PolyCert: Polymorphic Self-Optimizing Replication for In-Memory Transactional Grids

Maria Couceiro, Paolo Romano, Luis Rodrigues

INESC-ID
Instituto Superior Tecnico, Universidade Tecnica de Lisboa

WDTM - 22 February 2012

Index

Introduction

- In-memory transactional data grids are an alternative to relational distributed databases
- Key/value store data model
  - Spurred out of the NoSQL movement
- In-memory storage
  - Durability via replication
- Higher performance, scalability and elasticity

- Example of target applications: distributed transactional memory

Introduction

- Transactional Memory is a powerful paradigm to develop concurrent applications
- Programmers only need to identify sequences of instructions that access/modify concurrent objects
- Results: more reliable code and shorter development time

Introduction

Replication

- Key mechanism to ensure data durability in case of failures
- Algorithms inspired in the replication of database systems

- Different protocols behave differently according to the workload
- Static configurations may lead to sub-optimal performances

We need a dynamic solution capable of guaranteeing the best performance in any possible scenario

Introduction

Replication

- Key mechanism to ensure data durability in case of failures
- Algorithms inspired in the replication of database systems

- Different protocols behave differently according to the workload
- Static configurations may lead to sub-optimal performances

We need a dynamic solution capable of guaranteeing the best performance in any possible scenario

Introduction

Replication

- Key mechanism to ensure data durability in case of failures
- Algorithms inspired in the replication of database systems

- Different protocols behave differently according to the workload
- Static configurations may lead to sub-optimal performances

We need a dynamic solution capable of guaranteeing the best performance in any possible scenario

Index

Certification Protocols

- Transactions execute locally
- When they are ready to commit, a message is atomically broadcast to the network
- Replicas validate the transaction when this message is received
  - A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set)
- The transaction is committed or discarded based on the outcome of the validation

Protocol dependent

- Message contents
- Validation process

Certification Protocols

- Transactions execute locally
- When they are ready to commit, a message is atomically broadcast to the network
- Replicas validate the transaction when this message is received
  - A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set)
- The transaction is committed or discarded based on the outcome of the validation

Protocol dependent

- Message contents
- Validation process

Certification Protocols

- Transactions execute locally
- When they are ready to commit, a message is atomically broadcast to the network
- Replicas validate the transaction when this message is received
  - A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set)
- The transaction is committed or discarded based on the outcome of the validation

Protocol dependent

- Message contents
- Validation process

Certification Protocols

- Transactions execute locally
- When they are ready to commit, a message is atomically broadcast to the network
- Replicas validate the transaction when this message is received
  - A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set)
- The transaction is committed or discarded based on the outcome of the validation

Protocol dependent

- Message contents
- Validation process

Certification Protocols

- Transactions execute locally
- When they are ready to commit, a message is atomically broadcast to the network
- Replicas validate the transaction when this message is received
  - A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set)
- The transaction is committed or discarded based on the outcome of the validation

Protocol dependent

- Message contents
- Validation process

Certification Protocols

- Non Voting
- Bloom Filter Certification
- Voting

Non Voting

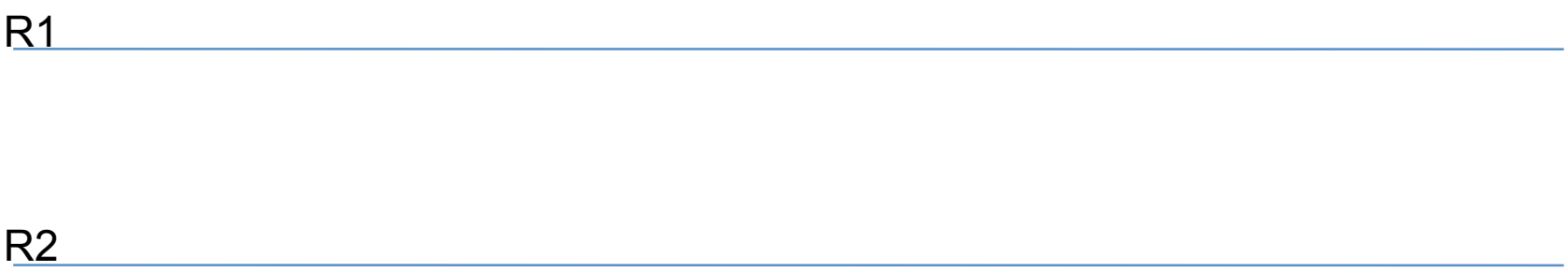


Figure: Example of the execution of the Non Voting Protocol.

Non Voting

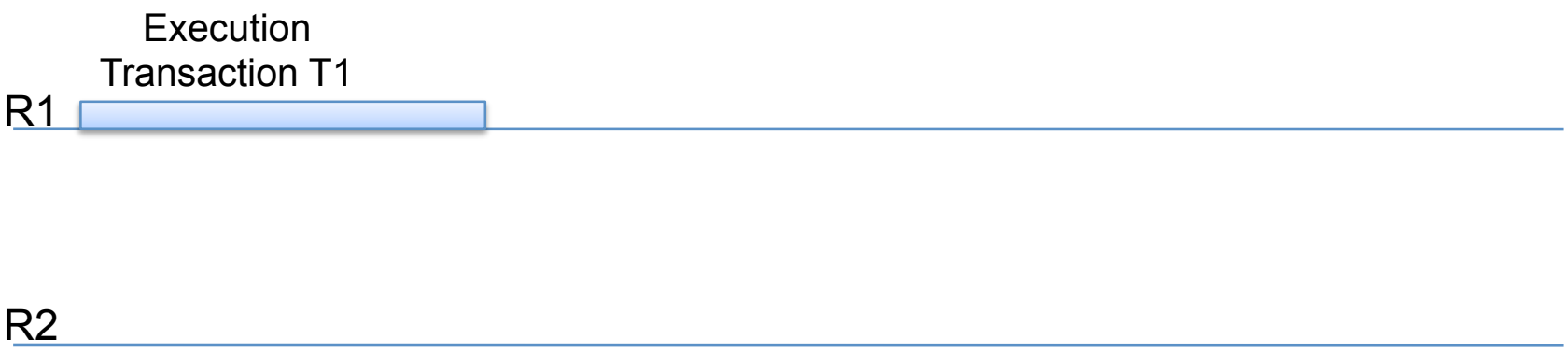


Figure: Example of the execution of the Non Voting Protocol.

Non Voting

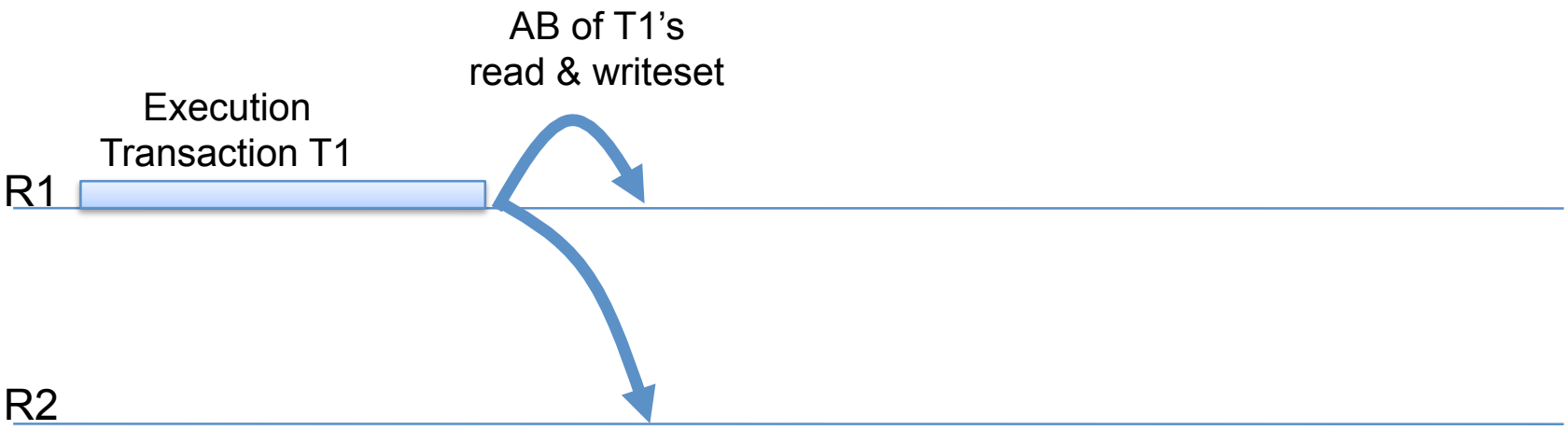


Figure: Example of the execution of the Non Voting Protocol.

Non Voting

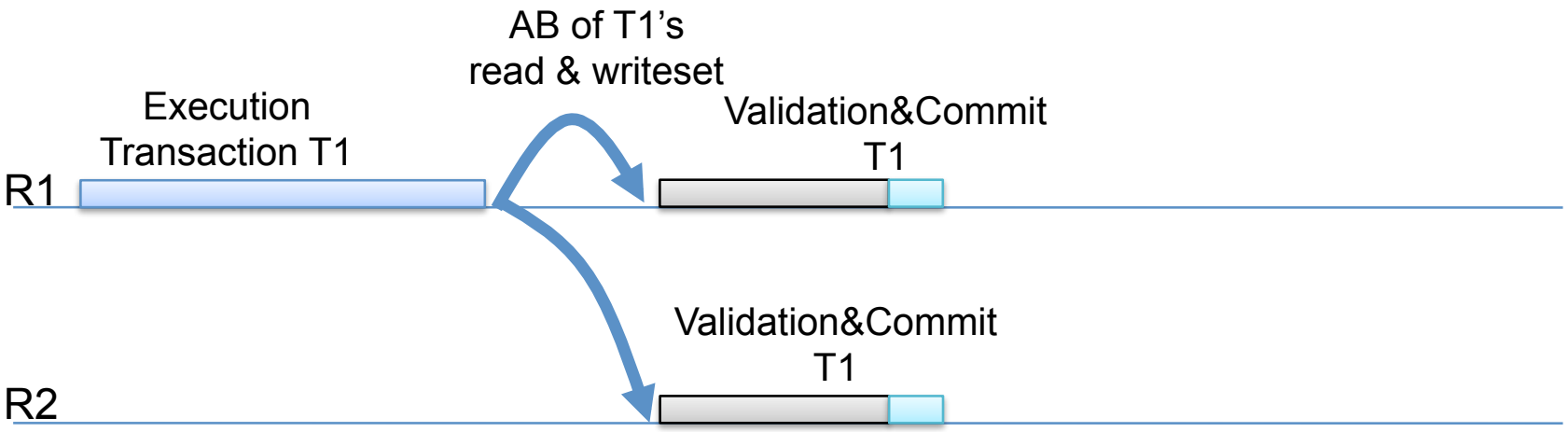


Figure: Example of the execution of the Non Voting Protocol.

Non Voting

Message

Read and write set

Validation

Each replica validates using the received read set

Pros:

- Simple validation process

Cons:

- Potentially large messages

Non Voting

Message

Read and write set

Validation

Each replica validates using the received read set

Pros:

- Simple validation process

Cons:

- Potentially large messages

Non Voting

Message

Read and write set

Validation

Each replica validates using the received read set

Pros:

- Simple validation process

Cons:

- Potentially large messages

Bloom Filter Certification

Bloom filters:

- Space-efficient data structure for test membership queries
- Probabilistic answer to “Is elem contained in BF?”
  - No false negatives: a “no” answer is always correct
  - False positives: A “yes” answer may be false
- Compression is a function of a (tunable) false positive rate

Bloom Filter Certification

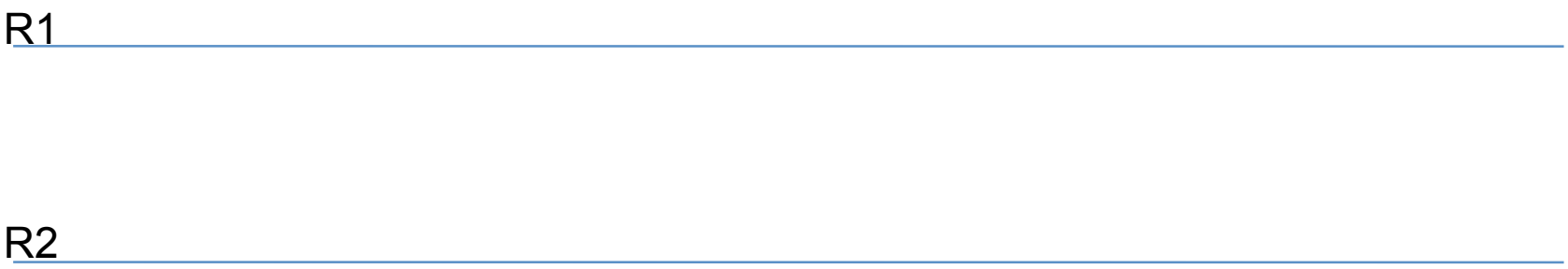


Figure: Example of the execution of the Bloom Filter Certification Protocol.

Bloom Filter Certification

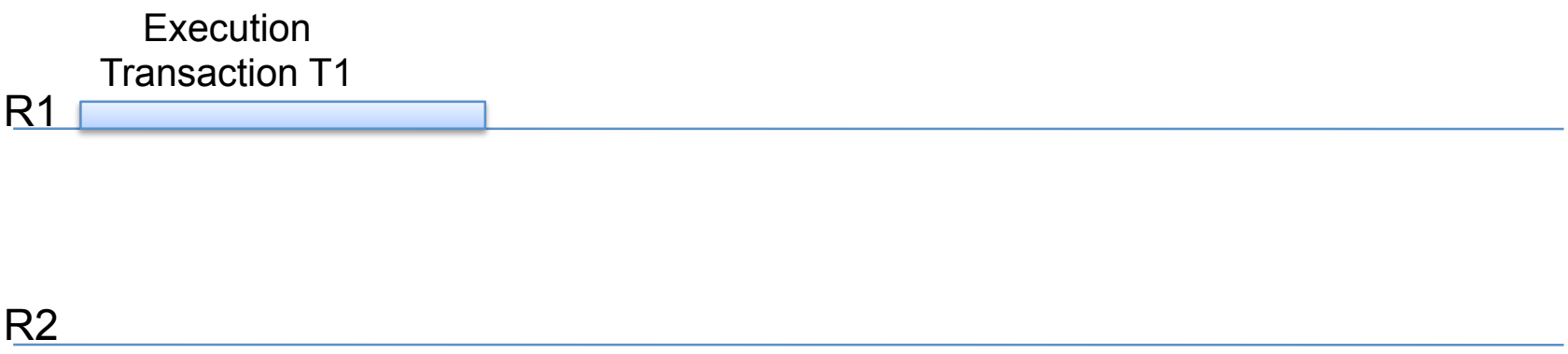


Figure: Example of the execution of the Bloom Filter Certification Protocol.

