Making Fast Databases

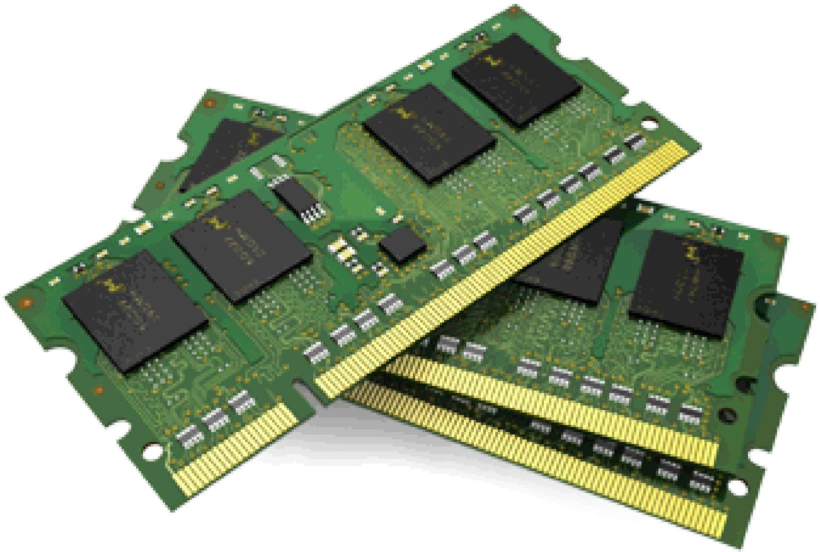
FASTER

@andy_pavlo

BROWN

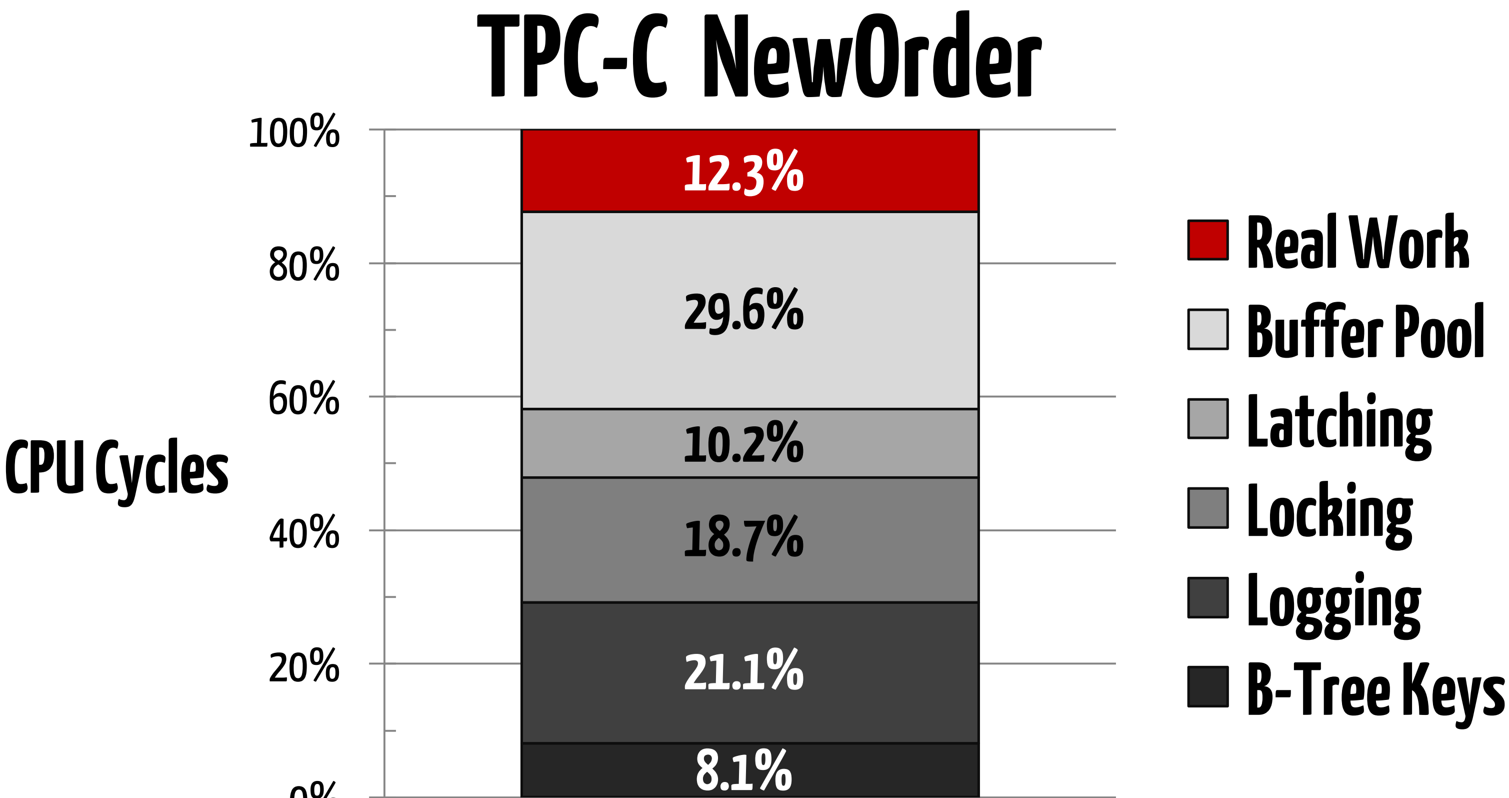
Yale University
Columbia University
April 2012


BLUEGRASS



Fast
+
Cheap

Legacy Systems



OLTP Through the Looking Glass, and What We Found There
SIGMOD 2008

OLTP Transactions

Fast

Repetitive

Small

H-Store

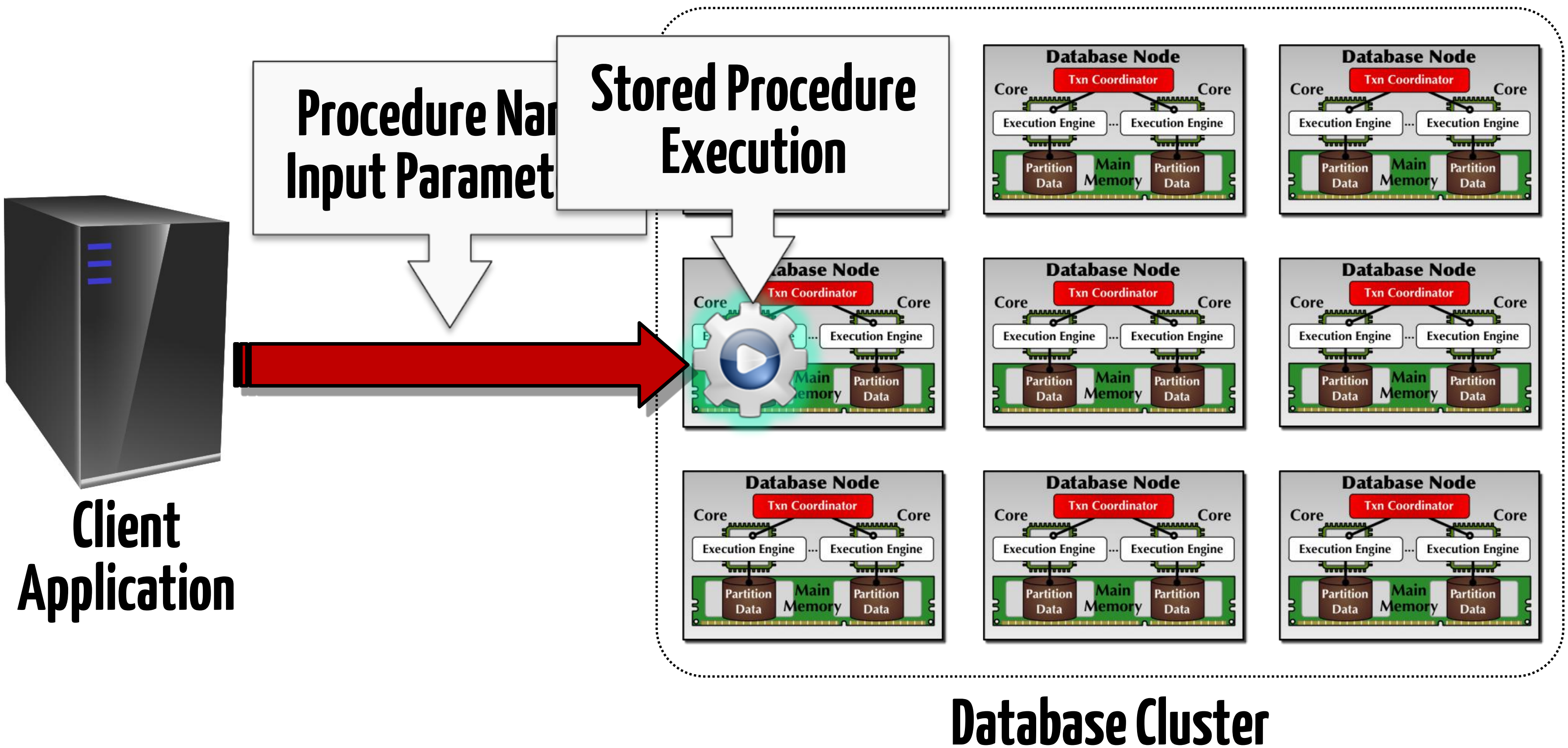
Main Memory • Parallel • Shared-Nothing Transaction Processing

