Real World Big Data – Achieving 100k Transactions per Second with a NoSQL Database

Hsiao "Shao" Su – Senior Performance Engineer

Ken Tune – Senior Consultant

Why we are here…

- Customer wants:
  - NoSQL
  - ACID
  - Linear scale
  - High update rate
  - Nearly no text…
- We need to come up with something new…

Outline

- MarkLogic
- New Challenges
- New Techniques – how we addressed these challenges

What is MarkLogic

What is MarkLogic?

- Database
  - Non-Relational
  - ACID
  - MVCC
- Search Engine
  - Web Scale (Big Data)
  - Realtime Update

Data Model

Key	Value
/hello/world	<trade> <trader_id>8</trader_id> <time>2012-02-20T14:00:00</time> <instrument>IBM</instrument> … </trade>
/book5293	It was the best of times, it was the worst of times, it was the age of wisdom, …
/2012-02-20T14:47:53/01445…976	.mp3 .avi [your own binary format]

MarkLogic™

Inverted Index

MarkLogic™

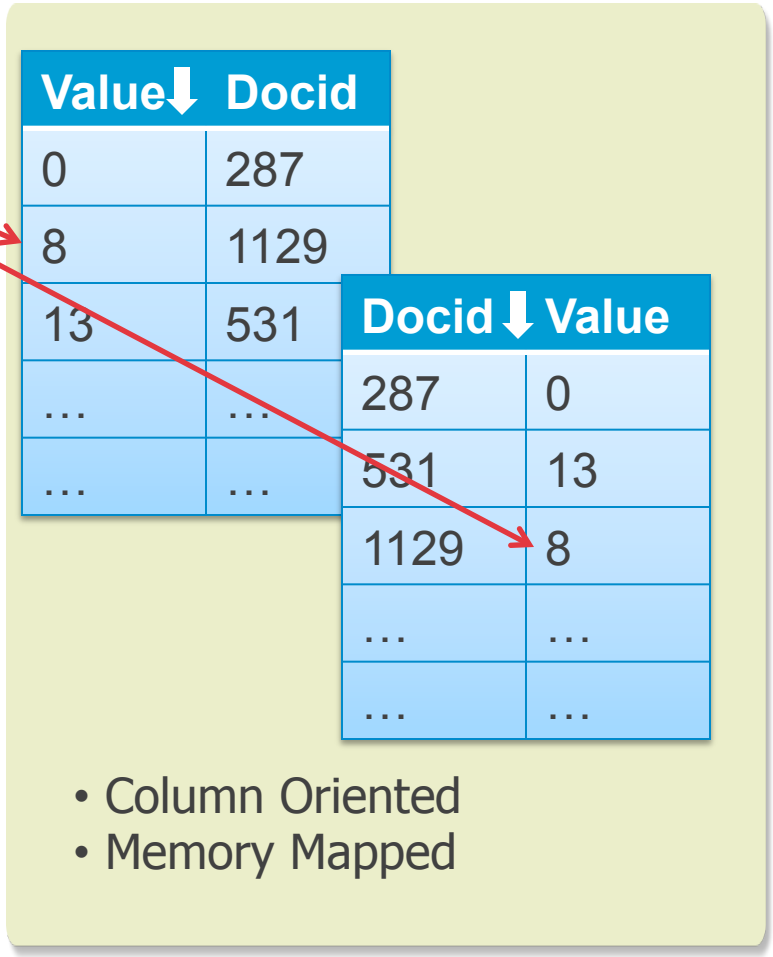
Range Index

Rows

```
<trade>
  <trader_id>8</trader_id>
  <time>2012-02-20T14:00:00</time>
  <instrument>IBM</instrument>
  …
</trade>
```

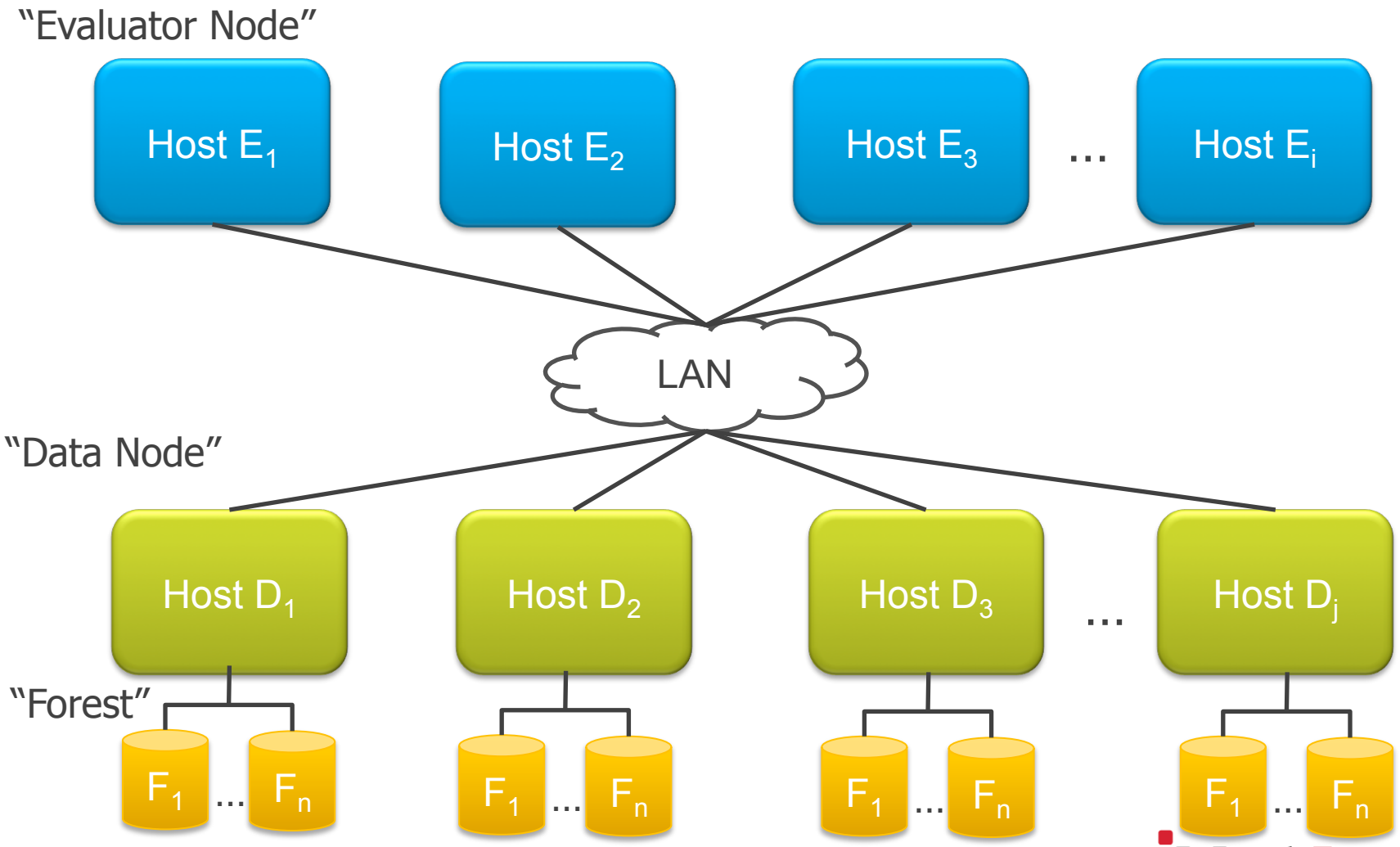
```
<trade>
  <trader_id>13</trader_id>
  <time>2012-02-20T14:30:00</time>
  <instrument>AAPL</instrument>
  …
</trade>
```

```
<trade>
  <trader_id>0</trader_id>
  <time>2012-02-20T15:30:00</time>
  <instrument>GOOG</instrument>
  …
</trade>
```

