Global Scale Web 2.0 Data Platforms
Sharding with Oracle Database

Wei Hu
Senior Director, High Availability Technologies
Oracle Corporation
THE FOLLOWING IS INTENDED TO OUTLINE OUR GENERAL PRODUCT DIRECTION. IT IS INTENDED FOR INFORMATION PURPOSES ONLY, AND MAY NOT BE INCORPORATED INTO ANY CONTRACT. IT IS NOT A COMMITMENT TO DELIVER ANY MATERIAL, CODE, OR FUNCTIONALITY, AND SHOULD NOT BE RELIED UPON IN MAKING PURCHASING DECISION. THE DEVELOPMENT, RELEASE, AND TIMING OF ANY FEATURES OR FUNCTIONALITY DESCRIBED FOR ORACLE'S PRODUCTS REMAINS AT THE SOLE DISCRETION OF ORACLE.
Outline

• What is sharding
  • Use of MySQL in sharding
• Challenges in managing sharded databases
• Why Oracle is best for sharding

• Goals
  • Presentation is targeted at customers who have decided to deploy a sharding architecture
  • It is NOT meant to describe the pluses and minuses of a sharding architecture
Sharding Architecture

- Sharding is an application-managed scaling technique using many (hundreds / thousands of) independent databases
  - Data is split into multiple databases (shards)
  - Each database holds a subset (either range or hash) of the data
  - Split the shards as data volume or access grows
  - Shards are replicated for availability and scalability
- Sharding is the dominant approach for scaling massive websites
  - One social networking site shards across 1800 databases
Sharding Architecture – cont.

- Sharding is used in custom applications that require extreme scalability and are willing to make a number of tradeoffs to achieve it
  - Application code dispatches request to a specific database based on key value
  - Queries are constrained - simple queries on shard-key
  - Data is denormalized to avoid cross-shard operations (no joins)
- This is called Horizontal Sharding
A Variant of Sharding Architectures

- Each database holds all the data
- Request dispatched to a specific database based on read/write, key value
- Updates go to one database, changes are replicated to the other databases. The other databases are available for reads
- Provides read scalability
- Can be combined with horizontal sharding so that each shard is replicated to a different degree
- Main benefit is that you do not need to reshard
Sharding and MySQL

• Database sharding is often deployed in conjunction with MySQL because of its scalability limitations
• In fact, sharding is a database independent architecture
• People become used to the basic functions offered by MySQL and don’t realize what’s possible
  • Smart phones, iPods, and GPS are ‘luxuries’ that one cannot live without after you’ve used them
• Many of Oracle’s capabilities developed to manage large scale data are directly applicable to sharding
• The large number of databases required in sharded architectures *multiplies* the benefits of advanced database technology
Many Customers Shard using Oracle

- Leading online auction site
- Leading online retailer
- Leading software as a service company
- Leading online gaming site
- Leading business networking site
- Leading online media and search company
- Leading online music and video vendor
Sharding Challenges

- Schema Changes Happen
- Failures and Disasters Happen
- Data Corruptions Happen
- Hardware Changes Happen
- Application Changes Happen
- Growth Happens
- Complexity Happens
Schema Changes Happen

- Schema changes are a necessary part of growing & successful websites
- The ideal sharding solution should enable you to change schemas, tables, configurations, .. *everything*, online with high performance, painlessly
- From the status pages of a major social networking site [not running Oracle]
  
  *We’re still seeing some slowness ... We’ll be making some schema changes to our slave databases tomorrow afternoon to help minimize the issue.* [Saturday]
  
  *We’re gathering today to make some minor changes on our slave databases … to help speed queries.* [Sunday]
  
  *We’re going to make a change to the databases which will take the service down for a bit. ... Thanks for your patience!* [Later on Sunday]
MySQL - Offline Schema Changes

- MySQL does not support online schema changes
  - Schema changes are also slow since they make complete copies of affected tables
  - DDLs lock out updates for the entire duration
- The pain is multiplied when you have to make these changes to many databases
- You have a painful choice
  - Take a total outage by making the change to all the shards at once, or
  - Do it shard-at-a-time and take a lengthy rolling downtime
- You still need to
  - Schedule the downtime
  - Be there on Sunday or 2AM to do the work
    - Unless you are global where there is no downtime window
  - Apologize and endure complaints from your loyal user community
Oracle - Online Schema Changes

• With Oracle, many table and index DDL operations are online
  • Create new index, move index, rebuild, and coalesce index
  • Online index build with NO pause to DML
  • Reorganize and redefine tables online
  • Invisible indexes prevent premature use of newly created indexes
  • Online add column
  • Fast add column with default value (no need to update all rows to default value)
  • Add partitions, constraints online

