Characterizing the Performance and Energy Efficiency of In-memory Storage Systems

Y. Taleb, S. Ibrahim, G. Antoniu, T. Cortes*
Kerdata, Inria, France
* Barcelona Supercomputing Center, Universitat Politècnica de Catalunya, Spain

25/06/2017

It’s time for low latency!

amazon

1s slowdown in page loading =
$1.6 billion loss in sales each year

Google

250ms slowdown =
~3 billion less searches every year

https://www.fastcompany.com/1825005/how-one-second-could-cost-amazon-16-billion-sales

Storage shifting to main memory

Web applications

Caching

BigData processing

In-memory computations

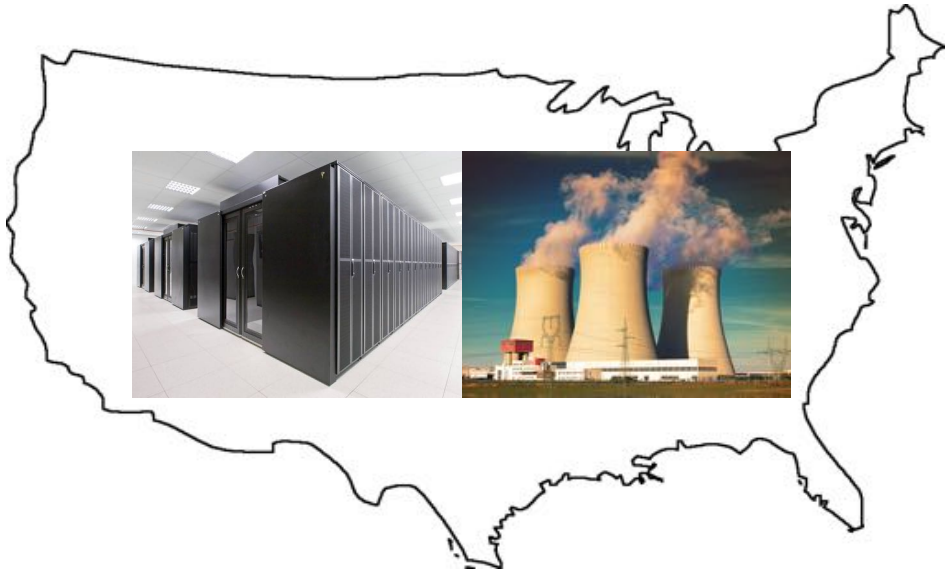
Flink

APACHE Spark™

ALLUXIO

Energy concerns

According to a study, by 2020 datacenters will consume the equivalent of 50 power plants, in the US <u>only</u>



25% to 40% of server's **total energy consumption** due to **DRAM**

https://www.nrdc.org/resources/americas-data-centers-consuming-and-wasting-growing-amounts-energy

Aniruddha N. et al. 2010. Rethinking DRAM design and organization for energy-constrained multi-cores. In *Proceedings of the 37th annual international symposium on Computer architecture* (ISCA '10). ACM, New York, NY, USA, 175-186.

Outline

- Context
- **Characterizing the performance and energy efficiency of the RAMCloud storage system**
  - The RAMCloud storage system
  - Methodology
  - Results
- Conclusion

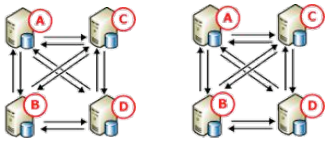
Characterizing Performance and Energy Efficiency of in-memory storage

How to identify sources of performance/energy inefficiency?

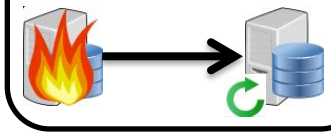
Study each feature of a system separately

Scalability

Replication



Fault-tolerance



. . .

We need a representative system …

Characterizing Performance and Energy Efficiency of the RAMCloud storage system

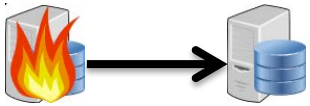
The RAMCloud storage system

Low Latency

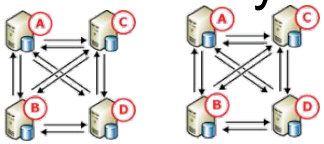
Scalability

Strong Consistency

Availability



Durability



. . .

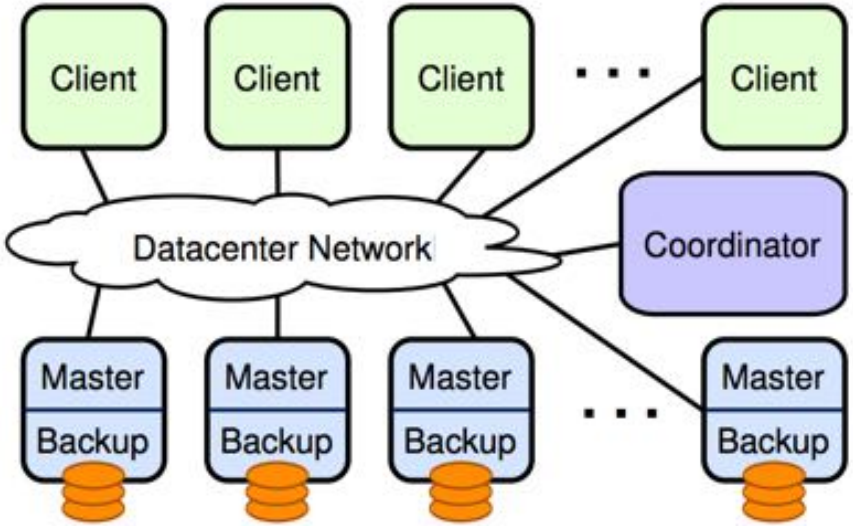
The RAMCloud storage system

General purpose in-memory key-value store, developed at Stanford

Keeps all data in DRAM
→disk for replication only

Relies on high performance networks (e.g. Infiniband)

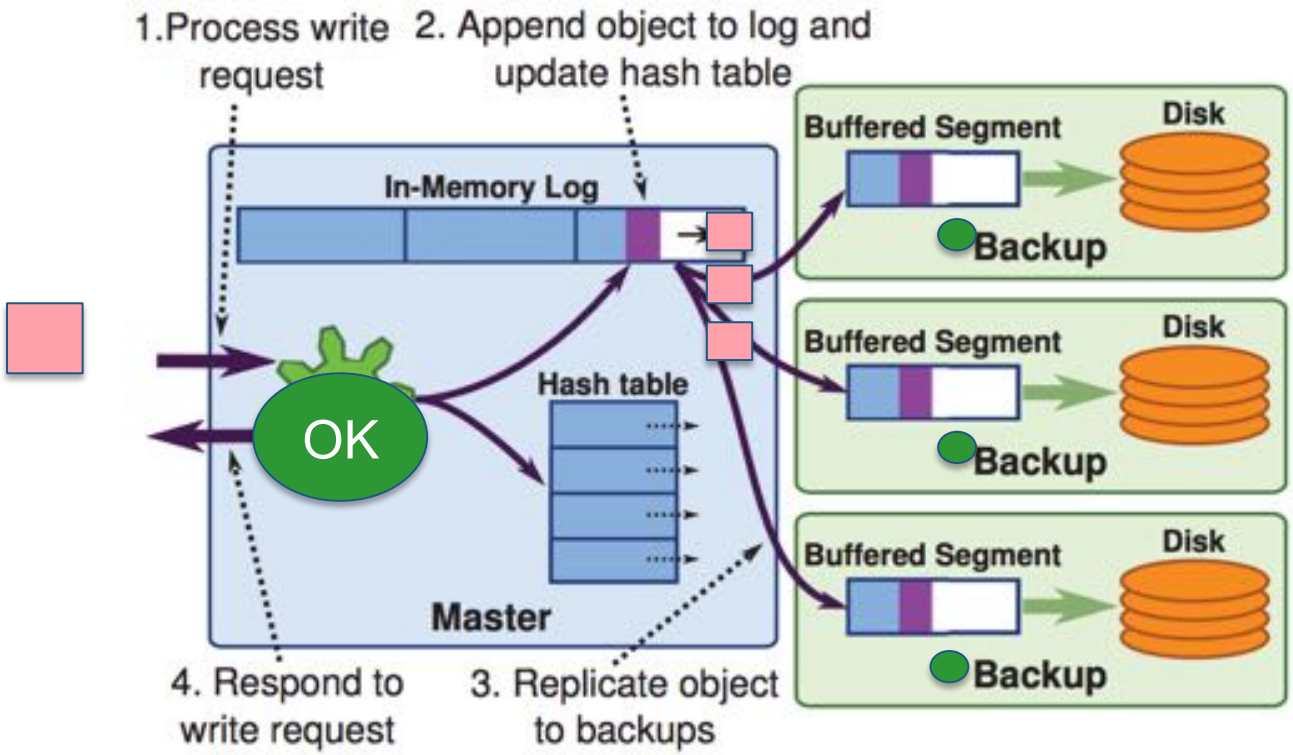
Log-structured memory and disk → high memory efficiency



