

Automatic Database Partitioning in Parallel OLTP Systems

@andy_pavlo



BROWN

SIGMOD
May 22nd, 2012









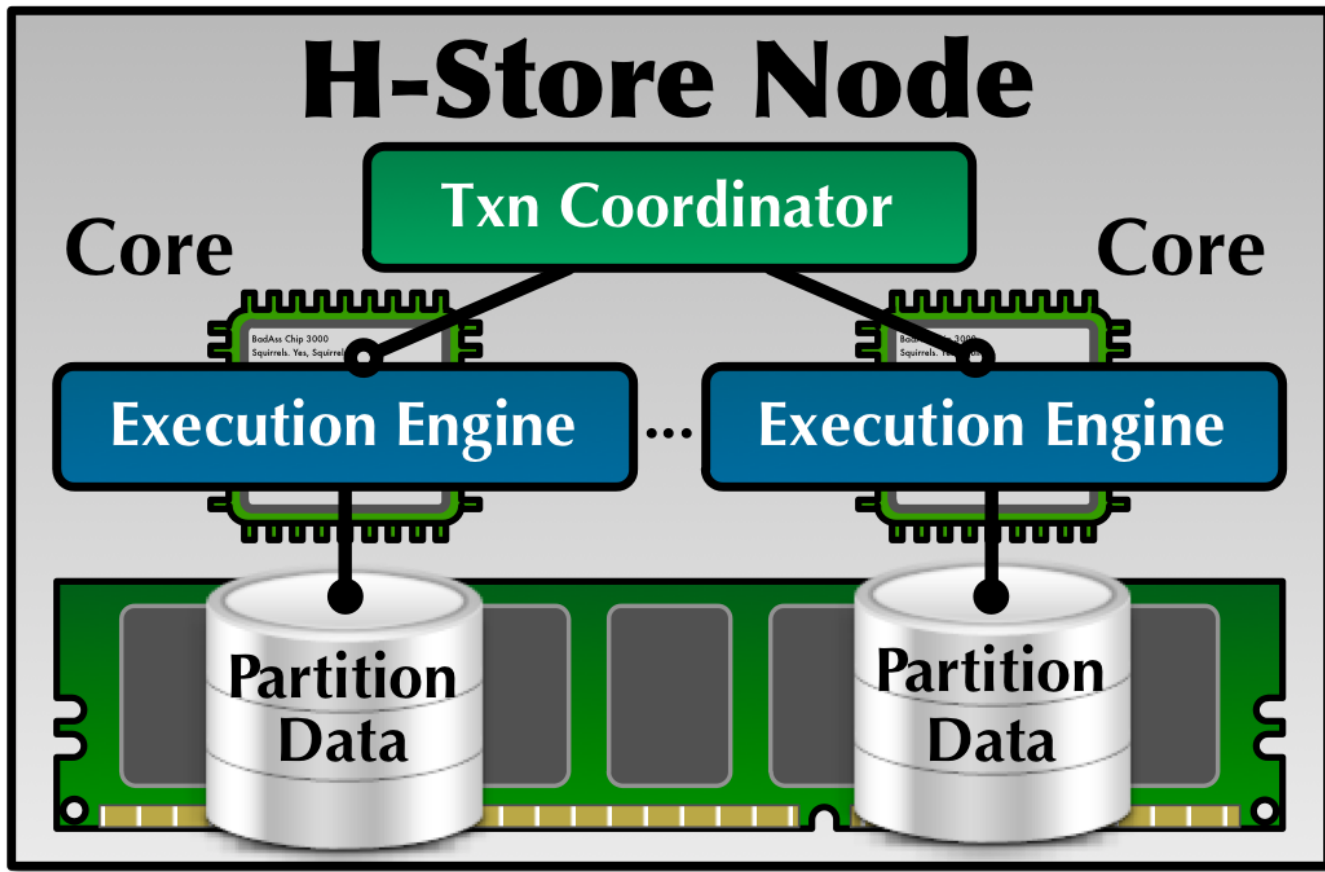
-Store

Main Memory • Parallel • Shared-Nothing