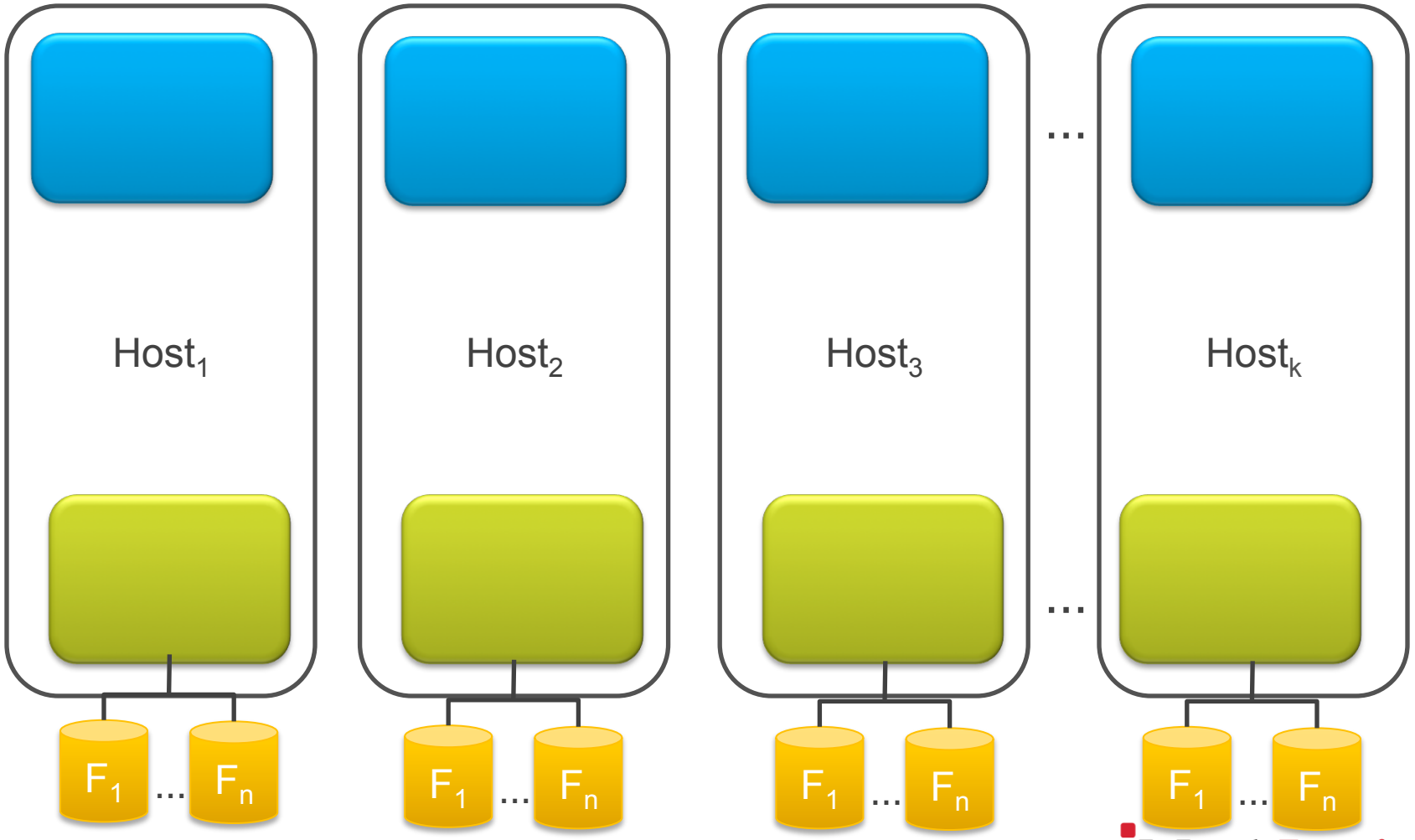


MarkLogic™

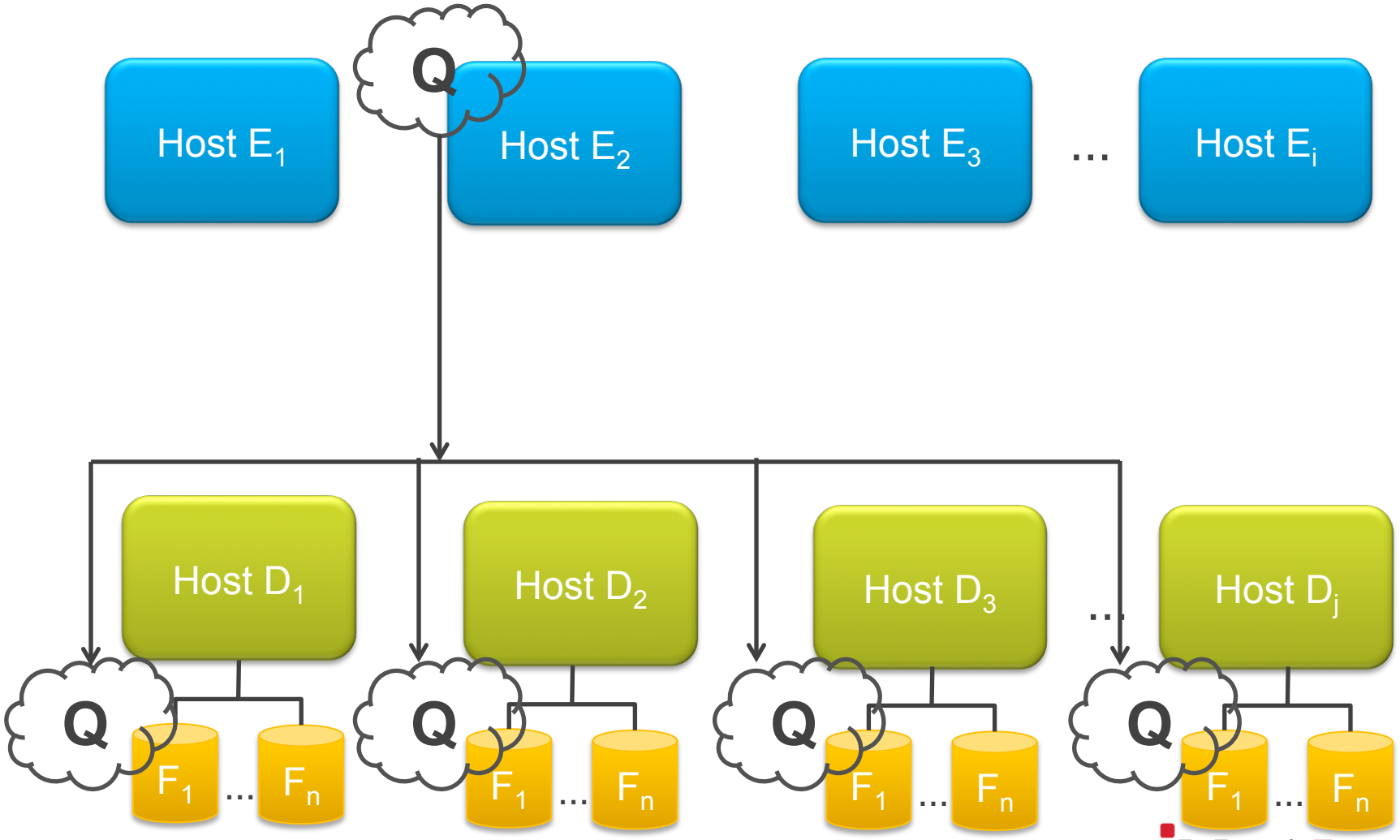