Credit: D. Ongaro et al

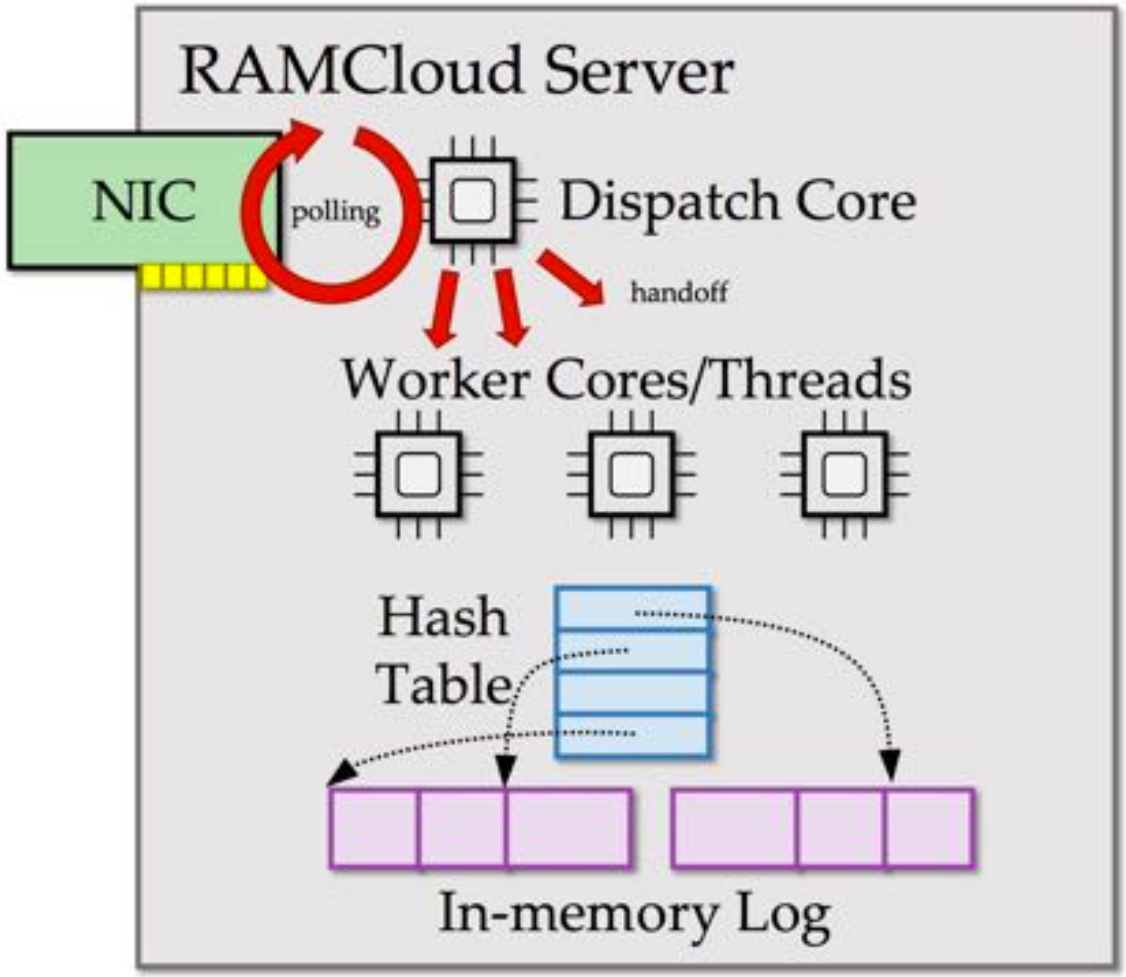
The RAMCloud storage system

Primary-backup replication
-> guarantees linearizability



Credit: D. Ongaro et al

RAMCloud dispatch



How do we proceed?

Industry standard: *Yahoo*! Cloud Serving Benchmark

Deployed RAMCloud on GRID'5000 experimental testbed
Infiniband, PDUs (power measurement), etc.

What are we studying?

- Energy efficiency at peak performance
- Replication overhead
- Crash recovery
- Read/Write workloads
- Client's location/network impact

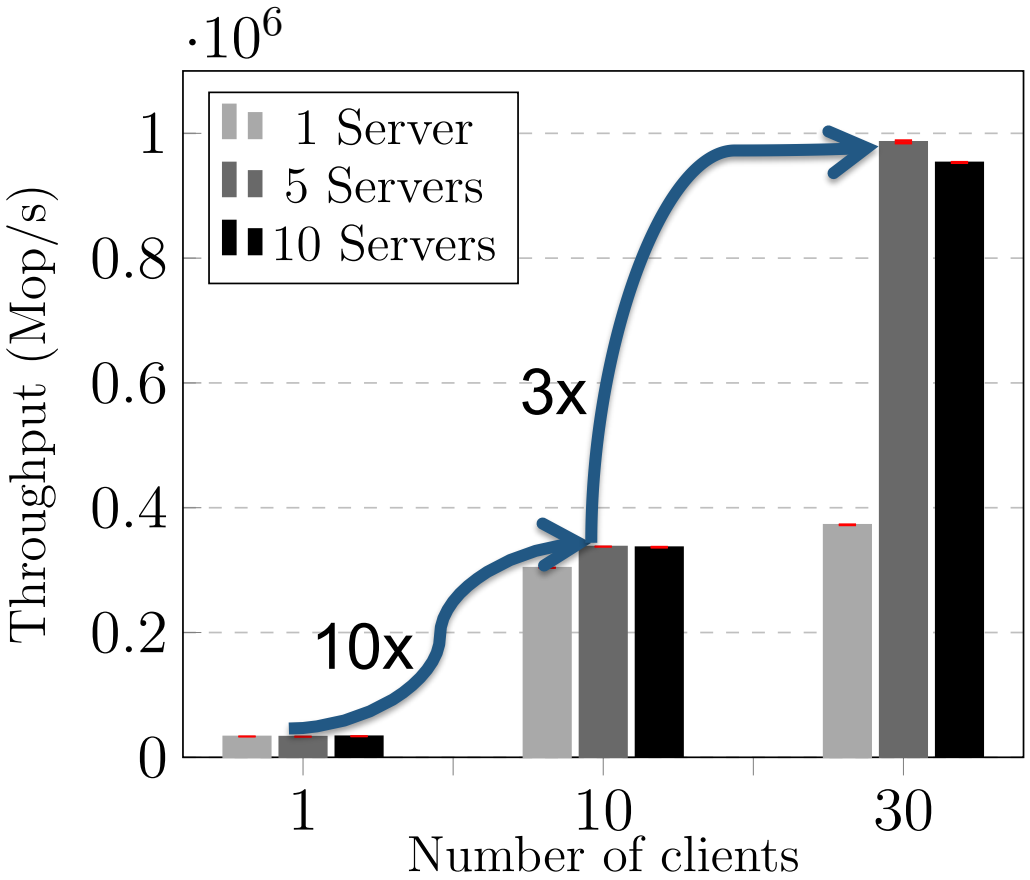
The energy efficiency of peak performance

Config

Read-only
No replication
Dataset 5M 1KB objects
10M req/Client

Findings

Scalable



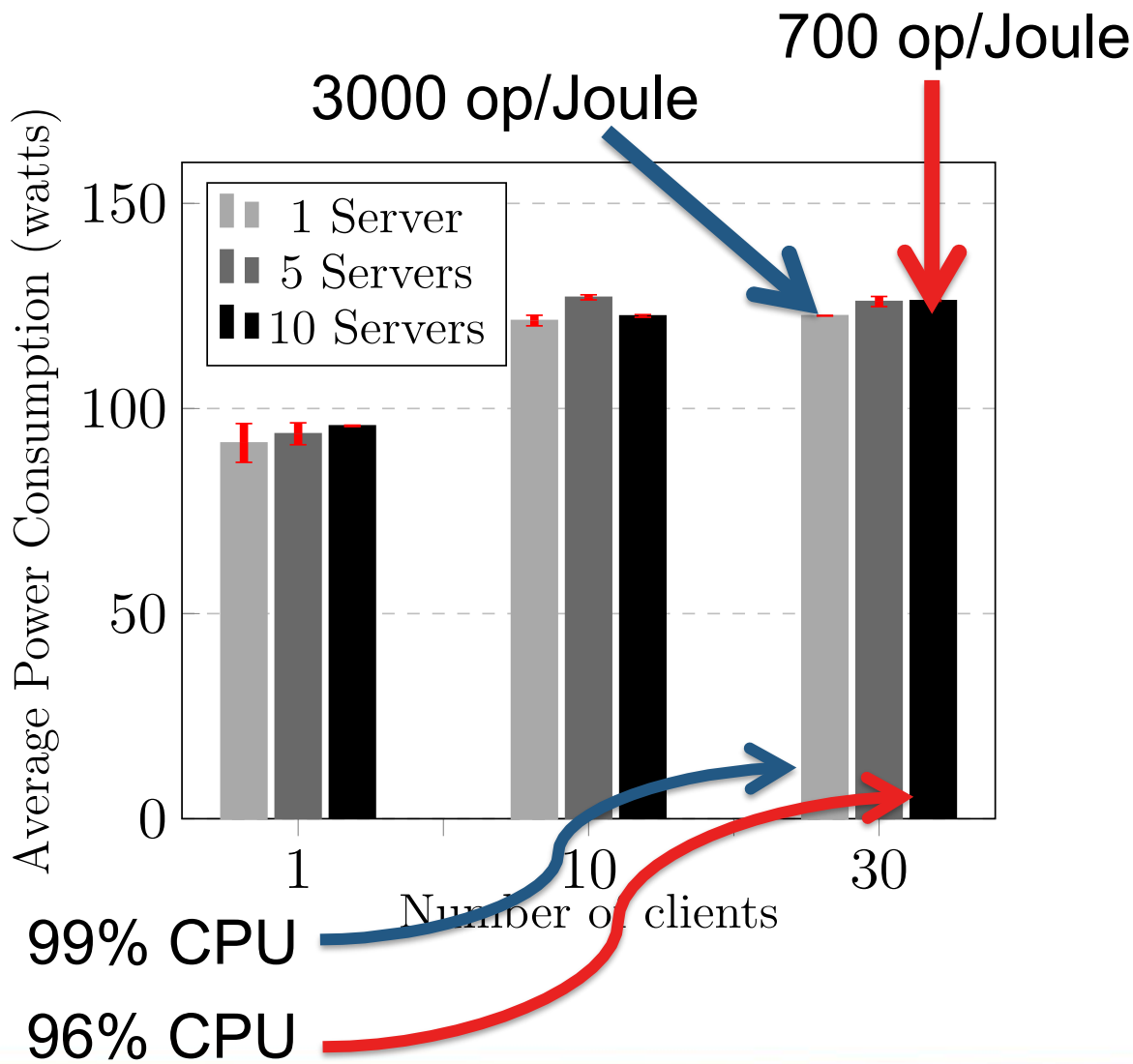
The energy efficiency of peak performance

Config

Read-only
No replication
Dataset 5M 1KB objects
10M req/Client

Findings

Scalable
Non-proportional power (& energy)
Reaches max CPU before peak performance



What are we studying?