Bloom Filter Certification

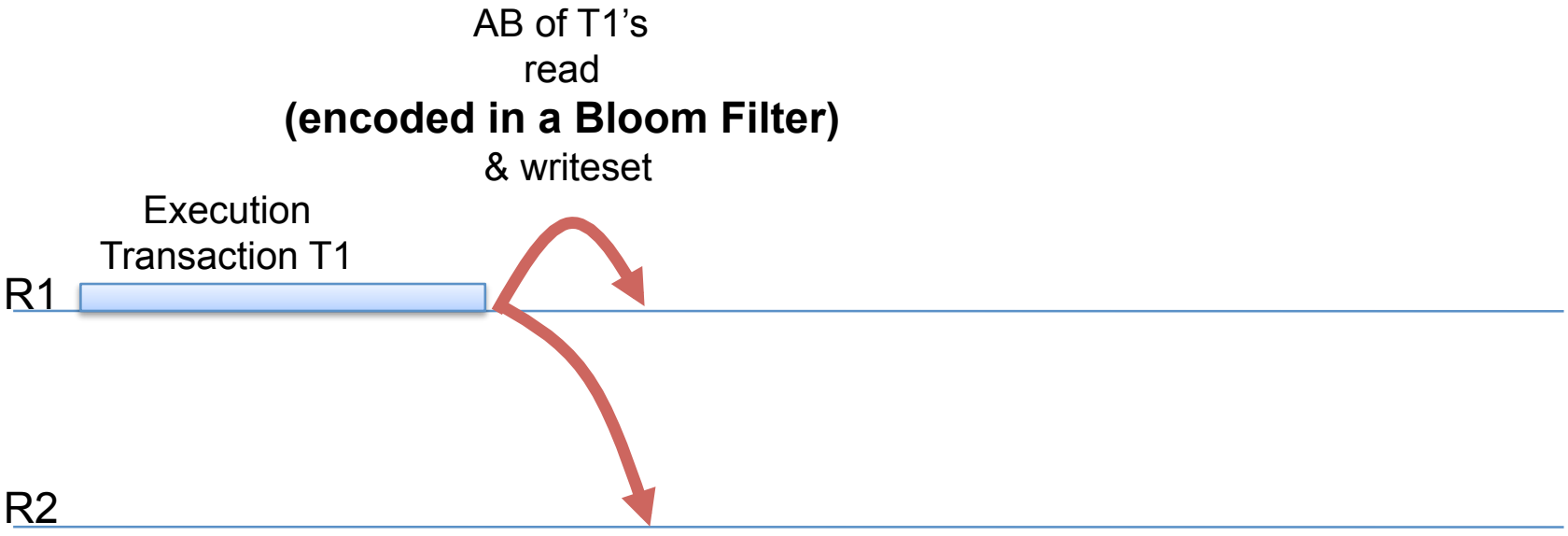


Figure: Example of the execution of the Bloom Filter Certification Protocol.

Bloom Filter Certification

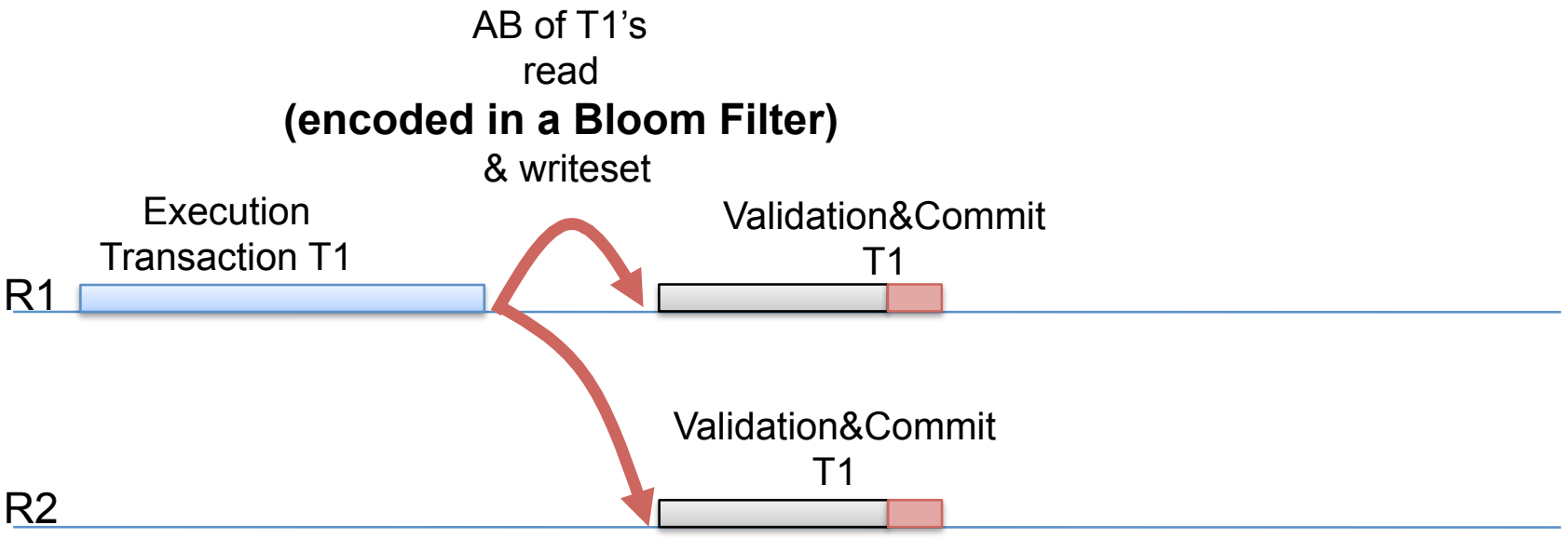


Figure: Example of the execution of the Bloom Filter Certification Protocol.

Bloom Filter Certification

Message

Read set encoded in a Bloom filter and write set

Validation

Test if any items written by concurrent transactions are in the Bloom filter

Pros:

- Reduce network traffic:
  - 1% false positive up to 30x message compression

Cons:

- False positives
  - additional (deterministic) aborts

Bloom Filter Certification

Message

Read set encoded in a Bloom filter and write set

Validation

Test if any items written by concurrent transactions are in the Bloom filter

Pros:

- Reduce network traffic:
  - 1% false positive up to 30x message compression

Cons:

- False positives
  - additional (deterministic) aborts

Bloom Filter Certification

Message

Read set encoded in a Bloom filter and write set

Validation

Test if any items written by concurrent transactions are in the Bloom filter

Pros:

- Reduce network traffic:
  - 1% false positive up to 30x message compression

Cons:

- False positives
  - additional (deterministic) aborts

Voting

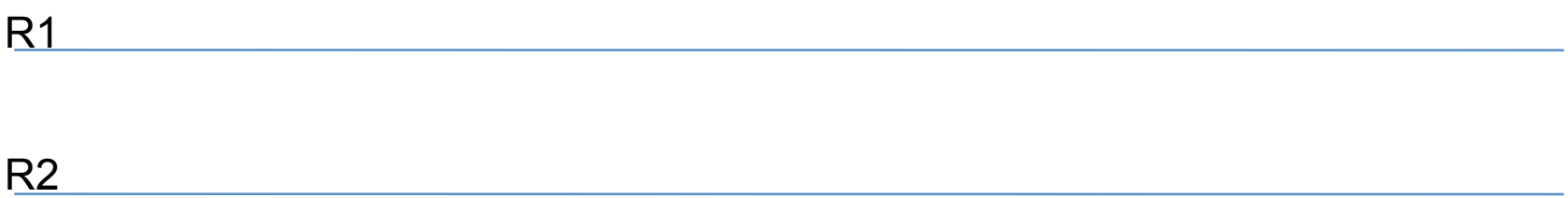


Figure: Example of the execution of the Voting Protocol.

Voting

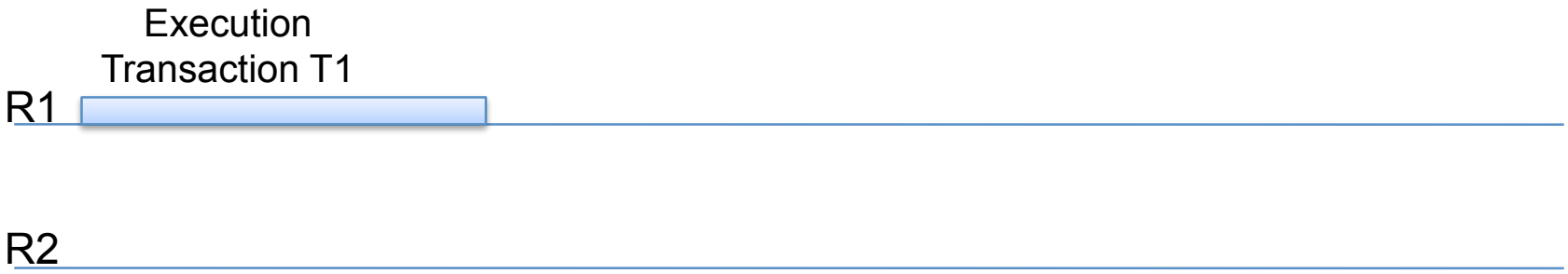


Figure: Example of the execution of the Voting Protocol.

Voting

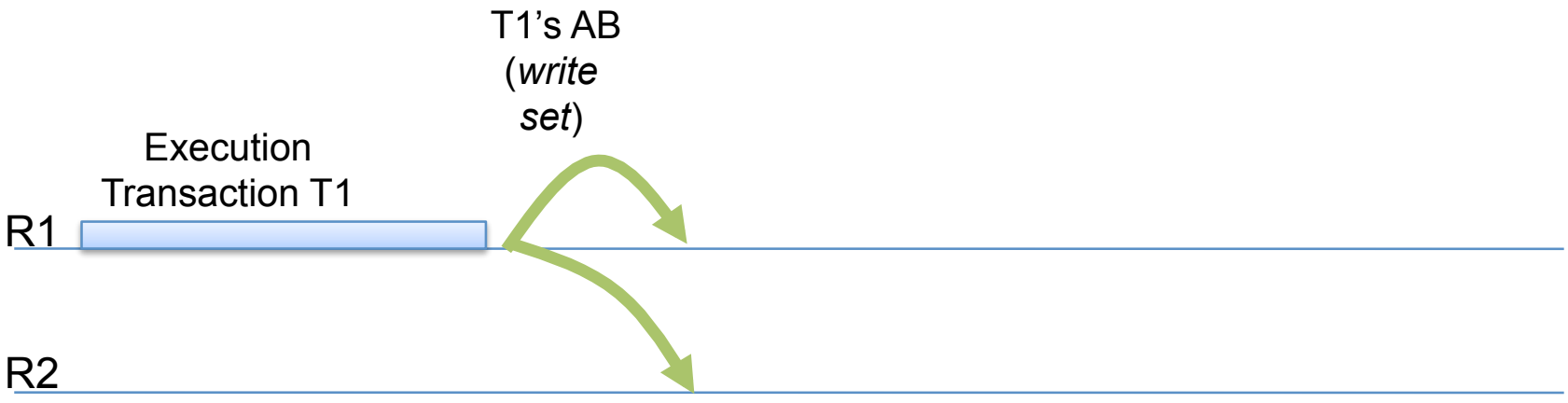


Figure: Example of the execution of the Voting Protocol.

Voting

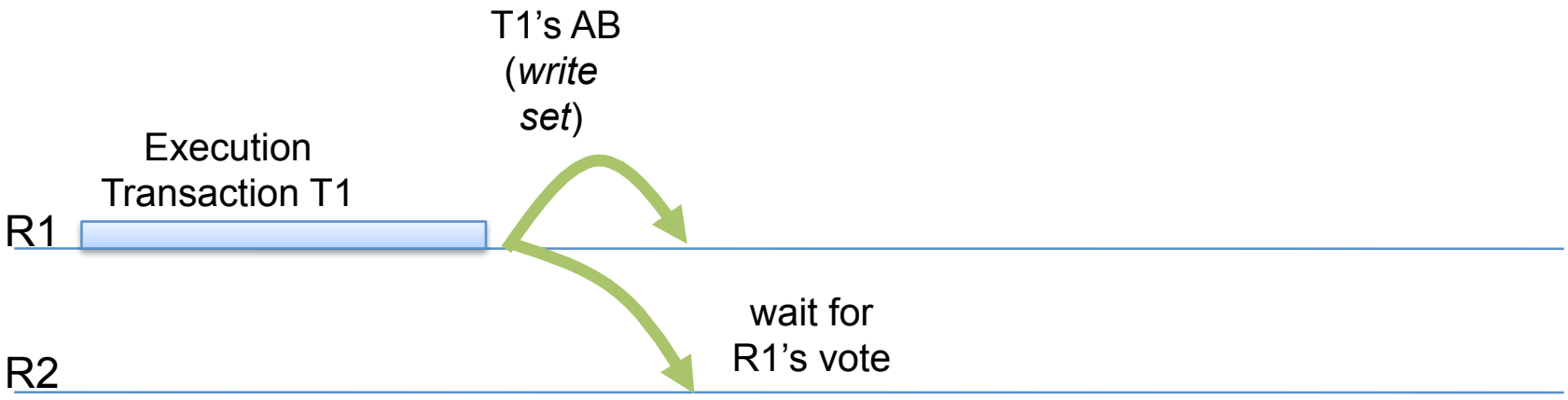


Figure: Example of the execution of the Voting Protocol.

Voting

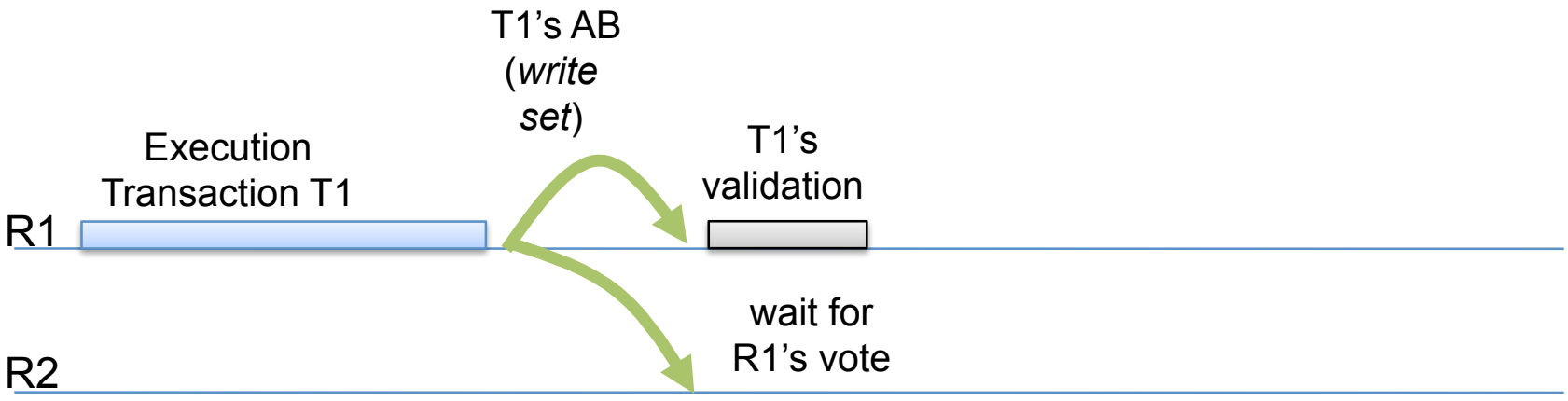


Figure: Example of the execution of the Voting Protocol.

