Disk Subsystem Performance

Considerations for Databases

Clay Isaacs – Advisory Technical Consultant
EMC Corporation

© 2004 EMC Corporation. All rights reserved.

Agenda

- Terminology
  - Just a few…
- Setting Expectations
  - Why do we tune? Do we have to?
  - Just what does tuning entail?
- Top Ten Tips
  1. Know Your I/O
  2. Choosing a RAID Type
  3. Disk Count
  4. Choose the Right System
  5. LU Distribution and MetaLUNs
  6. Cache Allocation
  7. Managing Cache
  8. Cache Page Size
  9. Logical Unit Settings
  10. Stripes

© 2004 EMC Corporation. All rights reserved.

Terms

- **Alignment** – Data block addresses compared to RAID stripe addresses
- **Coalesce** – To combine multiple smaller I/O into one larger I/O
- **Concurrency** – More than one application or thread writing to a LUN or disk at the same time
- **Flush** – Data in write cache written to disk
- **Locality** – Multiple I/O requested from a reasonably small area on the disk (same MB or GB)
- **RDBMS** – Relational Database Management System
- **Multipathing –** concurrent paths to the same storage LUN

© 2004 EMC Corporation. All rights reserved.

Setting Expectations

What do we tune? And do we have to?

- Tuning is mostly upfront design
  - Choosing an appropriate system, disk count, and RAID type
  - Some array settings which will change behaviors at the margin
- Do I have to tune my design?
  - A modest investment in planning will certainly pay off
- Do I have to tune the storage system?
  - Our "out of the box" settings are designed to satisfy the majority of workloads encountered
  - Clients with unusual workloads can get better performance with a modest effort
    - Some extreme workloads require adjustments
      - High percentage sequential
      - High percentage random, etc.
  - Once acceptable latency requirements met, additional tuning/disk/etc will not help!

© 2004 EMC Corporation. All rights reserved.

Setting Expectations

What does tuning entail?

- Planning ahead is still the most effective technique
  - It is hard to “fix” a problem if the problem is:
    - Suboptimal data layout
    - Wrong RAID type for the I/O profile
  - Full use of physical resources is important
    - No reason for idle drives or processors
- Planning puts the tuning in the right order
  - Some adjustments must be made before committing data
    - Selection of disks, RAID type, metaLUN design
  - Some adjustments are available online
    - Cache settings all can be done while in production

© 2004 EMC Corporation. All rights reserved.

Performance tuning – know your I/O

- What you need to know
  - Predominant I/O size
  - Read/Write ratio
  - Random vs. Sequential mix
  - Data sets that have “linked” contention
    - RDBMS distributed tables
    - Multiple components of the same DB
    - Snapshots/Clones
  - Application type
    - OLTP? DW? Etc.
  - IO or Bandwidth Requirements
  - Vendor suggestions or best practices

© 2004 EMC Corporation. All rights reserved.

Recall from 101: I/O Patterns

- Key things to know:
  - Logs are always sequential
  - Logs are always synchronous (written on commit or log buffer overrun)
  - In most cases – app performance is dominated by read performance
  - ***Not gospel – rather what is frequently observed***

Operation	***Random/Sequential***	***Write/Read***	***Size***
OLTP – Log	Sequential	Write (exception – recovery)	512 bytes – 64KB
OLTP – Data	Random	Read/Write	8KB
Bulk Insert	Sequential	Write	Multiple of 8KB up to 128KB
Read Ahead (DSS, Index Scans)	Sequential	Read	Multiple of 8KB up to 256KB
Backup	Sequential	Read/Write	1MB
Restore	Sequential	Read/Write	64KB
Reindex (read phase)	Sequential	Read	Multiple of 8KB up to 256KB
Reindex (write phase)	Sequential	Write	Multiple of 8KB up to 128KB
CREATE DATABASE	Sequential	Write	512KB

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

Choosing a RAID type

- Random I/O
  - **RAID 1/0** is the performance champ for random I/O
  - **RAID 5** will perform—for an equal number of *disks*—very close to RAID 1/0 in read-heavy (80%) environments
    - At an equal *capacity*, RAID 1/0 will offer much higher performance
    - If ratio of writes is above 20%, RAID 1/0 will be the best bet

