



Oracle on HP Storage



Jaime Blasco

EMEA HP/Oracle CTC
Cooperative Technology Center

Boeblingen - November 2004

ORACLE®



© 2003 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.

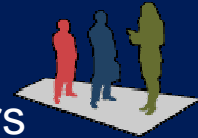
Agenda



- Oracle recommendations on HP Storage
- 10g Automatic Storage Management tests
- 3TB/hour Online Backup



- Located at HP in Boeblingen/Germany, Sophia Antipolis/France and Oracle Reading/UK
- Oracle and HP employees in one team
- Founded in spring 1994
- Technology consulting for partners / customers
- Evaluation and tests of new products/features
- Technology transfer to and from US labs
- Technical pre-sales assistance
- Customer specific Oracle Database, Oracle AS and Oracle E-Business Suite benchmarks



Agenda

- Oracle recommendations on HP Storage
- 10g ASM tests
- 3TB/hour Online Backup



- sort database objects by size and expected I/O volume
- implement high-volume objects in striped volumes with various stripe-widths
- squeeze other objects in where they fit
- monitor performance of individual disks and objects
- play 'chess' with file placement to find best performance

'SAME' technique

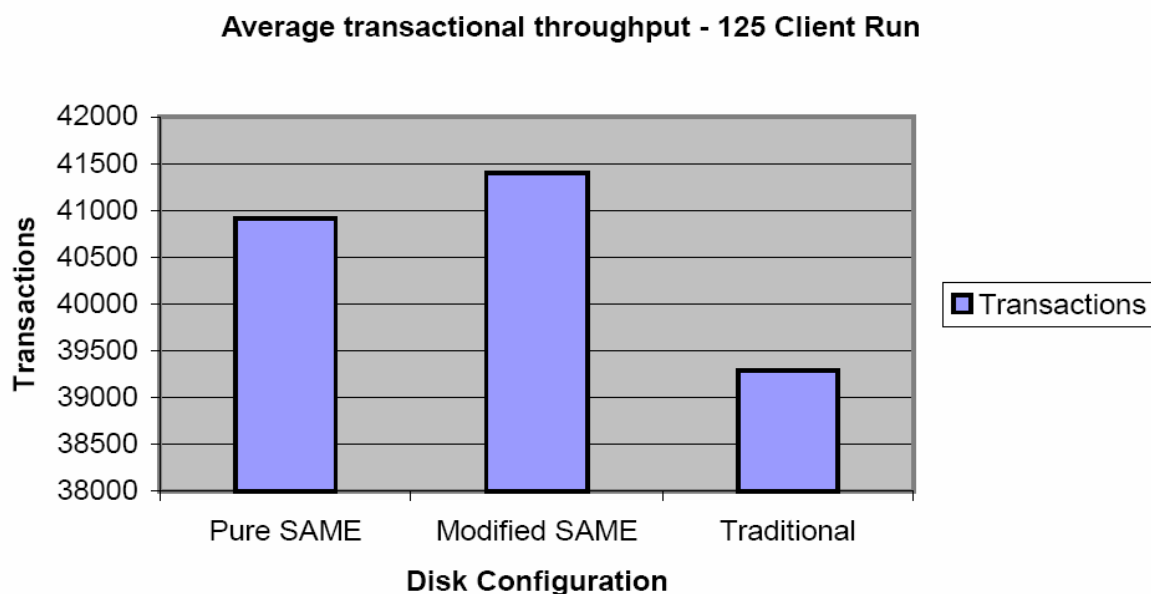
(proposed by Oracle's Juan Loaiza)

Stripe And Mirror Everything

- stripe all files across maximum # disks
- use 1MB stripe size
- use mirroring for high availability
- place "hot" files on outer edge of disks
- keep it simple

- large I/Os minimize impact of disk head movement
- outer edges of disk have larger capacity and greater transfer speed
- very wide stripe-set allows full I/O throughput capacity to help all transactions
- no need to consider characteristics of individual files/tables/transactions

