Scalable Storage Configuration for the Physics Database Services

Luca Canali, CERN IT

LCG Database Deployment and Persistency Workshop

October, 2005

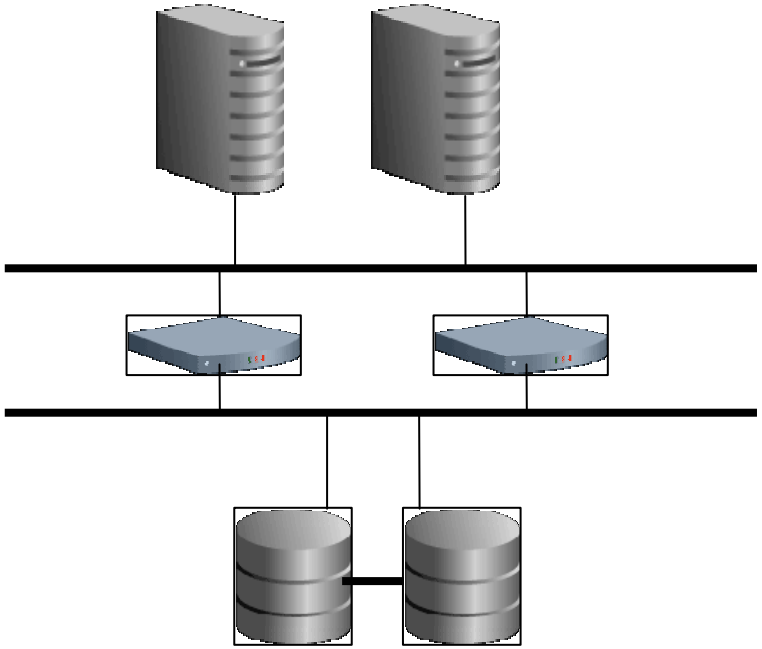
Outline

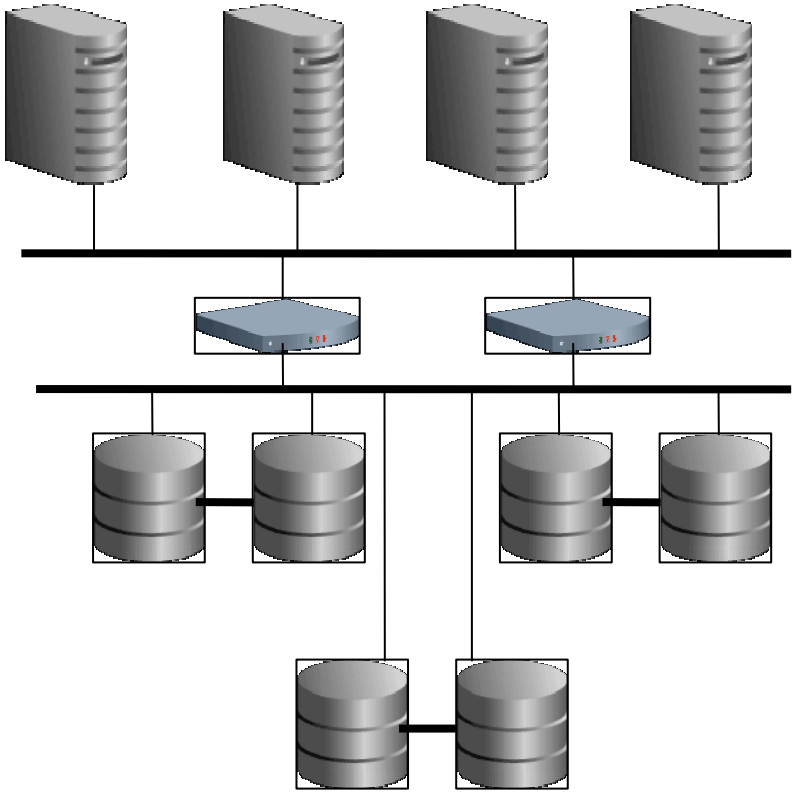
- In this talk I will discuss
  - Storage configuration for scalable database services
    - Main challenges
    - Best practices
  - An implementation of scalable storage
    - Impacts on DB logical to physical mapping
    - Performance and resource allocations
  - How we can help you to size new database projects or to scale up existing applications
    - Performance testing
    - Benchmark data

Oracle Scalable Architecture

Goal: A database infrastructure that provides the required system resources to the end-users and applications.