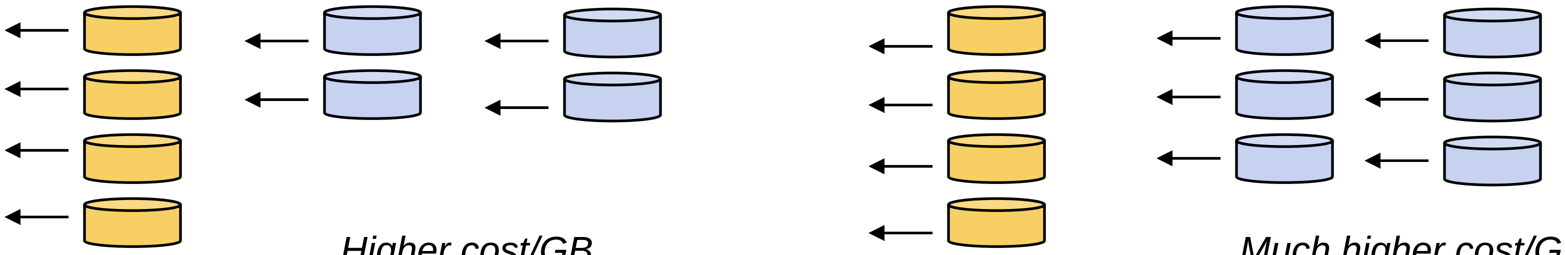
High Ratio of Reads to Writes

Equivalent Spindles: Neck and Neck

RAID 5 3+1 RAID 1/0 2+2

Equivalent Capacity: RAID 1/0 is best

RAID 5 3+1 RAID 1/0 3+3:



Higher cost/GB

Much higher cost/GB

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

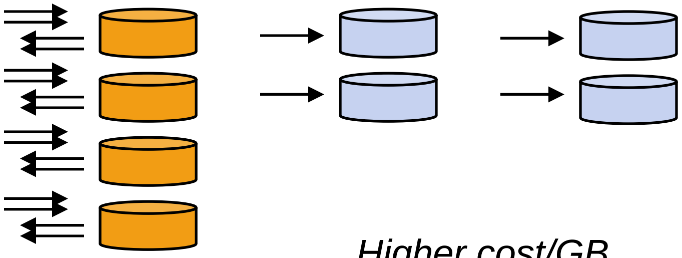
Choosing a RAID type (cont.)

- Random I/O
  - RAID 1 and 1/0 require that two disks be written for each host write
    - Total I/O = Host Reads + 2 * Host Writes
  - RAID 5 requires four operations per host write
    - A RAID 5 write requires 2 reads and 2 writes
    - Total I/O = Host Reads + 4 * Host Writes
    - We do one large stripe write if data is sequential or large (Modified RAID 3, "MR3")

High Ratio of Writes

Parity RAID operations increase disk load

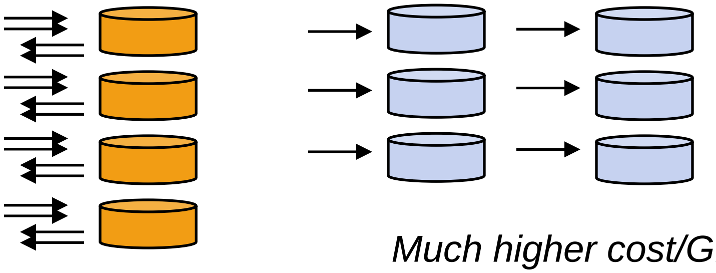
RAID 5 3+1 RAID 1/0 2+2



Higher cost/GB

Equivalent Capacity: RAID 1/0 is best

RAID 5 3+1 RAID 1/0 3+3:



Much higher cost/GB

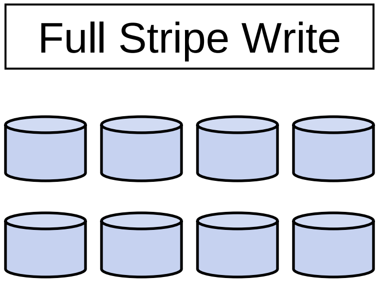
© 2004 EMC Corporation. All rights reserved.

