Clustrix

“Speed. Scale. Simplicity.”
History

• Founded 2006
• San Francisco based
• ~60 employees
• CEO and President Robin Purohit
  – former HP Vice President
Overview

• focuses on seamless scalability
• distributed shared-nothing architecture
• “moves the query to the data, not the data to the query”
• full ACID compatibility
• uses Sierra Database Engine
“slicing” data

• separate tables into “slices”
  – replicate slices for redundancy
  – number of slices determined by table size
  – placement of the slices and their replication factor is dynamic
  – determination of data to slices can be range or hash based
Query execution

• Incoming query translated into “query fragments” by Sierra Planner
• fragments are sent to the node that contain the data for that fragment
• page-level locks (i.e. no global table or row-level)
  – data consistency handled by MVCC
Example 1 (read)

Query:
SELECT * FROM T1 WHERE uid=10;

Node 1
Sierra Planner
Session

Node 2
Read(uid=10)
T1

Node 3

Schema:
T1:
uid int
gid int
PRIMARY KEY ("uid", "gid")

Infiband Interconnect

1 2 3 4
Example 2 (join)

Query:
SELECT uid, name FROM T1 JOIN T2 ON T1.gid=T2.gid WHERE uid=10;

Schema:
T1:
uid int
gid int
PRIMARY KEY (uid, gid)

T2:
gid int
name varchar
PRIMARY KEY (gid)
Example 3 (complicated)

Query:
SELECT * FROM T1 WHERE uid<100 AND gid>10 ORDER BY uid LIMIT 5;

Sierra Planner

Sierra Execution Engine

Node 1

Node 2

Node 3

Session

Infiniband Interconnect

T1

T1_1

T1_2

Read(uid<100)

Filter(gid>10)

Limit(5)

Sort

Limit(5)

Read(uid<100)

Filter(gid>10)

Limit(5)

Schema:

T1:
uid int
 gid int
PRIMARY KEY (uid, gid)
Example 4 (insert)

Query:
INSERT INTO T1 VALUES (10, 20);

Schema:
T1:
- uid int
- gid int
- KEY `i` (gid)
- PRIMARY KEY (`uid`, `gid`)

Sierra Planner

Sierra Execution Engine

Infiniband Interconnect
Conclusions

• Benefits
  – distributed database with no tuning necessary
  – balanced I/O
    • Clustrix is purchased by buying Clustrix boxes
    • Infiniband interconnect
  – scale by adding a new Clustrix box
    • near-linear by 3\textsuperscript{rd} party benchmarks
  – “drop in” replacement for MySQL

• Users
  – The Ladders, AOL, PhotoBox