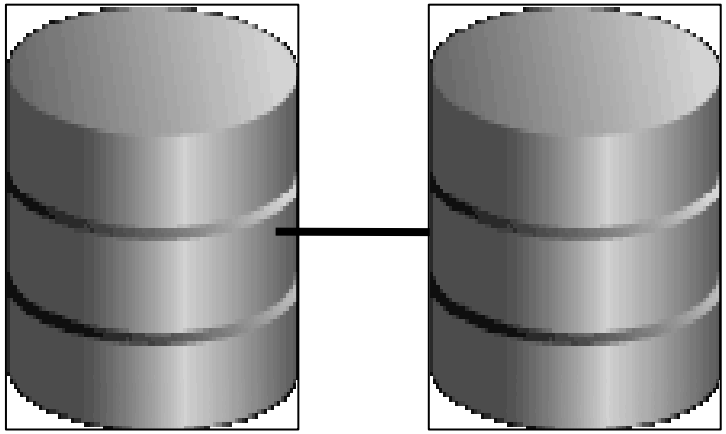
How: A modular architecture that can scale up to a large number of components





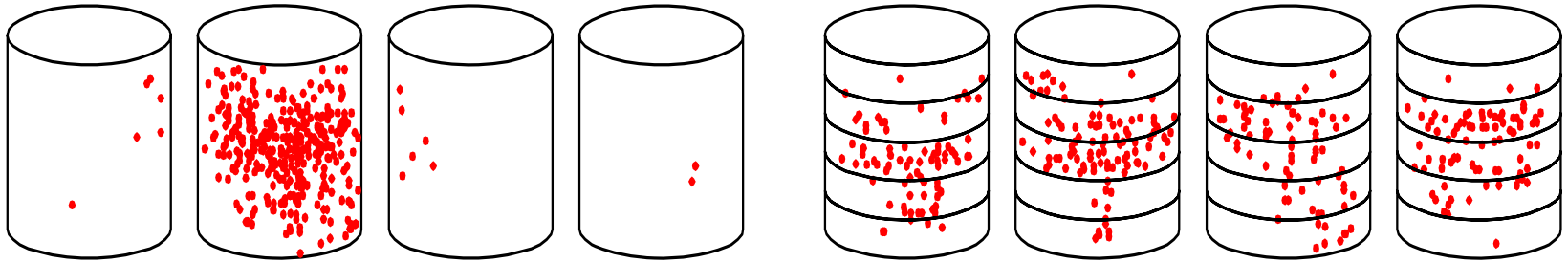
RAID 1: HA Storage

- Mirroring
  - 2-Way mirroring (RAID 1) protects against single point of failures
  - Can be used to redistribute I/O load (performance)



RAID 0: Scalable Performances

- RAID 0 (Striping) automatically redistributes files across multiple disks.
- Performance and scalability are increased
- Error resiliency is decreased



Unstriped Disks **Striped Disks**

Mechanical and Geometrical constraints

- The external part of the disk provides
  - More throughput
  - Less latency

S.A.M.E Strategy

- Goal: optimize storage I/O utilization
- S.A.M.E. (Stripe And Mirror Everything) Strategy
  - Built on the concepts of RAID 1 + 0
  - Proposed by J. Loaiza (Oracle) in 1999
  - Replaces “old recipes”: manual balancing across volumes
- Need a Software or Hardware Volume Manager
  - ASM is Oracle’s solution with 10g “S.A.M.E. out of the box”
  - Other solutions available from different vendors require configuration

Storage Configuration Guidelines

- Use all available disk drives
- Place frequently used data at outer half of disk
  - Fastest transfer rate
  - Minimize seek time
- Stripe data at 1MB extents
  - Distribute the workload across disks
  - Eliminate hot spots
  - Optimum sequential bandwidth gained with1MB I/O
- Stripe redo logs across multiple drives
  - Maximize write throughput for small writes
  - Smaller stripe size (128KB) and/or dedicated disks
- Use cache on the controller
  - 'Write-back' cache
  - Battery-backed cache