Performance Tuning

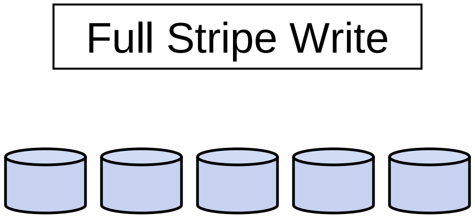
Choosing a RAID type (cont.)

- High Bandwidth (large, sequential I/O)
  - RAID 5 better than RAID 1/0
    - RAID 1/0 is good, but RAID 5 is a bit faster
    - Fewer drives to synchronize: N+1 not N+N
  - RAID 3 is now much more effective
    - Now you can use write cache with RAID 3
    - RAID 3 with ATA drives yields near-FC bandwidth
  - RAID 1/0 is best when workload is mixed (some sequential, some random)

RAID 1/0 4 + 4

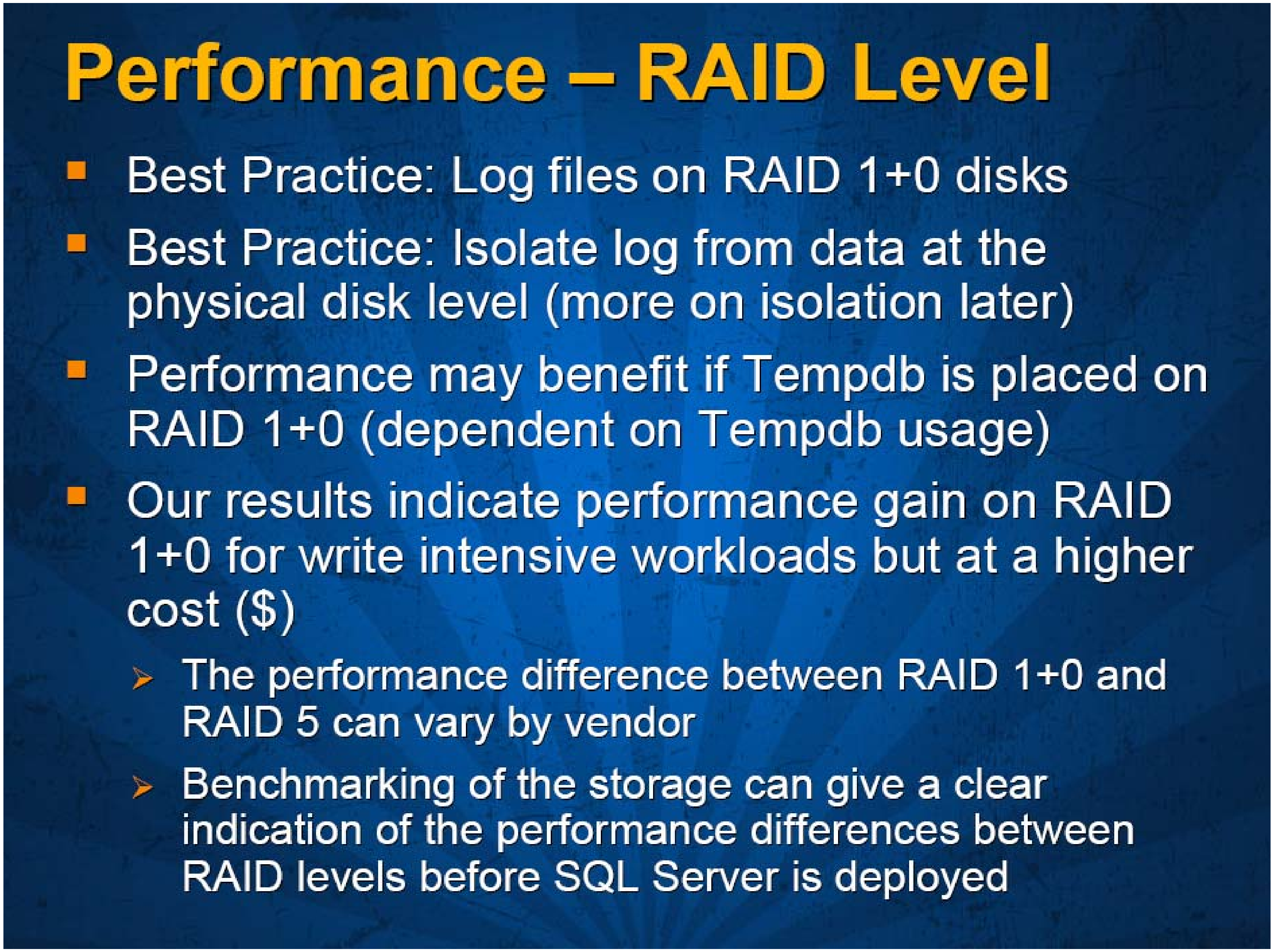


RAID 5 4+1



© 2004 EMC Corporation. All rights reserved.

Microsoft Best Practices for SQL Storage



Performance – RAID Level

- Best Practice: Log files on RAID 1+0 disks
- Best Practice: Isolate log from data at the physical disk level (more on isolation later)
- Performance may benefit if Tempdb is placed on RAID 1+0 (dependent on Tempdb usage)
- Our results indicate performance gain on RAID 1+0 for write intensive workloads but at a higher cost ($)
    - The performance difference between RAID 1+0 and RAID 5 can vary by vendor
    - Benchmarking of the storage can give a clear indication of the performance differences between RAID levels before SQL Server is deployed

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

Determine your disk count

- If your data is host-centric we have to convert to disk IOPS
  - SAR, Perfmon, etc.
- A host write may cause one or more disk I/O
  - Sequential access will likely result in fewer, as I/O are coalesced
  - Random access may result in more I/O
- Convert host load to disk load based on your RAID type
  - RAID 5: Total I/O = Host Reads + 4 * Host Writes