Transaction Processing

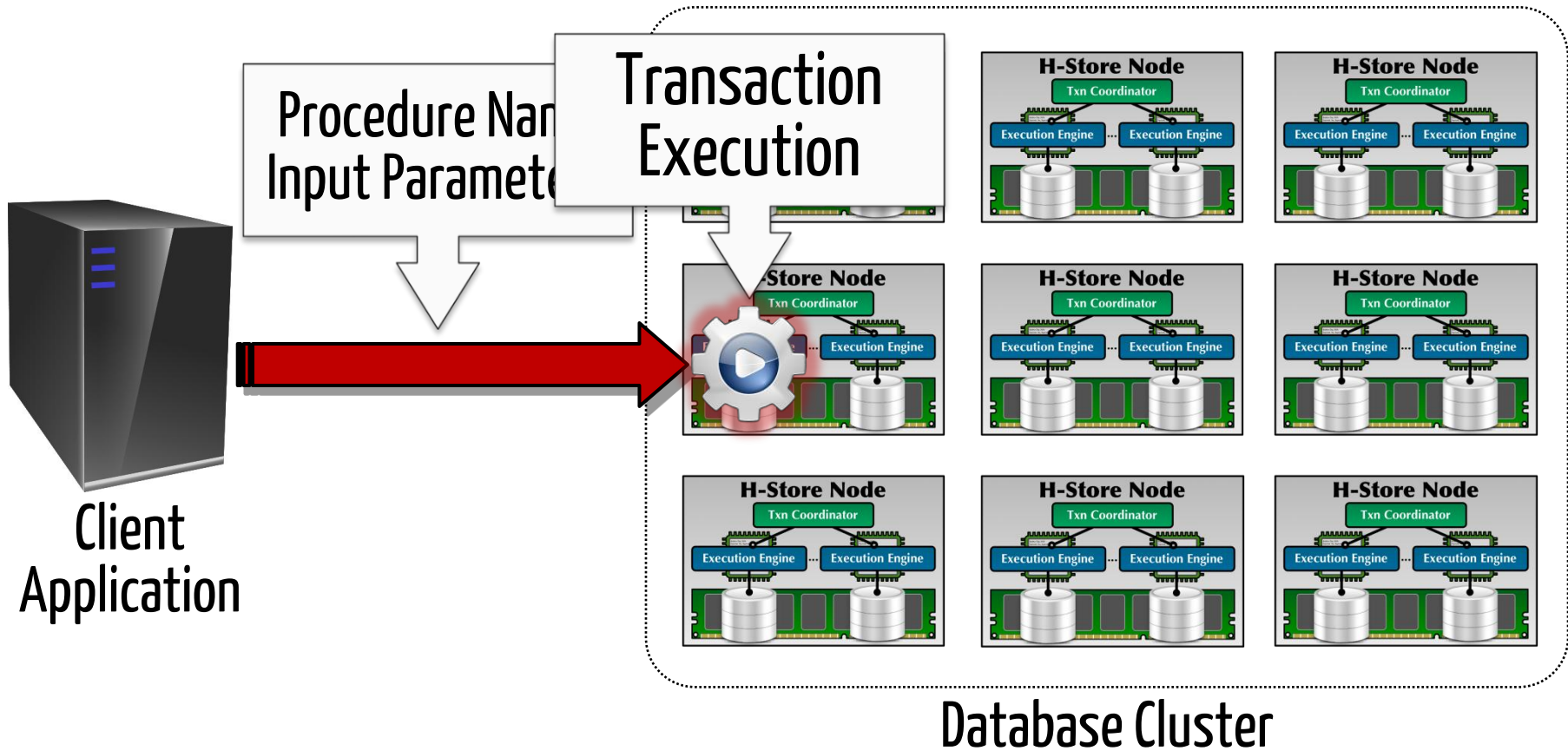
H-Store: A High-Performance, Distributed
Main Memory Transaction Processing System
Proc. VLDB Endow., vol. 1, iss. 2, pp. 1496-1499, 2008.



