IBM VSE/ESA I/O Subsystem Performance Considerations

Wolfgang Kraemer
Hanns-J. Uhl

VSE Product Mgmnt
Dept 3221
71032-14 Boeblingen
WKRAEMER at DEVM
wkraemer at de.ibm.com

Update 2001-07-15

Contents

Contents

Contents

Contents

Contents

Notes

Notes

All information contained in this document has been collected and is presented based on the current status.

It is intended and required to update the performance information in this document.

It is the responsibility of any user of this VSE/ESA I/O document

- to use the latest update of this document
- to use this performance data appropriately

This document is unclassified and intended for VSE customers.

These VSE performance documents are also available from Internet via the VSE/ESA home page

http://www.ibm.com/servers/server/zseries/os/vse

(http://www.ibm.com/s390/vse/ former URL)

Starting with the VSE/ESA 2.4 documentation, these documents are also available on the VSE/ESA CD-ROM kit SK2T-0060.

The following documents are also available via Internet, in Adobe Reader format (.PDF):

'IBM VSE/ESA 1.3/1.4 Performance Considerations'
'IBM VSE/ESA V2 Performance Considerations'
'IBM VSE/ESA Turbo Dispatcher Performance'
'IBM VSE/ESA I/O Subsystem Perf. Considerations' (this document)
'IBM VSE/ESA VM Guest Performance Considerations'
'IBM VSE/ESA Hints for Performance Activities'
'IBM VSE/ESA TCP/IP Performance Considerations'
'IBM DFSORT/VSE Performance Considerations'
'IBM VSE/ESA CICS Transaction Server Performance'
'IBM VSE/ESA V2.5 Performance Considerations' (new)
'IBM VSE/ESA Performance on xSeries (NUMA-Q) Enabled for S/390'

The files are
VE13PERF.PDF, VE21PERF.PDF, VE21TDP.PDF, VEIOPERF.PDF, VEVMPERF.PDF, VEPERACT.PDF, VETCPPER.PDF, VESORTP.PDF, VECICSTS.PDF, VE25PERF.PDF, VEXEFSP.PDF

Notes ...

Disclaimer

This document has not been subjected to any formal review or testing procedures and has not been checked in all details for technical accuracy. Results must be individually evaluated for applicability to a particular installation.

Any performance data contained in this publication was obtained in a controlled environment based on the use of specific data and is presented only to illustrate techniques and procedures to assist to understand IBM products better.

The results which may be obtained in other operating environments may vary significantly. Users of this document should verify the applicability of this data in their specific environment.

The above disclaimer is required since not all dependencies can be described in this type of document.

Any statement in this document on 'supported by VSE x.y' does NOT replace the official statement in the Announcement. They represent the authors' technical view and do NOT consider any releases out of service and do NOT imply any guarantee beyond the pertinent Announcement.

Acknowledgements

Thanks to all who contributed directly or indirectly, be it by measurements, suggestions or in other ways.
Specific thanks is due to Chris Saul and Bill Worthington, SSD San Jose.

All mistakes and inaccuracies in this document are owned by us.

Please, as in the past, contact us if you have

- suggestions or questions regarding this document
- questions on VSE/ESA performance, not covered in any of the VSE/ESA performance documents

Wolfgang Kraemer, WKRAEMER at DEVM or wkraemer@de.ibm.com
Hanns-J. Uhl, UHL at DEVM or huhl@de.ibm.com

IBM VSE Development, Boeblingen Lab, Germany

Notes ...

Base Documents

This document essentially deals with VSE/ESA I/O performance aspects. It applies to all VSE/ESA releases, but e.g. the scope of ECKD support has always been increased from release to release.

It has been composed of existing charts from the other VSE performance documents and newly arranged and enhanced.

The VSE/ESA performance documents (see a previous foil) are also available to any IBM person, as part of the VE12PERF/VE13PERF/VE21PERF PACKAGEs on the same IBMVSE TOOLS disk. Contact your IBM representative to retrieve a copy for you by entering the following CMS command:

```
TOOLS SENDTO BOEVM3 VMTOOLS IBMVSE GET VExxPERF PACKAGE
```

These documents contain references to further VSE performance documents.

Trademarks

The following terms included in this paper are trademarks of IBM:

ES/9000	ESA/390	System/390	SQL/DS	PR/SM
VM/ESA	VSE/ESA	ESCON	ECKD	RAMAC
Multiprise	Magstar	Seascape		

Trademarks of other companies:

EXPLORE/VSE	Legent Corporation / Computer Associates
TMON/VSE	Landmark Corporation
ADABAS	Software AG, Germany
R/2	SAP AG, Walldorf, Germany
CACHE/VSE	Blueline Software Corporation
BIM VIO	Ben I. Moyle Corporation
OPTI-CACHE	Barnard Systems Incorporated
StorageTek, Iceberg, SnapShot	Storage Technology Corporation

Glossary

Glossary

CFW	Cache Fast Write A 3990-3/6 function which can be used for volatile data
CSS	Channel Subsystem An ESA architectural term for the total I/O subsystem. Also IOS is used for I/O Subsystem
DCME	Dynamic Cache Management Enhanced An MVS SMS function to dynamically cache on data set level
DFW	DASD Fast Write A 3990-3/6 extended caching function
DIM	Data in Memory A concept to store as much data as possible/reasonable in processor storage
EMIF	ESCON Multiple Image Facility Sharing of ESCON channels between PR/SM LPARs
ESS	Enterprise Storage Server An I/O subsystem with multiple platform attachment.
IDRC	Improved Data Recording Capability A data compaction feature for for tape subsystems
LIC	Licenced Internal Code Microcode as part of the H/W
LSR	VSAM Local Shared Resources A VSAM buffering method which allows that different files share the same buffers (Data, Index)
NSR	VSAM Nonshared Resources A VSAM buffering method with separate buffers per file
NVS	Non Volatile Storage
RAID	Redundant Array of Independent/Inexpensive Disks
RAMAC	RAID Architecture with Multilevel Adaptive Cache
RDF	Regular Data Format A 3990-6 exploited caching bit, for CKD/ECKD tracks with equal length records and w/o (H/W) keys

General References

Some General References

The following are general references for further performance information in the context of VSE I/O performance.

Introduction to Non-Synchronous Direct Access Storage Subsystems
GC26-4519-0, 01/90

Extended Count Key Data and Non-Synchronous DASD I/O
GC23-3571-00, 07/91

DASD Tuning and VSE Performance, by Thurman Pylant,
GUIDE 81, New Orleans, 11/91
As DASDTUN PACKAGE on the IBMVSE tools disk

VSAM Performance Tuning for the Experienced VSAM Tuner
VSE/ESA Techn. Conf. 05/95 Atlanta, by Dan Janda, __ pages

IBM Storage Subsystem Enhancements, G246-0011-00,
ITSO San Jose, 07/92, 206 pages (3990 and 9340)

S/370 and S/390 DASD Reference Chart, by Bill Worthington
DASDCHT package on MKTTOOLS disk
(IBM INTERNAL USE ONLY)

VSE/ESA Performance Management and Tuning, by Bill Merrow
1993, 385 pages, Mac Graw-Hill, J. Ranade IBM Series, ISBN 0-07-041753-9

Balanced Systems and Capacity Planning, GG22-9299-04,
by P.T. Borchetta and Ray J. Wicks, 08/93, 125 pages

IBM Storage Systems Update for VSE and VM, by Bill Worthington,IBM
VM/VSE Tech Conf Reno, NV, 05/98

IBM Storage Systems Update, by James Cosentino, IBM.
WAVV Albany, NY,09/98

For device specific references, refer to the individual reference lists at the begin of the individual chapters of this document.

All references to documents (e.g. Presentation Guides) on IBM disks cited here are intended for use by your IBM representative in discussions with you on individual products. Contact him if specific need arises.

Refer also to the VSE/ESA information on the Internet:
http://www.ibm.com/s390/vse/

Introduction and Overview

PART A.

Introduction and Overview

VSE/ESA Performance/Capacity Evolution

VSE/ESA Performance/Capacity Evolution

```
                                                  ESA-mode only
                                                + Turbo Dispatcher
                                                + H/W Data Compression
                                                + POWER Perf.Improvmnts
                                                + VSCR
                                                + VTAM 4.2
                                                + LE/VSE (incl. PL/1)
                                                + Many new fctns
                                                    ___________________
                                                   |
                     Even more Real/Virt.Storage   |  VSE/ESA 2.1-2.3
                       "    "  Private Space       |
                     + 31-bit Applications/Buffers |  CMOS exploit.
                     + 31-bit Internal Fctns       |  + new fctns
                     + Data Spaces                 |
                     + Virtual Disk                |
                     + CICS Data Tables            |
                     + Extended Caching Fctns      |
                          _________________________|
                         |
More Real Storage        |      VSE/ESA 1.3, 1.4
  "  Address Spaces      |  -> ESA Exploitation,
  "  Total Virt. Storage |      mainly for
  "  Partitions          |       - VSCR
  "  Private Space       |       - DIM
+ Dynamic Channel Subs.  |
+ ESCON and ECKD support |
     ____________________|
     |
     |  VSE/ESA 1.1, 1.2
     |   -> entry to ESA,
     |      w/o 31-bit
     |
     |
     |
_____|

VSE/SP
```


ESA Exploitation Basics

- **Do NOT run VSE/ESA 1.3 and up with the same setup and parameters as you did before**
  (except for temporary migration reasons)
  - **Take the chance to exploit ESA for YOUR benefits**
    (apart from VSCR)

    1. to SAVE I/Os

    - follow the concept of DIM
    Refer to the charts on DIM in base document

    'Fastest I/O is no I/O'

    2. to SPEED UP I/Os
    I/Os you cannot save

    - use faster I/O attachments/devices
    ESCON, DASD caching, dynamic path selection by H/W
    - use better setup of I/Os (ECKD, blocking)

    3. to MORE OVERLAP I/Os
    - set up more concurrent partitions/tasks
    E.g. more concurrent batch partitions

- **It is wise to apply these ESA concepts for your own benefits**
  Refer to the VSE/ESA Exploitation Checklist of 'IBM VSE/ESA 1.3/1.4 Performance Considerations'
- **VSE/ESA with 31-bit may need much more CPU-time if the 'detuned' I/O setup from the 24-bit environment is kept**

I/O Subsystem Performance Tasks

I/O Subsystem Performance Tasks

- **I/O Trouble Shooting (Sporadic)**
  - **Analyze and solve sporadic performance surprises**
    Mostly after configuration or setup changes
    Mostly for partial workloads or jobs or txns
    The faster the problem can be solved the better, problem often not customer specific
- **I/O Performance Management (Short term)**
  - **I/O tuning with existing total load and I/O configuration**
  - **Mostly systematic search for I/O bottlenecks, based on regular monitoring**
  - **Changing data set distribution, caching, buffering ...**
    Customer specific actions, following general tuning rules
- **I/O Capacity Planning (Long term)**
  - **I/O planning for future growth and requirements**
  - **Recognizing and extrapolating long term workload trends**
    I/O rates and relative intensities, buffer sizes, load changes ...

What is This Document For?

What is This Document For?

- **How to optimally exploit IBM I/O attachments**
- **How to achieve 'best msec per I/O'**
  - for individual situations/jobs/tasks
  - overall
- **Show how channel programs have been optimized in subsequent VSE/ESA releases**
  - VSE/ESA 1.2 was the first release with ECKD support
  - VSE/ESA 1.3 supported the Extended Caching Functions
  - VSE/ESA V2 has the broadest spectrum of optimal I/O support
- **Provide some hints for optimal channel programs**
  For those
  - who have to setup channel programs (VSE developers, including vendors)
  - who want/need understand performance impacts
- **Give guidance how to proceed in case of specific I/O performance problems ('Trouble Shooting')**
  - Optimal I/O buffering ...
  - Optimal setup of DIM ....

  is considered in the base documents
- **Provide insight into the anatomy of I/O times to understand/tune I/O subsystems**
- **Describe general and I/O subsystem specific performance relevant items**

Impact of I/O on Overall Performance

„ **Overall System Performance**
(i.e. ET/RT at a given total system load)
depends on

CPU-time component
I/O-time component
Other resources (locks on user/system res.)

Includes queueing (wait) times.

```
                       Overall Runtime
    (Batch Elapsed Time ET or Tx Response Time RT)
                             |
     ----------------------------------------------
     |                       |                     |
 CPU-time             Total I/O-time          Other delays
     .                       |                     .
     .      ------------------------               .
     .         |                       |
               |               ===============
            # I/Os             Msec/IO or IORT
               .               ===============
               .
               .
            (DIM)
```

```
  Batch job  Elapsed Time  or  Online Tx  Response Time:


  CPU    CPU        Other         Total I/O-time
  wait   time       delays        (#IOs x IORT)
 |.....|----------|........|===========================|


 Other delays may be
  - waiting for a locked resource
  - caused by VM, not dispatching the VSE guest

 Simplified figure does not show potential CPU/IO overlap
```


I/O Performance Problems

I/O Performance Problems

In case of performance problems,

í **High(er) ET/RT may be caused by high(er) total time spending in the I/O Subsystem:**

More I/Os
and/or
Slower I/Os
(higher device service time)

This document deals with the 2nd aspect

í **In any case, reducing the number of I/Os via DIM (Data In Memory) is the most effective way:**

'Fastest I/O is NO I/O'

I/O Response Time Information

Ù **Performance Monitors**
To determine I/O response times (and its components), a VSE system monitor is required, e.g.

EXPLORE/VSE	from Computer Associates
TMON/VSE	from Landmark

Ù **VSE SIR SMF command**
For trouble shooting only, VSE/ESA 2.1.x. and up.
Refer to chart D3 'I/O Response Times from SIR SMF'

Ù **Estimates**
In exceptional cases approximate total I/O reponse time can be estimated (e.g. in single thread, neglecting (seldom) IO/compute overlap)

I/O Response Time in a Nutshell

IORT (at a single glance)

```
 IORT      =  Wait_in_VSE_channel_queue

           +  Wait_in_Channel_Subsystem

(Uncached) +  Device_seek + Rot.delay  + RPS_miss_time
           +  Device_Channel_transfer_time

 (Cached)  +  %Cache-miss x Cache_miss_resolution_time
           +  Channel_transfer_time
```

All components of IORT are being discussed in detail in section 'I/O Response Time Component Analysis'

Refer also to that section

- if required/beneficial on top of the hints given here for 'trouble shooting'
- if you want or need to do a more systematic 'performance tuning'

What are unacceptable I/O Response Times?

„ **User Definition**

'Any value that is worse than my expectation'
'Any value that is worse than what I had before'

„ **Technical Definition**

'Any value, be it average or individual, that is technically not explainable with the real potential of the I/O attachment with optimal channel programs'
'Any value that simply is caused by a too high I/O-rate to a device or by a too high path utilization'

For rough values, you simply can expect, refer to Foil D2 'Rough Values for Device Service Time Components'

How to Proceed in Case of High IORT

To observe

„ **Always relate msec/IO to KB/IO**
or to the function performed, (e.g. VTOC rest-of-cylinder search)

„ **For track oriented sequential operations, calculate the #revolutions per read/written track**
Applies epecially to native (non-simulated) devices

Things to Differentiate when IORT is High

To faster localize and solve the problem ...

Try (if possible) to differentiate between

„ **Single and Multi-thread**

It simplifies the analysis, if the msec/IO problem already occurs in single thread
(only 1 task/partition is issuing I/Os to the disk)

„ **Logical and Physical Device Utilization**

S/390 logical devices (seen by the S/W) may be mapped into/to different physical devices or HDDs, e.g. RAMAC or Internal Disk.

In any case, the S/390 logical device utilization (from S/W monitors) should not be too high, since only 1 SSCH can be active per S/390 logical device.

In such a case, as holds already for cached devices, physical HDD utilization is not directly visible from S/390 S/W, but may be a bottleneck if the HDD utilization is high.

For RAID-5 this should not be a specific problem due to load balancing via RAID-5 across all HDDs of a drawer.
For Internal Disk (RAID-1) several S/390 volumes reside on a single 9G HDD (and its mirror), thus this can be a problem.

How to Proceed in Case of High IORT ...

Things to Differentiate (cont'd)

- **READ and WRITE**

  WRITEs may even be subdivided into 'Format' and 'Update' WRITEs, of special impact for cached subsystems

- **Different program products or type of DASD accesses (random, sequential, ...)**

- **VSE and VM/ESA CP**

  I/O timings and caching bits may be different, e.g.

  - a higher VSE I/O Response Time may indicate high CP overhead (check VM CCW translation, refer to VM/VSE performance doc.)
  - VM/CP may not allow a VSE guest to use certain cache functions (if so, correctly set the guest parameters in VM)

- **Statistical overall values vs individual single values**

  Performance monitor results vs an individual trace entry

- **The following data may give more insight in case of an IORT problem:**

  ```
  IOSQ        Time    Time waiting in VSE channel queue
  PENDing     Time    Time between SSCH and first CCW executed
  CONNect     Time    Channel connected to 'device'
  DISConnect  Time    Channel disconnected from 'device'
  ```

  The anatomy of I/O Response Times is widely discussed in the section 'I/O Response Time Component Analysis'

IORT Checklist

How to Proceed in Case of an IORT Problem?

Checklist predominantly is for trouble-shooting, not for standard I/O performance tuning (discussed in part D)

Spectrum and Sequence of Checks		
	DASD Attachment	
To check	Cached	Uncached
Problem in Single or Multi-thread:		
1. VSE device type (VOLUME cuu)	X	X
2. Actual cache settings (VM+VSE)	X	-
3. Cache hit ratio(s)	X	-
4. VM settings of I/O relevance	X	X
5. S/W and H/W levels (PTFs...)	X	X
6. ECKD channel programs	X	X
7. Cache bits/Mask byte in DX CCWs	X	-
8. EREP incidents	X	X
9. IOCDS definitions	X	X
10. Sector value settings	X	X
Problem in Multi-thread only:		
11. Device utilizations (logical/physical HDD)	X	X
12. Channel/path utilizations	X	X
13. Cache sizes and hit ratios	X	-
- Sequence of checks is suggested, not mandatory - Sector value settings is really last, done only if lost revolutions are the only chance		

Checks in case of I/O Degradation

Checks in case of I/O Degradation

To Check 1:

- **Check the actually used device type**

  All simulated 3380s, 3390s for the RAMAC family and the Internal Disk must be ADDed for performance (and functional) reasons as ECKD.

  VSE/ESA 2.1 shows the device type in the VOLUME command display, before VSE/ESA 2.1 it is part of the 'PUB-table'.

  If device type is 3380='6C', find out why. May be you ADDed it as '3380,EML' what you should only do temporarily to prove a vendor product deficiency (see separate item)

  (This is also true for uncached attachments)

  Contact the vendor if channel programs are setup by the vendor. Contact IBM if reason is not obvious.

Checks in case of I/O Degradation ...

If I/O Degradation Occurs for Cached Attachments

To Check 2:

- **Check that all cache settings are active/enabled**

  (depends on subsystem type)

Applicable for	3390-3/6 with any DASD	9345 cached (no DFW)	RAMAC Array Subsys a)	9390 with RAMAC3	RVA RSA	Multi. Int. Disk b)
Basic caching						
Device level ACTIVE	x	-	-	x	x	-
Subsys level ACTIVE	x	-	-	x	x	-
DASD F. Write ACTIVE	x *	-	-	x *	x	x **
NVS ON	x	-	-	x *	x	x **

- Basic Caching on device and subsystem level (Device/Subsystem Caching) required both for READ and WRITE
- DFW (device level) and NVS (subsystem level) are for WRITEs
- Settings are done in CU via Set Subsystem Mode '87' CCW

* Especially check after RAMAC device replacement/migration

a) DFW done fully transparently, VPD nonsync mode must be set in H/W (RAMAC Array Subsyst) (i.e. none of the synchronous modes must be set). Carefully check EREP. If drawer battery low, no DFW is being performed. Service alert message not obvious

b) DFW available as IDFW or IDRFW. Only the ',STATUS' command can be used to query Internal Disk

** Setting only possible, if at all, in u-code

- If -for I/O subsystems which always use the cache- caching is not enabled, an immediate de-stage may occur

For VM/VSE, checks must be done not only from VSE's but also from VM's view.

Checks in case of I/O Degradation ...

Checks in case of I/O Degradation (cont'd)

To Check 2 (cont'd)

For VM:

To query cache status, use:

```
QUERY CACHE cuu      Gives 'available for subsystem'
                      plus 'activated for device'
QUERY DASDFW cuu     Gives DFW status 'active/inactive'
QUERY CACHEFW cuu    Gives CFW status 'active/inactive'
QUERY NVS cuu        Gives 'available/unavailable'
```

To set cache status, if required, use also:

```
SET CACHE DEVICE ON cuu  Activates device level caching
SET CACHE SUBSYS ON cuu  Activates subsyst. lvl caching
SET DASDFW ON cuu        Activates DFW on device level
SET CACHEFW ON cuu       Activates CFW on subsystem lvl
SET NVS ON cuu           Makes NVS available on subsystem
```

VM may enforce BYPASS CACHE if e.g. the NOCACHE option is defined in the MINIOPT directory statement for the Partial Pack Minidisk.

VM never changes cache related settings in a DEFINE EXTENT CCW for a DEDicated device

The ability to change/set caching status from a guest depends on the control level defined for the device, and on the type of device usage:

- o DEVCTL/SYSCTL for Full Pack Minidisks (FPMs) and DEDicated devices (DEDs)
- o For Partial Pack Minidisks (PPMs) guests cannot change e.g. caching status.

Caching status and statistics can be obtained for any type of disks.

Be aware that under VM, VSE caching statistics for PPMs represent the values for the total device.

Checks in case of I/O Degradation ...

Checks in case of I/O Degradation (cont'd)

To Check 2 (cont'd)

For VSE:

To query cache status, use:

```
CACHE UNIT=cuu,STATUS   and   CACHE SUBSYS=cuu,STATUS
```

In case of SUBSYS=cuu,

- any cuu at the same logical subsystem can be used
- also the configured and available cache sizes are shown

To set cache status, if required, use:

```
CACHE UNIT=cuu,ON         CACHE SUBSYS=cuu,ON
CACHE UNIT=cuu,FAST,ON    CACHE SUBSYS=cuu,NVS,ON
```

> If uncached or cache setting is OK, ...

With some probability, no native/optimal VSE ECKD channel programs are used for the non-sync attached devices.
In the very end this only can be proven by CCW traces in VSE.

Refer to 'VM/ESA Planning and Administration', SC24-5750
'VM/ESA CP Command and Utility Reference', SC24-5773

Checks in case of I/O Degradation ...

To Check 3:

- **Total Cache Hit ratios should be >80% (ROT)**

For certain individual types of I/O requests it can approach even 100% (e.g. update or format WRITEs for 'RDF tracks').

Add more cache if this is the problem.

See 'Cache Size Considerations' in IORT Component Analysis part

To Check 4:

- **VM settings of I/O relevance**

These settings affect the msec/IO seen by VSE.

If msec/IO (and/or number of I/Os) seen from VM differs from earlier results,

- make sure that VM MDC (minidisk caching) is in effect the same way as before
  QUERY MDC MDI cuu
  (check MDC hit ratios and remaining number of physical I/Os via VM Monitors e.g. RTM/ESA, VMPRF)
- make sure that (real) device is not throttled
  QUERY THROTTLE

If the msec/IO delta seems to stem from more CP-time per I/O,

- check the DASD definitions in VM (FPM, PPM and DEDs)
- check the SIE (I/O) assist status
- check the VM CCW translation status/stats

Refer e.g. to the document
'IBM VSE/ESA VM Guest Performance Considerations'

To Check 5:

- **Refer to official S/W APARs/PTFs and to H/W EC-level patches**

For first APAR overview, refer to this document.
In any case, IBMs RETAIN is to be consulted via Level-1 support

Checks in case of I/O Degradation ...

To Check 6/7:

- **Check channel programs (READ & WRITE)**

  - use of ECKD
  - caching bits/mask byte in the DX CCWs

Use SDAID to trace I/O operations

```
SDAID
OUTDEV T=cuu-tape
TRACE IO UNIT=cuu OUTP=(TOD CCWD=16)
TRACE SSCH UNIT=cuu OUTP=(TOD CCWD=16)
READY
STARTSD
//////////////
STOPSD
ENDSD
```

Usually CCWD=16 byte is sufficient for performance analysis.
To limit the amount of output, you also may specify
AReа=partition-id
OPTion=OCcurrence=1:50 (e.g. for first 50 occurences)

SDAID output also shows blocksize and target location.

```
Look for CCW '63' = DEFINE EXTENT DX
         CCW '47' = LOCATE RECORD LR  (native ECKD)

CCW= addr. 63------ .. DATA= MmGgDddd 000000Xx Bbbbbbbb Eeeeeeee
CCW= addr. 47------ .. DATA= -------- -------- CcccHhhh Rr------

  Mm       Mask byte (Chart B8)
  Gg       Global Attributes byte (Chart C8)
  Dddd     amount of data:  0400 =  1K, 1000 = 4K, 4000=16K
  Xx       Global Attribute Extended byte
  Bbbbbbbb begin of the Extent address
  Eeeeeeee end      - " -
  CcccHhhh target track, Rr target record  (search argument)
```

For VSE Supervisor-converted CKD/ECKD channel programs:

```
 IO   trace  shows the CKD channel programs only,
             NOT how it was converted
 SSCH trace  shows the actual used (converted) CCWs
```

For IBM VSE/VSAM and VSE utilities, the DASD cache bit usage is as shown on Chart C20 and C21 in the General DASD Caching part.

Checks in case of I/O Degradation ...

Checks in case of I/O Degradation (cont'd)

To Check 11:

„ **Check device utilizations**

Refers to logical volumes and/or physical HDDs, depending on the type of I/O subsystem.

Is only relevant for concurrent access from multiple partitions.

High Device Service Time (DST):

Caused usually by a too high I/O rate to a logical/physical volume.

> Reduce the number of I/Os to such files (I/O Blocking)

> Relocate the files to another volume/HDD

High IOSQ Time in VSE Channel Queue:

Caused by a too high logical device utilization, seen from VSE.

> Proceed as above

> Use PRTYIO settings if device utilization cannot be reduced and Online I/O has to be preferred

Refer to

- 'VSE/ESA Workload Balancing' in the VSE/ESA V1.3/1.4 doc,
- 'PRTYIO Usage Hints' in 'Part D' of this document.

I/O Tuning Logical Flow

I/O Tuning and Rules-of-Thumb (ROT)

Any values for ROTs hold e.g. for a 15 min peak load

A) Reduce #I/Os

- Apply Data In Memory (DIM)
- Discussed in the base documents

B) Check I/O response times

- Use a VSE performance monitor (best), or SIR SMF
- Check IORT values
- Proceed in case of 'High Msec/IO' as described before

C) Check device utilizations

(multi-thread only)

- Check device utilizations from the monitor runs
- Balance device utilizations and/or increase number of devices

ROT: Any (physical or logical) device utilization should not permanently exceed 30%

I/O Tuning Logical Flow ...

I/O Tuning and Rules-of-Thumb (cont'd)

High device utilization harms if more than 1 task/partition issues I/Os to that device

No S/390 operating system issues 2 concurrent SSCHs to the same physical/logical device

Actual (real) device utilizations for cached devices cannot be determined exactly by S/W (cache hides device).

But wait time in VSE until IO is issued only depends on the device utilization seen by VSE

D) Check (channel) path utilizations

(multi-thread only)

- Use data from monitor runs
- Balance channel utilizations and/or add channels/paths
- Item is less critical in case of multi-path

ROT: Max. recommended channel path utilizations:

		Uncached	Cached
Single path	Parallel ESCON	25% 35%	30% 40%
Multi-path (2-/4-way)	Parallel ESCON	40% 50%	45% 55%

- All values are rough estimates only (relative values are the main message here)
- Actual overall values may be higher, especially for batch workloads
- Balanced multi-path is fine, but not so important (Channel Subsystem dynamically selects paths)
- Probability 'at 50% 4-way path utilization ALL paths busy' is 6%

- Make sure that all paths are active: Use the IBM processors path utilization H/W display, or check inactive paths via 'STATUS cuu'

I/O Configuration Rules

Remark

More
- ROTs
- tuning rules (also for cached I/O subsystems)

are contained in chapter D 'I/O Response Time Component Analysis'

General Rules

„ **Adjust the I/O capacity (performance-wise) to the increased speed of your processor**

ESCON channels
Cached I/O attachments
Number of paths (channel and device paths)
Number of physical devices ('HDDs')
Number of logical(=simulated) devices ('CUUs')

Use as many simulated devices (RAMAC family) as possible in order to reduce wait time in operating system

„ **Do not mix DASD and tape on same channel**

ESA I/O Channel Subsystem

ESA I/O Channel Subsystem

- **All channel path related data is handled by H/W (Channel Subsystem CSS, or IOS)**
- **Performance/capacity functions**

Dynamic Path Selection (DPS)

Any path out of up to 8 associated with the target device can be selected to initiate an I/O.

The CSS uses a rotation order for the initiation of I/O requests to a device.
Also you can select a preferred path which the CSS always tries first.