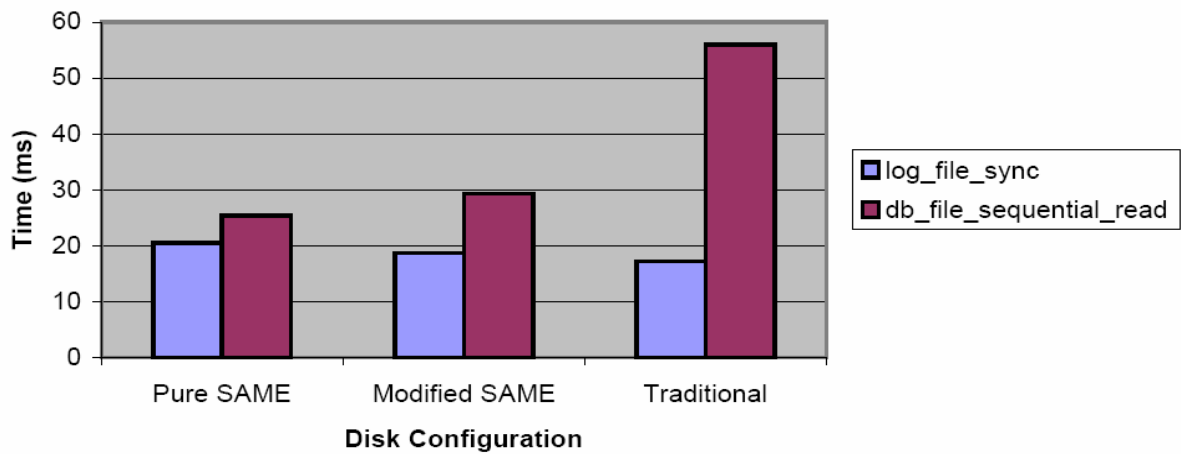
SAME Methodology on HP Storage



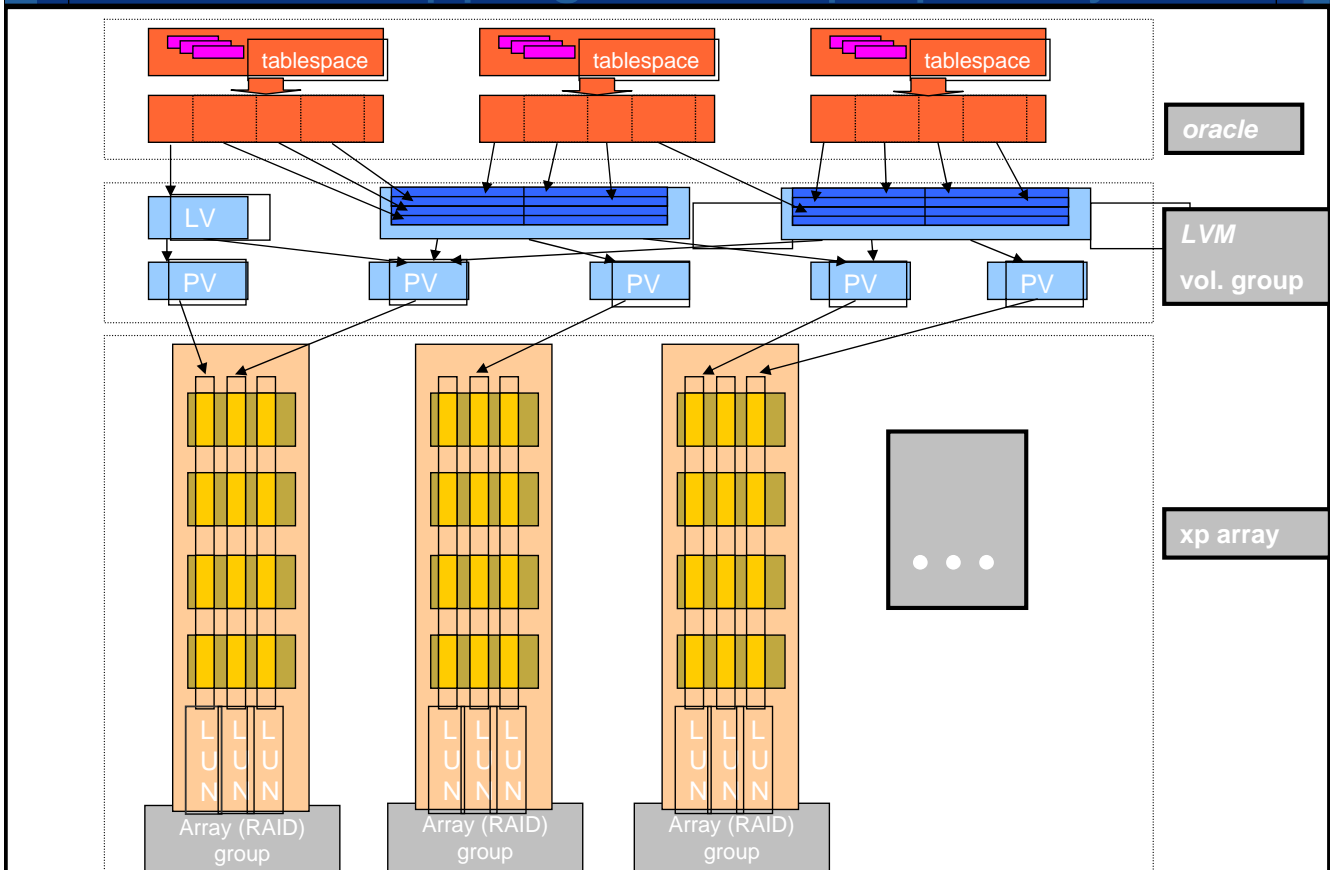
SAME Methodology on HP Storage



Average I/O Response Times - 125 Client Run



PV mapping to the hp xp array



- decide “stripeset width” (# array groups)
 - four to eight AGs recommended
 - span ACP pairs
 - same disksize/RAID
- create one volume group per stripeset
 - use LUSE to coalesce LUNs (reduce # devs)
- create multiple logical volumes per VG
- (create filesystems on top of LVs)
- allocate Oracle objects among stripesets
- map Oracle files to LVs or to filesystem files

RAW or File System

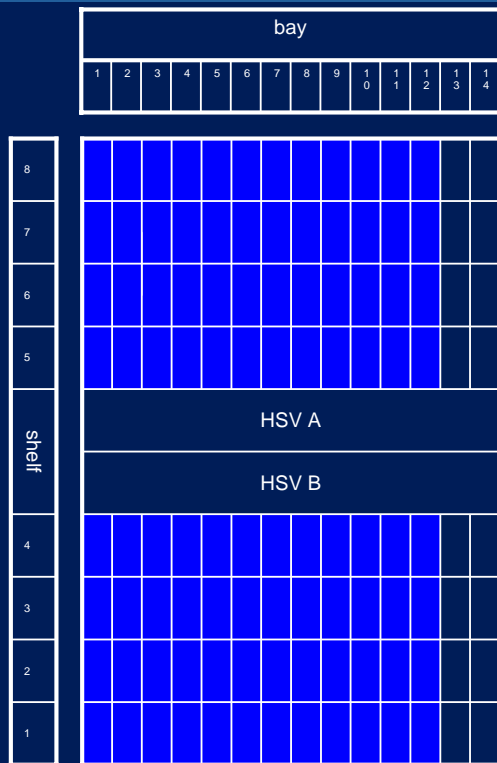
- raw performance
 - double-buffering
 - asynchronous writes (one DBWR process)
 - no unnecessary file overhead
 - no unnecessary file locking
- File System:
 - hp-ux buffer cache size
 - Modified VXFS mount options:
 - + Undo segments and Temp datafiles:
mincache=dsync, convosync=dsync
 - + Datafiles and Redo log files:
mincache=direct, convosync=direct

- Configure enough LUNs to allow I/O load balancing
- Redo logs: written sequentially and in a flush → should be written to the array cache
- Use a dedicated LUN (different VG) for the redo logs to avoid the log writer waiting on the same SCSI-queue with other I/O requests
- Archive logs physically separated from dbf and redo logs (security and performance)
- Oracle colleagues approach: 10 spindles/CPU
- Stripe size: 64k, 256k, 1M, 4M, 16M, 32M (?)