Voting

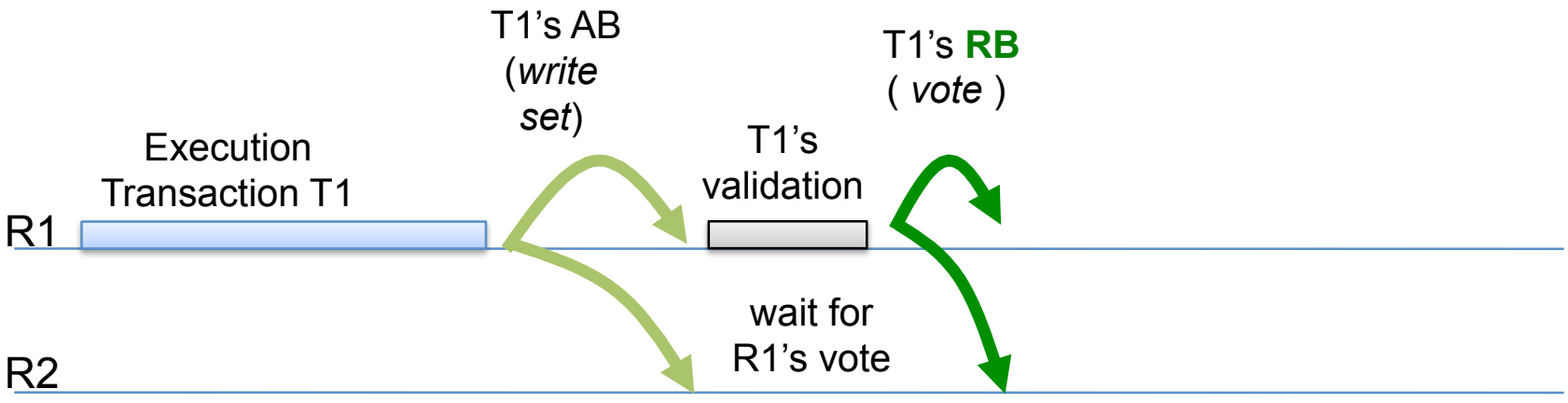


Figure: Example of the execution of the Voting Protocol.

Voting

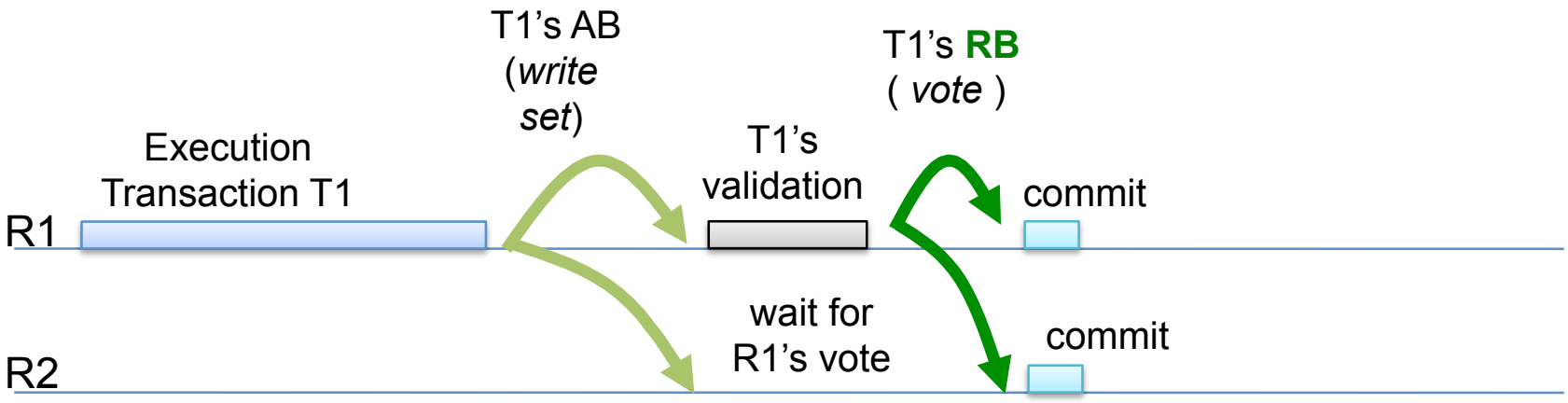


Figure: Example of the execution of the Voting Protocol.

Voting

Message

Write set

Validation

- Only the replica that executed the transaction can validate it
- When it receives this message
  - Checks if read set is valid
  - Sends the outcome to all replicas (Reliable Broadcast)

Pros:

- Short messages

Cons:

- Two communication steps

Voting

Message

Write set

Validation

- Only the replica that executed the transaction can validate it
- When it receives this message
  - Checks if read set is valid
  - Sends the outcome to all replicas (Reliable Broadcast)

Pros:

- Short messages

Cons:

- Two communication steps

Voting

Message

Write set

Validation

- Only the replica that executed the transaction can validate it
- When it receives this message
  - Checks if read set is valid
  - Sends the outcome to all replicas (Reliable Broadcast)

Pros:

- Short messages

Cons:

- Two communication steps

Throughput Comparison

Bank Benchmark

- Synthetic benchmark simulating transfers of funds
- Fixed read set sizes: 1, 1.000, 100.000
- No conflicts

Throughput varies greatly with the protocol used

Throughput Comparison

Bank Benchmark

- Synthetic benchmark simulating transfers of funds
- Fixed read set sizes: 1, 1.000, 100.000
- No conflicts

Throughput varies greatly with the protocol used

Throughput Comparison

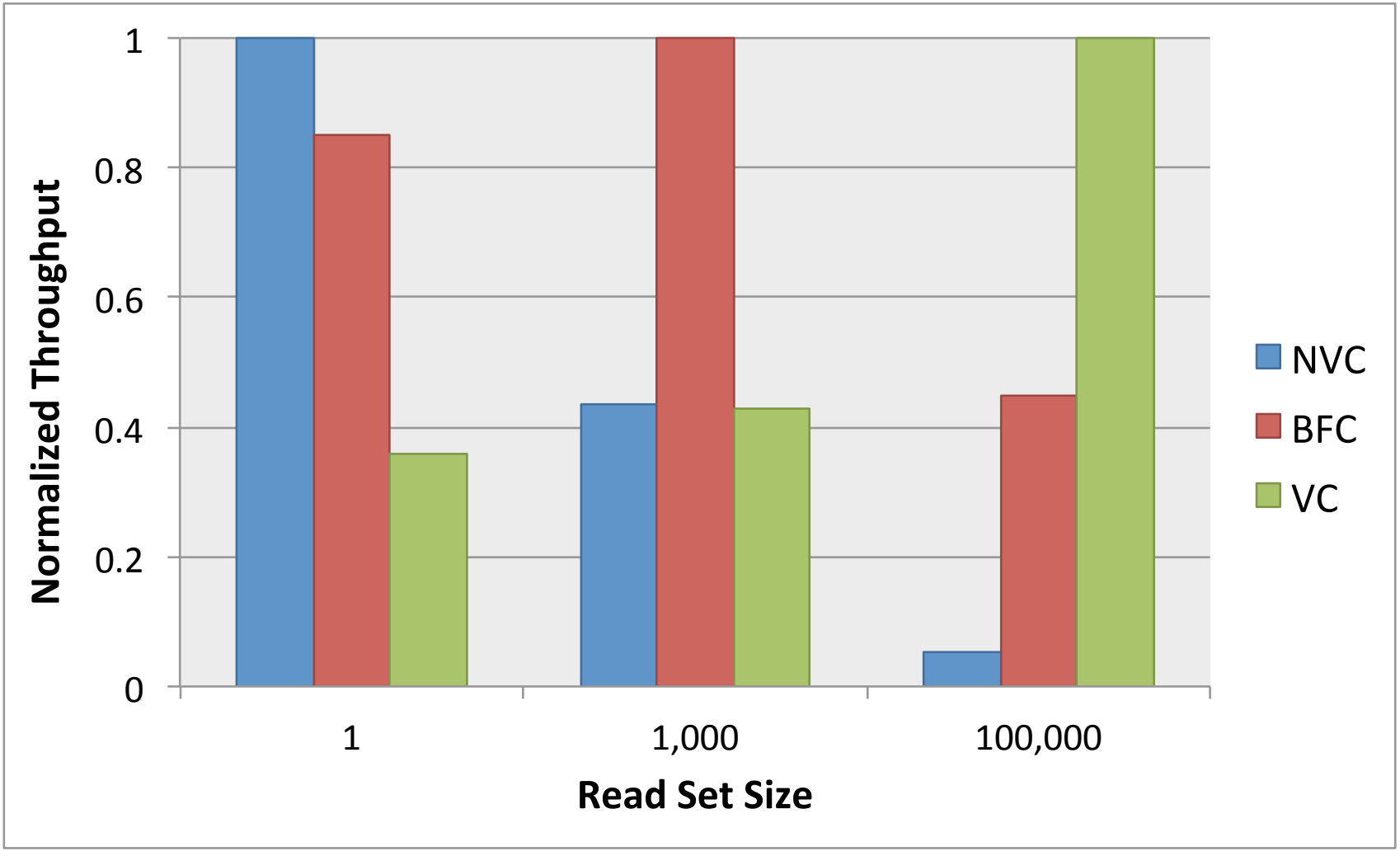


Figure: Throughput of three certification strategies with different read-set sizes.

Read Set Distribution

- Performance strongly depends on the size of the read sets
- Real applications exhibit very heterogeneous workloads

STMBench7

- Benchmark for Transactional Memories
- Complex benchmark with very heterogeneous transactions
- Operations that manipulate a graph with a significant number of objects strongly interconnected

Read Set Distribution

- Performance strongly depends on the size of the read sets
- Real applications exhibit very heterogeneous workloads

STMBench7

- Benchmark for Transactional Memories
- Complex benchmark with very heterogeneous transactions
- Operations that manipulate a graph with a significant number of objects strongly interconnected

Read Set Distribution

- Performance strongly depends on the size of the read sets
- Real applications exhibit very heterogeneous workloads

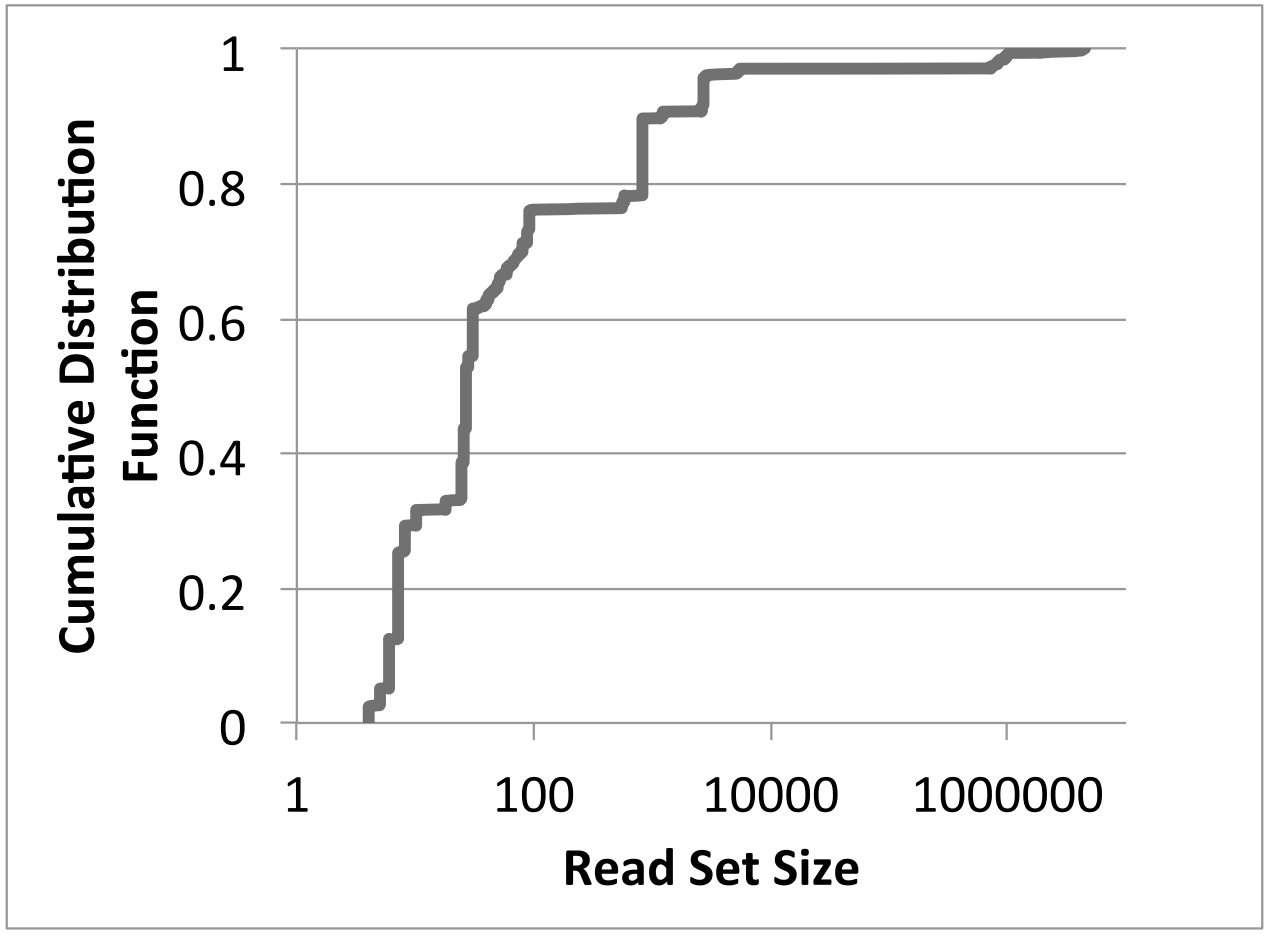


Figure: Distribution of transaction read set size in the STMBench7.

