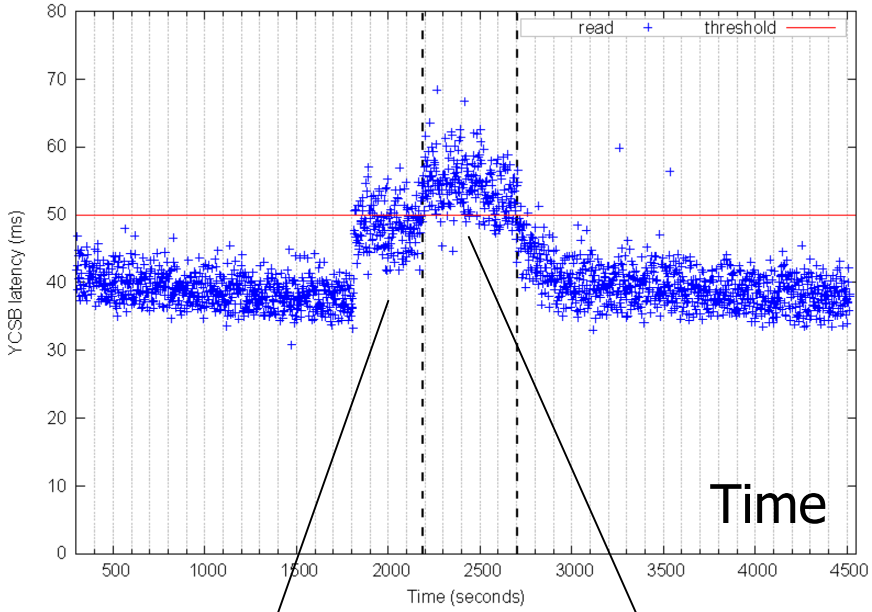
Adaptation in distributed NoSQL data stores

Kostas Magoutis

Department of Computer Science and Engineering
University of Ioannina, Greece

Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)

Workload variations lead to SLO violations



Latency

SLO

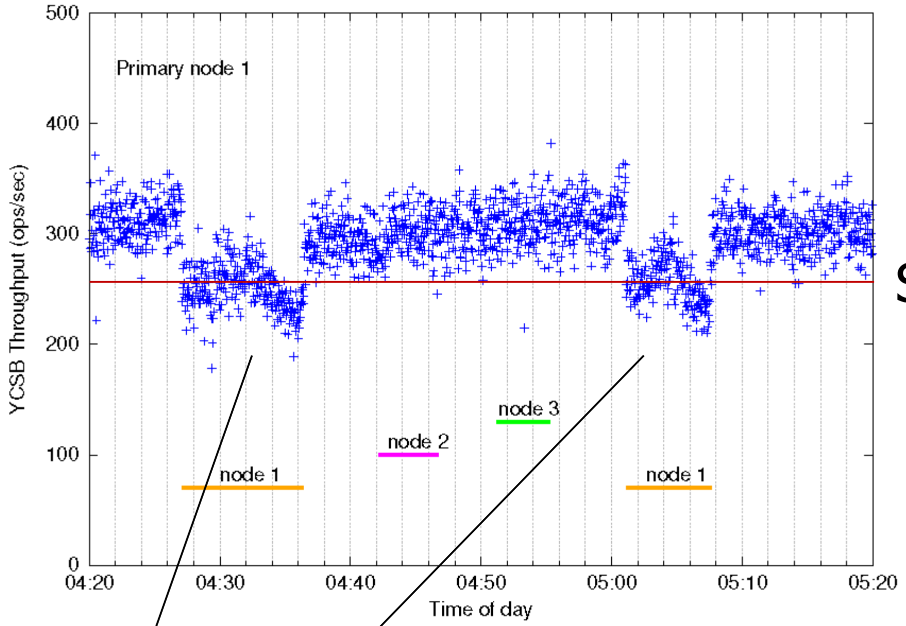
Time

Impact of workload increase

Adapt by adding a node

However, rebalancing has an impact

Background tasks lead to SLO violations



Throughput

SLO

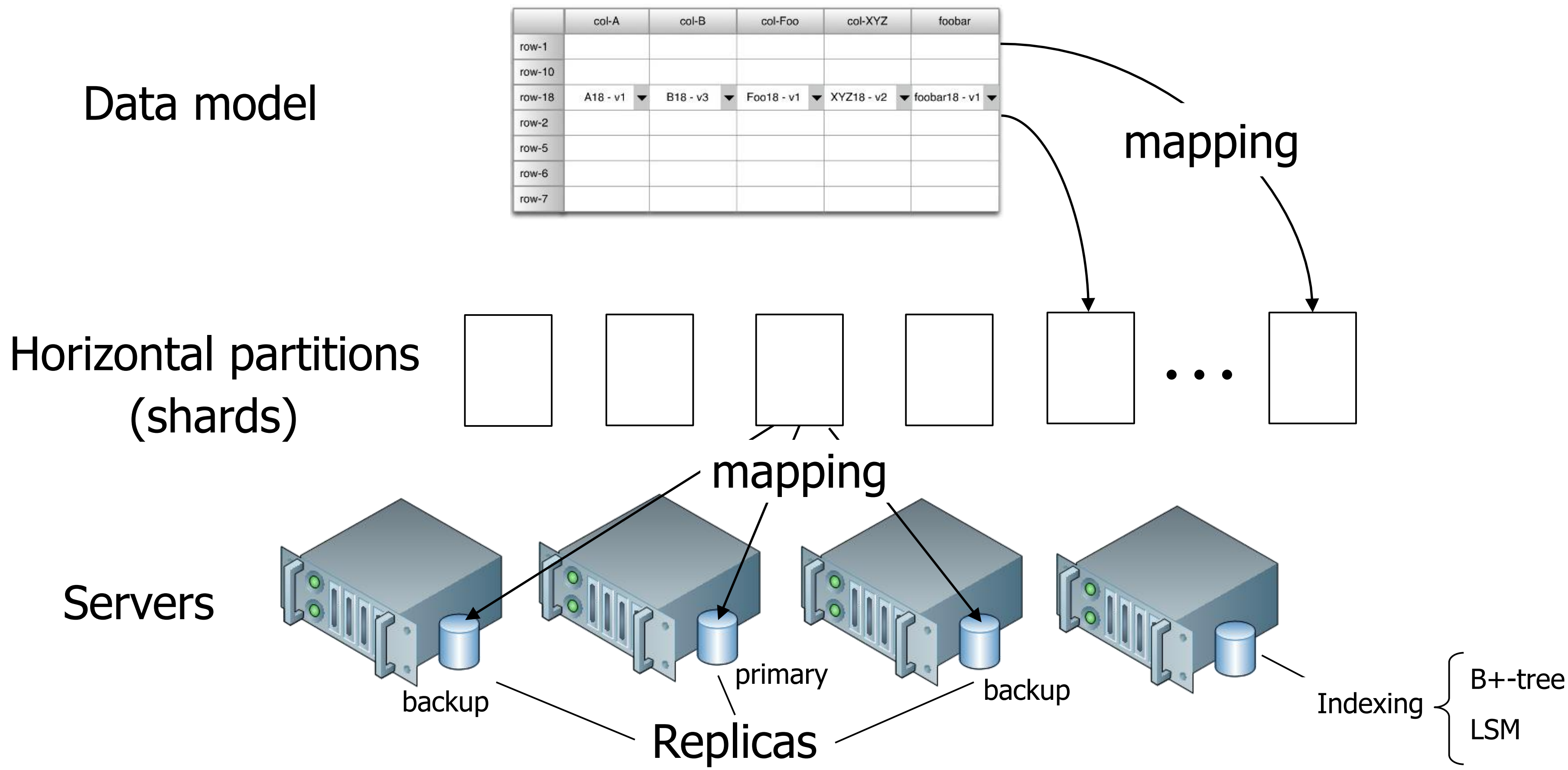
Impact of background-task induced overload at leader node

Adapt by reorganizing replica groups (changing leader)

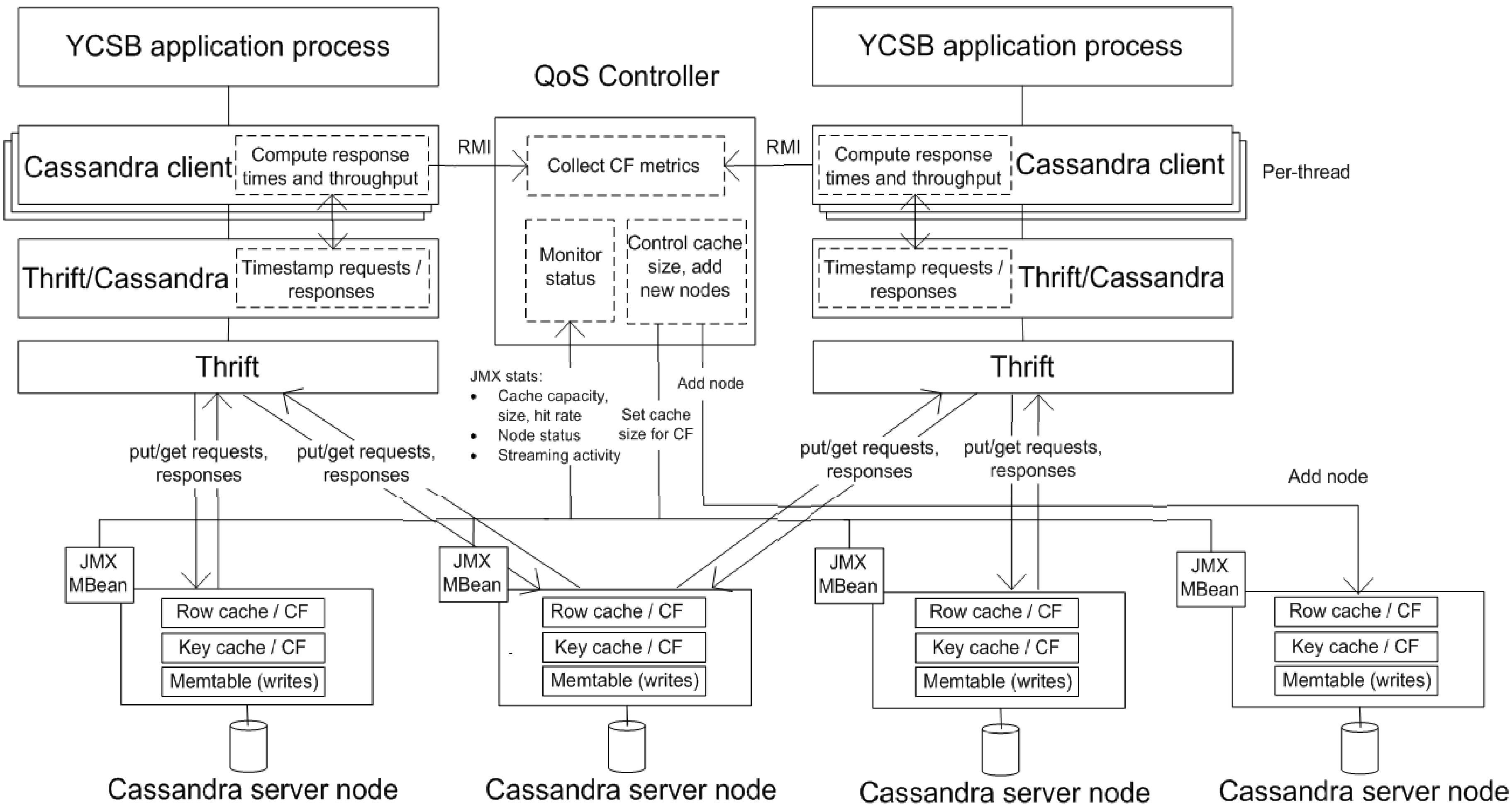
Agenda

- Workload or resource variations $\Rightarrow$ SLO violations
  - Need to adapt to maintain SLO
  - Examples: Elasticity, rebalancing, reconfiguration
- Feedback-loop based adaptation
  - Performance modeling via systematic measurements
  - Importance of fast, light rebalancing actions
- Adapting via overhead-hiding operations
  - Replica group leadership change
  - Hide overhead at the leader