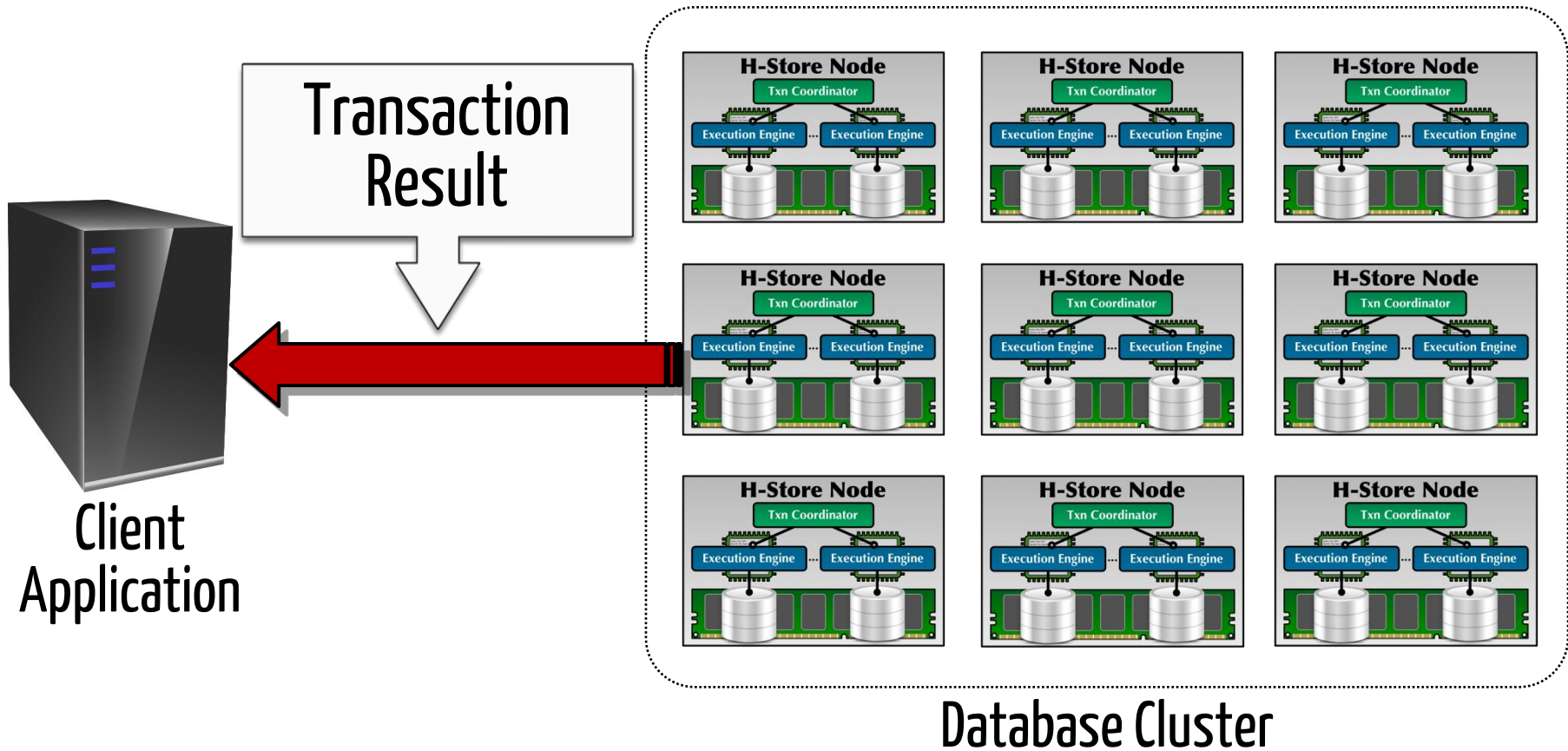
H-Store

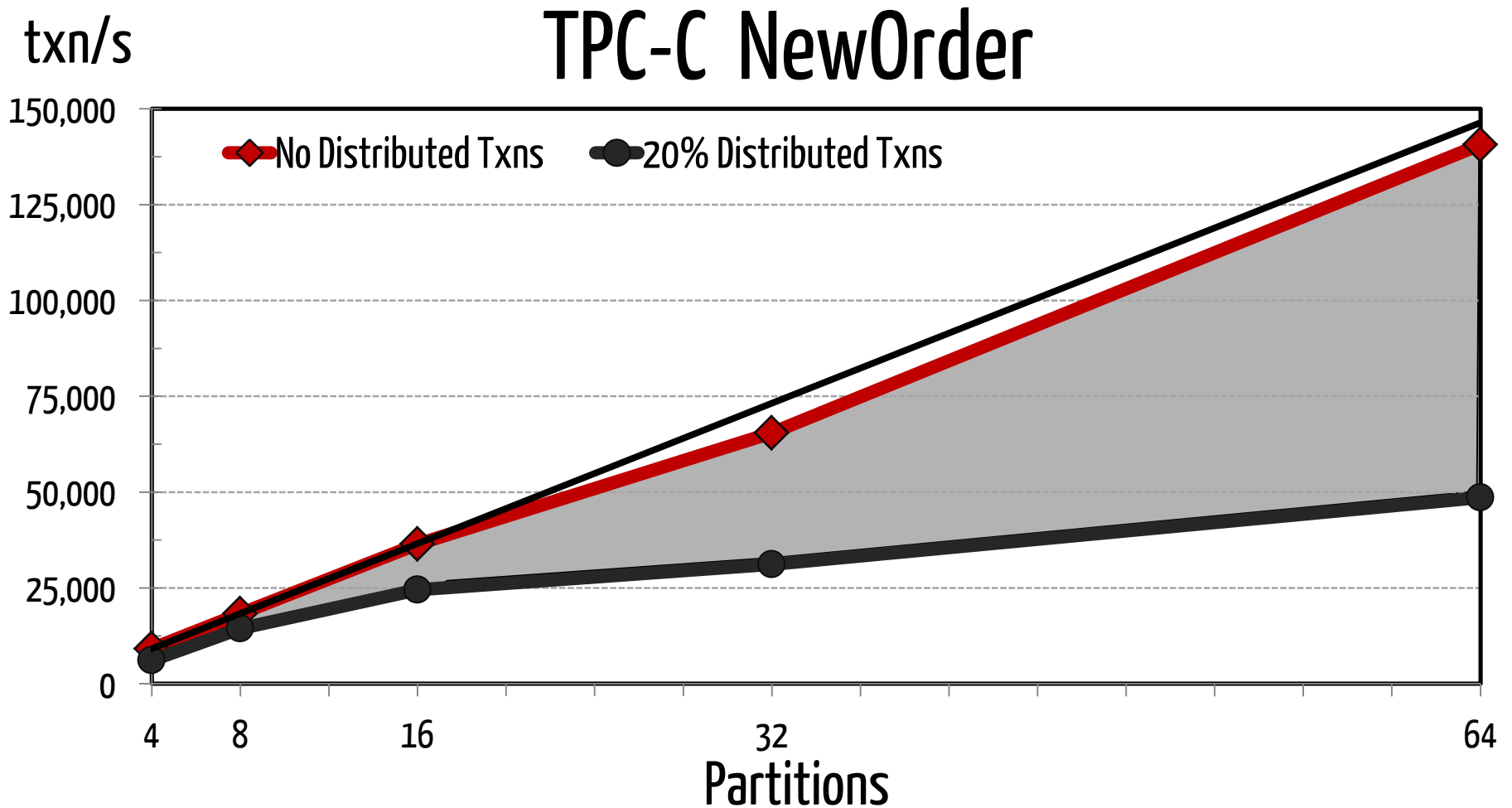


H-Store



H-Store









Horticulture

Automatic Database Design Tool
for Parallel Systems

Skew-Aware Automatic Database Partitioning
in Shared-Nothing, Parallel OLTP Systems
SIGMOD 2012