Cluster



Cluster – same binary

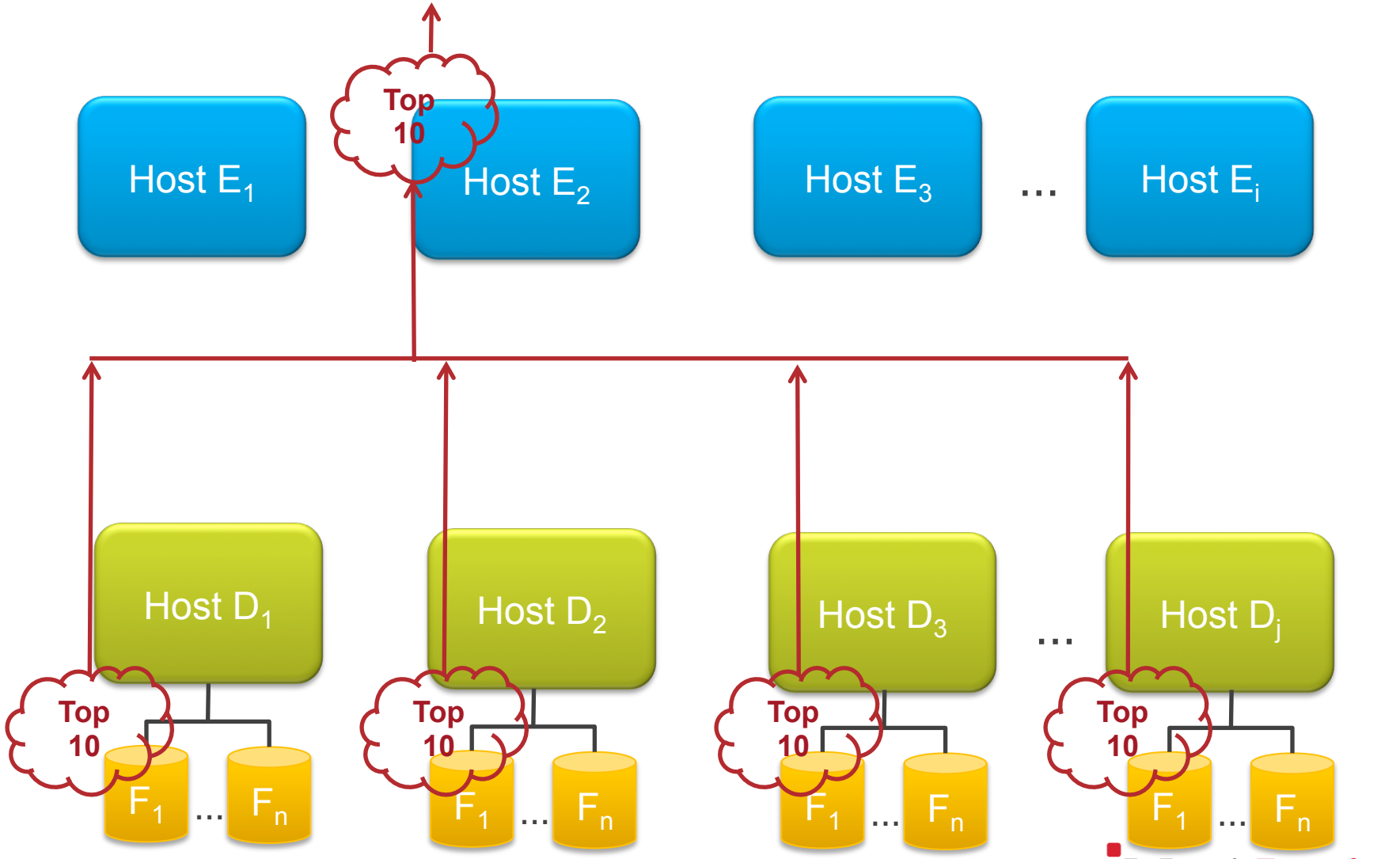


Query Evaluation – "Map"