NoSQL data stores: overview



QoS architecture for NoSQL data stores



Maria Chalkiadaki and Kostas Magoutis, Managing Service Performance in the Cassandra Distributed Storage System, in *Proc. of 5th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2013)*

Provisioning methodology

- Prediction of service capacity requirements
- Tables of measured performance results
  - Response time
  - Throughput

	Workload: W; Server type: S				
# Servers / # Clients	1	2	3	4	...
clients_1	r_1, t_1	r_2, t_2	r_3, t_3	r_4, t_4	r_5, t_5
clients_2	r_1, t_1	r_2, t_2	r_3, t_3	r_4, t_4	r_5, t_5
clients_3	r_1, t_1	r_2, t_2	r_3, t_3	r_4, t_4	r_5, t_5
...	r_1, t_1	r_2, t_2	r_3, t_3	r_4, t_4	r_5, t_5

Provisioning methodology

QoS specification

- 100% reads
- Zipf distribution
- Load: 512 threads
- Resp. time: 35ms

ZIPF-100% READS: AMAZON M1.SMALL

UNIFORM-100% READS: AMAZON M1.MEDIUM

ZIPF-100% READS: AMAZON M1.MEDIUM

# Clients \ # Servers	UNIFORM-100% READS: AMAZON M1.SMALL 2	3	4	5	6	7
128	25.4, 4.8	22.17, 6.14	17.61, 7.06	15.76, 8.17	12.78, 9.55	12.24, 10.4
256	51.28, 5.16	51.12, 4.88	40.94, 6.46	33.24, 7.8	26.6, 9.4	22.7, 10.7
512	116.9, 4.42	83.14, 5.58	70.27, 7.70	54.73, 9.25	44.24, 10.6	44.46, 10.6

# Clients \ # Servers	ZIPF-100% READS: AMAZON M1.SMALL 2	3	4	5	6	7
128	23.37, 5.55	19.7, 6.66	15.62, 7.72	13.98, 8.81	12.18, 10.17	10.75, 11.6
256	49.51, 5.47	37.92, 6.87	32.3, 8.5	25, 9.65	22.4, 11.1	18.01, 11.14
512	102.23, 5.21	76.15, 6.59	61.01, 8	51.45, 9.8	44.01, 10.9	$\boxed{34.6, 12.2}$

Exploring the accuracy of different regression approaches

- Interpolation exhibits 70-80% (avg) prediction accuracy in most cases – can we improve on this?
- Evaluate prediction accuracy using more advanced regression methods
  - Multivariate adaptive regression splines (MARS)
  - Support vector regression (SVR)
  - Artificial neural networks (ANN)

Flora Karniavoura and Kostas Magoutis, A Measurement-based Approach to Performance Prediction in NoSQL Systems, in *Proc. of 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*

Overall results

Predict performance for different cluster sizes

Regression Method	Case 1 Accuracy	Case 2 Accuracy	Case 3 Accuracy	Average
MARS	96.93	97.81	98.82	97.85
SVR	93.97	92.42	96.11	94.16
ANN	92.11	89.85	89.22	90.39

Flora Karniavoura and Kostas Magoutis, A Measurement-based Approach to Performance Prediction in NoSQL Systems, in *Proc. of 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*

Overall results

Predict performance for different load levels

Regression Method	Case 1 Accuracy	Case 2 Accuracy	Case 3 Accuracy	Average
MARS	96.93	97.81	98.82	97.85
SVR	93.97	92.42	96.11	94.16
ANN	92.11	89.85	89.22	90.39

Flora Karniavoura and Kostas Magoutis, A Measurement-based Approach to Performance Prediction in NoSQL Systems, in *Proc. of 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*

Overall results

Predict performance for different update settings

Regression Method	Case 1 Accuracy	Case 2 Accuracy	Case 3 Accuracy	Average
MARS	96.93	97.81	98.82	97.85
SVR	93.97	92.42	96.11	94.16
ANN	92.11	89.85	89.22	90.39

Flora Karniavoura and Kostas Magoutis, A Measurement-based Approach to Performance Prediction in NoSQL Systems, in *Proc. of 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*

Overall results

MARS provides better accuracy in all test cases

Regression Method	Case 1 Accuracy	Case 2 Accuracy	Case 3 Accuracy	Average
MARS	96.93	97.81	98.82	97.85
SVR	93.97	92.42	96.11	94.16
ANN	92.11	89.85	89.22	90.39

- MARS provides excellent accuracy
- SVR, ANN involve tuning (kernel, activation function)

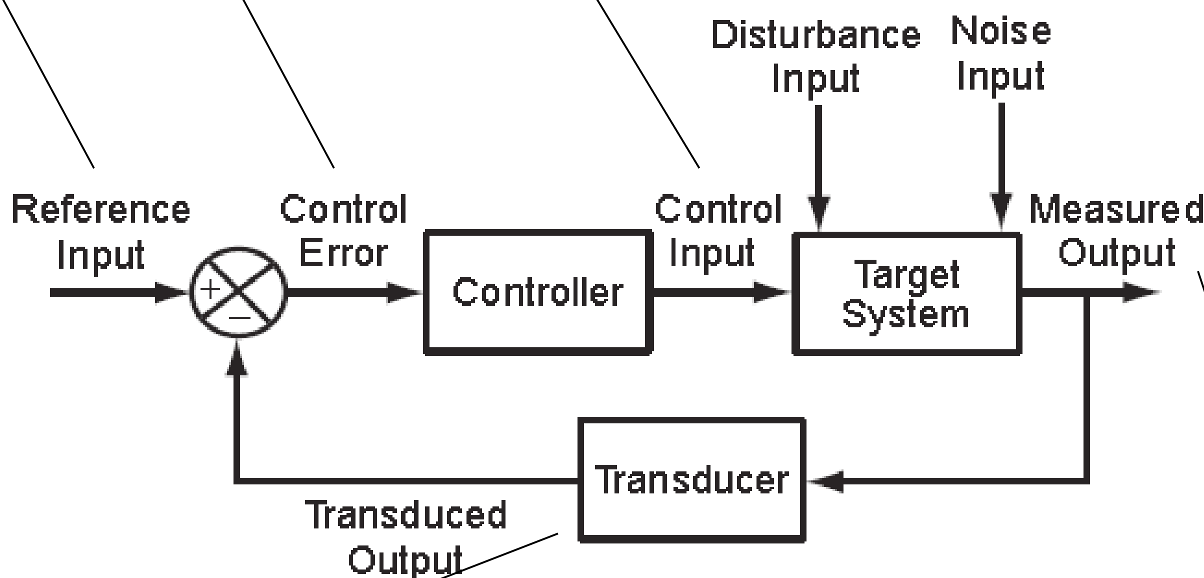
Flora Karniavoura and Kostas Magoutis, A Measurement-based Approach to Performance Prediction in NoSQL Systems, in *Proc. of 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*

Feedback-based control

r(k): desired value of measured output, e.g., 66% CPU utilization

difference between reference input and measured output

u(k): setting of parameter(s) that manipulate the system



y(k): measurable characteristic of target system (e.g. CPU)

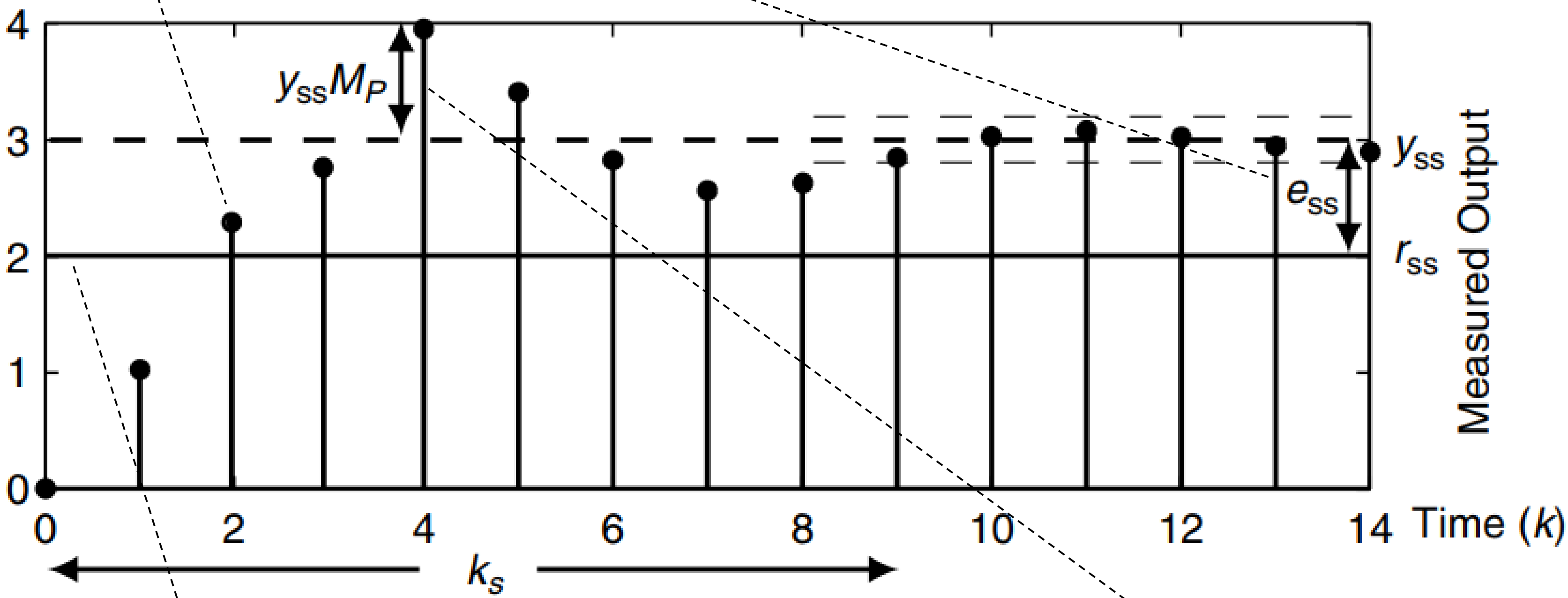
Transform the measured output so that it can be compared to reference input (e.g., smoothing)

