TSM Oxford Symposium
Disk Tuning with TSM

Dave Canan, IBM Advanced Technical Support
ddcanan@us.ibm.com

Tivoli. software

e business software

Agenda

- Cost vs. Performance: Considerations when sizing the TSM environment
- TSM Disk Tuning Basics
- How TSM does I/O
- Utilities that can be used for examining disk IO with TSM
- TSM Disk Tuning and AIX
- Recommendations for TSM and Various Disk Technologies/Subsystems
- Issues still needing further study
- Appendix – ATS Disk Study

IBM Software Group

Cost vs. Performance: Considerations When Sizing the TSM Environment

Cost vs. Performance: Considerations When Sizing the TSM Environment

- Customer needs to consider pros and cons of cost vs. performance
- Any backup application does heavy I/O. TSM is no exception. If this is not planned for, you may have performance issues.
- Better performance comes at an additional cost.
- Trend in industry is to have larger capacity disks. Customer must decide whether they should purchase more disks of a smaller size vs. fewer disks of a larger size. This comes with a trade-off in performance
- Best practice is to isolate TSM components as much as possible from each other.
- In most cases, some portion of the disk must be left vacant to achieve maximum performance
- For larger TSM environments, we recommend that customers dedicate a complete disk subsystem to TSM

Cost vs. Performance: Considerations when sizing the TSM environment

- Data management software (TSM) should not be a last minute “add-on” for a project. Proper capacity and performance sizing should be a priority. This will save money in the long run.
- If capacity and performance planning is done by a department other than the TSM group, that organization should be involved with the TSM planning from the beginning of the project.
- Don’t let the storage management group dictate what you have to use. You don’t just need space, you need performance.

IBM Software Group

TSM Disk Tuning Basics

Oxford University TSM Symposium 2007

TSM Disk Tuning Basics

- Poor disk setup **WILL** impact performance !!
  - Advanced functionality of some devices should reduce RAID impact
- DIO (direct I/O) now the default in TSM 5.3 + for disk storage pools (not for Linux)
  - Filesystem selection is now less important (exception Solaris and UFS)
  - With the advent of DIO, file system I/O is now more efficient
- Sharing disk with other apps **WILL** impact performance.
- Be careful when multiple TSM servers are sharing a disk subsystem
  - Treat them as separate applications
- For some disk subsystems, cache settings can be set separately for read and write
- Disk monitoring tools are available for performance tuning

TSM Disk Tuning Basics

- Disk not the only component to consider. All I/O factors must be considered together:
  - I/O characteristics of client workload
  - Disk drive capacity and % utilized
  - Disk and/or filesystem fragmentation
  - Disk drive speed (rpm) (example, SATA vs. FC)
  - Processor speed of disk subsystem
  - Disk subsystem cache size and settings
  - Number and type of disk adapters
  - RAID type
  - Data layout
  - How the disks are “wired”
  - Type of Filesystem used (JFS, JFS2, RLV, etc.

How TSM does I/O

The Nature of I/O Behavior with TSM

- Three separate components of TSM architecture
  - TSM database
  - TSM Recovery Log
  - TSM Storage Pools
- Need to be considered separately for setup and tuning
  - Each has a different I/O access behavior
  - Each has different cache requirements

I/O Behavior for the TSM Database (1 of 2)

- Access pattern for DB is **random** during most operations. Because of random IO, use the fastest disks you have for the DB.
- Access pattern is **sequential** during DB backup
  - DB volumes read one at a time, from beginning to end of DB volume
  - Volumes read for backup in the order in which they were defined to TSM
- DB volumes will initially be filled to EOV before going to next volume
- DB reorganization moves data to fewer volumes (may not be desirable for performance)
- With DS4800, RAID5 is good for DB. Pay attention to cache settings for DB

I/O Behavior for the TSM Database (2 of 2)

- Look at IOPS for how the TSM DB is performing. If IOPS are not being met, you should consider adding additional volumes for the database
  - Other FC disks can handle roughly 150 IOPS before queuing occurs
  - Iostat or filemon can be used to determine the number of IOPS occurring against volumes
  - Server instrumentation DB threads DB also give count of actual I/Os
- If possible, sacrifice space and use more disks for best performance
- Create between 4 and 16 DB volumes for use with TSM
- Blocksize for I/O varies, I/O to DB does NOT use Direct I/O
  - 4KB for all operations except during TSM database backup
  - 256KB reads done during TSM database backup

I/O Behavior for the TSM Recovery Log

- Access to recovery log is always sequential
- I/O blocksize used is always 4KB, I/O to log does NOT use direct I/O
- Good candidate for RAID0 (if using TSM mirroring). If not using TSM mirroring, then you could mirror with RAID1
- Sacrifice disk space for best performance. If you can't sacrifice a volume for the log, it is better to place with storage pools rather than with DB.
- Number of log volumes not important; may want to have 2 for easy of maintenance.
- TSM recovery log normally only written to except at TSM initialization.

I/O Behavior for the TSM Storage Pools

- TSM storage pools are generally written to in the order in which they are defined to TSM. This becomes irrelevant once many tasks are executing.
- Understand the disk technologies when designing the storage pools. Some SATA-1 drives do not support NCQ, SATA-IO (also known as SATA-3.0) drives do.
- Two “rules-of-thumb” for how many storage pool volumes to define on disk. In order of priority, consider the following:

1. For any given logical volume (LV), you should have no more storage pool volumes defined within that LV than you have hdisks in the LV. (Example: if you had 6 hdisks striped together in a logical volume, you would define no more than 6 storage pool volumes within that LV.)

2. Try to have as many storage pool volumes defined in a storage pool as you have simultaneous backup sessions to that storage pool

I/O Behavior for the TSM Storage Pools

- TSM storage pools always read/written using 256KB blocksize
- TSM storage pools should be sized (# of physical disks required) based on throughput requirements.
  - Generally, FCS adapter is limiting factor (200MB/sec per adapter)
- With TSM 5.3, Direct I/O is enabled by default for disk storage pool volumes (If desired, DB and Log volumes can be mounted with DIO option under AIX)
  - TSM 5.3 removed this option from documentation (AIXDIRECTIO YES|NO)
  - Because of direct I/O now being used, vmtune/vmo/ioo tuning is no longer necessary for storage pool volumes. It still may provide some minor benefit for TSM DB database backup

I/O Behavior for the TSM Storage Pools

- TSM storage pools are written to sequentially (in 256KB blocks)
  - Try and keep all pieces of I/O path (sector size, segment size, PP size, etc.) at least this big or larger.
- Consider how many storage pool volumes you have per physical disk.
  - If there are too many, then the sequential I/O in effect becomes random because of disk head movement (thrashing)
  - Command queuing helps here.
  - Some research from hardware storage indicates more than 3 sequential streams to a disk causes the I/O to be more random in nature.