Horticulture

DDL

CUSTOMER

ORDERS

ITEM

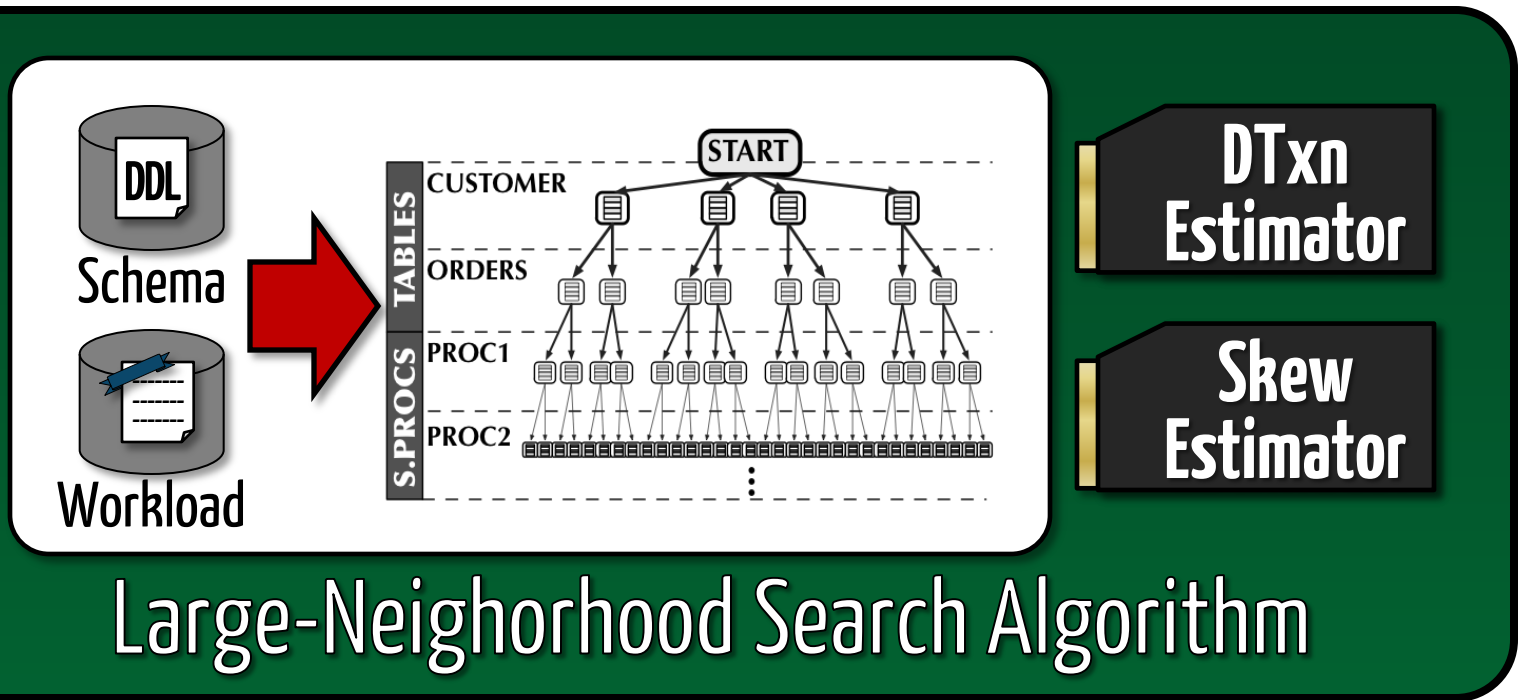
NewOrder

```
SELECT * FROM WAREHOUSE  
WHERE W_ID = 10;
```

```
SELECT * FROM DISTRICT WHERE  
D_W_ID = 10 AND D_ID = 9;
```

```
INSERT INTO ORDERS (O_W_ID,  
O_D_ID, O_C_ID,...) VALUES  
(10, 9, 12345,...);  
⋮
```