T. Abdelzaher *et al.* "Introduction to control theory and its application to computing systems," Performance Modeling and Engineering, Springer, 2008

Behavior of a stable system

Measured output, eventually converges to $y_{ss}=3$

Steady-state error $e_{ss}=r_{ss}-y_{ss}=-1$



Reference input r_{ss} changes from 0 to 2

Settling time k_s

Maximum overshoot

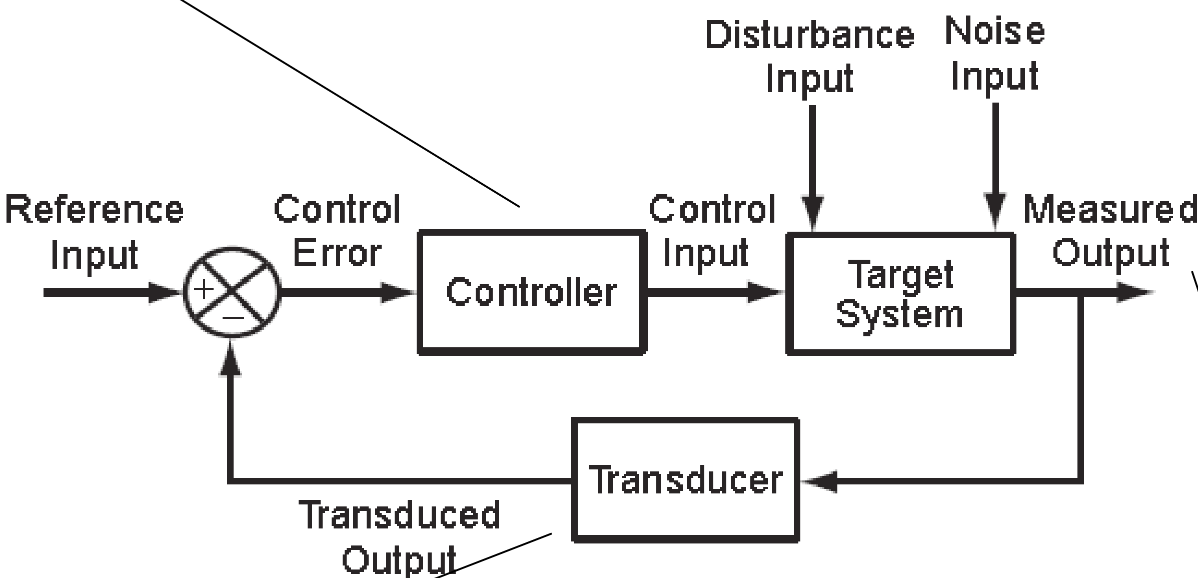
Goals: Stability, Accuracy, Short settling times, does not Overshoot (SASO)

T. Abdelzaher *et al.* "Introduction to control theory and its application to computing systems," Performance Modeling and Engineering, Springer, 2008

Integral control

Integral controller: provides incremental adjustments to u(k)

$$u(k+1) = u(k) + K_I e(k)$$



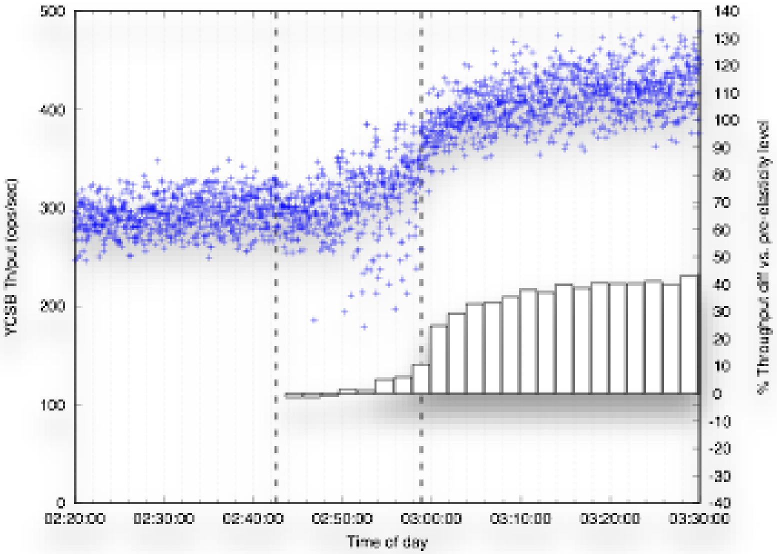
y(k): measurable characteristic of target system (e.g. CPU)

Transform the measured output so that it can be compared to reference input (e.g., smoothing)

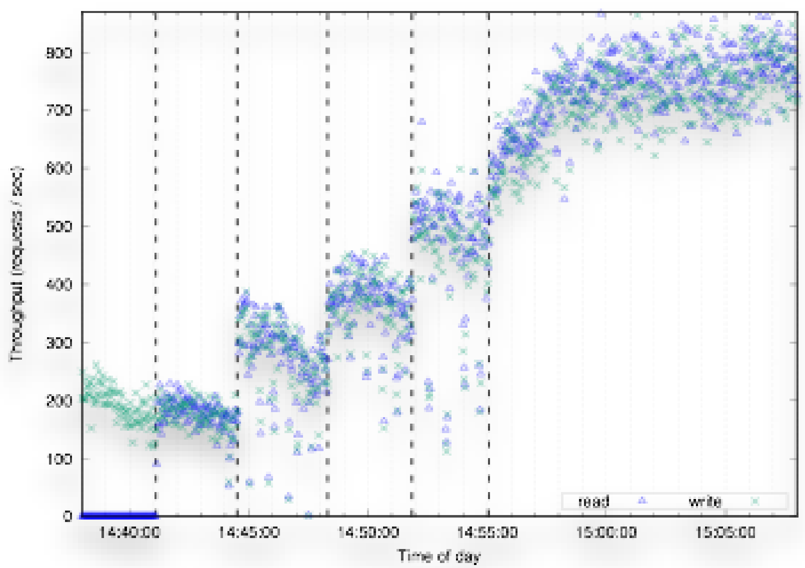
T. Abdelzaher *et al.* "Introduction to control theory and its application to computing systems," Performance Modeling and Engineering, Springer, 2008

Reducing the impact of data rebalancing via <u>incremental elasticity</u>

Results in smoother elasticity action



Processing capacity at joining node

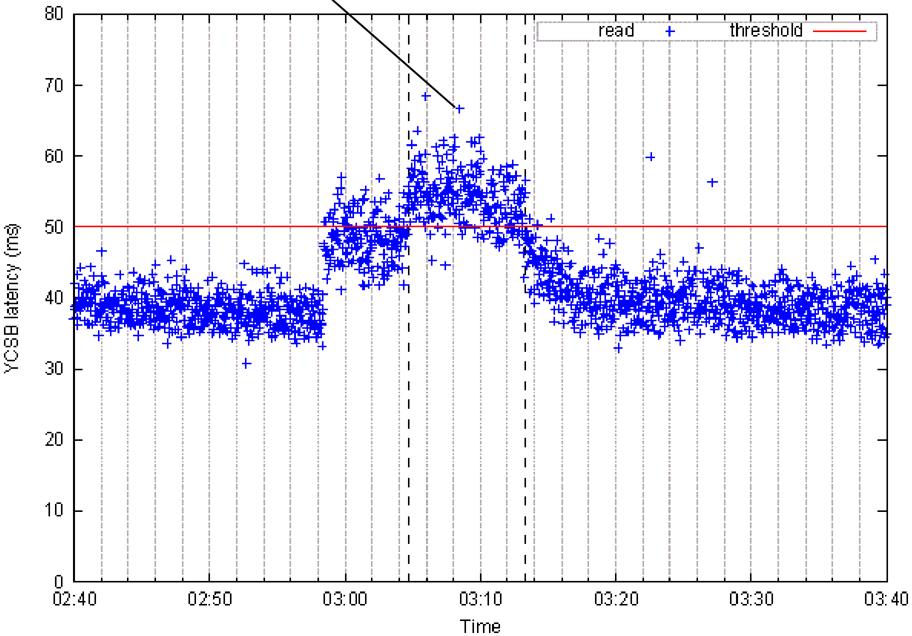


Antonis Papaioannou and Kostas Magoutis, Incremental elasticity for NoSQL data stores, in *Proc. of 36th Symposium on Reliable Distributed Systems (SRDS 2017)*, Hong Kong, China, September 27-29, 2017 (full paper)

Antonis Papaioannou and Kostas Magoutis, Incremental elasticity for NoSQL data stores, in *Proc. of 37th IEEE International Conference on Distributed Computing Systems (ICDCS 2017)*, Atlanta, GA, USA, June 5-8, 2017 (poster)