SAME: Benchmarks best practices 5 steps approach

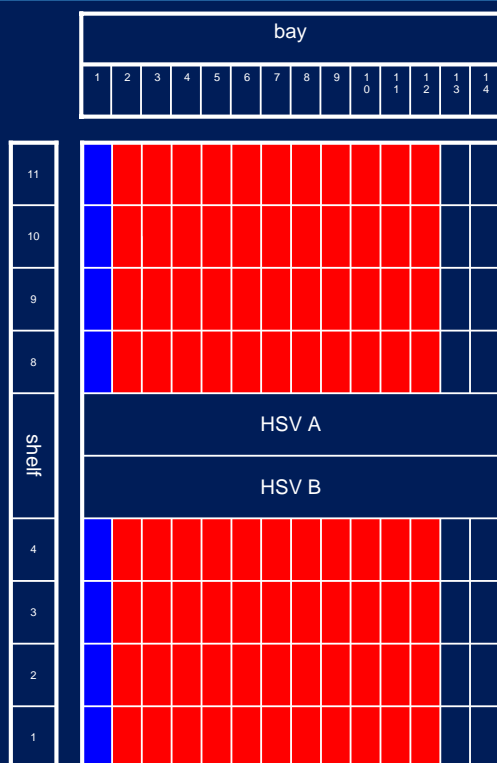
1. VG1 over at least 4 I/O channels
2. VG2 for redo logs (queue depth) over same channels
3. VG2 for redo logs over new I/O channels
4. Put the redo logs on a JBOD (15 disks)
5. Use Cache LUN for redo logs

EVA layout: one-disk-group configuration



- All database components are in the same disk group
- Multiple Oracle installations are in the same disk group

EVA layout: two-disk-group configuration, parallel installations



•Oracle

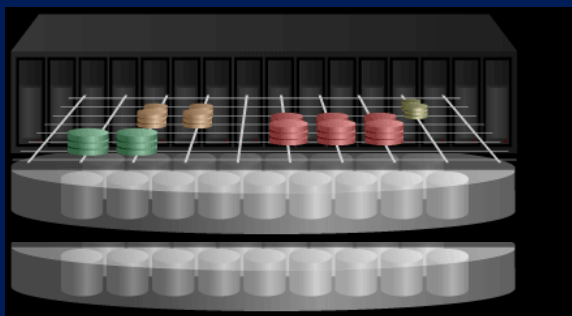
Archived redo
 Mirrored online redo
 Executables

Database data
 Original online redo

- Oracle recommendations on HP Storage
- 10g ASM tests
- 3TB/hour Online Backup



Oracle 10g ASM Automatic Storage Management



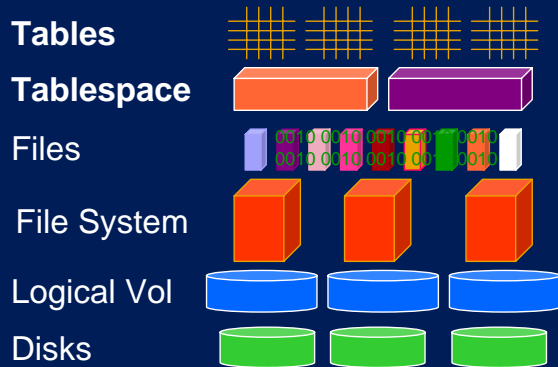
- Dynamically Provision and Tune Storage
- Portable, high performance storage system
- Eliminates need for conventional file system and volume manager
- Automatic mirroring
- Automatic I/O tuning
 - Stripes data across disks to balance load



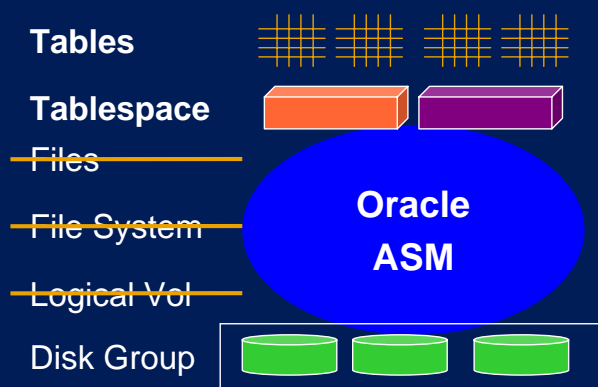
The Operational Stack



TODAY



ASM



“The best way to lower mgmt costs is to remove complexity”

Traditional vs ASM - Setup



- | | |
|--|--|
| <ol style="list-style-type: none">1. Determine required storage capacity2. Install Volume Manager, File System3. Architect data layout to avoid hot spot4. Create logical volumes5. Create file systems6. Install database7. Create database | <ol style="list-style-type: none">1. Determine required storage capacity2. Install ASM3. Create Disk Groups4. Install database5. Create database |
|--|--|

- 3 different environments for the tests
 1. Database using ASM
 2. Database using RAW devices and HP's Logical Volume Manager (LVM)
 3. Database using File System with default and modified mount options
- Hardware Setup:
 - HP server rp7410 8 CPU's (750Mhz), 8GB RAM
 - HP Storage Virtual Array VA7410 with 90x73GB disks
- Software Setup:
 - Oracle Database 10g beta 2 (10.1.0.1.0) on HP-UX 11i