IBM Software Group

Utilities that can be used for Examining Disk I/O with TSM

Tivoli. software

e business software

TSM Disk Tuning Basics – Utilities to Measure Performance (TSM Server side)

- For TSM Server, you can use:
  - Nmon utility (AIX/Linux)
  - Iostat (UNIX) (look for disks > 20% busy)
  - Filemon command (provides detailed perf info at a file level) (AIX)
  - Fileplace command (provides fragmentation information) (AIX)
  - Server instrumentation trace – average disk response time, InstTput rate
  - Client Instrumentation trace – high % of DataVerb and EndTxn time might indicate an disk issue on the server
  - Tops/monitor utility
  - Expire inventory, backup DB processes
  - Perfmon (Windows)

TSM Disk Tuning Basics – Utilities to Measure Performance (Client side)

- For TSM Client, you can use:
  - Nmon utility (AIX/Linux)
  - Iostat (UNIX) (look for disks > 20% busy)
  - Filemon command (provides detailed perf info at a file level) (AIX)
  - Fileplace command (provides fragmentation information) (AIX)
  - Client instrumentation trace (high % File I/O or Process Dirs = Red Flag)
  - Tops/monitor utility
  - Perfmon (Windows)

TSM Disk Tuning Basics – The NMON "Weighted Average"

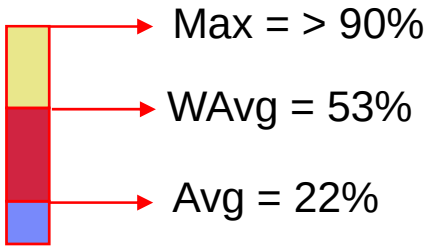
Snapshot	% Busy	
1	0	Disk was inactive
2	0	
3	0	
4	0	
5	25%	Disk was active
6	35%	
7	90%	
8	80%	
9	35%	
10	0	Disk was inactive
11	0	
12	0	

The **Average** disk utilization during the 12 snapshots was :

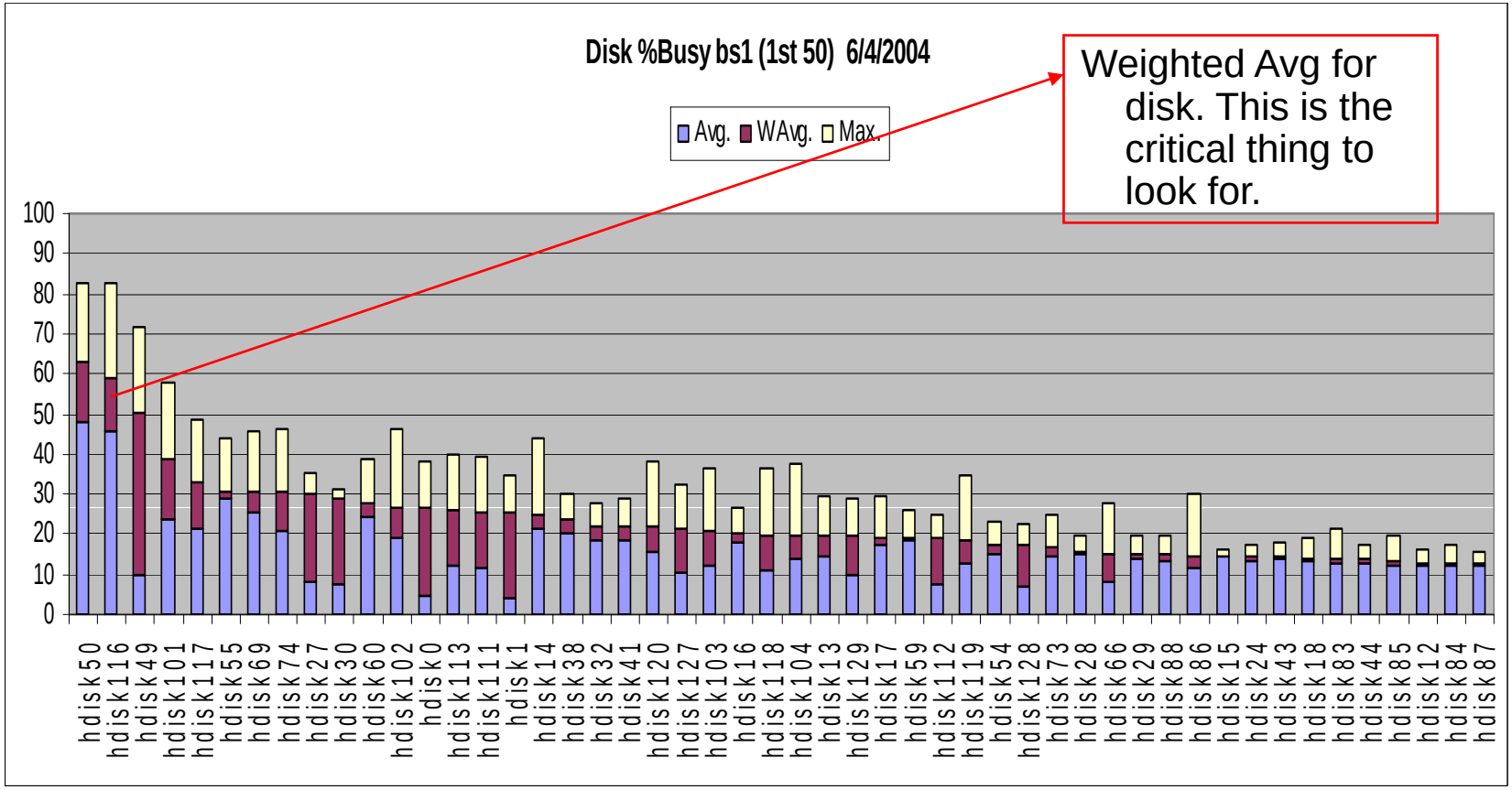
(25+35+90+80+35)/12 = **22%**

The **Weighted Average** disk utilization for the time the disk was active was:

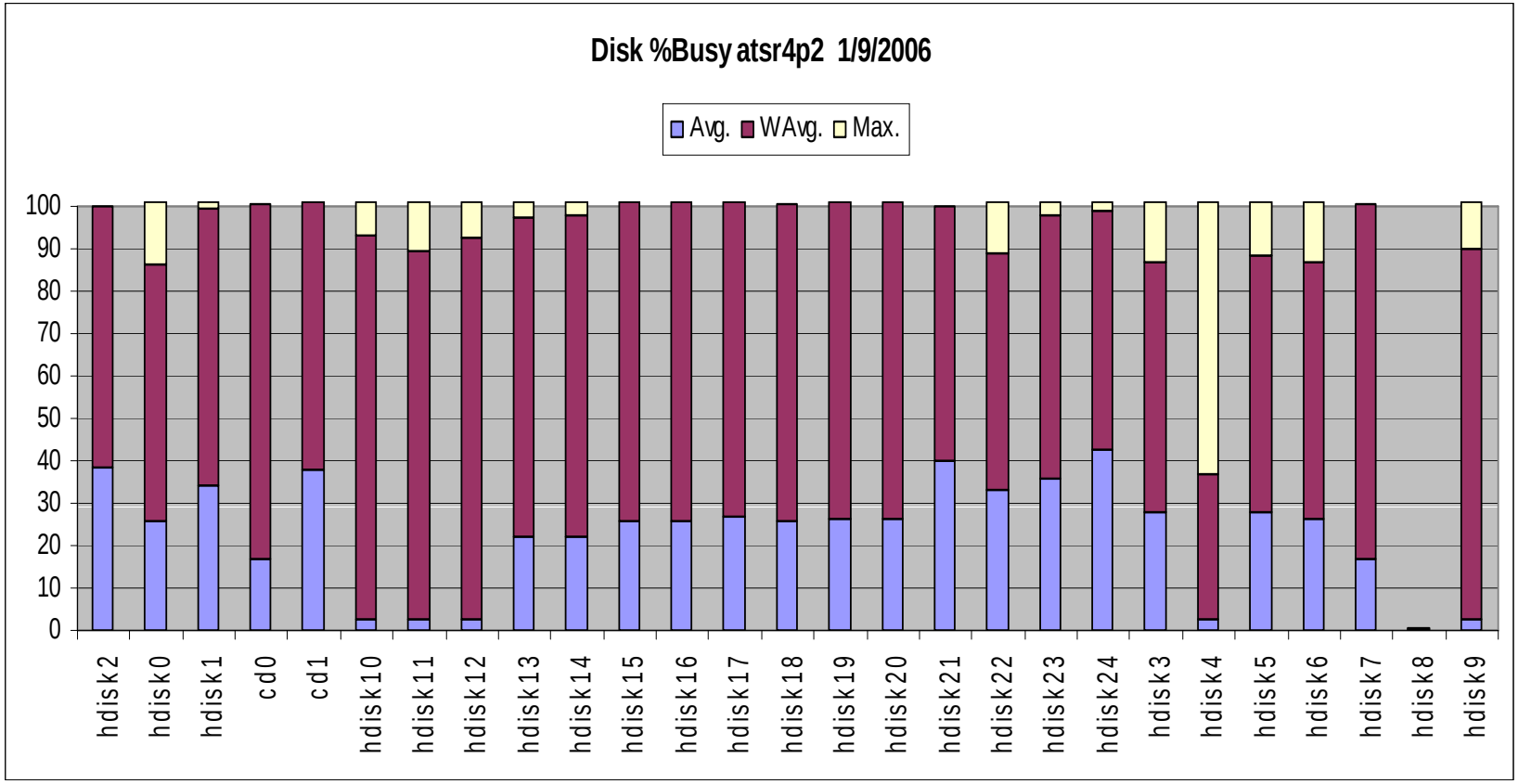
(25+35+90+80+35)/5 = **53%**



TSM Disk Tuning Utilities – NMON Example #1 (Some disks having high I/O)



TSM Disk Tuning Utilities – NMON Example #2 (Heavily Loaded System)



TSM Disk Tuning Utilities – TSM Selects for DB backup and Expiration Performance

```
select activity,cast((end_time)as date) as "Date", -
(examined/cast((end_time-start_time) seconds as
  decimal(18,13))*3600) "4KB Pages Backed Up/Hr" -
from summary where activity='FULL_DBBACKUP' and -
days(end_time)-days(start_time)=0

select activity,cast((end_time)as date) as "Date", -
(examined/cast((end_time-start_time) seconds as
  decimal(18,13))*3600) "Objects Examined/Hr"  -
from summary where activity='EXPIRATION' and -
days(end_time)-days(start_time)=0
```

TSM Disk Tuning Utilities – Select Example for DB Backup Performance

ACTIVITY	Date	4KB Pages Backed Up/Hr
FULL_DBBACKUP	2007-02-26	12726000
FULL_DBBACKUP	2007-03-05	10299600
FULL_DBBACKUP	2007-03-12	9691200
FULL_DBBACKUP	2007-03-19	12330000

Rule of Thumb – 8MB/sec is “acceptable” base for DB backup performance. (8MB/sec = 28 GB/hr)

Last column is in # of 4KB pages per hour, we are looking for this number to be greater than ~ 2000000

If rule of thumb not being met, this may indicate a problem

TSM Disk Tuning Utilities – Select for Expiration Performance Example

ACTIVITY	Date	Objects Examined/Hr
EXPIRATION	2007-02-22	1828800
EXPIRATION	2007-02-23	1674000
EXPIRATION	2007-02-24	1972800
EXPIRATION	2007-02-25	1425600
EXPIRATION	2007-02-26	2084400
EXPIRATION	2007-02-27	1782000
EXPIRATION	2007-02-28	2926800
EXPIRATION	2007-03-01	2829600
EXPIRATION	2007-03-02	2613600
EXPIRATION	2007-03-03	2336400
EXPIRATION	2007-03-04	2052000
EXPIRATION	2007-03-05	2138400
EXPIRATION	2007-03-06	1584000
EXPIRATION	2007-03-07	2314800
EXPIRATION	2007-03-08	2584800
EXPIRATION	2007-03-09	2466000
EXPIRATION	2007-03-10	2311200
EXPIRATION	2007-03-11	2617200
EXPIRATION	2007-03-12	2156400
EXPIRATION	2007-03-13	1692000

Rule of Thumb – 3,800,000 Objects Examined / Hr

Note: this can indicate disk tuning **may** be needed **or** that the DB is fragmented or that additional hardware may be needed for the database to perform properly. .

TSM Disk Tuning Utilities - Server Instrumentation Traces

42 ms for avg disk read!! (stgpool volume)

6MB/sec!! (Look here 1st)

33 ms for avg read!! (DB Volume)

Count/Tottime yields 30 IOPS

```
------------------------------ Disk Migration Example -------------------

Thread 56 DiskServerThread parent=0 (AIX TID 1798311)  13:47:34.766-->14:03:20.744
/dev/rdiskpool_lv3
Operation       Count   Tottime  Avgtime  Mintime  Maxtime      InstTput  Total KB
-------------------------------------------------------------------------------
Disk Read       13835   584.982    0.042    0.000    3.332       6012.2   3517004
Disk Write         14     0.087    0.006    0.000    0.010      32162.6      2804
Thread Wait     13784    360.45    0.026    0.000   73.961
Unknown                           0.459
-------------------------------------------------------------------------------
Total                   945.978
                                                                 3720.8   3519808

------------------------------ Backup TSM DB Example -------------

Thread 1 LvmDiskServer parent=0 (AIX TID 2064489)  11:51:21.126-->12:05:18.174
Operation        Count  Tottime   Avgtime  Mintime  Maxtime   InstTput  Total KB
-------------------------------------------------------------------------------
Disk Read        24093  793.630    0.033    0.000    2.622    6732.1   5342796
Disk Write         125    0.279    0.002    0.000    0.050    1788.7       500
Thread Wait      23983   42.133    0.002    0.000   22.656
Unknown                   1.004
-------------------------------------------------------------------------------
Total                   837.047
                                                                 6383.5
   5343296
```

TSM Disk Tuning Utilities - Client Instrumentation Trace Example

```
Final Detailed Instrumentation statistics

Elapsed time:   318.117 sec

Section       Total Time(sec)  Average Time(msec)  Frequency used

-------------------------------------------------------------------
Client Setup          5.297          5297.0                   1
Process Dirs          0.050             6.3                   8
Solve Tree            0.000             0.0                   1
Compute               0.280             0.0               16396
Transaction           0.411             0.0               49219
BeginTxn Verb         0.000             0.0                   3
File I/O             11.885             0.7               16406
Compression           0.000             0.0                   0
Encryption            0.000             0.0                   0
CRC                   0.000             0.0                   0
Delta                 0.000             0.0                   0
Data Verb           298.322            18.2               16396
Confirm Verb          0.250            35.7                   7
EndTxn Verb           0.551           183.7                   3

-------------------------------------------------------------------
```

94% of time spent here!!