H-Store: A High-Performance, Distributed Main Memory Transaction Processing System

VLDB vol. 1, issue 2, 2008

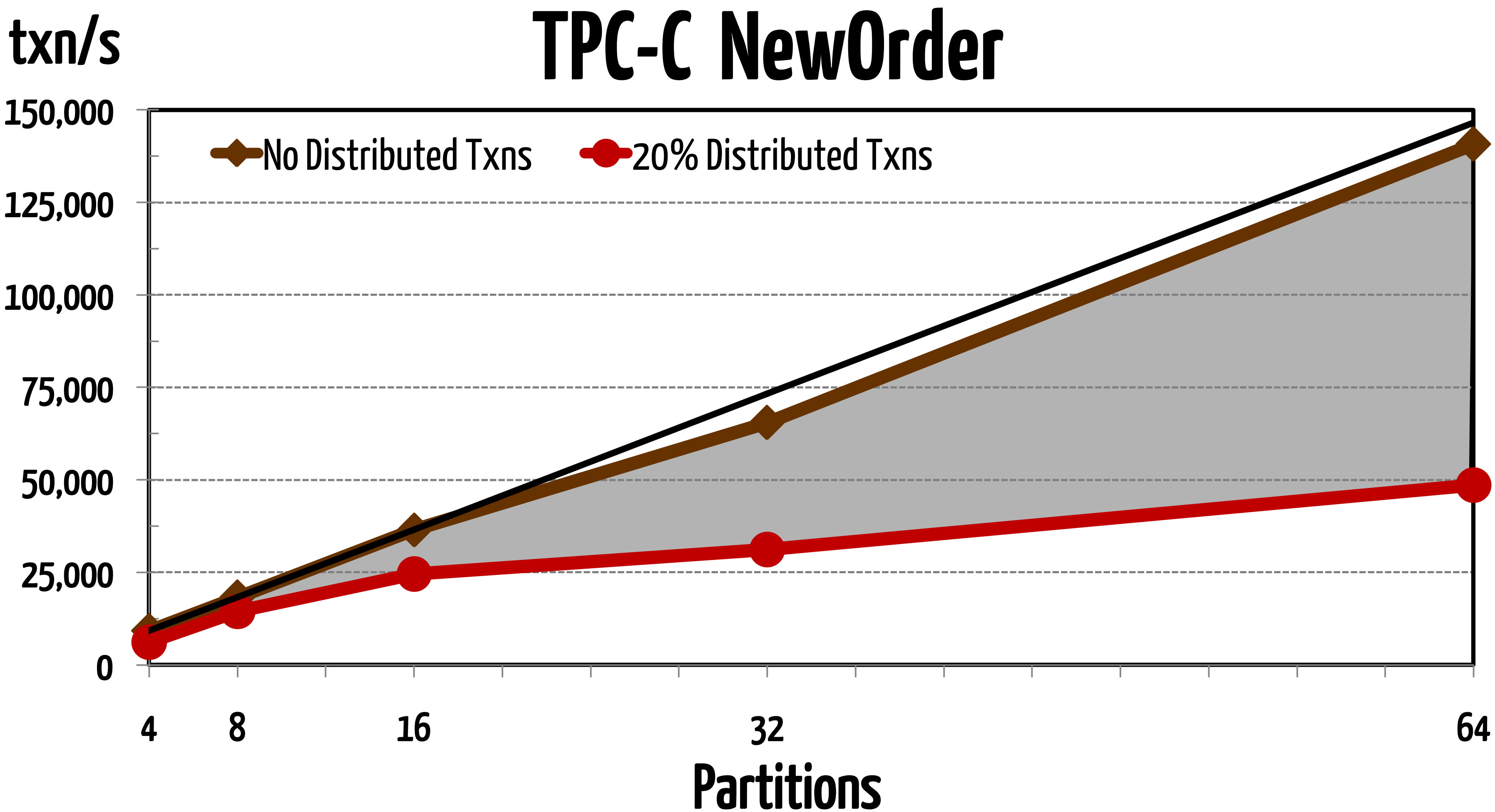

H-Store: A High-Performance, Distributed Main Memory Transaction Processing System

H-Store


Procedure Nar
Input Paramet
Stored Procedure Execution
Database Node
Txn Coordinator
Core
Execution Engine
Partition Data
Main Memory
Client Application
Database Cluster


January 2012

H-Store


TPC-C NewOrder
txn/s
150,000
125,000
100,000
75,000
50,000
25,000
0
No Distributed Txns
20% Distributed Txns
4
8
16
32
64
Partitions

Optimization #1:

Partition database to reduce the number of distributed txns.

Skew-Aware Automatic Database Partitioning in Shared-Nothing, Parallel OLTP Systems

SIGMOD 2012


Skew-Aware Automatic Database Partitioning in Shared-Nothing, Parallel OLTP Systems

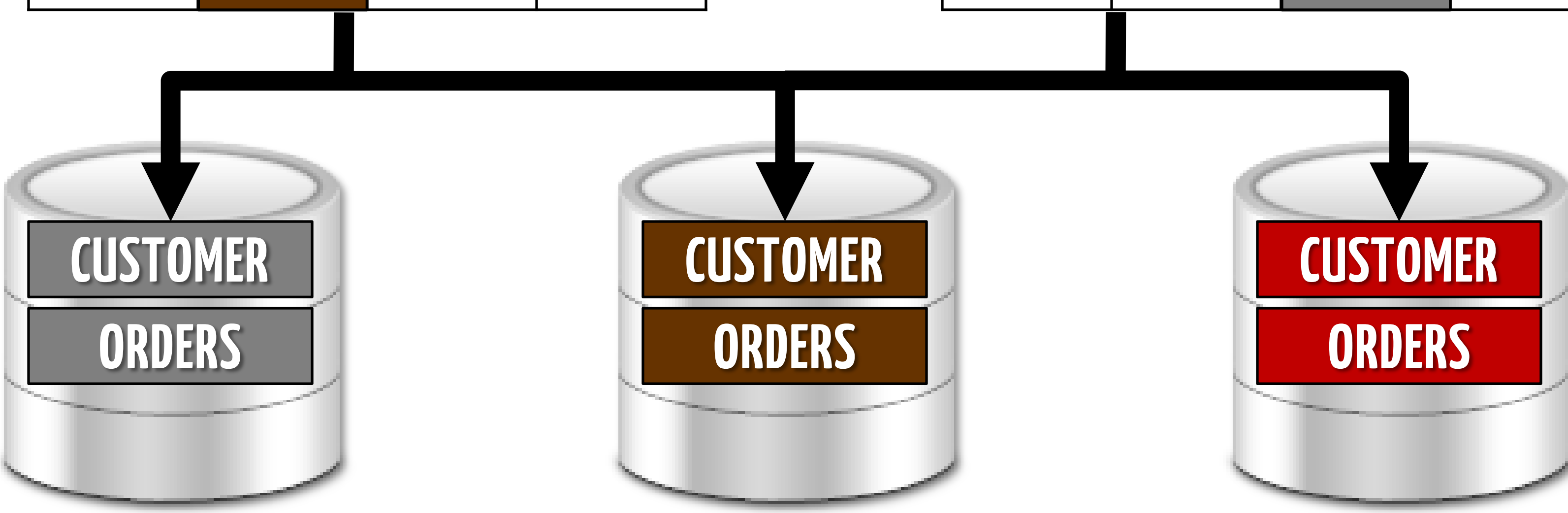
Horticulture

CUSTOMER

c_id	c_w_id	c_last	…
1001	5	RZA	-
1002	3	GZA	-
1003	12	Raekwon	-
1004	5	Deck	-
1005	6	Killah	-
1006	7	ODB	-