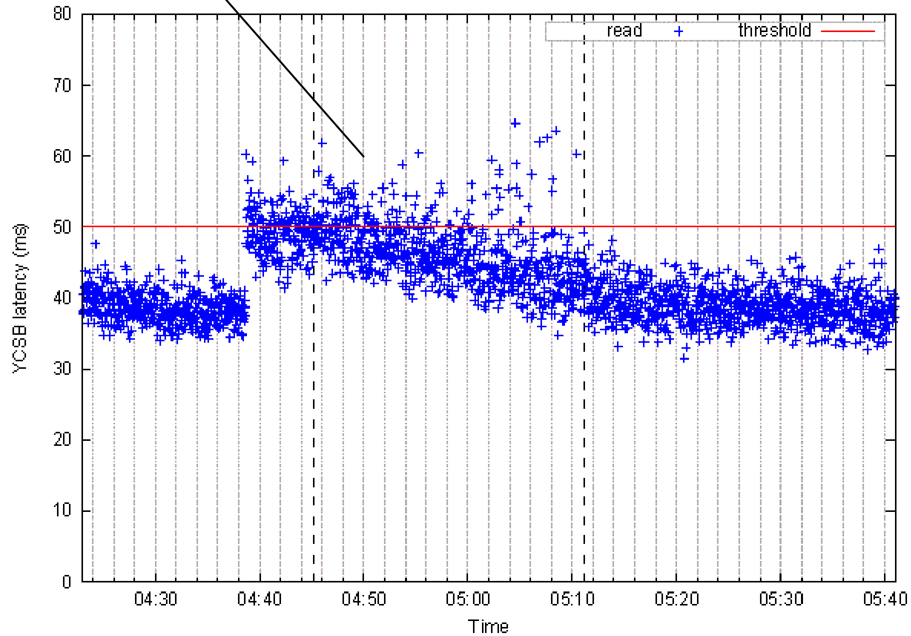
Impact on response-time SLO

Further response-time increase



(a) Parallel streaming

Smoother transition to new state

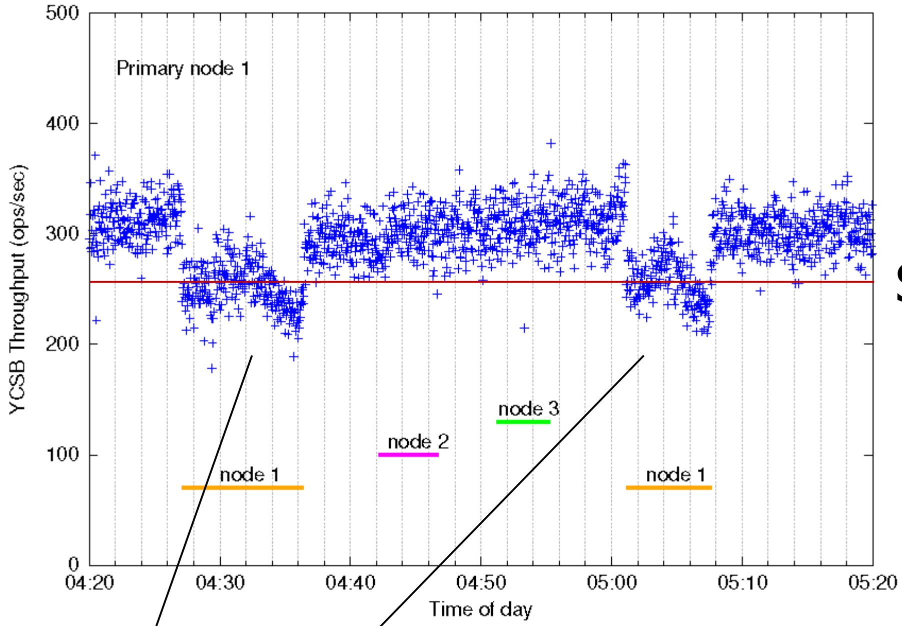


(b) Incremental streaming

- YCSB Workload B (95%-5%), SLO 50ms
- Load surge 20->30 YCSB threads
- Elasticity action 5 mins after surge

Adapting to impact of background tasks

Throughput



SLO

Impact of background-task induced overload at leader node

Adapt by reorganizing replica groups (changing leader)

Replica-group leadership change as a performance enhancing mechanism

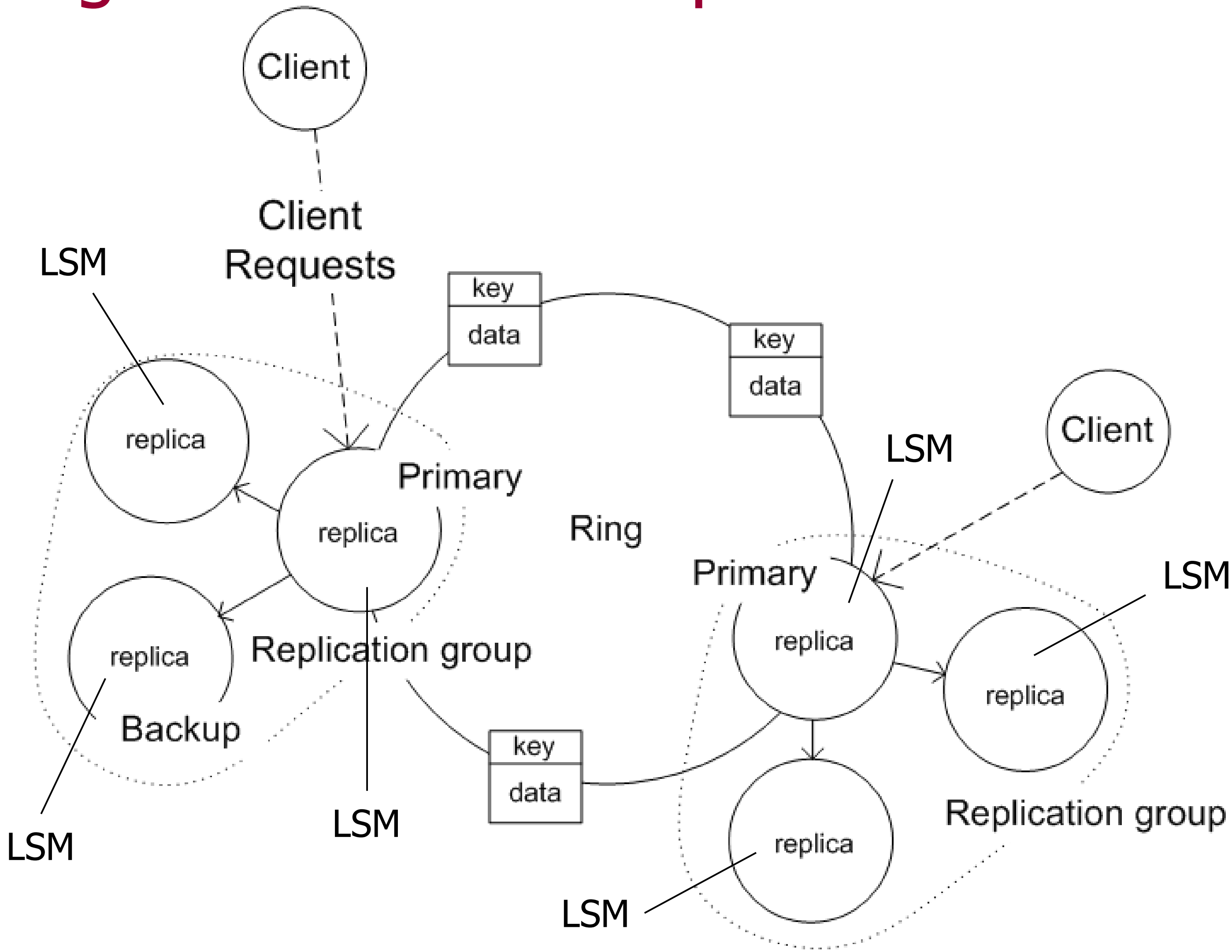
- Proactive replica group reorganizations provide rapid remedy to upcoming performance issues
  - Lightweight adaptation actions

- Replica group management increasingly possible via programmable APIs in NoSQL data stores
  - Examples: MongoDB, RethinkDB (both primary-backup)

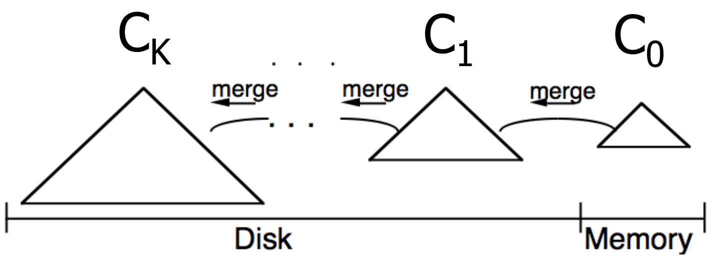
A. Papaioannou, K. Magoutis, "Replica-group leadership change as a performance enhancing mechanism in NoSQL data stores", 38th IEEE International Conference on Distributed Computing Systems (ICDCS'18), Vienna, Austria, Jul 6-9, 2018

P. Garefalakis, P. Papadopoulos, K. Magoutis, "ACaZoo: A Distributed Key-Value Store based on Replicated LSM-Trees ", Proc. 33rd IEEE Symposium on Reliable Distributed Systems (SRDS'14) 2014, Nara, Japan, Oct 6-9, 2014. **Best Student Paper**

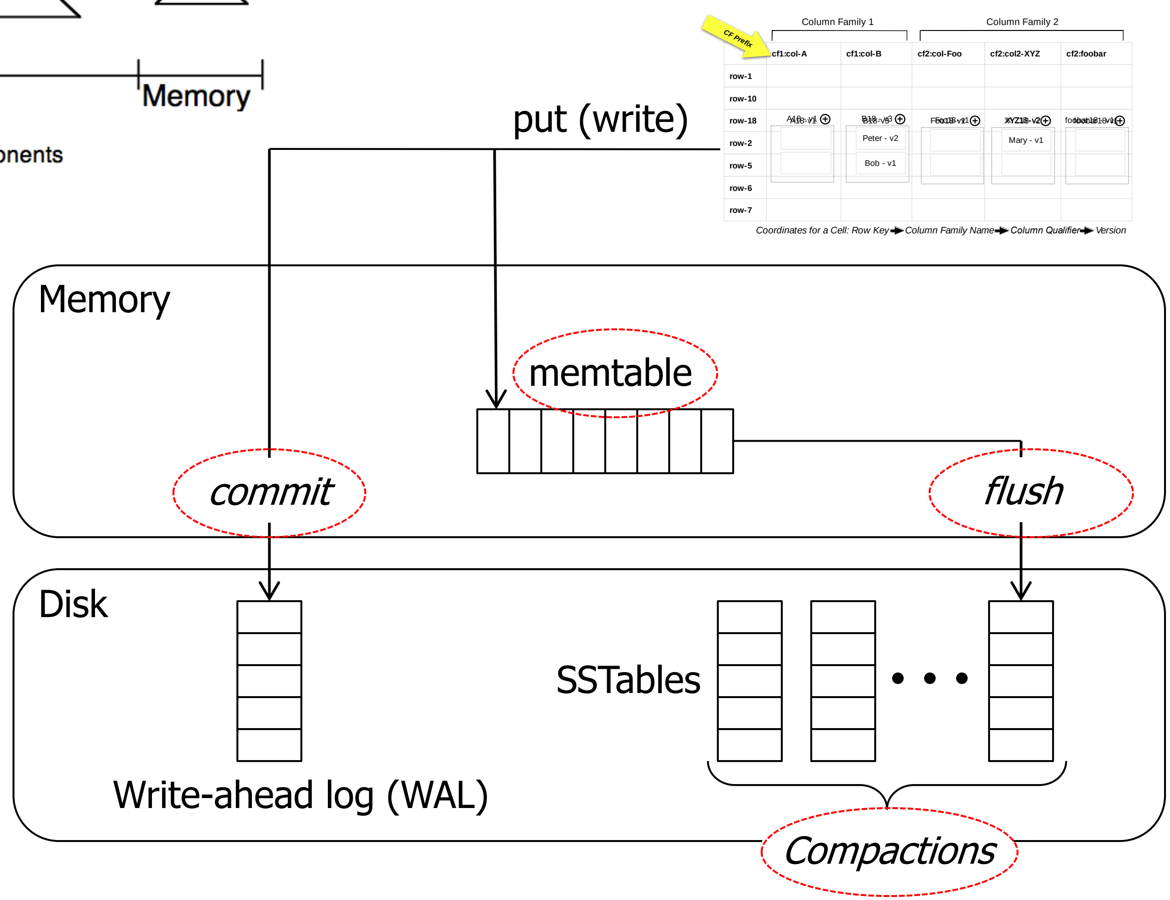
High-level view of replicated data store