Horticulture



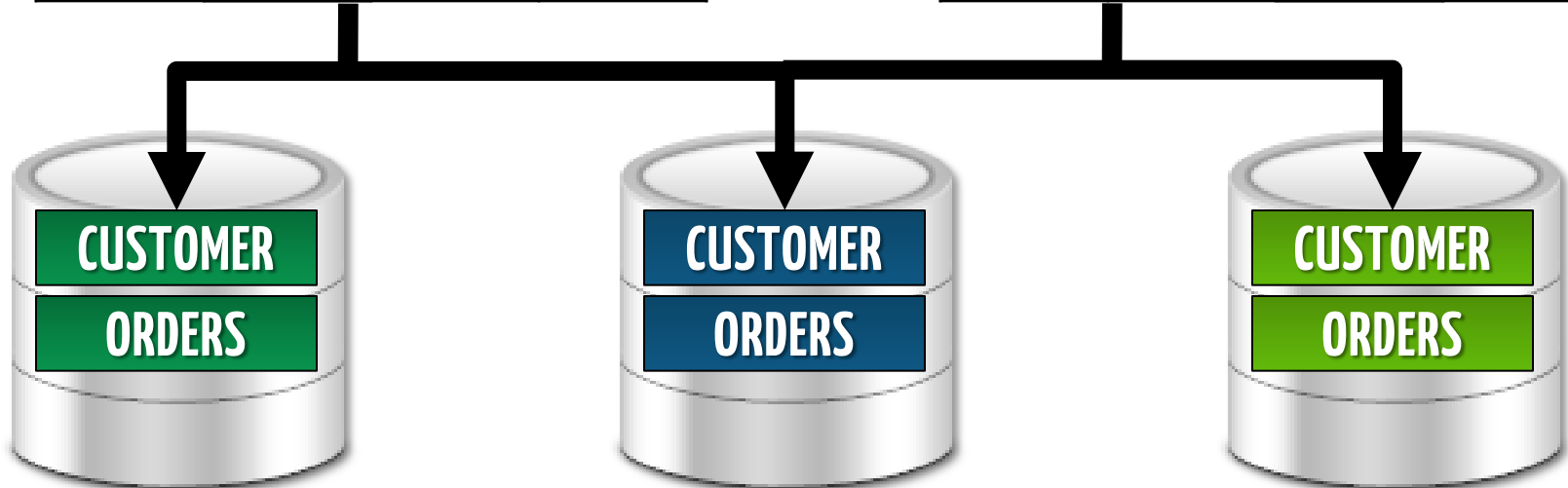
Horticulture

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	-

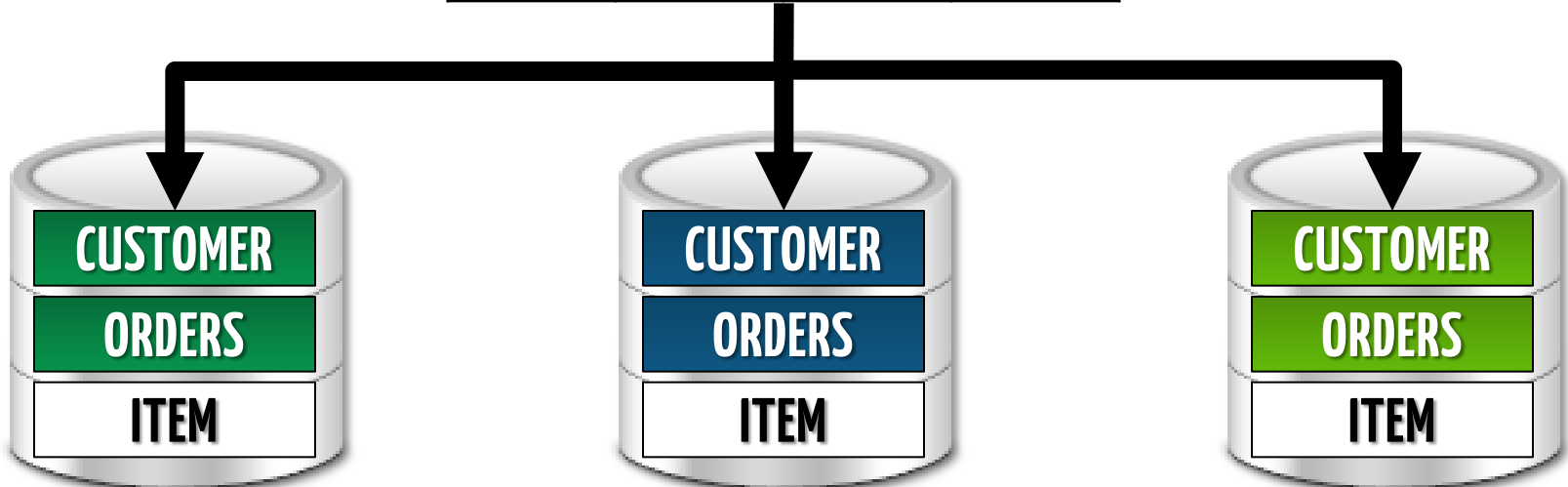




orticulture

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

CUSTOMER

c_id	c_w_id	c_last	...
1001	←	RZA	-
1002	←	GZA	-
1003	←	Raekwon	-
1004	←	Rak	-
1005	←	Killa	-
1006	←	OP	-



Horticulture



Client Application

? ? ?
NewOrder(5, "Method Man", 1234)



Horticulture



Client Application

NewOrder(5, "Method Man", 1234)