- ✓ Energy efficiency at peak performance
- Replication overhead
- Crash recovery
- Read/Write workloads
- Client's location/network impact

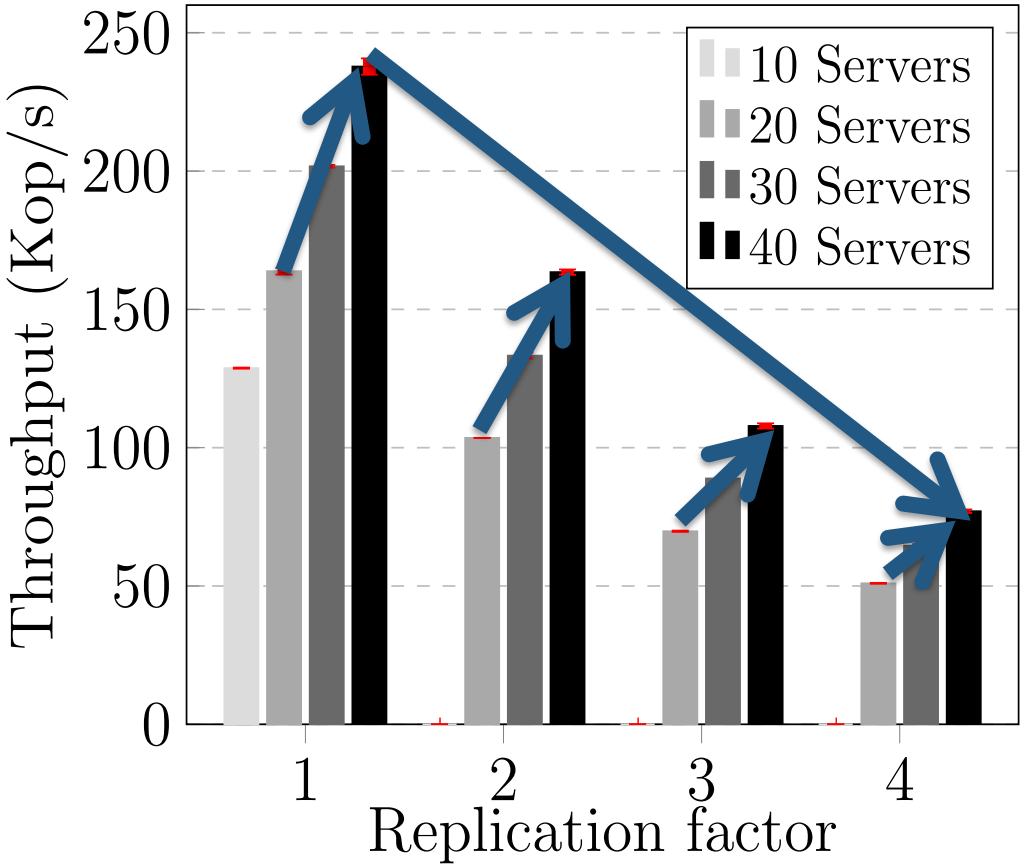
Replication's Impact - Throughput

Config

Read/write (50/50)
Replication (1 to 4)
Dataset 100K obj
100K req/Client
60 Clients

Findings

2/3 less throughput from 1 to 4 replicas



Most of the time servers are waiting for ACKs from backups

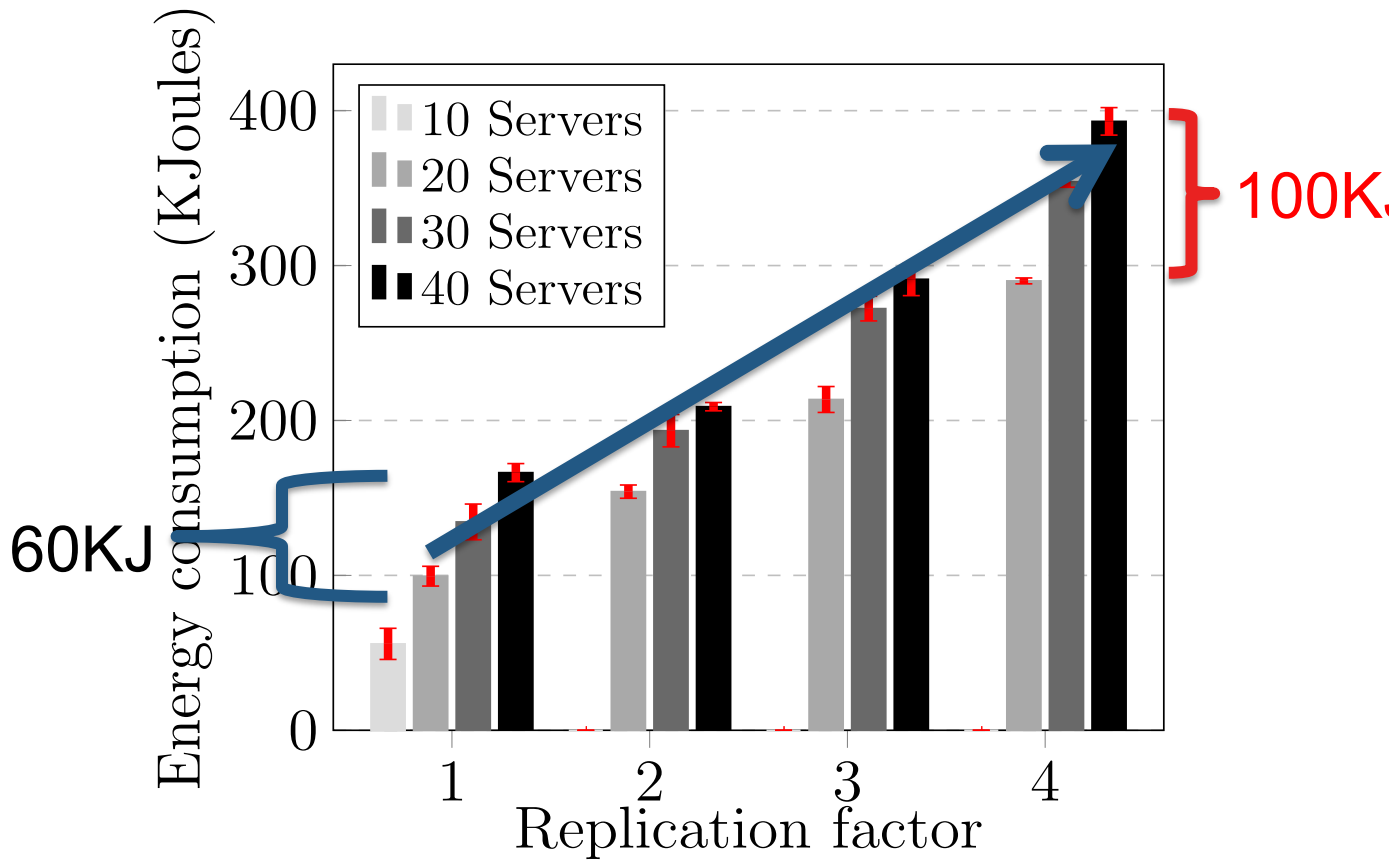
Replication's Impact - Energy

Config

Read/write (50/50)
Replication (1 to 4)
Dataset 100K obj
100K req/Client
60 Clients

Findings

2/3 less throughput from 1 to 4 replicas and 3.5x more energy consumed ->waiting for ACKs from backups



What are we studying?

- ✓ Energy efficiency at peak performance
- ✓ Replication overhead
- Crash recovery
- Read/Write workloads
- Client's location/network impact