Client side

Server/NW Side

TSM Disk Tuning Utilities - Filemon Example 1

Busiest logical volumes disks reported during run period

```
Most Active Logical Volumes
------------------------------------------------------------------------
  util  #rblk  #wblk   KB/s  volume                    description
------------------------------------------------------------------------
  0.96      0 123723608 82139.1  /dev/lv-stg-c            /stg-c
  0.90      0 123722072 82138.0  /dev/lv-stg-d            /stg-d

Most Active Physical Volumes
------------------------------------------------------------------------
  util  #rblk  #wblk   KB/s  volume                    description
------------------------------------------------------------------------
  0.65      0 61860520 41068.7  /dev/hdisk8              N/A
  0.65      0 61861672 41069.4  /dev/hdisk5              N/A
  0.59      0 61861552 41069.4  /dev/hdisk7              N/A
  0.55      0 61861936 41069.6  /dev/hdisk6              N/A
```

Busiest physical volumes disks reported during run period

TSM Disk Tuning Utilities - Filemon Example 2

```
VOLUME: /dev/hdisk8  description: N/A
writes:                  241743   (0 errs)
  write sizes (blks): avg    255.9 min     max    256 sdev      4.5
  write times (msec): avg    2.023 min   0.237 max 1545.563 sdev
  13.474
  write sequences:       241743
  write seq. lengths:    avg   255.9 min       8 max     256 sdev
  4.5
seeks:                   241743   (100.0%)
  seek dist (blks):      init 34114816,
                 avg 6316189.4 min    1008 max 81789184 sdev
  13621796.3
  seek dist (%tot blks):init 24.64729,
                 avg 4.56332 min 0.00073 max 59.09109 sdev 9.84148
time to next req(msec): avg    2.911 min   0.260 max 1553.078 sdev
  19.307
throughput:              41068.7 KB/sec
utilization:             0.65
```

Blksize of write

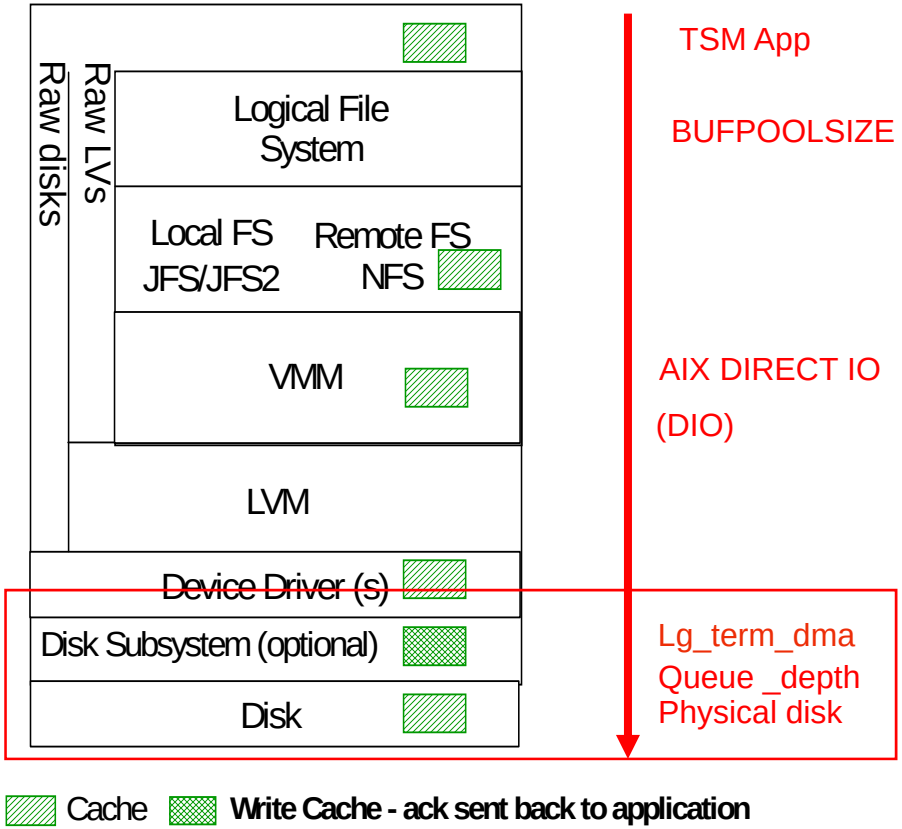
How far for avg seek

% Busy

IBM Software Group

TSM Disk Tuning and AIX

The AIX I/O Model and TSM



- At top is application layer. TSM uses cache buffers to avoid DB I/O
- Direct I/O being used with TSM 5.3 (no VMM) for stgpool volumes
- Queues exist for HBAs and hdisks
  - Queue_depth is # of outstanding IO requests that can be done at a disk level
  - Lg_term_dma is a memory area on the FC adapter used to store IO commands and data
- Subsystem R/W Cache
- Disk memory to Store commands

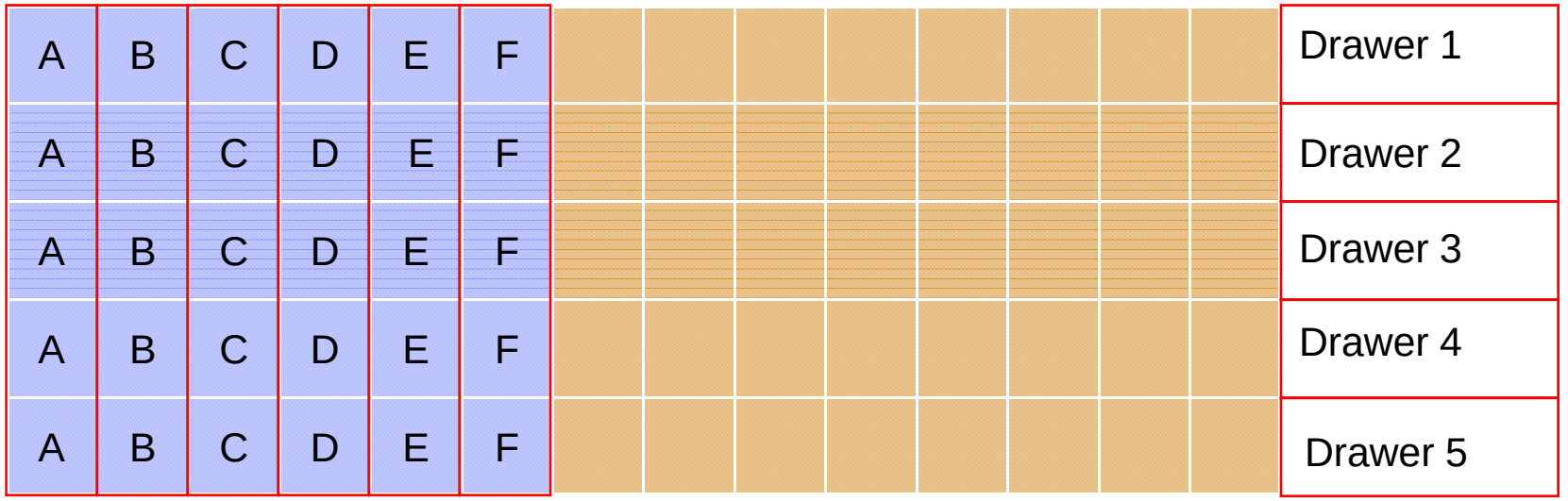
TSM Disk Tuning and AIX

- When designing the Storage Pools, try to spread them out across as many disks as you can. Two ways you can do this when using TSM Server on AIX: (See diagrams on following 2 pages)