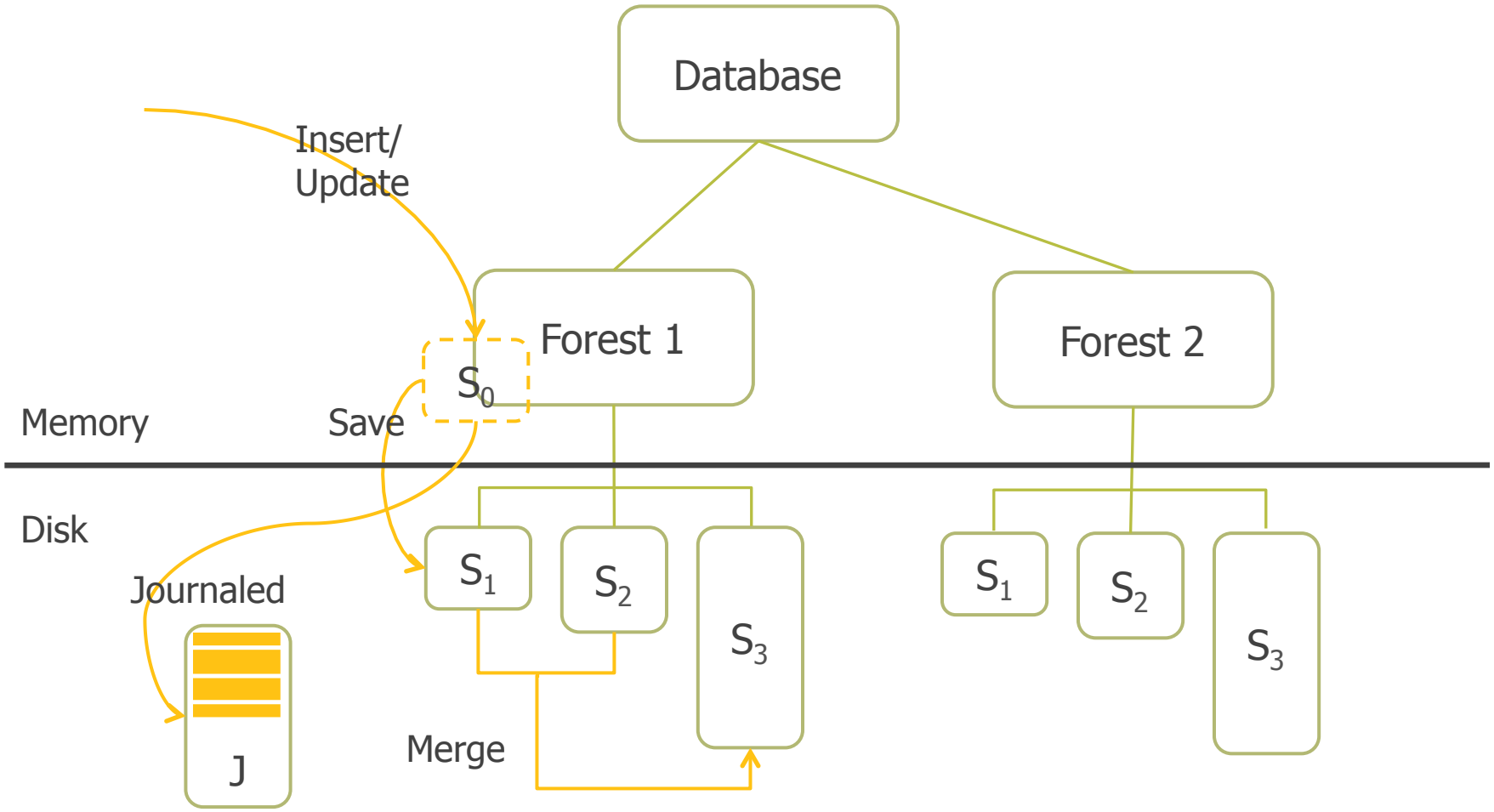
MarkLogic™

Query Evaluation – "Reduce"



MarkLogic™

Save-and-merge (Log Stuctured Tree Merge)



MarkLogic™

New Challenges (Top Tier Bank)

New Challenges (Top Tier Bank)

- Scale – 4.6bn Trades in a Four year window
- Trade complexity
- Increasingly expensive hardware
- Increasing cost of maintenance
- Need to shorten time to market

MarkLogic™

Why did MarkLogic Server make sense?

- Scalable (to petabyte level) architecture
- Schema agnostic and document-centric
- It runs on commodity hardware
- Anecdotal evidence shows the TCO is far lower for a MarkLogic cluster, as is the DBA:node ratio
- It allows far greater agility

MarkLogic™

Other considerations

- ACID
- Enterprise features (real time replication)
- Monitoring capability
- Reliability

One More Thing

- In order to make sure they were truly prepared for the future we had to show them MarkLogic would scale linearly even at the limits of the wildest projections.

- This is why we showed them MarkLogic could cope with an input stream of over 100,000 documents per second.

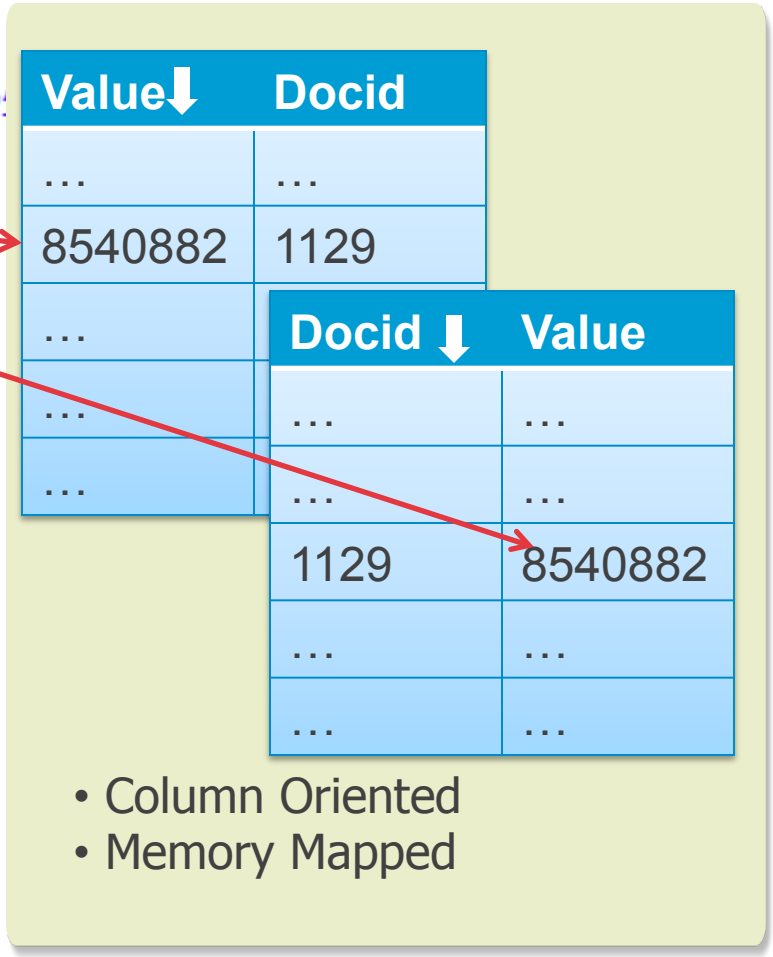
New Techniques – how we addressed these challenges