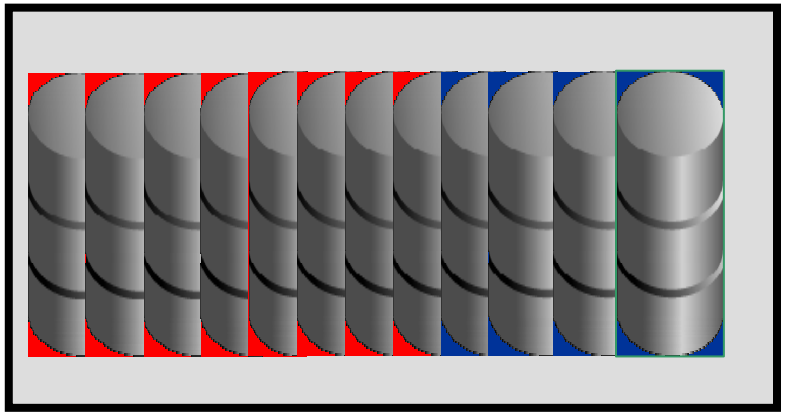
Oracle's ASM Main Features

- Mirror protection:
  - 2-way and 3-way mirroring available.
  - Mirror on a per-file basis
  - Can mirror across storage arrays
- Data striping across the volume:
  - 1MB and 128KB stripes available
- Supports clustering and single instance
- Dynamic data distribution
  - A solution to avoid 'hot spots'
  - On-line add/drop disk with minimal data relocation
  - Automatic database file management
- Database File System with performance of RAW I/O

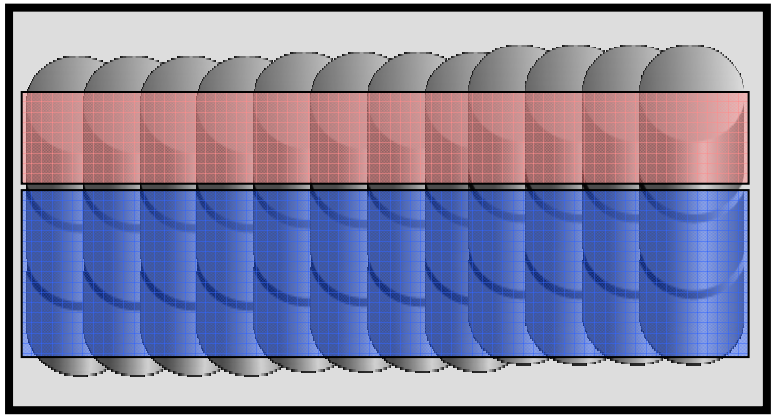
ASM's Configuration – Examples

- ASM is a volume manager, its output are disk groups (DG) that Oracle databases can mount to allocate their files

DATA-DG | RECOVERY-DG



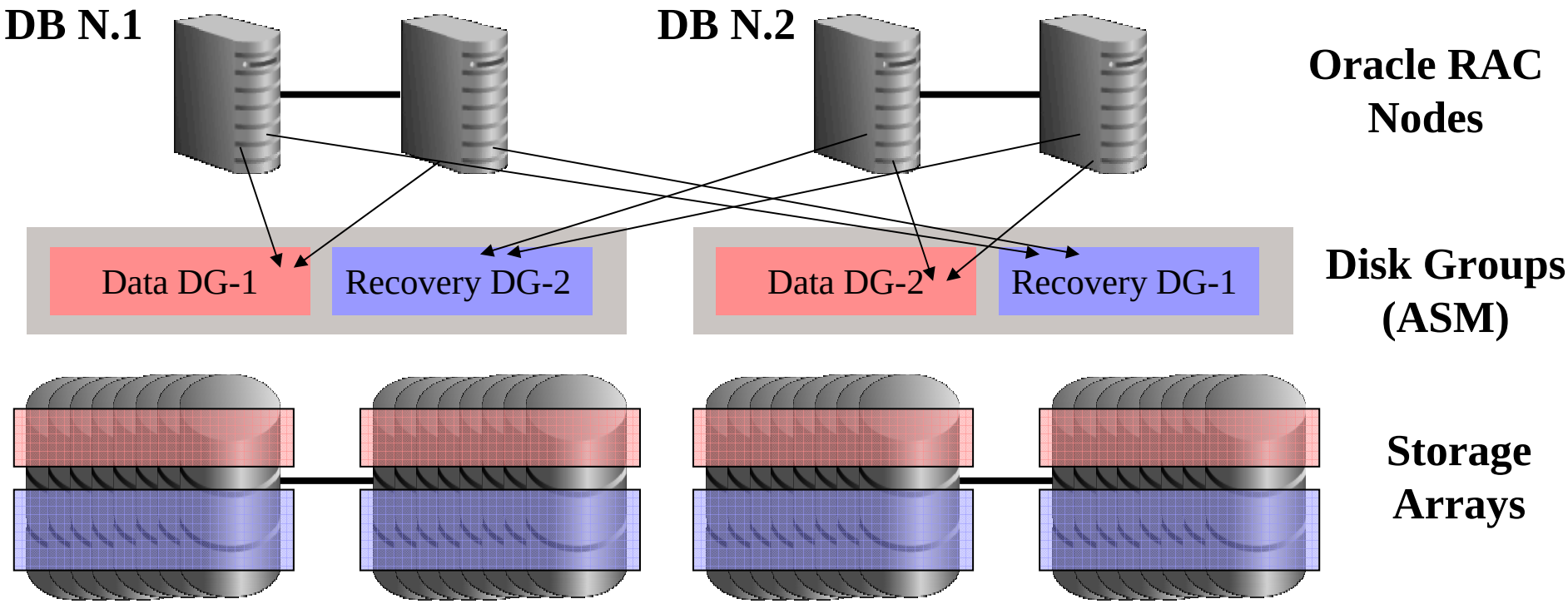
Config 1: Disk groups created with dedicated disks