PolyCert

- Protocol choice heavily influences the system throughput
- PolyCert:
  - the co-existence of protocols
  - to predict the most appropriate per-transaction

Index

Architecture

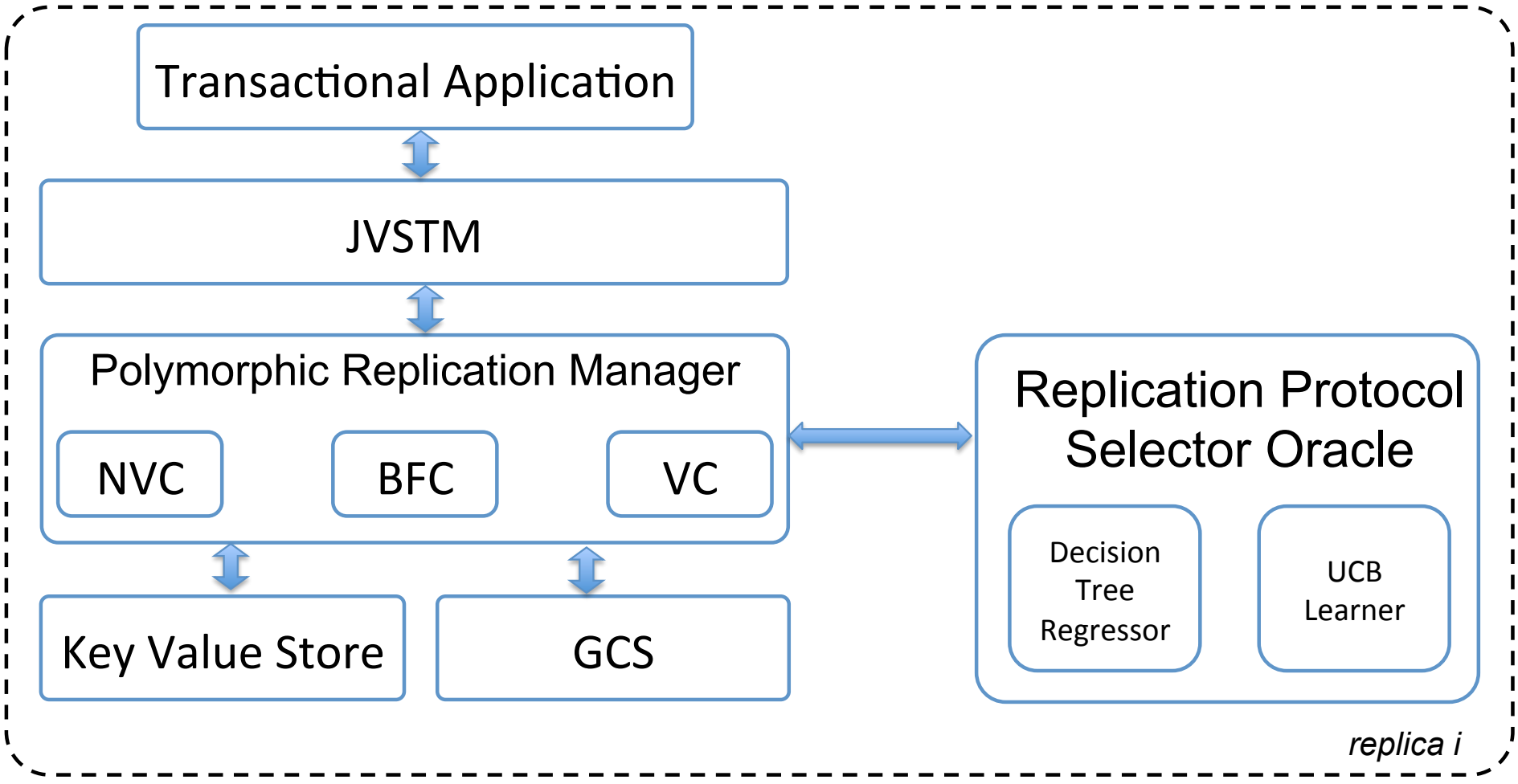
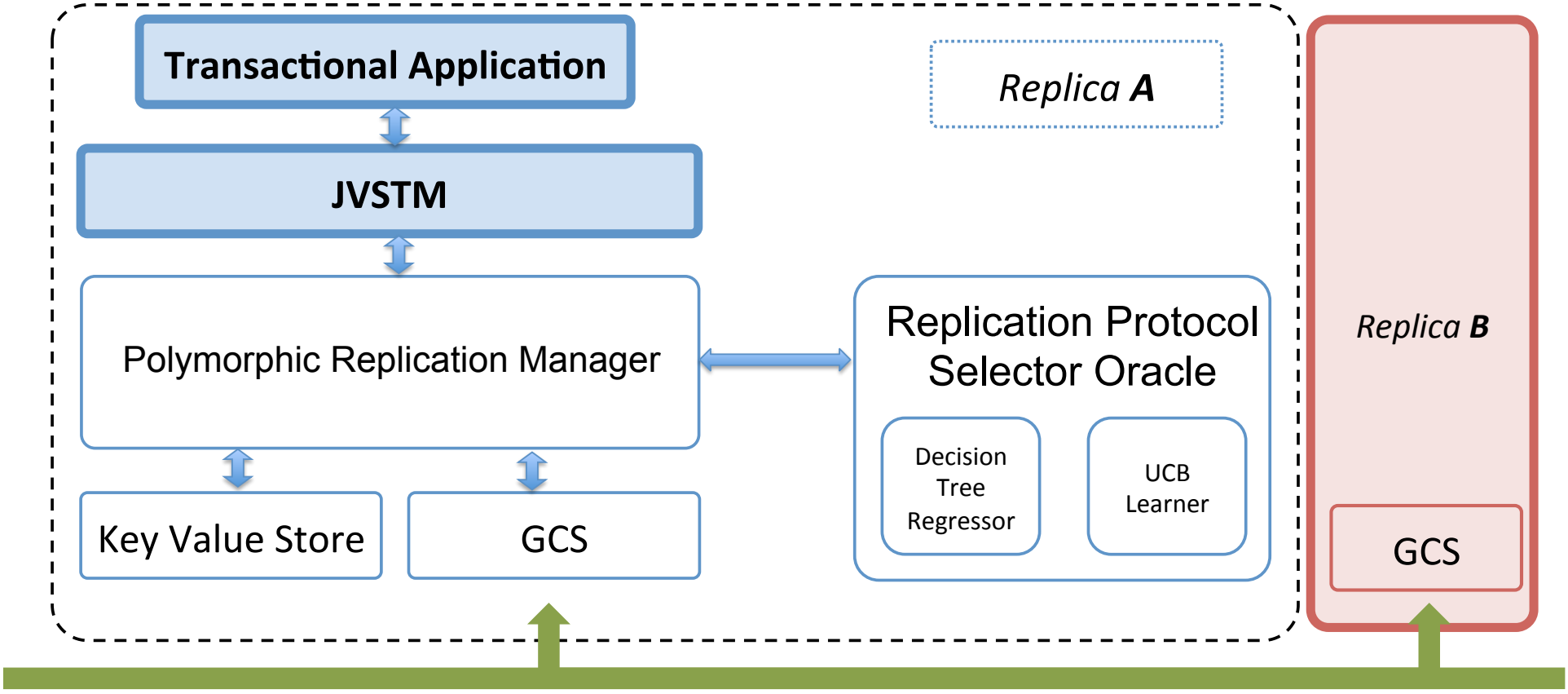
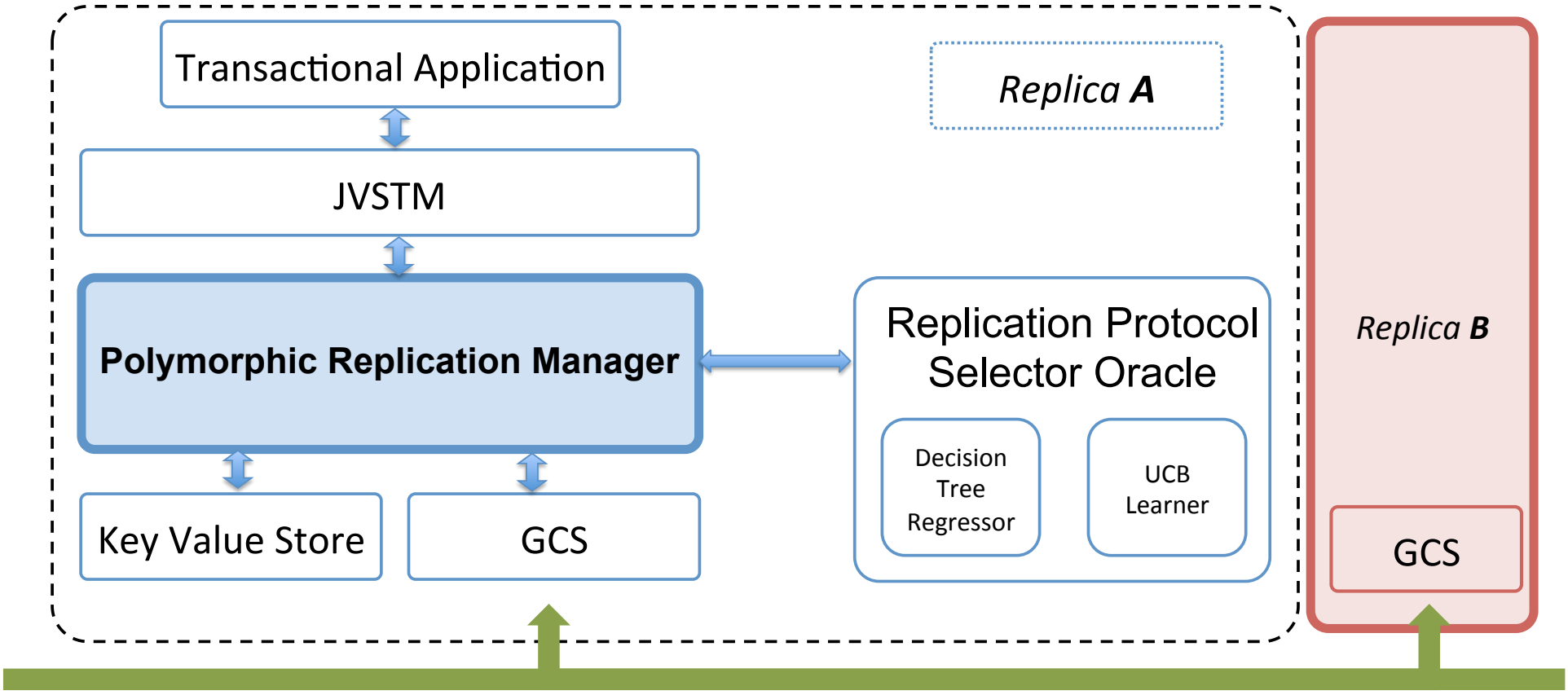


Figure: Architectural Overview (Single Node Perspective)

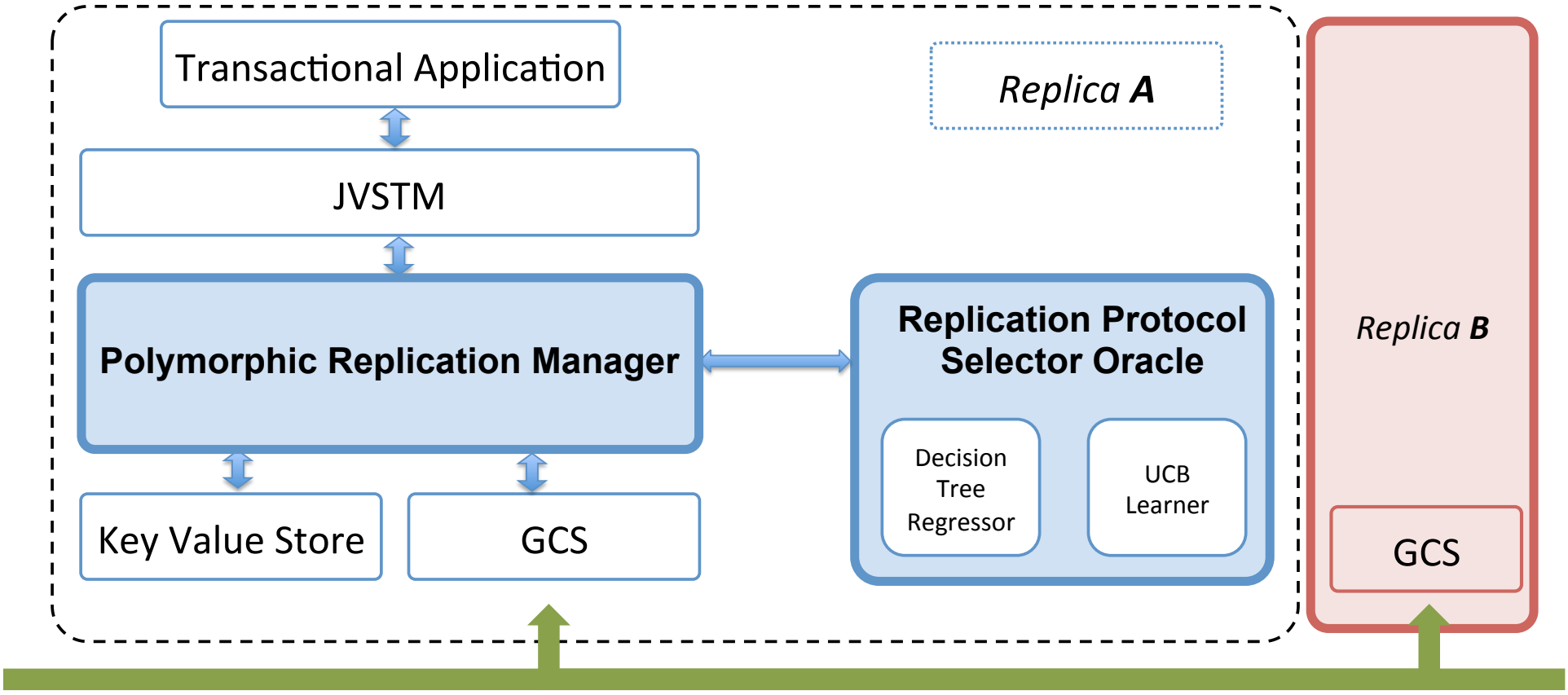
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

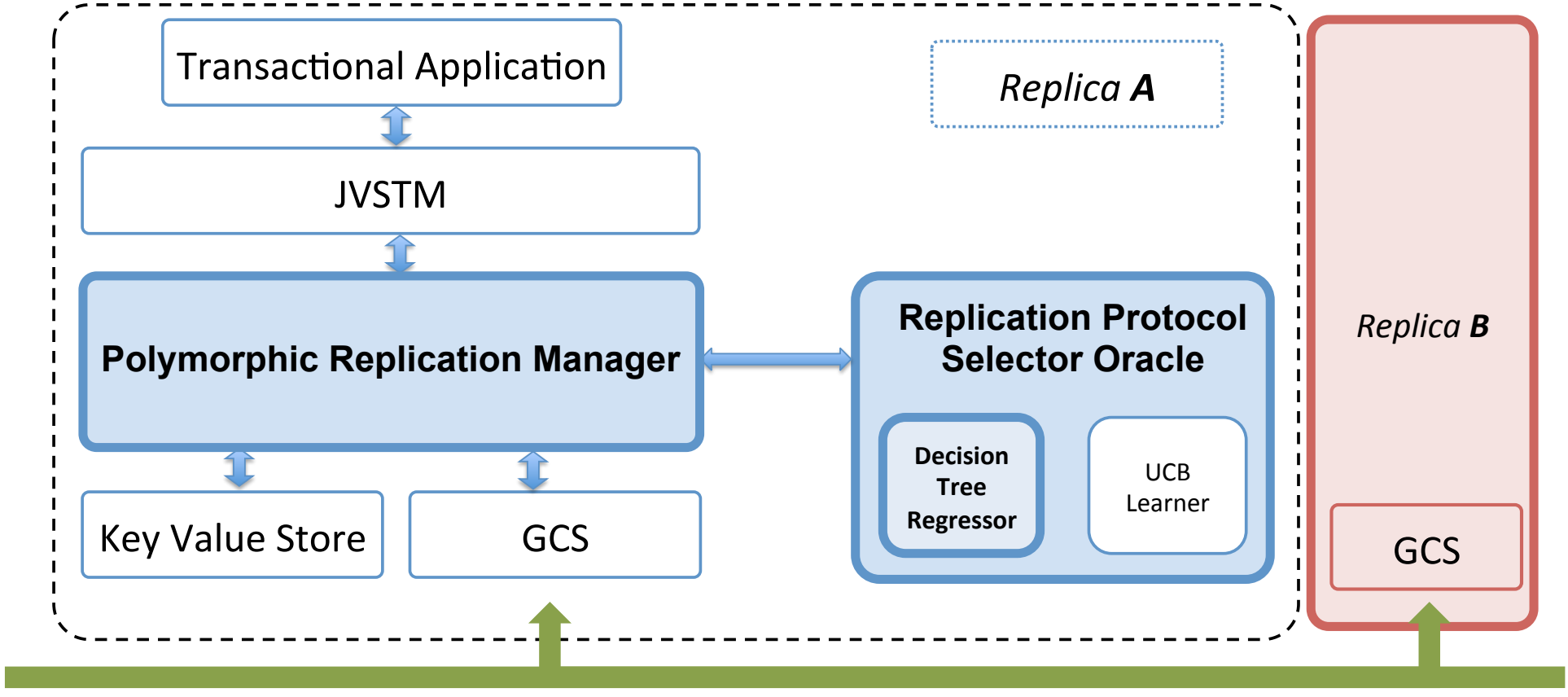
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