Log-structured merge (LSM) trees

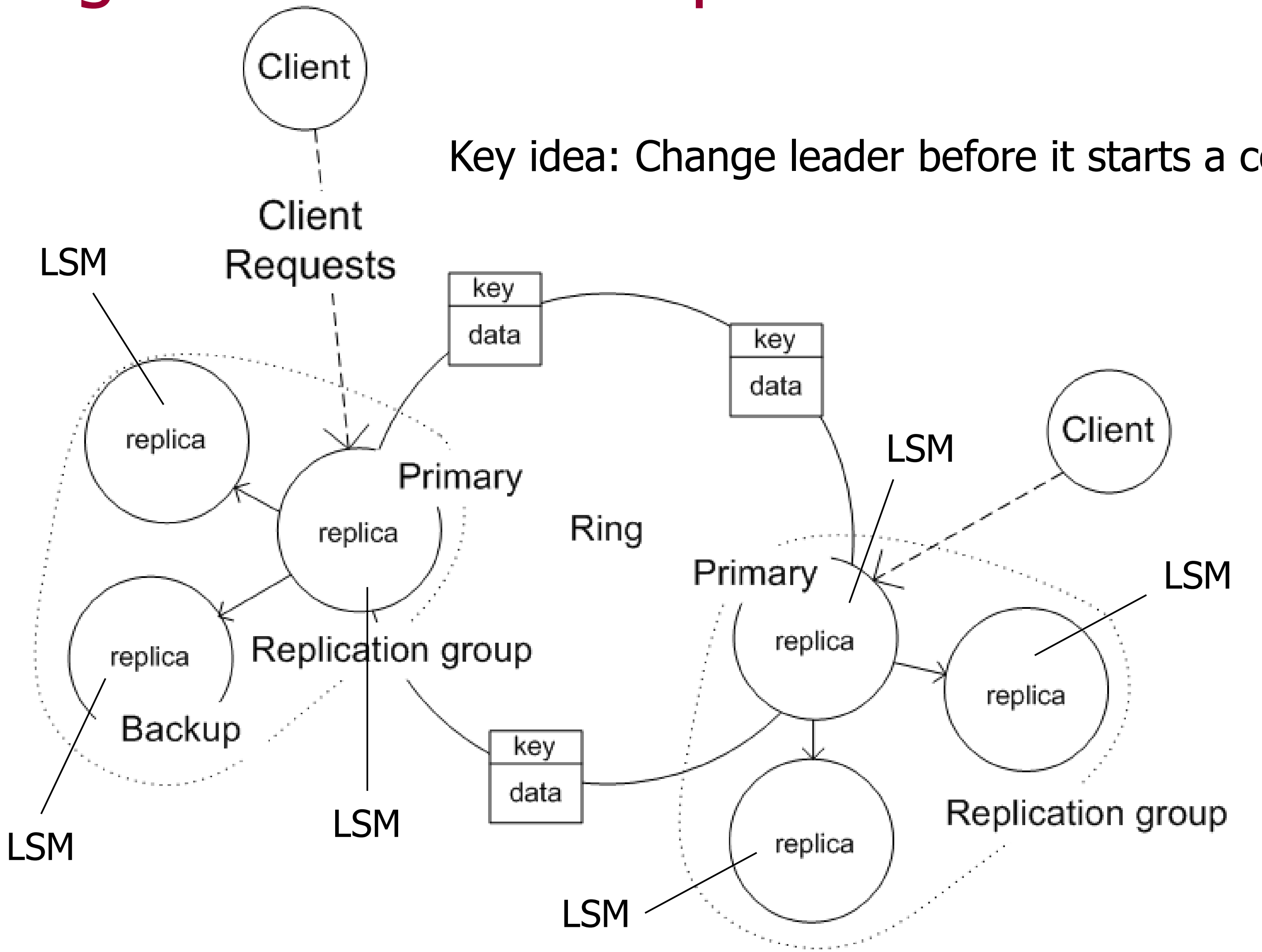


An LSM-tree of K+1 components

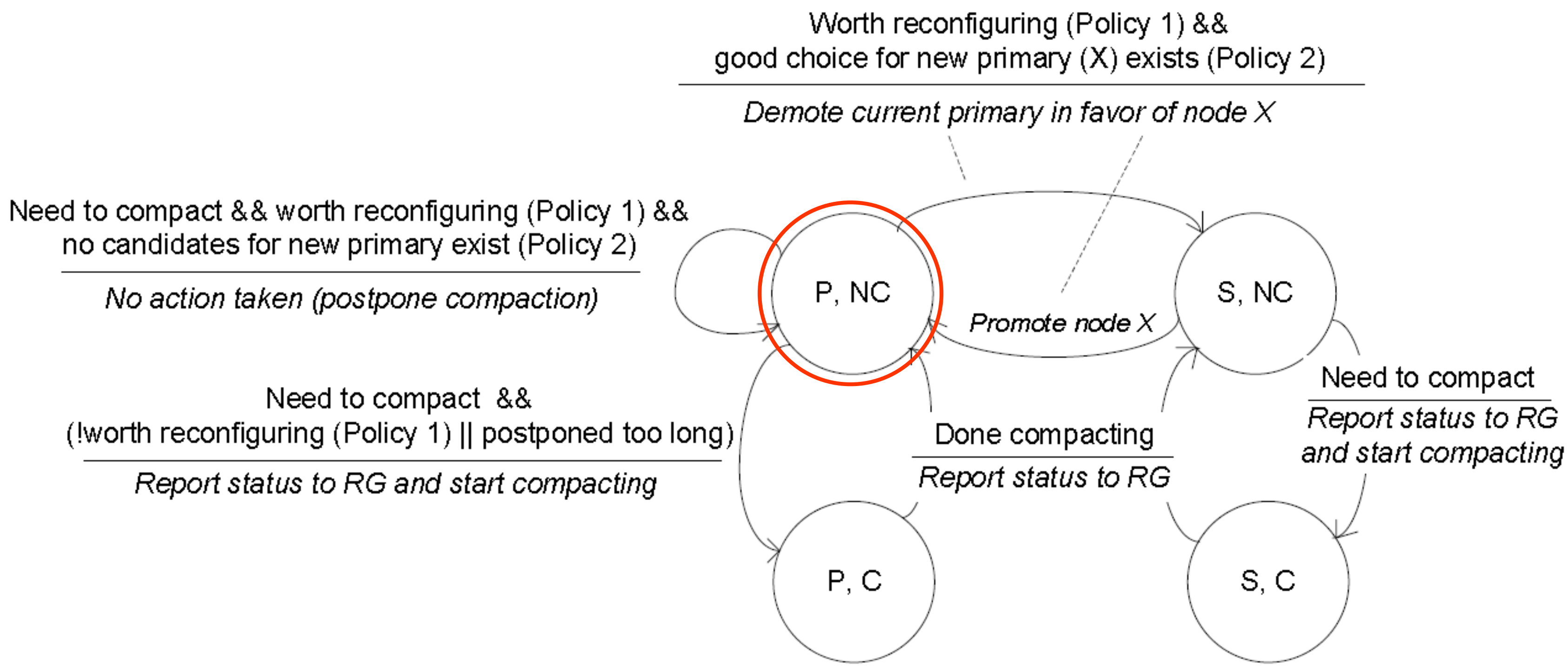


High-level view of replicated data store

Key idea: Change leader before it starts a compaction



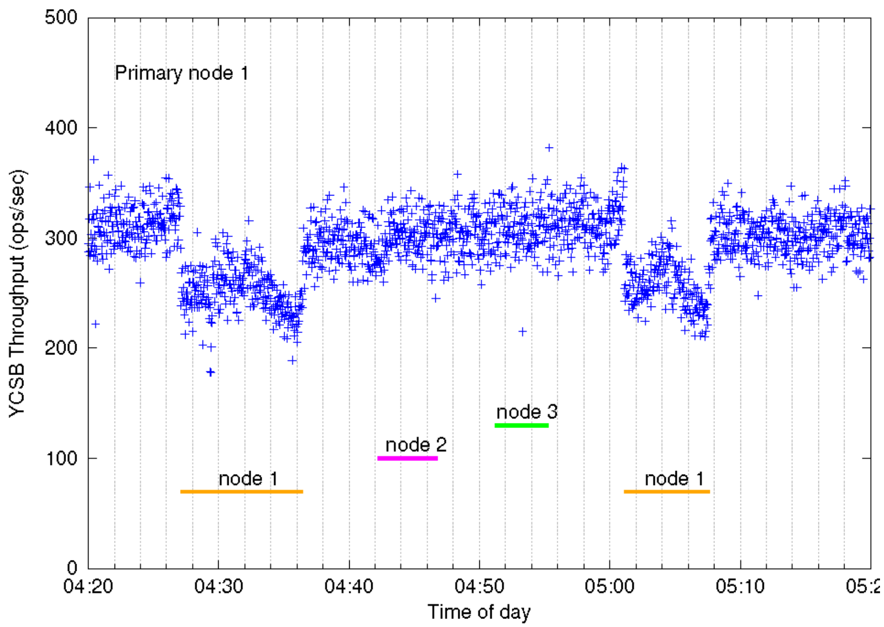
When to change a leader, whom to elect



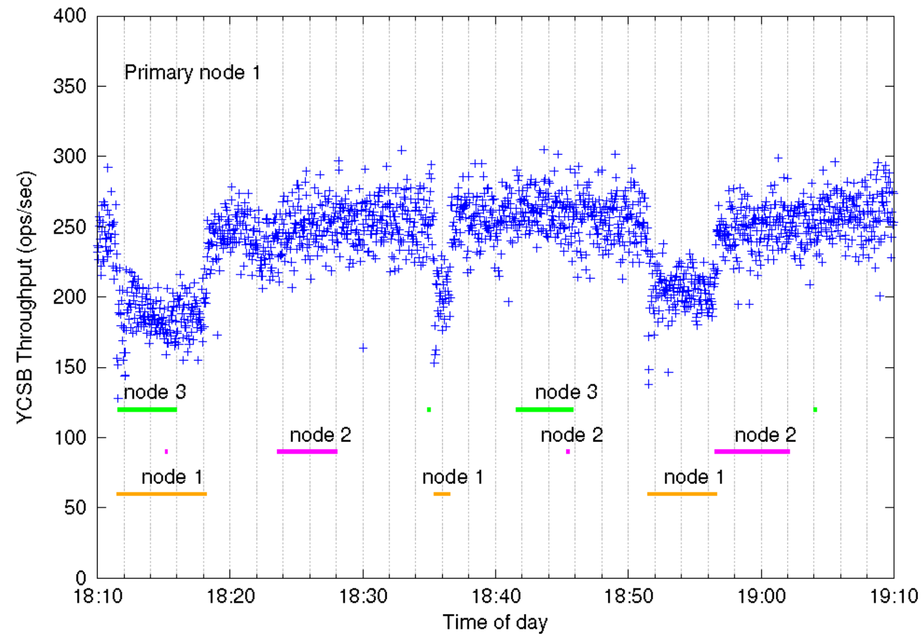
A. Papaioannou, K. Magoutis, "Replica-group leadership change as a performance enhancing mechanism in NoSQL data stores", 38th IEEE International Conference on Distributed Computing Systems (ICDCS'18), Vienna, Austria, Jul 6-9, 2018