• Oracle schema changes are also fast
  • Everything done in parallel
    • Parallel create/alter table, parallel create/alter cluster, parallel create/alter index, parallel create/alter materialized view, …

• Sleep better at night, reclaim your weekends, spend less time apologizing to your users
Sharding Challenges

- Schema Changes Happen
- Failures and Disasters Happen
- Data Corruptions Happen
- Hardware Changes Happen
- Application Changes Happen
- Growth Happens
- Complexity Happens
Failures and Disasters Happen

- Shards need to be replicated for availability and disaster recovery
- The ideal replicated shard would be easy, keep copies consistent, handle failures automatically, support zero data loss, and offer disaster protection

- Major social networking site down for 12 hours in 2006 due to power outage
- A data center’s backup generators did not start after a power surge knocked out primary power. … Many websites affected…
MySQL Replication is Terrible

- MySQL replication limitations
  - Only async log shipping which can lose data in case of failure
  - Slaves can return inconsistent data
  - Statement based replication has correctness issues & row-based replication is immature
  - Replication is slow (high overhead on each reader, slaves are single-threaded)
  - No support for failover between master (primary) & slaves (backup)
  - Does not handle failure conditions such as missing or damaged logs
  - Storage engine and replication state may become inconsistent after a crash
  - Bringing a failed master back requires copying the database
  - …
  - the slaves get out of sync in totally different ways. … the slaves don’t even match each other after a little while – Baron Schwartz “High Performance MySQL”, 2007
  - All of the scripting, management daemons, and ease of use scenarios will not solve its problems … – Posting on LiveJournal.com
MySQL Customers Engineer Their Own Solutions

- Google had to significantly change MySQL replication to make it work
  - MySQL/Sun did not agree to take back all changes
  - Fork from mainstream with obvious support issues

- Facebook also had to make “improvements” to MySQL replication
Oracle Has Time-Tested Replication

- Oracle supported replication since Rel 7 (1994)
  - Enough time to provide a complete product
  - Enough time to solve the complex timing issues and error conditions that come with distributed systems
  - Enough time to incorporate feedback from thousands of customer deployments
- Active Data Guard provides out-of-the-box replication optimized for failover, disaster recovery, read scalability, and reporting
- Oracle Streams provides highly-customizable, general-purpose replication
Oracle Active Data Guard - For Availability

• Active Data Guard is Complete
  • Single command for planned switchover, automatic unplanned failover, online creation of standby, zero-data loss capability, no rebuild on failover, extensive instrumentation, and graphical UI
  • No custom scripts or software required

• Active Data Guard is Easy to Manage
  • Schema changes and structural database changes (e.g., datafile add/remove) propagate automatically to the standby databases

• Active Data Guard is Fast and Reliable
  • Redo-based replication is deterministic, divergence-free, more complete, & an order of magnitude faster than traditional logical replication
  • Redo apply is done in parallel
  • It works!! Can push hundreds of megabytes of redo per minute, average data gap is less than 1 second (test measurement granularity)
  • Grant McAlister, Principal Database Engineer, Amazon.com
Active Data Guard – Works in the Presence of Failures

- Automatically detects and resolves missing logs
- No need to recreate the primary after failover
  - Data Guard automatically eliminates inconsistent updates that can result in data divergence
- Automatically monitors the state of the primary database
- Automatic failover on preconfigured error conditions (old primary automatically rejoins configuration after recovery)
- Comprehensive checks to detect and prevent data corruptions resulting from disk and I/O failures
Active Data Guard – Also for Scaling Reads

- Can be easily deployed in a reader farm configuration
  - Updates to the primary DB are replicated to multiple reader DBs
- Guarantees readers transactionally consistent view of the data even as newly-replicated changes are being applied
- This makes the reader databases much more useful – for testing, for backups, for reporting, etc.
Oracle Streams for Sophisticated Replication Needs

- Active-Active Replication with Conflict Detection and Resolution
- Can handle subset of data, extra columns, renamed tables, data transformations, DDLs, heterogeneous versions and platforms …
- Deterministic and divergence-free
- High performance parallel apply technology without compromising transactional atomicity or consistency
- Online Instantiation – no quiesce of source required

- we can easily change schema, table and even column names between our replicas without having to do any custom coding — Michael Margetts, DBA, Myriad Genetics
Sharding Challenges

- Schema Changes Happen
- Failures and Disasters Happen
- Data Corruptions Happen
- Hardware Changes Happen
- Application Changes Happen
- Growth Happens
- Complexity Happens
Data Failures Happen All the Time

• With increasing data volume and complex IO subsystems, data corruptions are inevitable

• Low-cost, commodity hardware increases problem
  • Schroeder and Gibson, *Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?*, 2007
  • Elerath, *Hard Disk Drives: the Good, the Bad, and the Ugly!* 2007