Config 2: Disk groups created by 'horizontal' slicing

Proposed Storage Configuration

- Proposed storage configuration:
  - High availability - Allows backups to disk
  - High performance - Allows clusterware mirroring (10.2)
  - DBs have dedicated resources



FAQ 1: Datafiles

- Do I need to worry on the number and names of the datafiles allocated for each tablespace?
- "Traditional" storage allocation across multiple volumes:
  - Requires a careful allocation of multiple datafiles across logical volumes and/or filesystems
  - Datafile-to-filesystem and filesystem-to-physical storage mappings have to be frequently tuned
- S.A.M.E. storage, such as Oracle ASM, provides balanced I/O access across disks
  - There is NO NEED, for performance reasons, to allocate multiple datafiles per tablespace.
  - 10g new feature "bigfile tablespace" allows for tablespaces with a single datafile that can grow up to 32 TB (db_block_size=8k)

FAQ 2: Data and Index Tablespaces

- Do I need dedicated tablespaces for indexes and tables?
- Separation of indexes and tables has often been advised to:
  - Distribute I/O
  - Reduce fragmentation
  - Allow separate backup of tables and indexes
- S.A.M.E. storage, such as Oracle ASM, provides balanced I/O access across disks
  - No performance gains are expected by using dedicated tablespaces for INDEXes and TABLEs.
- Additional Notes:
  - Tablespaces fragmentation has little impact when using locally managed TBSs and automatic segment space management (9i and 10g)
  - Very large database can profit from using multiple tablespaces for admin purposes and logical separation of objects