ORDERS

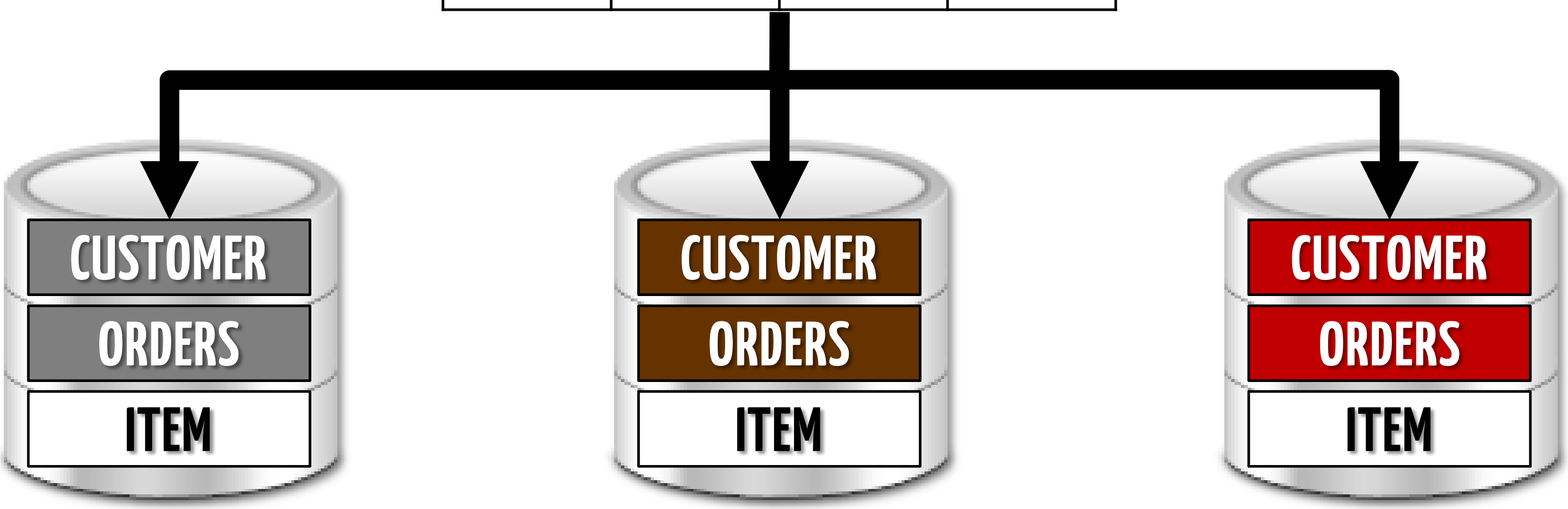
o_id	o_c_id	o_w_id	…
78703	1004	5	-
78704	1002	3	-
78705	1006	7	-
78706	1005	6	-
78707	1005	6	-
78708	1003	12	-



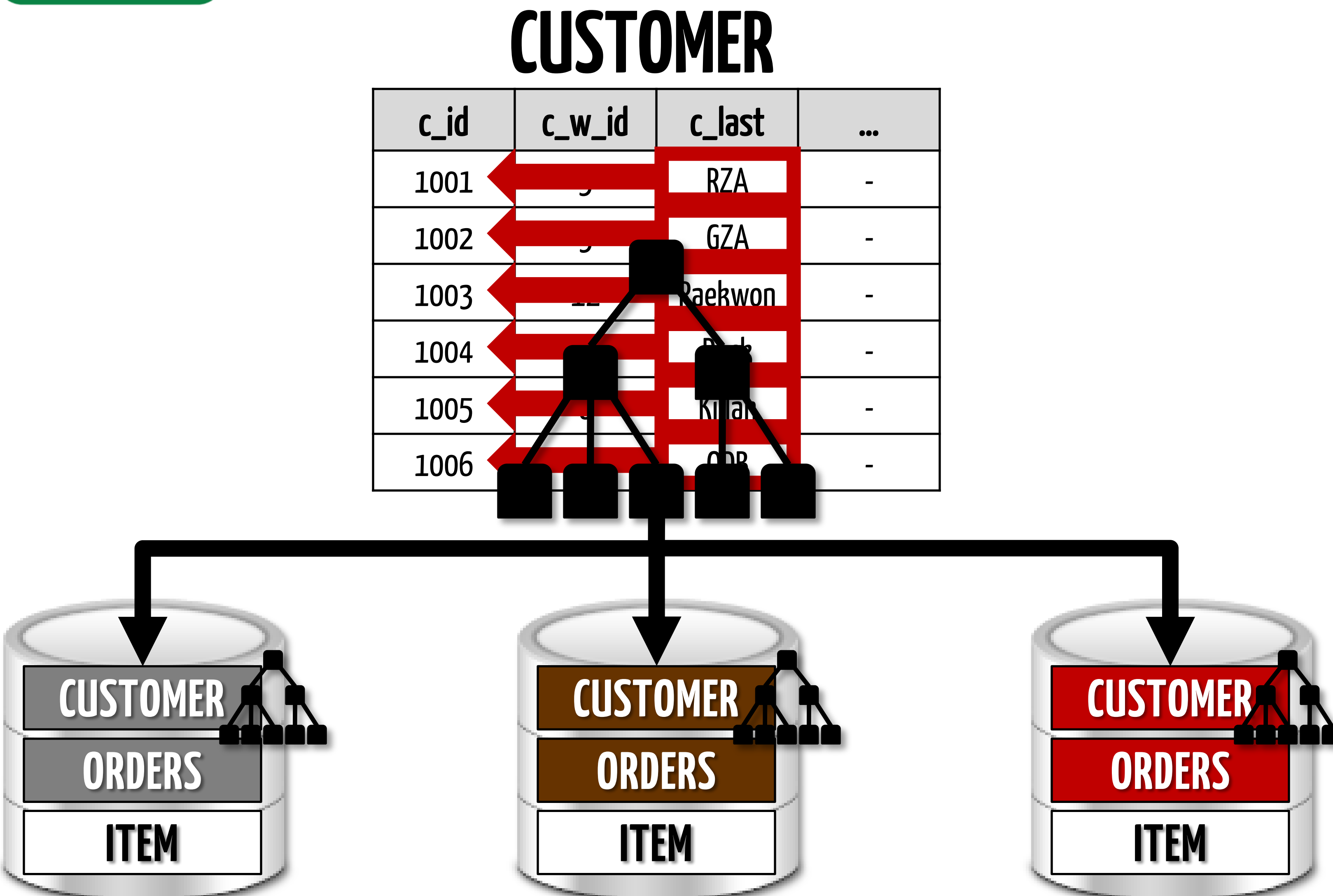
Horticulture

ITEM

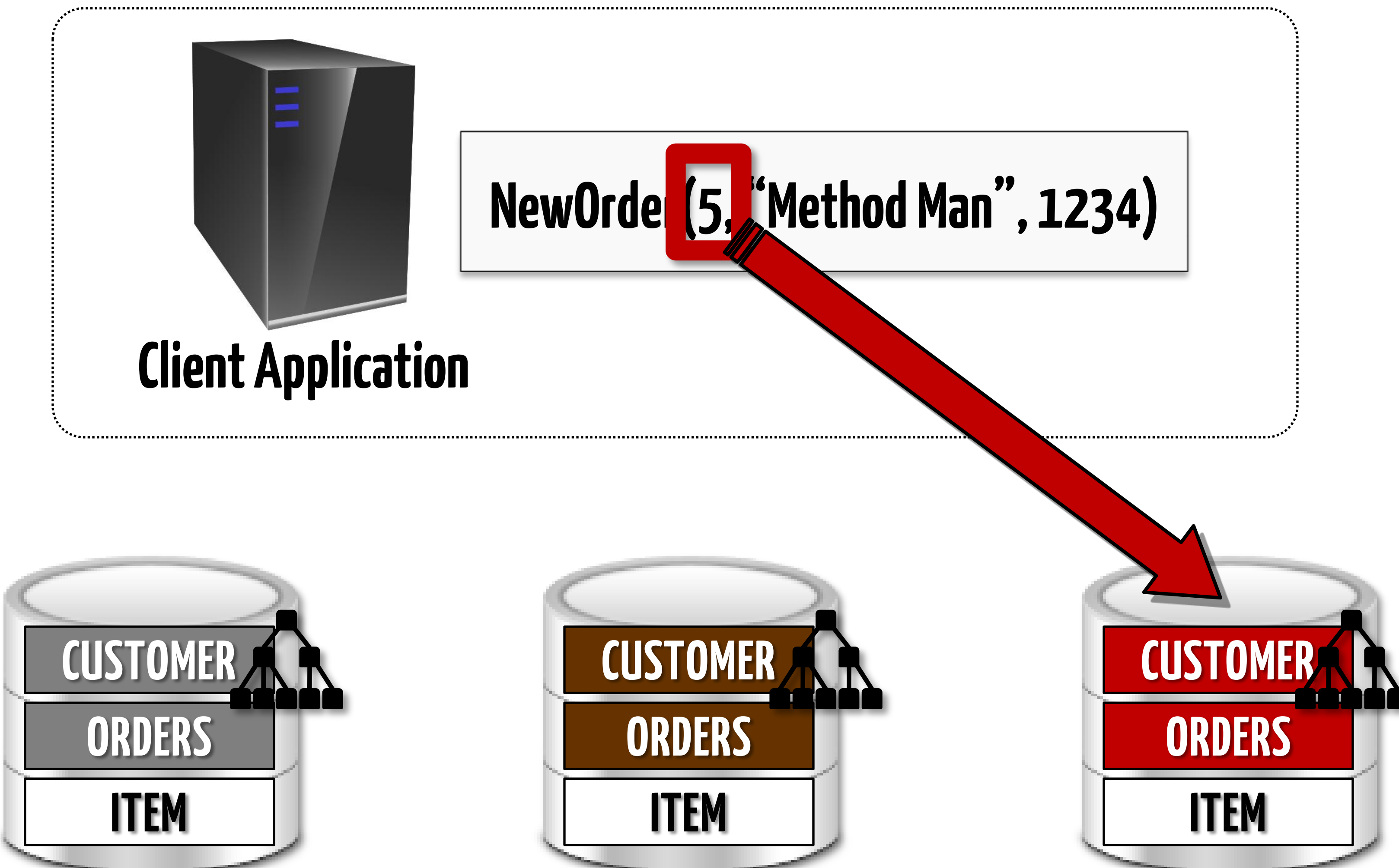
i_id	i_name	i_price	...
603514	XXX	23.99	-
267923	XXX	19.99	-
475386	XXX	14.99	-
578945	XXX	9.98	-
476348	XXX	103.49	-
784285	XXX	69.99	-