Techniques

- Range Indexes (for positions)
- Batching
- In-forest Eval

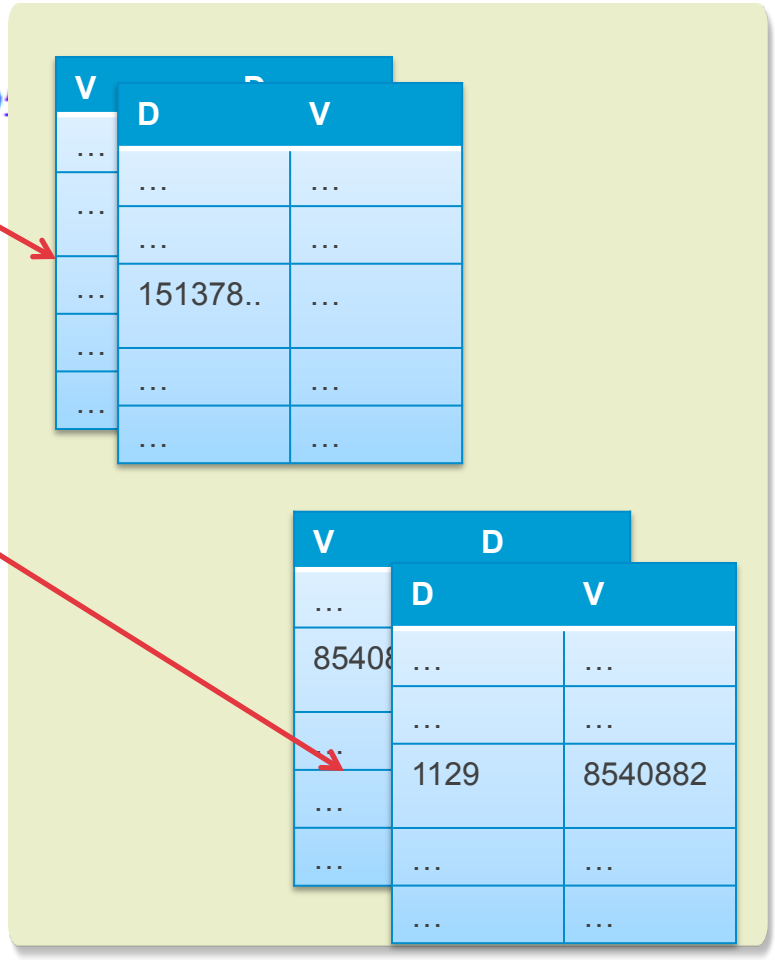
Range Index (for position)

```
- <trade>
    <uri>2011-7-17567379666848873227</uri>
    <rollup book-date-instrument="1513787203889360769
    <quantity>8540882</quantity>
    <quantity2>11193.71</quantity2>
    <instrument>Liz Claiborne</instrument>
    <tradedate>2011-03-13-07:00</tradedate>
    <depot>DM</depot>
    <book>679</book>
    <settledate>2011-03-17-07:00</settledate>
  </trade>
```



Range Index (for position)

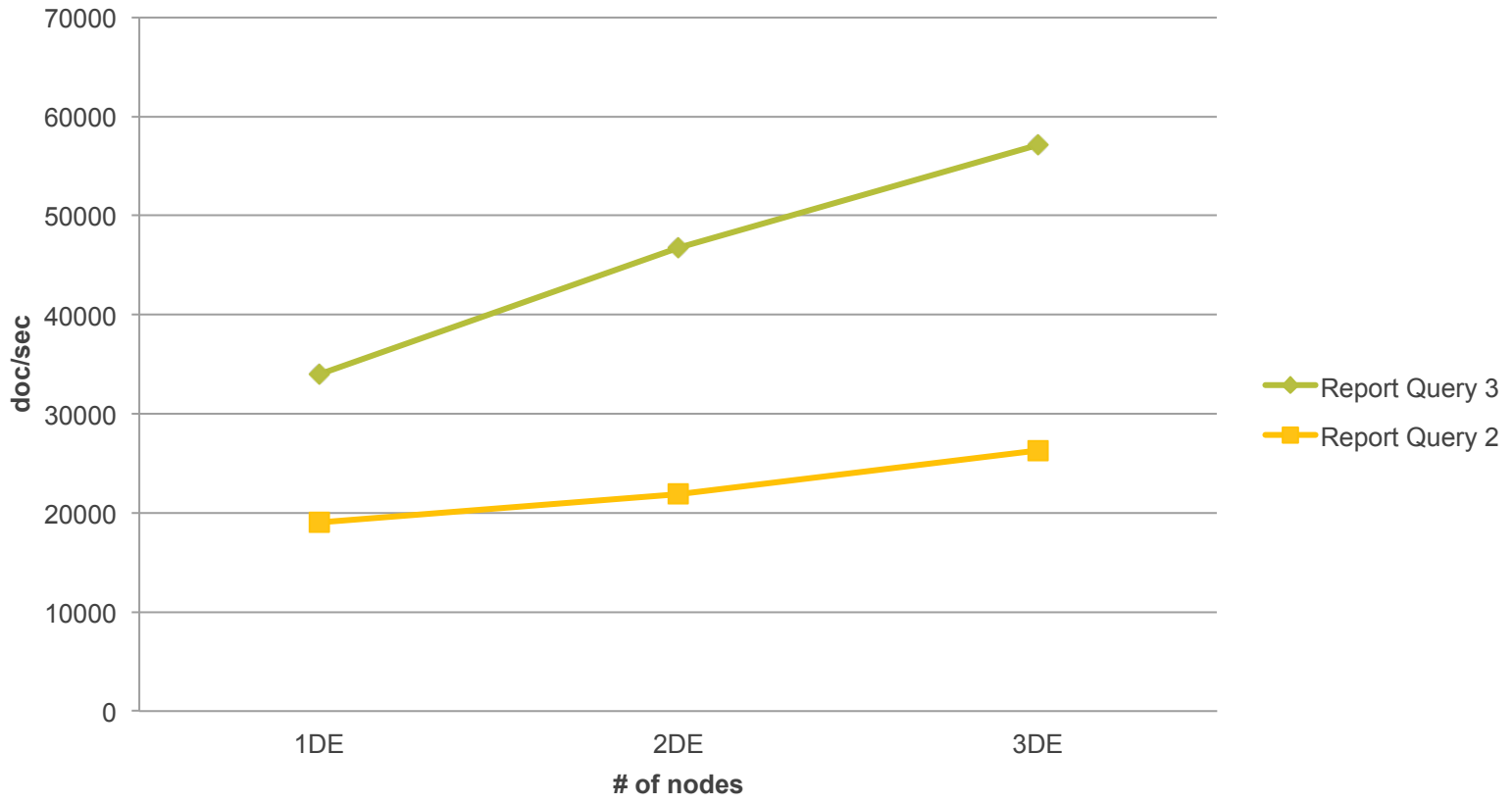
```
- <trade>
    <uri>2011-7-17567379666848873227</uri>
    <rollup book-date-instrument="151378720388936076
    <quantity>8540882</quantity>
    <quantity2>11193.71</quantity2>
    <instrument>Liz Claiborne</instrument>
    <tradedate>2011-03-13-07:00</tradedate>
    <depot>DM</depot>
    <book>679</book>
    <settledate>2011-03-17-07:00</settledate>
  </trade>
```



- Co-occurrences:
  - Find pairings of book-date-instrument and quantity
- Fast Aggregate:
  - sum up the above
- Group-by, in a column-oriented, in-memory database