Oracle 10g ASM test: setup

- For each database we used 4 LUN's with 150GB
- LVM stripe size = ASM default template:
 - 1MB for datafiles / temp files / undo
 - 128KB for control files and redo logs
- loaded data was TPC-C like in a 300 GB tablespace
- Goal: 3 environments as comparable as possible, not necessarily tuned
- Test queries:
 - select count (*) with parallel degree 8 and full scan hint
 - create table with parallel degree 8 as select degree 8

Oracle 10g ASM test: results



| | RAW | |
|------------------------|-----------|--------------|
| | Peak MB/s | Time Min:sec |
| Select full scan | 180 | 4:37 |
| Create table full scan | 200 | 0:51 |

Oracle 10g ASM test: results



| | RAW | | FILE default | |
|------------------------|-----------|--------------|--------------|--------------|
| | Peak MB/s | Time Min:sec | Peak MB/s | Time Min:sec |
| Select full scan | 180 | 4:37 | 110 | 5:40 |
| Create table full scan | 200 | 0:51 | 145 | 1:02 |

Oracle 10g ASM test: results



| | RAW | | FILE default | | FILE opt | |
|------------------------|-----------|--------------|--------------|--------------|-----------|--------------|
| | Peak MB/s | Time Min:sec | Peak MB/s | Time Min:sec | Peak MB/s | Time Min:sec |
| Select full scan | 180 | 4:37 | 110 | 5:40 | 140 | 4:52 |
| Create table full scan | 200 | 0:51 | 145 | 1:02 | 140 | 1:12 |

Oracle 10g ASM test: results



| | RAW | | FILE default | | FILE opt | | ASM | |
|------------------------|-----------|--------------|--------------|--------------|-----------|--------------|-----------|--------------|
| | Peak MB/s | Time Min:sec | Peak MB/s | Time Min:sec | Peak MB/s | Time Min:sec | Peak MB/s | Time Min:sec |
| Select full scan | 180 | 4:37 | 110 | 5:40 | 140 | 4:52 | 180 | 4:34 |
| Create table full scan | 200 | 0:51 | 145 | 1:02 | 140 | 1:12 | 200 | 0:50 |

- Impressive performance & manageability
- ASM Performance equal to RAW
- ASM much better than file system (optimized)
- Additional unique ASM features

Agenda

- Oracle recommendations on HP Storage
- 10g ASM tests
- **3TB/hour Online Backup**