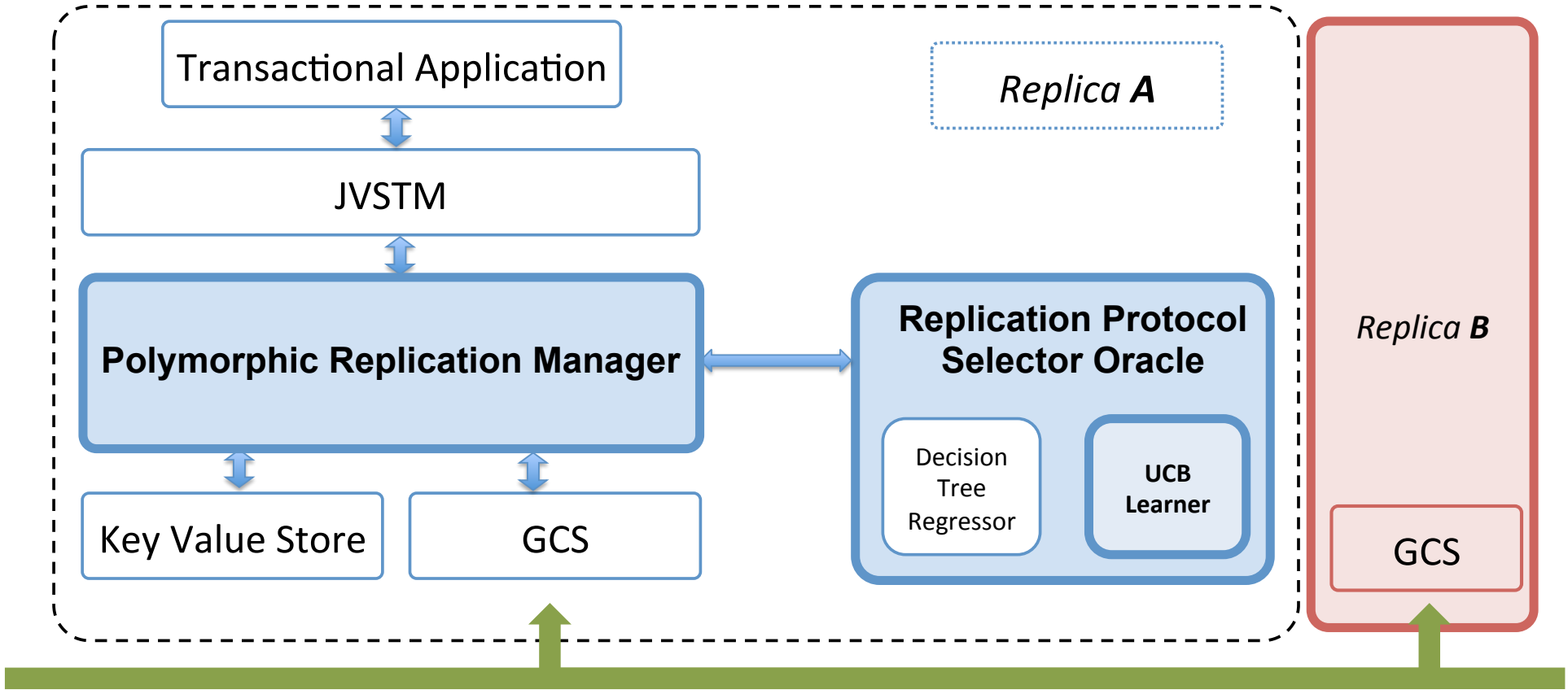
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

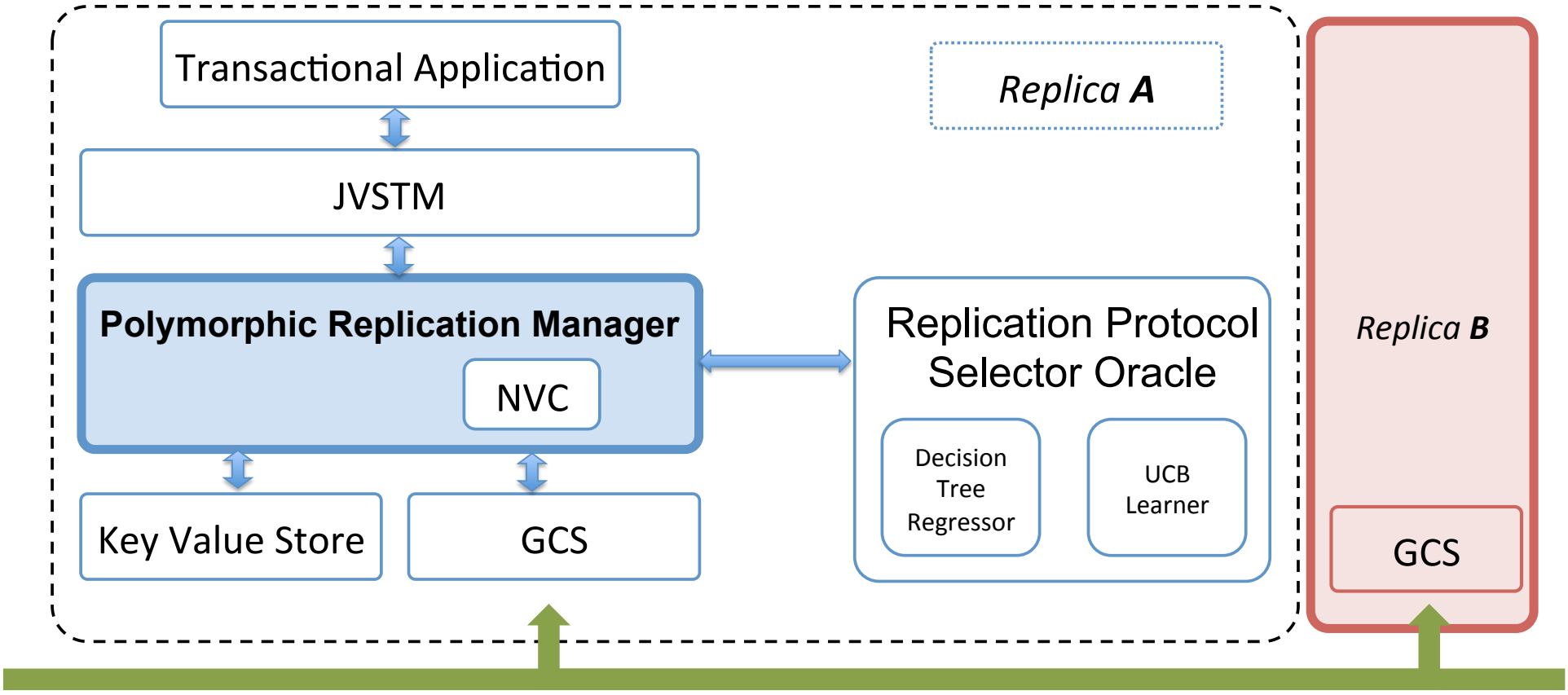
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

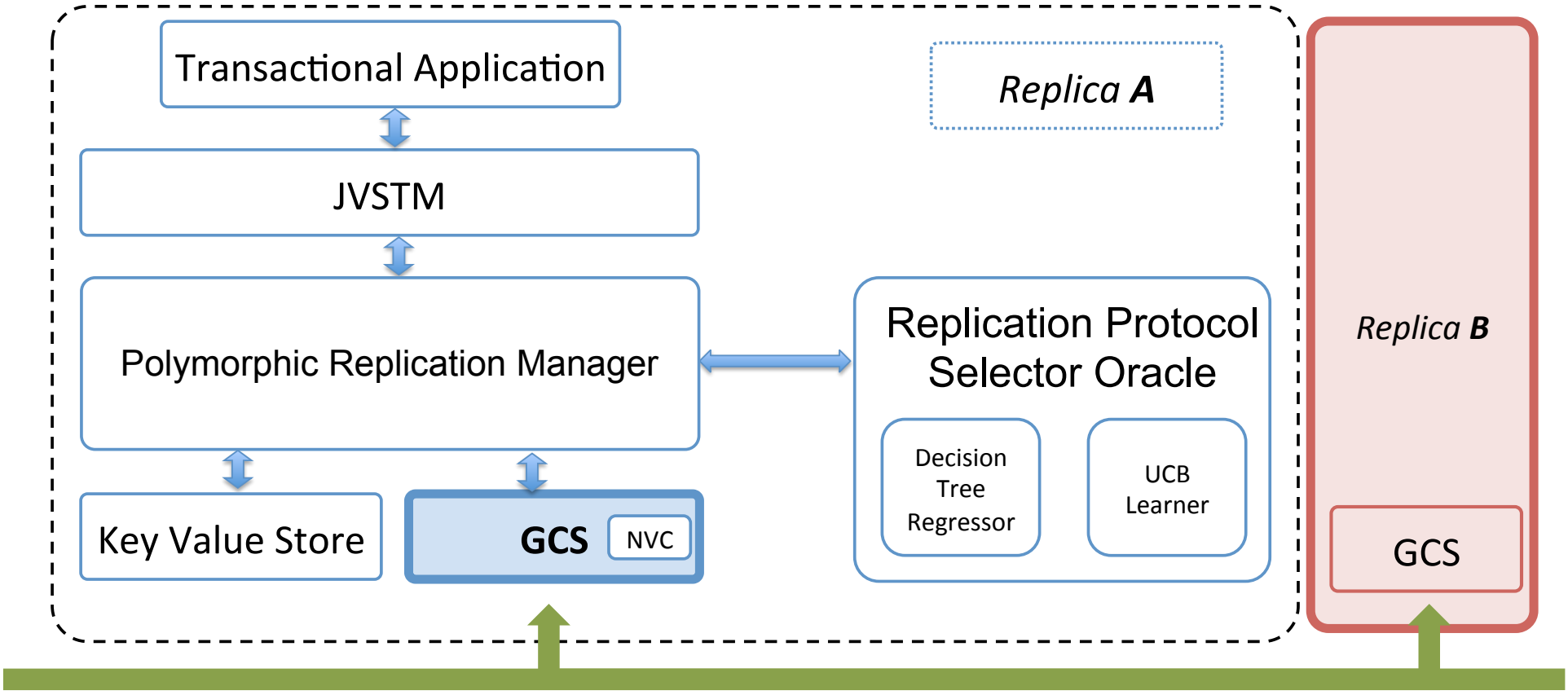
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

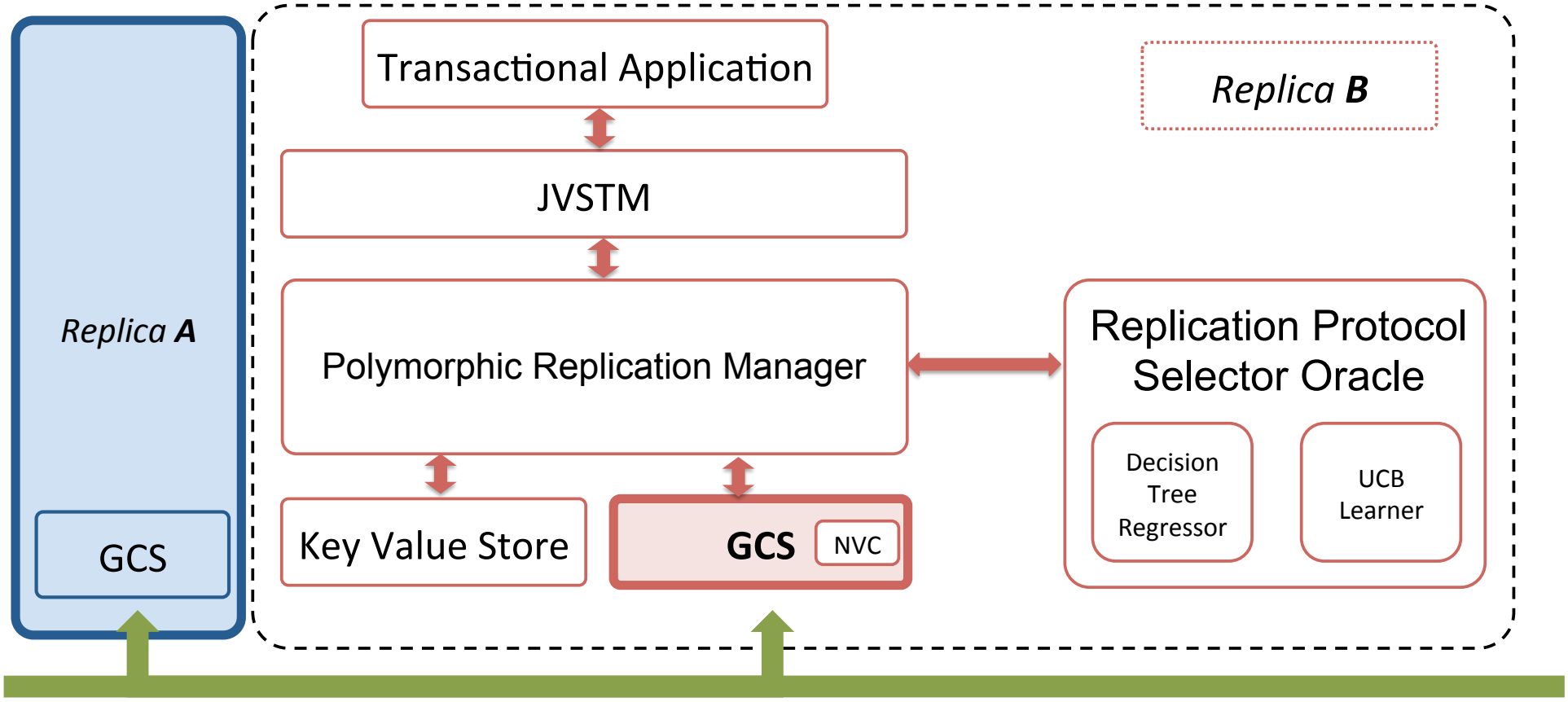
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
NVC
Key Value Store
GCS
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