• More insidious than outright disk failures are latent data corruptions
Data Corruptions Happen

• The ideal sharding solution should detect corrupted data, prevent them from being written, and repair the data without downtime

• SAN JOSE, Calif. (AP) - Business was buzzing on Internet ... site ... on Monday, but that didn't prevent investors for pushing the company's stock down 18 percent as fallout continued from last week's 22-hour outage

• The site was down for at least 36 hours due to data corruption on the Hosting Company's server. .. it took them forever to get it all straightend out. …sorry for not being there for you. – Online Discussion Forum not running Oracle

• You can replace computers and other hardware. You cannot replace your data
Rapid Recovery from Data Failures

- Oracle Data Guard protects your data against corruptions and lost writes
- Oracle supports UltraSafe mode for even higher data integrity
- Oracle Flashback technologies allow you to recover and repair data without having to restore a backup
- Oracle also offers high performance backup and recovery
  - Incremental backups do not need to scan the data
  - Online block level recovery
## Sharding Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schema Changes Happen</td>
</tr>
<tr>
<td>Failures and Disasters Happen</td>
</tr>
<tr>
<td>Data Corruptions Happen</td>
</tr>
<tr>
<td><strong>Hardware Changes Happen</strong></td>
</tr>
<tr>
<td>Application Changes Happen</td>
</tr>
<tr>
<td>Growth Happens</td>
</tr>
<tr>
<td>Complexity Happens</td>
</tr>
</tbody>
</table>
Hardware Gets Faster and Cheaper

- The ideal sharding solution should take full advantage of state-of-the-art hardware
  - You should be able to pick your hardware based on cost-effectiveness without being constrained by the software

- We are not using the multi-core CPU efficiently, MySQL doesn’t scale that well after 4-8 cores … - Michael ‘Monty’ Widenius, MySQL Fellow, 2008
Oracle Scales on Multicores

- The number of cores per chip is growing
  - Intel just announced 6 core Xeon chips. This means a dual-socket system has 12 cores and quad-socket systems has 24 cores
  - Their roadmap shows even more cores coming
- Oracle has demonstrated scalability all the way to 256-core systems
  - Sun E25K/M9000, HP Superdome, IBM Regatta, Fujitsu PRIMEPOWER
  - 20 years of experience
  - World record SAP SD benchmark: Oracle on a 256-core SMP
- This scalability gives you the architectural flexibility to pick the number of CPUs/cores that is most cost effective for your environment
Oracle Scales to Terabytes of Memory

- Oracle is architected to run with any amount of memory
  - Customers routinely run with hundreds of gigabytes
  - Benchmarks run with terabytes of memory
  - You don’t have limitations such as running out of memory with many active threads or not using enough memory when running with few threads

- All algorithms that require scanning all the memory have been removed
  - Checkpointing, DDLs, LRU, etc.
Mid-Tier Caches get Added

• The ideal sharding architecture integrates seamlessly with mid-tier caches (e.g., Memcached)

• These caches can become inconsistent
  • Direct updates or loads to the database
  • With geographically distributed database and memcache servers, need to update the cache based on replicated changes
  • Need to invalidate cache entries as data changes
    • With MySQL, you must use heavy-weight database triggers to identify all the updated rows, reverse-engineer the primary key from the SQL statements, or change MySQL to log additional data
    • Oracle LogMiner directly returns the primary keys of all changed rows from the redo logs. This can be used to refresh data in front end caches
Sharding Challenges

- Schema Changes Happen
- Failures and Disasters Happen
- Data Corruptions Happen
- Hardware Changes Happen
- Application Changes Happen
- Growth Happens
- Complexity Happens
Application Changes Happen

- Websites start out simple, and become complex. It is very difficult to simulate the workload and access patterns.
- The ideal sharding architecture should allow application changes to be introduced rapidly and safely.

- There are still a few kinks [that we] are trying to iron out. We appreciate your patience as we try to fix these issues. … please come back in a few days. Thanks! — Leading Social Networking Site not running Oracle
Test Using Real Workload

- Oracle Real Application Testing allows you to recreate actual production database workload in test environment
  - No test development required
  - Replay workload in test with production timing
  - Analyze & fix issues before production
  - Can be used for schema and database changes
Comprehensive Performance Diagnostics

• Oracle has a rich and mature set of diagnostics developed over decades
  • Automatic Workload Repository to capture and compare performance statistics over time for troubleshooting
    • What happened last night?
  • Automatic Database Diagnostic Monitor (ADDM) proactively analyzes database for performance troubleshooting
  • Active Session History (ASH) to capture fine-grained information about performance anomalies
  • Also a comprehensive set of wait events, stats, v$ views

• Find the one or two line performance bugs that are just killing your database
Sharding Challenges