  - RAID 1 and RAID 1/0: Total I/O = Host Reads + 2 * Host Writes

Example: HOST LOAD: 5,200 Random IOPS, 60 % Reads

RAID 5
Disk Load = 0.6 * 5,200 + 4 * (0.4 * 5,200)
= 3,120 + 4 * 2,080
= 3,120 + 8,320
= 11,440 IOPS

RAID 1/0
Disk Load = 0.6 * 5,200 + 2 * (0.4 * 5,200)
= 3,120 + 2 * 2,080
= 3,120 + 4,160
= 7,280 IOPS

© 2004 EMC Corporation. All rights reserved.

Layout Goals

- Layouts should attempt to support the required I/O profile
- In general acceptable performance metrics are:
  - Data Files

< 6 msec	Ideal
6 – 20 msec	Acceptable
> 20 msec	Needs resolution, poor user experience

  - Log Files

< 1-2 msec	Great
2 – 6 msec	Acceptable
6 – 15 msec	Investigate
15 – 20 msec	Will start to impact scalability
>20 msec	Needs resolution

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

Determine your disk count (cont.)

- Disk drives are a critical element of CLARiiON performance
  - Use the rule of thumb to determine the number of drives to use
- Rules of Thumb for Drive Performance are below
  - These are a conservative *starting point* for analysis, not the absolute maximums!

	10 K rpm	*10 K rpm*	*15 K rpm*	*15 K rpm*	*SATA II*
IOPS	100 IOPS	120 IOPS	150 IOPS	180 IOPS	80 IOPS
Bandwidth	10 MB/s		13 MB/s		10 MB/s (RAID 3)

Example: HOST LOAD: 5,200 Random IOPS, 60 % Reads

RAID 5
Disk Load = 11,440 IOPS

RAID 1/0
Disk Load = 7,280 IOPS

11,440 / 100 = 114 drives ← **10 K rpm Drives** → **7280 / 100 = 72 drives**

11,440 / 150 = 76 drives ← **15 K rpm Drives** → **7280 / 150 = 49 drives**

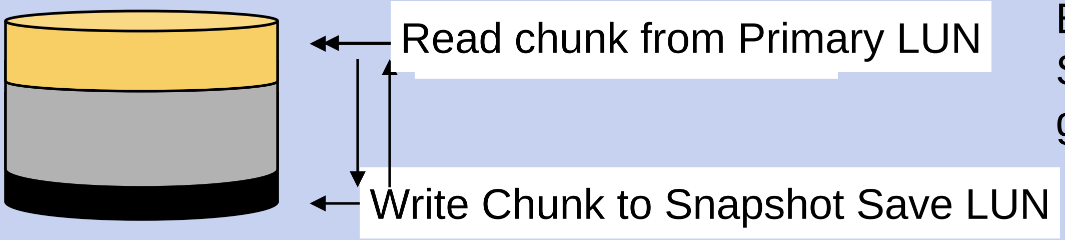
© 2004 EMC Corporation. All rights reserved.