Experimental results

Standard MongoDB RocksDB



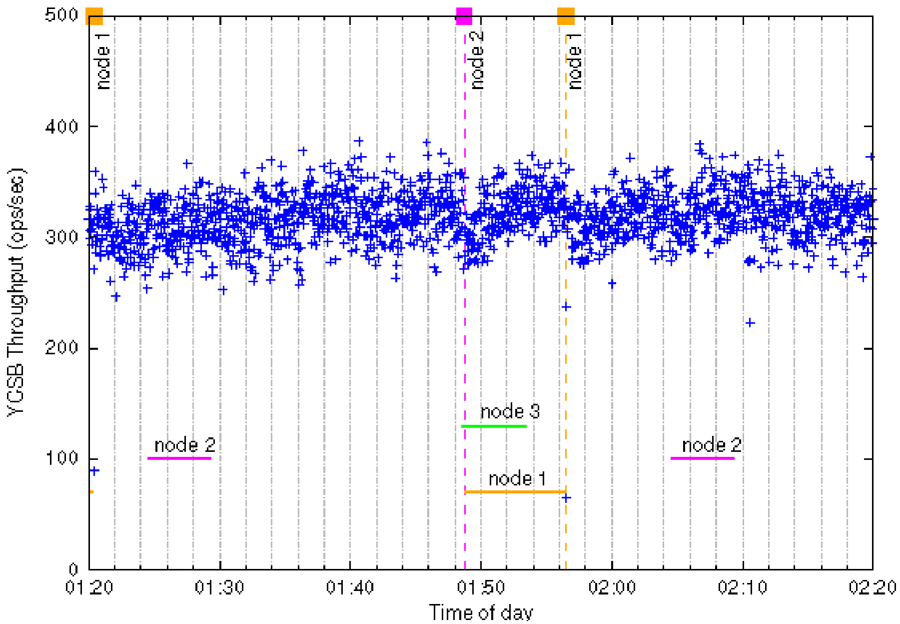
90% reads, 10% writes

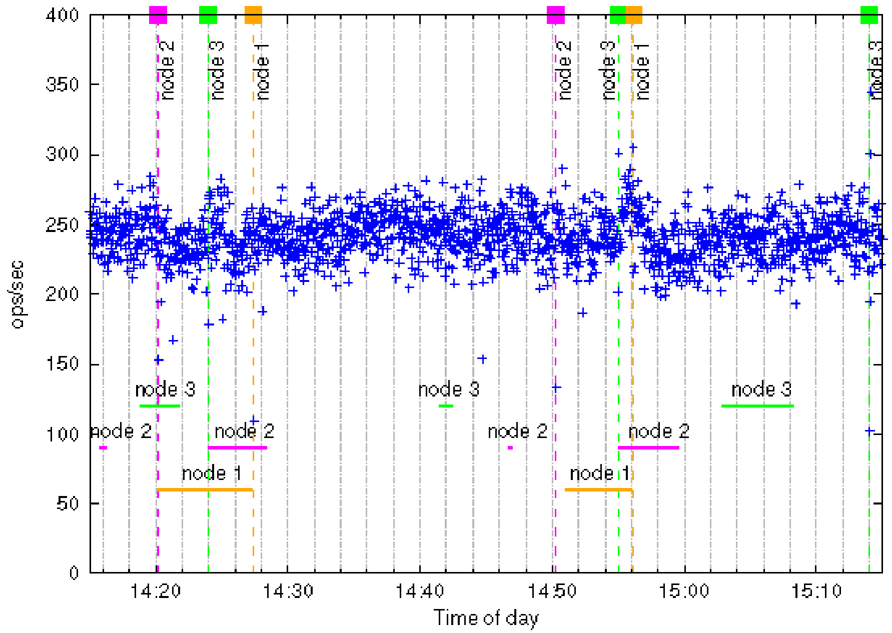


50% reads, 50% writes

Experimental results

MongoDB RocksDB with leadership changes

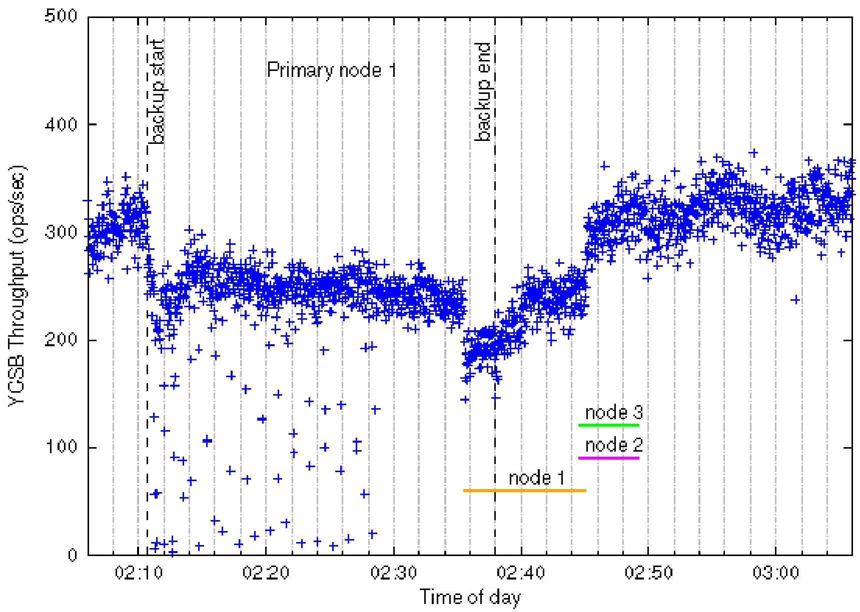




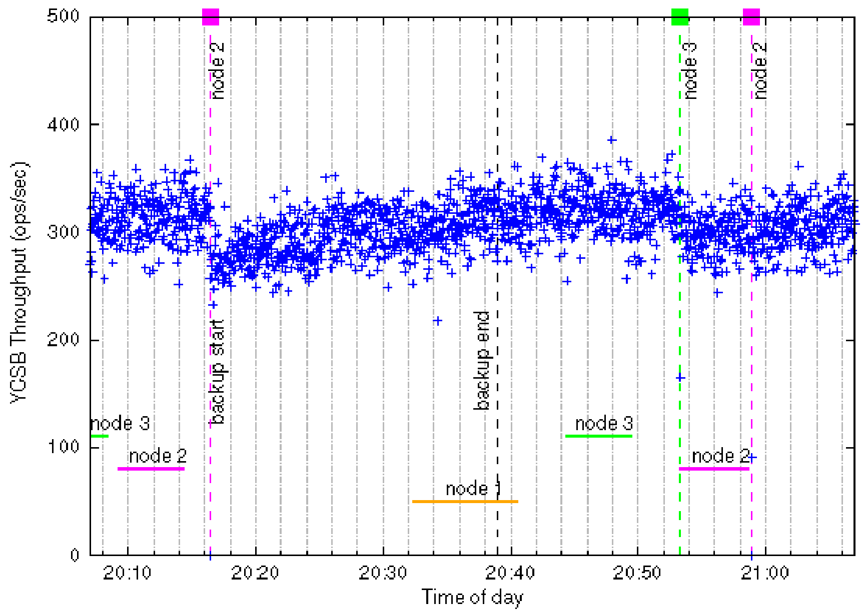
90% reads, 10% writes

50% reads, 50% writes

Data backup

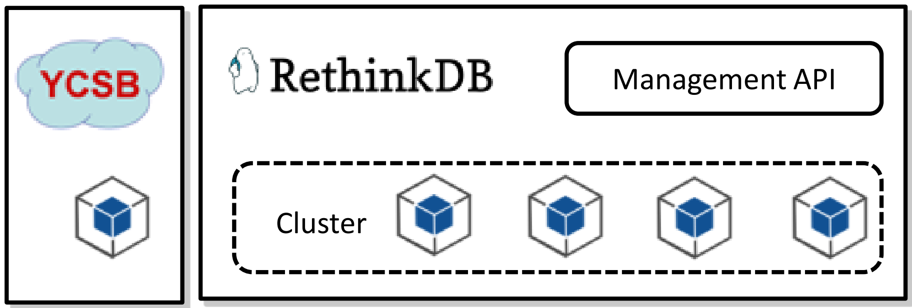


MongoDB RocksDB

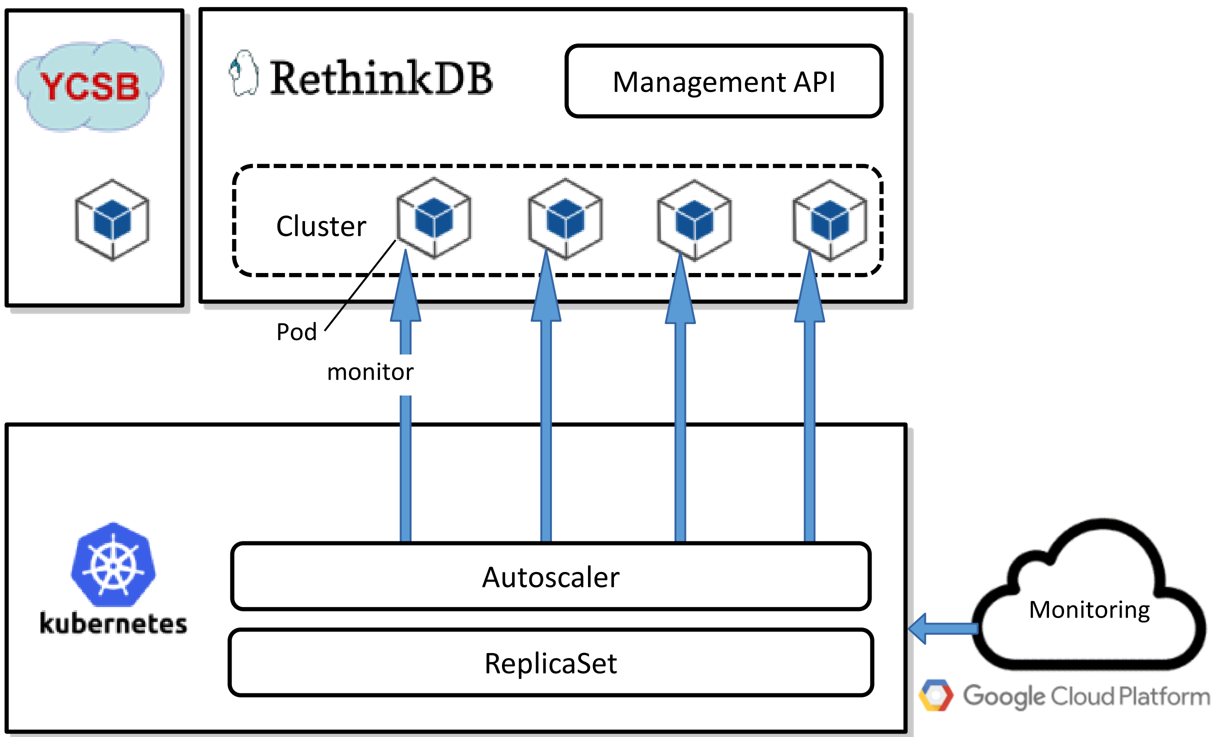