In S/370-mode, only up to 2 alternate channel paths could be defined, but this had to be done in the operating system itself.
Also, the selected channels had to be consecutive, what is not required for the CSS.

Since the CSS knows and handles the status of all channels, there are no cases of 'channel busy' as they existed for S/370 I/O.

Dynamic Path Reconnect (DPR)

Any channel path out of up to 8 can be selected to perform the actual transfer of data, not just the path the operation was started.

An RPS miss (if non-cached) only occurs when all eligible paths are busy.

On S/370, the data had to be transferred on the same channel on which the I/O operation was initiated.

> ESA operating system can handle different paths for a single I/O

Support of up to 256 channels/CHPIDs

S/370-mode only allowed up to 16 physical channels

ESA I/O Channel Subsystem ...

- **ESA I/O Performance Benefits**

  Reduced utilization of channels
  Mostly for ESCON

  Higher I/O capacity for a given number of channels
  or
  Better I/O service times at same I/O-rate per channel

  Reduced tuning/balancing requirements in case of channel bottlenecks

- **Performance Results**

  with the Channel Subsystem are history and documented e.g. in 'MVS/XA I/O Performance Considerations'.

  For example: At 33% channel utilization, the use of DPR provided an 11 msec (or 39%) improvement in the I/O response time.
  Using DPR also permitted the channels to run at about 20% (absolute) higher utilization and still maintain an I/O response time between 17 and 21 msec (3380 Std).

- **Actual performance benefits depend on:**

  - Channel path utilization
  - Number of channel paths
  - Setup of the device paths from the control unit to the devices with DLSE (Device Level Select Enhanced, provides 4 independent and simultaneous data transfer paths to a single DASD string)
  - Relative I/O intensiveness of a workload
  - Processor type

General ESCON Statements

Big Functional Benefits

Distance

Cabling

Configuration flexibility
Dynamic (=logical) connections between ESCON channels and CUs via
- EMIF and/or
- ESCON director(s)

> Savings of channels, cables and CU channel ports

Performance Benefits

by higher data transfer rate

are limited, since the 'msec per I/O' only partly depend on the channel transfer rate

are more visible if effectively cached

would also show up with fast (uncached) devices

directly show up if channel becomes a capacity bottleneck

allow less channels if ESCON

Please also distinguish between 17 MB/sec channels and 10 MB/sec channels (e.g. elder 9121s)

ECKD vs CKD Channel Programs

PART B.

ECKD vs CKD Channel Programs

- **What is ECKD?**
- **Why ECKD?**
- **Performance Relevance**
- **Scope of Usage by VSE**
- **More Hints for Optimal ECKD Use**
- **ADDing Disk Devices**

ECKD Basics

What is ECKD?

- A channel command (CCW) architecture to optimally use new device attachments/devices
  - Predictive, to avoid CCW chaining in gap

    DEFINE EXTENT CCW (DX, hex63)

    - Defines extent limits on operations that follow
    - Provides block-size value
    - Specifies cache controls

    LOCATE RECORD CCW (LR, hex47)

    Specifies
    - Location of first record (incl. sector value)
    - Number of records
    - Type of operation
  - Optimal for 'Non-synchronous operation'

    Data are transferred from device to real storage in a stepwise fashion, at individual speeds
  - New commands

    READ TRACK
    More multi-track CCWs
    ...
- ECKD channel programs are handled on an extent basis, in CKD channel programs each CCW is handled independently

ECKD Basics ...

Why ECKD?

For more info refer to 'VSE/ESA 1.2 Performance Considerations', Chart PB06

- Avoid performance degradation for WRITE CCWs (if uncached) for non-synchronous attachments (ESCON)
  - Maintain device performance of todays DASDs i.e. no performance improvement vs CKD
- Use specific performance beneficial new ECKD commands (e.g. Read Track, or more multi-track CCWs)
  - Performance improvements possible
- Provide performance optimal device support of devices with increased device data rates
  - Exploit fast, non-sync attached devices optimally
- Allow optimal S/W control of cached subsystems

Allow 'smaller' gaps on track (higher track exploitation)

No need anymore for command chaining of CCWs on the fly while gap is passing by.

This is transparent to S/W.

ECKD for Performance Reasons

ECKD channel programs required performance-wise ...

- Whenever the DASD attachment is non-sync and w/o write cache

  3380s & 3390s at ESCON, 9345s
- Whenever a DASD cache must be optimally exploited/controlled by S/W

  - 3990-3, 3990-6 cached subsystems, incl. RAMAC Array DASD, 9390s (*)
  - 9345 cached subsystems
  - RAMAC Array Subsystem (*)
    RVA, RSA
    Internal Disk

  (*) Even newer cache implementations benefit from bit settings

Notes

- Caching bits and their functions are described in the part 'DASD Caching in General'
- The newer the VSE/ESA release, the more cache performance functions are exploited

Functional Reason for ECKD

- Nonsync attachments (split CE and DE)
  The CU causes the channel to present split CE and DE.
  Not new in general, but not exploited for DASD WRITEs in the past
- Potential data integrity exposure if ADDed as CKD

  (exception conditions reported at DE, but application has made no precautions to force to be posted only at DE time)

ECKD Usage in VSE/ESA

VSE/ESA Components using Native ECKD

- VSE/VSAM
- LIBRarian and FETCH/LOAD
- FAST COPY
- Page Manager
- Lock Manager
- Hardcopy file

Release details are discussed later

Notes

- Native ECKD is used by VSE/ESA only if device type is '6E'

  For all type of non-synchronous and/or cached DASD attachments, VSE should run with 'device type' '6E'!

  6E: ECKD channel programs (3380, 3390, 9345, RAMAC, ID)
  6C: CKD channel programs (3380 only)

  For 3390s and 9345s, VSE/ESA uses native ECKD channel programs, independent of the attachment.

  See also the discussions to this subject in the RAMAC Array Subsystem section
- Products and ECKD

  If any vendor product does not yet use optimal ECKD channel programs, contact the vendors. They are aware of this need.

  Naturally, this also applies to IBM products
- Do NOT ADD devices as ',SHR', except if required for sharing data across VSEs

ECKD Usage in VSE/ESA ...

Conversion Routine

- **Dynamic CKD/ECKD Conversion Routine is provided by the VSE/ESA Supervisor**

  For device type 6E, the VSE/ESA CKD/ECKD conversion routine attempts to convert CKD channel programs into ECKD

  Naturally, any sector values are taken over from SET SECTOR to LOCATE RECORD.

But ...

- **The VSE CKD/ECKD conversion routine cannot convert all types of channel programs**

  E.g. self-modifying CKD channel programs (RRRRRRG!).
  This is to be assumed if CCB Byte 3 Bit 7 is ON
  (Applies to EXCP with or w/o DTFPH)

  Such channel programs remain CKD, an thus produce performance problems

- **Not all CKD channel programs can be fully translated to ECKD**

- **The VSE CKD/ECKD conversion routine never can set individual caching bits**

  Refer also to
  'Smarter CKD/ECKD Conversion Routine'
  in VSE/ESA V2 Performance Considerations document

More Hints for Optimal ECKD Use

More Hints for Optimal ECKD Channel Programs

No ECKD channel program can/should be setup for a specific I/O subsystem. If, for any reason, a very performance beneficial CCW is only available on a specific (newer) I/O subsystem, this can be treated separately on top, after having dynamically checked the availability before usage.

Any chance for an improvement should be taken, i.e. all settings should be done always, if of potential benefit somewhere

Define Extent (DX) Hints:

- **Specify the caching bits correctly/optimally**

  This will, depending on the type of I/O subsystem, enhance I/O performance, if cached.

  Set RECord caching only if really cache unfriendly.
  Set SEQuential caching if sequential.

  (DX parameter list: Byte 1, Global Attributes, bits 3-5)

  Refer to 'DASD Cache Strategies' chart in the next section

  Using CFW became less important meanwhile:
  - ignored for all RAID-5/6 I/O subsystems
  - NVS sizes (if separate from cache) increased

  (DX parameter list: Byte 1, Global Attributes, bit 6)

- **Specify Regular Data Format (RDF), if record format is fixed**

  This is beneficial for some I/O subsystems
  (DX parameter list: Byte 7, Global Attributes Extended, bits 0-1, must be 01)

More Hints for Optimal ECKD Use ...

Optimal ECKD Channel Programs (cont'd)

- **Specify Standard Record 0, if R0 is not misused**

  This is beneficial especially in RAID-5/6 I/O subsystems
  (DX parameter list: Byte 7, Global Attributes Extended, bit 5, should be 1, i.e. 0s in data field)

- **Careful select the DX Mask Byte setting**

  The Write Control bits should be as restrictive as possible
  (DX parameter list: Byte 0, Mask Byte, Bits 0-1 'WRITE Control').

  The combination '11' must be avoided, since it may force 'Bypass Cache' for some types of I/O subsystems.

  ```
  Use  01  for pure READ channel programs,
       10  to permit update writes only
       00  for all WRITEs except WRITE HA and WRITE R0
  ```

- **Use a single DX CCW in an ECKD channel program**

  This avoids, at least in some non-cached subsystems, a lost revolution

- **Specify always the same DX extent limits for sequential I/Os**

  The extent limits should be constant (= as big enough) as possible, since optimal (clustered) de-staging and sequential pre-staging may end at the upper extent limit specified in the current CCW program.

More Hints for Optimal ECKD Use ...

Optimal ECKD Channel Programs (cont'd)

General Hints:

- **Whenever possible, maximise the amount of data in a single channel program**

  - KB/IO value reasonably large, if possible
  - Use multi-track CCWs or READ/WRITE >1 track in a single channel program

- **Whenever doing FORMAT-WRITEs, try to format full (logical) tracks**

  - avoids padding rest-of-track in specific cases

Locate Record (LR) Hints:

- **Specify correct sector values in LR CCW**

  No sector value ('FF') or a wrong sector value may result in higher utilizations for channel and/or CU in case of non-cached I/O subsystems.

  This may also apply to some elder cached I/O subsystems.

  This problem may also occur, if in CKD channel programs no RPS is used, or wrong sector values are given.

  For cached I/O subsystems the sector value should have no impact
  - for READ or WRITE hits
  - for I/O subsystems doing full track staging only (RVA)

More Hints for Optimal ECKD Use ...

Optimal ECKD Channel Programs (cont'd)

Ù **Make sure the LR count field is correct**

It must contain the EXACT count of the #records to be transferred in the 'LR domain' (#records transferred in the following CCWs to the next LR or to the end of the channel program).

It makes NO DIFFERENCE how many CCWs are used to traverse a record on DASD.

Basics:

A DASD record consists of a

- Count field,
- Key field (optional), and
- Data field.

```
       |--------------- DASD record ---------------|
        cnt    key        data
  R1   |____| |_________| |___________________________|
         |         |                 |
         |         |                 |_____ data FIELD of a record
         |         |
         |         |______________________ key FIELD of a record
         |
         |________________________________ count FIELD of a record


       |--------------- DASD record ---------------|
        cnt    data
  R1   |____| |_______________________________________|
         |                           |
         |                           |_____ data FIELD of a record
         |
         |________________________________ count FIELD of a record
```


More Hints for Optimal ECKD Use ...

LR count field (cont'd)

Some Examples:

If you intend to read ... ,

your channel program will consist of
- a LR CCW

followed by ...

- 5 consecutive count fields, and no intervening key or data fields:

  5 Read Count CCWs, with no Read Key, Read Data, or Read Key and Data CCWs. The LR count value must be 5.

- 3 consecutive Data fields, and none of the intervening Count or Key fields:

  3 Read Data CCWs, with no intervening Raed Count or Read Key CCWs. The LR count value must be 3.

- 1 count field using a Read Count CCW, then read the Key and the Data fields of that record with a separate Read Key and Data CCW:

  1 Read Count CCW followed by one Read Key and Data CCW. Because only one DASD record will be processed, the LR count value must be 1. That's right ONE. Although more than one data transfer CCW, is used (Read Count, then READ Key and Data), only one DASD record is processed.

- The count field of one record, then the data field of that and 17 subsequent records:

  1 Read Count CCW followed by 18 Read Data CCWs. In this example, the LR count value is 18, as 18 records on DASD will be traversed, with the count and data fields for the first record being read and the data fields being read for the remaining records.

The 'Introduction to Nonsync' and the '3990 Product Reference' publications contain detailed information on ECKD channel programming with which anyone creating DASD channel programs needs to be familiar.

ADDing VSE DASDs

ADDing VSE DASDs

„ **Always use 'ADD cuu, ECKD' in VSE/ESA, if attachment knows ECKD,**

except DY41099 is not or cannot be applied (a general rule, since nearly always required for performance reasons)

ADD cuu,ECKD	for 3380 | at any 3990 (-2,-3,-6) +for 3390 | at any RAMAC attachment | at Internal Disk
" " "	for 9345 even when uncached

- ECKD support not available before VSE/ESA 1.1 (the newer the release the more ECKD is exploited)
- Make sure DY41099 is included in VSE/ESA to avoid 'imprecise ending' for ECKD (avoid risk of data integrity) Standard since VSE/ESA 1.2.1, applicable since 1.2.0
- ADD cuu, 3380 (without ,EML) may be sufficient for 3380s at any 3990 to get also ECKD channel programs (if VSE forces ECKD)
- Refer also to the RAMAC Array Subsystem ADDs page

„ **ADDing devices as CKD (3380) for nonsync attachments (all newer I/O subsystems) ... may result in unrecoverable I/O errors**

„ **RAMAC Array Subsystem was/is the only newer I/O subsystem which allowed to set synchronous mode ('CKD'), but at cost of performance.**

Refer to the RAMAC Array Subsystems ADDs page

ADDing VSE DASDs ...

ADDing VSE DASDs (cont'd)

„ **VSE never overrules ',EML' for ADDing a device**

'EML' is only required if, for whatsoever reason, customer wants to avoid that with 'TYPE=SENSE' the device type is reset to the sensed type

í **Never use '3380,EML' since this forces CKD channel programs (device type '6C'), even if the attachment knows ECKD**

This is especially true for 3380s in RAMAC environments

„ **VSE DASD Recognition**

The following table shows how VSE will see a 3380 DASD for non-sync attachments depending on how the device is defined (ADDed) to VSE.

Note that some of the combinations should not be used; this is just to show the VSE action for these combinations:

ADD statement	Attach-ment	VSE sees device as	PUB	DTF	PROBLEM/ Notes
ADD cuu,ECKD	nonsync	ECKD	6E	xx	
ADD cuu,3380,EML	nonsync	CKD	6C	0C	split CE/DE
ADD cuu,3380	nonsync	ECKD	6E	xx	*1

*1 Nonsync VPD mode causes device to be defined as ECKD (Msg 1I71I)

í **If device type is 6E, ADD was performance-wise optimal**

ADDing VSE DASDs ...

Potential Vendor Program Deficiency

„ **Some vendor products are/were sensitive to the device type code**

```
Some programs look for a '6C' in the PUB and may not recognize the
'6E' that VSE will place there for ECKD devices.

Some programs look e.g. for a '0C' in the DTF and may not recognize
the '0E' for 3380 ECKD.
Similarly, this also applies for other ECKD DTF device type codes
like '04' for 9345, '26' for 3390-1, '27' for 3390-2, '24' for
3390-3, '32' for 3390-9.
```

„ **Some vendor products use the 6C indication to identify 3380 device type for track capacity info**

```
VSE has a GETVCE service that can be used for this, eliminating
the need to interpret the PUB content.
```

> Device type 6C is disadvantageous for any nonsync ECKD DASD, be it real devices or RAMAC emulated devices, cached or uncached

- Some vendor products may not be able to handle separate presentation of Channel End w/o Device End
- VSE will put 6C in the PUB for a device for a 3380 ADD statement with EML:

ADD cuu,3380,EML

should be only used as a temporary circumvention

- ADD cuu 3380,EML will result in degraded performance, by using
  - device type '6C' (0C in DTF)
  - and CKD channel programs, not ECKD

> Customer should contact the vendor to request a fix to recognize 6E (for device type identification and/or track capacity calculation), both for all nonsync real and RAMAC emulated devices

DASD Caching in General

PART C.

DASD Caching in General

Overview

Ù **DASD Cache**

- **Functions**
- **Strategies**
- **Benefits**
- **Statistics**

Principal DASD Cache Benefits

Principal DASD Cache Benefits

Ù **Significantly faster msec/IO**

Fast 'DASD response' in case of a cache hit

> The higher the hit ratio, the better is performance:

- READ hits via Basic Functions or Record Caching
- WRITE hits only with Extended Functions and non-volatile storage

Ù **Faster processing of I/O intensive workloads (faster user response times or batch elapsed times)**

Ù **Dependencies**

Workload characteristics
- I/O intensity
- R/W ratio
- Access patterns
- ...

Cache functions supported
- Basic Caching
- Extended Functions
- ...

Configuration dependencies
- CPU
- I/O configuration
- Cache size
- ...

DASD Cache Terms

Some terms:

'Staging'

Loading of a cache with data from DASD (DASD/Cache transfer)

'De-staging'

Writing of cache data to DASD (Cache/DASD transfer)

'Rest-of-track staging'

(General amount of data to be staged for Track Caching)

The unit of transfer from DASD to cache generally is the requested record plus 'rest-of-track'.

Only if subsequently a lower indexed record of the same track is needed and not in the track slot ('front-end-miss'), the track is completed by reading all records not yet in cache. This is better than a pure track oriented strategy.

Rest-of-track staging does NOT elongate I/O response times for misses, since the staging of additional data is done after the I/O request is complete. It only occupies still the physical device and the device path, which may cause contention for other accesses.

On newer control units, also 'Record Caching' may be available

'Least Recently Used, LRU'

(General selection rule for the data to be de-staged)

In general those data are discarded from cache, to which the latest access was the 'least recently used' of all 'tracks', i.e. which for the longest time has not been referenced.

An exception from this rule is SEQuential data, see later.

Principal DASD Cache Functions

Principal DASD Cache Functions

```
                |     | Data in cache   |  Data not in cache
                |     |     (hit)       |      (miss)
----------------+-----+-----------------+-----------------------
Read  Caching   |Cache| |--|            |  |------------|-----|
                |     |    DEf          |                   EOx
                |NVS  |                 |
                |     |                 |
                |DASD |                 |  |============|.....|
(Reads only)    |     |                 |              DEn
----------------+-----+-----------------+-----------------------
Basic Write     |Cache| |--|            |  |-|
        Caching |     |                 |
                |NVS  |                 |     (no staging w/o DFW)
                |     |                 |
                |DASD | |==|=======|    |  |============|
(Writes only)   |     |            DEn  |              DEn
----------------+-----+-----------------+-----------------------
DASD Fast Write |Cache| |--|            |  |------------|-----|EOx
  (DFW)         |     |    DEf          |
                |NVS  | |--|            |
                |     |         later   |
(Writes, Reads  |DASD |      |::::::::| |  |============|.....|
 same as above) |     |                 |              DEn
----------------+-----+-----------------+-----------------------
Cache Fast Write|Cache| |--|            |  |------------|-----|EOx
                |     |    DEf          |
                |NVS  |                 |
                |     | later(if at all)|
(Writes, Reads  |DASD |      |::::::::| |  |============|.....|
 same as above) |     |                 |              DEn
---------------------------------------------------------------

EOx = end of last staged record
    = EOT = End of Track (in case of Track Caching)
    = EOR = End of last Record of VSE channel pgm (Record Caching)

DEn = 'normal', DEf = 'fast' device end to S/W

==== immediate physical access to DASD
:::: later (asynchronous) physical access to DASD
.... DASD (and cache) activity for rest-of-track staging

For DASD Fast Write and Cache Fast Write a FORMAT WRITE (Write CKD)
is always a hit (in contrast to an 'UPDATE WRITE' w/o RDF)
```


Principal DASD Cache Functions ...

'Basic Cache Functions'

```
Read (Only)      Caching of READ operations.
       Caching   In case of a read miss the track is read from DASD
                 to cache and (except 3880-13) concurrently via the
                 channel.

Basic Write      Caching of WRITE operations.
       Caching
                 a)If data in cache,
                   cache and DASD update are started together,
                   DE when DASD finished ('Forked Write').
                 - performance benefit if later on a READ finds the
                   data in cache.

                 b)If data not in cache,
                   record is written to DASD but NOT to cache.
                 - never causes cache load (staging).
```

'Extended Function Fast Write'

```
DASD Fast Write  For all write intensive files, full data integrity.
   (DFW)         Can be set off by a bit in DEFINE EXTENT.

                 a)If data in cache,
                   cache updated and data saved in NVS with imme-
                   diate DE.
                   Physical write to DASD done only when required
                   (may be a long time later) when a 'de-staging'
                   occurs or when a 'commit' is requested in the
                   PERFORM SUBSYSTEM FCT-CCW.

                 b)If data not in cache
                   b1) UPDATE WRITE (or READ) is done on/from DASD
                       with late DE and cache is being loaded.
                   b2) FORMAT is done in cache and NVS with imm.DE.

Cache Fast Write Only for temporary data e.g. work files, which
    (CFW)        need not reside on DASD.
                 Data are lost in case of power failure.
                 Can be set on by a bit in DEFINE EXTENT.

                 Same as DASD Fast Write, but w/o NVS.
                 Written to DASD only if 'de-staging' required or
                 if it is disabled in the SET SUBSYSTEM MODE-CCW.

 Formally DFW and CFW can also be used for read channel programs,
 but then it is identical to Read Caching.
```


Staging and Destaging

Staging and Destaging

- **Data are loaded from DASD into cache (staging) at a ...**
  - **READ cache miss**
  - **DASD Fast Write miss**
  - **Cache Fast Write miss**
- **(Updated) Cache/NVS data are written to DASD (de-staging) ...**
  - **at a DASD Fast Write hit (later)**
  - **at a Cache Fast Write hit (later, if at all)**
  - **at 3990-3/6 restart, or later**
    NVS data, not yet on DASD, after power loss
  - **when caching is turned off**
  - **at any time**
    any candidate data, when cache/NVS slots are needed

DASD Cache Strategies

DASD Cache Strategies

Normal (LRU) Caching
(General Caching Strategy)

```
   Stages only rest-of-track into cache,
   replaces (de-stages) LRU track of whole cache.
   This is the standard way of caching, unless otherwise set by
   S/W in each individual CCW chain.

   Most appropriate for general access.
```

Sequential Access Caching
(Special Caching Strategy)

```
   Stages rest-of-track into cache plus subsequent full tracks.
   The 3990-3 pre-stages up to 5 tracks into the cache,
   the 3990-6 up to 1 cylinder (15 tracks).
   (Tracks accessed with Sequential Access Caching are sooner
    candidates for being de-staged).

   Appropriate for sequential file processing.
```

Inhibit Cache Loading

```
   Does not allow to load any new tracks into the cache
   (avoids unnecessary load of data into cache).
```

Bypass Cache

```
   Does not allow to use the cache at all.

NOTES:
   All S/W settings of these modes are done with special bits in
   the DEFINE EXTENT CCW and are valid only until end of chain.

   For CKD channel programs, only Normal Caching and DASD Fast Write
   can be used (no DEFINE EXTENT CCW available).

   For CKD to ECKD converted channel programs the same applies as for
   CKD (no caching bits are set in DEFINE EXTENT CCW).
```

DASD Cache Strategies ...

DASD Cache Strategies (cont'd)

„ **Track Caching vs Record Caching**

2 principal caching methods used:

Method	Staging	
Track Caching	Rest-of-track	Traditional caching
Record Caching (Access)	Record only	Complementary method, for cache unfriendly data (3990-6, RAMAC Array Subs. RSA-2, 9390s)
- Adaptive Caching dynamically uses both methods		

„ **Global Atttributes in DEFINE EXTENT CCW**

Settings of bits 3-5 of Byte 1

```
000 (C0)  Normal Cache Replacement
001 (C4)  Bypass Cache               BYP
010 (C8)  Inhibit Cache Load         ICL
011 (CC)  Sequential Access          SEQ
101 (D4)  Record Access              REC
```

- Values in (Gg) are usually seen in SDAID I/O traces, if bits 6 and 7 (described below) are 0:
  DATA= MmGg...
  (Mm= mask byte, Gg= Global Attributes byte)
- REC cannot/needs not be combined with other settings (no combination possible/required)

Settings of bit 6 of Byte 1

0 Do not use Cache Fast Write
1 Use Cache Fast Write

Settings of bit 7 of Byte 1

0 Allow DASD Fast WRITE
1 Inhibit DASD Fast WRITE

Cache Options Hierarchy

Cache Options Hierarchy

1. **Highest level:**

   H/W defaults

2. **Medium level:**

   H/W defaults, altered by SET SUBSYSTEM MODE

   - by VSE/ESA (native or guest)
   - by VM/ESA

3. **Lowest level:**

   Global Attribute Setting in DEFINE EXTENT (DX)

   - DX not in CCW chain
   - DX added by VM
   - DX used by VSE

The scope of the lowest level is always a single CCW chain, whereas the other levels are on a 'permanent' basis.

DASD Cache Performance Benefits

DASD Cache Performance Benefits

Ù **Reduction of virtual space requirements for CICS**

Virtual storage for a transaction is released earlier, if response time faster

Ù **Improved elapsed times for I/O intensive loads**

Any type of DASD I/Os

Up to about 70% / 105% throughput increase (basic caching only / with DFW)

in the example below

Sample Calculation

Ù **Assumptions**

- Any type of batch job or transaction, here single thread considered
- 12.5 sec total CPU-time on a 9221-150
- 4000 I/O operations to disk, un-overlapped, READ/WRITE ratio 4/1, i.e. 3200 READs, 800 WRITEs
- 20 msec average per DASD I/O, 3 msec at cache hit
- 70% basic caching hit, 70% DFW hit ratio (very conservative)

Ù **Calculation Results**

	CPU-time CPUT	I/O-time IO	Elapsed time ET	ET reduction	Rel. thruput
No caching	12.5 sec	80.0 sec	92.5 sec	Base	1.0
Basic caching (ESA 1.1/1.2)	12.5 sec	41.9 sec	54.4 sec	41%	1.7
Basic caching +DFW (ESA 1.3)	12.5 sec	32.4 sec	44.9 sec	53%	2.05

í **70%/105% more single batch or single thread throughput**

DASD Cache Performance Benefits ...

Ù **Calculation Details**

- Elapsed Time ET = CPUT + (unoverlapped) IO
- CPU-time (same for all cases)

  CPUT = 12.5 sec
- IO-time without 3990-3/6 cache:

  IO = 4000 x 20 msec = 80 sec
- IO-time with basic caching only (VSE/ESA 1.2)

```
IO = 0.3 x 3200 x 20 msec  (19.2 sec)
   + 0.7 x 3200 x  3 msec  ( 6.7 sec)
   +        800 x 20 msec  (16.0 sec)  = 41.9 sec
```

- IO-time with full caching (VSE/ESA 1.3)

```
IO = 0.3 x 3200 x 20 msec  (19.2 sec)
   + 0.7 x 3200 x  3 msec  ( 6.7 sec)
   + 0.3 x  800 x 20 msec  ( 4.8 sec)
   + 0.7 x  800 x  3 msec  ( 1.7 sec)  = 32.4 sec
```

Ù **Throughput Increase Sensitivity Factors**

Benefits of I/O caching are ...

```
higher if e.g.  - more I/O or more write intensive
                - removing a device bottleneck
                - processor faster

lower  if e.g.  - workload less I/O dependent
                - multithread (several batch partit.
                  or concurrent transactions)
                - channel is/becomes a bottleneck
```

DASD Cache Performance Benefits ...

