more efficient at cost/operation and energy/op

  → efficient for high throughput applications
- high latency for cross-DC replication → no gain for writes, still efficient for reads

# Discussion points



- 3TB
- 18K Euro
- 410K IOPS

- [illegible] Euro
  1000k IOPS

# Discussion points

- If there is a need, why are there no PCI-e RAMDrives used?
- If we don't have durability (security for crashes) do we need it?