HP achieves world record backup and restore



3.16TB/hr Oracle online backup

1.23TB/hr Oracle restore



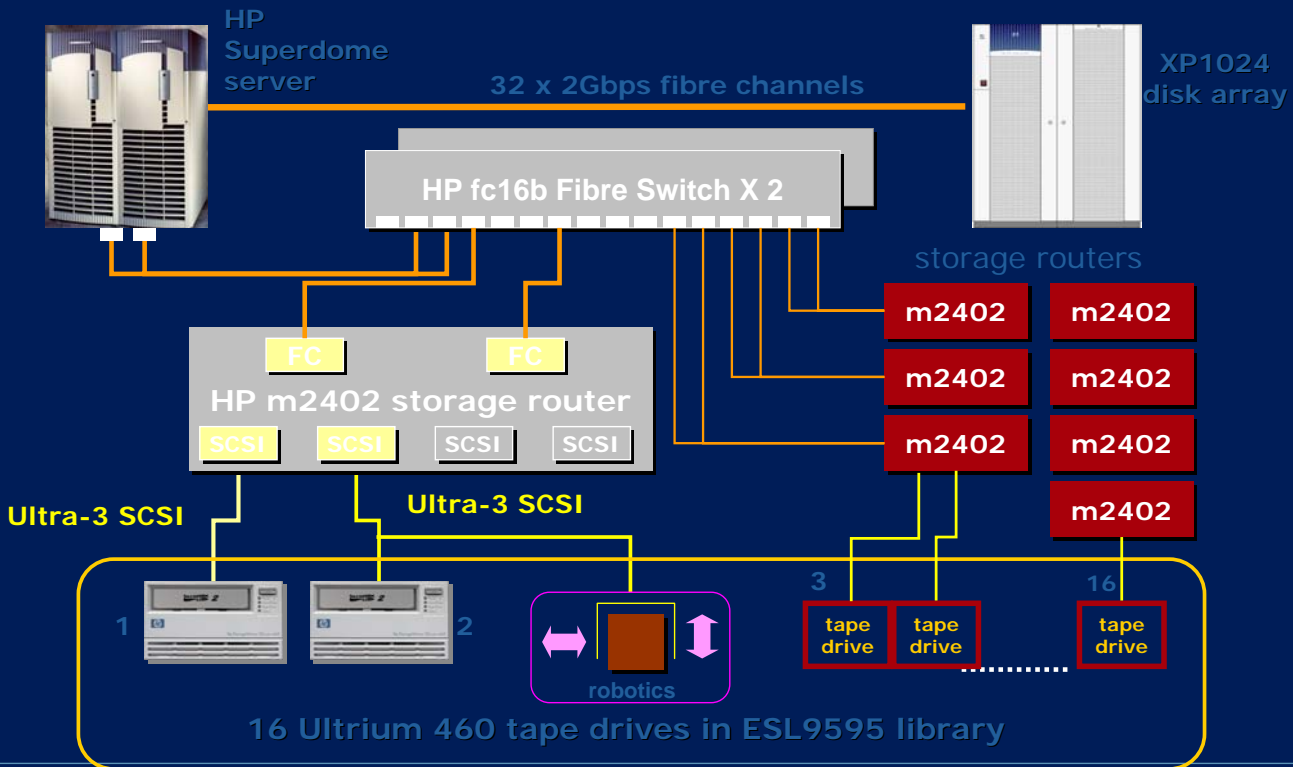
achieved with the following equipment:

- Superdome with 32CPU's
- XP1024
- HP Tape library (ESL9595) with 16 Ultrium2 drives
- Oracle 9iR2 (9.2.0.2.0) for recovery catalog and target database
- OpenView Data Protector 5.0
- 4.3TB database on file system using the TPC-H standard setup: 8 tables partitioned over ~150 Tablespaces and ~ 260 datafiles.