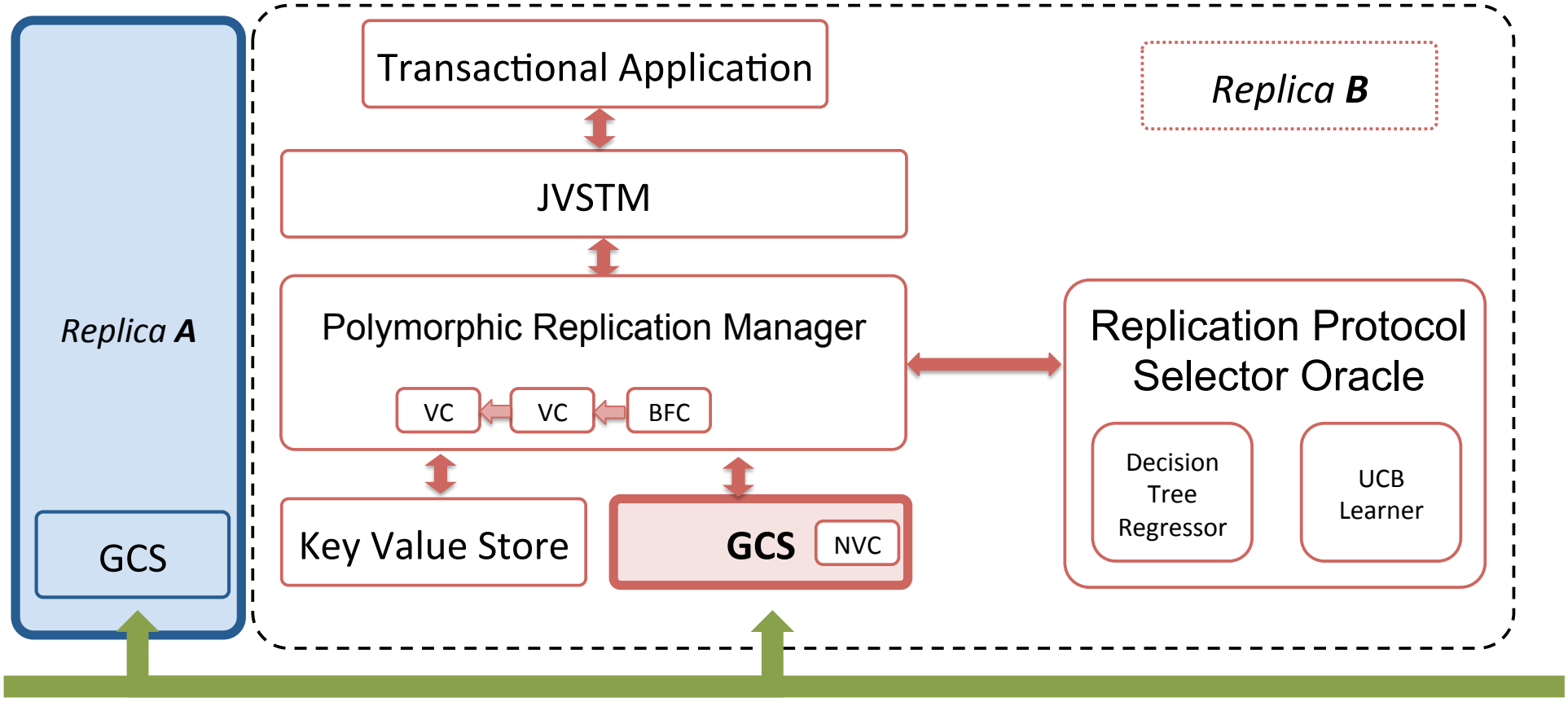
Protocol


Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
NVC
Replica A
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner
Replica B
GCS

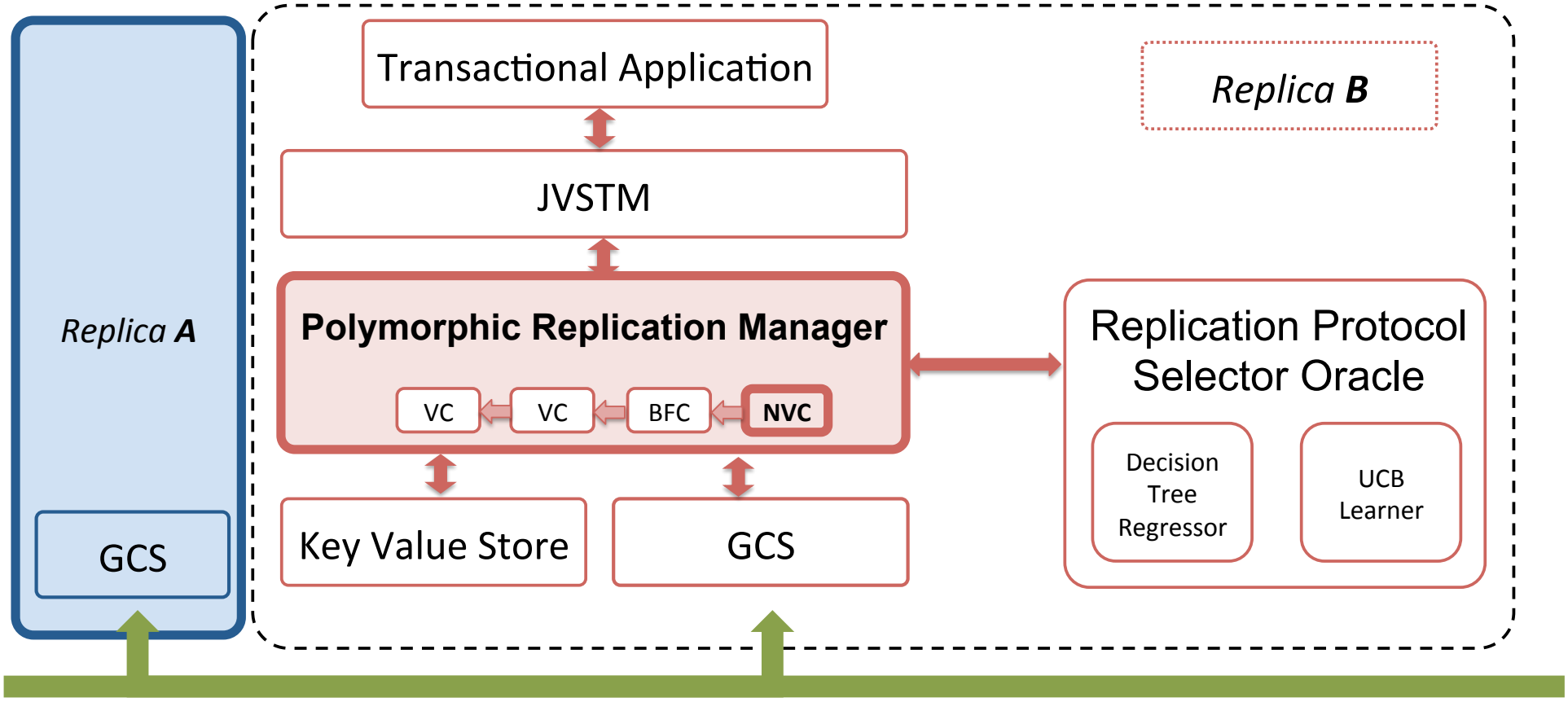
Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
NVC
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

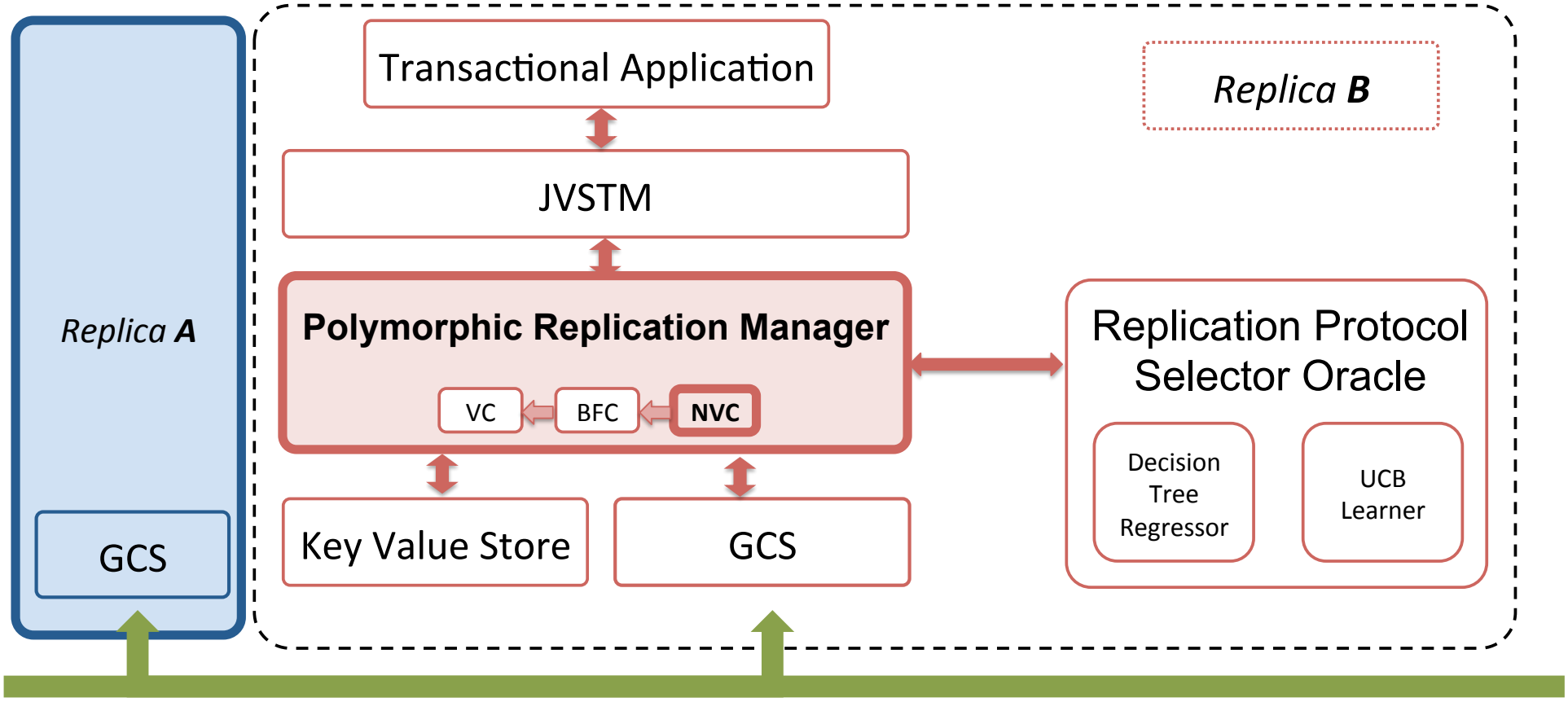
Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
VC
VC
BFC
Key Value Store
GCS
NVC
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

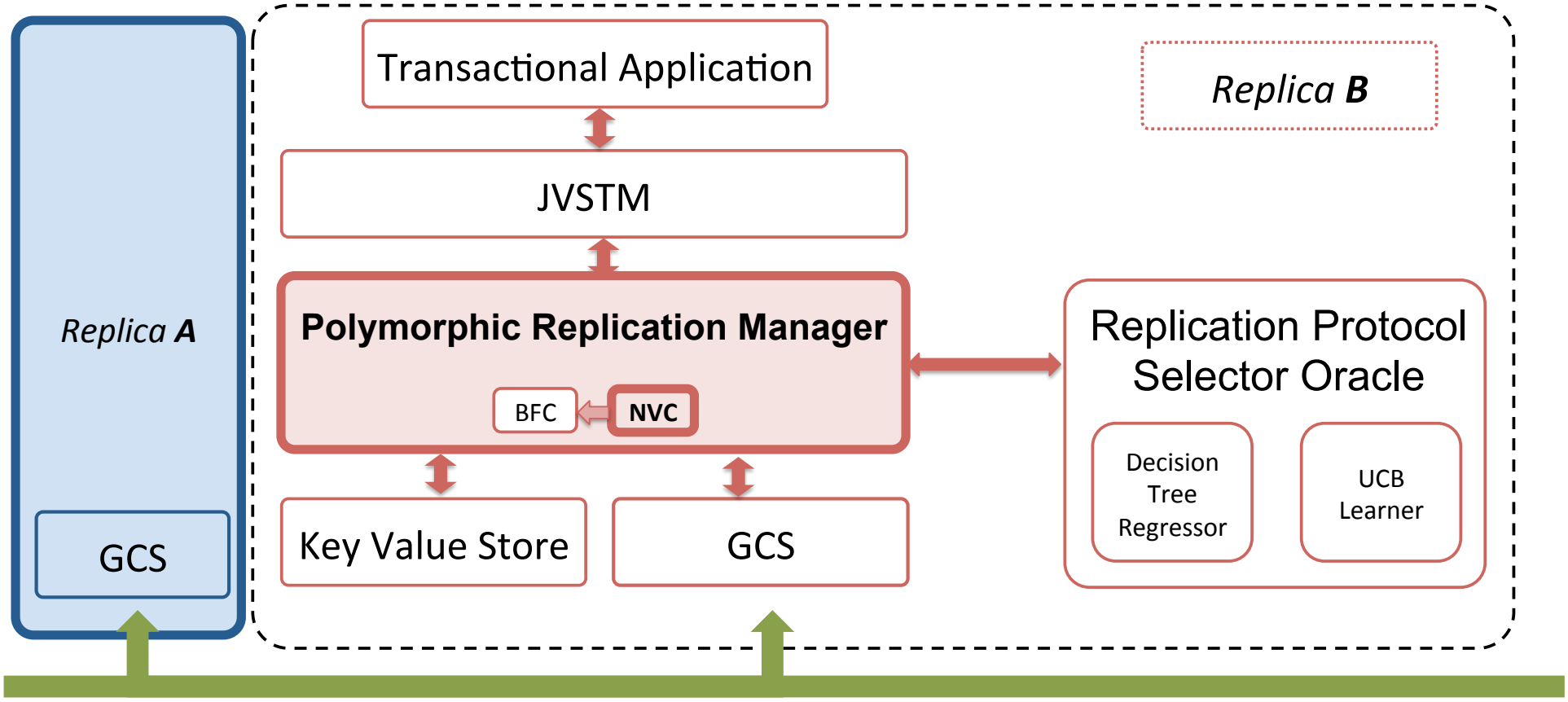
Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
VC
VC
BFC
NVC
Key Value Store
GCS
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