  1. Use TSM to do the I/O spreading across the disks.

  2. Use AIX LVM to do the I/O spreading across the disks.

TSM Storage Pools with AIX – “Spreading the I/O” Using TSM

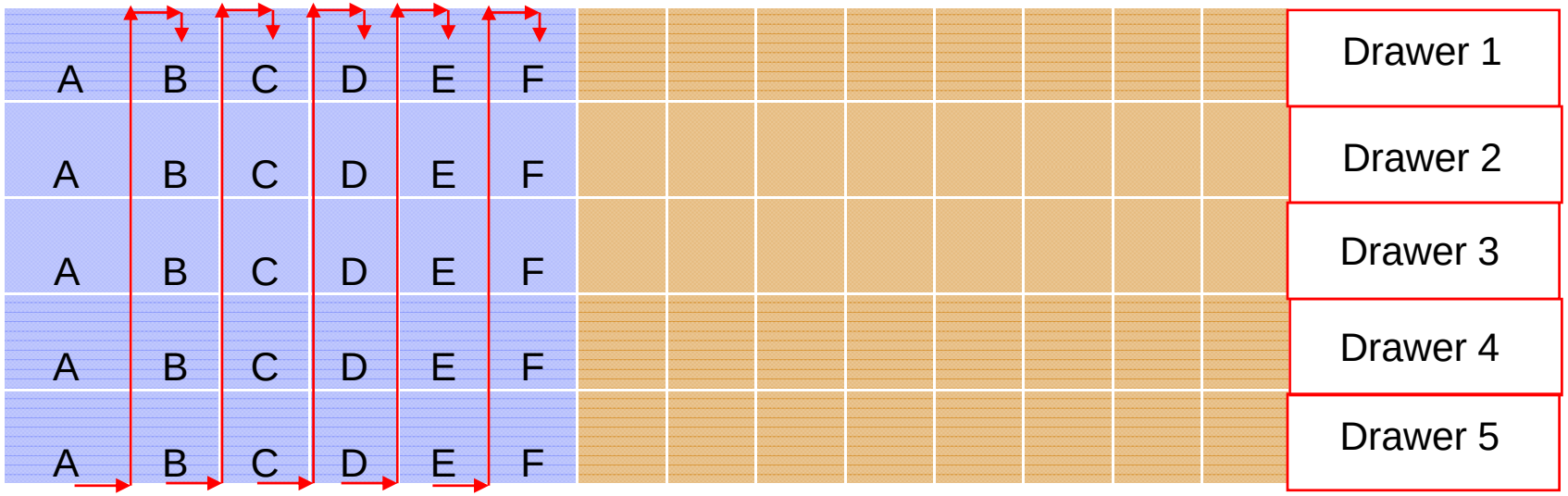


Technique #1 – using TSM to “spread the I/O”. In this diagram, we have:

- 6 LUNs of 4+1 RAID5 (LUNS A-F)
- Each LUN has 1 AIX Logical Volume and 1 Stgpool volume

Each storagepool volume spread across 5 volumes

TSM Storage Pools with AIX – "Spreading the I/O" Using AIX LVM



Technique #2 – using AIX to "spread the I/O". In this diagram, we have:

- 6 LUNs of 4+1 RAID5 – 6 Hdisks to AIX (LUNS A-F)
- LVs created with LVM striping, maximum spread across hdisks
- LV has 6 hdisks volumes and 6 TSM Stgpool volume defined in it

Total of 6 storage pool volumes spread across 30 volumes

IBM Software Group

Recommendations for TSM and Various Disk Technologies/Subsystems

Tivoli. software

e business software

Recommendations TSM and Disk Subsystems

- Raid types and use with TSM
- DS4xxx Recommendations
- SATA Recommendations
- SVC (San Volume Controller) and Storage Pools Recommendations
- SVC (San Volume Controller) and TSM DB/Log Recommendations

Mirroring / RAID 1 / RAID 5 / RAID 10

- TSM mirroring recommended over hardware mirroring
  - Partial writes can potentially corrupt TSM DB
  - Hardware mirroring is usually faster
  - Hardware mirroring comes at an additional cost.
- RAID0 is ideal for the recovery log (when used with TSM Mirroring)
- Buy more cache, this helps offset the extra reads and writes that occur with RAID5. Buy as much cache as you can afford.
- RAID5 vs. RAID10: RAID10 provides better read performance, at a much higher cost.

Cache Recommendations for using TSM with the DS4xxx

- **Consider** disabling write cache mirroring (WCM) on DS4xxx – understand implications of doing this
- Start/stop cache flushing should be set to 50/50
- TSM Database
  - Write cache enabled
  - Read cache disabled (may want to enable it for DB backup purposes)
- TSM Recovery Log
  - Write cache enabled
  - Read cache disabled
- TSM Storage Pools
  - Read/Write cache enabled (pre-fetch or cache read-ahead = 1 for DS4xxx firmware level of 6.1 or higher)

Segment Size Recommendations for using TSM with the DS4xxx

- TSM database
  - 64KB segment size
- TSM recovery log
  - 64KB segment size
- TSM storage pools
  - Set the segment size equal to the stripe size

Recommendations for using TSM with the DS4xxx (Parameters)

- Because of storage pools perform heavy sequential IO, you should:
  - lg_term_dma attribute of the fcs adapter should be set to 0x800000
  - Note – this was moved to slide 42, and this slide will be removed.

Recommendations for using TSM with AIX 5.3

- **AIX tuning parameters (no outage or reboot required).**
  - Set lru_file_repage=0 via vmo
  - Set j2_maxPageReadAhead=128 via ioo
  - Set maxclient=maxperm=80% (default) via vmo
  - Set strict_maxclient=1 (default) via vmo
  - Set minfree=max(960, 120*#lcpus/#mempools) via vmo
  - Set maxfree=minfree + max page ahead x #lcpus/#mempools via vmo
  - Set lru_poll_interval=8 via vmo.

Recommendations for using TSM with the DS4xxx (AIX)

- **Change HDISK queue depth (for both FC and SATA disks) - *Requires Reboot.***
  - Queue_depth=2048/(number of LUNS from the DS4800 controller)
  - Command to change the hdisk queue depth is:

    chdev -l <hdisk#> -a queue_depth=<new value> -P

- **Change Fibre Channel Adapter parameters - *Requires Reboot***
  - For FCs for disk

    num_cmd_elems = 512
    max_xfer_size = 0x200000
    lg_term_dma= 0x800000

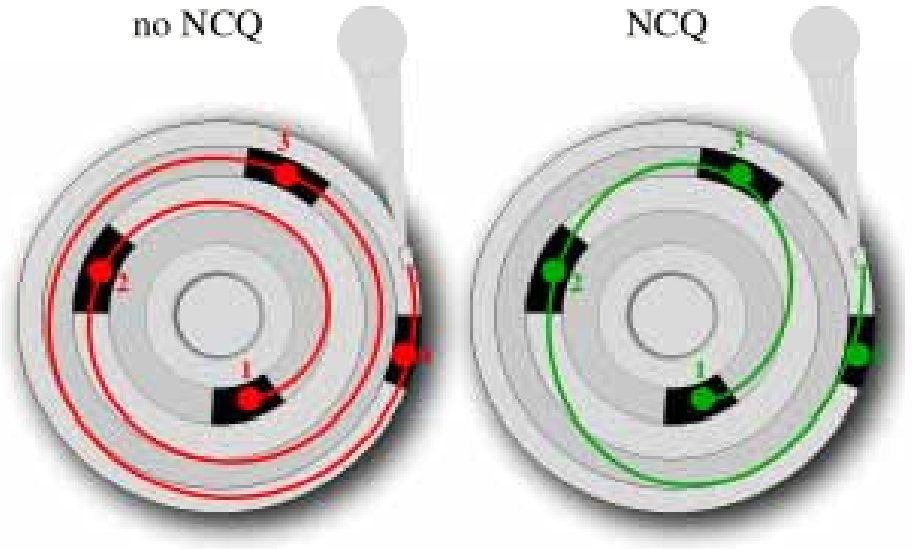
    command is: (should be all one line)
    chdev -l <fcs#> -a num_cmd_elems=512 -a
    max_xfer_size=0x200000 -P