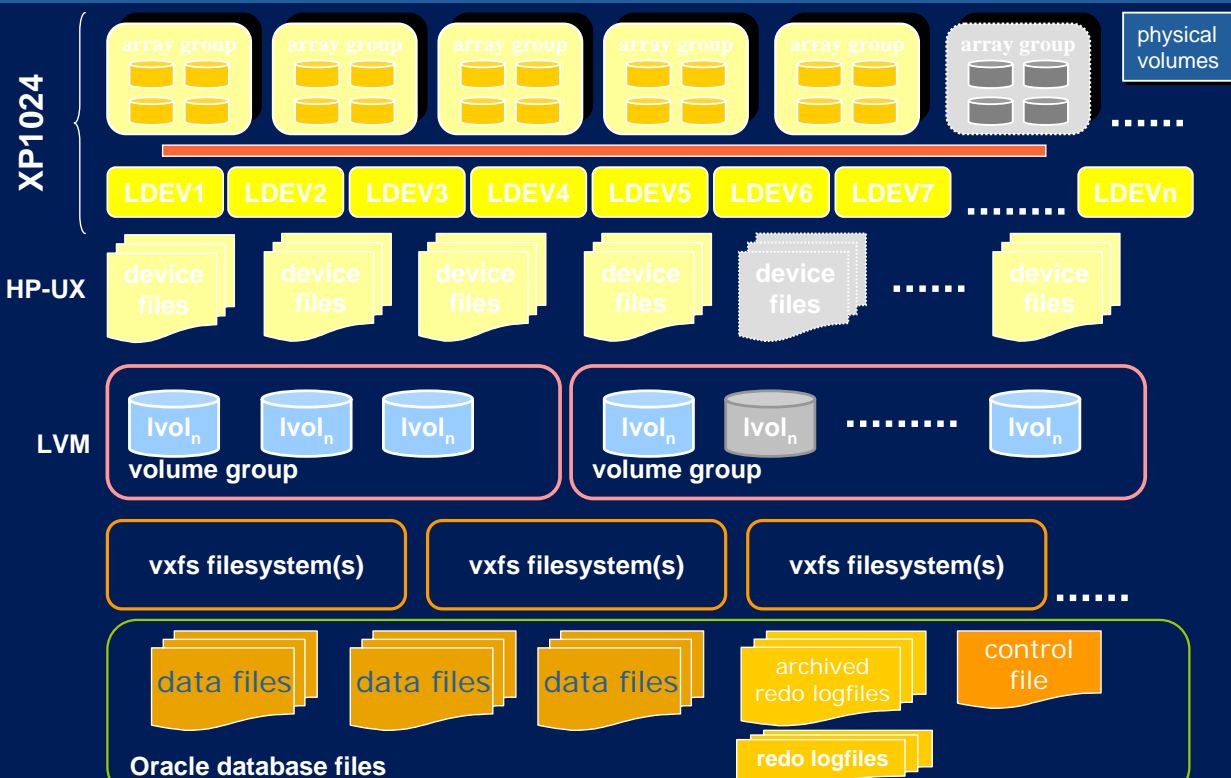
The team



Test lab - system diagram



HP-UX/Oracle9i



Generating the test database



- Test databases are available from the Transaction Processing Council TPC

<http://www.tpc.com>

- Used TPC-H dbgen tool
- Limit of 535GB per mount point as a result of the 8MB extent size. This is a limitation set by the HP-UX logical volume manager.

Test database structure



- 9TB of space allocated with 4.3TB of real data loaded
- 8 tables within the TPC schema
 - Line item 15billion records partitioned over 45 tablespaces
 - 3 datafiles per tablespace
 - Order : 4.5 billion records over 15 tablespaces
 - Partsupp : 2.4 billion records over 20 tablespaces
 - Part: 600 million records over 20 tablespaces
 - Customer : 400 million records over 20 tablespaces
 - Supplier: 30 million records over 20 tablespaces
 - Nation: 25 records in a single tablespace
 - Region: 5 records in single tablespace
- Total: 259 datafiles and 146 tablespaces

- concurrency
- filesperset
- blocksize
- maxopenfiles
- tape I/O slaves
- disk I/O slaves
- contofile autobackup

Results: 3TB/hr backup

| # tape drives | # processors | CPU idle | Data protector parameters | RMAN settings | backup performance |
|-------------------------------|--------------|----------|--|---|---|
| 16 in HP ESL9595 tape library | 32 | 34% | concurrency=1 (1 RMAN channel/tape drive) | disk I/O slaves=16 filesperset=no set maxopenfiles=1 tapeblocksize=256K backup_tape_io_slaves=disabled contofile autobackup=ON | 3.16 TB/hr (including tape load/unload time) |
| | | | | | 3.62 TB/hr (excluding tape load/unload time) |

Results: backup (16 way CPU)



| # tape drives | # processors | CPU idle | Data protector parameters | RMAN settings | backup performance |
|-------------------------------|--------------|----------|---|---|---|
| 16 in HP ESL9595 tape library | 16 | 15% | concurrency=1 (1 RMAN channel/tape drive) | disk I/O slaves=16 filesperset=no set maxopenfiles=1 tapeblocksize=256KB backup_tape_io_slaves=disabled controlfile autobackup=ON | 2.87 TB/hr (excludes tape load/unload time) |

Results: restore



| # tape drives | # processors | CPU idle | Data protector parameters | RMAN settings | backup performance |
|-------------------------------|--------------|----------|---|--|--|
| 16 in HP ESL9595 tape library | 32 | 45% | concurrency=1 (1 RMAN channel/tape drive) | disk I/O slaves=16 filesperset=no set maxopenfiles=1 tapeblocksize=1MB backup_tape_io_slaves=enabled controlfile autobackup=ON | 1.23 TB/hr (including tape load/unload time) |
| | | | | | 1.29 TB/hr (excluding tape load/unload time) |

•Restore times subsequently improved to 2.46TB/hr !

Lessons Learnt



- Backup performance determined in the main by disk subsystem performance.
- Use SAME for best Oracle file system performance
- Storage Router configuration could be simplified.
- Memory usage was low 19GB Max (of 256GB)
- Could have been achieved with 16 Processors (15% idle)

Lessons Learnt



- In a high performance environments set 1 RMAN channel per tape drive, don't use external "multiplexing"
- Restore performance subsequently increased to 2.4 TB/HR – restore 40-60% of backup rate.
- Data Protector 5.1 integration with Oracle is now even better – restore GUI.
- Only Ultrium technology has adaptive tape speed to optimize performance and increase media life.