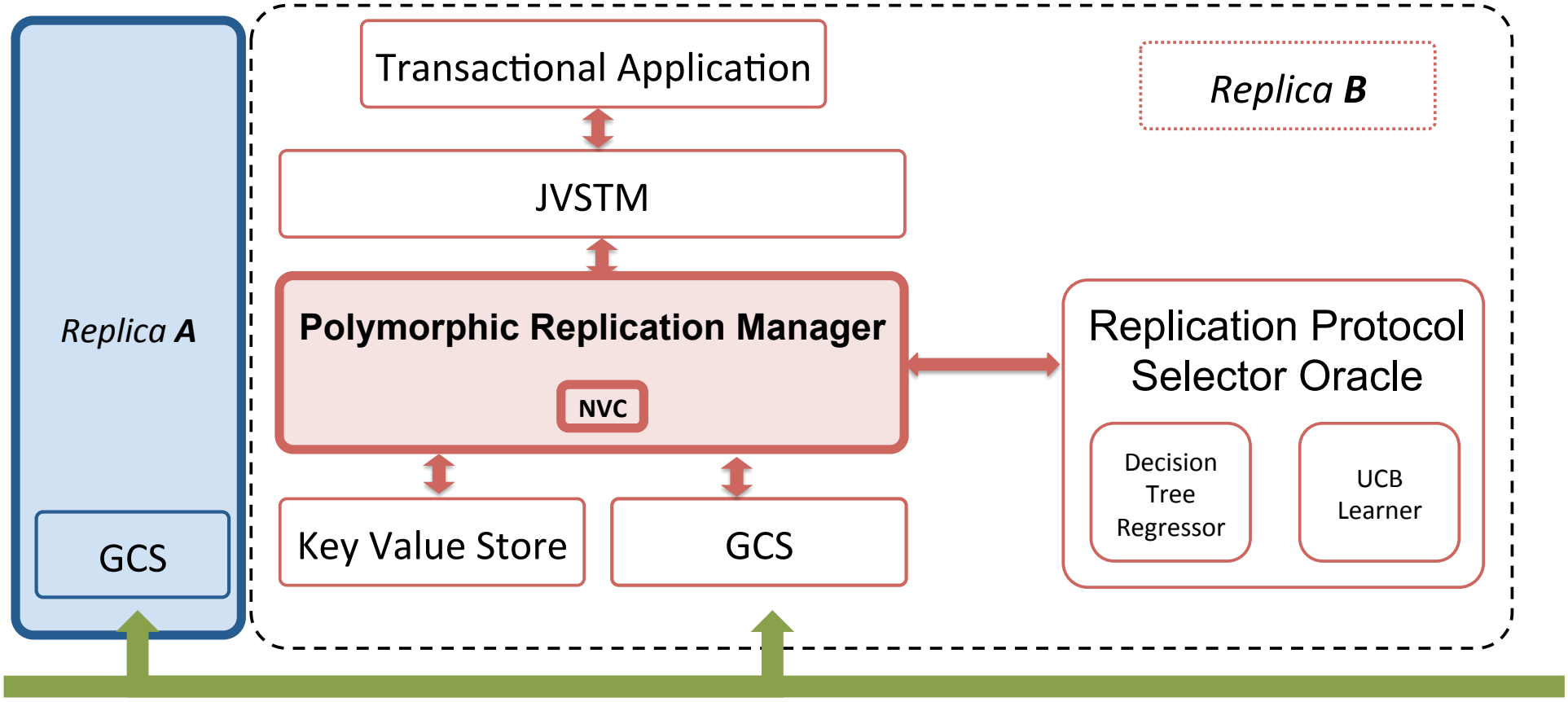
Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
VC
BFC
NVC
Key Value Store
GCS
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

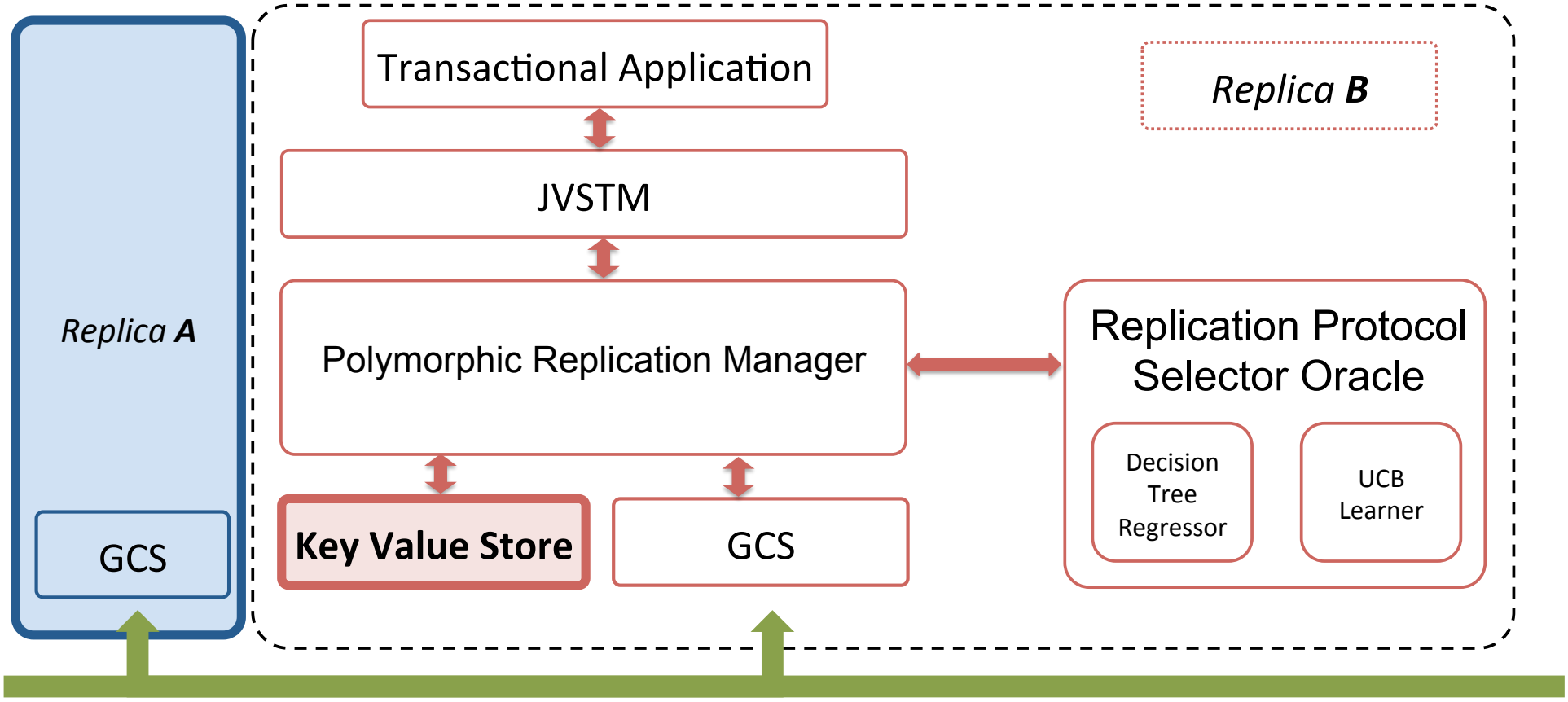
Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
BFC
NVC
Key Value Store
GCS
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
NVC
Key Value Store
GCS
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

Protocol


Replica A
GCS
Transactional Application
JVSTM
Polymorphic Replication Manager
Key Value Store
GCS
Replica B
Replication Protocol Selector Oracle
Decision Tree Regressor
UCB Learner

Index

Replication Protocol Selector Oracle

Two implementations:

- Off-line Machine Learning Technique: Decision Trees
  - **Pros:** No learning during the execution of the system
  - **Cons:** Computational intensive training phase
- On-line Reinforcement Learning: UCB
  - **Pros:** Adapts easily to change
  - **Cons:** Needs to learn while the system is running

Replication Protocol Selector Oracle

Two implementations:

- **Off-line Machine Learning Technique: Decision Trees**
  - **Pros:** No learning during the execution of the system
  - **Cons:** Computational intensive training phase
- On-line Reinforcement Learning: UCB
  - Pros: Adapts easily to change
  - Cons: Needs to learn while the system is running

Off-line Machine Learning Techniques

For each transaction:

- Predict size of AB message m for the various certification schemes
- Forecast AB latency for each message size.
  - Regression Decision trees
- Forecast the time for marshalling and validation for each protocol
  - BFC: forecast the time to build and populate the Bloom filter

Choose the protocol with the smallest commit latency

Off-line Machine Learning Techniques

For each transaction:

- Predict size of AB message m for the various certification schemes
- Forecast AB latency for each message size.
  - Regression Decision trees
- Forecast the time for marshalling and validation for each protocol
  - BFC: forecast the time to build and populate the Bloom filter

Choose the protocol with the smallest commit latency

Off-line Machine Learning Techniques

- Uses up to 53 monitored system attributes:
  - CPU
  - Memory
  - Network
  - Time-series
- Requires computational intensive training phase

Replication Protocol Selector Oracle

Two implementations:

- Off-line Machine Learning Technique: Decision Trees
  - Pros: No learning during the execution of the system
  - Cons: Computational intensive training phase
- **On-line Reinforcement Learning: UCB**
  - **Pros:** Adapts easily to change
  - **Cons:** Needs to learn while the system is running

On-line Reinforcement Learning

Each replica builds on-line expectations on the rewards of each protocol:

- no assumption on the rewards' distributions
- updates the knowledge of the oracle while the system is running

Tackles the exploration-exploitation dilemma:

- did I test this option sufficiently in this scenario?

On-line Reinforcement Learning

Each replica builds on-line expectations on the rewards of each protocol:

- no assumption on the rewards' distributions
- updates the knowledge of the oracle while the system is running

Tackles the exploration-exploitation dilemma:

- did I test this option sufficiently in this scenario?

On-line Reinforcement Learning

Distinguishes workload scenario solely based on read-set's size

- exponential discretization intervals to minimize training time

Optimization: DistUCB

Replicas exchange statistical information periodically to boost learning

On-line Reinforcement Learning

Distinguishes workload scenario solely based on read-set's size

- exponential discretization intervals to minimize training time

Optimization: DistUCB

Replicas exchange statistical information periodically to boost learning

Index

Evaluation

Four benchmarks:

- Bank Benchmark - 1
- Bank Benchmark - 1000
- Bank Benchmark - 100000
- STMBench7

Bank Benchmark - 1

- Compare the throughput of the certification protocols and PolyCert
- Read set: **1** item

PolyCert achieves a performance very close to the best static certification protocol

Bank Benchmark - 1

- Compare the throughput of the certification protocols and PolyCert
- Read set: **1** item

PolyCert achieves a performance very close to the best static certification protocol

Bank Benchmark - 1

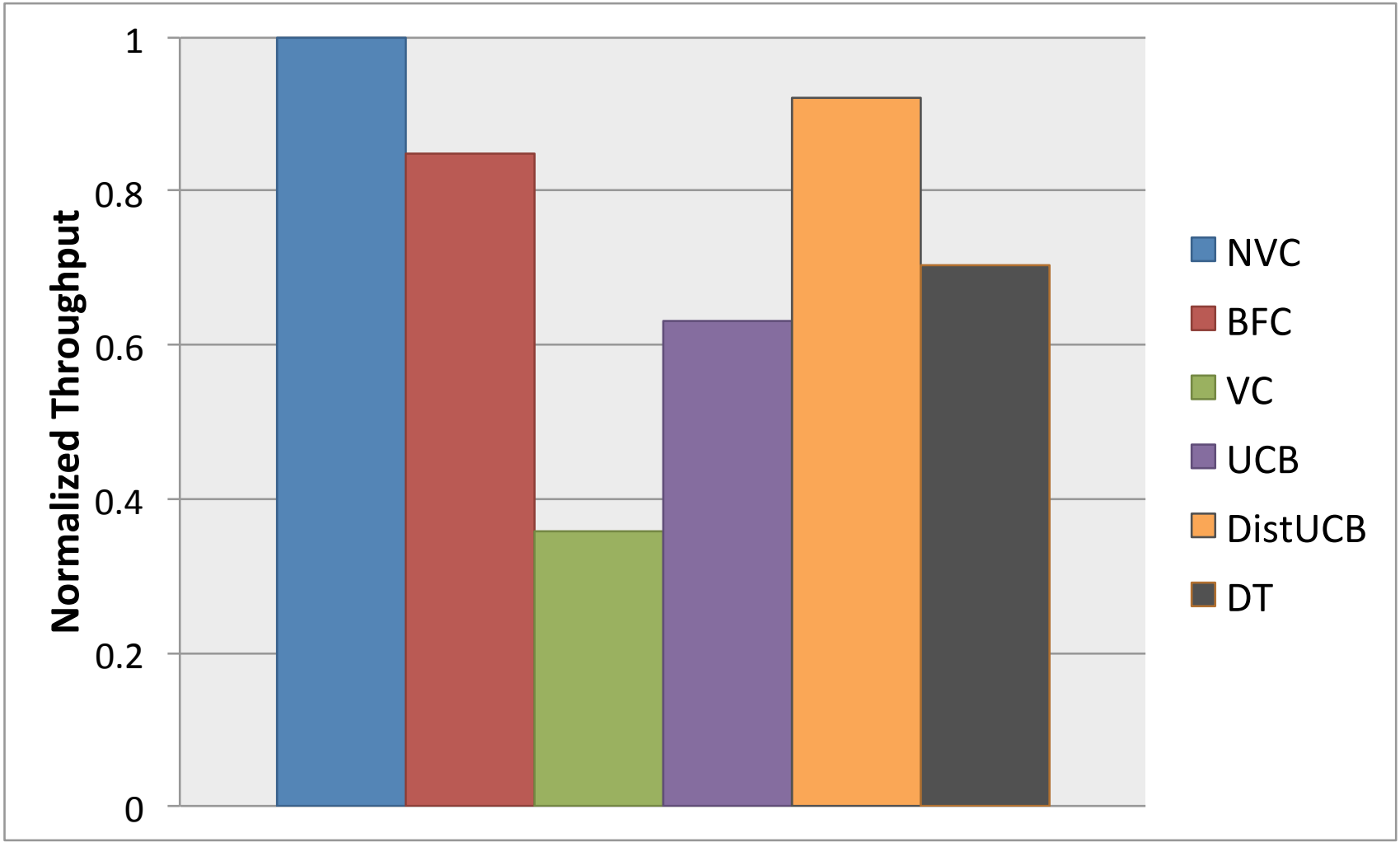


Figure: Normalized throughput of PolyCert and static protocols

Bank Benchmark - 1000

- Compare the throughput of the certification protocols and PolyCert
- Read set: **1000** items

PolyCert achieves a performance very close to the best static certification protocol

Bank Benchmark - 1000

- Compare the throughput of the certification protocols and PolyCert
- Read set: **1000** items

PolyCert achieves a performance very close to the best static certification protocol