Large-Neighborhood Search

Input



Schema

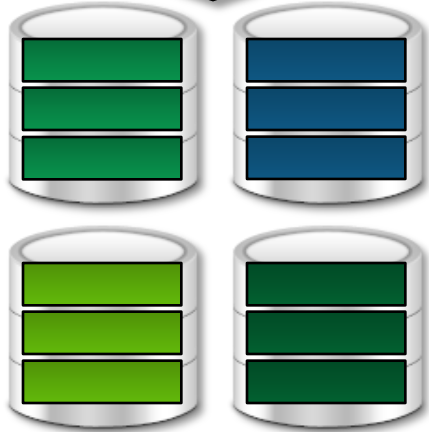


Workload

Best Design



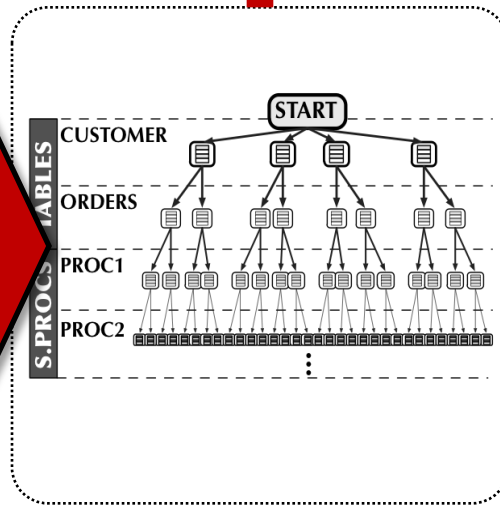
Restart



Initial Design



Relaxation



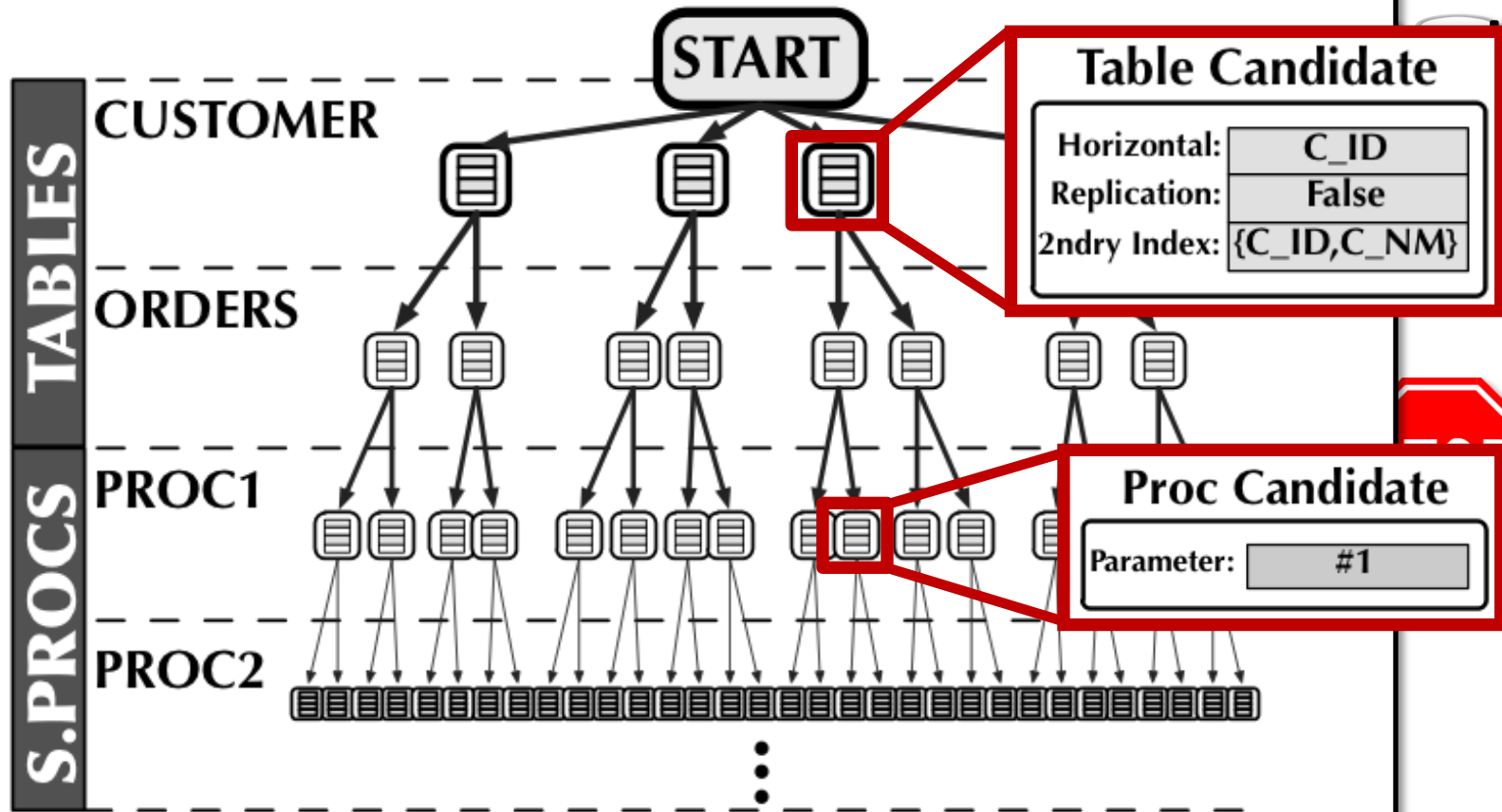
Local Search



Large-Neighborhood Search

Input

Best Design

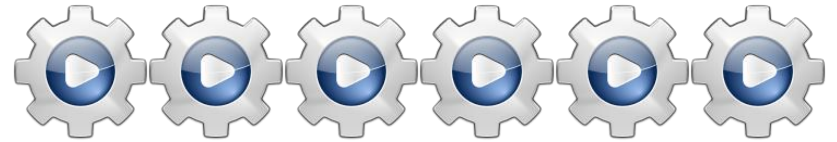
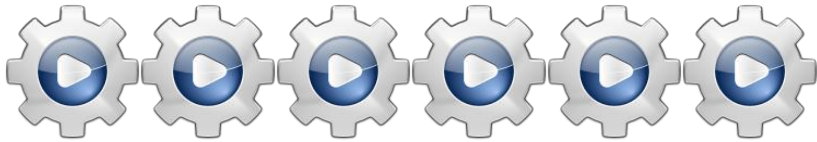