MarkLogic™

Query 3 (Fast Agg + Co-occurrence) Vs. Query 2

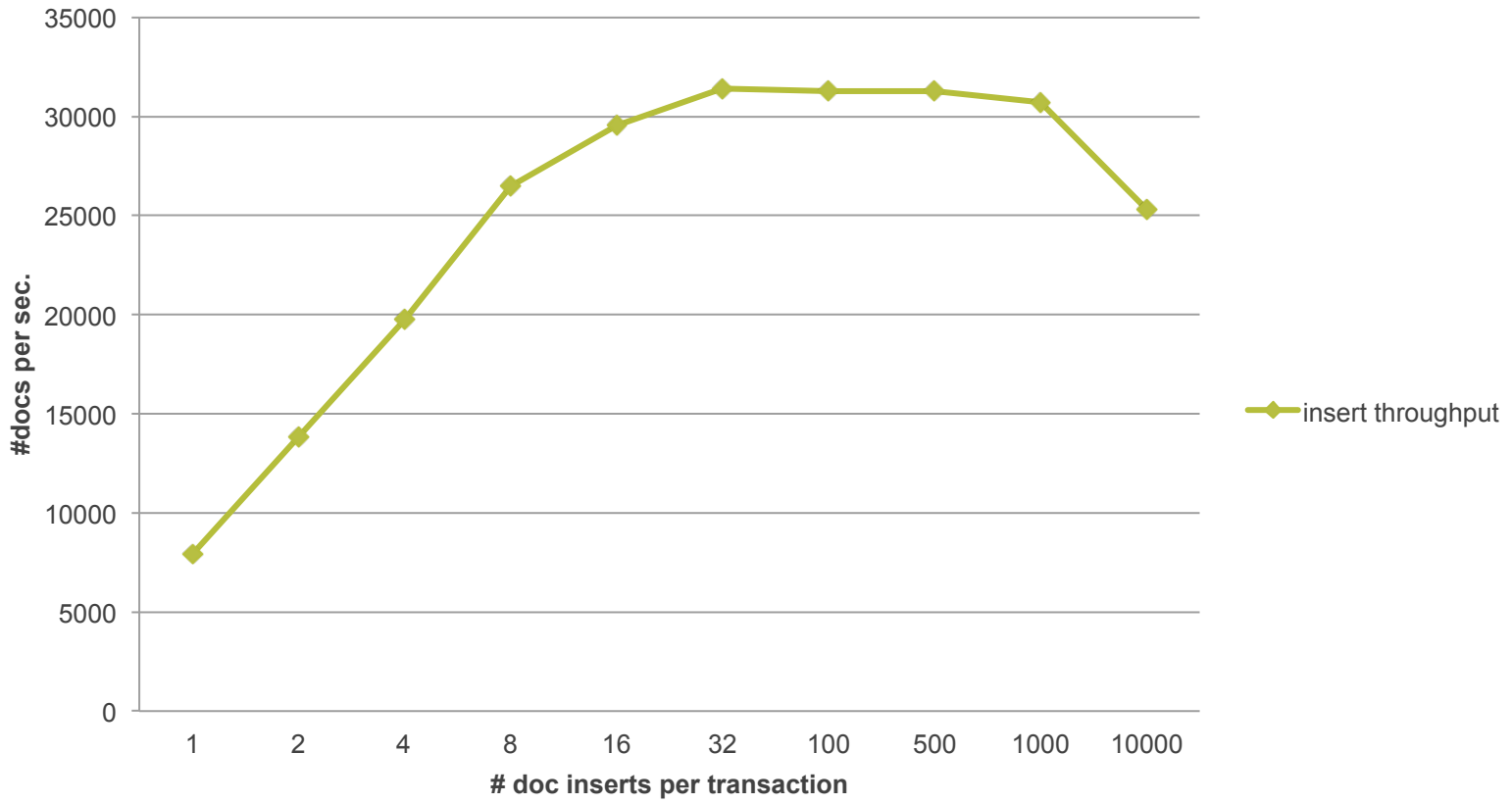


MarkLogic™

Batching

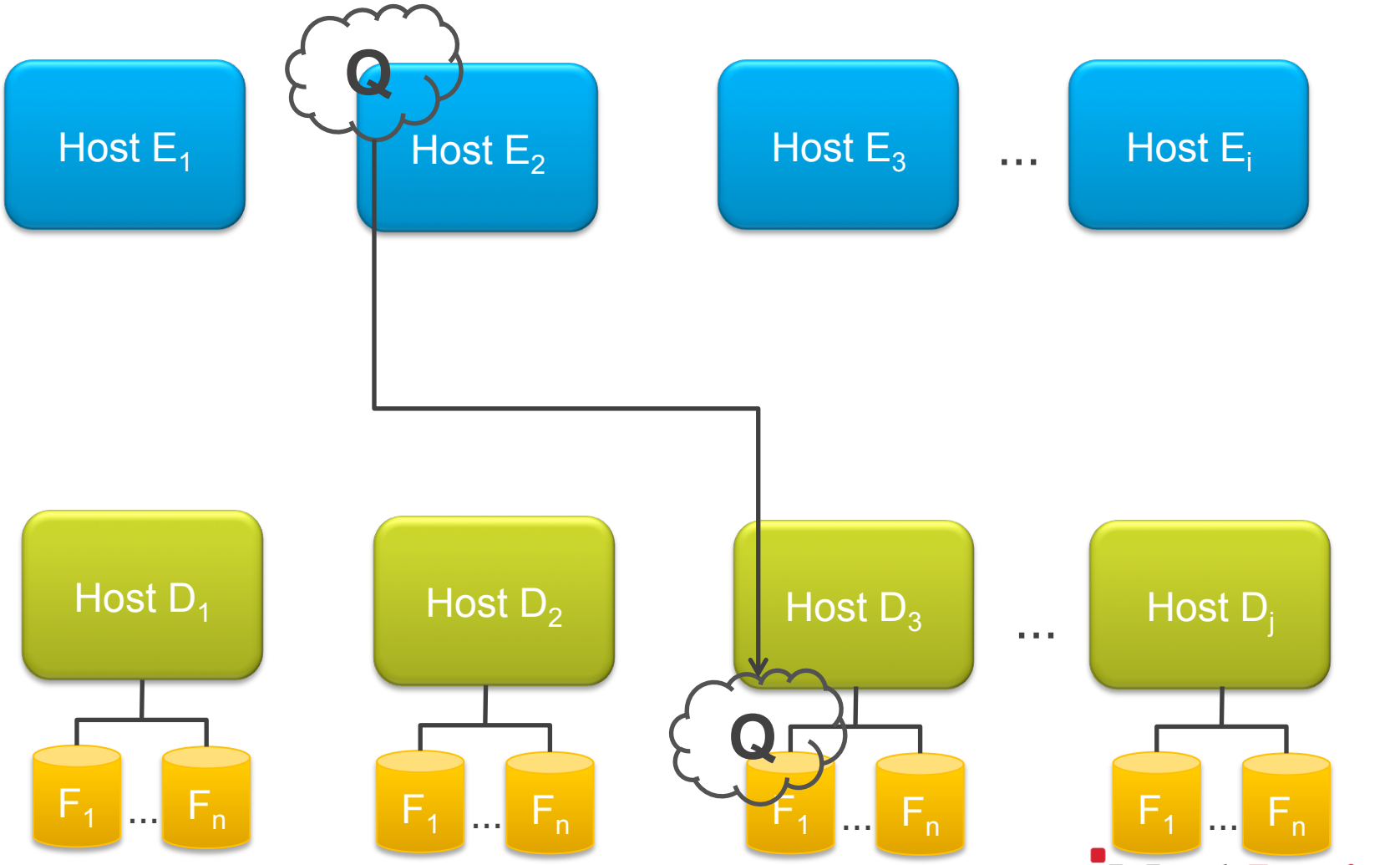
- A Simple Idea
  - Txn cost is high
  - Combine multiple insert/updates within a single txn