Performance Tuning

LUN distribution and MetaLUNs

- Distribute your I/O load evenly across your available disk drives
  - Bandwidth: Better to dedicate a few drives per process than share drives between processes
    - Fewer seeks, more sequential operation
  - IOPS: Share spindles with multiple processes
- But – Avoid linked contention
    - Primary and BCV on same RAID group
    - Primary and SnapView Save area on same RAID group
    - Application contention
    - Meta or Host Striping across the same RAID group

In Linked Contention, disk heads must move over large ranges of disk addresses in order to service I/O being requested at the same time.



Read chunk from Primary LUN

Write Chunk to Snapshot Save LUN

EXAMPLE: Snapshot Save area on same disk group as primary LUN

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

How to allocate cache (assuming it's tunable)

- Write Cache should get the largest share
  - Disk writes are more costly than reads
  - Most Windows environments are random except certain DB files
  - Larger the write cache, the better. Be careful of small NVRAM write caches
- Read Cache
  - Typical users need <= 100 MB Read Cache
  - Large Systems
    - 250 MB of read cache is plenty
    - Media/Massive DSS: 1 GB of read cache is about the max you can use
  - Why so little? We don't need to keep the data around long.
- Match cache page size to predominate I/O size for dedicated workloads.
  - Use defaults for mixed workloads (8KB in a Clariion)

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

FC or iSCSI?

- Fiber Channel
  - Highest bandwidth per connection (currently 4Gb/s)
  - Robust medium, non-shared
  - Highest cost (although getting cheaper)
  - Best for large I/O or high bandwidth designs
- **iSCSI**
  - Currently 1Gb/s per connection
  - Lowest cost (in some cases free, others same cost as FC)
  - Shared Medium – follow best practices (VLAN’s or physical network seperation
  - Works best for OLTP, small block I/O
  - iSCSI HBA vs MS iSCSI initiator
- Whichever choice, use a minimum of 2 connections per host. More if needed…
  - Add PowerPath for heavy I/O workloads for load balancing

© 2004 EMC Corporation. All rights reserved.

Performance Tuning

Stripes

- The default stripe element size is 64 KB (128 blocks)
  - Don't change it unless directed to
- Be careful with huge stripes (large RAID groups)
  - Requires a larger cache page for high bandwidth writes
  - Harder to get good bandwidth on write-cache bypass
  - Takes longer to rebuild
  - Better to use smaller groups and MetaLUNs/host striping
- Fix alignment issues when you begin
  - Windows is the primary concern
  - Use DISKPAR.EXE to align at 64 KB (see Best Practices white paper, for example). This can result in 20-30% performance improvement alone.
- Be careful of diminishing returns with large spindle counts and stripe widths. Start smaller and grow.

© 2004 EMC Corporation. All rights reserved.

Summary

1. Know Your I/O
2. Choosing a RAID Type
3. Disk Count
4. Choosing the Right System
5. LU Distribution and MetaLUNs
6. Cache Allocation
7. Managing Cache Space
8. Cache Page Size
9. Logical Unit Settings
10. Stripes

Revisit your settings periodically to insure you’re still on track.

© 2004 EMC Corporation. All rights reserved.

Some Good Reference Whitepapers

- EMC Clariion Best Practices for Fibre Channel Storage
- EMC Clariion Database Storage Solutions SQL Server 2000 Best Practices
- EMC Clariion Database Storage Solutions Best Practices for Microsoft SQL Server 2000 in SAN Environments
- Project REAL: SQL Server 2005 Storage Management and Availability
- The Essential Guide to Table Partitioning and Data Lifecycle Management
- SQL Server 2005: Storage Engine Best Practices and Enhancements (mikeruth@microsoft.com)

© 2004 EMC Corporation. All rights reserved.