Effective DASD Response Times

Example for 100 DASD I/Os, R/W Ratio of 4:1, 70% hit ratios

No caching: 20 msec

```
 _____________
|    CPU      |                                        _____
|             |                                      / 3390 \
|VSE/ESA 1.x  |        80 READs                      \ ____ /
|             |------------------------------------->|      |
|     &       |                                      |      |
|Applic. I/O  |------------------------------------->|      |
|             |        20 WRITEs                     \______/
|_____________|                                         |
   A                                                    |
   |                 100 responses from DASD            V
    ------------------------------------------------------
```

Basic Caching: 10.5 msec (.44x20 + .56x3)

```
 _____________
|    CPU      |                 3990-3/6
|             |                 _________             _______
|VSE/ESA 1.x  |                |         |           / 3390 \
|             |   80 READs     |70% hit  | 24        \______/
|     &       |--------------->|         |------>|   |      |
|             |                |no DFW   |       |   |      |
|Applic. I/O  |--------------->|-------->|------>|   |      |
|             |   20 WRITEs    |_________| 20        \______/
|_____________|                     |                   |
   A    A         56 cache hits     |                   |
   |    |---------------------------V                   |
   |              44 responses from DASD                V
    ------------------------------------------------------
```

Basic Caching + DFW: 8.1 msec (.3x20 + .7x3)

```
 _____________
|    CPU      |                 3990-3/6
|             |                 _________             _______
|VSE/ESA 1.3  |                |         |           / 3390 \
|             |   80 READs     |70% hit  | 24        \______/
|     &       |--------------->|         |------>|   |      |
|             |                |         |       |   |      |
|Applic. I/O  |--------------->|70% hit  |------>|   |      |
|             |   20 WRITEs    |_________|  6        \______/
|_____________|                     |                   |
   A    A         70 cache hits     V                   |
   |    |--------------------------                     |
   |              30 responses from DASD                V
    ------------------------------------------------------
```


VSE/ESA DASD Cache Statistics

VSE/ESA Cache Statistics for 3990-3/6

Likewise applies to RAMAC I/O subsystems, 9390s and others

„ CACHE REPORT command

```
CACHE UNIT=cuu,REPORT  Provides statistic counters for device cuu
```

```
                 |.....READ.......|.........WRITE...................
REQUESTS:        |TOTAL  CACHE-RD | TOTAL  CACHE-WRT  DASD FAST-WRT
                 |        (hits)  |         (hits)      (all)
--------------------------------------------------------------------
NORMAL           |  A1       B1   |   C1      D1          E1
SEQUENTIAL       |  A2       B2   |   C2      D2          E2
CACHE FAST WRITE|   A3       B3   |   C3      D3         N/A
TOTALS           |  A        B    |   C       D           E

REQUESTS: (read and write)
 INH. CACHE LOAD|        F1
 BYPASS CACHE   |        F2

DATA TRANSFERS:     DASD->CACHE       CACHE->DASD
                 |     (Stage)          (De-stage)
 NORMAL          |       G1                H1
 SEQUENTIAL      |       G2               N/A
```

- All counters above are reset at IML of the I/O subsystem, or (not recommended) when Subsystem Caching is set off
- READ channel programs are all channel programs w/o any WRITE (at least 1 SEARCH or READ)
- WRITE channel programs contain at least 1 WRITE
- Channel programs for SENSE and DIAGNOSE purposes or with RECORD CACHING are not included in the above counters
- CACHE-RD are all Read-I/Os which did not require any data movement from DASD (Read hits).
- WRITE TOTAL counters include DASD FAST WRITE and non-DASD FAST WRITE requests (e.g. CKD channel programs, without DEFINE EXTENT)
- CACHE-WRT are all I/Os (with at least 1 WRITE command) which performance-wise profited from the cache (Write hits, D1 and D2 via DASD FAST WRITE, D3 via CACHE FAST WRITE).
- DASD FAST-WRT counters include all requests (hits and misses) to devices, where DFW is NOT disabled.

VSE/ESA DASD Cache Statistics ...

VSE/ESA Cache Statistics (cont'd)

For NORMAL WRITE requests (i.e. no SEQUENTIAL, no CFW) it holds:
(All CKD channel programs belong to this 'NORMAL' category)

```
                         / hits      D1    |
      DASD FAST WRITE --                   |  E1    |
                         \ misses   E1-D1  |        |
                                                    | C1 --TOTAL
                         / hits      0 *)  |        |      NORMAL WRT
  non-DASD FAST WRITE ---                  | C1-E1  |
                         \ misses   C1-E1  |
```

*) WRITE hits only exist for DASD FAST WRITE.

If all WRITEs are DASD FAST WRITEs, C1 and E1 counters are identical.

Counters cannot be reset in VSE. Thus, data for a specific interval have to be calculated from the deltas of 2 displays.

Calculable Hit Ratios:

Read	DASD Fast Write	Cache Fast Write
B/A	(D1+D2)/E	D3/C3

- CACHE FAST WRITE and DASD FAST WRITE are exclusive, therefore N/A ('not applicable') is shown in the CFW line.
- INH. CACHE LOAD and BYPASS CACHE include Reads and Writes. They are not contained in the 'A' or 'C' counters.
- The G counters are the number of transfers to stage the cache from DASD. Tracks, being read ahead via sequential access caching are counted separately.
- Counter H1 designates all CACHE to DASD de-staging transfers (write from cache to physical device).
- For VM/VSE, these device related counters are only valid for DEDicated devices or Full Pack Minidisks, not for Partial Pack Minidisks
- To calculate subsystem hit ratios, only devices with caching ENABLED should be included, since only those devices are candidates (this holds for basic caching and DFW caching)

VSE/ESA DASD Cache Statistics ...

VSE/ESA Cache Statistics (cont'd)

„ 3990-specific side aspect

In very specific situations the following may be good to know for any real or simulated 3990 cached control unit:

The 3990 statistic counters count I/Os on the basis of 'Channel Operations'.

A channel operation is in nearly all cases identical with a SSCH, except the case where a channel program contains intermediate Define Extent or SEEK CCWs. Such a channel program is counted e.g. as a READ and a WRITE operation, if it would consist of 2 'CCW subchains'.

„ CACHE SUBSYS=cuu,REPORT Enhancement

APAR DY43697 for VSE/ESA 2.1 provides an enhancement for this command

Native VSE/ESA:

All data for the devices attached to this subsystem (and ADDed to VSE, also those ADDed with ,SHR) are accumulated in order to directly give summary data for all devices of a DASD subsystem.

VSE under VM:

The CACHE SUBSYS=cuu,REPORT command is treated as

```
CACHE UNIT=cuu,REPORT
```

since physical devices may be shared via minidisks.

VSE/ESA Cache Status Displays

VSE/ESA Cache Status Displays

„ CACHE STATUS command

Requesting status info:

```
CACHE UNIT=cuu,STATUS     For caching status info for device cuu

CACHE SUBSYS=cuu,STATUS                      ... for the subsystem
```

System responses:

```
   DEVICE CACHING STATUS: ACTIVE
         DASD FAST WRITE: ACTIVE
```

```
SUBSYSTEM CACHING STATUS: ACTIVE
        CACHE FAST WRITE: ACTIVE
           CACHE STORAGE: CONFIG.    ...... 234881024
           CACHE STORAGE: AVAIL.     ...... 234881024
              NVS STATUS: AVAILABLE
```


2-Stage vs 1-Stage Cached Subsystems

2-Stage vs 1-Stage Cached Subsystems

- 2-stage means that all channel transfers are detached from device (i.e. the cache must be used for ANY transfer)

  Here, the subsystem cache is addressed. RAMAC 1/2/3 has also a drawer cache, used e.g. for parity handling

	1-stage	2-stage
I/O Subsystem	3990-3, 3990-6 w/ native DASD. RAMAC 1/2/3 *)	RVA, RSA
Subsystem cache may be req'd for	-	RAID-5, track conversion
Channel Xfers for misses	Device speed *)	Channel speed
Start of Xfer at misses	When 1st byte arrived from DASD	After all req. data are in subs. cache
Inh.Cache Load, Bypass Cache	Fully honoured	May just result in 'early cache discard'
*) RAMAC formally is 1-stage but where 'device speed' is drawer cache speed and thus = channel speed		

Refer to scheme on next foil

2-Stage vs 1-Stage Cached Subsystems ...

Cache Data Flow

```
                Hits                       Misses
              1/2-stage         1-stage             2-stage
               - - - - -     - - - - - - - - - - - - - - - - - -

 --------    READ   WRITE     READ    WRITE     READ    WRITE
|        |    A      |         A       |            A       |
| Real   |    |      |         |       |            |       |
| Storage|    |      |         |       |            |       |
|        |    |      |         |       |            |       |
 --------     |      |         |       |            |       |
    |         |      V         |       V            |       V
   ---  ......|......|.........|.......|............|.......|.....
  |Ch |       |      |         |       |            |       |
  |ann|       CS     CS        DS      DS           CS      CS
  |el |      =ChnlSpeed       =DeviceSpeed         =ChnlSpeed
  |   |       |      |         |       |            |       |
   ---  ......|......|.........|.......|............|.......|.....
    |'Upper   |      |         |       |            |       |
    |  I/F'   |      |         |       |            |       |
 -------- ....|......|.........|.......|............|.......|.....
| Cache  |    |      |         |       |...         |       |
| and/or |    |      V...      A       |  A     ____|   ____|...
| NVS    |    -         |      |       |  |     A        A     |
 -------- ..............|......|.......|..|.....|........|.....|..
    |'Lower             |      |       |  |     |        |     |
    |  I/F'             R     R+S      R  S    R+S      R+S    R
  -----   ..............|......|.......|..|.....|........|.....|..
 /      \               |      |       |  |     |        |     |
| 'DASD' |              |      |       |  |     |        |     |
|(+drawer|              V      |       V  |     |        |     V
| if so) |          Async  Sync    Write Async  Sync   Sync  Async
 \      /          Destage Stage   Thru  Stage  Stage  Stage Destage
   -----  ..........................................................
```

R or S means transfer of the Requested record and/or
the Staging unit (whatever it is in a specific case)

Cache Behavior Factors

Caching Behavior Factors for Workloads

„ Reasons for DASD cache hits:

Probability of 're-reference'

Re-referencing a record already referenced recently, a READ after a WRITE, or any other combination

Occurs more often for
- VTOC accesses (MVS)
- index components of data bases
- VSAM index (if index not in storage), less often for VSAM data

Locality of track reference

Referencing a record in the same track. Track itself is random at first reference

It is assumed that overall this track locality plays a bigger role than re-referencing of records, even in a multi-thread environment

Track sequentiality

Referencing subsequent tracks
Cache provides
- track blocking (1 track per I/O)
- also higher rate (if sequential bit set on top)

„ Efficiency of cache usage

Overall hit ratio depends also on ...
- size of the cache
- type of caching (track vs record)

„ Traditionally 'cache unfriendly' data

- Very small locality of track reference
- No track sequentiality
- Only minor increase of hit ratio with higher cache size

Certainly, with cache sizes becoming larger and larger, 'cache unfriendliness' may reduce.

DASD Cache Bit Settings

DASD Cache Bit Settings for VSE/VSAM

Status as of VSE/ESA 1.3.x with the VSAM PTF for 3990-6 Enhancements (APAR/PTF DY43072/UD90363, dated 03/94).

TYPE OF ACCESS	CACHE HANDLING	REMARK
Write I/Os to WRITECHECK files	BYP	
Read I/Os to WRITECHECK files	Normal	
Replicated index set I/Os	REC	**
Noreplicated index set I/Os	Normal	
Complex channel programs (>1 LR domain, except for WRITECHECK) This covers mainly CA-splits and sequential read-ahead from highly scattered CAs	Normal	
Format-writes in SPEED (LOAD) mode (includes REPRO if SPEED)	SEQ	was BYP
Read I/Os on behalf of GET (SEQ,NUP,FWD) (e.g. REPRO for INFILE w/o ENV parameter) (includes DL/1 IMAGECOPY)	SEQ	ACB access
Write I/Os on behalf of PUT (SEQ,NUP,FWD) (if WRITECHECK not in effect) (e.g. REPRO for OUTFILE w/o ENV parameter) (includes DL/1 RECOVERY) (includes SQL/DS 3.5 Online ARCHIVE)	SEQ	ACB access
I/O for (DIR,NUP) or (DIR,UPD) for ESDSs opened with MACRF=(CNV,UBF), except BLDINDEX work files, includes - SQL/DS * - DL/1 data component, UNLOAD/RELOAD	REC***	**
ALL OTHERS (includes REPRO if not SPEED)	Normal	

* For SQL/DS ARCHIVE use SQL/DS 3.5 with VSAM controlled buffers

** REC means record cache, applicable to 3990-6 Enhanced, and Internal Disk. Mostly superseded by adaptive caching and larger cache sizes.

*** Default since VSE/ESA 2.4 or DY44796 is Normal. Can be controlled via a SYSCOM bit.

- All I/Os specify 'Regular Data Format' RDF (except I/Os to a mixed data/sequence set of an IMBED KSDS) -> 3990-6 KSDSs should not be defined with IMBED

DASD Cache Bit Settings ...

DASD Cache Bit Settings for VSE Utilities

UTILITY	FUNCTION	CACHE HANDLING	REMARK
VSAM B/R	BACKUP to tape/disk: Source disk READ Data Index Target disk WRITE Data Index RESTORE from tape/disk: Source disk READ Data Index Target disk WRITE Data Index	 SEQ Normal SEQ * Normal SEQ Normal SEQ * Normal	 was BYP was BYP
VSAM REPRO	Source file READ Target file WRITE: SPEED RECOVERY **	SEQ SEQ Normal	
LIBRarian	BACKUP/COPY/RESTORE/LIST/CATALOG (for data, not index blocks) + LIBRM GET/PUT Source disk READ Target disk WRITE	 SEQ SEQ	 was ICL was BYP
FAST COPY	DUMP Volume/File (OPT=1) (OPT>1) RESTORE COPY Volume/File	SEQ SEQ SEQ SEQ/ICL	 was ICL was BYP
DSF	INIT	BYP	

- Any cache settings needs ECKD channel programs
- ICL means Inhibit Cache Load
- Settings apply to all cached I/O subsystems

* VSAM APAR/PTF DY43138/UD49025 uses SEQ (03/94)

** If cluster defined with RECOVERY or cluster not empty

- SEQ Setting for LIBR and FCOPY OPT>1 in VSE/ESA 2.3

Some ECKD Caching Bits and VSE Releases

Some ECKD Caching Bits and VSE Releases

	SEQ (Sequential)	RDF (RegDataForm)	REC *6 (RecordCache)
Used by VSE/... SP 4.1 *1	- (!)	-	-
ESA 1.1 *1	x	-	-
ESA 1.2 *1	x	-	-
ESA 1.3/1.4	x	VSAM only	VSAM
ESA V2	x	x	VSAM
Beneficial for ... Cached 9340 DASD	yes	no *5	no *5
3990-3 +any DASD	yes	no *5	no *5
3990-6 +any DASD	yes/no *2	yes	yes/no *3
RAMAC Subsystem	yes	yes	no *3
9390/RAMAC 3	no *2	yes	no *3
RAMAC Virt.Array	yes/no *2	no *5	no *5
RAMAC Elec.Array	no *4	no *4	no *4
RAMAC Scal.Array	yes	no *5	no *5
RAMAC Sc.Array-2/-3	yes	yes	no *3
Multipr.Int. Disk	yes	no *4	yes

*1 No DFW available w/o VM. Required for RAID-5
*2 3990-6 since 06/96, RAMAC 3 and RAMAC Virtual Array can detect sequential access (Sequential Detect)
*3 Dynamic record caching, controlled via PTT or similar fct
*4 Function not required/beneficial
*5 Function not available, bits ignored
*6 Record caching (used by VSAM for DL/1 and SQL databases) less beneficial for larger/newer cache sizes

„ **ECKD channel programs**
- required/highly beneficial for optimal cache control

- often required to avoid usual CKD WRITE performance degradation for nonsync attachments

I/O Response Time Component Analysis

PART D.

I/O Response Time Component Analysis

The following are references for this subject

DASD Performance Analysis Using Modeling, by Thomas Beretvas, IBM Corporation, now Beretvas Performance Consulting
Computer Measurement Group (CMG) Proceedings 86, 12/86, pp 749-760

A classic paper, still of interest for many reasons

MVS/ESA RMF Version 4 -Getting Started on Performance Management- LY33-9174-00, 12/93
OS/390 RMF Performance Management Guide, SC28-1951-00, 09/96

RMF oriented tuning books, refer to Chapter 5 'Analyzing I/O Activity'

Understanding Cached DASD I/O Performance, by Thomas Beretvas, IBM Paper 10/91

Balanced Systems and Capacity Planning, by R.T. Borchetta and Ray J. Wicks, IBM WSC Technical Bulletin, GG22-9299-04, 08/93, 125 pages

DASD Performance and Capacity Planning Class, by Thomas Beretvas, Beretvas Performance Consulting, Kingston, NY. Tel 914-339-5897

A very good, competent and extensive course on I/O subsystems for MVS. Includes also non-IBM devices.

Overview and Summary

Some Basic Relationships

```
IORT = IOSQ + DST                    (1)
DST  = PEND + DISC + CONN            (3)
DISC = Seek + Latency + RPS_delay    (4)
CONN = PROT + XFER                   (6)
```

- For simplicity, assume uncached (physical) devices for the moment

Rough Values for Device Service Time Components

Rough values for Achievable DST components (msec)

PEND	DISC *1	PROT	XFER *2 uncached or cache miss f.non-RAMAC	XFER *2 cache hit or cache miss for RAMAC and 2-stage
1.0 (unshared)	12-15 (uncached)	1.0 (parall)	2.0 (parallel)	2.0
2.0 (shared devices)	HRx(12-15) (cached devices)	1.5 (ESCON chnls)	2.0 (ESCON channels)	0.5

*1 DISC depends on phys. device and path availability
*2 XFER assumes about 8K blocksize, 3390 DASDs

All timings assume usual device/path utilizations

I/O Response Times from SIR SMF

I/O Response Times from SIR SMF

„ Introductory Remarks

- The VSE SIR command is documented e.g. in 'Hints and Tips for VSE/ESA'. Edition 2 was sent to all VSE customers. Edition 4 is available since 05/2000. SIR info is also on the Internet, go via the VSE/ESA home page.
- SMF stands for Subsystem Measurement Facility (standard on all IBM S/390 processors for all I/O subsystems). SMF counters are maintained by H/W in Measurement Blocks inside VSE, provided SMF was set ON. Even on processors w/o these statistics being maintained, VSE creates the data by its own.
- This VSE function is only intended for sporadic trouble shoot. It only provides I/O data for short intervals, since wrapping (overflow) of I/O counters may occur.
- Make sure that you use the latest SIR/SIR SMF version. APAR DY44442 was only the first version.

Recently, an extension of this command has been made available via APAR DY44841, in order to allow to get msec/IO values seen and directly measured by VSE itself:

```
SIR SMF,VSE
```

These results are independent of SMF counters in the I/O subsystem, but, naturally would include e.g. any time in VM or other intercepting programs.

„ SIR SMF Syntax

```
SIR SMF      displays status of SMF (if inactive)

SIR SMF=ON   activates/starts measurement interval
             (if OFF)

SIR SMF      displays current accumulated I/O measurement
             data (if active)
SIR SMF=cuu  displays only data for device 'cuu'

SIR SMF=OFF  deactivates/ends measurement interval
```


I/O Response Times from SIR SMF ...

„ SIR SMF Output (Example 1)

```
SIR SMF
  DEVICE    I/O-CNT   QUEUED   CONNECT   DISCONN   TOTAL
  ---------------------------------------------------------
   240       3572      0.215    2.452     7.832    10.499
   ...
```

„ SIR SMF Output (Example 2)

```
SIR SMF=182
  TIMINGS FOR  182  BASED ON        624 I/O INSTRUCTIONS

  QUEUED    PENDING  CONNECT   DISCONN   DEV.BUSY  TOTAL
  ---------------------------------------------------------
   0.128     0.128    6.272     8.704     0.000    15.232
```

- All times are in msec
- I/O-CNT is the #I/Os since SMF was started. Before VSE/ESA 2.3, SMF data from H/W were used for that, which wrap at 64K (2-byte counters). Starting with VSE/ESA 2.3, VSE maintains internal 3-byte counters (which wrap only at 16M)
  -> You may issue additional SIR SMF displays, to better be able to detect wrapping before VSE/ESA 2.3. Re-calculate e.g. TOTAL I/O times with corrected counters
- QUEUED is composed of IOSQ (wait in VSE), and also includes PEND (except where explicitly shown)
- DEV.BUSY is the time between CE and DE, and, where shown, is NOT included in DISCONN
- Timing details are from the I/O subsystem (SMF): PEND, CONNECT, DISC, except IOSQ and DEV.(still) BUSY, which are from VSE. SMF timer values are wrapped after accumulated 153 hours.
- TOTAL values are from VSE by adding up the detail values. If SMF does not deliver the details, TOTAL times are determined by VSE alone.

Cont'd

I/O Response Times from SIR SMF ...

„ SIR SMF Output (Example 2) (cont'd)

- Displayed data are only reset when a new interval is started with SIR SMF=ON, after SIR SMF was set off
- VSE Virtual Disks:
  All time is CONNECT time
- MDC cached VM minidisks:
  SMF data are setup by VM, and include MDC hits and the data for MDC misses from the I/O subsystem
- The SIR SMF=cuu output must coincide with the same line in the SIR SMF output (TOTAL, and CONNECT)
- SIR SMF=ON overhead is about 0.5% CPU-time for I/O intensive average total loads

Concepts for I/O Intensiveness

Concept of 'Access Density'

```
                          DASD_IO_rate
   Access Density = ---------------------         (8)
                        GB_DASD_installed
```

Access Density

- For MVS the range of 0.4 ... 2.3 IO/sec per GB was observed. For VSE/ESA even with DIM, statistical figures might be higher
- Reduces slightly over time (say around 10% per year), more when DIM is being applied
- Is a constant for an installation (with given data buffer setup and application mix)

Concept of 'Relative I/O Content' (RIOC)

Relative I/O content is a measure for the 'relative I/O intensiveness' of a workload.

```
                    DASD_IO_rate
   RIOC  = ---------------------------------       (9)
            Unit of consumed processor power
```

Relative I/O Content

- Same remarks as for Access Density above
- In the IBM tool CP90, for CPU power so-called 'M-values' are used (very roughly 1 'MIPS' corresponds to about M=50)
- Just as rough description, you may use another metric which may be called

```
      KI/IO  = average #instructions in K per DASD-IO

   For VSE, say
        10 KI/IO    is very I/O intensive,
        30 KI/IO    is avg  I/O intensive
```


I/O Response Time

I/O Response Time (IORT)

Ù **Definition and formula**

```
 Wait in VSE     Time from SSCH to I/O-
 channel queue      interrupt to VSE
                                             I/O
             SSCH                          intrpt
|..............|-------------------------|
                                                Terms used:
    IOSQ       +     Device Service Time       (MVS/RMF)
      = I/O RESPonse Time                      (  "    )

    Q-TIME     +     ACTIVE/UTILTIME           (EXPLORE/VSE)
      = I/O SERVice Time                       (    "      )

    Chq Time   +     ...                       (TMON/VSE)
      = I/O Service Time                       (    "   )
```

```
             IORT = IOSQ + DST       (1)
```

Ù **Cached devices**

- Device Service Time DST refers only to a 'logical device', which is the only thing
  - S/W can 'see',
  - IOSQ depends on
- DST depends, naturally, on the cache hit or miss ratio

Basic Queuing Relation ('Little's Law')

For any 'system' in steady state:

```
 'Avg_population' = thruput_rate x 'avg_time_in_system'
e.g.
   Queue_length         = arrival_rate x queuing_time
   Single_server_util.  = arrival_rate x avg_service_time
   DASD_utilization     = IO_rate x DASD_service_time
```

E.g. 20 IO/sec to a device with 15 msec/IO gives 300 msec/sec = 30% device utilization

IOSQ Wait in Channel Queue

Wait in channel queue (IOSQ)

Ù **Queuing theory formula**

```
                 u
   IOSQ = K x ------- x DST     (2)
                1-u
```

- K depends on the type of service time distribution, of this 'M/G/1 queuing model' and may vary for DASDs between 1.0 and 2.0
- u is the overall utilization of the real/logical device in multi-thread (IO/sec x DST, see 'Little's law')

Formula (2) applies both to uncached and cached attachments, if interpreted adequately.

If IOSQ is too high, usually the device utilization is too high. You may use the PRTYIO command to prefer e.g. Online partitions if problem is caused by unavoidable concurrent Batch I/Os

Ù **A Rough Rule-of-Thumb (ROT)**

```
   Device_Util. <  25% .... 33%       (ROT A1)
or
   IOSQ / DST   <  1.3 .... 1.5       (ROT A2)
                (uncached...cached)
```

For cached subsystems with high hit ratios (DST << 10 msec), IOSQ plays a bigger role. ROT A2 is similarly important, though sometimes (workload dependent) harder to fulfill

For fast utilities (which work with double-buffering and 2 CCBs), IOSQ may be higher, since, in order to get fast turnaround time between I/O interrupt and next SSCH to the same device, a subsequent I/O is placed into the channel queue as early as possible to obtain maximum single device throughput

IOSQ Wait in Channel Queue ...

IOSQ hint for simulated devices

Ù **The more simulated (=logical) devices are used, the lower may be IOSQ for any I/O subsystem**

```
The earlier   an SSCH is isssued to the I/O subsystem,
  - " -       might the physical operation(s) be started,
  - " -       might the DE be presented to the operating system.
```

(The case where this would not help is when your I/O subsystem is totally overloaded anyhow, i.e. it could not manage more concurrency)

Í **Simulate the smallest (or a small) device type**

to split total I/O load on as many physical/logical devices as possible.

This can be done
- even more flexibly when the I/O subsystem also supports small volumes of any size
- most flexibly with the RAMAC Virtual Array with its Virtual Disk Architecture (VDA)

Type of simulated volumes	RAMAC Array DASD	RAMAC Array Subsys	9390 with RAMAC3	RAMAC Virt. Array	RAMAC Scal. Array	Multi. Int. Disk
- 3380-	3339cyl	K	trackf	J,E,K	J,K	J,E,K
- 3390-	3	3	3	1,2,3 (9)	1,2,3 ,9	1,2,3 ,9
- Any #cyls	no	no	no	no	yes*	no**
Total # volumes (max)	4/drawer	128	256 8/draw	256 (1024)	256/ 512	256 8/HDD

```
Disk sizes:   3380-J    885 cyl      3390-1    1113 cyl
                  -E   1770 cyl          -2    2226 cyl
                  -K   2655 cyl          -3    3339 cyl
                                         -9   10017 cyl
 * RSA volumes are allocated with 5 cyls minimum
   and increments of 5 (RSA-2 9) cyls ('FlexVolumes')
      -> allows many small dedicated volumes
** Int. Disk with 'Residual Volumes' of misc. size
```

Device Service Time

Device Service Time (DST)

Ù Definition and formula

DST is the time, the I/O subsystem needs from the issued SSCH until I/O operation is completed

```
SSCH                                                          IO
   | Wait in IOS                                              intrpt
   V for path, CU,    'Device only busy'   'Channel+device
     shared device                          busy'
   |.............|-------------------------|=============|
                                           |=============|
       PEND             DISConnect            CONNect
```

```
DST = PEND + DISC + CONN      (3)
```

Ù PEND time

PEND   is usually low in orderly configured ESA systems.

If device unshared, and CU w/o bigger contention, PEND time only depends on

- the average utilization and number of channel paths (including links between ESCON director and CU) eligible for I/O initiation
- the average channel path connect time.

Ù PEND Time ROT

Approximate real life data for PEND may be up to 1 msec (if unshared) and up to 2 msec (shared):

```
              <  e.g. 1 msec  (unshared)
      PEND                                  (ROT B)
              <  e.g. 2 msec    (shared)
```

Check PEND values contained in performance monitor outputs

Device Service Time ...

Device Service Time for Cached Devices

To simplify the formulae, PEND for cached devices is omitted here. If it should be required, add it like IOSQ when determining IORT

Ù Traditional (1-stage) cached subsystems

(Transfer of data across channel at device speed in case of miss)

Effective Device Service Time is a weighted average:

```
DST_eff = HR x DST_hit + MR x DST_miss         (3a)
```

DST_hit   is equal to CONN = PROT+XFER,
DST_miss is equal to DISC for uncached devices, see next foils.

Since DST_miss is much higher than DST_hit (say 15..20 msec vs 3 msec), it is important to have a high hit ratio HR

Ù 2-stage cached subsystems

(Data are first transferred to the subsystem cache 'stage 1'; and then transferred via the channel 'stage 2')

(-> Transfer via channel is always at cache speed, but starting only when all requested data are in cache)

```
DST_eff =  (PROT+XFER) + MR x DST_miss       (3b)
```

PROT   is the protocol overhead time,
XFER   here is the transfer time out of cache (see CONNect time)
DST_miss is harder to calculate, often a drawer cache is involved and RAID-5 interleaving to multiple physical HDDs, which usually have also a device level cache

Again, a low miss ratio is important for fastest I/O

Device Disconnect Time

Device Disconnect Time (DISC)

DISC is that time of the duration of an I/O operation where (neglecting PEND) only the 'device' is busy w/o occupying the channel path

```
DISC   = Seek + Latency + RPS_delay      (uncached)
                                                     (4)
DISC_eff = MR x (cache_miss_resolution)  (cached)

         (MR = overall miss ratio, Read+Write)
```

Latency   in average is rot_time/2 (except for sequential access)

RPS_delay   is individually n x rot_time (n=0,1,2..), (if individual I/Os would be traced).

```
                 u
RPS_delay =  -------- x rot_time              (5)
               1-u
```

u is the probability that at an arbitrary instant the required 'path' is occupied by other activities. 'Path' may be even 2-way or 4-way.

In the most simple case of 1-way pathing for uncached devices, it holds u = channel utilization by other activities

Average RPS_delay/rot_time			
Avg. util per path	1-way pathing	2-way pathing	4-way pathing
20%	0.25	0.05	0.01
30%	0.45	0.15	0.05
50%	1.0	0.50	0.20

ROT (uncached devices):

```
DISC     <  e.g. 15 msec           (ROT C1)
```


Device Disconnect Time ...

Device Disconnect Time for Cached Devices

Ù Cached devices

For cached devices above definition applies to a 'logical device'

Cache miss resolution   is similar to DISC time in case device would not be cached, but must include any 'lower-interface' contention, caused by staging and destaging for other I/O operations.