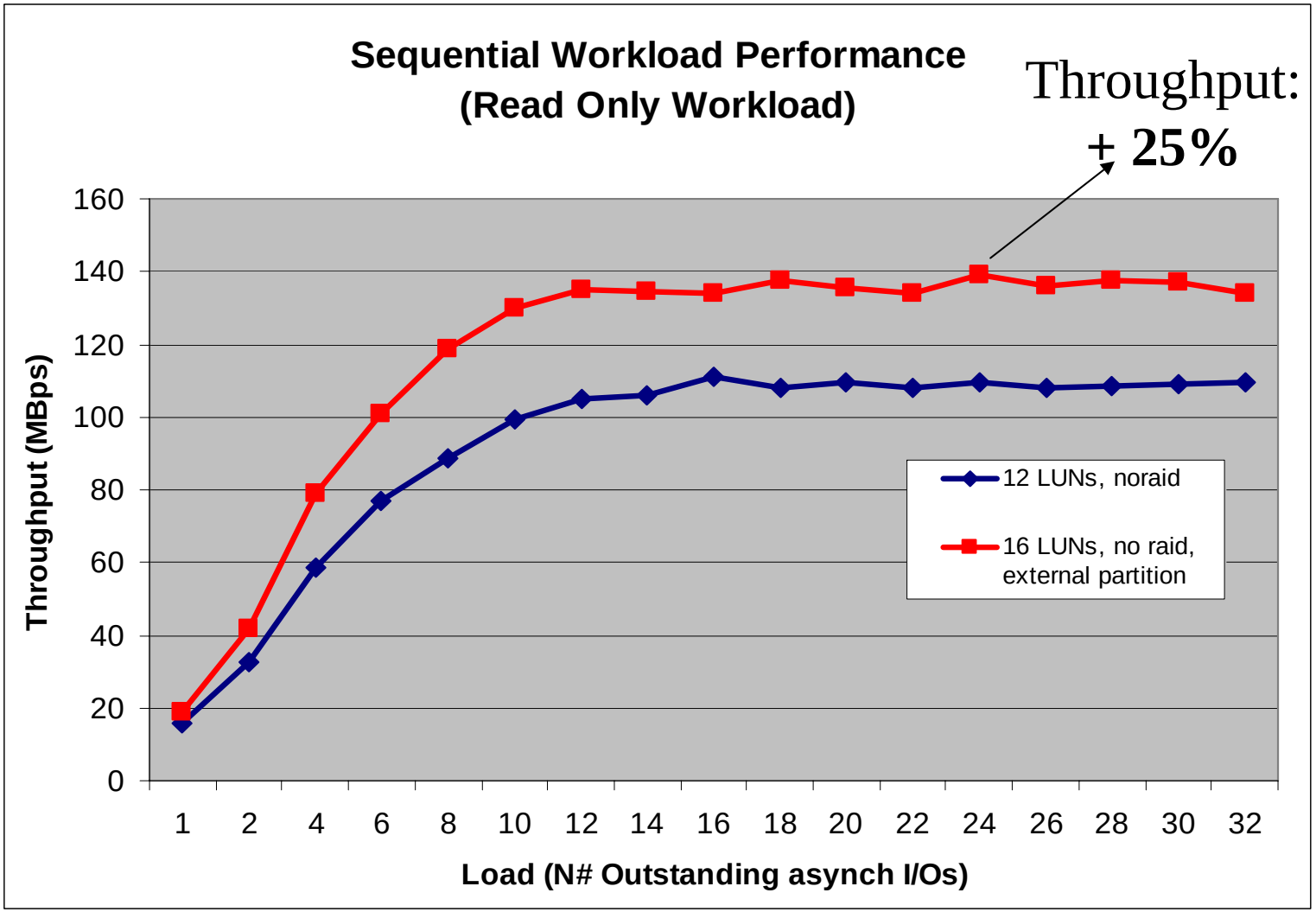
FAQ 3: Sizing Storage

- Storage sizing for database should not take size as the first requirement
  - Bandwidth and performance metrics are bound to the number of disk spindles
  - Magnetic HD technology has improved the GByte/$ ratio
  - The rest of HD technology has not seen much improvements in the last 5 years (since 15K rpms HDs)
- Sizing for storage requirements should
  - Be based on stress test measurements
  - Past performance measurements on comparable systems
  - New projects can leverage benchmark data.
- Extra HD space is not wasted
  - Can be used to strengthen the B&R policy with Disk Backups