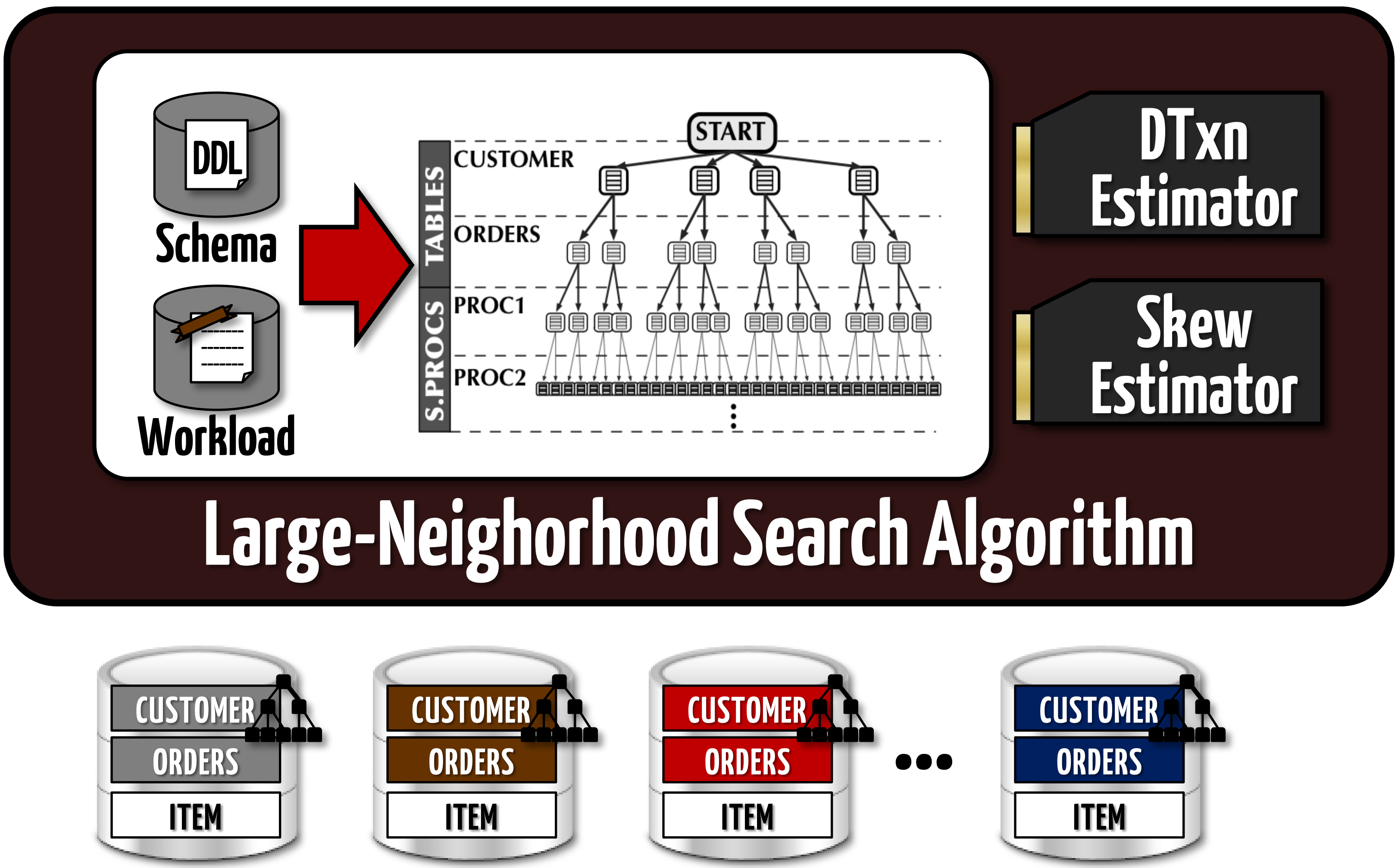
Horticulture



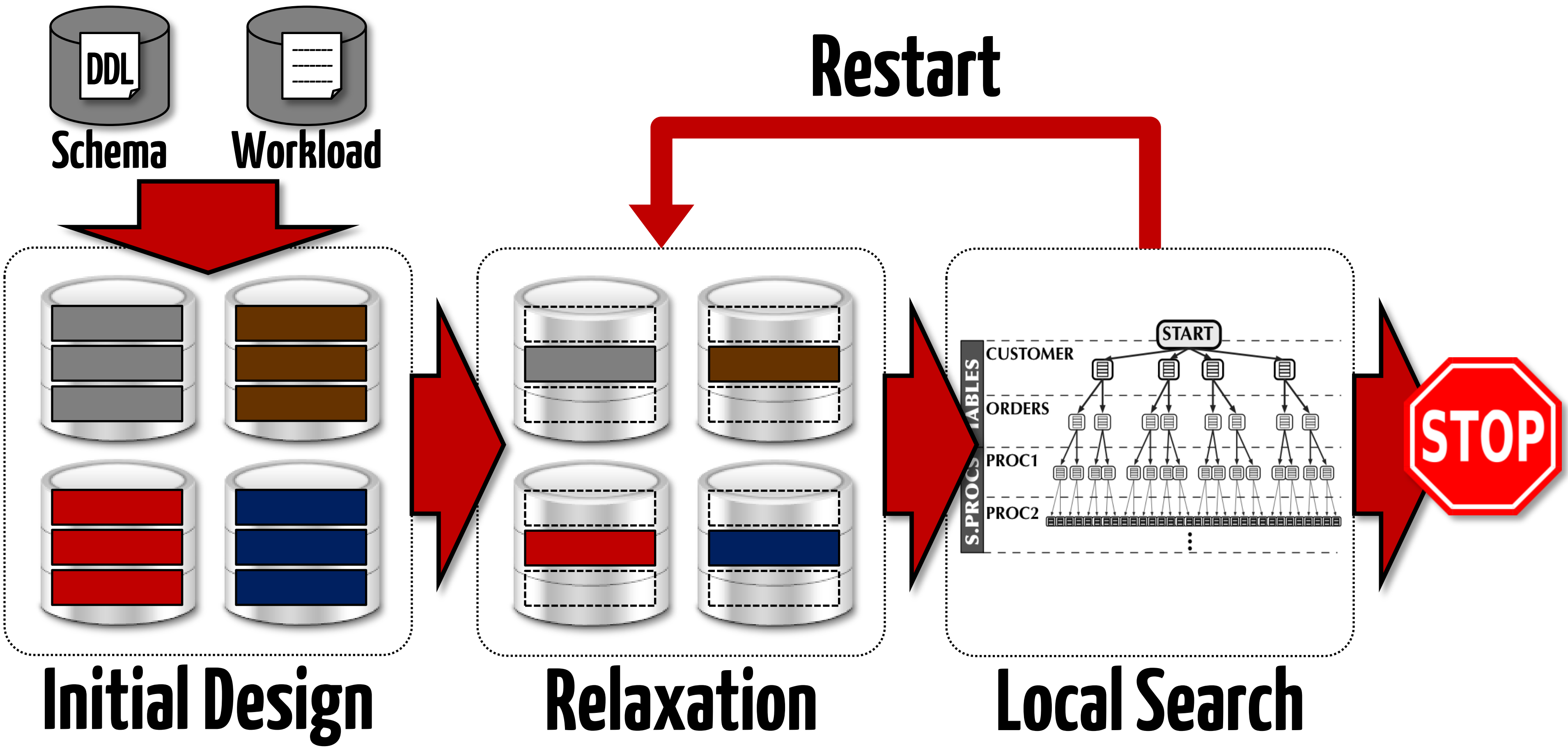
Horticulture



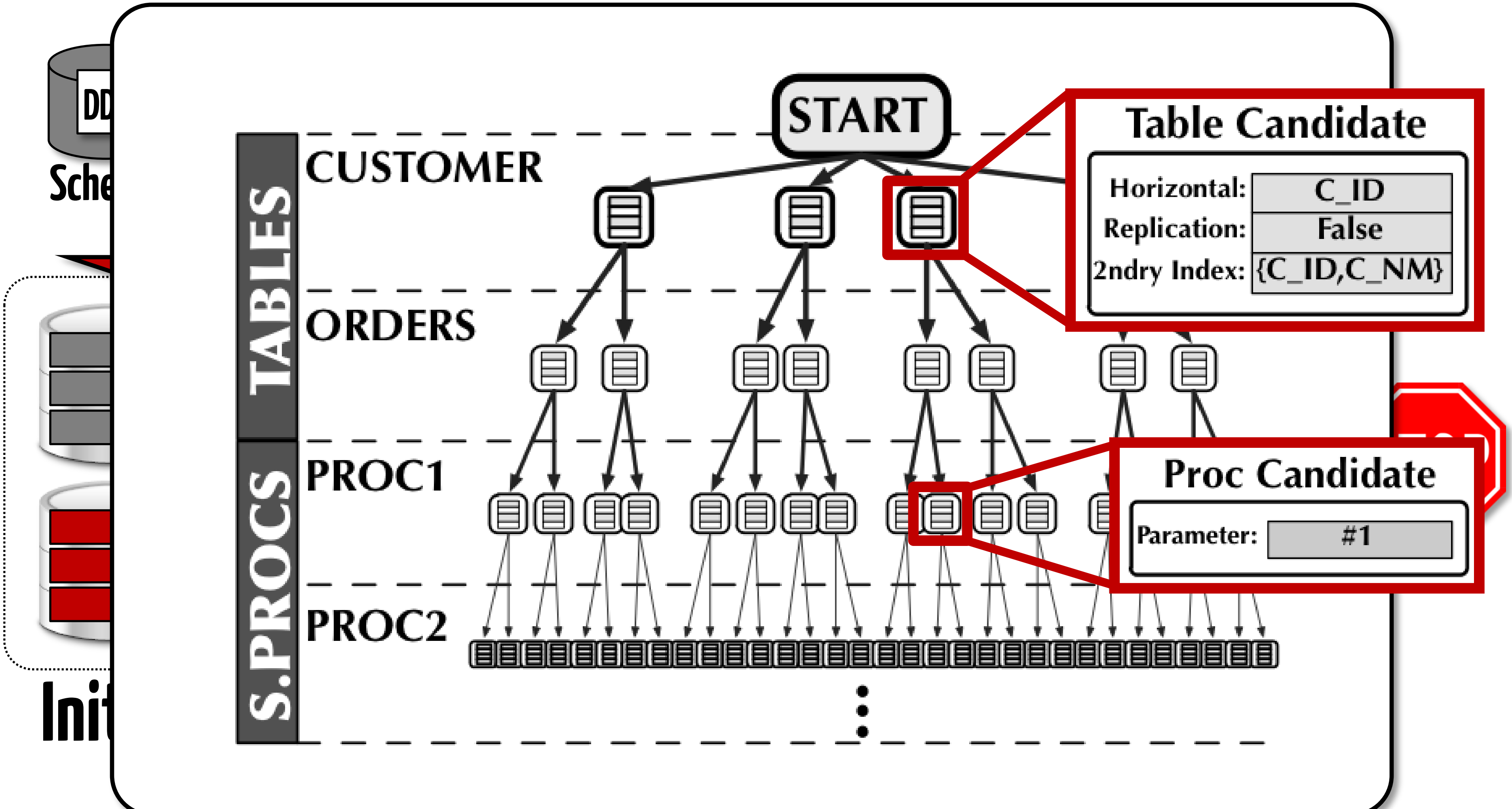
Horticulture



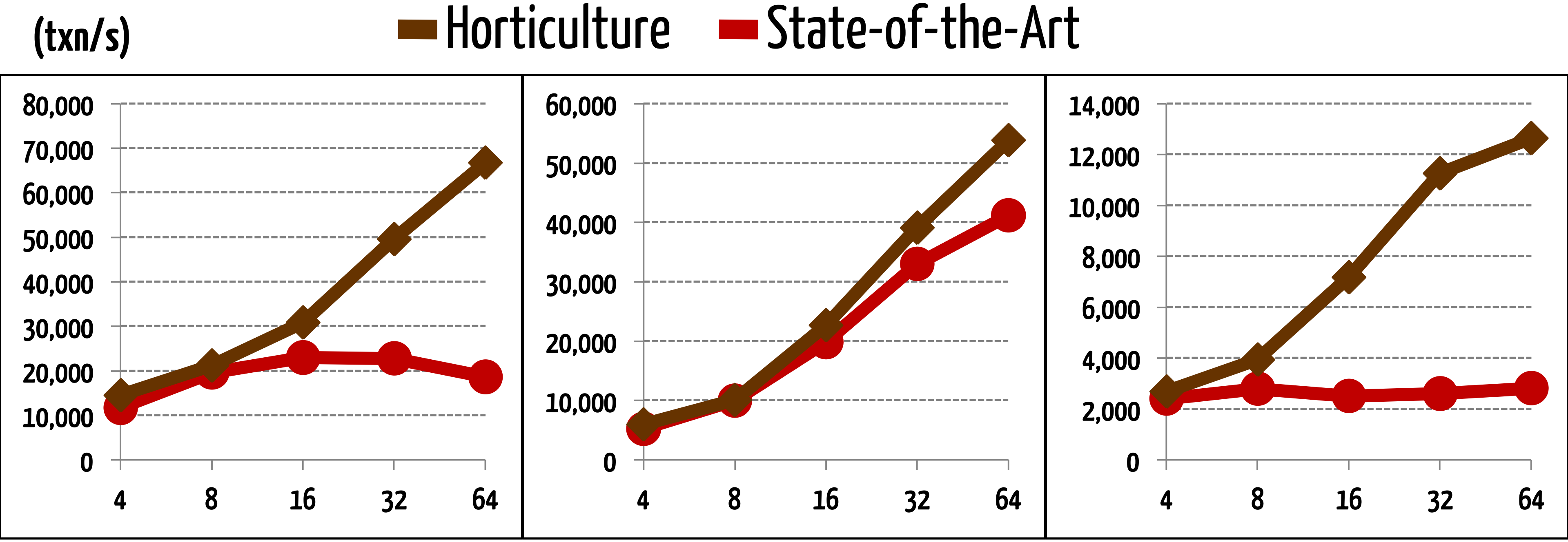
Large-Neighborhood Search