(Additional staging of data into cache in case of track caching does NOT elongate miss resolution time of an individual request)

```
      A  DISC (msec)
      |
      |        x
  8 -            x           3990-3 sample results from '92
      |             x
      |                x
  4 -                    x
      |                    x
      |                     x
   ----|----|----|----|----|---> hit ratio
      .2   .4   .6   .8   1.0
```

Í I/O caching essentially reduces DISC time, which is THE major part of Device Service Time

ROT (cached devices):

```
DISC-eff     <  MR x 15 msec         (ROT C2)
```

Device Connect Time

Device Connect Time (CONN)

Definition and formula

CONN is the time of the I/O required to transfer the data across the channel path (including protocol overhead)

```
CONN   = PROT + XFER             (6)
```

PROT is usually a very small time for initiating channel transfers (here non-overlapped part only). It varies with conditions (channel program, caching) and is higher for ESCON than for parallel channels (orders of magnitude is 1 to 1.5 msec, depending on situation and I/O subsystem). It improves with faster technology.

May be roughly determined with performance monitors via CONN, when XFER time is calculated. (Actually, a small part of the protocol-time PROT happens at the begin of the I/O initiation)

XFER depends on actual transfer speed across the channel path:

```
XFER = Blocksize/Xfer_speed       (7)
```

XFER also includes key SEARCH time to localize a key field on track in case of SEARCH KEY CCWs

CONNect time for uncached devices		
Blocksize	Parallel chnl (PROT=1msec)	ESCON chnl (PROT=1.5msec)
2.1K	1.5 msec	1.6 msec
4.2K	2 msec	1.7 msec
8.5K	3 msec	2.0 msec
12.7K	4 msec	2.25 msec
17 K	5 msec	2.5 msec

Device Connect Time ...

Device CONNect time

Cached devices

- PROT time plays a much bigger role due to cache hits and higher transfer speeds
- Data transfer across the channel is mostly at channel (=cache) speed, except for cache misses in traditional cached subsystems.

CONNect times for (2-stage) cached ESCON I/O Subsystems (times in msec)			
Blocksize	MR=0	MR=0.1	MR=0.3
4.2K	1.75	1.83	1.98
8.5K	2.0	2.15	2.35
12.7K	2.25	2.48	2.93
17 K	2.5	2.8	3.4
25.5K	3.0	3.45	4.35
- Values from Tom Beretvas			

ROT (uncached and cached devices):

```
CONN    <  e.g. 4 msec            (ROT D)
           (avg blocksize of 8K)
```


Data Transfer Speeds

Transfer Speeds

The following transfer speeds apply for transfers via channels:

Xfer_speed =
- **channel_speed (usually)**
- **device_speed (cases * below)**

Device speed cases (*):

- non-cached devices
- Read or WRITE misses in traditional cached subsystems (physical 3990-3/6s with real 3380/3390s, 9345s)

2-stage subsystems always transfer at cache=channel speed, even in case of cache misses (RVA, RSA, and all RAMACs seen from transfer speed, refer to the chart in part C)

Channel Speeds

```
                    4.2 MB/sec     Parallel Channels
                   (3.0 MB/sec     in old days)
Channel_speed =
                   17.0 MB/sec     ESCON
                  (10.0 MB/sec     e.g. old 9121s)
```

Higher ESCON speed is partly compensated by higher protocol time (as is a tendency in workloads and DIM exploitation)

ESCON performance benefits for higher blocksizes only

Cache Size Considerations

Cache Size Considerations

General Hit Ratio Curve

```
     A  Hit ratio HR
     |
1.0 -  --------------------||--x-----------
     |                       x
     -             x
     |           x
     -       x
     |                            The higher the cache size,
     -    x                       the higher the hit ratio
     |
     - x
     |
     x--------------------||--------------> Cache size
```

Reducing Miss Ratios (MR)

Increase cache size by factor F to reduce MR by factor of 2

Bruce Mc Nutt, IBM SSD: F=8
Tom Beretvas, Beretvas Consulting: F=4 sufficient
(Observations/Estimates from MVS)

Example: MR=30% (HR=70%) at 128M
MR=15% (HR=85%) at F x 128M cache size

Required Cache Sizes (ROT)

```
2 views:

 -  1 MB cache  per 1 IO/sec DASD I/O rate (HR=80%)  (ROT F1)

 -  1 MB cache  per 1 GB installed DASD capacity     (ROT F2)
    (= 0.1% cache_to_backstore_ratio)
```

Both rules coincide if a system has an 'Access Density' of 1 IO/sec per installed GB DASD.

The ratio of 'active' and 'passive' DASD data is very installation dependent.

If modelling tools are available with actual customer statistics as input, this would be the best predictor for performance

Cache Size Considerations ...

More Cache Size ROTs

Consider that such ROTs may change long term with the change of technology and H/W cost

- **Also observe specific model dependent cache size recommendations, if given specifically (ROT F3)**

  For cached I/O subsystems with

  - high miss resolution time
  and/or
  - a smaller 'lower interface' bandwidth,

  it is required and desirable to achieve good IORTs thru a higher hit ratio via a larger cache size.

- **Do not select very small cache sizes for I/O subsystems which use part of cache storage for storing track related data**
  (e.g. count fields, hit ratios, etc ... in systems with adaptive caching: 3990-6, RAMAC Array Subsystem, 9390 ...)

- **Use all above ROTs (F1, F2, F3), select a reasonable compromise**

NVS Size ROTs

- Applies if NVS is separate from cache
- Very dependent on R/W ratio
- Model dependent recommendations apply (if available)
  (Required size also depends on 'lower I/F bandwidth')

Candidate Devices for I/O Tuning

Candidate Devices for I/O Tuning

For I/O Performance Tuning, specific attention may be attributed to DASDs with the following characteristics.

Only devices are of interest for tuning with
- not too low device I/O rate
- not too low utilizations.

High DASD-utilization (e.g. >25% ... 30%) (ROT A1)

You may proceed in the order of descending

'Response_Time_Volume'= IORT x IO/sec

(e.g. in msec/sec), which is some measure of tuning potential

- **A) Uncached Devices**

  1. **High IOSQ time (e.g. > 5 msec) (ROT A2)**
     Reduce device utilization
     - Move data sets
     - Reduce DST (PEND+DISC+CONN)
  2. **High PEND time (e.g. > 1 or 2 msec) (ROT B)**
     Check
     - whether device is shared
     - utilization of specific path
  3. **High DISC time (e.g. > 15 msec) (ROT C1)**
     Check for
     - high SEEK times (file placement)
     - high RPS misses (= lost revolutions)
       caused by high path utilization from other devices
  4. **High CONN time (e.g. > 4 msec) (ROT D)**
     Make sure that this is caused by higher blocksize, since RPS should be in effect and used

  2. 3. and 4. together result in DST > 21 msec

Candidate Devices for I/O Tuning ...

- **B) Cached Devices**

  To start, all symptoms as for uncached DASD apply

  Refer to last foil

  Cache specific symptoms:

  1. **High Miss (=low Hit) Ratio for cached DASDs**

     If possible, distinguish between READ and WRITEs:

     READ HR < 80%
     (ROT E)
     WRITE HR < 90% (<70% 3990-3)

     Achievable WRITE hit ratios are high for

     - Format WRITEs (all WRITE cached subsystems)
     - Update WRITEs (all RAMAC flavors, all 3990-6s, not for 3990-3s)

     For I/O subsystems with separate NVS, low WRITE hit ratios may be caused by too small NVS size.

     Low overall Hit ratios:

     - Check size of cache via Cache Size ROTs

  2. **High DISC_eff time (e.g. > MR x 15 msec)**
     Corresponds to ROT C2.

     At given miss ratio, tuning similar to 'uncached'

PRTYIO Settings for I/O Priorities

PRTYIO Settings for I/O Priorities

- **Technical Background**

  VSE/ESA schedules the I/Os (SSCHs) according to the following rules:

  - Only 1 I/O is allowed to be started/active on each device
  - All 'System-I/Os' get 'headqueue' priority by using SVC15 and thus are initiated first
  - 'Non-System-I/Os' use SVC0 before entering the Channel Queue
  - The VSE Channel Queue is searched in a 'rotating PUBSCAN' to initate the I/Os
  - If more than 1 request for the same device is ready to be initiated (mostly from different partitions)...

    the sequence of I/O initiation is
    (independent of the partition dispatching priority):

    - According to PRTYIO, if set
    - On First In First Out (FIFO) base, else

- **Purpose**

  Flexibly prioritize the sequence of I/O initiations to the same volume in case of volume contention:

  e.g. prefer Online (CICS) I/Os over batch I/Os to the same volume in case of volume contention

- **Performance Results**

  Runs with Mixed Online and Batch Production loads (using files on the same volumes) showed:

  Using PRTYIO to favor CICS Online I/Os vs Batch I/Os resulted in the specific case in

15% CICS RT improvement
at only 1% reduced Batch thruput

PRTYIO Settings for I/O Priorities (cont'd)

Ù **Recommendations**

PRTYIO can only have an effect in case of volume contention.

In any case, it is promising to try to reduce volume contention, if possible.

Ù **More Usage Hints**

Remember that the priority sequence is REVERSE to the specification in PRTY for the partition dispatching priorities.

Any set/sequence of partitions can be given, separated by commas (,) or equal signs (=).

Dynamic partition classes can be specified, BUT only as a whole ('DYNC'), not as individuals.

```
Example:   PRTYIO F1,F3,F4=F5,DYNC

           Highest:    F1
           Next lower: F3
           Next lower: F4 and F5, treated in FIFO
           Next lower: All dynamic partitions
           Lowest:     All remaining partitions in FIFO
```

It is possible and convenient to put the PRTYIO 'AR (attention routine) only' command into the startup procedure for the BG partition, or into a separate POWER job:

```
// EXEC DTRIATTN,PARM='PRTYIO ...'
```

PRTYIO should be used in the BG startup only after the START F1 and STOP statements, in order to allow dynamic partitions to be included.

PART E.

9340 DASD Attachments

Features of Cached 9340 Subsystems

„ **32 or 64 MB (volatile) cache storage**

„ **Read caching only**

Plus 'Basic Write Caching', but no DASD Fast Write as with 3990-3/6

The following caching bits are exploited

SEQuential (up to 2 seq. tracks in cache, no read-ahead)

BYPass Cache

Inhibit Cache Load

„ **Dynamic, adaptive cache management, controlled by Licenced Internal Code (H/W)**

48 KB track slots

End-of-track staging only

Disabling caching for those data that will not benefit (hit ratio permanently monitored and caching decisions adjusted)

Switching caching off on device level only by CE (in case of trouble shooting)

Í **No S/W control required (but ECKD channel pgms)**

No dynamic caching (on file level) required: 'Self tuning'

No cache statistics available, just fast DASD response times

Performance Patches for Cached 9340s

Mandatory H/W patches:

„ **Microcode level EC 486392**

Fixes cache domination by sequential applications

Sequential bit in DEFINE EXTENT was not correctly used to limit number of sequential tracks in cache to 2

„ **H/W Patch E6392AC**

EC-level EC486392 and up mandatory

Required if S/W sets REC CACHE bits

REC CACHE bits are used,
e.g. by VSAM PTF UD90363 (std since VSE/ESA 1.3.5)

- for DL/1 data component
- for SQL/DS

This H/W patch avoids that VSAM's use of the Record Caching bits (beneficial for 3990-6) is not misinterpreted by cached 9345s, what then resulted in lower cache hits.

Under VM/ESA 1.2.2, this patch requires also APAR VM59317 (PTF UM27166).

3990-6 I/O Subsystem

PART F.

3990-6 I/O Subsystem

- Enhancements
- RDF
- Record Caching/Adaptive Caching
- Exploitation by VSE
- Recommendations
- PPRC

3990-6 General Remarks

General Remarks

- **Purpose**

These 3990-6 charts were setup in order to highlight VSE/ESA specific performance-related aspects of the 3990-6 Enhancements, announced 03/94, and 03/96.

For a full discussion of the functions refer to the official 3990-6 documentation.

- **More Info on 3990-6**

3990-6 Storage Control Enhancements
Ivory letter 194-051, dated 03/01/94

IBM 3390-6 Technical Information
3990MOD6 package on MKTTOOLS disk
(UNCLASSIFIED)

IBM 3990-6 Record Cache Performance Improvements
3990PERF package on MKTTOOLS disk
(IBM INTERNAL USE ONLY)
(Available to your IBM representative, for discussion with you)

3990-6 Record Cache I Performance Results
WSC Flash 9422.2, Doc-ID OZSG023395, 06/94
(IBM INTERNAL USE ONLY)
(Available to your IBM representative, for discussion with you)

3990-6 Large Cache/NVS Performance
WSC Flash 9416.1, Doc-ID OZSG023379, 04/94
(IBM INTERNAL USE ONLY)
(Available to your IBM representative, for discussion with you)

3990-6 Storage Control and RAMAC Array Family Enhancements
Performance White Paper
3990ENWP package on MKTTOOLS disk, 03/96

Solving Performance problems with the 3990-6 Record Cache,
Jeff Berger, IBM, SHARE 83 Session 5068, 08/94

Subject documents include all enhancements, including those with special value for high-end oriented installations.

They also contain quantitative performance results for selected environments. Be aware of the dependency of such performance data and improvements from the workload and the specific environment.

3990-6 Summary

- **3990-6 Control Unit offers increased capacity and additional functions/performance for 'less' or 'non-cache-friendly' applications**

  2x64 Logical Paths maximum (instead of 2x8 for 3990-3)

- **VSE/ESA supports all 3990-6 functions being S/W transparent**

  - bigger cache and NVS sizes
  - faster internal processing and transfer

    Includes faster de-staging of data out of NVS (important for RAMAC Array DASD)

  - Adaptive Caching ('Record Cache II')

- **VSE/ESA 1.3 via PTF also supports**

  - the new Record Mode ('Record Cache I')
  - the 'Regular Data Format'

    (SQL/DS data bases and DL/1 data component)

This PTF, naturally, has been integrated into the newer VSE releases

3990-6 Enhancement Summary

3990-6 Storage Control Enhancements (Summary)

- **Bigger storage sizes**

  (up to 4 GB cache, up to 128 MB NVS, 07/96)

  Faster internal processing and transfer

  Higher overall throughput potential vs 3990-3

- **'Record Mode' (or 'Record Cache I')**

  Of benefit for certain cache unfriendly data, especially at higher accesses/sec per MB cache.

  (Was available before Adaptive Caching = Record Cache II, 01/95)

  - to exploit cache benefits without staging/caching cache inefficient data

- **'Adaptive Caching' (or 'Record Cache II')**

  Flexible cache management

  Standard on all models, CE can switch it off for trouble shooting

  - to dynamically select optimal caching strategy for each track
  - to offload system programmer from tuning activities

Key to both functional enhancements is ...

- **'Regular Data Format' (RDF)**

  Is more an internal function
  - S/W can set bits in DEFINE EXTENT CCW
    (VSE settings discussed under 'VSE/ESA 1.3.x Exploitation')
  - Adaptive Caching can find out RDF property

  - to get 100% Update-WRITE hits for RDF tracks
  - to make Record Mode feasible

Record Caching

'Record Cache (mode)' or 'Record Access'
(or 'Record Cache I')

- **Specified in the DE CCW, valid only for RDF tracks**

  Staging only of the requested record
  no EOT staging as for track caching
  Staging at READ misses, all (RDF DFW) WRITEs are hits
  Cache not wasted for cache-unfriendly data
  Savings in cache storage depends on cache space management. Suited for data sets or volumes with poor caching characteristics
  BUT: CCW must specify record cache mode
  Requires proper determination of cache unfriendly data by S/W, if subsystem cannot determine that dynamically

- **A 'new complementary cache mgmnt algorithm'**

S/W support required (RDF and REC cache bits)

'Adaptive Caching' (or 'Record Cache II')

- **Dynamic switching between record cache and track caching (track individual)**
  3990-6 determines internally (after each IML), which of the tracks benefit from track caching (vs record caching)

- **A combination of 'old' and 'new cache mgmnt algorithms'**

No S/W support required on top of 3990-3 support, Licensed Internal Code only.

Available since 01/95

Regular Data Format (RDF)

Regular Data Format (RDF) Rationale

Most of the records (data fields) of CKD/ECKD volumes are fully formatted with equal length records, without (H/W) key fields (all count fields with same counter).

> Such type of tracks are called here 'RDF tracks'.

- **No RDF (e.g. 3990-3)**

  DASD Fast Write (DFW) requests with 'FORMAT-Write' CCWs oriented to record 0 could and were treated always as hits, even if the track was not in cache ('predictable writes'), resulting in an immediate device end.

- **3990-6 RDF Extensions**

  Cached WRITEs

  A fast device end can also be given to an 'UPDATE-Write', after a record is written into cache and NVS: formerly called 'Quick Write'.

  So all UPDATE-Writes to an RDF track are hits (if DFW ON and NVS ON), provided sufficient NVS is available.

  This extension is valid both for
  - track caching (End-Of-Track (EOT) staging as for 3990-3, but for such requests no staging occurs)
  - record cache mode

  Cached READs

  For RDF tracks, a fast READ device end can be given:

  - if only the referenced record is in cache and NVS (which stems from an RDF-Write in RECORD CACHE mode).
    -> Use of the RDF bit in READ channel programs does not harm, but has no effect.

  - if the record with residual track is in cache (as was always done on 3990-3 with TRACK CACHING without RDF)

  If there is a READ miss, the RPS Miss Avoidance may enhance physical access to the disk (ESCON channels, for parallel channels '3990 Enhanced Mode' is required)

Regular Data Format (RDF) ...

Regular Data Format (cont'd)

- **Selected Background Info**

  - Data will be written to DASD as soon as NVS storage is required, or as soon as CU is less busy.

  - Modified data are copied from the cache into NVS immediately.

  - For channel programs with 'FORMAT-Writes' (WRITE CKD), the RDF bit may be set, but, naturally, cannot bring benefits. This is already a hit as long as oriented to record 0.

  - RDF may also be of benefit for ECKD channel programs with Inhibit Cache Load (UPDATE-Write in case of track caching gives a hit). The frequency depends on the application.

  - For READ or WRITE channel programs with Bypass Cache honoured there may be no benefit by RDF.

  - UPDATE-Writes mostly are done after a READ. So, RDF benefits exist for those situations where a READ
    - was not done before
    - has ocurred 'long' ago.

- **RDF in DEFINE EXTENT CCW**

  RDF bits (byte 7 bit 0-1 in Global Attributes EXTENDED)

  To set the RDF bit also means that the records are unkeyed with standard record zero (R0) on each track.

- **Note**

  System programmers do NOT have to deal with any S/W bit settings in CCWs

  This is done by the access methods and only of direct interest for those
  - setting up own channel programs
  - responsible to understand performance implications

New Cache bit Settings and Effects

New S/W Cache bit Settings and Effects

Update-Writes	Caching Bit Combination		
RDF Record Access	- -	X -	X X
Track Caching only	Miss possible	QW	QW
Track Caching + Record Access	Miss possible	QW	QW +benefits if cache unfriendly
Adaptive Caching	QW	QW (immediate)	QW (immediate)
QW = 'Quick Write' for RDF track Update-Writes			

- **RDF bits beneficial even with 3990-6 Adaptive Caching**

  Traditional track caching mode:
  - higher overall DASD Fast Write hit ratios (Quick Write for RDF Update-Writes)

  Record cache mode:
  - enabling this mode, benefits for cache unfriendly data

  Adaptive Caching:
  - allows immediate decision of Quick Write

- **Record cache bits to be used with care**

  For subsystems with Adaptive Caching, record caching is enforced, even if adaptive caching might have decided to do track caching for certain times and tracks

  For really cache unfriendly data only, especially at higher accesses/sec per MB cache size

  Be aware that with higher avg cache sizes, definition of cache friendliness may shift

IBM 3990-3/6 H/W Defaults for Caching

IBM 3990-3/6 Standard (H/W) Defaults for Caching

Function	Default
Subsystem Caching	ON
Basic Write Caching	ON
DASD Fast Write	OFF *5
NVS	OFF *5
Cache Fast Write	ENABLED *1
All devices cached	YES
Normal (LRU) Caching	ON
Seq. Access Caching	ENABLED *2
Bypass Cache	ENABLED *2
Inhibit Cache Load	ENABLED *2
Record Caching	ON (3990-6)
Adaptive Caching	ON (3990-6)
Sequential Detect	ON (3990-6) *3
Support Fac.Mgmt Opt.	OFF(3990-6) *4

*1 S/W must provide the pertinent bit setting in the DEFINE EXTENT of every chain to be effective

*2 Always available. S/W must provide bit combination in each DEFINE EXTENT to be effective

*3 Sequential Detect is a new function as of 06/96 (refer to separate chart)

*4 New options as of 05/96, may need IBM assistance

*5 May have changed meanwhile to ON, check in any case

3990-6 Exploitation by VSE/ESA

3990-6 Exploitation by VSE/ESA 1.3 and up

- **Adaptive Caching support**
  3990-3 support mostly sufficient, but new bit settings beneficial
- **Record Cache Mode support**
- **RDF set in ECKD channel programs**

  Required for Record Cache Mode, beneficial for Adaptive Caching.

  The VSAM APAR/PTF for VSE/ESA 1.3.x is DY43072/UD90363.

  Under VM/ESA 1.2.2, this PTF requires APAR VM59317 (PTF UM27166)

  This VM fix avoids that VM Fast CCW translation is aborted and the standard VM CCW translation is used instead.

  VSAM, represents 70 to 80% of all I/Os
  Medium potential

  The following holds for VSE/ESA 2.1 and up:

  LIBRarian and FETCH/LOAD
  Some potential, overall
  Page Manager
  Small potential, if paging
  Lock Manager
  Small potential, since update occurs shortly after read
  HardCopy support
  Was CKD in VSE/ESA 1.3 with 2K blocks, 4K in 2.1.
  Small overall potential by RDF
- **Support of 3990-6 'Enhanced Mode' (2.1 only)**

3990-6 Exploitation by VSE/ESA ...

3990-6 Functions for VSE/ESA

3990 Model 6 VSE/ESA Support

	1.1/1.2	1.3/1.4	2.1/2.x
FUNCTION:	==== 3990 BASIC MODE ====		
1GB/2GB Cache (3GB/4GB 06/96) *	YES	YES	YES
32/64MB NVS (128MB 06/96) *	NO	YES	YES
RDF (Regular Data Format) *	NO	PTF ***	YES
Record Mode (Record Cache I) *	NO	PTF ***	YES
Adapt. Caching (Rec. Cache II)*	YES	YES ****	YES
Dual Copy Enhancements *	NO	YES $	YES
XRC (Extended Remote Copy)	NO	NO	NO
PPRC (Peer-to-Peer Remote C.)	NO	NO	YES
Addt'l functions:	==== 3990 ENHANCED MODE ====		
CUIR (CU Initiated Reconfig.)	NO	NO	YES
RPS Miss for Parallel Ch. *	NO	NO	YES

```
*      Performance related function
***    VSE/VSAM only
****   Record Cache II is optimally exploited if RDF set
YES    Software level supports this item
PTF    Software level plus PTF(s) supports this item
NO     Software level does not support this item
-      3990 Enhanced Mode is set in the H/W (CE)
$      Dual Copy for 3990-6 and VSE/ESA 1.3 only supported
       in 3990 BASIC MODE
```

Info on the 07/96 39390-6 enhancements is contained in the documents

- 'IBM 3990-6 and RAMAC Array Family Enhancements'
  Performance White Paper, 03/96, 17 pages
- 3990ENWP document on MKTTOOLS,
  available to your IBM representative

Further 3990-6 Enhancements

3990-6 Sequential Detect

Available since 06/96

- **Detects sequential processing of disk areas**

  Invoked, if >3 cylinders are read sequentially
- **Provides 'sequential' pre-staging benefits as with setting of SEQuential 'bit' for ECKD**

  BUT records are left in cache for normal LRU replacement to allow reuse

  - **Beneficial when no SEQuential bit is set**
    and access is at least short term/partially sequential
  - **SEQ bit setting by S/W is still beneficial**
    in order not to flood cache with sequential (non-reused) data
- **Scope of I/Os potentially affected**

  - SEQ bit setting for IBM channel programs in VSE is not done
    - for any non-convertable CKD channel programs
    - for (normally few) CKD-ECKD converted channel programs
    - for I/O accesses usually random, but in specific cases done sequentially.

    Naturally, also vendor data base products may apply.
  - For ECKD channel programs with BYP or ICL set (provided BYP or ICL are set to be ignored globally, refer to 3990-6 SF options chart):

    Sequential Detect is also in effect.
    e.g. LIBRARIAN BACKUP and RESTORE
    FAST COPY DUMP (OPT>1) and RESTORE and COPY FILE
  - For ECKD channel programs with RECord caching set, Sequential Detect is also enabled

Further 3990-6 Enhancements ...

3990-6 Sequential Detect (cont'd)

- **Function in new u-code is active by default**

  Can be switched off at the service panel

  - Function not retrofitted to RAMAC Array Subsystem, but also contained in 9390 and RVA

- **More information**

  - 3990 Sequential Detect Enhancement, WSC Flash 9633, 06/96
  - IBM RAMAC 3 Array Storage, ITSO Red book, SG24-4835 (12/96), p93

Further 3990-6 Enhancements ...

3990-6 Support Facilities Mgmnt Options

Available since 06/96 as LIC update, but not mandatory
'Extends cache mgmnt functions beyond general purpose workloads'

- **Ignore Bypass Cache**
- **Ignore Inhibit Cache Load**
- **Sequential LRU Processing**

  Avoids early discard of tracks, beneficial in case of early re-use by same or other task

- **NVS Destage Threshold Freespace**

  Reserves NVS storage for 'clustered WRITEs'

- **Customer specific assessment is required**

  Options can have also negative impact, depending on workload: e.g. the first 3 options may require big cache sizes

- **Activation**

  Request assistance by IBM Tucson Engineering (Options now also available via standard Sevice Facility interface)

  May be done e.g. for 'off-shift' loads only, but control unit must be re-IMLed

- **Increased flexibility for specific I/O loads**

More detailed info is contained in:

- 3990 Support Facility Cache Management Options, WSC Flash 9618.1, 05/96

VSE/ESA Caching Recommendations

VSE/ESA 3990-3/6 Caching Recommendations

- **No Adaptive Caching available or installed:**

  3990-3 or elder 3990-6

  Note that S/W cache control is on DEVICE base, NOT on FILE base

- **Cache all 'important' DASD volumes (files),**

  especially those with

  - files having high read/write ratio or many WRITEs to RDF tracks (3990-6)

  or

  - files with read hit ratio of 70% or better

  or

  - VSE Lock File
  - frequently used catalogs or libraries or read intensive VSAM files
  - frequently read VSAM index components

  - VSAM parameters REPLICATE, IMBED not reasonable for cached files

- **Start with a trial to cache all volumes, observe results. If not OK, ...**
- **Restart with most important volumes, add additional ones in a controlled manner**
- **Consider rearranging file distribution across DASDs, if appropriate (not req'd for RAMAC logical volumes)**

VSE/ESA Caching Recommendations ...

VSE/ESA 3990-6 Caching Recommendations

- **Adaptive Caching (Record Caching II) available:**

  Also applies to RAMAC Array Subsystem, RAMAC 3, RSA-2, ...

  - **3990-6 with Adaptive Caching automatically selects caching status on track base**
  - **Provides performance benefits vs all non-adaptive solutions**

    Especially for 'smaller' cache sizes

  - **Offloads system programmers from**

    Speculations on cache friendliness of files

    Moving files around

    Decisions to cache a volume or not

- **New 3990-6 u-code (and nearly all newer I/O subsystems) provide a Sequential Detect function**

Support of Pinned Data for Cached 3990s

Pinned Data for DASD Fast Write (DFW)

- Pinned data occur, when DFW data (in cache/NVS) cannot be written to DASD

They are

- de-staged automatically to DASD as soon as possible
- discarded
  - by a special subsystem IML
    (with 'activated' REINIT)
  - e.g. by the CACHE SUBSYS=cuu,REINIT command
- kept at a power failure, since still 48 hours in NVS

- VSE/ESA 2.1 provides the ability to display pinned tracks in the NVS of the 3990 control unit:

  Extension of the CACHE,UNIT=cuu, STATUS command display

```
PINNED DATA FOR: CYL= ...  TRK = ...
```

  Function retrofitted as PTF to VSE/ESA 1.3/1.4

- Further details

  Refer to '3990 Operations and Recovery', GA32-0133

Peer-to-Peer Remote Copy (PPRC)

PPRC Function

Synchronously 'dual copy' data to a remote disk:

Real-time continuous data shadowing, used for

- Disaster recovery
- DASD migration

- Remote disk of same type, attached to a remote 3990-6
  Includes 3390-3s of a RAMAC 2 Array DASD configuration
- Both 3990-6s are connected via ESCON links
  Up to 20 km distance and more
- Implemented in 3990-6 H/W (LIC)
  Includes RAMAC 2 Array DASD and RAMAC 3
- PPRC for RAMAC Virtual Array (RVA)
  Announced 11/98. Two RVA Models T82, connected via PPRC link

VSE Implementation/Support

- ICKDSF commands, no change in VSE/ESA
- SPE PTF on top of ICKDSF16
  APAR PN66541, PTF UN88673
- Supported by VSE/ESA 2.1 and up

  VSE/ESA PTFs UD50230/UD50231 (APAR DY44407) required

PPRC Performance

Scenario

1. WRITE data into cache/NVS
   of the local/primary storage control
2. Generate Channel End
   to free primary channel
3. WRITE same data to cache/NVS
   of the remote storage control (from primary 3990-6)
4. Generate Device End
   to be presented to the 'application'

   - Data transferred 'cache to cache', no processor involvement
   - DASD Fast Write required for performance reasons

PPRC Performance

- Few msec more effective device response times for WRITEs

  Time to transfer data into 2nd cache/NVS and to signal back
  (4 to 6 msec for 4K data and up to 100 meters, w/o addt'l queuing)

  - Some performance degradation for WRITEs

PPRC Performance ...