Bank Benchmark - 1000

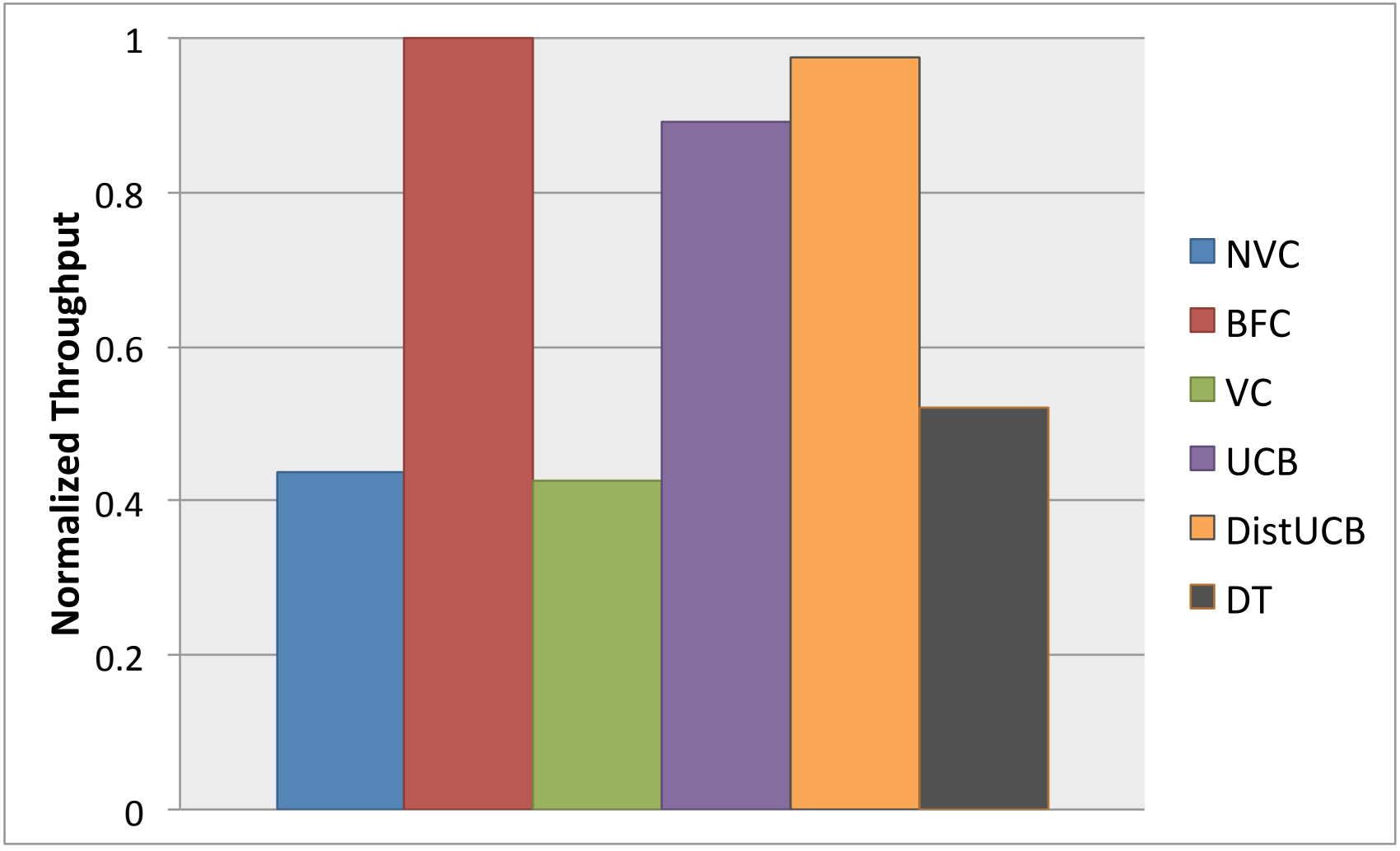


Figure: Normalized throughput of PolyCert and static protocols

Bank Benchmark - 100000

- Compare the throughput of the certification protocols and PolyCert
- Read set: **100.000** items

PolyCert achieves a performance very close to the best static certification protocol

Bank Benchmark - 100000

- Compare the throughput of the certification protocols and PolyCert
- Read set: **100.000** items

PolyCert achieves a performance very close to the best static certification protocol

Bank Benchmark - 100000

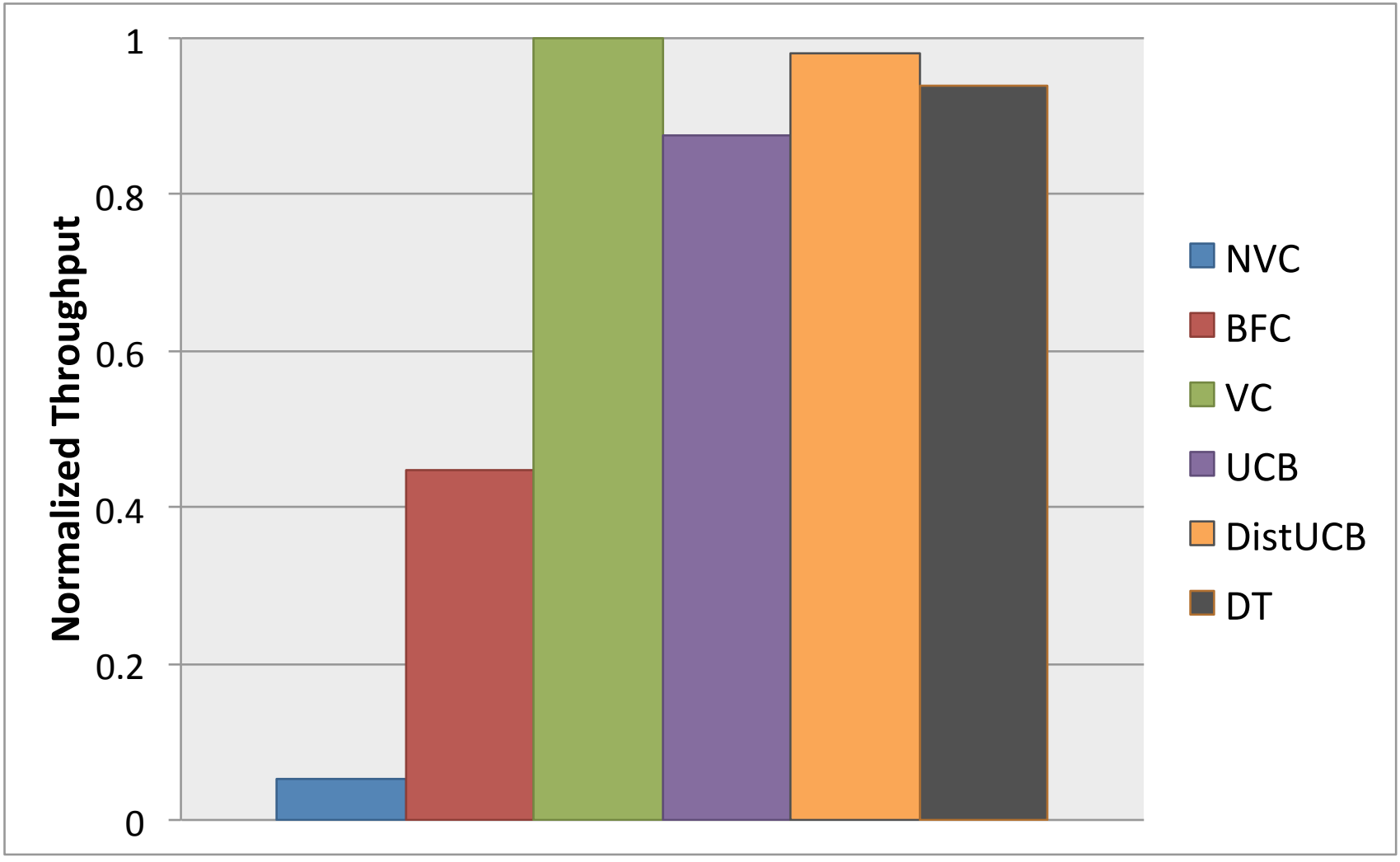


Figure: Normalized throughput of PolyCert and static protocols

Bank Benchmark - Highlight

- Compare the evolution of the throughput of UCB and Distributed UCB when the workload changes
- Read set: 100000 items

Distributed UCB converges faster than UCB

Bank Benchmark - Highlight

- Compare the evolution of the throughput of UCB and Distributed UCB when the workload changes
- Read set: 100000 items

Distributed UCB converges faster than UCB

Bank Benchmark - Highlight

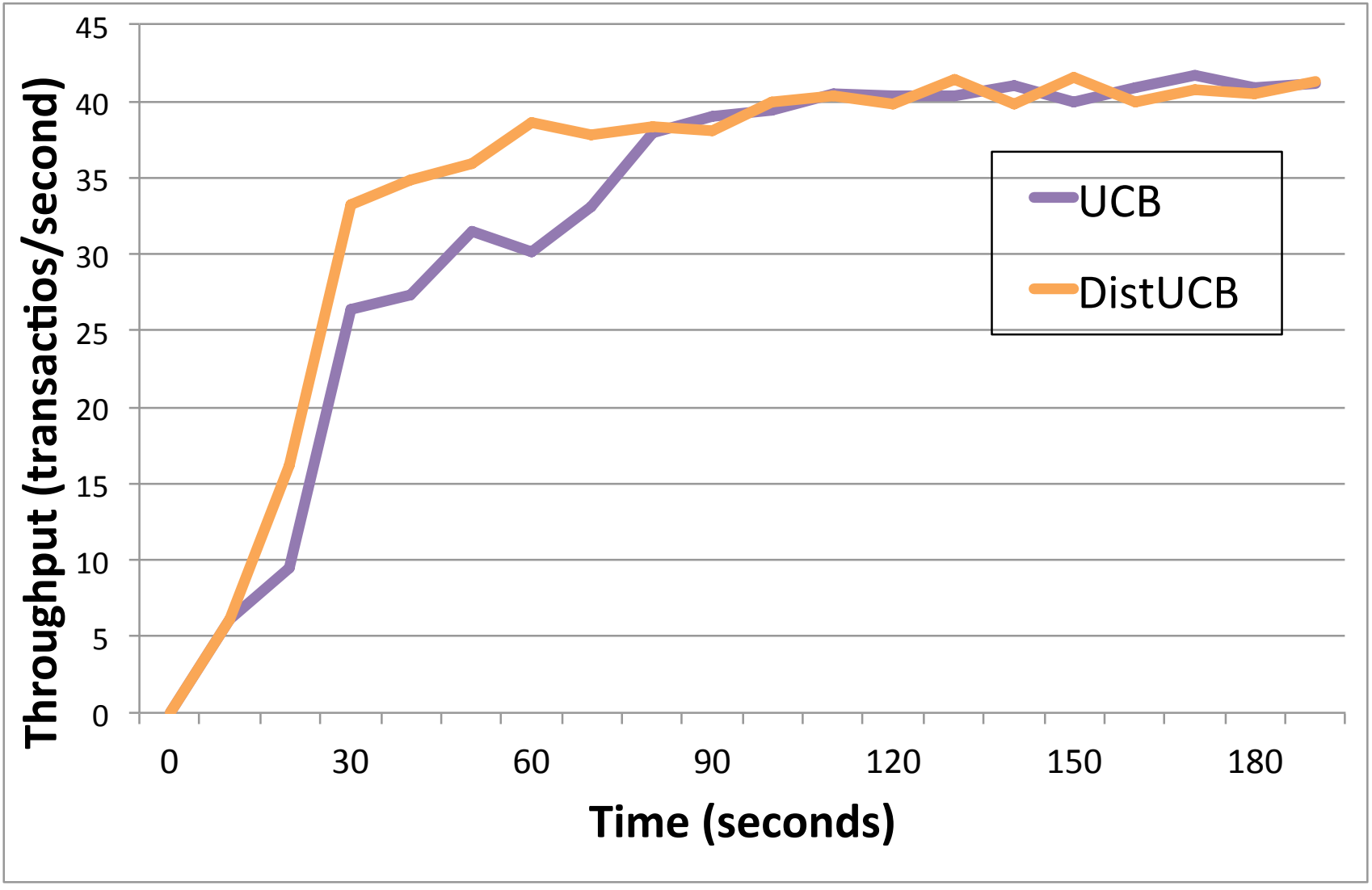


Figure: Evolution of throughput over time with UCB and DIST-UCB

STMBench7

- Compare the throughput of the best performing static certification protocol with PolyCert

PolyCert's throughput is higher than the best performing static certification protocol

STMBench7

- Compare the throughput of the best performing static certification protocol with PolyCert

PolyCert's throughput is higher than the best performing static certification protocol

STMBench7

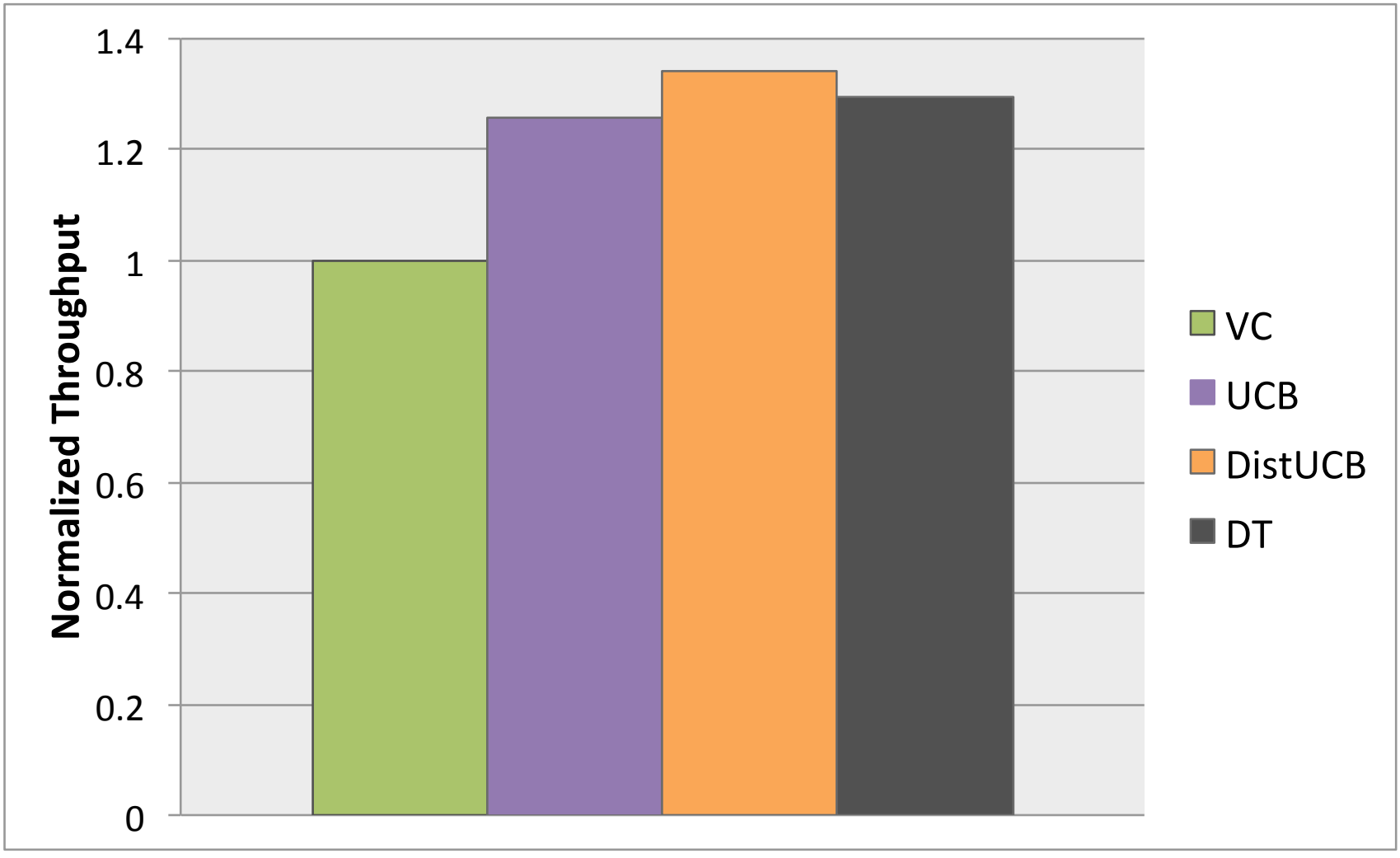


Figure: Normalized throughput of the adaptive and VC protocols

Index

Summary

- PolyCert: Polymorphic Self-Optimizing Certification
- Allows the co-existence of multiple certification protocols
- Machine-learning techniques to determine the best certification strategy per transaction
- Logic associated with the on-line choice of the replication strategy encapsulated into a generic oracle
- Achieves speed-ups when compared to static protocols
- Increases the robustness of the replicated data platform

Thank You