MongoDB RocksDB with leadership changes

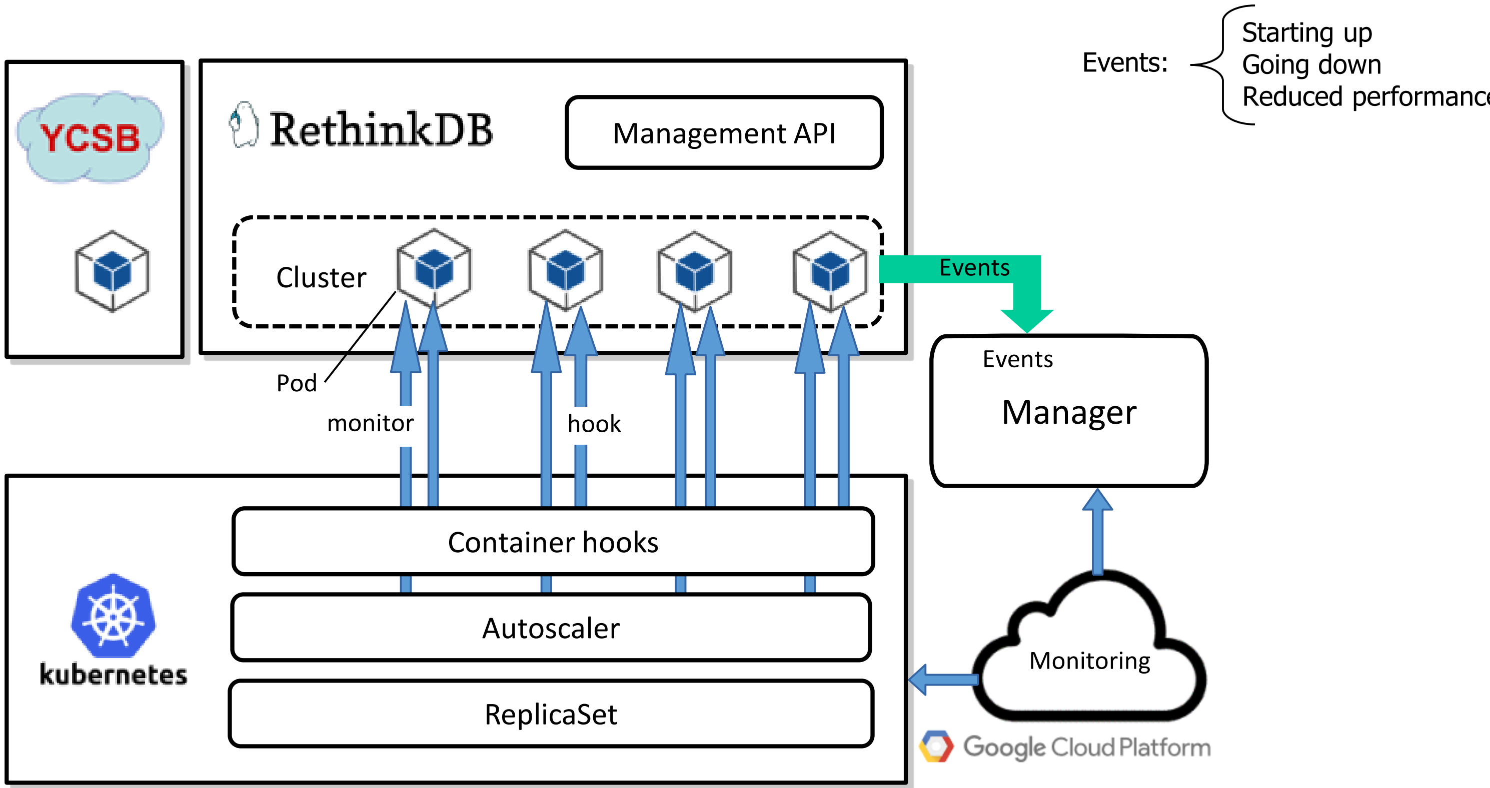
Cross-layer management of data stores



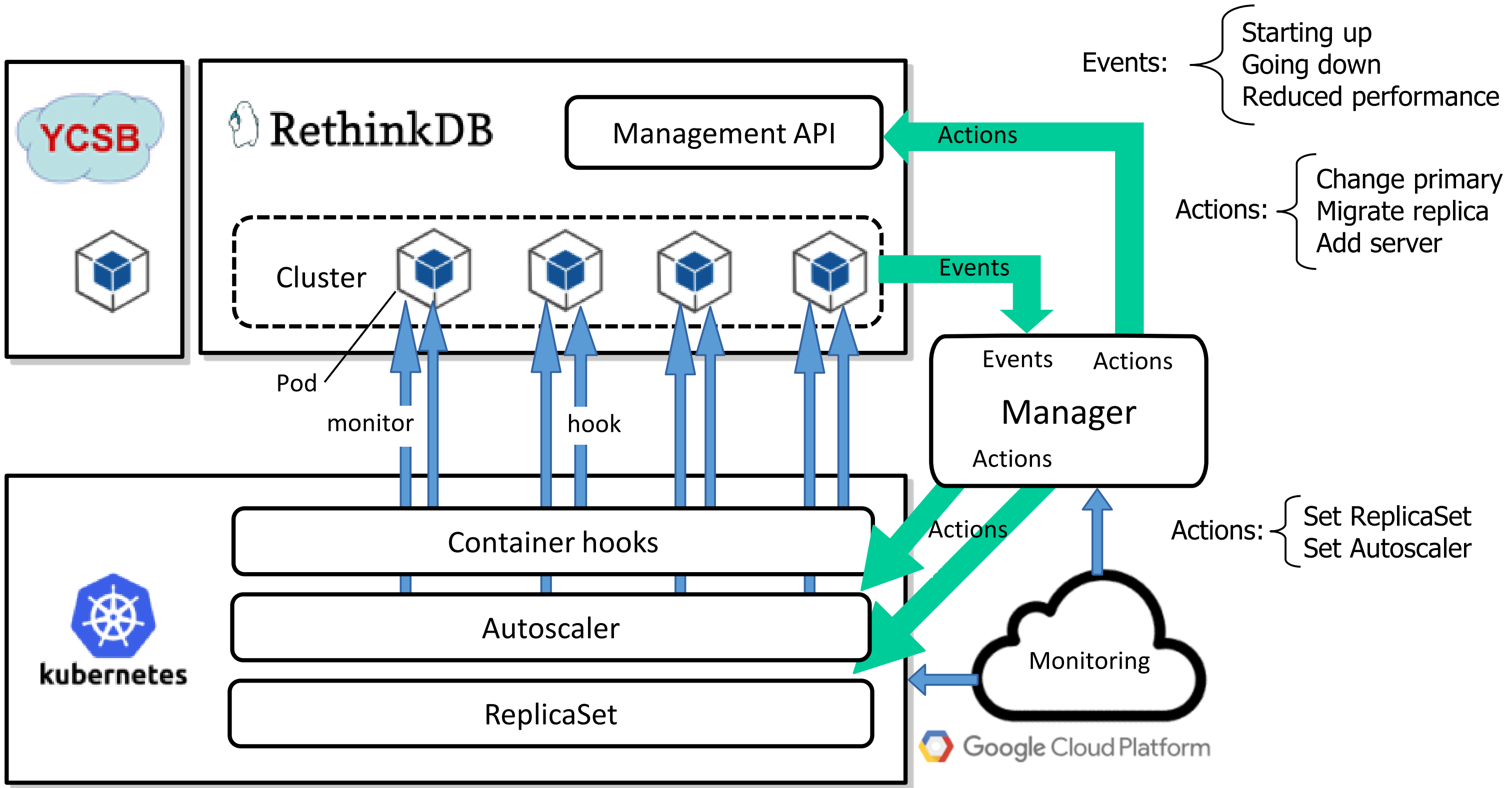
Cross-layer management of data stores



Cross-layer management of data stores



Cross-layer management of data stores

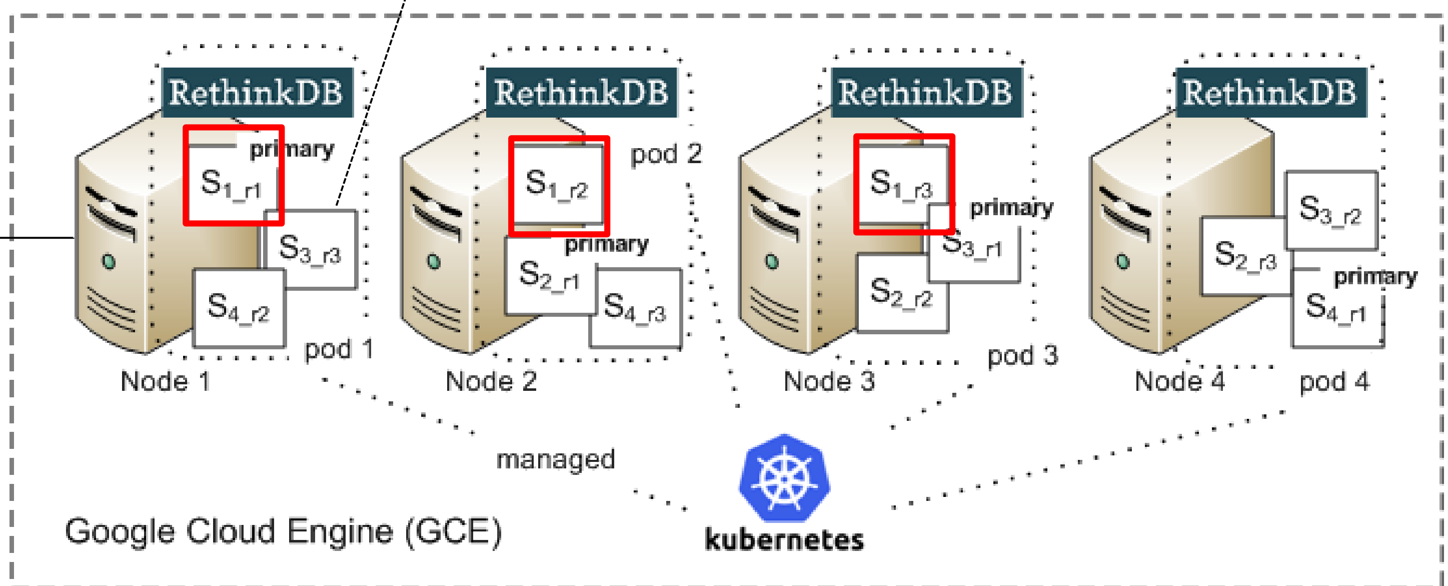


E. Bekas, K. Magoutis, "Cross-layer management of a containerized NoSQL data store", *15th IFIP/IEEE International Symposium on Integrated Network Management (IM 2017)*, 8-12 May 2017