Large-Neighborhood Search



Throughput

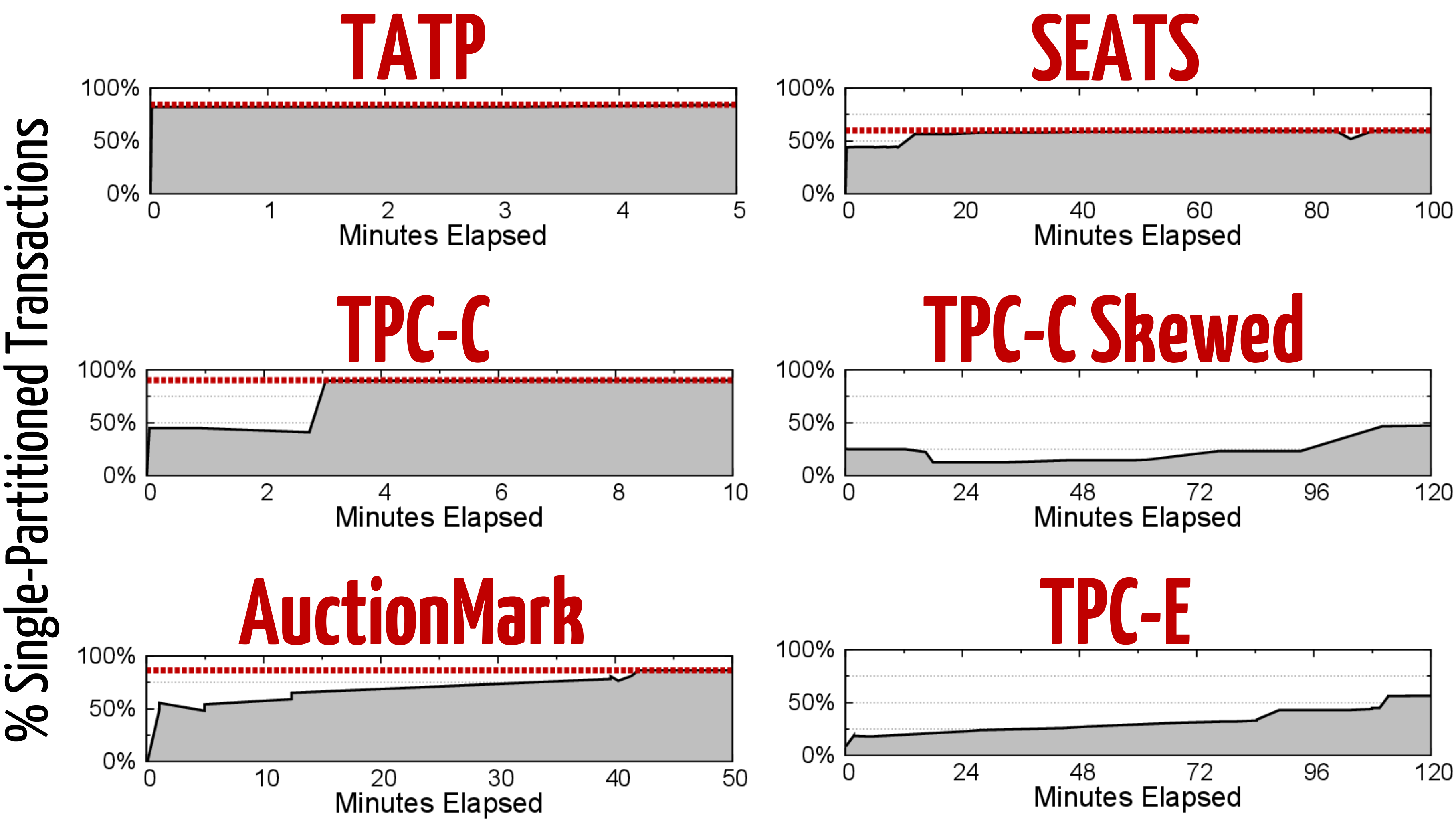


TATP
+88%

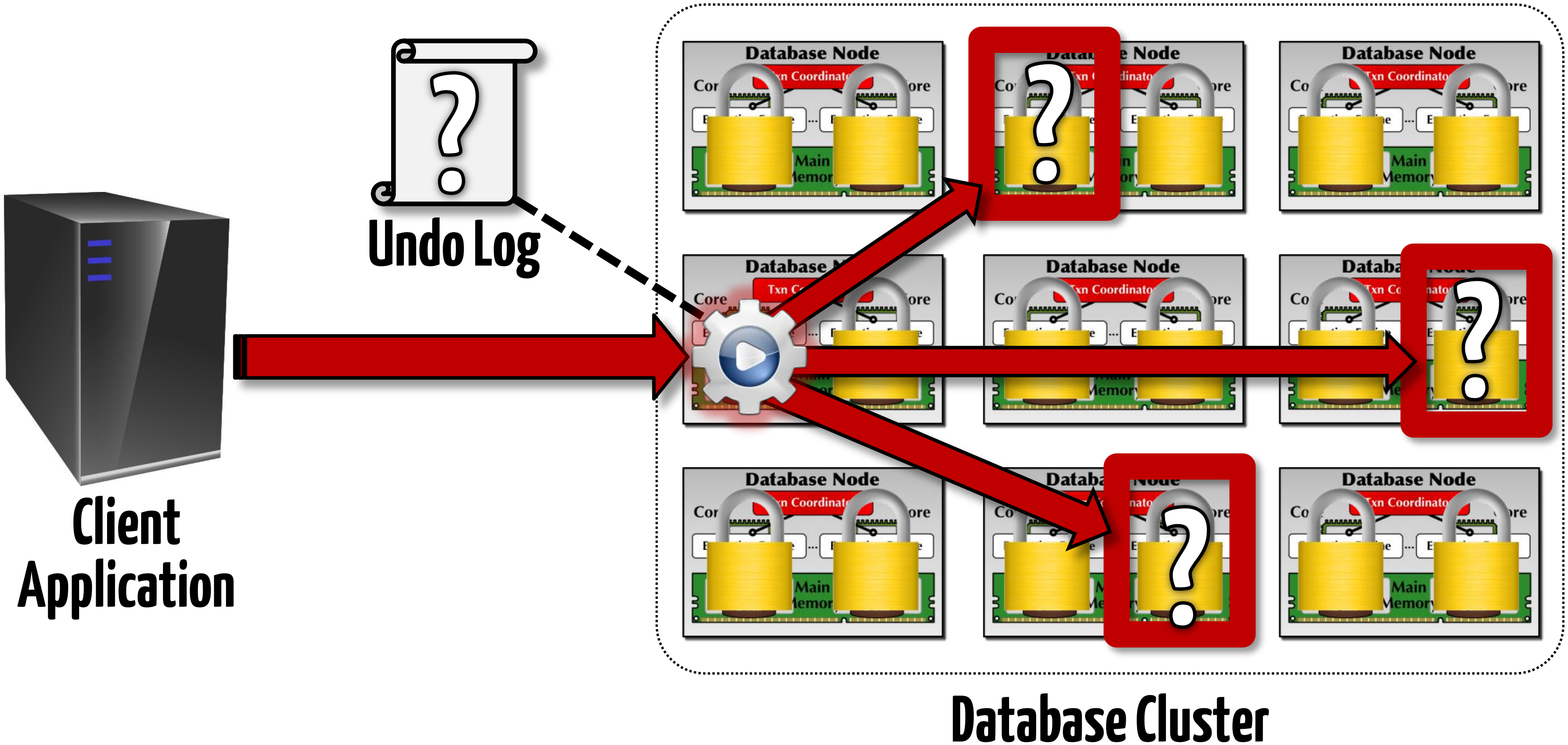
TPC-C
+16%

TPC-C Skewed
+183%

Search Times



H-Store



Undo Log
Client Application
Database Node