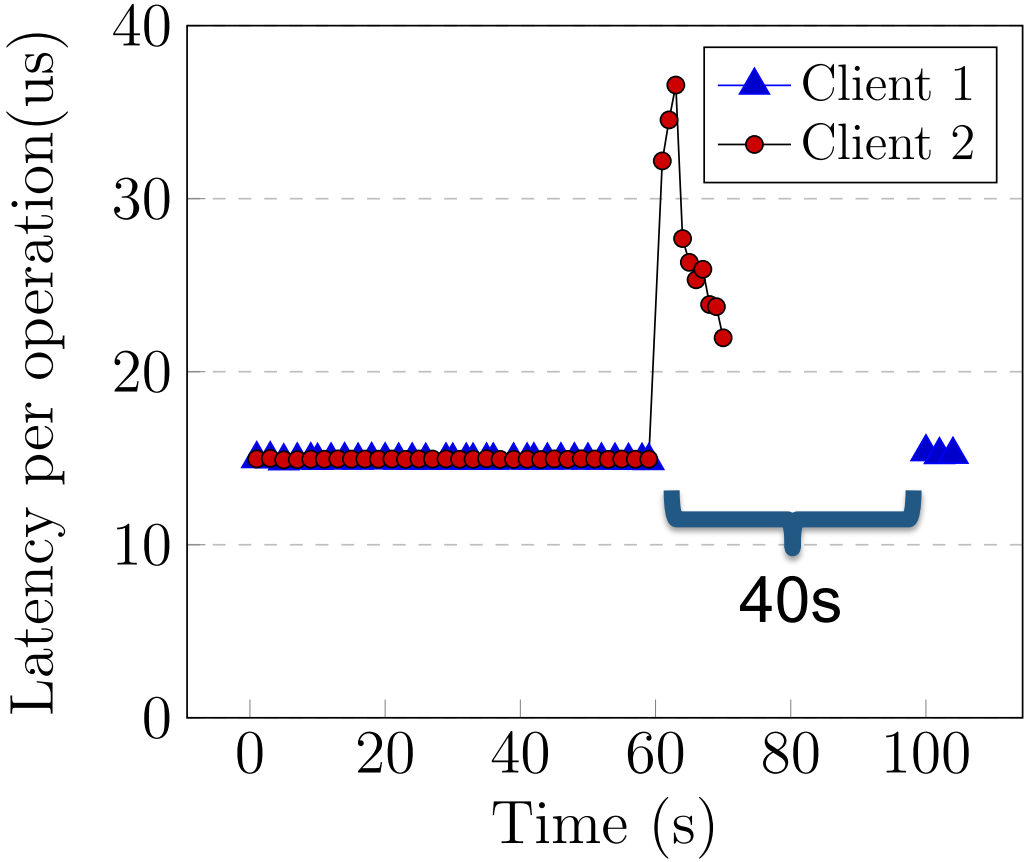
Crash Recovery

Config

10 Servers
1GB/server
After 60s idle, kill a server
2 Clients in parallel

Findings

Between 1.4X to 2.4X increase in latency
Unavailable data during recovery



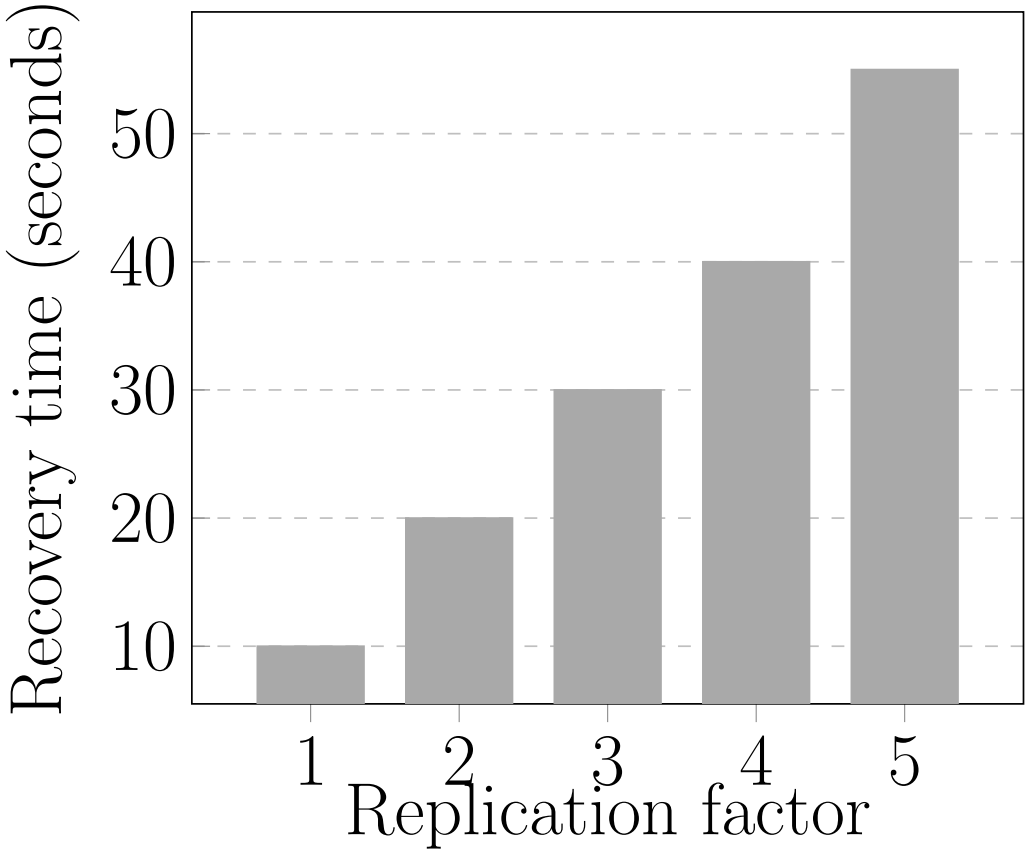
Crash Recovery

Config

10 Servers
1GB/server
After 60s idle, kill a server
2 Clients in parallel

Findings

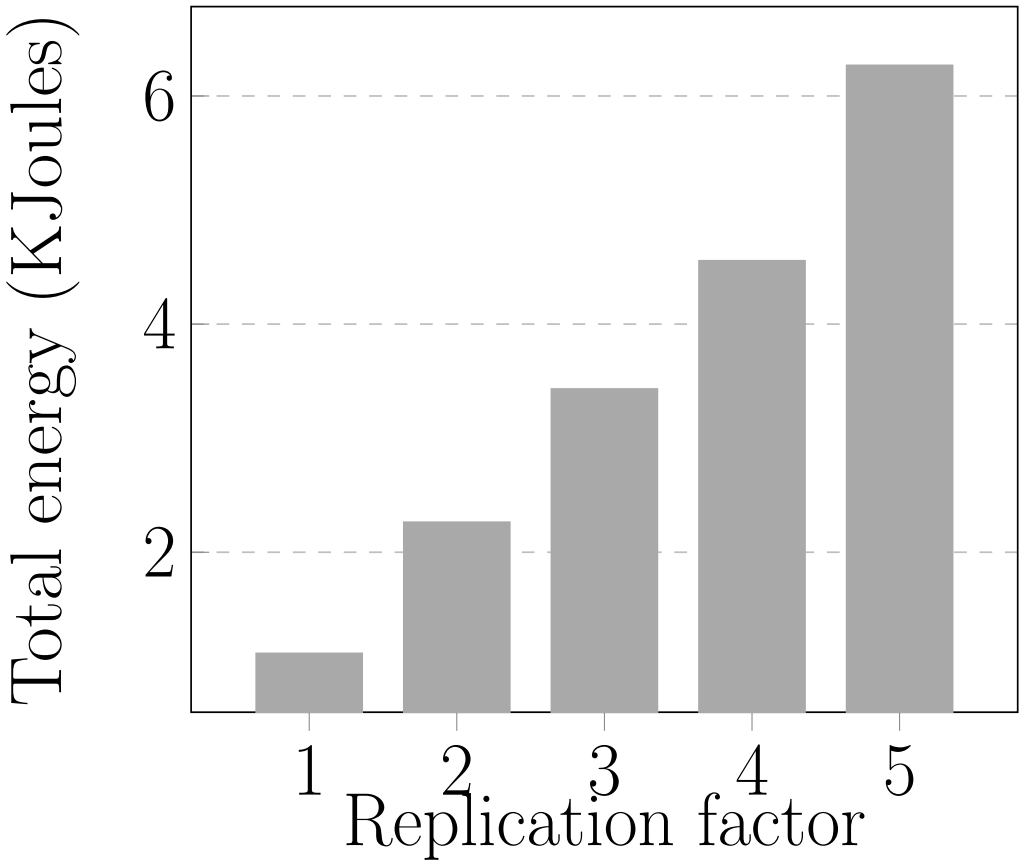
Between 1.4X to 2.4X increase in latency
Unavailable data during recovery
Linear increase in recovery time



Crash Recovery

Config
10 Servers
1GB/server
After 60s idle, kill a server
2 Clients in parallel

Findings
Between 1.4X to 2.4X increase in latency
Linear increase in recovery time and energy consumption



Replication increases crash recovery duration and energy consumption

What are we studying?

- ✓ Energy efficiency at peak performance
- ✓ Replication overhead
- ✓ Crash recovery
- Read/Write workloads
- Client's location/network impact

} More details on the paper

Outline

- Context
- **Characterizing the performance and energy efficiency of the RAMCloud storage system**
  - The RAMCloud storage system
  - Methodology
  - Results
- Conclusion

Conclusion

In-memory storage Challenges!

It's time for low latency!

amazon

1s slowdown in page loading = $1.6 billion loss in sales each year

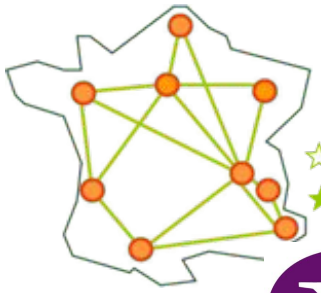
Google

250ms slowdown = ~3 billion less searches every year

Energy concerns

According to a study, by 2020 datacenters will consume the equivalent of 50 power plants, in the US only

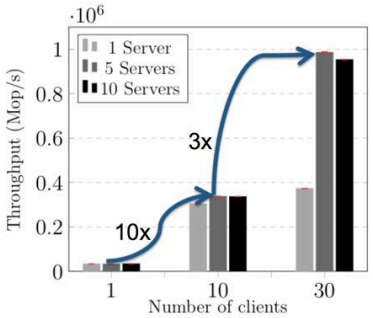
25% to 40% of server's **total energy consumption** due to **DRAM**



RAMCloud

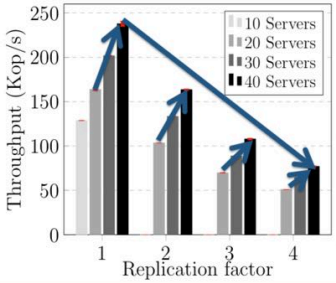
Grid'5000

Y!



Scalable Read-only

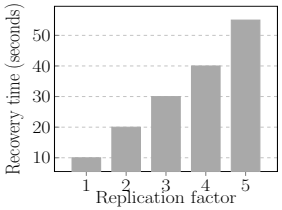
None energy-proportional



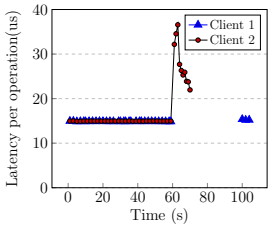
Most of the time servers are waiting for ACKs from backups

Replication overhead

Throughput and Energy!