A. Papaioannou, D. Metallidis, K. Magoutis, "Cross-layer management of distributed applications on multi-clouds", *13th IFIP/IEEE International Symposium on Integrated Network Management (IM 2015),* Ottawa, Canada, May 11-15, 2015

Experimental testbed

shard = horizontal partition

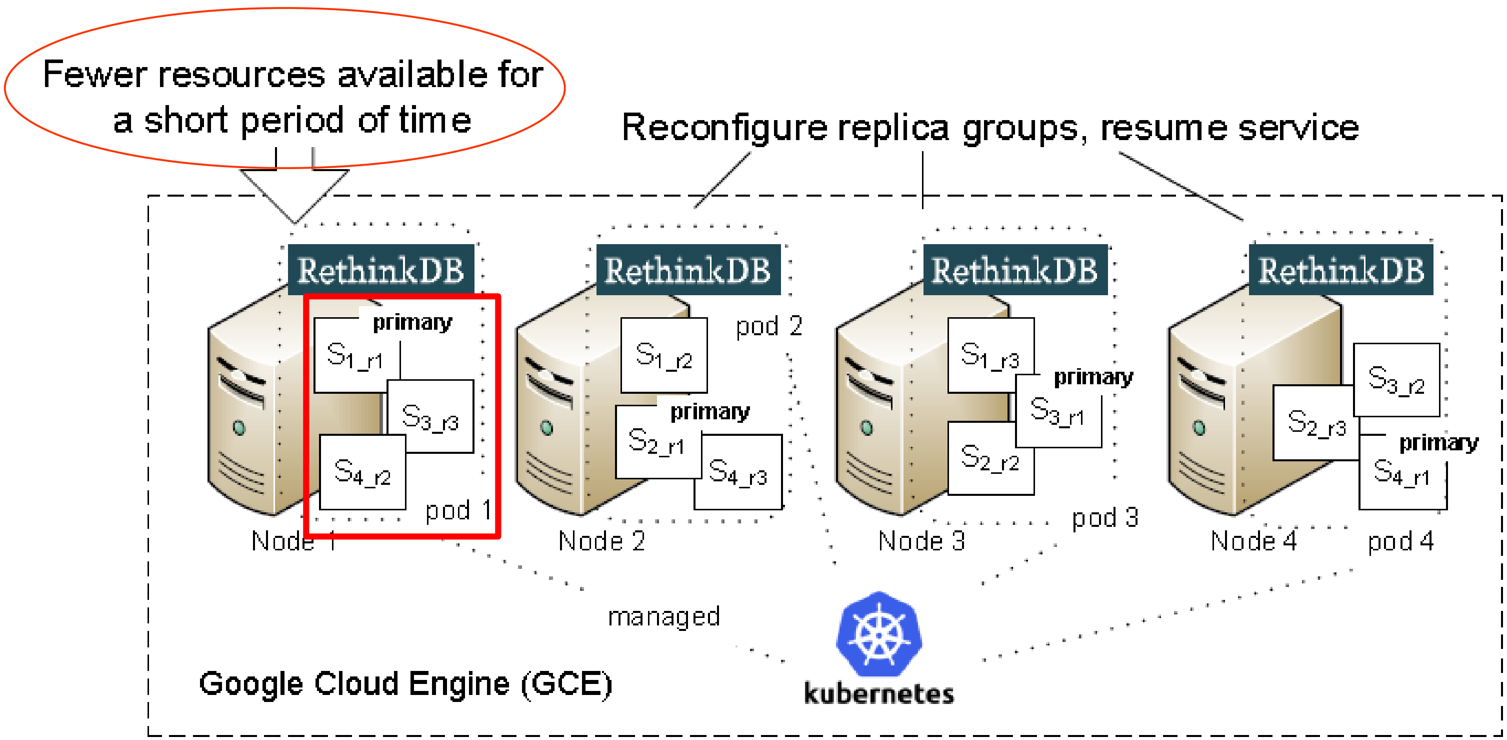


2 vCPUs 2.6GHz Intel Xeon E5
13GB RAM
SSD

YCSB settings

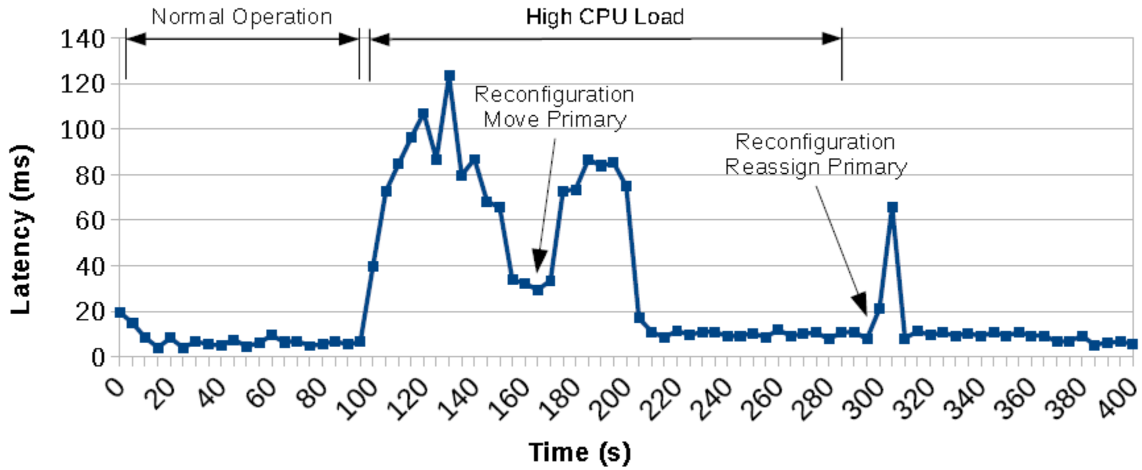
- 50-50 reads/writes
- 16 client threads
- target 1000 operations/sec

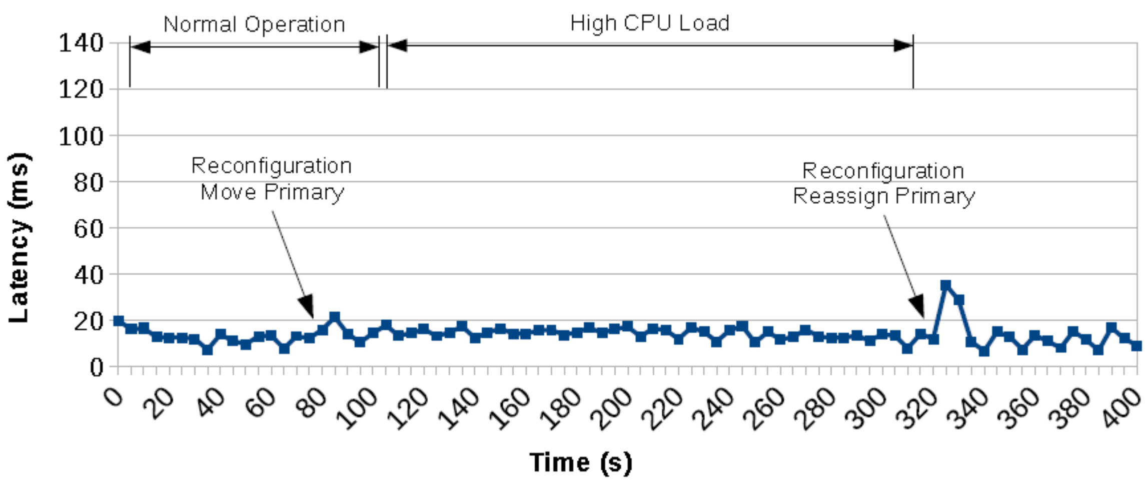
Temporary offload via RG reorganization



- Move S_1 primary out of Node 1

Temporary offload via RG reorganization





Summary

- Proactive replica group reorganizations provide rapid remedy to upcoming performance issues
  - Lightweight adaptation actions

- Functionality previously unavailable as infrastructure-level events invisible to NoSQL middleware
  - Richer feedback useful: How long is impact expected to last?

References

- A. Papaioannou, K. Magoutis, "**Replica-group leadership change as a performance enhancing mechanism in NoSQL data stores**", 38th IEEE International Conference on Distributed Computing Systems (ICDCS'18), Vienna, Austria, Jul 6-9, 2018
- Antonis Papaioannou and Kostas Magoutis, **"Incremental elasticity for NoSQL data stores"**, in *Proc. of 36th Symposium on Reliable Distributed Systems (SRDS 2017)*, Hong Kong, China, Sep 27-29, 2017
- Flora Karniavoura and Kostas Magoutis, **"A Measurement-based Approach to Performance Prediction in NoSQL Systems"**, in *Proc. of 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*
- Antonis Papaioannou and Kostas Magoutis, **"Incremental elasticity for NoSQL data stores"**, in *Proc. of 37th IEEE International Conference on Distributed Computing Systems (ICDCS 2017)*, Atlanta, GA, USA, June 5-8, 2017
- E. Bekas, K. Magoutis, **"Cross-layer management of a containerized NoSQL data store"**, in *Proc. of 15th IFIP/IEEE International Symposium on Integrated Network Management (IM 2017)*, 8-12 May 2017
- A. Papaioannou, D. Metallidis, K. Magoutis, **"Cross-layer management of distributed applications on multi-clouds"**, *IFIP/IEEE International Symposium on Integrated Network Management (IM 2015),* Ottawa, Canada, May 11-15, 2015
- P. Garefalakis, P. Papadopoulos, K. Magoutis, **"ACaZoo: A Distributed Key-Value Store based on Replicated LSM-Trees"**, Proc. 33rd IEEE Symposium on Reliable Distributed Systems (SRDS'14) 2014, Nara, Japan, Oct 6-9, 2014. **Best Student Paper**
- Maria Chalkiadaki and Kostas Magoutis, **"Managing Service Performance in the Cassandra Distributed Storage System"**, in *Proc. of 5th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2013)*, Bristol, UK, December 2-5, 2013

Questions?

FORTH
Institute of Computer Science

UNICORN

H2020 GA no. 731846 EU project