Txn Coordinator
Main Memory
Database Cluster

Optimization #2:

Predict what txns will do before they execute.

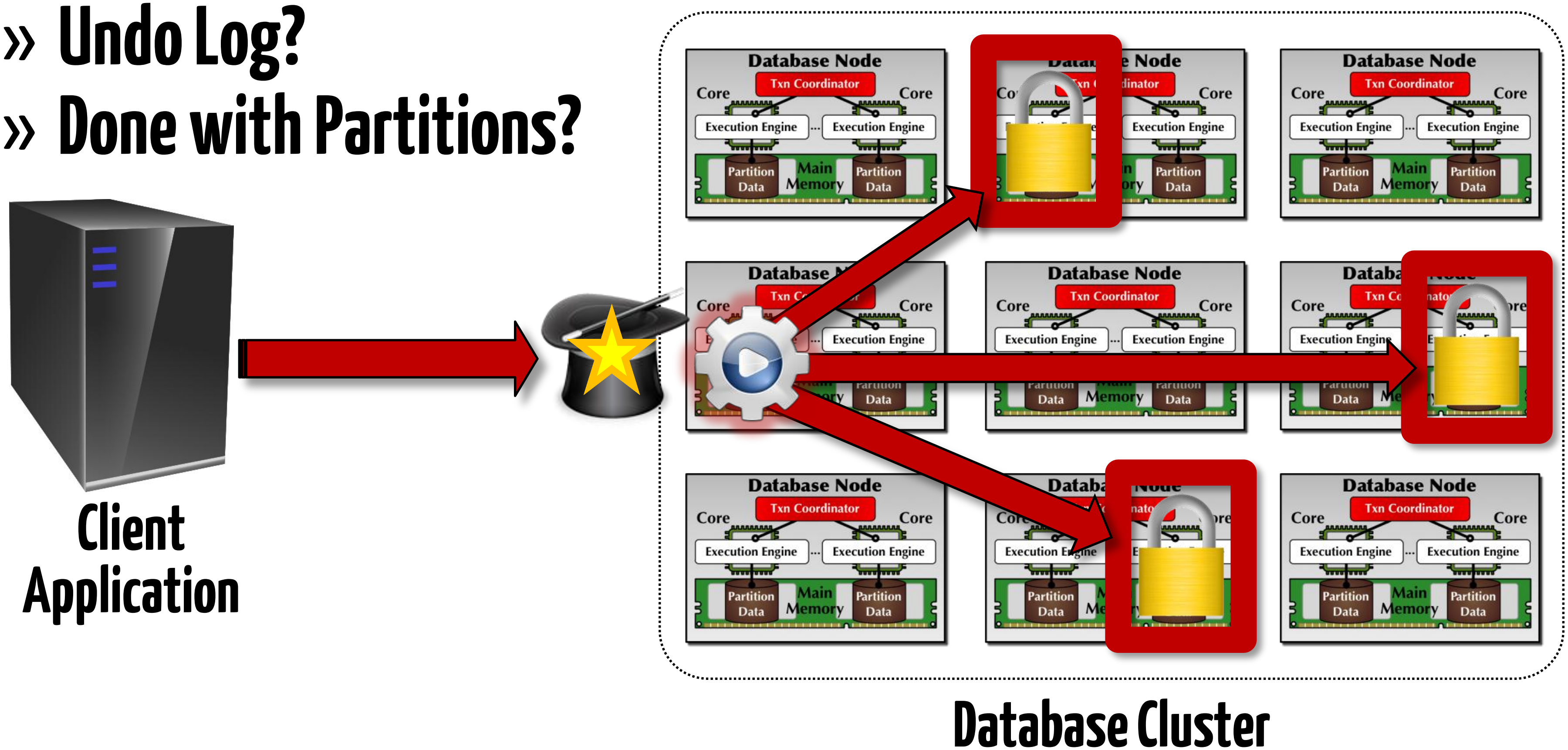
On Predictive Modeling for Optimizing Transaction Execution in Parallel OLTP Systems