Recovery time & energy

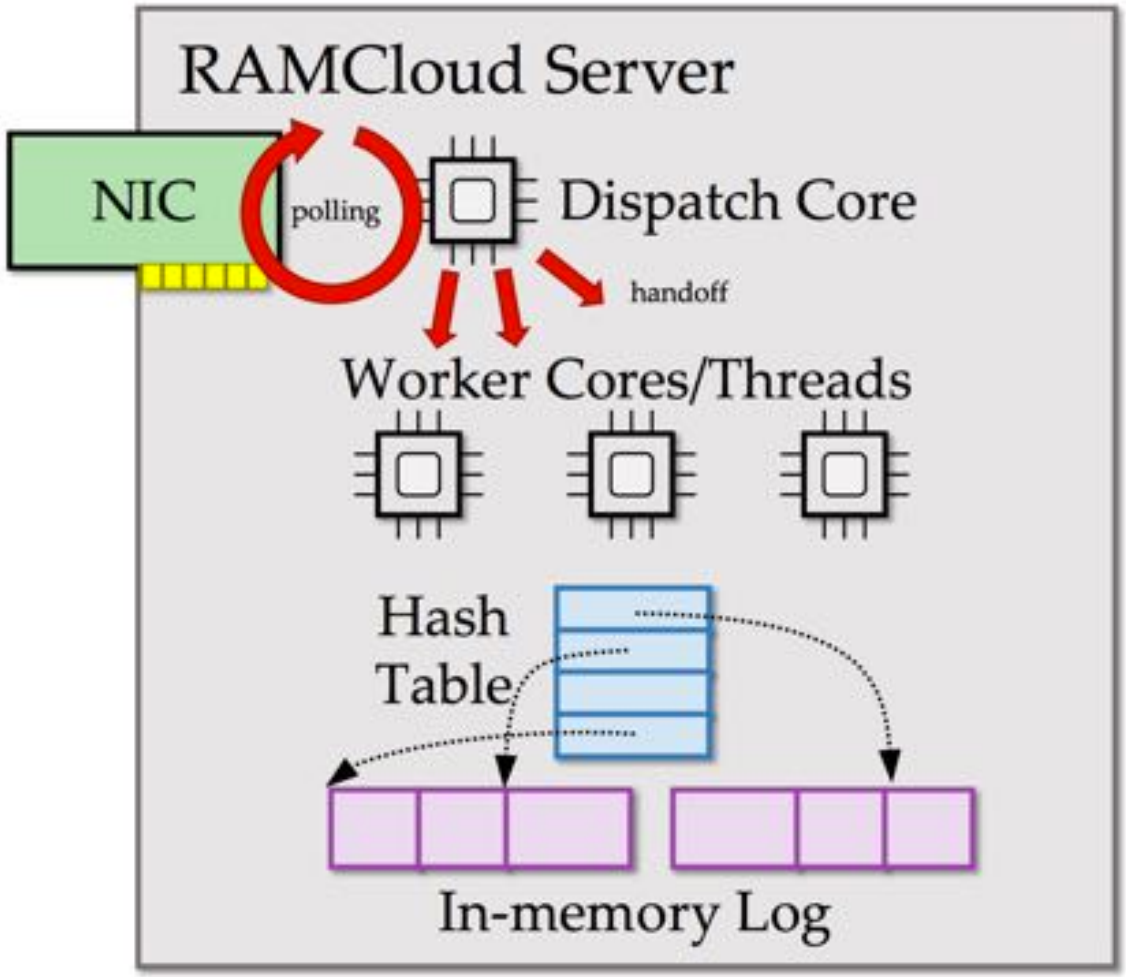


Data availability

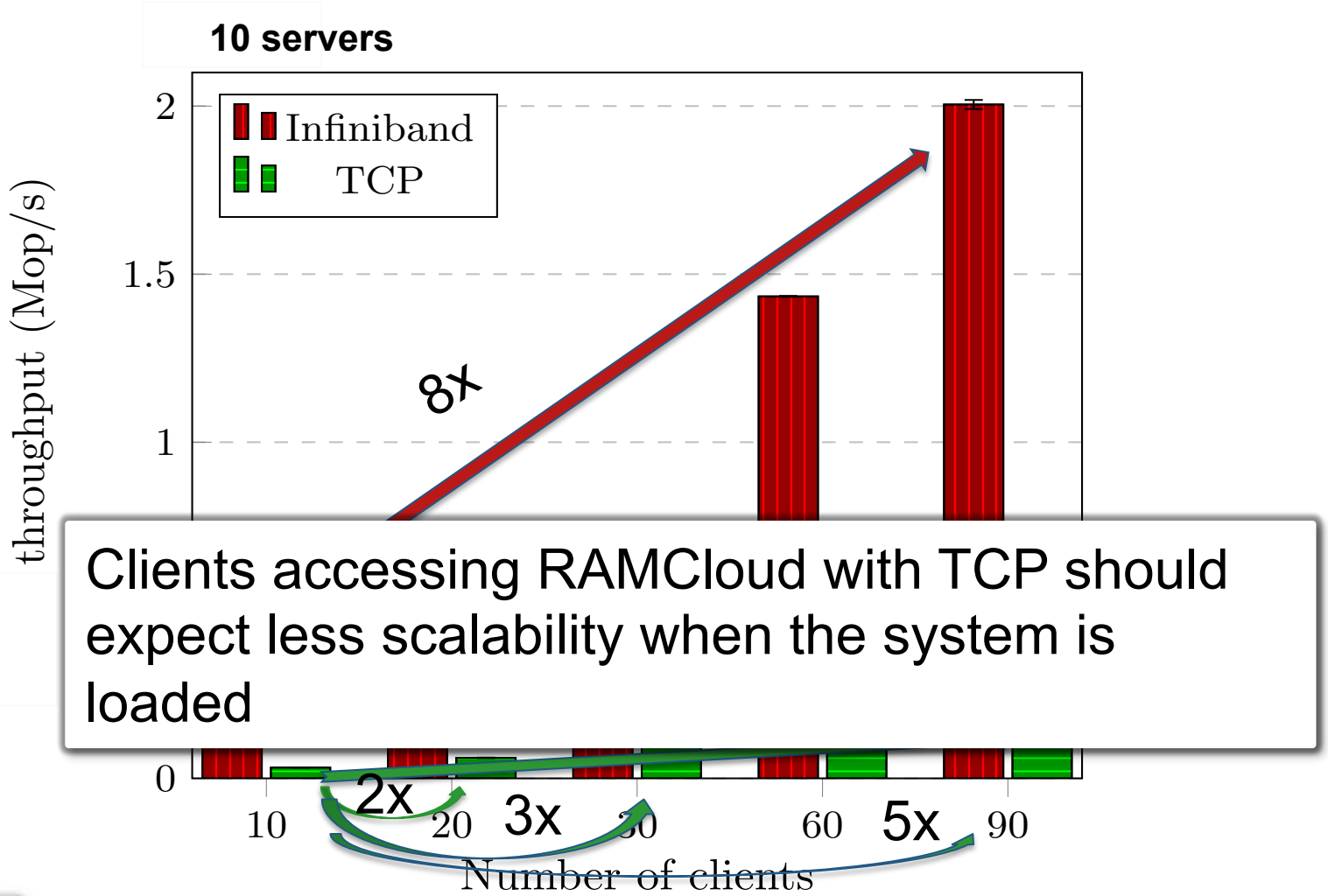
Thank you !

Backup

RAMCloud dispatch



Scalability

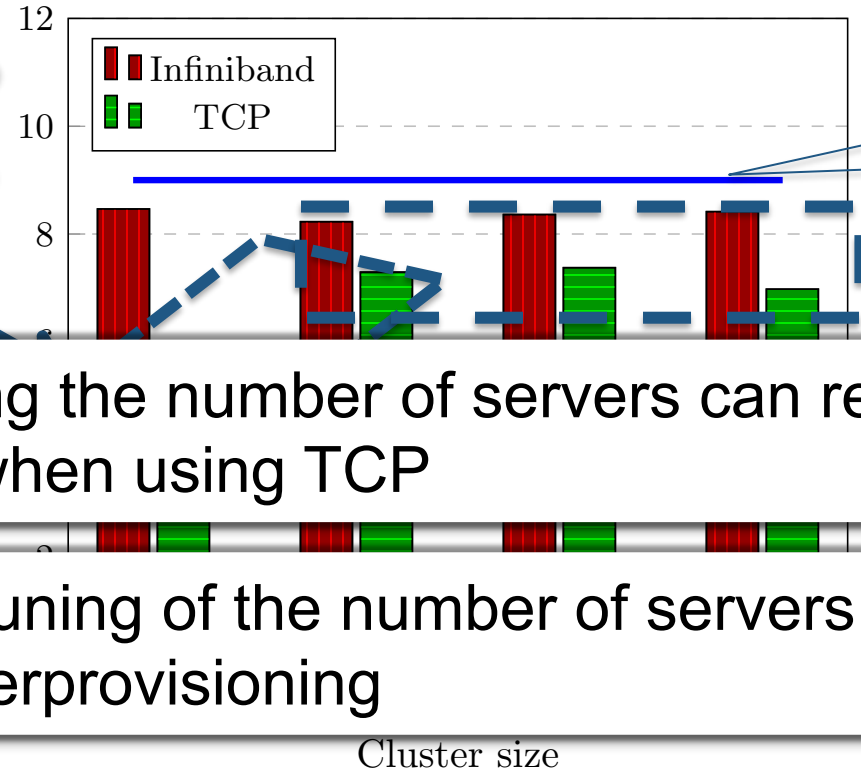


Clients accessing RAMCloud with TCP should expect less scalability when the system is loaded

Is it useful to add servers?