Overall Findings from DS4800 Study

- Configuration using Vertical LUNs out performs Horizontal LUNs
- Configuration using striped LUNS Across LV used with DEVCLASS=FILE sequential volumes outperforms DEVCLASS=DISK random volumes
- Mirror write consistency impacts performance
- Mixing TSM components impacts performance of backups
- Segment size for storage pool LUNs should be equal to stripe size
- Sequential files out performs random disk
- DS4800 cache settings are different depending on how the LUN is used

SATA disk

- Recommend using fastest disks available for use with DB and Recovery Log
  - SATA-1 disks are generally 7200 RPM and can handle fewer IOPS
- SATA technology is still improving.
- Earlier generations had limitations
  - NCQ – (Native Command Queuing) optimizes order of I/O requests to disk
  - Adapter Compatibility



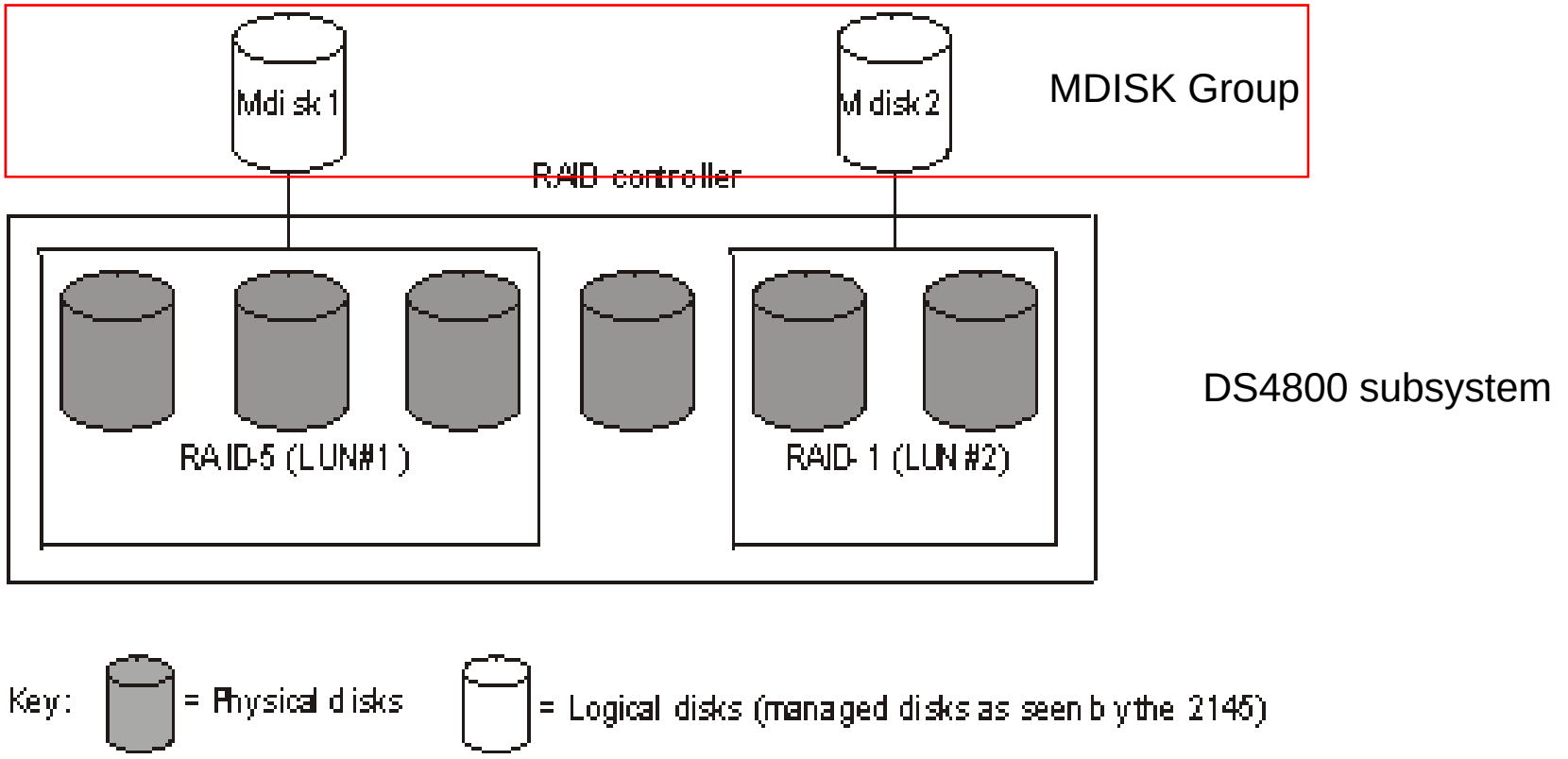
SATA disk

- Check your throughput requirements when evaluating SATA technology for use with storage pools.
- Good for less frequently accessed TSM type storage (HSM for example) or Tier-2 type storage.
- Design the storage pools so that you attempt to keep storage pool IO sequential all the way through to the hard disk drive
  - Keep a 1-1 relationship if possible between LUN-Logical Volume-File System-Storagepool volume

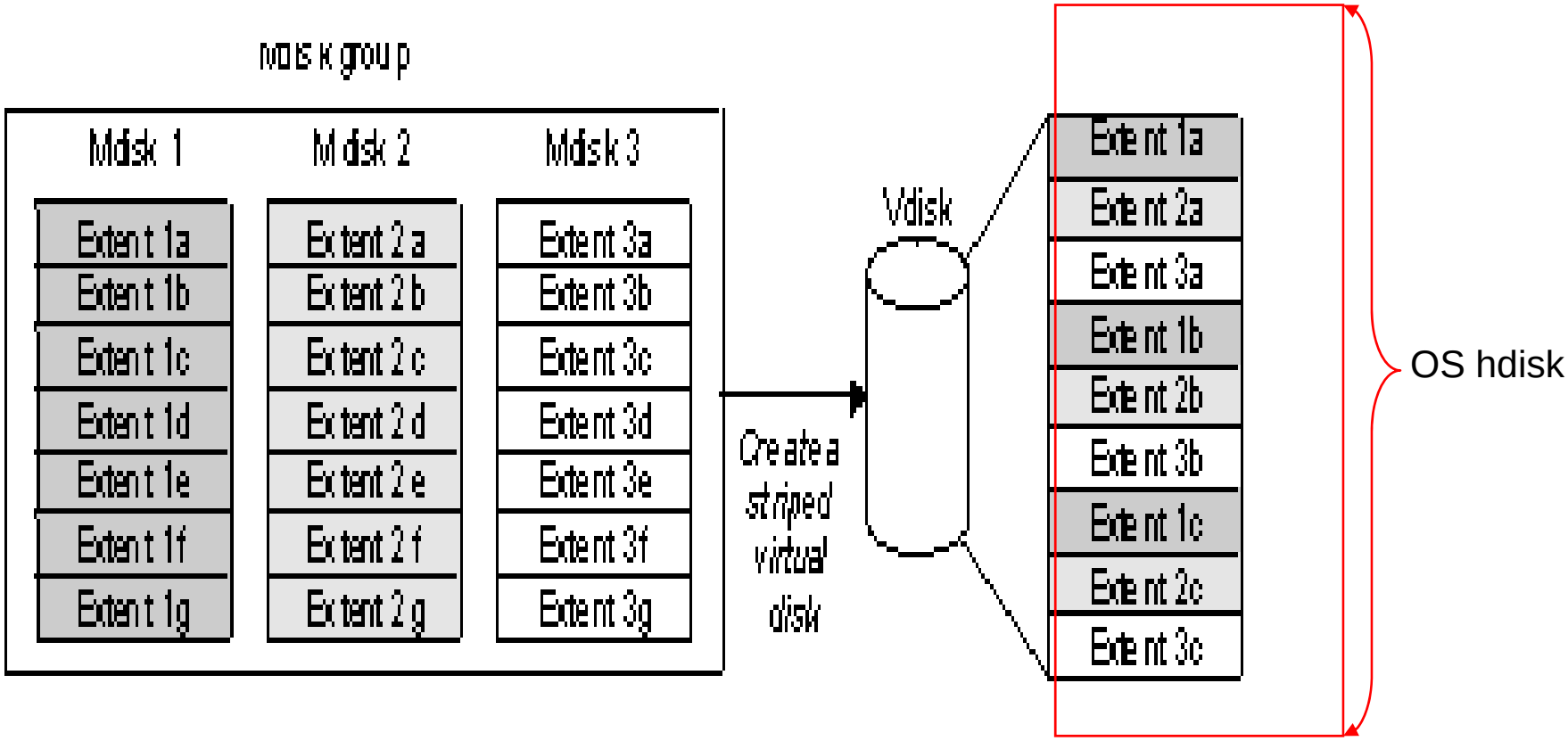
SVC (San Volume Controller)