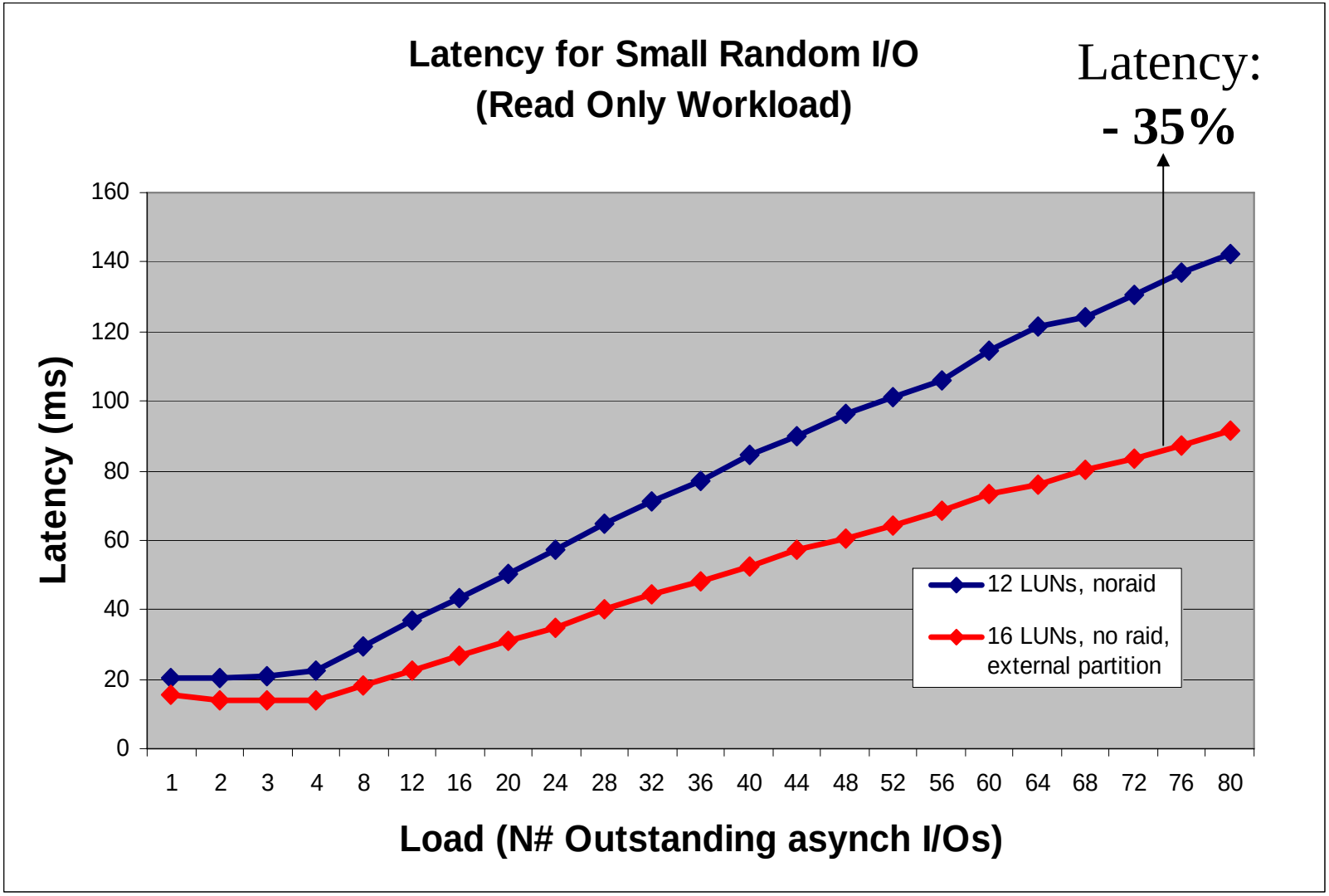
IO Benchmark Measurements

- Benchmark data can be used for
  - Sizing of new projects and upgrades
  - Performance baseline, testing for new hardware
- The following metrics have been measured:
  - Sequential throughput (full scans)
  - Random access (indexed access)
  - I/O per second (indexed access)
  - Metrics are measured as a function of workload
- Other test details
  - Benchmark tool: Oracle's ORION
  - Infortrend Storage array: 16 SATA x 400 GB disks, 1 controller and 1 GB cache