Scalability → Ratio of throughput from 10 to 90 clients

Better scaling, i.e., the system is more sensitive to concurrency when accessed from TCP

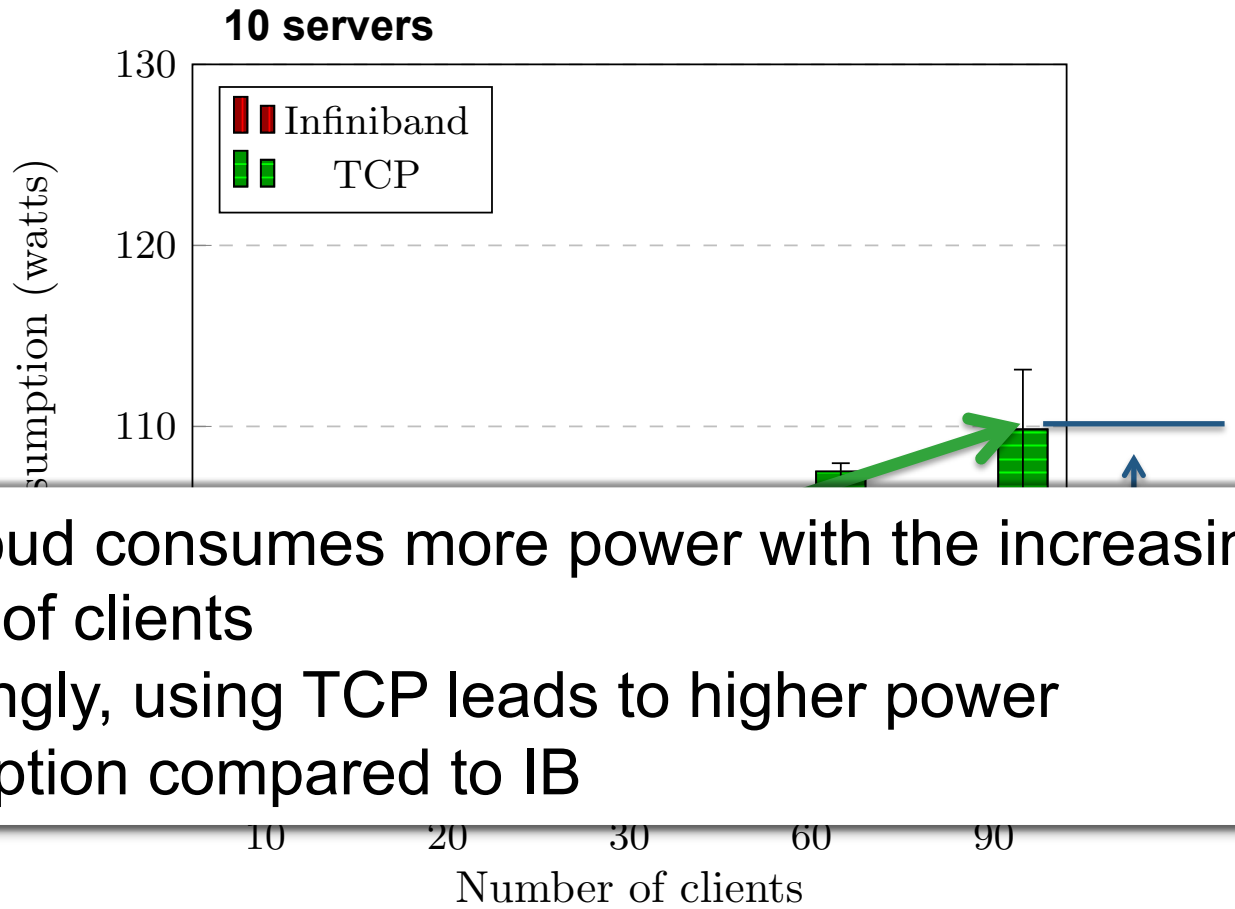


Perfect scalablility

Increasing the number of servers can result in better scaling when using TCP

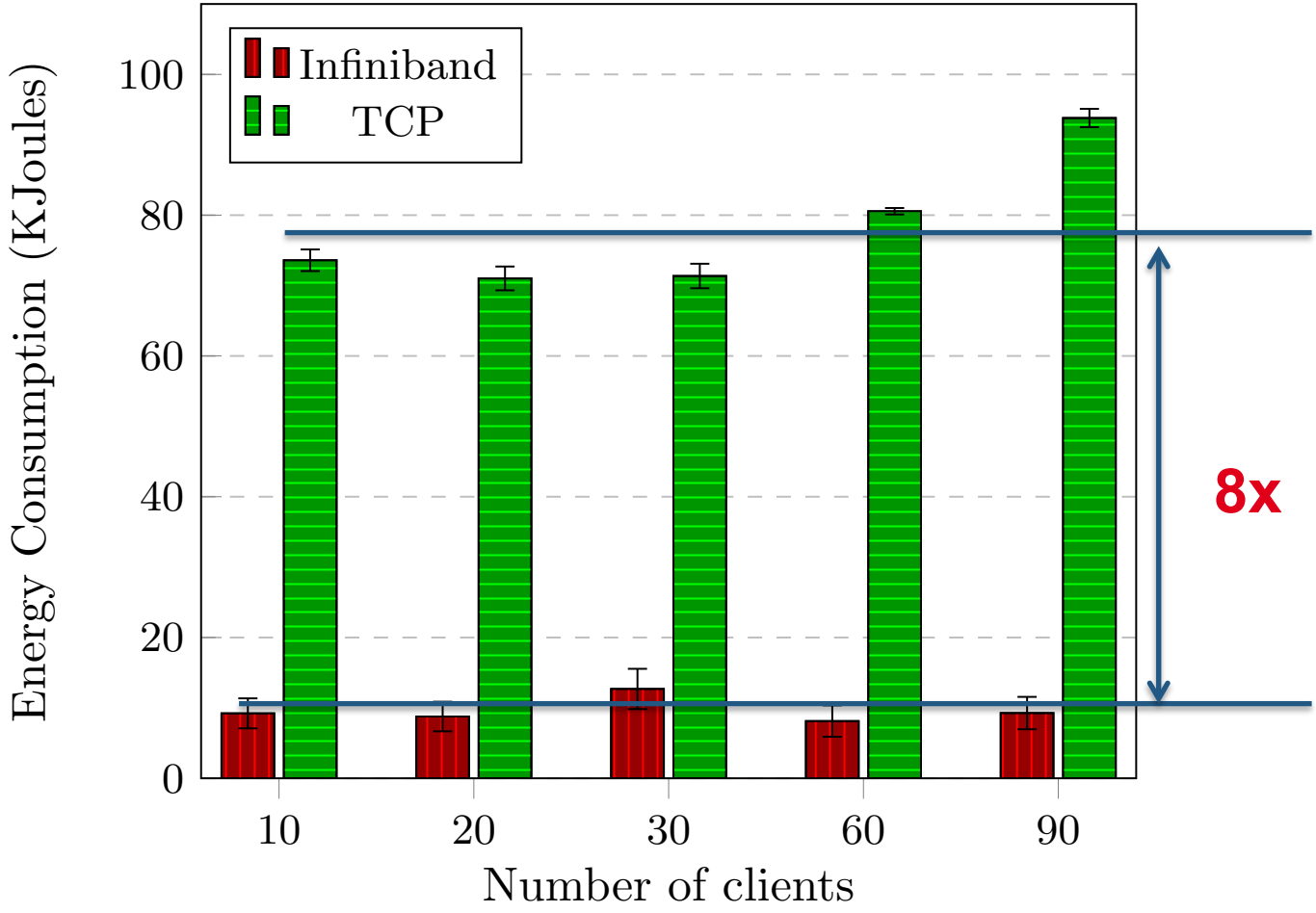
Careful tuning of the number of servers is needed to avoid overprovisioning

How about power consumption?

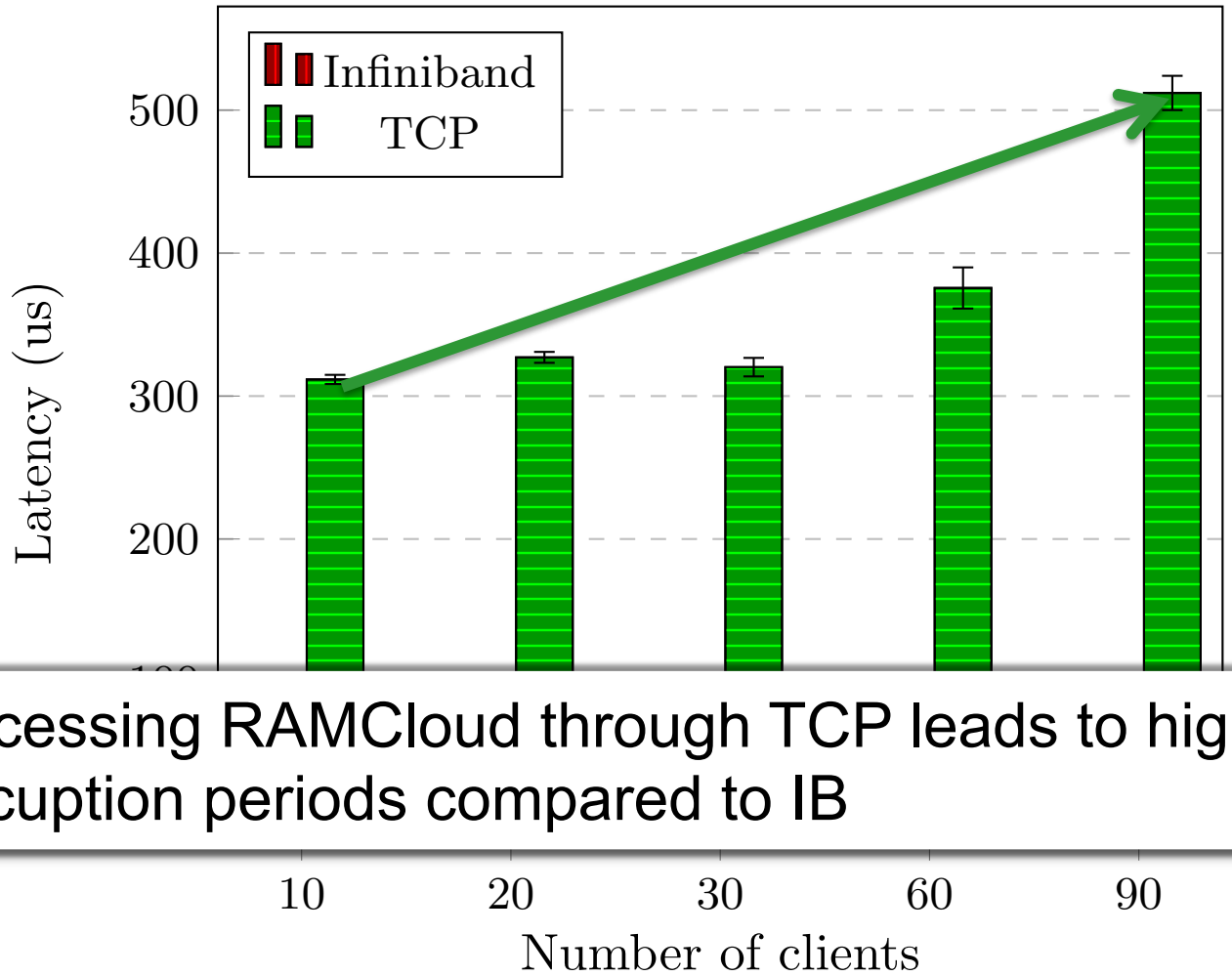


RAMCloud consumes more power with the increasing number of clients
Surprisingly, using TCP leads to higher power consumption compared to IB

The impact on energy consumption



Why RAMCloud consumes more power and energy with TCP?