Transaction Size and Throughput



MarkLogic™

In-Forest Eval



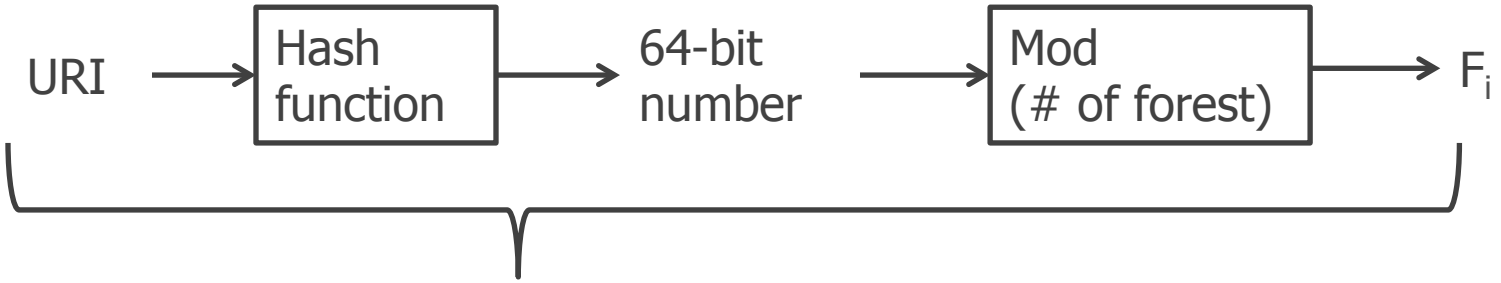
Insert Mechanics

1) New URI+Document arrive at E-node
2) ***URI Probe – determine whether URI exists in any forest**
3) *URI Lock – write locks are created
4) URI Assignment – URI is deterministically placed in Forest
5) Indexing
6) Journaling
7) Commit – transaction complete
8) *Release URI Locks – D nodes are notified to release lock

*** Overhead of these operations increases with cluster size**

Deterministic Placement

4) URI Assignment – URI is deterministically placed in Forest



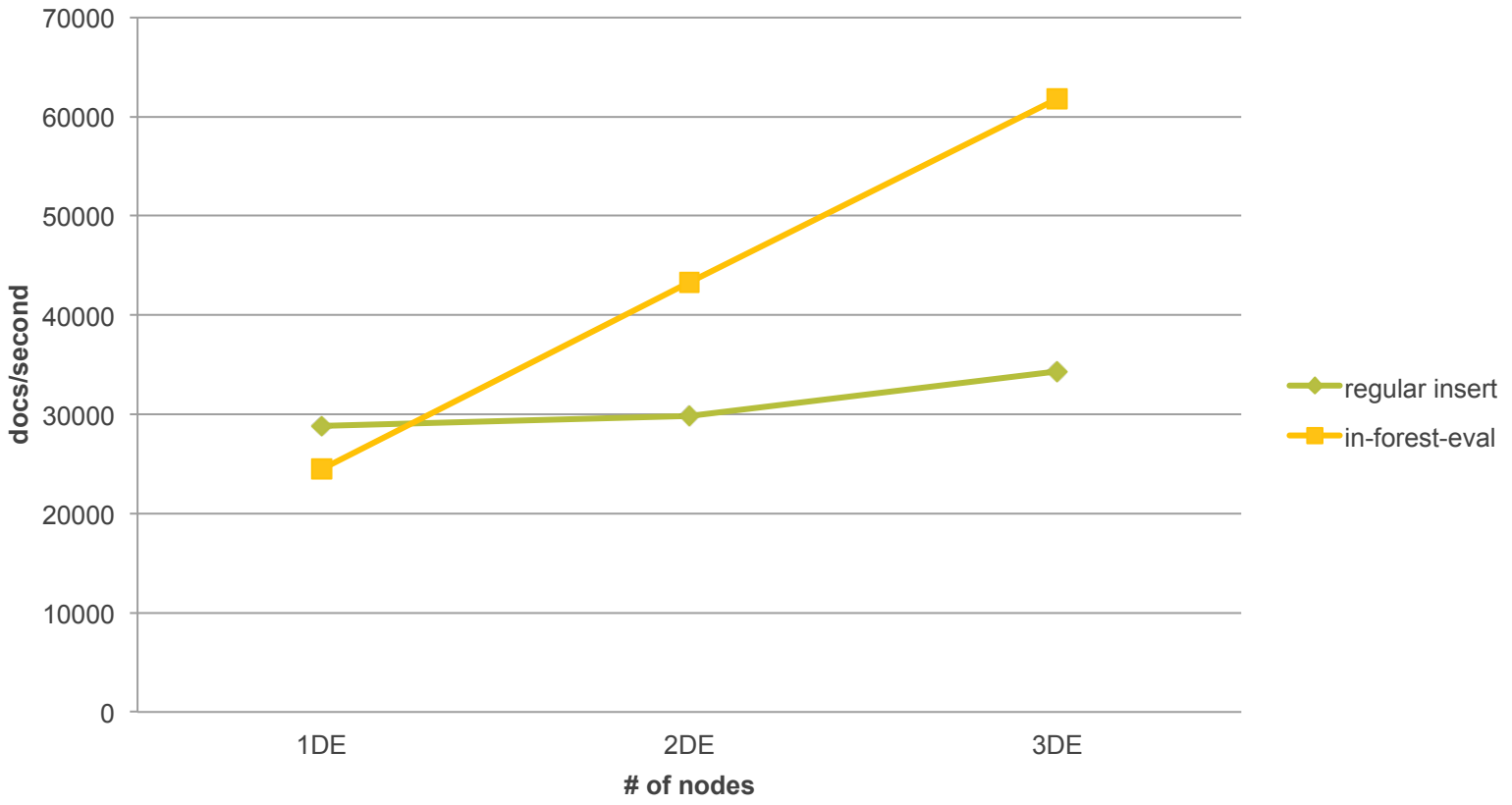
- Done in C++ within server
- But…
  - Can also be done in XQuery outside of the server
  - Also, server does allow queries to be evaluated against only one forest…

Improved Insert Mechanics

1) New URI+Document arrive at E-node
2) Compute F_i using XQuery, ask server to evaluated the insert query only against F_i
3) URI Probe – F_i Only
4) URI Lock – F_i Only
5) URI Assignment – F_i Only
6) Indexing
7) Journaling
8) Commit – transaction complete
9) Lock Release - F_i Only