VLDB, vol 5. issue 2, October 2011

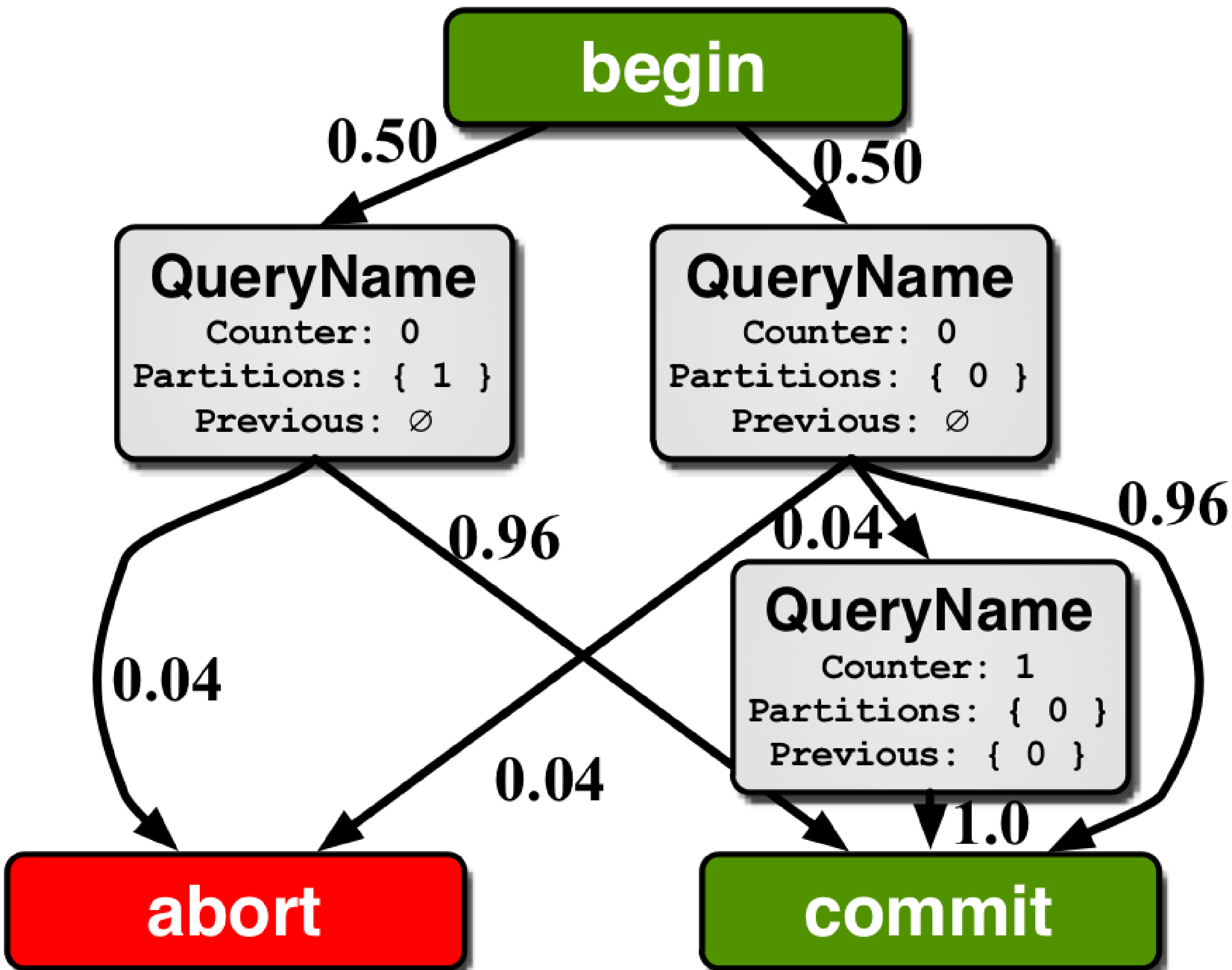

On Predictive Modeling for Optimizing Transaction Execution in Parallel OLTP Systems

Houdini

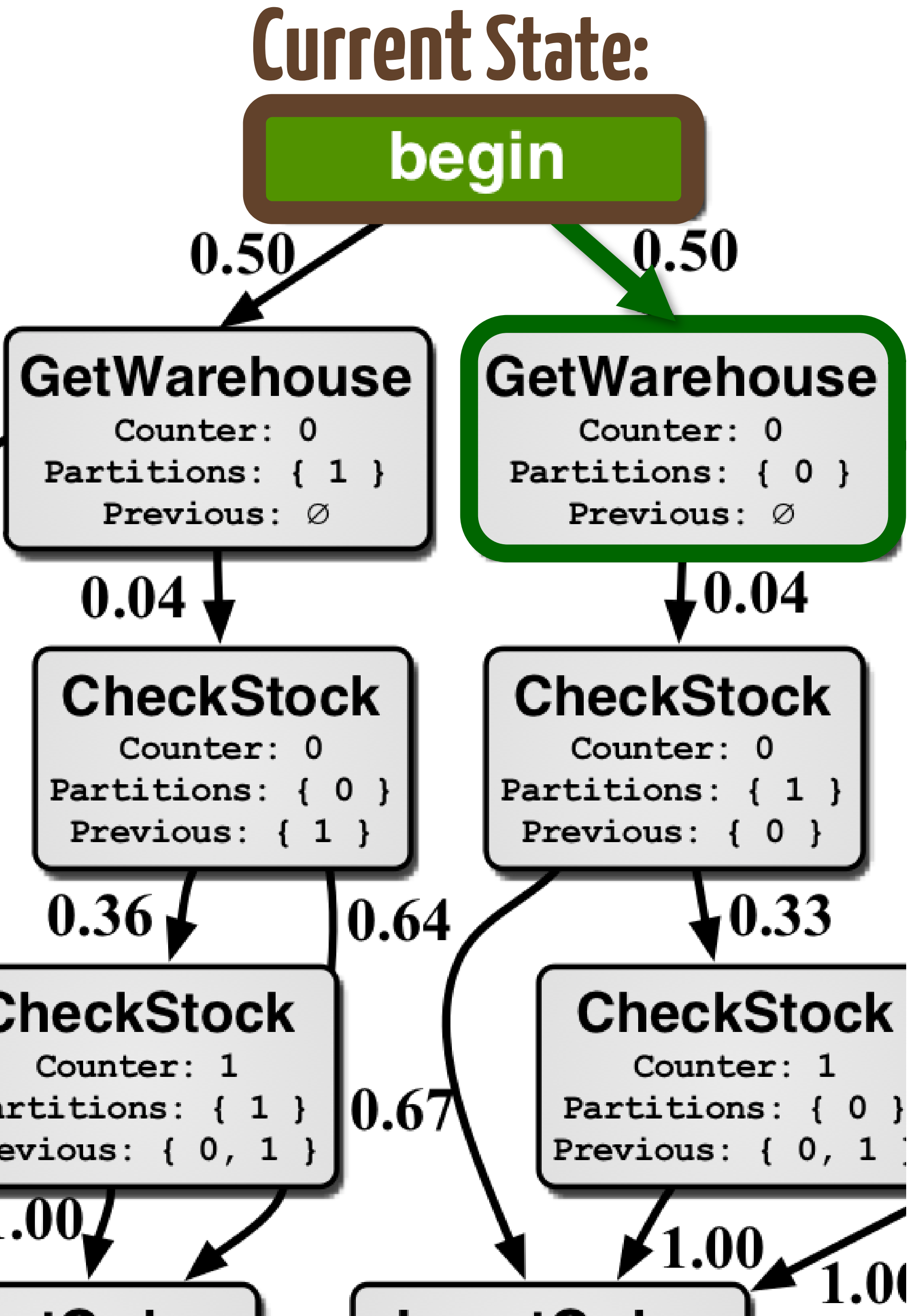
- » Partitions Touched?
- » Undo Log?
- » Done with Partitions?