Configuration Recommendations

Recommendations apply to 3990-6

- Double the cache/NVS sizes from what is currently installed
  (256M cache and 16M NVS is absolute minimum)
- 4 ESCON host channels should be connected to each 3990-6
- 4 ESCON paths should be connected to each 3990-6,
  2 ESCON paths sufficient if w/o a great deal of sequential WRITE activity
- Configure the secondary storage control identically to the primary
- PPRC stress cases (requiring more resources)
  - R/W ratio < 3:1
  - Transfer block size > 12K
  - Peer-to-peer distance > 9 km
  - WRITE I/O rate > 200 IO/sec

More Info on PPRC

- More info is contained
  - in a document available from your IBM representative:

    'IBM 3990-6 Storage Control, Remote Copy Services Performance'
    03/14/96, 36 pages
    White Paper, MKTTOOLS document 3990RCWP

  - in the 'Red Book'
    'Planning for IBM Remote Copy', GG24-2595-00
    ITSO San Jose, 12/95, 333 pages
    (XRC and PPRC, focus is on MVS)
  - in 'Migrating to RAMAC 2' by Bill Worthington
    VM/ESA and VSE/ESA Technical Conference, Orlando, 06/96
- PPRCOPY commands are described in a further MKTTOOLS document called PPRCDSF

RAMAC Array Family

PART G.

RAMAC Array Family

- **RAMAC Array DASD**
- **RAMAC Array Subsystem**
- **RAMAC Array Storage (RAMAC 3)**

RAMAC Virtual Array Storage

RAMAC Electronic Array Storage
RAMAC Scalable Array Storage
are discussed in separate parts

RAMAC Array Family -Contents-

Here RAMAC also applies to RAMAC 2 and RAMAC 3

Index

RAMAC = RAID Architecture with Multilevel Adaptive Cache

General Remarks

General Remarks

The following charts on the RAMAC Array Family have been setup mostly from a VSE performance view. They do not and cannot replace the extensive official documents available on the MKTTOOLS disk:

'Announcement Overview Presentation Guide'
RAMAOG package on MKTTOOLS

'RAID Primer' White Paper
RAIDRAB package on MKTTOOLS

'RAMAC Dynamic Sparing Paper', 06/95
DYNAMIC package on MKTTOOLS

'RAMAC Array Family Performance White Paper'
06/95 (__ pages)
RAMWP package on MKTTOOLS

These documents are available for you through your IBM representative.

Available as ITSO Red Book:

'IBM RAMAC Array Family', GG24-2509, ITSO Center San Jose, -00, 12/94, 168 pages
Also as GG242509 package on MKTTOOLS

Documents that also reflect the RAMAC 2 announcement of 06/95:

'RAMAC 2 Array Products Performance'
June 09, 1995 (11 pages)
RAM2PERF package on MKTTOOLS
Obsoleted by the RM2BENCH update

'RAMAC 2 Array Products Performance Update'
October 18, 1995 (11 pages)
RM2BENCH package on MKTTOOLS

'IBM RAMAC Array Family Additions (RAMAC 2) Presentation Guide'
October 31, 1995 (95 pages)
SG244563 package on MKTTOOLS

3990-6 Storage Control and RAMAC Array Family Enhancements
Performance White Paper
3990ENWP package on MKTTOOLS disk, 03/96

General DASD Issues

Overall Criteria

- **Cost**
- **Capacity**
- **Performance**
- **Environmental characteristics**
- **'Migrateability'**
- **Attachability**
- **Reliability**
- **Availability**
- **Serviceability**

Means to improve 'R A S' vs 'Base Mode'

- **Dual (I/O) Systems**
  Mirroring on (I/O) system level
- **System Checksum (by S/W)**
  - Uses host CPU resources (OS/400)
  - Stop when error detected
- **Mirroring (RAID-1)**
- **RAID (>1)**

For RAID, refer e.g. to

- next charts
- DASD Array Tutorial GG66-3201
- The RAID Primer (RAIDRAB on MKTTOOLS)
- A High-End RAID Perspective (RAIDHEP on MKTTOOLS)
- RAID Basics by G.H.Cox, Enterprise System Jnl 07/94, pp50-55
- Performance Implications of RAID by YiPing Ding and Subhash Agrawal, Enterprise System Jnl 11/96, pp64-69

RAID Overview

RAID Principles (Overview)

Redundant Array of Independent/Inexpensive Disks

- **Data and/or parity are located on different devices**
  - **Single device failure still allows to access data**
  - **At least WRITEs require access to >1 device**

RAID Level	1	2	3	4	5
Mirroring of a total volume (identical data pattern, no parity) = Dual Copy	X	-	-	-	-
Striping/Interleave Increment: (*) User data distributed across >1 (phys.) devices on	n/a				
- Bit		X	-	-	-
- Byte (*)		-	X	-	-
- Record,'sector' (*)		-	-	X	-
- Track, 'segment' (*) ...level		-	-	-	X
Parity data on	n/a				
- multiple (phys.) devices		X	-	-	X
- extra (phys.) device		-	X	X	-
Arm movement of all devices (access)					
- as ONE arm (synchronized)	n/a	X	X	-	-
- independent		-	-	X	X

* Striping or Interleave increment varies depending on implementation

- RAID-0 stripes data across mult. disks, w/o redundancy -> no 'RAID'
- RAID-1 to RAID-5 provide virtually identical data protection
- Providing a spare device may be an additional implementation feature
- RAID-6 is a RAID-5 implementation with dual parity (2 concurrent device failures allowed)

- **RAID-1: Data Mirroring**
- **RAID-2,3: Parallel Access Arrays**
- **RAID-4,5: Independent Access Arrays**

RAID Overview ...

Parity Aspects

- **Parity schemes**
  - **Even parity**
  - **Odd parity**
  - **Error Correction Code (ECC)**

    A more sophisticated use of parity, written with data

- **Parity bases**

  Parity data may be retrieved/stored in data segments:

  bit, byte, multibyte, record, block

- **Parity usage**

  Reading parity data:

  Concurrently to READs (Parallel Access Arrays)
  Separately, but overlapped to data (Independent Access Arrays)

  Writing parity data:

  Always on physical device separated from associated data

RAID Advisory Board news

- For a recent change in classification of disk storage systems according to extraordinary data protection and availability criteria, refer to

  http://www.raid-advisory.com/criteria.html

RAID Performance Implications

Some RAID Performance Implications

- **Penalty for RAID with parity**

  Additional (phys.) I/Os (*) to handle parity data:

Random WRITE accesses:	Sequential format-WRITEs:
READ old data WRITE new data READ old parity (*) WRITE new parity (*)	Write new data Write new parity (*)

Assuming DASD Fast Write is active, the effect of these additional I/Os is on Read and Write misses of other I/Os and depends on

- degree of possible/implemented parallelism (RAMAC does allow a completely overlapped calculation of parity and fully overlapped parity WRITEs)
- type of WRITE implementation
  - WRITE cache (NVS, RAMAC)
  - combining/grouping WRITEs (9337)
- I/O capacity of the physical HDD array

- **Performance Consequences**

	READs WRITEs	Suited for
RAID-1	No penalty No penalty	WRITE intensive (50% DASD space)
RAID-2	All devices required/accessed	-
RAID-3	All devices occupied (1 arm), but R/W in parallel	Huge seq. files
RAID-4	Good Parity device bottleneck 1 WRITE at a time	High R/W ratio
RAID-5 RAID-6	Good Penalty (see above) R/W in parallel	Tx oriented app's

RAID-5 Benefits

RAID-5 Benefits

Refers e.g. to the RAID-5 implementation of IBM RAMAC

- **No outage from a single disk failure**

  With RAID-6 (RVA) even 2 disks may fail concurrently

- **Failed HDA can be replaced while system up**

- **Data on new HDA is rebuilt automatically while system is running**
  - No host resources required
  - I/O performance degradation during recovery and rebuilt

- **RAMAC Dynamic Sparing Option**
  - HDA can be removed, replaced, reformatted, and good data built while system has full access to data in the drawer. When HDA has been rebuilt, the drawer will sense its presence.

- **System remains available for customer use**

RAMAC Array Family

Summary

Announced 06/94 and 06/95

- **'RAMAC (2) Array DASD'**

  Attaches to 3990-3 and (RAMAC 2 only) -6
  Appears as 3390-3s (or 3380s, 07/96)
  Suited as 3390 replacement or coexistence
  'High Availablity and Performance'

  Performance-wise requires support of DASD-Fast-Write (DFW), not supported before VSE/ESA 1.3

- **'RAMAC (2) Array Subsystem'**

  Attaches to all processors/type of channels
  Appears as 3390-3s or (B13 drawers only)
  3380-Ks or 3380s (B23 drawers, 09/96)
  Suited as 3380 replacement
  'More cost effective solution'

- **Common Characteristics**

  RAID-5 plus Dynamic Sparing
  - Minimized planned and unplanned outages
  - Non-disruptive drawer maintenance

  4 3.5" SCSI-2 disk drives per drawer
  - Each drive with 512K device level buffer

  Drawer caches
  - Battery protected: 'Non-volatile' until data have been written to DASD

  Up to 16 drawers/rack
  (Up to 90/180 GB with B13/B23 drawers)

RAMAC Array Family ...

RAMAC (2) Array DASD ('9391')	RAMAC (2) Array Subsystem ('9394')
DASD units, attach to 3990-6 and 3990-3 (not RAMAC 2) DLSE (4 path access), coexistence with 3390s	DASD subsystem, direct attach to 'all' channels: 3 and 4.5 MB parallel, ESCON (ES/9000,3090,308x,4381,4341,9370, S/390 9672 Rx1 to Rx3 models)
Logical (S/W) view: Fully transparent as	
3390-3 | (2 / B13 drawer) 3380 | (4 / B23 drawer)	3990-2/3390-3 (2/B13 drawer) (4/B23 drawer) or 3990-2/3380-K (3/B13 drawer) (4/B23 drawer) + CACHE ...,REPORT cmd accepted
Physical view:	
9391 Array Rack - - 2..16 9392 Array Drawers	9394 Array Controller 4/8 channels 64 MB-2 GB cache 2..16 9395 Array Drawers
64M Drawer cache (battery protected = non-volatile) 4 3.5" Disk Drives (HDA), SCSI-2 (FBA) Each HDA: 2 GB (B13), 4.0 GB (B23 drawer) 512K HDA cache separate path to drawer cache	
Caching functions:	
All 3990-3/6 functions	Nearly all 3990-3/6 functions
RAID implementation: RAID-5 + Dynamic Sparing Option + Dynamic Disk Reconstruct (B23 RAMAC 2 only)	
S/W Support: (Native device type support required)	
VSE/ESA (1.2*), 1.3, 2.1	VSE/SP 4.1$, VSE/ESA 1.1$ (3380-K) VSE/ESA 1.2$, 1.3, 2.1, 2.2
ICKDSF 16 required/highly recommended (APAR PN62330, PTF UN68459) EREP 3.5 + SPE required	
* RAMAC Array DASD-DFW requires VSE DFW support, not available before VSE/ESA 1.3 $ RAMAC Array Subsystem-DFW ERP is done fully transparently. WRITE hit device ends given only when data in 9395 drawer cache	

RAMAC Array Family ...

RAMAC Family, 9392/9395 Drawer Types

B13 Drawer	B23 Drawer
4 x 2 GB HDAs	4 x 4 GB HDAs (Ultrastar XP)
Simulated Devices, part 1/2	
2 x 3390-3 3 x 3380-K (Array Subsystem only)	4 x 3390-3 (Array Subsystem Only)
Attaches to	
RAMAC Array DASD at 3990-3/6 RAMAC 2 Array DASD at 3990-6 RAMAC Array Subsystem RAMAC 2 Array Subsystem	- RAMAC 2 Array DASD at 3990-6 - RAMAC 2 Array Subsystem
Simulated Devices, part 2/2	
2 x 3380 vols	4 x 3380 vols (09/96)
Attaches to	
RAMAC Array DASD at 3990-6 only (07/96) 1 vol = 3339 3380-cyl (3380-K=2655 cyl)	

HDA Physical Performance Characteristics

	3380-K	3390-3	HDA 0664 in 9392/9395 B13 drawer	HDA 34320 in 9392/9395 B23 drawer
GB/actuator	1.89	2.83	2.0 (tot) 1.4 (net)	4.0 (tot) 2.8 (net)
Tracks/cyl.	15	15	15	15
Cyl./act.	2655	3339	1668	3339
Avg Seek (msec)	16.0	15.0	9.4/11.4 (r/w)	8.0/9.5 (r/w)
Revolution	16.7	14.2	11.2	8.4
Device Data Rate MB/sec	3.0	4.5	5.2	>>5.2
Devicecache	no	no	512K	512K

RAMAC Array Family ...

Other Performance Aspects

- **Each drawer cache operates independently**
- **Asynchronous drawer staging and destaging (data and parity)**
- **Multiple concurrent transfers per drawer cache**
  1 for each logical volume on HDA + 1 per HDA: 6/7 for B13 drawer, 6/8 for B23 drawer
- **Cache bit settings also exploited on drawer level, if beneficial**
  No additional cache statistics on drawer level provided
- **High 'cache-to-backstore-ratio'**
  e.g. 0.5% up to 2% (RAMAC 1 Array Subs.) vs 0.1% (3990-3)
- **CKD/FBA conversion in drawer cache is very fast**
- **Parity data only exist within drawers and HDAs**
- **512K HDA caches used for**
  - **RPS miss avoidance (read/write)**
  - **sequential prestage read data**
- **Automatic load balancing for all logical volumes of a drawer**
  The distribution of tracks of a volume across all HDAs ('3 across, 5 down') gives automatic I/O load balancing across all PHYSICAL HDAs of a drawer.
  Allows e.g. an easy transition from VM Partial to Full-pack minidisks.

  Holds for RAMAC Array DASD at 3990-6 and for RAMAC Array Subsystem

RAMAC Array Family ...

Overall Performance

- The newer the VSE release, the more cache performance functions are supported (Native ECKD channel programs required)
- 'Under most circumstances, RAMAC (1) can offer significant performance improvements, in both response time and throughput, when compared to 3990 subsystems commonly installed today (1994).

  This should be generally true for typical 3990 cache sizes (32-128MB) and workloads with reasonable cache characteristics (>30% read hits).'
- Performance Measurement Results in ...

  'RAMAC Array Family Performance White Paper'

  For RAMAC 2 (discussed later) refer to RAM2PERF PACKAGE
- Notes

  Measurement results shown are for MVS workloads (IMS, TSO, DB2).

  DASD/IO response times shown as function of I/O rates are very similar for VSE, since ...

  VSE/ESA itself uses optimal ECKD channel programs

  - RAMAC does not care by which operating system an SSCH was issued

RAMAC Array Family ...

RAMAC Array Family cache sizes

- Select at least 128 MB RAMAC (1) cache, if possible use 256 MB on the subsystem level

  64 MB in general is too small for READ and WRITE caching.
  Note that a fixed part of the cache is used for control information (refer to PTT)

- Select at least 256 MB RAMAC 2 cache

  Required due to twice the DASD capacity.

  Refer to RAMAC 2 charts

Predictive Track Table (PTT)

- Predictive Track Table is a table with entries for each logical track, in order to optimize performance
- Size of PTT is reducing effective cache size by some amount

  This is of specific interest for RAMAC Array Family, if only 64M cache size is selected
- PTT is being built for RAMAC Array Subsystem at startup and when new drawers are added (a minor impact).

  (Similar aspects apply to 3990-6).

RAMAC Array DASD PTFs

VSE/ESA R0 Performance PTFs (RAMAC Array DASD)

Applies primarily to 3990-3, but also to 3990-6 (R0 means record 0) benefit from it, or RVA

- Performance impact

  Add'tl physical I/Os to R0-record for parity

  Is independent of the type of operating system, but impact depends on the track layout (blocksize)
- Areas of performance impact

  RAMAC Array DASD 9391/9392 attached to 3990-3.
  Also to 3990-6, but only until track was referenced once, i.e. the 'predictive track table' (PTT) entry exists for the track.

  RAMAC Array Subsystem does not need this, since PTT built faster

  Format Writes (i.e. not update writes)
  - e.g. VSAM Initial Load and CA-splits, file extensions or Restores, SAM writes ...
- Performance PTFs

  APAR DY43335, PTF UD49325/49332

  Setting of 'Regular R0 Data Format' (Byte 7 Bit 5) in DX by supervisor (for all VSE components) and in SA-FASTCOPY (not required if Write Track Operation set in LR)

  Applicable to VSE/ESA 1.2 (pre-req UD90367/90368) and VSE/ESA 1.3 (pre-req UD49219/49220). VSE/ESA 1.4/2.1 both include that PTF. Make sure VM APAR VM60996 is applied for VM MDC
- Corresponding R0 PTFs for ADABAS from SAG

```
                 R0              Sector correction
V 5.2.6  AD26048 AD26049
V 5.2.7  AD27001 AD27002
V 5.3.2  AD32024 AD32050         AD32038
V 5.3.3  AD33005                 AD33017
V 5.3.4                          AD34004
V 6.1.2  AD10001                 AD12008
V 6.1.3 + 6.2.1 (PTFs are intergated)
```

The sector correction PTFs apply to all DASD types, contact SAG

RAMAC Array DASD PTFs ...

VM/ESA R0 Performance PTFs (RAMAC Array DASD)

The following is a list of PTFs required for optimal VM/ESA performance with RAMAC Array DASD (and other I/O subsystems, like RVA-2).

Refer to latest VM/ESA documentation for updates and non-performance related PTFs
(Status here is as of 03/09/95).

This list includes APARs for which PTFs may not exist yet, in order to show that a problem area has been identified

Product	APAR	PTF	Description
VSE/VSAM for VM 2.2.0 " " " 2.1.0	VM58884 "	UV90734 UV90733	R0 fix for 9391 DASD
VM/ESA 1.2.2 VM/ESA 1.2.1	VM59200	UM27170 UM27169	R0 fix for TDSK
VM/ESA 1.2.2 VM/ESA 1.2.1	VM59119	UM27058 UM27057	R0 fix for CMS FORMAT

VSAM B/R Performance PTF (RAMAC Array DASD)

- APAR DY43414 (PTF UD49333) for VSE/ESA 1.3/1.4

  This PTF sets the beginning of the extent address in the DEFINE EXTENT CCW for VSAM B/R to the begin of the current extent, in order to allow an optimal sequential de-staging for 3990 type of cached control units during RESTORE

  Refer to next foil

RAMAC Array DASD (9391)

RAMAC Array DASD and Intensive Sequential Writes

- **Volume RESTORE is a very MB intensive write activity**
  (or e.g. LIBRARIAN FORMAT CKD Library)
- **3990-3/6 DASD Fast Write (DFW) is a pre-req for RAID-5 write operations**
- **RAID-5 Write penalty will be hidden as long as data can be written immediately into NVS, but ...**

  If NVS fills up after many MBs written, the write hit ratio may go down:

  Write caching cannot be as effective as it normally is, since cache/NVS size exhausted by being 'physical device bound', i.e. the physical device(s) itself becomes the bottleneck
- **3990-6 has a more sophisticated DFW implementation than 3990-3**
  (NVS destaging implementation for RAMAC Array DASD)
- **Conclusions**
  - **Bigger NVS sizes will help for such cases**
  - **RAMAC Array DASD with intensive sequential writes may show lower performance for 3990-3 than for 3990-6**
  - **Performance results for this kind are not representative for overall DASD performance**
  - **3990-6 is much better suited than 'old' 3990-3**

RAMAC Array Subsystem (9394)

Statistics for Subsystem Cache

- **CACHE UNIT=cuu,REPORT provides cache statistics for device cuu**
  - **Same as for 3990-3 and 3990-6**

    Refer e.g. to 'DASD Caching in General' part
  - **Caution:**

    BYPass Cache counter erroneously is increased, if record caching is set
    (e.g. for DL/1, SQL/DS by VSAM)

    In RAMAC Array Subsystem u-code level B482658 this reporting problem is solved
  - **CACHE UNIT=cuu,STATUS NOT accepted**

    Not required since all functions in H/W enabled by default
- **VM/ESA statistics for RAMAC Array Subsystem**

  RAMAC Array Subsystem cache statistics can be obtained by VM/ESA, but only if the following VM APARs/PTFs have been applied:

```
VM59200  UM27169    for VM/ESA 1.2.1
         UM27170    for VM/ESA 1.2.2

VM59341  UM27152    for VMPRF 1.2.0/1.2.1
```

  The latter PTF is for VMPRF and required to format the cache statistics

RAMAC Array Subsystem (9394) ...

RAMAC Array Subsystem Performance Hints

- **Blocksize of 4K or larger is optimal.**

  For smaller blocksizes, RAMAC Array DASD with 3990-6 is the best suited alternative
- **For WRITE hits, about 3 to 4 msec more time is required vs 3990-3/6**

  Data have to be transferred to the non-volatile drawer cache first

  This may become a problem, if e.g. logging was already a critical point with 3990-3/6 cached subsystems, using DFW (e.g. the SAP R/2 Update transaction or CICS logging)
- **Sequential bits highly beneficial for massive sequential operations**

  RAMAC Array Subsystem avoids that massive sequential data w/o a SEQuential indication in DEFINE EXTENT can flood the cache.

  With SEQuential bit, all data are fully cached, but discarded early

VSAM PTF for Format WRITEs

- **Make sure VSAM PTFs UD49763 are installed**
  Solves APAR DY43836

RAMAC Array Subsystem Microcode

Very recent observations with a VSE customer in Italy:

U-code level EC 29119 showed much better I/O performance than level EC 29118c

(especially sequential performance for READ seemed to be affected).

Contact IBM to get info on latest available u-code level

RAMAC Array Subsystem (9394) ...

RAMAC Array Subsystem ADDs

ADD statement 'device type' values ADD cuu, xxx(,SHR) for RAMAC Array Subsystem volumes

	Emulation Mode		
	3990-2/ 3380-K		3990-2/ 3390-3
VSE Release	parallel	ESCON	parallel+ESCON
VSE/SP 4.1	3380 (c)	n/s	n/s
VSE/ESA 1.1	3380 (c)	n/s	n/s
VSE/ESA 1.2.0 (a)	3380 (c)	n/s	n/s
VSE/ESA 1.2.0 (b)	ECKD	ECKD	ECKD
VSE/ESA 1.2.1-3	ECKD	ECKD	ECKD
VSE/ESA 1.3/1.4	ECKD	ECKD	ECKD
VSE/ESA 2.1 & up	ECKD	ECKD	ECKD

n/s Not supported
(a) Without PTF for APAR DY41099
(b) With " " " "
(PTF for APAR DY41099 is standard since 1.2.1)
(c) Synchronous_2 mode must be set.
Normal and default setting is non-sync mode.
See also WSC Flash 9553.3 '9394 Synchr. Settings' and WSC Flash 9507 'VSE Considerations for RAMAC Array Subsystems (9394)'

- Do not use the ADD 'EML' parameter, except for the case 'Vendor program deficiency'. 'EML' does NOT specificly apply to RAMAC
- In all cases above, device type '6E'(ECKD) used, except where ADDed as 3380
- Holds also for VM/VSE dedicated devices, for VM/VSE minidisks see VM requirements

Consequences of ADD command variations (e.g. VSE/ESA 1.3, 3380-K)

ADD cuu, 3380	Gives CKD channel programs, except if overruled at IPL time (overruling depends on attachment)
ADD cuu, 3380,EML	Never use, forces CKD!
ADD cuu, ECKD	Gives ECKD channel programs, if RAMAC (device type not changed by VSE)
ADD cuu, ECKD,EML	Forces ECKD channel programs

RAMAC Array Subsystem (9394) ...

The following detailed info (very similar to WSC flash 9507) has been written for the information of technical specialists and is courtesy of Axel Pieper (VSE, Boeblingen) and Bob Shomler (RAMAC, San Jose).

Additional Considerations for 3380-K Emulation:

RAMAC Array Subsystem, including 3380-K format, is designed to run with ECKD channel programs. For VSE/ESA release levels that have ECKD support, all RAMAC Subsystem DASD including 3380-K format should normally be defined to VSE by 'ADD cuu,ECKD'.
However, some vendor program products and applications that work with 3380-K are sensitive to the device type code in VSE's physical unit block (PUB). These programs look for a '6C' and may not recognize the '6E' that VSE will place there for ECKD devices.

Other software may have a similar sensitivity to the device type in VSE's DTF. The DTF may contain a device type of 0C or 0E, with the 'C' or 'E' being set the same as the 'C' or 'E' in the PUB. VSE will put the value from the DASD's Read Device Characteristics data in the DTF when the device is recognized as ECKD. This will be 0E for 3380-K. VSE will put 0C in the DTF for 3380 when it is not recognized as an ECKD device. VSE determines that a DASD is ECKD or CKD based on the ADD statement and the DASD's indication that it is capable of nonsync operation. These combinations are shown in the table at the end of this section.

(Note: The 6E in PUB and 0E in DTF will be present for any nonsync-capable ECKD DASD, real devices and RAMAC emulated devices.)

An 'ADD cuu,3380,EML' statement may be used to force 6C into the PUB and 0C into the DTF to enable this software to run. However there are some corollary effects and requirements that will be explained here.

An effect of having a 6C in the PUB is that VSE/ESA will generate CKD channel programs rather than the more efficient ECKD channel programs, and it will not convert application CKD channel programs to ECKD as it will do for an ECKD device.

VSE/ESA also will separately present Channel End and Device End status on (CKD) DASD write operations, with the application being posted complete at Channel End time. For ECKD devices (PUB = 6E), VSE will hold Channel End to be presented together with Device End to the application, posting the application I/O complete at Device End time (Device End posting).

Some application software will not be able to properly handle separate presentation of Channel End without Device End. This may be software that also requires a 6C in the PUB entry, or there may be software not sensitive to 6C/6E that could have this problem.

RAMAC Array Subsystem (9394) ...

The RAMAC Array Subsystem can present some operation exceptions only at at Device End time, as is the case for any nonsynchronous DASD. Thus if it is necessary to force VSE to CKD mode (via 'ADD 3380,EML') then it also will be necessary to activate Device End posting for those DASD. Device End posting in VSE can be accomplished by one of three means:

1. Device End Posting can be explicitly requested by an application.
2. RAMAC Subsystem synchronous-2 VPD mode may be set, but this will be effective only for RAMAC subsystem attached via parallel channels.
3. Device End posting will be automatic for devices recognized as ECKD.

Option (1) may not be feasible for existing applications.

Option (2) does not help ESCON configurations, and on parallel channels the additional channel connect time can impact system performance, Option (3) would be the ideal choice, but is denied by the application software requirement for 6C in the PUB (or 0C in the DTF).

The remedy for these applications is an update for these programs to accept 6E in the PUB entry and 0E in the DTF. This will allow 3380 format DASD to be defined and operate as ECKD nonsync (ADD cuu,ECKD), avoiding CKD channel program and extended connect time performance impacts, and using the VSE Device End posting inherent in VSE ECKD support. The customer should contact the vendor or application maintainer to request an update to recognize 6E/0E (for device type identification or track capacity calculation), both for real nonsync devices and RAMAC emulated devices.
Until the application software can be updated to recognize 6E/0E, a workaround for a parallel channel attached RAMAC Subsystem is to ADD the DASD as 3380,EML and set synchronous-2 in RAMAC Array Subsystem VPD.
The only workaround for ESCON is to define the DASD as 3380,EML, set the subsystem VPD to nonsynchronous, and either use option (1) above -- request Device End posting (if feasible for the application) or request a temporary VSE patch as described below.

VSE development can provide a temporary patch for VSE/ESA 1.3 and 1.2 to force Device End posting for 3380 devices until application software can be updated to work with devices defined as ECKD, and recognize 6E in the PUB entry and 0E in a DTF. If needed, this should be requested from VSE development by software PMR for systems that have all of the three following conditions:

RAMAC Array Subsystem is attached via ESCON channels, or the performance impact of synchronous-2 operation prevents successful system operation

Any software running on that system requires a 6C PUB value or 0E DTF device type to run, and

Application-requested Device End posting is not (or cannot be) used.

RAMAC Array Subsystem (9394) ...

