DMF DB Replacement

Arun Ramakrishnan
Tech Lead, Storage Software

sgi

RDM (RAIMA Data Manager)

- Record based data-store maintained by SGI.
- Maximum of 4 billion stored records with 1 copy and 2 billion with 2 copies of data.
- Supports custom DDL with composite primary keys.
- Snapshots are supported but involve physical copy of table data.
- Indexes are used to locate a record quickly.
- Custom query language implemented using dm*adm tools.
- Data can be imported/exported in text format.
- Dmdbcheck is used to verify database consistency.

RDM contd ...

- Allows concurrent reader and single writer access at the database level.
- No explicit support for foreign keys, table joins etc.
- Multiple record operations can be batched together to improve throughput.
- Operations are journaled in order to facilitate recovery in the event of a crash.
- Single server, no replication.
- DMF data maintained across daemon, volume and chunk databases.

DB Choices

- MySQL
  - Not BSD licensed. Issues with redistribution in a commercial product.
  - Oracle
- NoSQL databases
  - Raima usage model needs classic SQL semantics
  - So port to NoSQL cumbersome & disruptive
- Postgres!!

PostgreSQL : New DB Stack

- Fundamentally scalable design.
  - 50 billion objects (tested)
  - 4 TB table size.
  - 1 TB indexes.
  - 32 concurrent clients aggregating 200 TPS on ultra large databases.
  - Fully ACID complaint with tuple level locking.

PostgreSQL : New DB Stack contd...

- Rich abstraction of data storage
- Data model specified using rich DDL.
- Support for common C oriented data types.
- Support for esoteric data types
  - IP addresses
  - UUIDs (used to store BFIDs)
  - Arrays
  - Timestamps (used to store time-related fields)
  - Key/Value pairs

PostgreSQL : New DB Stack contd...

- Rich support for hot (online backups)
  - Dedicated backup server support for backups.
  - Backups dont block concurrent access to the database.
  - Streaming replication of journals to remote locations for DR.
  - PiTR (Point In Time Recovery) for incremental hot backups.
  - Recovery to any point in time from last good base backup.

- Extensive support for high availability.
  - Full support for warm standby with read-only query support on the standby.
  - Support for multiple levels of cascading replicas.

PostgreSQL : New DB Stack contd...

- Language agnostic data abstraction model.
  - Stored procedures can be written in SQL, PL/pgSQL, Python, R and even Javascript.
  - Extensive support for ODBC and JDBC clients.
- Rich ecosystem of tools and plugins.
  - PhpPGAdmin : Web based DB monitoring and administration.

PostgreSQL : New DB Stack contd...

- Barman : Hot incremental backup and recovery.
- Repmgr : Master / Slave clustering of databases.
- Postgres-XC : Experimental distributed multi-master db clustering.

PostgreSQL : DMF Implementation

- Concurrent operation with RAIMA
  - All db write operations (insert, update and delete) executed on both RAIMA and PostgreSQL.
  - Queries are satisfied by PostgreSQL for scalability with support for complex data joins.
  - DMF code interfaces to db preserved to minimize non database code impact.
  - Stored procedures hide the data layout complexity and return simple tuples to users with same columns as RAIMA.
- Support for additional use cases.
  - Additional fields added to Postgres tables to support non DMF use cases and analytics.
  - Prevent locking of entire database and maintain availability during all normal dmf operations.
  - Support logical snapshots of data to avoid redundant tuple copies.

PostgreSQL : DMF Implementation status

- Dmscanfs recursive mode enhanced to capture paths, handle and bfid in the core database.
- Dmhdelete ported to use direct logical snapshots of the database.
- Dmaudit being ported over to avoid redundant data copying and also scale to larger managed object counts.
- Misc user tools like dmvoladm, dmdump etc ported to support Postgres.

PostgreSQL : DMF Implementation contd...

- Optimize data layout.
  - Data is normalized to prevent duplication of column values across tables.
    - Object attributes like bfid, inode number and size are stored in a single table.
    - Chunk table refers to these objects by means of foreign keys.
  - All library server tables stored in a single database cluster with logical namespaces.
  - Large columns are automatically stored in compressed format.
  - Indexes maintained on frequently used query columns.

PostgreSQL : DMF Implementation status

- Dmscanfs recursive mode enhanced to capture paths, handle and bfid in the core database.
- Dmhdelete ported to use direct logical snapshots of the database.
- Dmaudit being ported over to avoid redundant data copying and also scale to larger managed object counts.
- Misc user tools like dmvoladm, dmdump etc ported to support Postgres.

PostgreSQL : DMF QA Status

- Full stims with concurrent Postgres and Raima access running in pDMF cluster since Oct' 13.
- Integrated backup scheme support added to backup Postgres schemas and database clusters.
- Support scripts developed to import existing customer RAIMA databases into Postgres automatically.
- Normal DMF admin operations like tape merges, hard deletes and dskfrees tested in detail on the new cluster.