- Still need more research in this area – some experience through performance related PMRs.
- Strive again for keeping the sequential throughput performance for storagepool volumes and high IOPS for database recovery log volumes

SVC (San Volume Controller) Overview



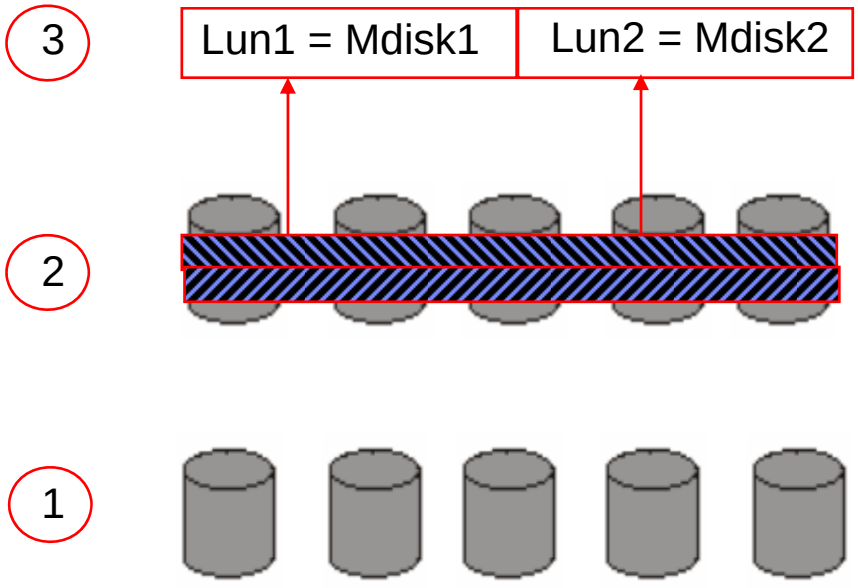
SVC (San Volume Controller) Overview



SVC + DS4xxx And StoragePools (SATA drives)

- Goal is to maximize the opportunity to maintain sequential performance and read ahead capability.
  - No more than 2 LUNS per SATA array
  - For 2 arrays, this would mean 4 mdisks
- DS4xxx - Stripe size / segment size: 256K
- DS4xxx - Read cache = YES
- DS4xxx - Write Cache=YES
- DS4xxx - Write Cache mirror=NO
- DS4xxx - Read Prefetch=YES. (value > 1)
- SVC – Read Cache = YES
- SVC – Write Cache = YES