Additional Notes:

Do not set synchronous-1 in VPD. VSE/ESA 1.2.x and VSE/ESA 1.3 without the fix for APAR DY43207 will insert a 6C in a 3380 PUB entry as a result of RAMAC Array Subsystem VPD synchronous-1 or synchronous-2 being set.

Synchronous-1 should not be set in VPD, since 6C in the PUB will inhibit ECKD merged Channel End and Device End posting, which will cause a potential data integrity exposure when exception conditions are reported with Device End.

For programs that presently use the 6C in the PUB to identify 3380 device type for track capacity information, VSE has a GETVCE service that can be used for this purpose, eliminating the need to interpret PUB content.

VSE ECKD Recognition, PUB, and DTF Values

The following table shows how VSE will see a 3380 RAMAC Array Subsystem DASD based on VPD mode and how the device is defined (ADDed) to VSE. Note that some of the combinations should not be used; this is just to show the VSE action for these combinations:

Note	ADD statement	DASD VPD	VSE sees device as	PUB	DTF	PROBLEM
	ADD cuu,ECKD	nonsync	ECKD	6E	xx	
	ADD cuu,3380,EML	nonsync	CKD	6C	0C	split CE/DE
(1)	ADD cuu,3380	nonsync	ECKD	6E	xx	
(2)	ADD cuu,ECKD	sync-2	ECKD	6E	xx	
(3)	ADD cuu,ECKD	sync-2	CKD	6C	0C	
	ADD cuu,3380	sync-2	CKD	6C	0C	
	ADD cuu,3380,EML	sync-2	CKD	6C	0C	
(2)	ADD cuu,ECKD	sync-1	ECKD	6E	0C	
	ADD cuu,ECKD,EML	sync-1	ECKD	6E	0C	

Notes:
(1) Nonsync VPD mode causes device to be defined as ECKD (MSG1I71I)
(2) VSE 1.3 with fix for APAR DY43207
(3) VSE 1.2.x and VSE 1.3 without fix for APAR DY43207, VSE will force 3380 (MSG0I71I)
xx device dependent (0E for 3380s, 24/26/27/34 for 3390-3/1/2/9s)

RAMAC and some Vendor Products

RAMAC and some Vendor Products

Vendor/Product	Rel	Comments
ALTAI		
ZEKE	4.0.B	- DTF does not recognize 6E
	4.1.C	fix available
CA		
DATACOM/DB	8.0	- fix available
	8.1	- fix available
VOLLIE	4.3	- no fix, out of service
	5.0	- fix available (phase OLLE6100)
RAMIS	7.1.0	fix: RA71174D,RA71178D,RA71180D RA71206D,RA71207D,RA71208D
IDEAL	2.1	- use VLSBKUP (not VLSUTSE) for Backup
CINCOM		
SUPRA	1.2.5	fix: 942139,942140
	1.2.6	fix: 942139,942140
	1.3.5	fix: 942139,942140 + patch
	2.6.0	fix: 942139,942140 + patch
PHOENIX		
FALCON D/E	14.1	- no fix...install 16.0
SAG		
ADABAS	5.x,6.x	refer to RAMAC Array DASD PTFs
Others?		

This table is updated based on available info.
It can only be a hint for faster problem solution.
Let us know if something is missing or wrong.

Thanks to all who provided patches and solved our common customers problems.

Contact the vendor for maintenance.

RAMAC 2 Specific Performance Remarks

Refer also to RAM2PERF PACKAGE on MKTTOOLS, available through your IBM representative

A General High Capacity HDD Issue

- **At same total GB, higher msec per I/O may occur,**
  increasing the probability of being device bound through more IO/sec per HDD
  - **This aspect is mostly true for**
    - **- uncached attachments**
      Cached devices also suffer for loads if hit ratio small
    - **- HDDs without striping**
      (if logical device is smaller or identical to physical volume)
      RAMAC RAID-5 suffers less from that effect

RAMAC 2 Aspects

- **This potential device/HDA effect only marginally applies to RAMAC 2**
  - **Much of this effect can be hidden by big cache sizes**
  - **RAMAC RAID-5 automatically balances I/Os across physical HDAs**
  - **RAMAC 2 HDAs (B23) are slightly faster than its predecessors (B13)**
- **Naturally, much too high I/O rates may impact device response times**

RAMAC 2 Specific Performance Remarks ...

Major RAMAC 2 Performance Findings (vs RAMAC)

- **To achieve comparable performance, a larger amount of controller cache must be provided in some cases**
- **At very large sizes of controller cache, RAMAC 2 performance may exceed that of RAMAC, despite twice as much data in each B23 drawer**
- **At same big cache sizes, RAMAC 2 can outperform RAMAC in specific cases,**
  provided the same amount of active data is stored on each drawer
- **In practice, comparable performance to RAMAC**

RAMAC 2 Cache Size Recommendations

Full (!) configurations (180 GB)

RAMAC 2 configuration	Cache size
Array DASD at 3990-6	256 MB
Array Subsystem model 2	256 MB
Array Subsystem model 3	128 MB (*)
3990-6 NVS size	> 16 MB often beneficial

* More internal paths available, 512 MB would be too costly

RAMAC 3 Array Storage

RAMAC 3 Array Storage -Summary-

Announced 09/96

A RAMAC 3 Array Storage subsystem consists of

- the RAMAC 3 Storage Control and
- 1 or 2 RAMAC 3 Storage Frames

- **RAMAC 3 Storage Frame**

  9392 B33 drawer

  Fast IBM Ultrastar 2XP 9.1 GB HDD
- **9390 Storage Controls**

  Offer complete suite of 3990-6/RAMAC 2 functions

  Includes the new Sequential Detect and PPRC

  Increased 'lower interface' bandwidth

  ESCON attachment only

  RAMAC 3 also attachable to 3990-6
- **45.4 to 726 GB, with 22.7 GB increments (1 drawer)**

  2 to 32 drawers, 2x RAMAC 2 capacity
- **VSE/ESA support (as for 3990-6)**

  VSE/ESA 1.3/1.4 (Basic Mode, default)
  VSE/ESA 2.1/2.2 (Also Enhanced Mode).

  Also 3380s must be ADDed as ECKD

RAMAC 3 Array Storage ...

RAMAC 3 Array Storage -More details-

- **9390 Storage Control models**

  Single and Dual CU model
  Only attaches RAMAC 3 disks

Model	-001	-002
Cache size NVS size #ESCON channels #addresses #Storage Frames	256M-4G 32,64,128M 4,8,12,16 128 1	2 x '-001'

- **RAMAC 3 Storage Frame (9391-A30)**

  Contains 2 to 16 drawers

  New High Speed Device Adapter (4 HSDAs)

  11.9 MB/sec data rate at lower interface, between storage control and drawer (3x RAMAC 2)

  Applies also to 3990-6 attachment (LIC update)
- **9392 B33 Drawer**

  4 9.1 GB HDDs, representing a RAID-5 array
  22.7 GB effective capacity
  Up to 8 logical volumes (3390-3)
  plus 3380 track format function
  64 MB non-volatile drawer cache

RAMAC 3 Specific Performance Remarks

Performance Aspects

„ **High Concurrency**
All 4x16=64 HDDs of a frame can transfer (READ/WRITE) concurrently

„ **Ultrastar 2XP 9.1 GB disk drives (HDDs)**

- Media data rate *	10.2 to 15.4 MB/sec
- Drive cache	1 MB (not 512 KB)
- Latency	4.17 msec
- Min seek (Read)	0.5 msec
- Avg seek (R/W)	8.5/10.5 msec
* (more data in the outer zones) -> 1.25 times the data rate of Ultrastar XP	

„ **Enhanced NVS Management in 3990-6 and 9390 with Branching WRITEs**

When large amounts of sequential data are written (be it via 'Seq. Bit' or via Sequential Detect)...

the second copy of the data is directed into the nonvolatile drawer cache (instead of filling the NVS).
This avoids NVS full conditions for other (random) WRITEs

Transfer from the ESCON channel into subsystem cache and drawer cache is being done in a 'branching WRITE' manner.

„ **Up to 8 logical volumes per drawer**
On RAMAC 1/2, only 4 logical volumes were possible

RAMAC 3 Size Recommendations

# Vols	Total GB	MINIMUM Cache	MINIMUM NVS	RECOMMENDED Cache	RECOMMENDED NVS
<=32	<=90	256M	32M	512M	32M
32-64	90-180	512M	32M	1-2G	64M
64-128	180-360	1G	64M	1-4G	64M
128-256	360-720 *	2x 1G	2x 64M	2x(1-4G)	2x 128M

- RAMAC 3 Array Storage at 9390-001 or 3990-6
* Configuration requires 9390-002 or 2 3990-6

RAMAC 3 Performance (vs RAMAC 2)

RAMAC 3 Array Storage -Sequential Performance-

Measured Elapsed Times

Workload	RAMAC 2 Array DASD	RAMAC 3 Array
VSAM KSDS Seq READ (w/o Seq. Detect)	22 min	not applicable
VSAM KSDS Seq READ (w/ Seq. Detect)	13 min	7 min
QSAM Seq. WRITE	16 min	10 min

- 8 volumes active simultaneously across 2 drawers, 1500 cylinders used for each volume
- 18 Mb/sec ESCON channels
- 4K blocksize for VSAM READ, 27K for QSAM WRITE
- OS/390 performance results
- VSE/VSAM can set SEQ indication for KSDS also SAM with ACB access (VSAM), but not SAM with DTF access (BAM)

í **Greater than 2 times sequential throughput**

Small block sequential WRITEs are up to 3 times faster

RAMAC 3 Performance (vs RAMAC 2) ...

RAMAC 3 Array Storage -Random Performance-

Configuration	Throughput	Response Time
2x 3990-6, each 512M Cache 32M NVS 180G RAMAC 2	2x 275 IO/sec @ 15 msec RT Total 550 IO/sec	Each 14 msec RT @ 250 IO/sec Total 500 IO/sec
1x 9390-001 1G Cache 64M NVS 360G RAMAC 3	580 IO/sec @ 15 msec RT	13 msec RT @ 500 IO/sec
1x 9390-001 4G Cache 128M NVS 360G RAMAC 3	770 IO/sec @ 10 msec RT	7 msec RT @ 500 IO/sec

- OS/390 DB2 workload was used
(Example for cache-unfriendly, random access)
with 3:1 R/W ratio
17 KB avg transfer/IO
29% Sequential stage
63%/83% cache hit ratio @ 1/4 GB total cache

í **'Equal or better performance at double capacity'**
(vs RAMAC 2, Array DASD or Subsystem)

More details on performance are contained in the RAMAC 3 White Paper (RAM3PERF), available to your IBM representative

More Info on RAMAC 3

More Information

For more information, refer to

IBM RAMAC 3 Overview (Presentation Guide), 09/96, 23 pages
Available to your IBM representative (MKTTOOLS)

IBM RAMAC 3 Array Storage Product Announcement, 96-09-10

IBM 3390-9390 Storage Control Introduction (updated), GA32-0098-08

IBM RAMAC 3 Array Storage (technical presentation), ITSO San Jose Red Book, SG24-4835-00, 12/96, 210 pages

This is a really excellent book for technically interested people

IBM RAMAC 3 Array Storage -Continuing Performance Enhancements- White Paper, by John Bacho, 96-10-09, 31 pages

As RAM3PERF on MKTTOOLS

RAMAC 3 Array Storage Spec Sheet, G2256688,
As G2256688 on MKTTOOLS

RAMAC Array Family, RVA

PART H.

RAMAC Array Family, RVA

„ **RAMAC Virtual Array Storage 2**

„ **RAMAC Virtual Array Storage 2 Turbo**
Includes the Models X-82 and X-83

RAMAC Array DASD
RAMAC Array Subsystem
RAMAC Array Storage (RAMAC 3)
are discussed in the previous part

RAMAC Electronic Array Storage
RAMAC Scalable Array Storage
are discussed in the next part

RAMAC Virtual Array Storage Overview

Overview on RVA Foils

- RVA Summary
- RVA Models
- More RVA General Performance Aspects
- RVA-2 Turbo Specifics
- RVA-2 Turbo Performance

- General Log-Structured File Aspects
- RVA IXFP Program
- IXFP DDSR for VM/VSE

- RVA SnapShot 'Instant' Copy
- IXFP/SnapShot for VM/VSE
- IXFP/SnapShot for VSE/ESA

Detailed Technical Info

- ITSO Redbooks
  - 'IBM RAMAC Virtual Array', ITSO Redbook, SG24-4951-00, 07/97, 475 pages
  - 'RAMAC Virtual Array, PPRC and IXFP/SnapShot for VSE/ESA', ITSO Redbook, SG24-5360-00 (01/99)

To get ITSO red books, refer to
http://www.redbooks.ibm.com

RAMAC Virtual Array Storage 'RVA' (9393)

RAMAC Virtual Array Storage -Summary-

Product evolved from StorageTek's Iceberg 9200 Disk Array Subsystem.

More info is contained in

- 'IBM RVA Storage Introduction', GC26-7168
- 'IBM RVA Planning, Implement. and Usage Guide, GC26-7170

„ **Complements RAMAC Array Family**
- High capacity and performence
- High availability
- Low cost
- Lowest footprint (RVA-2)

„ **Up to 1680 GB (effective) capacity**
(assumes a 3.6:1 compression ratio)
Refer to RVA Models Summary

„ **Managed Array of Independent Disks**
- Dynamic assignment of used storage
- RAID-6 (dual parity)

„ **Appears as 3380s or 3390s, at 3990-3 with DFW**
- Single/double/triple capacity volumes
- Flexible definitions
- Also 3380s must be ADDed as ECKD

„ **VSE Native support documented**
(refer to FLASH 97-030 as of June 30, 1997),
VSE/ESA 1.3 and up under VM
(1.3 is expected to run, as 1.4 is still comitted by IBM)

„ **IXFP/SnapShot for VSE/ESA is new in 09/98**

RAMAC Virtual Array Storage 'RVA' (9393) ...

'RAMAC Virtual Array -Summary- (cont'd)

„ **Managed Array of Independent Disks**

'Virtual Disk Architecture' (VDA), manages allocation of logical space/data to real space

No predetermined location of tracks
('homeless tracks')

- **'Log Structured File' system**
Track directory is in replicated cache and also on disk
(for safety reasons)

- **Automatic load balancing across physical HDDs**

„ **Built-in compaction and compression of data**
Done at channel level,
transparent to application and S/390 S/W
Typical compression factor is about 3.6
Sometimes even higher values for this modified LZ compression.
Lower values if data already software compressed.

„ **Up to 6 GB (effective) cache size**
(refer to RVA Models Summary)

„ **16 MB of (effective) NVS**
8 (=2x4) MB actual size

„ **Flexible definition of up to 256 logical volumes**
3380 and 3390 single to triple capacity volumes.
Up to 64 devices per up to 4 logical 3990-3 control units.
Use of many volumes easy (VDA), to reduce IOSQ.

On X-models, up to 1024 volumes are possible, plus 3990-9's

RVA 9393 Model Summary

RVA 9393 Model Summary

	RVA-1 RVA-2 -001 -002	RVA-2 Turbo -T42 -T82	RVA Turbo -X82 -X83
Announced	06/96 09/96	04/97	05/99 07/99
Effective disk capacity	160G -726G	160G 160G -726G -840G	160G 290G -840G -1680G
Max eff. cache	2G 3G	4G 6G	6G
#channels Paral. ESCON	16 8 or 16	- 8 or 16	- 16
#conc.chnl data X-fers	4	4 8	8
#log. devices	256	256	1024
#log.ESCON paths	32 ->128	128	128

- Increased number of log. ESCON paths was retrofitted to the 'non-Turbo' models
- X83 uses 9G HDDs, all others use 4.5G drives

Any RVA user/effective/nominal capacity is valid for
- an average disk compression ratio of 3.6:1
- the recommended 75% Net Capacity Load

Í **Performance/Capacity improved continuously over time**

„ **RAID-6 Disk Arrays**

2 types of disk arrays

```
 8 HDDs  = 5 +2P +1S      80/160 GB
16 HDDs  =13 +2P +1S     210/420 GB
```

2, 3, or 4 Disk Arrays in an RVA

RVA 9393 Model Summary ...

„ **3.5" HDDs/disks used**

Capacity	4.5 GB	9.1 GB
Model: IBM Ultrastar	2XP SCSI	9LP ? SSA
Rotational Speed RPM	7200	7200
Media data rate MB/sec (inner to outer)	10.2 to 15.4	11.5 to 22.4
Latency msec	4.2	4.2
Minimum SEEK msec	0.5	0.7
Avg SEEK msec	7.5	6.5
Actuator buffer KB	1024	1024

- 9 GB HDDs on Model X83 only

RAMAC Virtual Array Performance

More RVA Performance Aspects

„ **Effective usage of physical storage:**

Only actual data is stored

unallocated space does not reserve capacity
allocated space, but unused is not 'stored'

> space not occupied/reserved until data actually written

built-in data compression and compaction
Data compressed before entering cache,
4 independent LZ compression engines

CKD/ECKD inter-record gaps need not be 'stored'
Likewise applies to the RAMAC Array Family

Intelligent freespace collection and mgmnt

Messages regarding available space start at 85% utilization of physical space
(Net Capacity Load NCL)

„ **Fast and highly concurrent data transfers**

Up to 4/8 concurrent channel transfers, plus up to 8 operations w/o transfer
(e.g. interpret channel programs, initiate cache miss resolution...)

Up to 14 concurrent transfers between cache and disk arrays
Unit of staging/destaging is a full (logical) track

RAMAC Virtual Array Performance ...

Other Performance Aspects (cont'd)

„ **RAID-6 Dual Parity Architecture**

- Allows 2 simultaneous disk failures in 1 array (5+2, 13+2)
- Updated data and parity always written to a new location, thus reducing RAID-5/-6 WRITE penalty

„ **Self tuning arrangement of data:**

Tracks of a single logical volume are spread across physical HDDs, reducing the effect of 'hot spots'

Likewise applies to the RAMAC Array Family

„ **Sequential detect function**

„ **Highly efficient destaging**

Updates are done to a new physical location
Destaging tracks are collected into groups and bulk transferred

„ **RVA-2 performance is equivalent to RVA with latest u-code**

In fact, the following is the situation meanwhile (10/96):

Average (random) performance is better, sequential throughput is about 10% higher.

RAMAC 3 may give very slightly better performance than RVA-2 for comparable configurations

Refer also to the 'white paper':
'An Overview and Comparison of RVA-2, RAMAC 3 and RSA Performance'
as RAMFAM package on MKTTOOLS disk, 96-11-04, 19 pages,
available to your IBM representative

RVA-2 Turbo

Here only additional deltas to RVA-2 are addressed

General

„ **9393 Turbo Models announced 04/97**

- same sizes for cache/NVS/GBs on disk
- ESCON only
- SOD for a function similar to PPRC on 3990-6 and 9390 (Function announced 98-11-03)

„ **Turbo shared memory with faster access time**

More internal concurrency and CU internal paths

T42 with up to 4 concurrent data transfers
T82 with up to 8 concurrent data transfers

> faster disk service times at low loads

> higher maximum throughput

Performance

„ **Up to 50% higher throughput (T42 vs 002)**

Smaller RVA subsystems (210GB or less) will show smaller improvements

„ **Up to 20% higher throughput (T82 vs T42)**

Caused by more concurrent channel data transfers, appplies to loads with high blocksize or high hit ratios.

RVA-2 Turbo vs RVA-2 Performance

„ **Max. Sequential Throughput (MB/sec)**

PAWs Workload	RVA-2 9393-002	RVA-2 Turbo 9393-T42	Delta
Single Stream			
QSAM Read	5.5	5.9	+ 7%
QSAM Write	6.3	7.0	+11%
VSAM Read	2.8	3.0	+ 7%
VSAM Write	2.6	3.0	+15%
16 Streams			
QSAM Read	24	33	+38%
QSAM Write	19	29	+53%
VSAM Read	14	20	+43%
VSAM Write	8.8	13.8	+57%

Access Method:	VSAM	QSAM
- Block Size	4K	27K
- Blocks transferred per I/O	12	5
Measured Configurations:	9393-002	9393-T42
- ESCON Channels	8	8
- Effective DASD Storage	290G	420G

„ **RVA-2 Max. Random Access Throughput (IO/sec)**

RVA-2 Turbo can also provide substantially improved throughput for random access workloads:

PAWs Workload	RVA-2 9393-002	RVA-2 Turbo 9393-T42	Delta
Read Hit	1837	2738	+49%
Read Miss	479	719	+50%

Performance Assessment Workloads:
- 32 active volumes
- 4K block size - 100% Read Hit
- 12K block size - 100% Read Miss

Measured Configurations:	9393-002	9393-T42
- ESCON Channels	8	8
- Effective Cache	2G	2G
- Effective NVS	16M	16M
- Effective DASD Storage	290G	420G

RVA-2 Turbo vs RVA-2 Performance ...

„ **Maximum Random Access Throughput (cont'd)**

The following tests include all four database benchmarks from the ...

Performance Assessment Workloads (PAWs)

Cache Locality

They cover a range of cache locality, from cache friendly (typical read hit ratio 90 percent) to cache hostile (typical read hit ratio 40 percent).
The hit ratio for a benchmark run also depends upon cache size.

Volume Skew

One of the 4 workloads (the Cache Uniform workload) loads all of the active volumes equally; the other 3 are designed with realistically high levels of skew across the active volumes.

PAWs Workload	RVA-2 9393-002	RVA-2 Turbo 9393-T42	Delta
Cache Uniform	1184 IO/sec	1843 IO/sec	+55%
Cache Friendly	1450	1656	+14%*
Cache Standard	997	1462	+47%
Cache Hostile	619	864	+40%

Performance Assessment Workloads:
- 4K Block Size
- 48 Active Volumes

* The I/O activity for the Cache Friendly workload did not saturate the 9393-T42 subsystem. The 14% higher throughput was obtained with a significantly lower response time

Measured Configurations:	9393-002	9393-T42
- ESCON Channels	8	8
- Effective Cache	2G	2G
- Effective NVS	16M	16M
- Effective DASD Storage	290G	420G

RVA-2 Turbo vs RVA-2 Performance ...

„ **Minimum Random Access Service Time (msec)**

The following table indicates the minimum observed service times based on measuring a range of load levels:

PAWs Workload	RVA-2 9393-002	RVA-2 Turbo 9393-T42	Percent Improvement
Cache Uniform	5.9 msec	5.0 msec	-15%
Cache Friendly	3.1	2.6	-16%
Cache Standard	5.8	4.8	-17%
Cache Hostile	11.2	9.4	-16%

Performance Assessment Workloads:
- 4K Block Size
- 48 Active Volumes

Measured Configurations:	9393-002	9393-T42
- ESCON Channels	8	8
- Effective Cache	2G	2G
- Effective NVS	16M	16M
- Effective DASD Storage	290G	420G

Service Time = Connect + Disconnect + Pend

„ **More Information**

For more RVA performance information refer to

- the presentation 'RAMAC Virtual Array 2 -Enhancements Overview-' via the IBM INTRAnet Large Systems Storage home page http://w3.ssd.ibm.com/ramac
- RVA 2 Turbo 4-Path and Turbo 8-Path Performance by Bruce McNutt, IBM SSD, 04/97. Available to your IBM representative
- IBM Disk Storage Systems Performance Update, 09/97 RVA-2 Turbo 8-path, by Chris Saul, PERFUPD on MKTTOOLS. Available to your IBM representative

RVA Enhancements 05/98

RVA Enhancements 05/98

- **Announcement Contents**
  - **Maximum (effective) cache size of 4 GB**
    Of specific benefit for heavy data base workloads
  - **Improved u-code LIC 04.04.xx**
    Shorter pathlength of key functions,
    most apparent for loads with high READ hit ratios
- **Performance**

Read Hit Performance (100% READ hits)

Performance Metric	LIC 4.3	LIC 4.4	Delta
I/O rate at 2.2 msec RT	2334	2883	+24%
Max I/O rate	3081/sec	3785/sec	+23%

Performance assessment 100% read hit workload:
- 4K block size

Measured configurations:	LIC 4.3	LIC 4.4
- ESCON Channels	8	8
- Storage Capacity	420G	420G

Same cache size

Typical Database Performance

READ hits and READ/WRITE ratio typical of online data base loads.
Combined impact of shorter path lengths, as well as larger cache.

Performance Metric	LIC 4.3 3 GB	LIC 4.4 4 GB	Delta
I/O rate at 10 msec RT	1190/sec	1390/sec	+13%
RT at 1600 IO/sec	13.4 msec	11.4 msec	-15%

Performance assessment cache standard workload:
- 4K block size
- Read hit ratio at high load: 78%
- Read/write ratio: 3:1

Measured configurations: see above

RVA Enhancements

RVA Model T82 Enhancements (11/98)

- Announcement 98-11-03

- **PPRC (Feature 7001)**
- **Up to 840 GB effective capacity**
  Increments are 80, 130 and 210 GB

RVA Model T82 Enhancements: X82 (05/99)

- Announced 99-05-04

- **Up to 1024 logical devices (addresses)**
  Former limit for T82 and other models was/is 256
- **Emulation of 3390-9 volumes**
  Allows easier migration for these 8.5 GB volumes.
  Each 3390-9 reduces #UCBs by 3.
- **For performance reasons, avoid huge log. volumes**
  Higher wait time (IOSQ) in the operating system may occur,
  if device I/O rate and thus logical device contention is high.
  No performance disadvantage within RVA itself.
- **Faster subsystem controller**
  Includes faster microprocessors
- **Effective cache size of up to 6 GB**
  From 2G, in 512 MB increments.
  Of benefit for certain workloads (e.g. heavy data base).

RVA Enhancements ...

RVA Model X82 Performance

- **Max. Sequential Throughput (MB/sec)**

PAWs Workload	RVA-2 Turbo 9393-T82	RVA-2 Turbo 9393-X82	Delta
VSAM READs			
Single Stream	3.7	5.3	+43%
8 Streams	18.5	31.5	+70%
16 Streams	22.6	35.2	+56%

Access Method:	VSAM	
- Block Size	4K	
- Blocks transferred per I/O	12	
Measured Configurations:	9393-T82	9393-X82
- ESCON Channels	8	8
- Effective DASD Storage	420G	420G

- **Max. Random Access Throughput (IO/sec)**

PAWs Workload	RVA-2 Turbo 9393-T82	RVA-2 Turbo 9393-X82	Delta
Read Hit	3670	4890	+33%
Read Miss	873	1159	+33%
Cache Standard	1687	2204	+31%

Performance Assessment Workloads:
- 64 active volumes (96 for Cache Std)
- 4K block size - 100% Read Hit
- 12K block size - 100% Read Miss

Measured Configurations:	9393-T82	9393-X82
- ESCON Channels	8	8
- Effective Cache	4G	6G
- Effective NVS	16M	16M
- Effective DASD Storage	420G	420G

RVA Enhancements ...

RVA Model X82 Enhancements: X83 (07/99)

- Announced 99-07-27
- Here only deltas to Model X82 are listed

- **Use of 9G SSA HDDs**
  Vs 4.5G SCSI drives on all previous RVA models
- **Up to 1.68 TB (effective) capacity**
  Vs 726 GB/840 GB before, starts now at 290 GB.
- **Improved performance due to SSA HDDs**
  Refer to tables on next foil

RVA Enhancements ...