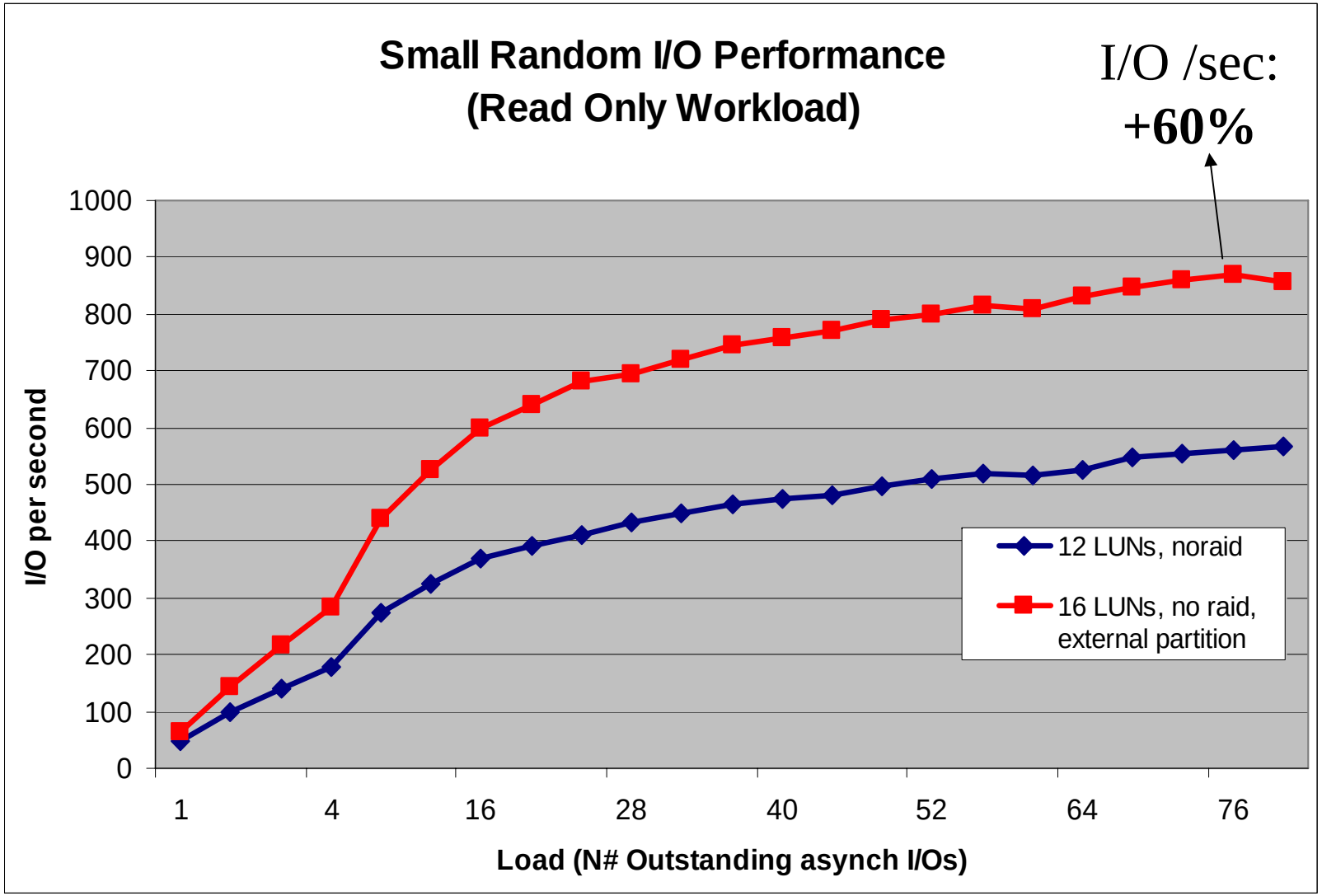
IO Benchmark Data



IO Benchmark Data



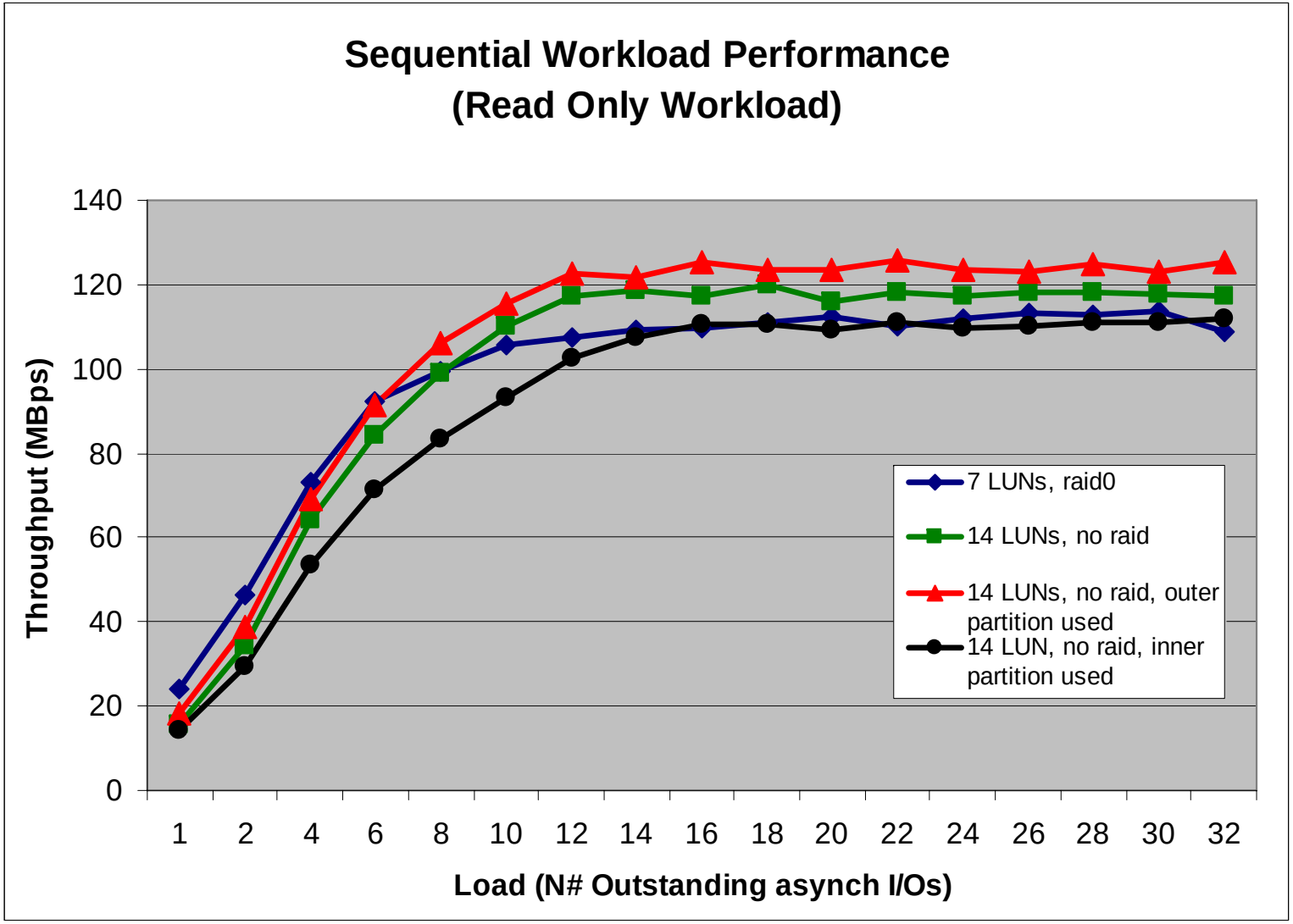
IO Benchmark Data



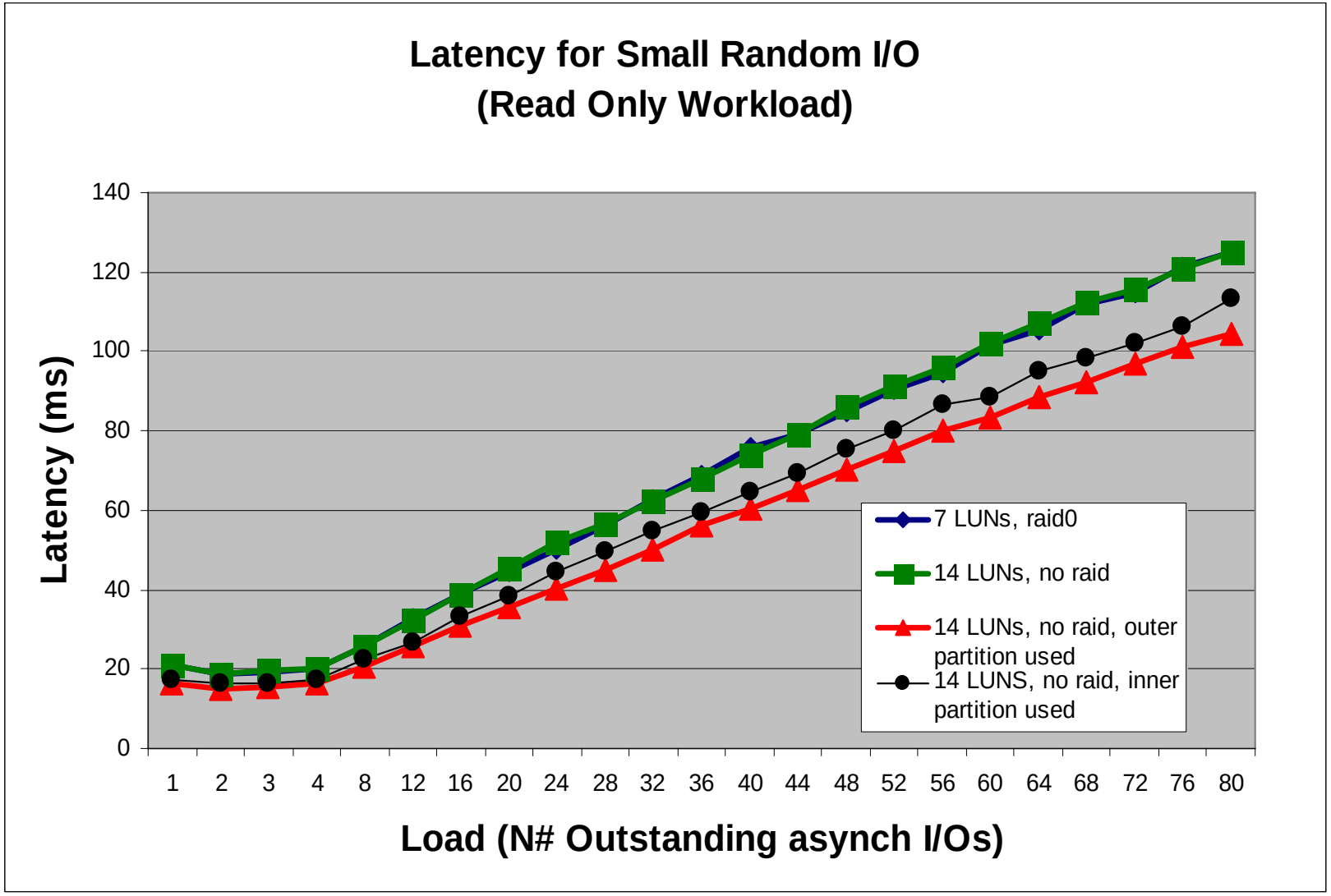
Conclusions

- The Database Services for Physics can provide
  - Scalable Database services
    - Scalable on CPU and Memory resources
    - Scalable on Storage resources
  - Sizing for new projects or upgrades
    - Stress/Performance testing
    - Integration and Validation Testing
    - Benchmark data for capacity planning

Additional Benchmark Data



Additional Benchmark Data



Additional Benchmark Data