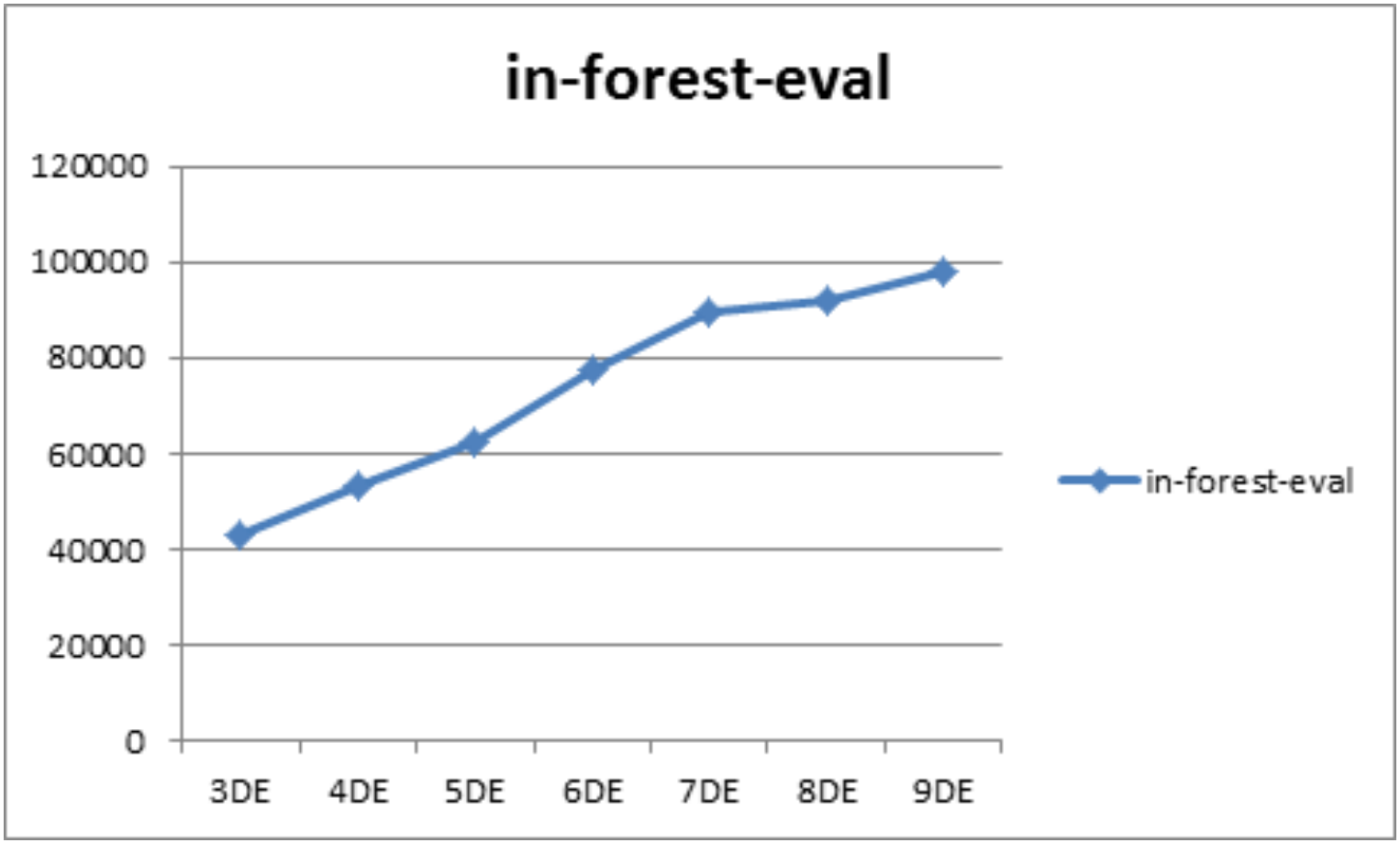
*** It's safe, if all inserts are done this way. (Same hash function, no collision)**

Regular Insert Vs. In-forest-eval



MarkLogic™

Extemporizing 100K (mixed cluster)



Tools for loading data

- InformationStudio
- RecordLoader via XCC
- ***XQSync via XCC**
- XQuery Built Ins
- ***Hadoop Connector**

*** Using the same technique!**

Some Takeaways

- We can do all these:
  - NoSQL
  - ACID
  - High Txn Rate
  - Linear Scale

Thank You!

hsiao.su@marklogic.com
ken.tune@marklogic.com


MarkLogic™

Trades and Positions

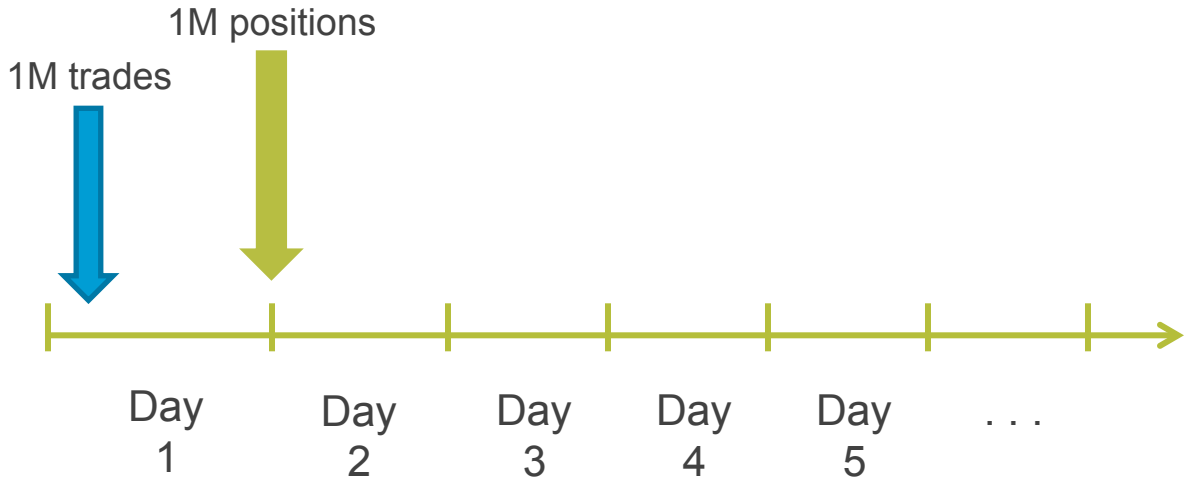
```
−<trade>
    <uri>2011-7-17567379666848873227</uri>
    <rollup book-date-instrument="15137872038893607695" book-date="16677327683893830249"/>
    <quantity>8540882</quantity>
    <quantity2>11193.71</quantity2>
    <instrument>Liz Claiborne</instrument>
    <tradedate>2011-03-13-07:00</tradedate>
    <depot>DM</depot>
    <book>679</book>
    <settledate>2011-03-17-07:00</settledate>
  </trade>

−<position>
    <rollup book-date-instrument="21098070446783181 20"/>
    <uri>P:2011-10-267-Fenner.xml</uri>
    <book>267</book>
    <instrument>Fenner</instrument>
    <businessDate>2011-03-25Z</businessDate>
    <positionDate>2011-03-24Z</positionDate>
    <quantity2>3</quantity2>
  </position>
```

Hash of <book>+<tradedate>+<instrument>

MarkLogic™

Trade Store Simulation

- 1 million trades per day
- followed by 1 million position reports at end of day
  - accumulates the trades of the current day + the prior day position for each "book:instrument" pair



Need to Achieve These…

- High-performance, High volume, Transactional, Mission-critical work load for the world's largest investment bank.
- No text.
- No hierarchy.
- Documents map to rows.
- Linear scale out