RVA Model X83 Performance

Here also compared to T82

„ **Max. Sequential Throughput (MB/sec)**

PAWs Workload	RVA-2 Turbo 9393-T82	RVA-2 Turbo 9393-X83	Delta
VSAM READs			
Single Stream	3.7	6.0	+62%
8 Streams	18.5	41.8	+126%
16 Streams	22.6	43.4	+92%

```
Access Method:                    VSAM
- Block Size                        4K
- Blocks transferred per I/O        12

Measured Configurations:       9393-T82  9393-X83
- ESCON Channels                   8         8
- Effective DASD Storage         420G      840G
```

„ **Max. Random Access Throughput (IO/sec)**

PAWs Workload	RVA-2 Turbo 9393-T82	RVA-2 Turbo 9393-X83	Delta
Read Hit	3670	5154	+40%
Read Miss	873	1336	+53%
Cache Standard	1687	2886	+71%

```
Performance Assessment Workloads:
- 64 active volumes (96 for Cache Std)
-  4K block size - 100% Read Hit
- 12K block size - 100% Read Miss

Measured Configurations:      9393-T82   9393-X83
- ESCON Channels                 8          8
- Effective Cache                4G         6G
- Effective NVS                 16M        16M
- Effective DASD Storage       420G       840G
```


Log-Structured File System Aspects

General Log-Structured File Aspects

More details are contained e.g. in

'The Design and Implementation of a Log-Structured File System'
by Mendel Rosenblum and John K. Ousterhout,
Univ. of California, Berkley.
ACM Transactions of Computing Systems Vol.10 #1, 02/92, pp26-52

Ù **Background**

„ **With larger caches, disk traffic will become dominated by WRITEs**

Ù **Benefits**

„ **Eliminates most SEEKs by writing changed data**
- **into a new location**
- **in a clustered manner**

„ **Eliminates RAID-5 write penalty**

Each updated track is written to a new location, together with the newly calculated parity

„ **Spreading data of a volume across all backing storage**

Further reduces the 'hot spot' effects (HDD load)

Ù **Challenges**

„ **Ensure that larger extents of freespace is always available**

„ **Do freespace collection and mgmnt effectively, w/o too much impact to update activity**

RVA Space Management

RVA Space Management

„ **Data on HDDs is treated in units of 'functional tracks'**

„ **Any update of a functional track is to a new location**
- All WRITEs are 'physical sequential'
- All READs are 'physically random'

„ **Types of physical disk space and transitions**

```
                     -------------------------
                    |    USED Space (NCL)     |
                    |                         |
                    |  Contains compressed    |
                    |  'functional tracks'    |
                     -------------------------
                     |   |                   A
              DDSR   |   | Rewrite a         |  Rewrite a track
                     |   |    track          |
                     V   V                   |
   --------------------------              ----------------------
  | UNCOLLECTED FREE Space    |            | COLLECTED FREE Space
|
  |                           | Freespace  |       (CFS)
|
  |Scattered free phys. tracks|----------->| Compact set of free
|
  |not yet collected          | Collection | phys. array cylinders
|
   --------------------------              ----------------------
```

> Avoid NCL>85%. Increased priority of CFS task may impact performance

Type of logical DASD data	Occupies phys. RVA disk space
VTOC occupied,unexpired Used tracks	Yes *1
" " Unused tracks	No
VTOC occupied, expired Used tracks	No, after DDSR
" " Unused tracks	No
VTOC free - -	No

*1 Preformatted disk space (e.g. initialized with 0's) compresses well and thus only occupies a small amount of bytes per track
- DDSR stands for the Deleted Data Space Release function
- 'Unused' tracks are tracks w/o any data on it

RVA Space Management

Perf. Benefits of LSA vs other RAMACs

Perf. Benefits of LSA vs other RAMACs

- **Virtual capacity can significantly exceed installed physical capacity**
- **Compression of data also saves phys. space**

  Besides less byte transferred on the 'lower interface'...
  Allocation of physical space for functional tracks is done in variable number of sectors
- **Compression of data saves phys. cache size and gives faster transfer of tracks between cache and HDDs**
- **De-Staging of changed tracks to phys. disk can be done very effectively**

  Homeless tracks allow clustering on phys. disk
- **Hot spots on log. volumes are spread across all HDDs of the I/O subsystem**

  Not only across 1 RAID-5 array
  - **As long as I/O rate per log. volume is not too high ...
    dataset placement on disks is uncritical**
- **Fast duplication of data possible w/o movement of data: 'SnapShot'**

RVA IXFP Program

IXFP program (OS/390, MVS, VM)

IXFP originally did not run under VSE, just under VM/ESA with IXFP/VM, or under OS/390.
Refer to the 09/98 announcement of 'IXFP/SnapShot for VSE/ESA'.

- **IXFP, beneficial for**

  Dynamic configuration

  Easier definition of volumes than from RVA operator panel

  Media acceptance test ('CE work')

  Test of phys. devices (e.g. at extensions)
  w/o tying up host/channel/controller resources
- **IXFP, required for**

  DDSR (Deleted Data Space Release) function: Reclamation of tracks from deleted files

  i.e. for obsolete data on tracks outside a valid VTOC extent
  - **All unallocated and unused space occupies no physical space, and is available for any of the 256/1024 'functional volumes'**

  DDSR is NOT required for updated tracks.

  It is recommended to use Dynamic DDSR all the time,
  plus Interval DDSR occasionally (e.g. once per week).
  Interval DDSR alone shows often STORED > ALLOC values.

  If DDSR would not be used or available, the capacity of an RVA is being lowered by such invalid files or DASD extents.
  This may, depending on installation, be from say 10% to 20%.

  Data collection and reporting

  Collect and log subsystem data similar to SMF/RMF

  SnapShot copy program

  Refer to separate foils

IXFP DDSR for VM/VSE

IXFP DDSR for VM used for VM/VSE

- **Deleting a VSE file does not free space**

  Deleting a VSE file on a VSE minidisk or a DEDicated device does NOT free the occupied space in the RVA, but these tracks/extents -naturally- tend to be reused later.
  - **Results in 10% to 20% higher occupied space (Net Capacity Load, NCL) for VSE**

    compared to the case where DDSR would be done.

  Only in case an ENTIRE minidisk would be deleted, the space would no more be occupied:

```
      IXFP/VM MINIDISK command
 or
      SIBVMRVA DDSRkill device-cuu

      (The SIBVMRVA utility enables a VM user to access
       RVA functions from a CMS REXX EXEC)
```

- **Limiting disadvantages**

  When DDSR for VSE is not available ...

  (i.e. IXFP/SnapShot for VSE/ESA is not installed)

- **You may put 'workfiles' on extra minidisks and 'VM-delete' them if capacity required**
- **Reuse VSE workspace extents as soon as possible**
- **Care for 10% to 20% higher usable capacity for the VSE owned share of the RVA**

The last 2 items also apply to VSE/ESA native.

RVA IXFP/SnapShot

SnapShot 'Instant Copy'

Requires IXFP 2.1

- **Duplicates data rapidly:**

  'Copy the pointers, not the data'.

  Create multiple independent views of the data, seen by S/W.
  Create only a physical copy when original/copied track is updated
  - **True 'point-in-time' copy**

  To SNAP a volume needs few seconds, e.g. 6 sec for a bigger volume
- **Dramatically reduces Backup times**

  No data is moved during snap
- **Eases creation of test data**
- **Operates at volume and data set levels**

  Be aware that copying a volume with e.g. VSAM files/catalogs needs specific attention if to be accessed by VSAM
- **Supported by MVS/ESA (OS/390)**
- **VM support announced 02/97**

  Refer to
  - the next foils and/or
  - the presentation 'IBM SnapShot Overview', available via the IBM INTRAnet Large Systems Storage home page http://w3.ssd.ibm.com/ramac
- **Usage Note**

  Before using SnapShot for a volume, think of potential consequences if any active user still would try to use a volume (and that VSE VSAM and VM MDC buffers must be flushed to real disk before).
  The same aspect applies to DDSR, just as for REAL disks.

IXFP/SnapShot for VM/VSE

IXFP/SnapShot for VM/VSE

IBM RAMAC SnapShot for VM/ESA V1 R1 (02/97).

For more info, refer to

- 'IBM RAMAC SnapShot for VM/ESA, Installing and Using SnapShot' Version 1 Release 1, SC26-7217-00 (02/97)
- 'The RVA and IXFP for VM/ESA', Presentation by Jack Flynn, IBM SSD, VM/VSE Tech. Conf. Kansas City, 05/97
- Implementing SnapShot, SG24-2241-00, ITSO San Jose Redbook, 11/97, 185 pages

„ **Snapping can be done**

- in a REXX program
- using the CMS command-line

„ **Principal Snap Considerations**

Data Duplication ('only')

- for creation of test data
- for 'data mining' purposes

Data might be usable, even if SnapShot ends with RC=4 ('fuzzy snap', since I/Os occurred during snap)

Backup

- VSAM / non-VSAM

Data consistency and recovery are vital

Simple data consistency mostly(?) seems to be achievable by avoiding I/Os during the few snap seconds, but in order to continue after a restore, a defined log entry point would have to be available.

í **SnapShot for VM (for VSE guest exploitation) is 'not easy' and limited**

IXFP/SnapShot for VM/VSE ...

IXFP/SnapShot for VM/VSE (cont'd)

„ **SnapShot of any VM minidisk**

'Snap' data from source to target minidisk linked/attached (independent of owner):

```
SNAP MINIDISK (SOUrce DEVice(cuu-s) TaRGet(DEVice(cuu-t))
```

„ **SnapShot of any volume known to VM**

```
  SNAP VOLUME (SOUrce DEVice(cuu-s)  TaRGet(DEVice(cuu-t))
or
  SNAP VOLUME (SOUrce VOLUME(volser) TaRGet(VOLUME(volser))
```

Device must be online to the VM attempting the snap operation (needs CP ATTACH and DEFINE MDISK):

Uses REAL device addresses, NOT intended for volumes attached to a user

It is not possible to SNAP VOLUME a DEDicated device for a VM guest (e.g. VSE) which currently is up.

The only way would be to IPL CMS under the same VSE guest ID, do a SNAP VOLUME and then re-IPL VSE.

Also, SnapShot VM does NOT allow to change the VOLID of the SNAPped volume (as does SnapShot for MVS).

„ **Also any contiguous subset of a minidisk or volume may be SNAPped:**

```
 .... FROM (cyl) FOR (ncyls)   additional SNAP parameters
```


IXFP/SnapShot for VM/VSE ...

IXFP/SnapShot for VM/VSE (cont'd)

„ **'Instant Format' uses an already 'SNAPped' empty minidisk (= pre-formatted)**

„ **Target 'functional volumes'**

- must be to already defined devices
- must be of same device type and model (geometry)

If a fast snap is not possible (since target volume is on dissimilar device types or not within the same RVA), a 'data mover' is called, equivalent to DDR

„ **Impact of SNAPped disks on back-end storage usage**

If NCL becomes a concern, you must consider
- the length of time SNAPped volumes are kept
- the amount of meanwhile updated data in source and/or target volume

„ **VOLID consideration**

SNAPping a volume under MVS allows to keep the VOLIDs, if COPYVOLID(YES) is used.

In any case 2 identical VOLIDs under the same VSE or operating system has known problems, if both are online.

IXFP/SnapShot for VSE/ESA

IXFP/SnapShot for VSE/ESA

Combines the most important IXFP and SnapShot functions in 1 product.

Available as a priced optional feature of VSE Central Functions.

Requirements:
- VSE/ESA 2.3.0 (incl. PTF for APAR DY44820) or higher or VSE/ESA 2.1/2.2 (incl. PTF for APAR DY44841, 06/99)
- RVA LIC level 03.00.00 or higher
- SnapShot feature 6001 of the RVA.

Performance Functions provided

„ **DDSR (Deleted Data Space Release)**

DASD space with data obsoleted via VTOC ('expired files') is freed, except files secured via DSF parameter.
In VSE, this is done upon operator request,
it is NOT 'Dynamic DDSR', which immediately frees storage.

```
 IXFP DDSR       Releases/Deletes all expired data on an RVA

 IXFP DDSR,cuu   Deletes a total volume (!) (only if device DOWN)
...(DSN=ds-name) or only a BAM data set
```

Do NOT specify the NOPROMPT option, unless you really are sure what you do.

„ **SnapShot 'Instant Copy' or 'Snap'**

Fast 'data duplication/replication' for

- logical volumes (incl. VM PPMs)
- cylinder ranges
- on a file basis (non-VSAM)

to an existing (ADDed) target volume

```
 IXFP SNAP,source : target  Snaps a total volume
                            to a target volume (must be DOWN,
                            can get another VOL1 label)
```

ESS FlashCopy is described in the VSE/ESA 2.5 document

IXFP/SnapShot for VSE/ESA ...

Performance Functions provided (cont'd)

„ **Display of RVA space utilization(s)**

Defined/allocated/physic. occupied space, capacity, compact. ratio

```
 IXFP REPORT      gives - Device Detail     Report
                        - Device Summary    Report
                        - Subsystem Summary Report
```

Space per logical device:

```
   All space here is in terms of 'functional space'.

   -- DEFINED -- (via volume def)
-----------------------------------
  | -- ALLOCATED --(in VTOC)-------------------------------
|
  || -- STORED --(needing phys.sp.)- -- UNUSED --(not ----- |
|
  |||                              |               mapped)||
|
  |||                              |                      ||
|
  |||                              |                      ||
|
  || ------------------------------ --------------------- |
|
  | -------------------------------------------------------
|

------------------------------------------------------------------

                                      STORED MBs
    Compaction/Compression Ratio =  ---------------
                                     PHYS.USED MBs
```

Space per subsystem:

```
   DEFINED              Sum of all volumes defined (functional MBs)

   DISK-ARRAY CAP.       Total RVA Capacity    (phys. MBs)

   FREE DISK-ARRAY CAP.  Free RVA phys. space  (phys. MBs)

                         FREE-DISK-ARRAY-CAP.
          NCL  =    1 -  --------------------
                            DISK-ARRAY-CAP.
```

• For official description, refer to

IXFP/SnapShot for VSE/ESA ...

- Memo to Current Licensees GI10-0487-00
- IXFP/SnapShot for VSE/ESA LPS, GC33-6630
- The description in the Internet (via VSE/ESA home page)
- 'RAMAC Virtual Array, PPRC and IXFP/SnapShot for VSE/ESA', IBM Redbook, SG24-5360-00 (01/99)

IXFP/SnapShot for VSE/ESA ...

Performance Benefits

IXFP functions provided via phase $IJBIXFP (<50K in SVA-24)

„ **DDSR gives lower Net Capacity Load (NCL),**
typically 10% to 20%

„ **SnapShot allows**

- continuation of Online work,
during backup to tape (from snapped data)

- easy creation of test data

- reduction of required batch window

„ **Copying is in seconds rather than minutes**

No actual movement of data is involved,
no host processor or channels tied up.

Snap time also depends on utilization of RVA subsystem.
Several snaps may be active concurrently

Benefits of IXFP/SnapShot VSE for VM/VSE users

VM/VSE users with IXFP/VM:

„ **Freeing of DASD space (DDSR)**
occupied by obsolete data within a total VSE volume

VM/VSE users with SnapShot/VM (includes IXFP/VM):

„ **Fast data duplication (SnapShot)**
- also for VSE partial volumes
- for total VSE volumes w/o VM DETACH

IXFP/SnapShot for VSE/ESA ...

DASD Space Considerations for SnapShot/VSE

How much DASD space do SnapShots require on top?

„ **Background info**

- DASD space on RVA is always allocated in units of tracks
- As soon as only 1 record is updated in an (original or snapped) track, a new track image is created (which is always written to a new position)
- Snapping data into a volume which already occupies DASD space for the same tracks does NOT increase requirements, even when original or snapped tracks are being updated

„ **Factors influencing addt'l DASD space**

- Number of logical cylinders/tracks snapped
- Physical DASD space already occupied for the target volume before snap
- Time to keep original and snapped data
- Resulting share of updated tracks by update activity to original and snapped tracks

Extreme cases:

- Original data are not updated: '0%'
No additional DASD space is required
(even when target volume has to be defined anew)

- All original data tracks are updated: '100%'
All snapped tracks cause that addt'l DASD space is required
(for the updated originals)

„ **Recommendations**

Implement SnapShot carefully, monitor NCL

Delete the snapped data via DDSR command, as soon as backed up to tape

RAMAC Array Family, REA and RSA

PART I.

RAMAC Array Family, REA and RSA

- **RAMAC Electronic Array Storage**
- **RAMAC Scalable Array Storage**

```
RAMAC Array DASD
RAMAC Array Subsystem
RAMAC Array Storage (RAMAC 3)

RAMAC Virtual Array Storage 2
                                  are discussed in previous parts
```


RAMAC Electr. Array Storage 'REA' (9397)

RAMAC Electronic Array Storage -Summary-

```
- Announced 09/96 (REA-1), 04/97 (REA-2)
- Supported by VSE/ESA 1.3/1.4 and up
  (no 'IXFP' required)
- Product evolved from STK 'Arctic Fox'
```

- **1/2/3/4 GB of electronic 'disk' storage**

  Battery backed nonvolatile cache storage, no hard drives involved, similar in principal to Solid State Devices from other vendors

- **Ultra high performance 'disk subsystem'**

  ESCON attachment is important, 48 hour of power makes copying to real disks unnecessary

- **Appears as 3380s or 3390s at 3990-2**

  Up to 256 logical volumes (REA-2: 512)

- **Designed for small data sets with permanent high WRITE activity**

  Especially when no chance for DIM exploitation is given.

  Examples are log files, VSE lock file (if native)

- **Upgradable to RSA (RAMAC Scalable Array)**
- **REA-2 (9379-A02, 04/97) with significantly improved performance vs REA-1 (9397-A01)**

  Enhancements to H/W and to LIC

  - **Ultra high performance for critical data sets**

RAMAC Electr. Array Storage 'REA' (9397) ...

RAMAC Electronic Array Storage ('REA')

- **Attached via parallel or ESCON channels**

  8, 12, or 16 channels

- **Mirrored cache**

  RAID-1 design allows non-disruptive cache card replacement and update ('hot pluggable components')

  1, 2, 3 or 4 GB of mirrored cache (i.e. up to 8 GB of physical cache)

  Non-mirrored cache available as RPQ

- **All I/Os are 'cache hits'**
- **Up to 16 concurrent host transfers (READ/WRITE)**
- **Dual 100 MB/sec internal busses, each with 4 channel directors**
- **Max. throughput scalable via #channels**

4 ESCON channels	up to 2000 IO/sec
16 ESCON channels	> 5000 IO/sec
16 ESCON channels REA-2	up to 10000 IO/sec

More Information

For more information, refer to

IBM RAMAC Electronic Array Storage, Introduction, GC26-7205-00

the presentation 'RAMAC Electronic Array 2 - Overview' available via the IBM INTRAnet Large Systems Storage home page http://w3.ssd.ibm.com/ramac

REA-2 Announcement Letter

RAMAC Scal. Array Storage 'RSA' (9396)

RAMAC Scalable Array Storage 'RSA' -Summary-

```
- Announced 09/96 (RSA-1), 04/97 (RSA-2), 02/98 (RSA-3)

- Product evolved from STK 'Kodiak'

- Supported by VSE/ESA 1.3/1.4 and up
  (no 'IXFP' required), 3380s must be ADDed as ECKD
```

- **Up to 1.4 TB (RAID-5) in a single frame up to 2.2 TB in total (RSA-3)**

	Total Capacity	#RAID-5 disk arrays	GB/array	HDDs / array	#DC pairs
RSA-1	278 ... 1394G *1	6 ... 30	46.5 G	5d + 1p	2..6
RSA-2	304 ... 1368G *1	4 ... 18	76 G	9d + 1p	2..6
RSA-3	629 ... 1258G *2	4 ... 8	157 G	9d + 1p	2..4
	... 2202G *3	+ ... +6	"	"	+2

```
 - 5d + 1p means 5 data +1 for rotating parity
*1 single frame for disks
*2 Model 300 Integrated Frame for CU + disks
*3 + Model 301 attached for disks
```

- **RAMAC Electronic Array as 'caching front end'**

  ```
  - Up to 16 Channel Adapters
  - Up to 512 logical ESCON paths (32 per ESCON port)
  - 1 to 4 GB (mirrored) cache
  - Two 100 MB/sec internal busses
  ```

  Refer to REA for details

- **Each device controller (DC) pair**

  ```
  - provides 4 SCSI-II F/W 20 MB/sec paths
    (Each 4 concurrent transfers, in total 24)

  - allows attachment of up to 5 arrays

  - provides 2x2M buffer memory (RAID processing)
  ```

RAMAC Scal. Array Storage 'RSA' (9396) ...

RSA -Summary- (cont'd)

„ **Further performance aspects**

Large Cache and fast 'lower interface'

- Large cache and high cache to disk bandwidth allow to absorb large bursts for WRITE data, w/o an emergency destage (as for small non-volatile caches or NVSs)

Fast disks

- 9.1 GB non-IBM disks used (RSA-1),
- IBM Ultrastar 2XP 9.1 GB disks (RSA-2)
- IBM Ultrastar 18XP 18.2 GB disks (RSA-3)

Highly concurrent data transfers

- Up to 16 (20, RSA-3) concurrent channel data transfers,
- up to 24 concurrent data transfers to RAID-5 arrays

„ **Simulated devices**

3380-J -K and 3990-1,-2,-3,-9 type of volumes.

Also 'FlexVolumes' with any number of cylinders (nx5, nx9)

RSA-1 Models

9396 Models (RSA-1)

„ **Control Unit (9396-001)**

- 1 to 3 disk cabinets per control unit

„ **Disk Cabinets (9396-002)**

- 1 or 2 independent RAID-5 domains
- up to 464 GB capacity
- up to 60 disk devices per cabinet

„ **Recommended configurations**

Disk cabinets	Quantity of Disk Arrays		Subsystem Capacity	Recommended Cache	ESCON
	278G	464G	(GB)	Size	Paths
1	1	-	278 GB	1 GB	8
	-	1	464	2	8
2	2	-	557	2	12
	1	1	743	3	12
3	3	-	836	3	12
	2	1	1022	4	16
	1	2	1208	4	16
	-	3	1394	4	16

„ **Incremental installation of additional ...**

ESCON paths, Cache size, Device controller, Disk arrays

RSA-2 Models

9396 Models (RSA-2)

„ **Control Unit (9396-200)**

4 3990 Storage Control images

„ **Storage Cabinets (9396-2XY)**

- 1 Storage Cabinet for 1 9396-200 control unit
- Y is # of included and required device controller pairs
- significant savings of space, power consumption vs RSA-1

9396 Model	Subsystem Capacity	Minimum Cache	ESCON Ports	'Performance'
242	304 GB	1 GB	8	High
252	380	1	8	"
263	456	1	8	"
273	532	1	8	"
284	608	2	12	"
294	684	2	12	"
2A5	760	3	12	"
2C6	912	3	12	"
2E6	1064	3	12	"
2G6	1216	4	16	"
2I6	1368	4	16	"
244	304	2	8	High+
255	380	3	8	"
266	456	4	8	"

High+ means higher hit ratios by larger cache

RSA-2 Performance

„ **Significantly improved performance vs RSA-1**

- Enhancements in H/W and LIC
- 512 vs 256 logical volumes
- Record-level caching, function is between RLC I and II (RDF bit settings enough for dynamic use)

Refer to the following results for OS/390, and/or to
- 'RAMAC Scalable Array 2 Performance', by Bruce McNutt, Available to your IBM representative (RSA2PERF PACKAGE on MKTTOOLS)
- the presentation 'RAMAC Scalable Array 2 - Overview' available via the IBM INTRAnet Large Systems Storage home page http://w3.ssd.ibm.com/ramac

RSA-2 vs RSA-1 Performance

„ **RSA Max. Sequential Throughput (MB/sec)**

PAWs Workload	RSA-1 9396-001	RSA-2 9396-200	Delta
QSAM Read	5.0	7.6	+52%
QSAM Write	6.0	7.2	+20%
VSAM Read	3.0	4.8	+60%
VSAM Write	2.4	3.5	+46%

Performance Assessment Workloads:
- 1 Sequential Transfer Operation ('Single Stream')

Access Method:	VSAM	QSAM
- Block Size	4KB	27KB
- Blocks transferred per I/O	12	5
Measured Configurations:	9396-001	9396-200
- Mirrored Cache	4GB	4GB

Valid for all 5 RSA-2 Measurement Tables shown:

Measured Configurations:	9396-001	9396-200
- ESCON Channels	16	16
- Device Controller Pairs	6	6
- Mirrored Cache	varies	

„ **Minimum Random Access Service Time (msec)**

The following table indicates the minimum observed service times based on measuring a range of load levels:

PAWs Workload	RSA-1 9396-200	RSA-2 9396-001	Percent Improvement
Cache Uniform	6.5 msec	4.0 msec	-38%
Cache Friendly	4.2	2.6	-38%
Cache Standard	6.4	3.9	-39%
Cache Hostile	9.6	7.1	-26%
100% Read Miss	25.2	22.1	-12%

12KB Block Size - 100% Read Miss Four Corners Load
4KB Block Size - Other Workloads

Measured Configurations:	9396-001	9396-200
- Mirrored Cache	4GB	4GB

Service Time = Connect + Disconnect + Pend

RSA-2 vs RSA-1 Performance ...

„ **RSA Max. Random Access Throughput (IO/sec)**

RSA-2 can also provide significantly improved performance for random access workloads with dramatically faster response times as well as higher throughput:

PAWs Workload	RSA-1 9396-001	RSA-2 9396-200	Delta
Read Hit	5700	8800	+54%
Read Miss	788	1256	+59%

Performance Assessment Workloads:
- 64 active volumes
- 4KB block size - 100% Read Hit
- 12KB block size - 100% Read Miss

Measured Configurations: 9396-001 9396-200
- Mirrored Cache 2GB 2GB

The following tests include all four database benchmarks from the PAWs Workloads, refer to RVA for description

PAWs Workload	RSA-1 9396-001	RSA-2 9396-200	Delta
Cache Uniform	3618 IO/sec	5103 IO/sec	+41%
Cache Friendly	3183	4249	+33%
Cache Standard*	2848	3062	+ 7%
Cache Hostile*	2155	2048	- 5%

Performance Assessment Workloads:
- 4KB Block Size
- 96 Active Volumes

* Licensed Internal Code used here did not include GA level enhancements which are anticipated to result in performance improvements for Cache Standard and Hostile workloads.

Measured Configurations: 9396-001 9396-200
- Mirrored Cache 4GB 4GB

RSA-2 vs RSA-1 Performance ...

„ **Sample Performance (IO/sec with msec/IO)**

PAWs Workload	RSA-1 9396-001	RSA-2 9396-200	Delta
Cache Uniform	1028 at 7.1 ms	3785 at 6.3 ms	+268%
Cache Friendly	850 at 4.5 ms	3386 at 4.5 ms	+298%
Cache Standard	425 at 6.4 ms	1702 at 5.6 ms	+300%
Cache Hostile	1494 at 21.7 ms	2048 at 19.9 ms	+37%

Performance Assessment Workloads:
4KB Block Size

Measured Configurations: 9396-001 9396-200
- Mirrored Cache 4GB 4GB

More Information

IBM RAMAC Scalable Array Storage Overview (Presentation Guide), 09/96, 32 pages. Available to your IBM representative (MKTTOOLS)

IBM RAMAC Scalable Array Storage System Architecture, 12/96, 9 pages. Available to your IBM representative (RSAARCH on MKTTOOLS)

IBM RAMAC Scalable Array Storage, Introduction, GC26-7212-00

IBM RAMAC Scalable Array Storage, Configuration and Performance, GC26-7210-00,
Update for RSA-2 available 2Q97

Performance White Paper (RSA-2 vs RSA-1)
Expected to be available on or before RSA-2 GA

'RVA 2 Turbo Spec Sheet', G2256675, as G2256675 on MKTTOOLS

IBM Disk Storage Systems Performance Update, 09/97
RSA-2 Enhancements, by Chris Saul, PERFUPD on MKTTOOLS.
Available to your IBM representative

RSA-3 Models

9396 Models (RSA-3)

GA was 98-05-29

„ **Integrated Frame Model 300 (9396-300)**

- up to 4 GB mirrored cache
- 4 DC pairs (+2 for Model 301)
- 1 or 2 disk arrays per DC pair
- 'sticky' cache as RPQ (for important volumes)
- up to 1258 GB in increments of 315 GB

„ **Model 301 Storage Frame (9396-301)**

- 2 additional DC pairs installed in Model 300
- adds up to 944 GB

í **Up to 2.2 TB total capacity**

„ **IBM Ultrastar 18XP disk drives for larger capacity and higher speed**

- 18.2 GB capacity
- 7200 RPMs, thus 4.17 msec avg latency
- 7.5 msec avg SEEK time
- Avg. Instantaneous Media data rate 18.3 MB/sec
- 2X larger device buffer

• Up to 1024 logical volumes

„ **Faster ESCON channel adapter**

„ **Significant performance improvements (vs RSA-2) for certain workloads**

Refer to next foil

RSA-3 vs RSA-2 Performance

„ **RSA Max. Sequential Throughput (MB/sec)**

PAWs Workload	RSA-2 9396-200	RSA-3 9396-30x	Delta
QSAM Read	7.6	9.8	+29%
QSAM Write	7.2	8.7	+21%
VSAM Read	4.8	6.7	+40%
VSAM Write	3.5	5.0	+43%

Performance Assessment Workloads:
- 1 Sequential Transfer Operation ('Single Stream')

Access Method:	VSAM	QSAM
- Block Size	4KB	27KB
- Blocks transferred per I/O	12	5

Valid for the RSA-3 Measurement Tables shown:

Measured Configurations:	9396-30x	9396-200
- ESCON Channels	16	16
- Device Controller Pairs	6	6
- Mirrored Cache	4 GB	4 GB

„ **READ Miss Performance**

	RSA-2 9396-216	RSA-3 9396-30x	Delta
Service time at min load	22.1 msec	18.7 msec	-15%
Max I/O rate	1256/sec	1429/sec	+14%

12KB Block Size - 100% Read Miss Four Corners Load

Measured Configurations: 9396-200 9396-30x
- Disks 180 (9G) 80 (18G)

More Info:

- RAMAC Scalable Array 3 Overview (Presentation Guide), 02/98. Available to your IBM representative (RSA3PG98 on MKTTOOLS)

RAMAC Family Performance Comparison

RAMAC Family White Paper Conclusions

'An Overview and Comparison of RVA-2, RAMAC 3 and RSA Performance'
as RAMFAM package on MKTTOOLS disk, 96-11-04, 19 pages,
available to your IBM representative

- **On-line (random) performance**

  RAMAC 3 vs RVA-2 vs RSA-1

  - **All 3 products can provide short response times, when configured for performance**
  - **All 3 products can provide an inexpensive, high-performance replacement for 3390/3380 devices**
  - **Among the 3 products, RSA provides by far the highest throughput capability**
    Lower maximum throughput does not mean that IORTs are higher for lower I/O rates (say < 1000 IO/sec)

- **Batch (sequential) performance**

  Each of the 3 products offers important sequential advantages

  - **RSA has highest aggregate sequential data rates**
  - **RVA-2 has highest single-stream WRITE**
  - **RAMAC 3 has highest single-stream READ**

RAMAC Family Performance Comparison ...