Accessing RAMCloud through TCP leads to higher CPU occupation periods compared to IB

Backup RDMA

RDMAs: a magic solution to everything

- Replication consumes CPU cycles on the backups
- Backups and storage servers are colocated, a lot of ‘wasted’ resources

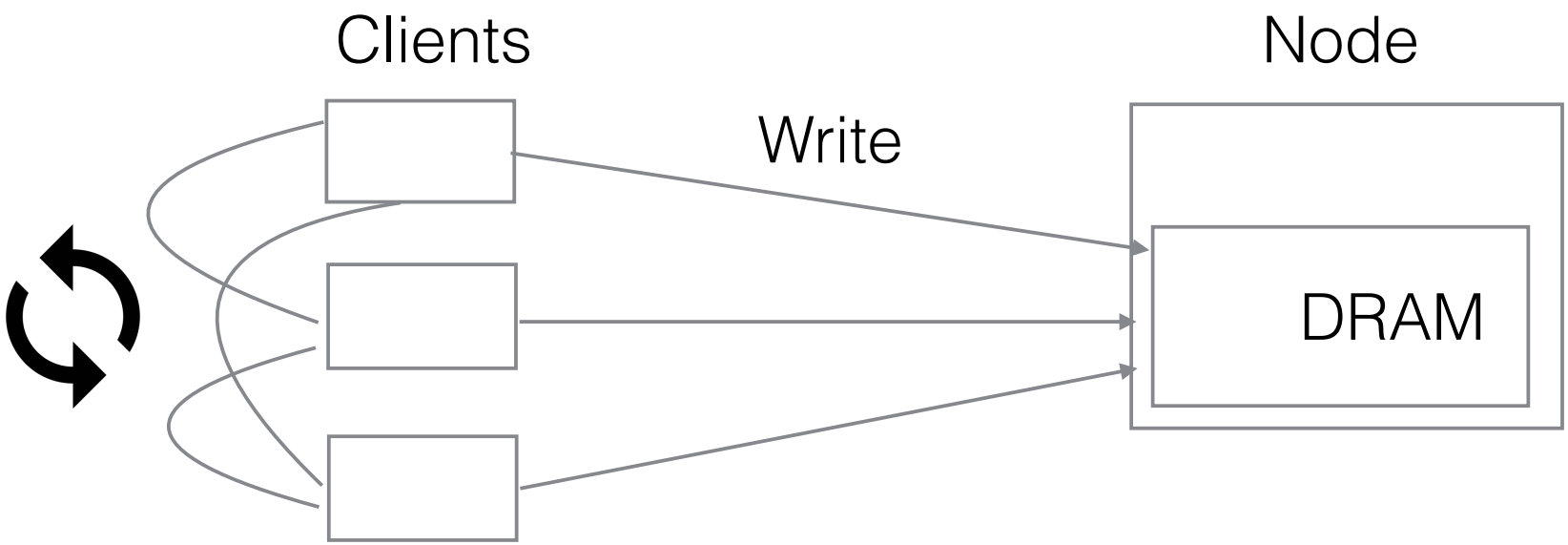
Remote-Direct Memory Access (RDMA): Network primitive to send data to a <u>contiguous</u> remote memory location

RDMAs: a magic solution to everything

Two-sided RDMAs: Kernel bypass, but still involves remote CPU during communication

One-sided RDMAs: Write/Read directly data to/from a remote memory location

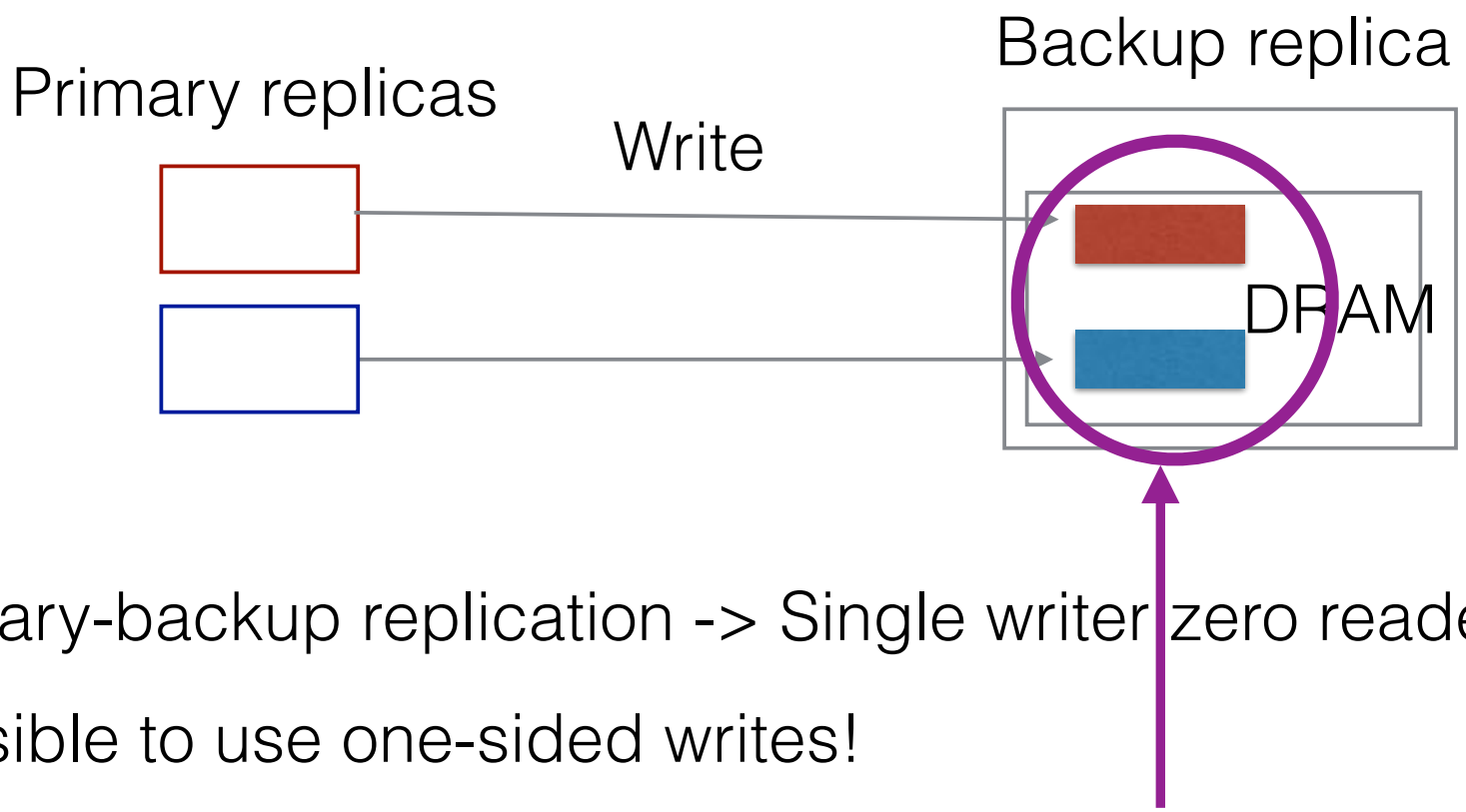
RDMAs: a magic solution to everything, **or not!**



Pilaf ATC'13

"We quickly discovered that using RDMA for all operations leads to complex and fragile designs"

Intuition



Primary-backup replication -> Single writer zero reader

Possible to use one-sided writes!

But wait… how does the receiver know about data?

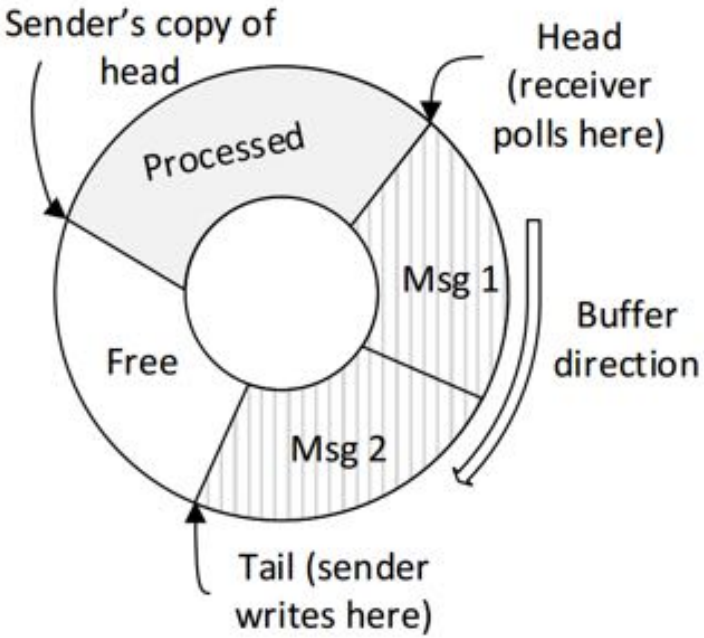
How to guarantee atomicity?