Cost Model

Distributed
Transactions

+

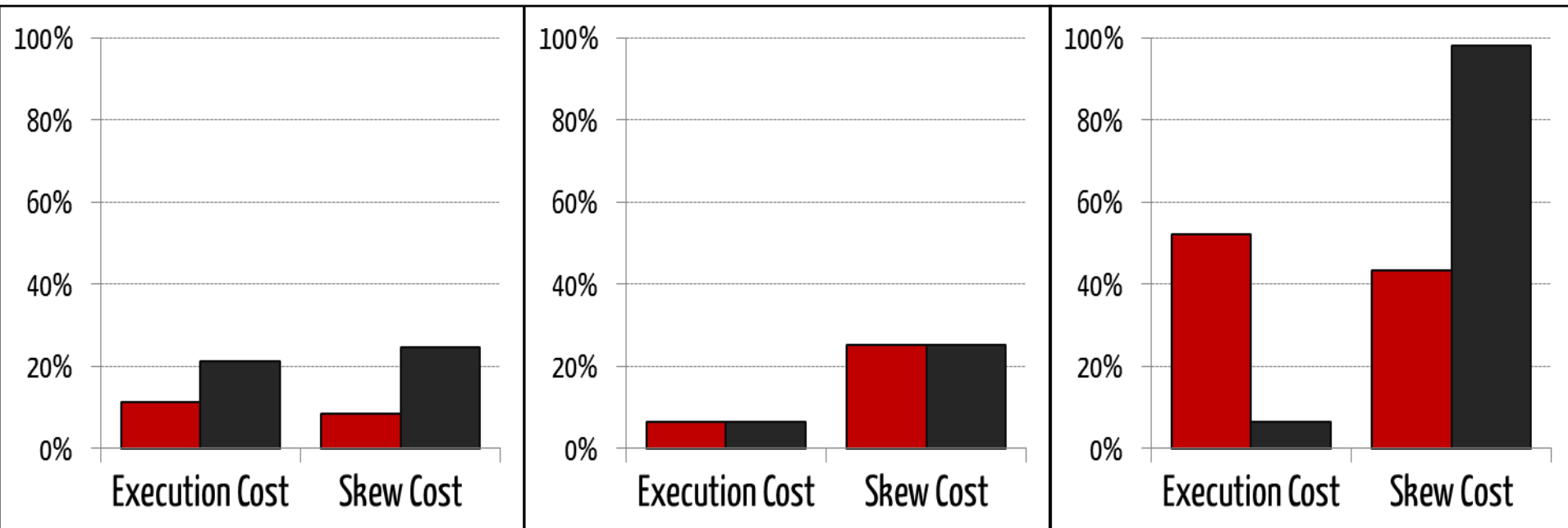
Workload
Skew Factor



Algorithm Comparison

(cost estimate)
lower is better

■ Horticulture ■ State-of-the-Art



TATP

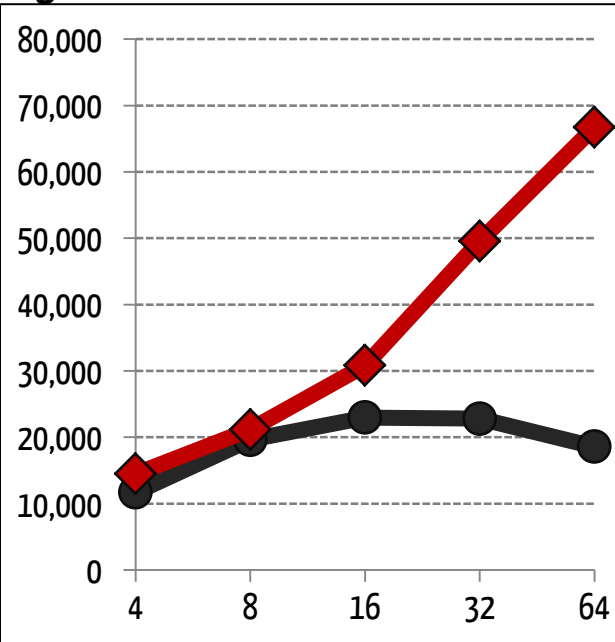
TPC-C

TPC-C Skewed

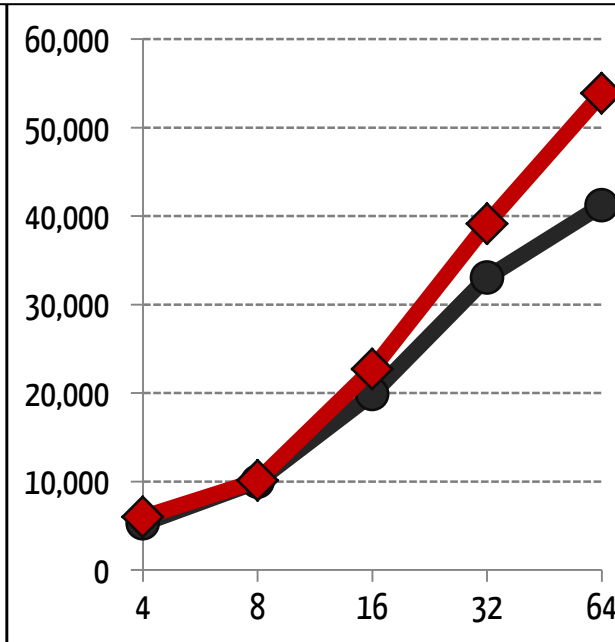
Throughput

(txn/sec)
higher is better

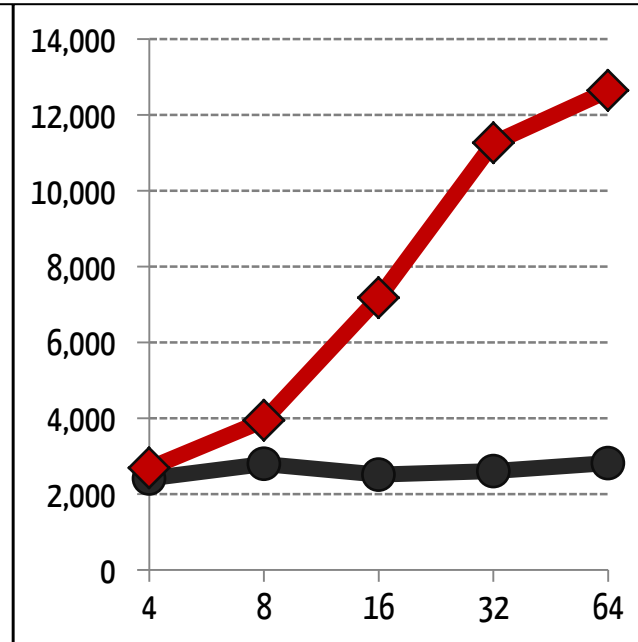
■ Horticulture ■ State-of-the-Art



TATP
+88%



TPC-C
+16%



TPC-C Skewed
+183%



Conclusion:

Dating scene is still difficult.

But partitioning your
database is now easier.



-Store

H-Store

Next Generation OLTP Database Research

[About](#) [Documentation](#) [Downloads](#) [People](#) [Projects](#) [Publications](#)

H-Store is an experimental main-memory, parallel database management system that is optimized for OLTP applications. It is a highly distributed, row-store-based relational database that runs on a cluster on shared-nothing, main memory executor nodes.

The H-Store project is a collaboration between [MIT](#), [Brown University](#), [Yale University](#), and HP Labs.



New Release (March 2012)

March 16th, 2012 at 9:36 pm

The H-Store project is releasing the latest version of its experimental OLTP database management system. This version contains a large number of speed and stability improvements. The next release will contain mostly new features.

Major highlights of this release:

1. Significant speed optimizations for single-partition transactions. [\[c7f1be4a\]](#) [\[c677abed\]](#) [\[28599fd\]](#) [\[0425253\]](#) [\[243f99d\]](#) [\[5abd178\]](#)
2. Fixed several memory leaks for aborted/read-only transactions. [\[d1d2723\]](#) [\[e429821\]](#)
3. Fixed several stability issues when transactions share lock-free data structures [\[4956be9\]](#)
4. Fixed predicate/aggregate pushdowns for distributed query plans [\[4956be9\]](#) [\[21e9f6d\]](#)
5. Restored the ability for transactions to return the proper status codes on failures that mimic VoltDB's protocol. [\[ae4e0e4\]](#) [\[7574e49\]](#)