Houdini



begin
0.50
0.50
QueryName
Counter: 0
Partitions: { 1 }
Previous: ∅
QueryName
Counter: 0
Partitions: { 0 }
Previous: ∅
0.96
0.04
0.96
QueryName
Counter: 1
Partitions: { 0 }
Previous: { 0 }
0.04
0.04
1.0
abort
commit



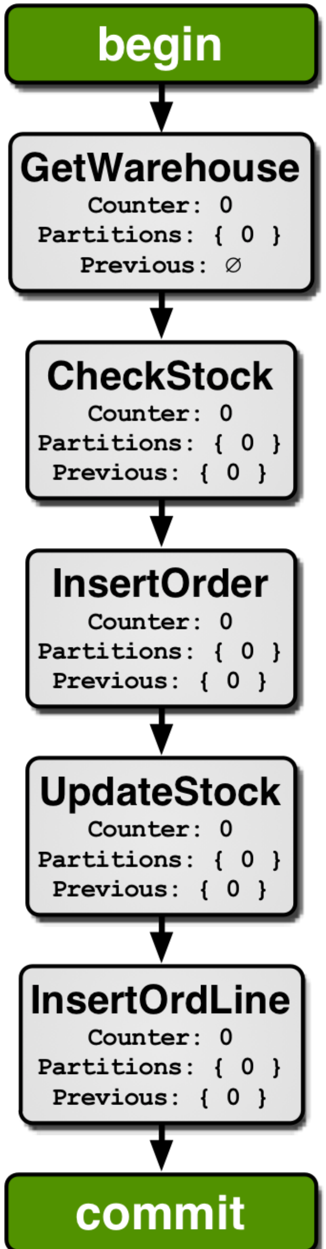
Input Parameters:

```
w_id=0
i_w_ids=[0,1]
i_ids=[1001,1002]
```

GetWarehouse:

```
SELECT * FROM WAREHOUSE
 WHERE W_ID = ?
```

Estimated Execution Path



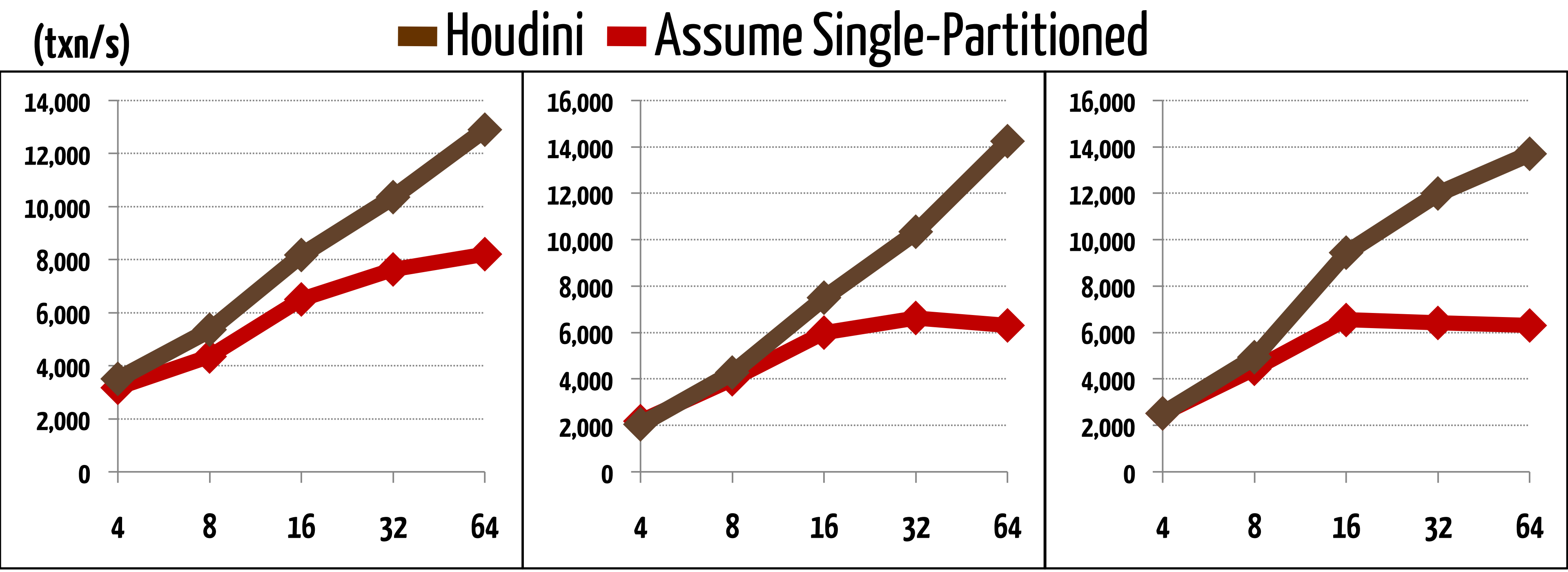
Input Parameters:

```
w_id=0
i_w_ids=[0,1]
i_ids=[1001,1002]
```

Transaction Estimate:

Confidence Coefficient:	0.96
Best Partition:	0
Partitions Accessed:	{ 0 }
Use Undo Logging:	Yes

Throughput

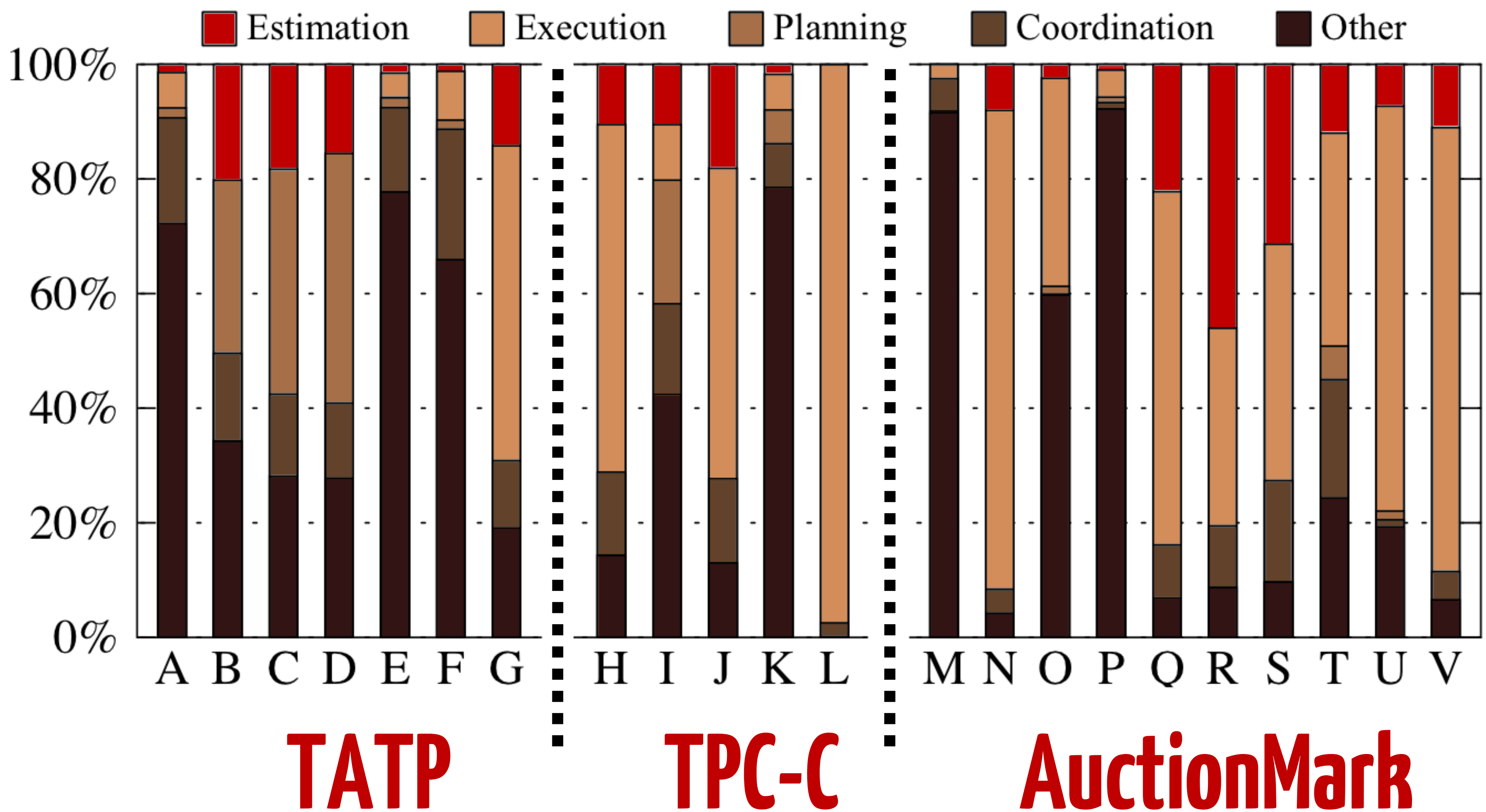


TATP
+57%

TPC-C
+126%

AuctionMark
+117%

Prediction Overhead



Conclusion:

Achieving fast performance is more than just using only RAM.

Future Work:

Reduce distributed txn overhead through creative scheduling.

h-store

hstore.cs.brown.edu

github.com/apavlo/h-store

Help is Available

+1-212-939-7064

Graduate Student Abuse Hotline

Available 24/7

Collect Calls Accepted