FaRM [SOSP'15,NSDI'14], HERD [SIGCOMM'14], HydraDB [SC'15]

one-sided writes to send "messages'

Receiver keeps polling memory for new messages

Extensive use of CPU



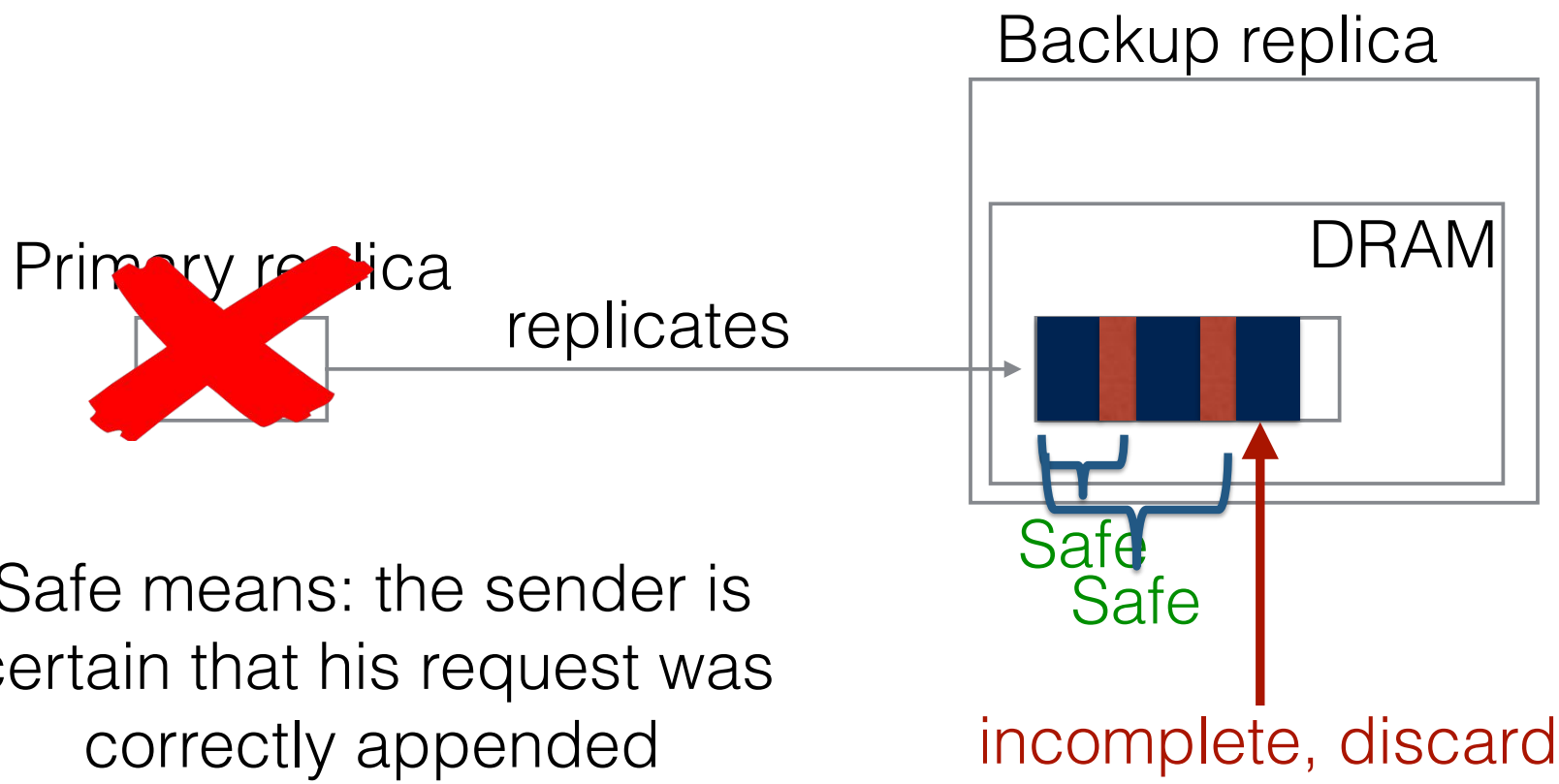
X

An RDMA-based atomic replication protocol

-When do we actually need to know about data?

Only when we need backup data, i.e. when a crash/fault happens

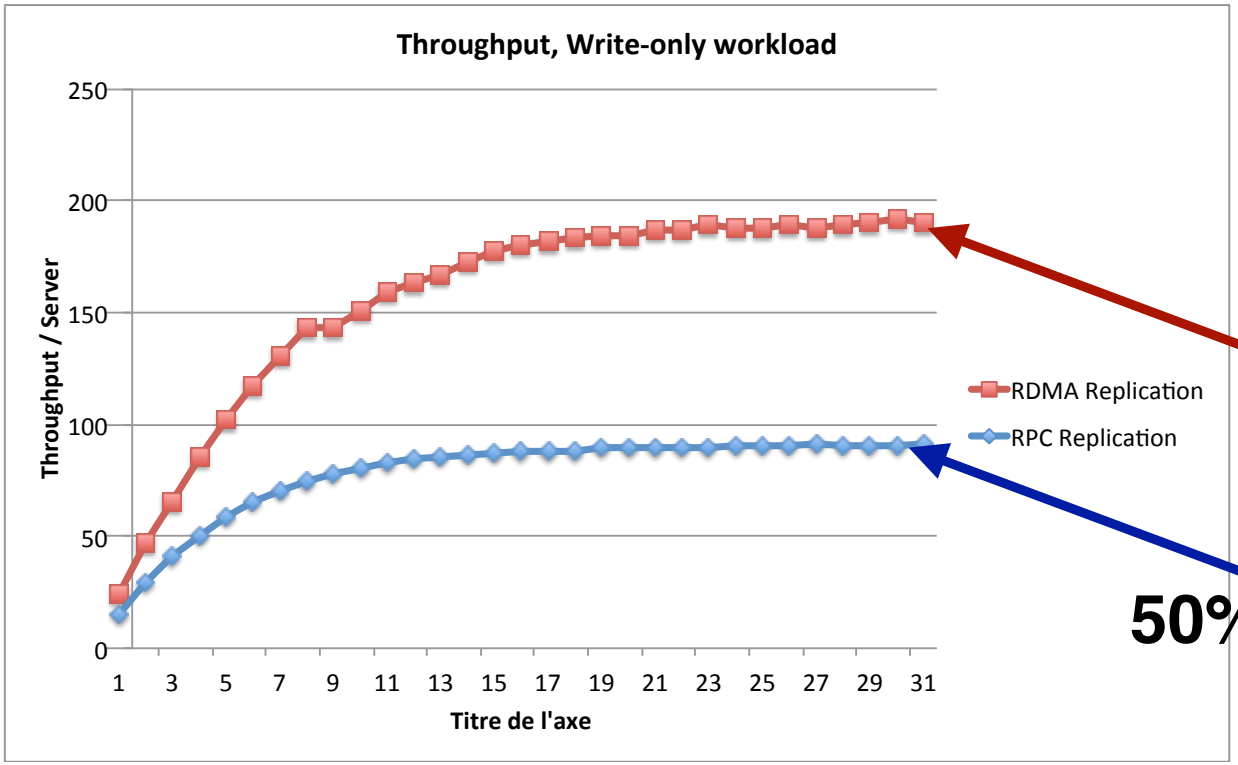
An Atomic RDMA-based Replication Protocol



Safe means: the sender is certain that his request was correctly appended

RDMA-based Replication in RAMCloud - Evaluation

4 Machines



3 Servers
3 Backups

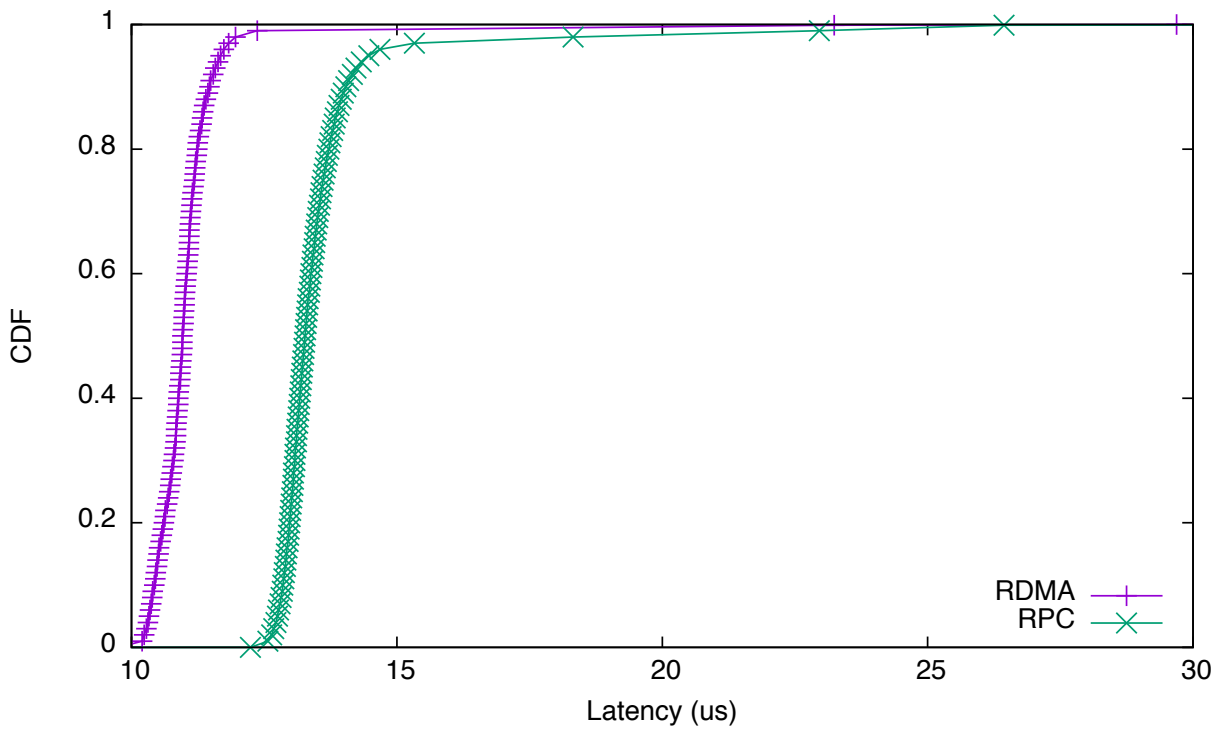
sustains max thput

50% decrease

X

RDMA-based Replication in RAMCloud - Evaluation

-CDF of write latencies (1 client - 3 servers)



~2-3x improvement in tail latency

X