SVC And Storage Pools (SATA drives) Sample customer

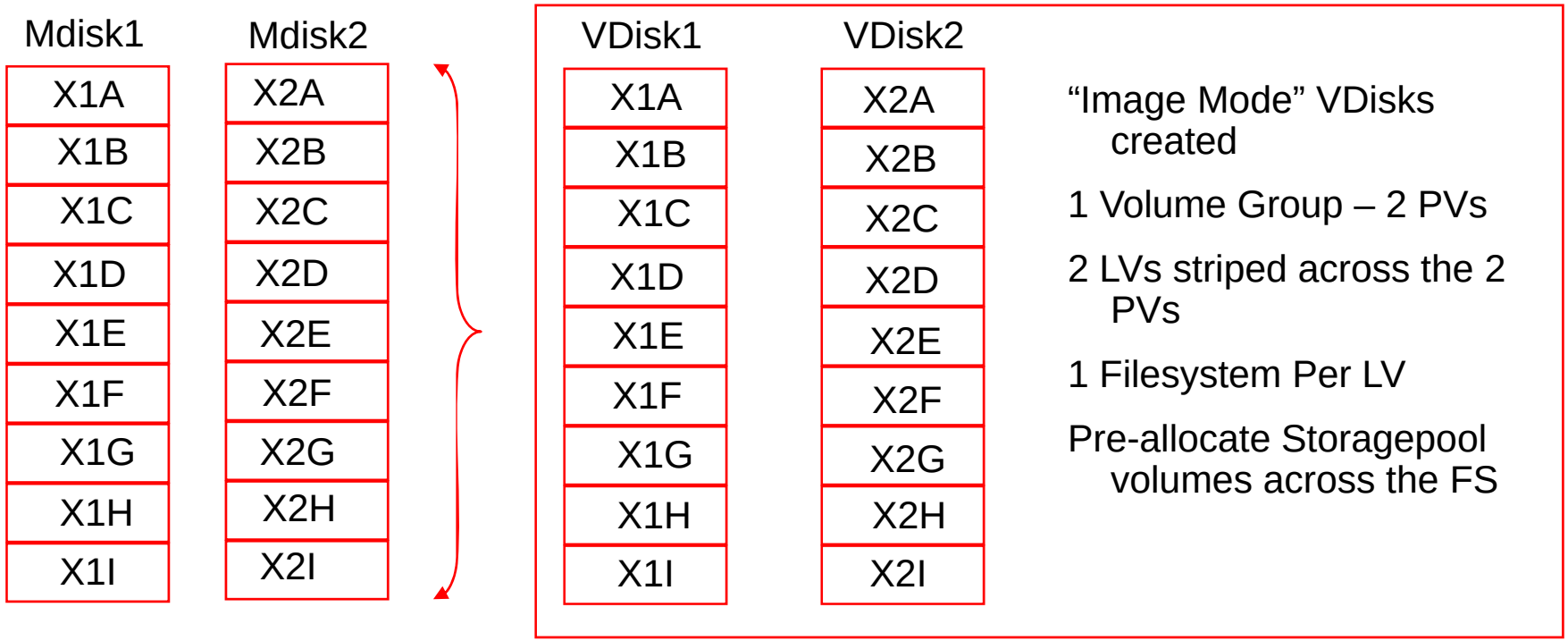


Step3: These 2 LUNS are presented to the SVC as 2 mdisks

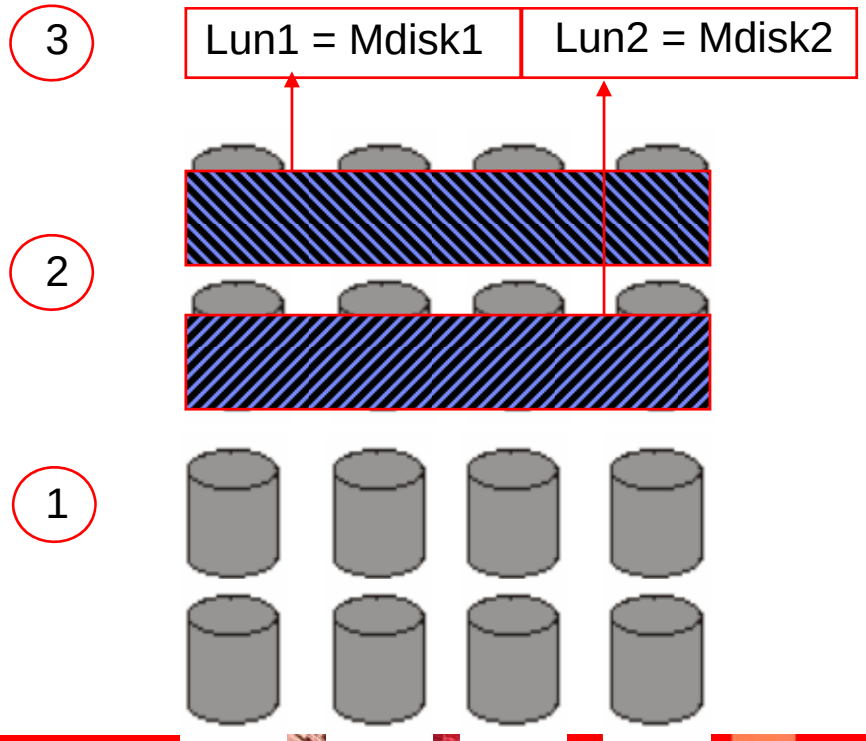
Step2: This array then has 2 LUNs carved from it, each 1TB in size)

Step1: Create 4+P 500GB RAID5 Array (~2TB)

SVC And StoragePools (SATA drives) Sample customer



SVC And Database/Log (Fibre drives) Sample customer (100GB Database)

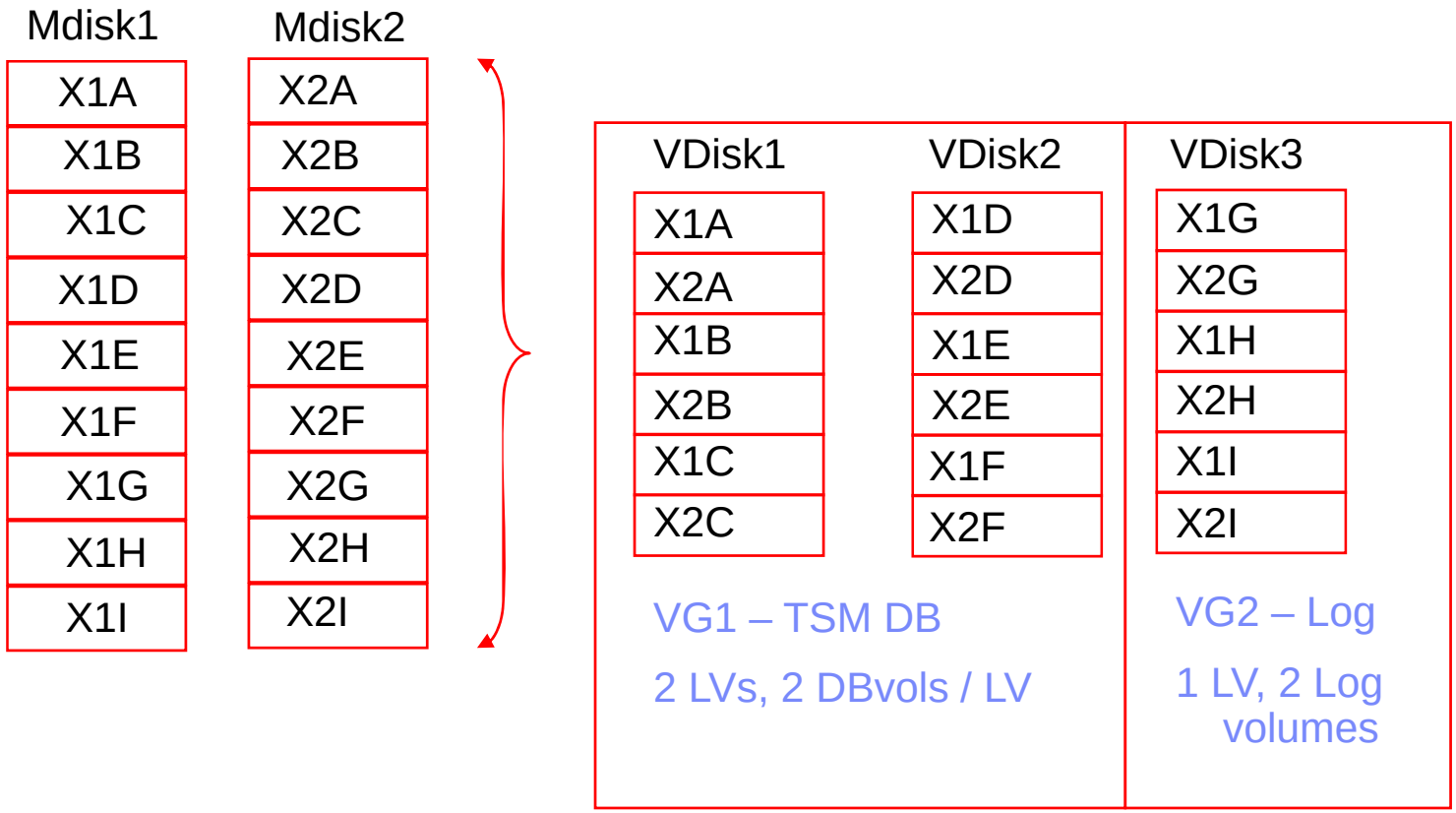


Step3: These 2 LUNS are presented to the SVC as 2 mdisks

Step2: This array then has 2 LUNs carved from it, each 400GB in size)

Step1: Create 2 3+P 146GB RAID5 Arrays (~800GB)

SVC And Database/Log (Fibre drives) Sample customer



SVC And Database/Log (Fibre drives)

- Again, goal is to maximize the **IOPS** to the fibre drives.
  - Consider wasting space on fastest drives to gain benefit of performance
  - Strive to isolate the components if possible or use extra space for install images or temp space
- DS4xxx - Stripe size / segment size: 64K
- DS4xxx - Read cache = NO
- DS4xxx - Write Cache=YES
- DS4xxx - Write Cache mirror=NO (Use TSM mirroring)
- DS4xxx - Read Prefetch=NO. (value = 0)
- SVC – Read Cache = NO
- SVC – Write Cache = YES

Issues Still Needing Further Study

Issues Not Addressed Yet

- RAW vs. JFS2 for Storage Pools
  - DIO might give better performance on backup, but may be hindrance on migration
    - Migration not studied in detail during study due to time
- Analysis of performance impact on using TSM mirroring
- Analysis of impact on using TSM with other applications
  - Has always been recommendation to isolate TSM on disk subsystem for DR
- Analysis of impact on sharing disks with other TSM Servers
- Extensive testing with different size arrays
- Need more testing with SATA-IO for use as Storage Pools

Questions?

TSM Disk Tuning
Reference Information

- Tivoli Storage Manager Performance and Tuning Guide
  - http://publib.boulder.ibm.com/infocenter/tivihelp/v1r1/index.jsp?toc=/com.ibm.itstorage.doc/toc.xml
- AIX 5L Performance Tools Handbook – SG24-6039
  - http://www.redbooks.ibm.com/redbooks/pdfs/sg246039.pdf
- NMON Tool (internal use only)
  - http://www.ibm.com/collaboration/wiki/display/WikiPtype/nmon