- Schema Changes Happen
- Failures and Disasters Happen
- Data Corruptions Happen
- Hardware Changes Happen
- Application Changes Happen
- Growth Happens
- Complexity Happens
Growth Happens

- The number of shards and overall data volume will both increase
- Working with lots of *anything* is inherently difficult
- The ideal sharding architecture should allow you to further partition each shard (database) into operational units for manageability, performance, and availability

*The partitioning implementation in MySQL 5.1 is still undergoing development and is not yet production-ready* – MySQL 5.1 Reference Manual
Manage Scale via Partitioning

- Oracle partitioning has been around for over 10 years and is widely used
- Oracle allows you to divide tables and indexes into multiple partitions that can be separately managed
  - Partitions can be specified in terms of key range values, dates, user-defined functions, etc.
  - Allows you to back up, bulk-load, index, or re-org just a partition instead of the entire shard (database)
- Partitions allow you to scale shards to the next level
  - Ten thousand shards can be cumbersome; but 100 shards each with 100 partitions is quite doable
- Oracle support 1 million partitions per table
Partitioning for Sharding

- A large software-as-a-service site uses partitions as logical shards to make it easier to re-shard across databases
- Additional use cases:
  - Organize data by date so that you can easily move older data to lower cost storage
  - Load data into one partition without affecting the rest of the database
  - Put inactive partitions into read-only mode so they do not need to be backed up
  - Combine partitions from multiple databases together within a single database so that off-the-shelf applications (e.g., reporting) can work against your data
- Since partitioning is application-transparent, it doesn’t affect your overall architecture or load balancing logic
Sharding Challenges

- Schema Changes Happen
- Failures and Disasters Happen
- Data Corruptions Happen
- Hardware Changes Happen
- Application Changes Happen
- Growth Happens
- Complexity Happens
Complexity Happens

- As you grow, your website becomes more complicated
- The ideal sharding architecture should manage many-as-one
Monitoring Many Databases

- Visualization is a proven way to manage large numbers of systems
- Grid Control presents a global view of your entire database farm, with monitoring, alerting, and point-and-click usability
- Help identify affected systems at a glance without having to first process hundreds/thousands of logfiles
  - No need to create custom scripts, telnet to individual machines, build your own tools

Nokia manages 500 *dissimilar* Oracle databases with 5 DBAs
Where Does this Leave Us?

- Sharding across many databases multiplies Oracle’s advantages a-thousand-fold
  - Make changes with no downtime
  - Replicate data for availability, scalability, and disaster recovery
  - Maintain data integrity
  - Provide flexibility to run on the most cost-effective hardware (1..1000 cores), with any size memory
  - Easily test your application before they are deployed
  - Partition your databases for manageability and performance
  - Monitor and manage many systems as one

- Get ahead of your competition by leveraging these capabilities today
- Oracle helps your business move at Web speed
HA Sessions, Labs, Demos From Oracle Development

**Mon, Sep 22**
- 2:30 pm - Database 11g: Next-Gen HA, Moscone South 103

**Tue, Sep 23**
- 9:00 am - Active-Active Data Centers, Moscone South 103
- 11:30 am - Sharding with Oracle, Moscone South 302
- 11:30 am - HA with Oracle VM, Moscone West 3024
- 1:00 pm - Active Data Guard, Moscone South 104

**Wed, Sep 24**
- 9:00 am - Fusion Middleware Grid HA, Marriott Nob Hill AB
- 11:30 am - RMAN Best Practices, Moscone South 103
- 1:00 pm - Database in the Cloud, Moscone South 305
- 5:00 pm - Data Guard & Real Application Testing, Moscone 102

**Wed, Sep 24 (contd.)**
- 5:00 pm - EM in Secure MAA, Moscone West 2001
- 5:00 pm - E-Business Suite HA, Moscone West 2002/04

**Thu, Sep 25**
- 9:00 am - Oracle Secure Backup, Moscone South 102
- 10:30 am - Streams Replication, Moscone South 102
- 12:00 pm - Rolling Database Upgrades, Moscone South 103
- 1:30 pm - Streams Performance, Moscone South 102
- 3:00 pm - Oracle Grid Computing, Moscone South 303
- 3:00 pm - E-Business Suite R12 MAA, Moscone West 2007
- 3:00 pm - Siebel MAA, Moscone South 308
- 3:00 pm - Fusion SOA HA & Scalability, Marriott Salon 14/15

**Hands On Labs - Thu, Sep 25**
- 10:30 - 11:30 am, 12:00 - 1:00 pm - Active Data Guard, Marriott Golden Gate A3

**DEMOgrounds, Mon-Thu**
- Active Data Guard, Streams, Oracle Secure Backup, RMAN/Flashback, MAA