6. H-Store's Fabric automatic deployment script for EC2 now supports multiple concurrent "virtual" clusters. This allows you to easily assign instances to independent clusters under a single AWS account. [\[189284e\]](#)
7. Restored preliminary support for VoltDB's regression test suite. [\[116f465\]](#) [\[c434fa2\]](#) [\[1945ba0\]](#)
8. Fixed building on OS X Lion [\[0328973\]](#)

The source code for this release can be downloaded via Github:

<https://github.com/apavlo/h-store/tarball/release-2012-03>

[Older Entries](#)

github

[Signup and Pricing](#) [Explore GitHub](#) [Features](#) [Blog](#) [Login](#)

apavlo / h-store

[Watch](#) [Fork](#) [33](#) [12](#)

Code Network Pull Requests 0 Issues 20 Wiki 1 Graphs

H-Store Parallel Main Memory OLTP Database System — [Read more](#)

<http://hstore.cs.brown.edu>

[ZIP](#) [HTTP](#) [Git Read-Only](#) <https://github.com/apavlo/h-store.git> [Read-Only access](#)

branch: master Files Commits Branches Tags Downloads

Latest commit to the master branch

Added some comments and TODOs for Alex to work deferrable stuff

apavlo authored 10 hours ago commit w79284193

h-store /

name	age	message	history
properties	3 months ago	Cleaned up callback mechanisms in BenchmarkComponent. Switched TPC-C ... [apavlo]	
scripts	21 days ago	Updated TPC-C mappings. Fixed BaseTestCase so that it uses the Abstra... [apavlo]	
src	10 hours ago	Added some comments and TODOs for Alex to work deferrable stuff [apavlo]	
tests	11 hours ago	Added new 'deferrable' commandline argument for 'hstore-prepare' to m... [apavlo]	
third_party	11 days ago	Merged from master [apavlo]	
tools	11 days ago	Merged from master [apavlo]	
classpath	2 months ago	Minor tweaks and importing CompressionService code from VoltDB to H-S... [apavlo]	
gitignore	3 months ago	Updated HStoreConfUtil to dump out HTML files to update the website w... [apavlo]	
gitmodules	4 months ago	Set up hstore-files submodule [c1e749]	
project	2 months ago	Added filters to hide unneeded files in the root directory in Eclipse [apavlo]	
COPYING	2 years ago	Repository overhaul. Moved all of src into trunk [apavlo]	

<http://hstore.cs.brown.edu>

<http://github.com/apavlo/h-store>