Volume Mapping Comparisons

- **Spreading of a Logical Volume**
  Any S/W only sees 'logical' volumes, be it

  - traditional 'physical' volumes, e.g. 'real' 3380/3390s:

    -> Data of 1 volume on 1 real DASD (HDD/HDA).
    Higher probability that a single real DASD is permanently overloaded (skew/hot spots)

  - RAMAC I/O subsystems with 'simulated' volumes and RAID5/6:

    -> Data of 1 volume are spread/striped across multiple HDDs.
    Low probability that 1 HDD is overloaded

I/O Subsystem	# of HDDs (plus HDD capacity) for 1 logical device to spread data	RAID
RAMAC 1/2/3	4 HDDs (2.2/4.5/9 GB)	RAID-5
RVA-2	8 or 15 HDDs (4.5 GB)	RAID-6
RSA-1/-2/-3	6/9/9 HDDs (9/18/18 GB)	RAID-5
Int. Disk	1(2 for READ) HDDs (9 GB)	RAID-1

- All HDDs with device buffer (RPS miss avoidance, both for READ and WRITE, seq. pre-staging)
- For Internal Disk, refer to separate foils

More Info on RAMAC Array Family

'An Overview and Comparison of RVA-2, RAMAC 3 and RSA Performance'
As RAMFAM package on MKTTOOLS disk, 96-11-04, 19 pages, available to your IBM representative

IBM RAMAC Family Performance Positioning, by John Bacho.
As RFAMPERF on MKTTOOLS

Storage Systems Alternatives for VSE and VM, by Bill Worthington.
VM/VSE Tech Conf, Kansas City, 05/97, Session 10D
(includes DASD selection criteria)

Value of RAMAC Family I/O Subsystems

Value of RAMAC Family I/O Subsystems

- **Performance**
- **RAID 5/6 Data Protection/Availability**
  Discussed in part G
- **Reconfiguration etc. during system up**
  Via RAID and/or duplication of H/W components
- **Automatic Load Balancing across phys. HDDs**
  Applies to log. volumes in same RAID array/drawer,
  for RVA even across the total I/O subsystem.

  Reduces/avoids disadvantages of hot spots, saves manual file placement for balancing (if possible at all)
- **Usage of Logical Volumes of any Type and Size**
  3380 and 3390 track/cylinder geometry,
  plus single/double/triple capacity volumes,

  RVA even with any number of cylinders,
  - for optimal performance of small/specific files
  - for potentially improved msec/IO times (less IOSQ)

Benefits for RVA only:

- **Savings of Physical GBs for Space not Occupied**
- **All Freespace is Common to All Logical Volumes**
- **Easy and Fast Data Duplication and Backup**
  Freedom
  - to determine instant of backup
  - when to copy backup data to tape cartridges
  - to copy multiple volumes on a single cartridge (provided the utility allows that)

Multiprise Internal Disk

PART J.

Multiprise Internal Disk

Multiprise 2000 ID
- Summary (Original & Enhanced ID)
- Performance Results (Original & Enhanced)
- ID Storage Hints
- Performance Hints
- IOCP Definitions
- Further References

Multiprise 3000 ID
- Summary
- Enhancements vs MP 2000 ID
- Performance

Multiprise 2000 Internal Disk

Multiprise 2000 Internal Disk -Summary-

A feature of the S/390 Multiprise 2000 processors, first announced 09/96, first available 01/97

- **S/390 I/O data are cached in real memory**

  32 MB up to 1 GB cache can be flexibly configured in 32 MB increments, 2 GB announced 05/98

  No WRITE caching first, but Internal Disk Fast Write since 09/97

- **From 18 up to 288 GB (576 GB 05/98) of user data, using DASD mirroring (RAID-1)**

  Increments are (mostly) in 4 HDDs (includes the mirrors).

- **Appears as 3380-K (+J,E) or 3390-1/2/3/9 volumes, ESCON attached to 3990-2**
  which on top accepts cache query commands

  - **Devices must be ADDed in VSE as ECKD**
    If not ADDed as ECKD, unrecoverable I/O errors may result (no VPD settings possible as for RAMAC Array Subsystem).

    Also, an additional WRITE performance degradation would occur

  - **VSE/ESA 1.2 is required at least**
    but release no more in service

- **Uses fast IBM Ultrastar HDDs, attached via internal SCSI-2 FW device adapters**

  - 9G 2XP HDDs originally, 9G 9LP and 18G 18XP since 05/98
  - Up to 8 logical S/390 volumes per HDD
  - 512 KB HDD cache (buffer) size

  - **No separate external control unit req'd**

Multiprise 2000 Internal Disk ...

Internal Disk -Summary- (cont'd)

- **The total DASD capacity determines the #HDDs and also their position in a drawer or cage**

- **Hot pluggable devices, with automatic resync after device replacement**

- **Int. Disk I/O processing on extra processor (SAP)**

  S/390 volume emulation is done on 'System Assist Processor'

  - **No impact on processor speed**

  Impact by moving of data within real memory is very small.

  Additional SAPs can be defined, using a spare S/390 u-processor. Refer to the WSC flash 9646

- **Internal capacity of 1 SAP is a high IO/sec figure**

  SAP utilization is shown on the System Activity Display (SAD) of the Multiprise 2000 H/W console

- **/370-mode guests (under VM/ESA or in /370 LPAR) originally not supported**
  (Restriction removed by 07/97 enhancements)

- **Originally targetted for non-shared data**

  Sharing of data between LPARs introduced 07/97

- **Bigger degradation for small Format WRITEs**

  Refer to separate bullet

Multiprise 2000 Internal Disk ...

Multiprise 2000 Internal Disk (cont'd)

- **Sequential Pre-staging**
  Up to 3 tracks are pre-staged, if in READ channel programs the sequential indication is set in the DX CCW

- **Unit of staging**

Staging	Used
Rest-of-track	in general (>90%)
Total track	if track format not 'predictable'
Record(s) only	if Record Caching indicated in channel program

- 'Predictable' track format allows direct calculation of the sectors on the HDD containing a desired record. For ID, only tracks are predictable with fixed size phys. records and full tracks

- **READs are done from the least busy HDD**

  Any data resides on 2 HDDs (RAID-1).

  This helps in case of higher READ-IO/sec rate if e.g. 2 busy logical volumes reside on the same HDD

- **256 logical devices in total**
  8 log. vols/HDD x 16 HDD/drawer x 2 drawers

- **S/W upgrades for Internal Disk**

  ICKDSF 16
  APAR PN86705 (PTF UN97485, UN97483 (SA))
  EREP 3.5
  APAR DY44343 (PTF UD50246)
  IOCP 1.5
  APAR DY44132 (PTF UD50041/UD50048 for VSE/ESA V1/V2)

Internal Disk Enhancements (06/97)

Internal Disk Enhancements (06/97)

Announced 97-06-09, available 97-08-31 with EC E26479 plus MCL 06

- **Internal Disk Fast Write (IDFW)**

  Done in the 512K device buffer of each HDD

  - The Internal Battery Feature (IBF) must be installed and fully operational (no IDFW for the 1-drawer -100 entry configuration).
  - IDFW must be enabled on the ID Control Unit Customization Panel for the CE, in case u-code level '92W'. (The CE can disable IDFW entirely on the system level, IDFW cannot be controlled by S/W)

  - **To be activated by CE, if u-code level <'98G'**
    Driver 98G is EC-level E26572 with MCL 06, shown via the Service Element (PC).

  IDFW is a limited DFW implementation:

  - 'Single thread' per HDD for random I/Os
  - Only up to 2 concurrent back-to-back sequential I/Os (each up to 1 logical track) to the same HDD can obtain fast DFW hits (early device end)

  -> 512K per HDD is ample
  -> IDFW not so effective e.g. for SQL/DS(DB2) Checkpointing e.g. for frequent small format WRITEs

- **Sharing of logical volumes across LPARs**
  LPARs must be within same processor

- **Simulation of 3380 models J, E, besides K**

- **S/370-mode allowed for VSE/ESA 1.3/1.4**

  - as a guest under VM/ESA
  - in a S/370 LPAR

Internal Disk HDD Configurations

Multiprise 2000 ID Drawers (-100 models)

```
- 1 Drawer (6/6/4 HDDs per SCSI bus, 3 SCSI buses)
                                                         Log.
  +-------------------------+                 GB   CHPIDs  CUs Drawers
  |  0   1   2   3   4   5  |                --------------------------
  |                         |                 18      2     1     1
  | --- --- --- --- --- --- | <-SCSI bus      36      4     2     2
  |                         |   (or CHPID)  54-144    6     3     2
  | --- --- --- --- --- --- | <-SCSI bus     162      8     4     4
  |                         |                180     10     5     4
  | --- --- --- ---         | <-SCSI bus   198-288   12     6     4
  +-------------------------+               (9G HDDs)

- 4, 8, 12, or 16 HDDs per drawer (also 2 if 'overflow')
- 1, 1x2, or 2x2 Device Drawers (and Adapters) total

- Total number of paths/logical CUs (includes mirrors) 2-12
```

Multiprise 2000 ID Cage (-200 models)

```
- 1 Cage (8 HDDs per SCSI bus, 4 SCSI buses)

   +---------------------------------+      GB  CHPIDs Log.CUs Cages
   |  0   0   1   1   2   2   3   3  |    --------------------------
   |                                 |     18-36    2      1      1
   | --- --- --- --- --- --- -8- -F- |     54-144   4      2      1
   |  |   |   |   |   |   |   |   |  |    162-180   6      3      2
   | -.- -.- -.- -.- -.- -.- -9- -E- |    198-288   8      4      2
   |  |   |   |   |   |   |   |   |  |   (9G HDDs)
   | -.- -.- -.- -.- -.- -.- -A- -D- |
   |  |   |   |   |   |   |   |   |  |
   | -.-.-.- -.-.-.- -.-.-.- -B-.-C- |
   |  |___|   |___|   |___|   |___|  |
   |                                 |
   +---------------------------------+
      SCSI1   SCSI2   SCSI3   SCSI4

Unit  00  08  10  18  20  28  30  38
addr.-07 -0F -17 -1F -27 -2F -37 -3F

- 1 Base and 1 optional I/O Expansion Cage total
- Total number of paths (includes mirrors) is 2 to 8
```

- **Mirroring is across SCSI buses**

  - even across drawers (if >1 drawer)

- A logical 3990-2 CU consists of a pair of mirrored SCSI buses

Internal Disk Enhancements (05/98)

MP 2000 Internal Disk Enhancements (05/98)

Requires LIC Level A2 u-code.

Driver DA2I with EC-level E26599 and needs MCL fix level 027.

- **Increased maximum cache size:**

  Up to about 2G (1920M) in real memory

- **Use of new HDDs**

  9G Ultrastar 9LP and 18G Ultrastar 18XP,
  with slightly better Avg Read time and MB/sec values.

Capacity	# 9G HDDs	# 18G HDDs	# HDDs tot
18 GB	4	-	4
36 GB	8	-	8
54 GB	12	-	12
72 GB	8	4	12
90 GB	4	8	12
108 GB	-	12	12
126 GB	4	12	16
144 GB	-	16	16
180 GB	-	20	20
216 GB	-	24	24
...	-	...	...
576 GB	-	64	64

- Table applies to NEWBUILD orders
- Increments are 18G (<144G) and 36G (>144G)
- For orders >288G, 2 cages will be used

- **Doubled total disk capacity:**

  up to 576 GB user data

Original Internal Disk Performance

Original MP 2000 Internal Disk Performance

These results apply to the original ID implementation

- **For VSE, Internal Disk performance was better than a READ-cached 9345 configuration**

  6 msec/IO vs 8 msec/IO (PACEX, R/W =1.52 = write intensive)

Int. Disk Performance on 2003-116				
Case	#HDDs	msec/IO	IO/sec	SAP util.
PACEX4	5	6.0	675	33%
PACEX8	10	6.0	1077	53%

- Load was about balanced across the HDDs
- Original ID, without IDFW

- **Originally not suited at all for high WRITE content**

  No RAID-5 WRITE penalty, but RAID-1 needs WRITE also to the mirror-HDD, which is started concurrently

  With IDFW better WRITE performance, but NOT as fast as traditional DFW for sequential access

- **High degradation for small Format WRITEs**

  Formatting <1 track per channel program (any I/O attachment) costs lost revolutions by padding the rest-of-track after the last CCW of such a channel program.
  These 'padding zeroes' are being transferred via the SCSI busses and must be done to 2 HDDs.

  Unfortunately IDFW cannot help too much to improve sequential access

Enhanced Internal Disk Performance

Enhanced MP 2000 Internal Disk Perf. Results

- **RANDOM I/O-Intensive Jobs**

Deltas achieved by IDFW				
	#HDDs	msec/IO change	I/O rates (IO/sec)	Thruput increase
PACEX1	1	4.67->3.34 (-28%)	218-> 289	+32%
PACEX4	5	5.68->4.18 (-26%)	689-> 880	+27%
PACEX8	10	5.96->4.63 (-22%)	1135->1346	+18%

- PACEXn means n times 7 jobs in n partitions
- R/W ratio of PACEX is 1.52 (i.e. 39.7% Write content)
- VSE/ESA 2.2.1 on 2003-116, 256 MB ID cache (08/97)

  - **Significant I/O time benefits with IDFW**

- **SEQUENTIAL Write Jobs**

Deltas achieved by IDFW			
	Type of Write	msec/IO change	Thruput increase
LIBR DEFine Lib ...	Format	-40%	+56%
LIBR REStore Sublib...	Update	-32%	+35%
FCOPY RESTORE VOLUME...	Format	-50%	+48%

- A thruput increase of 50% would correspond to a reduction of 33% in Job Elapsed time

  - **Significant I/O time benefits with IDFW, also for SEQuential writes**

Individual benefits of other jobs may vary, depending on the achieveable IDFW hit ratios
(IDFW is not as effective for massive sequential WRITEs as other DFW implementations)

MP 2000 Internal Disk Storage Hints

Cache and Real Storage Hints

- **Recommended 'Cache to Backstore Ratio':**

  about 0.1% and more

  (ratio of cache size to net user data)

  Some I/O workloads may benefit up to about a 0.5% ratio

  Note:
  This basic rule assumes that the GB on DASD are active 'to an average degree' of about 1 IO/sec per installed DASD GB.

  - **Use 1 MB cache for 1 GB of data, or more**

- **Do not 'steal' too much real storage from VSE/ESA**

  Applying DIM is most effective and saves also CPU-time.
  But no I/O benefit can compensate increased VSE paging if you 'over-DIM' compared to the available real storage.

  You are on the safe side if VSE/ESA exploits DIM, with

  up to 6 MB real storage/'MIPS'

  (Whatever 'MIPS' is and how it is often misused, refer to 'MIPS' in Turbo Dispatcher document)

  - Add at least 8 MB base for VSE
  - Do not forget the VM part, especially when
    - MDC is used for the VSE guest or
    - CMS applications on top.

  Even when using VM MDC, use the processor storage as central and not as expanded storage.

Cont'd on next pages

MP 2000 Internal Disk Storage Hints ...

Configuration of Processor Storage

Example holds for VM/VSE or VSE native without LPARs.

```
  ---       -------------          ---  'P' MB
   A       |             |          A
   |       |             |          |
   |       |  Expanded   |      Expanded  (optional, VM only)
   |       /             /       Storage
   |       |  'E' MB     |          |
   |       |             |          V
   |       |-------------|  ---    ---  'C' MB
   |       | HSA base    |   A      A
   |       |  'H' MB     |   |      |
   |       |      -------|  HSA     |
   |       |             |   |      |
Processor  | ID-cache    |   |      |
 Storage   |  'I' MB     |   V      |
   |       |-------------|  ---  Central
   |       |             |   A   Storage  (up to 2 GB)
   |       |             |   |      |
   |       |  VM/VSE     |  Real    |
   |       |  or         | Storage  |
   |       |  VSE native |  seen    |
   |       |             |   |      |
   |       |  'R' MB     |   |      |
   V       |             |   V      V
  ---       -------------   ---    ---    0 MB

       P = C + E = R + I + H + E
```

Optimal use of processor storage is user dependent

1. **Add up estimates for minimum sizes**

   R_min = 8MB + 6MB per consumed VSE-MIPS (VSE)
   + 16MB (VM/VSE, w/o CMS apps)

   I_min = 0.1% or more of installed GB DASD
   DO NOT ACCEPT 32M MANUFACTURING DEFAULT

   H_min = given value, start with 10M, check later
   E-min = 0 (only for VM, not recommended here)

2. **Distribute the rest appropriately**

3. **Observe your system (paging, cache hit ratios)**

MP 2000 Internal Disk Storage Hints ...

Cache and Real Storage Hints (cont'd)

- **Examples for Cache Size Calculations**

Multiprise processor storage sizes		
Processor	Min	Max
2003-102 to -107	128M	1G
2003-115 to -125	256M	4G
2003-126 to -156, 1C5	512M	4G
2003-203 to -207	128M	2G
2003-215 to -225	256M	2G
2003-227 to -257, 2C5	512M	4G

Size of Hardware System Area (HSA)

The HSA contains Licenced Internal Code (LIC) and configuration dependent control blocks, not available for program use. Its size varies, 16 MB and more real storage is occupied, depending on configuration (not counting Internal Disk cache).

For smaller processor storage sizes this has to be considered.

Refer to 'Multiprise System Overview', GA22-7152-01, pp 3-12, and/or use the HSA estimation tool provided.

Sample Calculations

128 MB total storage may be sufficient only:

e.g. for 54 GB of Int. Disk user storage (64 MB cache), 10 MB HSA, leaving 54 MB real for S/390, appropriate for about 54/6 = 9 'MIPS', reasonably 'DIMed', i.e. about a 2003-104.

256 MB total storage may be sufficient only:

e.g. 90 GB Internal Disk (96 MB cache), 10 MB HSA leaving 150 MB real for S/390, appropriate for about 150/6 = 25 MIPS, reasonably 'DIMed', i.e. about a 2003-115

Internal Disk Storage and VM MDC

In case VSE runs under a single VM and no data are shared with systems outside that VM:

Assign addt'l central storage to ID or to VM MDC?

The answer depends on your specific situation. Please consider...

- **Both provide excellent READ caching**

  WRITE caching must be done by IDFW

- **VM MDC cache processing is done in VM CP,**
  ID cache processing is done on the separate SAP

- **VM MDC caching can be controlled on minidisk level**

- **ID caching would honor 'Record Caching'**

  Mostly of benefit for smaller caches, but only if accesses not partly sequential

Internal Disk -More Insight-

MP 2000 Internal Disk -More Insight-

- **What is specific for ID, vs other I/O subsystems?**

  The fact that the READ-cache is in central storage (and thus closer to the S/390 programs) means

  - **In case of a READ hit, response is very fast**

    The responses for READ hits are as fast as VM MDC READ hits (no 'S/390 channel' is involved, just moving of data within central storage)

  The fact that the WRITE-cache is in the buffer of the HDD (and 'outside' of a pure ID implementation) means

  - **Real WRITE hits only occur when the HDD buffer is available to accept the data immediately**

    Refer to a previous foil for more details on the IDFW implementation

  - **RAID-1 is more vulnerable to hot spots (data are not striped across multiple HDDs)**

Critical VSE Performance Areas for ID

Critical VSE Performance Areas for MP 2000 ID

The following VSE specific areas may need attention, when migrating from a cached I/O subsystem with DFW to MP 2000 Internal Disk.

- **Do not expect in general that WRITE performance is as good as with your 'old' DFW I/O subsystem**

  Refer to the IDFW description in this document

- **SQL/DS (DB2 for VM/VSE) databases and DL/I**

  By default, so far VSAM used Record Caching (RC) for the I/Os to such types of ESDS files.

  Depending on the access pattern and ID cache size, better performance may be obtained by NOT using RC, since ID performance is very sensitive in that area.
  Consider also that RC I/Os are not included in the statistic counters provided by ID.

  Use VSE/ESA 2.4 or apply APAR DY44796 to VSE/ESA 2.3 in order to NOT use RC by default for such VSAM files.

- **SQL/DS (DB2 for VM/VSE) checkpointing**

  SQL/DS under VSE requires massive WRITEs at checkpoints.
  This is a stress case for IDFW.

  Refer to the IDFW description in this document

- **Formatting of data file extents (e.g. SQL/DS coldlogs or BAM files)**

  As with other I/O subsystems, formatting tracks may be inefficient, if much less than 1 track is formatted per I/O (By architecture, after the last format-write in a channel program the whole rest-of-track has to be erased).

  With ID, this may even hurt more, since 2 copies of the tracks (RAID-1) have to be written.

  For BAM/SAM files, use bigger physical records (BLOCKED) to define the file, or for FB files, overwrite RECSIZE in DLBL

MP 2000 Internal Disk Perf. Hints

I/O Channel/Device Hints

- **Use enough/more S/390 logical volumes**

  - to reduce IOSQ time
  - to potentially improve I/O response times

  Use as many S/390 volumes as you used to use w/o Internal Disk. Maybe use more if these disks were big with high I/O rates.

  A 3390-9 has the highest probability of causing excessive queuing in the channel queue of the operating system, thus prefer smaller (logical) volumes (e.g. 3390-3, maybe 3390-1).

  This reduces IOSQ waits in S/390 channel queue (less -logical- device contention) and offers more I/O concurrency to the Internal Disk I/O Subsystem. Refer to the User's Guide.

- **Careful avoid excessive accumulation of S/390 I/O**

  to any single HDD

  Up to 8 logical volumes per HDD

  to any set of 4/6/8 HDDs which share 1 SCSI bus

  This is a side consideration, for extreme cases only

  - **Place concurrently active data sets on different volumes that reside on different HDDs**

  This can be done by proper selection of the device number (cuu), see next item.

  WRITE is more exposed than READ to that, since ...

  - READs can be done from the original or the mirrored disk

  (no automatic striping (load balancing) as with RAMAC RAID-5).

MP 2000 Internal Disk Perf. Hints ...

I/O Channel/Device Hints (cont'd)

- **Volume placement (mapping)**

  To place a S/390 logical device on the next HDD, you have to re-start with a device number increased by 8 (vs the first device on that HDD)

  E.g. x00-x03 for 4 devices on 1st HDD,
  x08 for the 1st device on the 2nd HDD

  (This also applies if <8 logical volumes are defined per HDD)

- **Volume placement within an HDD**

  If you want or need to squeeze out the most of ID in terms of performance ...
  Assign the most busy volumes within an HDD to the lowest CUUs.
  The reason behind is that data rates are higher at the outer cylinders of the HDDs, where volumes are assigned first.
  Usually, I/O rates vary, so this may be less practical.

- **ADD all VSE DASDs as ECKD**

  This applies for functional reasons to any disk type (3380/3390).
  For performance reasons, applications may benefit from caching bit settings in ECKD channel programs.

  Performance-wise this aspect is OK, if the device type displayed by VOLUME cuu shows 6E.

- **Make sure ECKD channel programs are used**

  Using non-convertible CKD channel programs will lead to significant I/O performance degradations, especially for WRITEs.

  Refer to the ECKD vs CKD part of this document

VM/VSE Hint

- **Apply the PTFs for APARs VM60844/VM61046**

  PTFs are required to retrieve VM cache statistics for the Internal Disk (VM/ESA 1.2.2, 2.1.0, 2.2.0): UM28314/UM28315/UM28316

Internal Disk and IOCP

Internal Disk Definition in IOCP

Description holds for MP 2000 ID.
For MP 3000 ID, a new channel type 'DSD' must be used.

„ **CHPIDs with new type of channel path (ISD)**

```
CHPID  PATH=(..),TYPE=ISD
           ...
```

Defines Integrated System Device channel paths (internal SCSI).
The path-IDs are prescribed from the H/W configuration.

An ISD channel path can only be assigned to 1 control unit
(i.e. no daisy chaining of control units possible)

„ **CNTLUNIT definitions**

```
CNTLUNIT CUNUMBR= ...,PATH=(..,..),UNITADD=((00,48)),UNIT=3990
```

The 2 ISD channel paths go to 1 (logical) 3990 control unit
(must be the original path and the path to the mirror-HDDs)

UNITADD always must start at 00,
with a range of 48, (it may be 32 for only the '3rd' CU for -100s)

„ **IODEVICE definitions**

```
IODEVICE ADDRESS=(cuu,48),CUNUMBR=...,UNITADD=00,UNIT=3380
                                                    or3390
```

UNITADD must be 00, if only 1 IODEVICE statement is given per
CNTLUNIT

STADET=Y is the default for ISD channel paths

Refer also to the IOCP example on next foil

Internal Disk - IOCP Example

```
*****************************************************************
*  IOCP for Multiprise 2000-100s  Int. Disk  (Max. Configuration)  *
*     CU definitions for max. configuration:                      *
*       - 2 + 2 Drawers (2 original + 2 mirrored)                 *
*       - 32 + 32 HDD's                                           *
*       - max. 256 logical disks, here 3390s                      *
*****************************************************************
*  1st drawer
      CHPID PATH=(3C),TYPE=ISD
      CHPID PATH=(3D),TYPE=ISD
      CHPID PATH=(3E),TYPE=ISD
*  2nd drawer (= mirrored disks of 1st drawer)
      CHPID PATH=(38),TYPE=ISD
      CHPID PATH=(39),TYPE=ISD
      CHPID PATH=(3A),TYPE=ISD
*  3rd drawer
      CHPID PATH=(10),TYPE=ISD
      CHPID PATH=(11),TYPE=ISD
      CHPID PATH=(12),TYPE=ISD
*  4th drawer (= mirrored disks of 3rd drawer)
      CHPID PATH=(0C),TYPE=ISD
      CHPID PATH=(0D),TYPE=ISD
      CHPID PATH=(0E),TYPE=ISD
*** CU_0 *********************************************************
*   (connects max. 6 HDDs = max. 48 logical disks)
 CNTLUNIT CUNUMBR=4C00,PATH=(3C,38),UNITADD=((00,48)),UNIT=3990
 IODEVICE ADDRESS=(200,48),CUNUMBR=4C00,UNITADD=00,UNIT=3390
*
*** CU_1 ******************************************************
*   (connects max. 6 HDDs = max. 48 logical disks)
 CNTLUNIT CUNUMBR=4D00,PATH=(3D,39),UNITADD=((00,48)),UNIT=3990
 IODEVICE ADDRESS=(240,48),CUNUMBR=4D00,UNITADD=00,UNIT=3390
*
*** CU_2 ******************************************************
*   (connects max. 4 HDDs = max. 32 logical disks)
 CNTLUNIT CUNUMBR=4E00,PATH=(3E,3A),UNITADD=((00,48)),UNIT=3990
 IODEVICE ADDRESS=(280,48),CUNUMBR=4E00,UNITADD=00,UNIT=3390
*
*** CU_3 ******************************************************
*   (connects max. 6 HDDs = max. 48 logical disks)
 CNTLUNIT CUNUMBR=7800,PATH=(10,0C),UNITADD=((00,48)),UNIT=3990
 IODEVICE ADDRESS=(300,48),CUNUMBR=7800,UNITADD=00,UNIT=3390
*
*** CU_4 ******************************************************
*   (connects max. 6 HDDs = max. 48 logical disks)
 CNTLUNIT CUNUMBR=7900,PATH=(11,0D),UNITADD=((00,48)),UNIT=3990
 IODEVICE ADDRESS=(340,48),CUNUMBR=7900,UNITADD=00,UNIT=3390
*
*** CU_5 ******************************************************
*   (connects max. 4 HDDs = max. 32 logical disks)
 CNTLUNIT CUNUMBR=7A00,PATH=(12,0E),UNITADD=((00,48)),UNIT=3990
 IODEVICE ADDRESS=(380,48),CUNUMBR=7A00,UNITADD=00,UNIT=3390
***************************************************************
```


Further MP 2000 ID References

Further MP 2000 ID References

For more info on the Internal Disk, refer to

IBM S/390 Multiprise 2000 Server -Internal Disk Performance-
White Paper, IBM SSD.
INTDISK4 on MKTTOOLS, 08/98
Available to your IBM representative

Internal Disk for S/390 Multiprise 2000, G221-9010-00

Input/Output Configuration Program User's Guide and ESCON CTC
Reference, GC38-0401-04

IBM Multiprise 2000 Internal Disk Marketing Flash
by John Hopkins, IBM SSD, 11/22/96, 3 pages
Available to your IBM representative

9672 SAP Performance and Configuration Guide
WSC Flash 9646.5, 11/96, 4 pages
Available to your IBM representative

S/390 Multiprise 2000 Internal Disk Subsystem -User's Guide-,
SA24-4261-02

S/390 Multiprise 2000 Internal Disk Subsystem -Reference Manual-,
SA24-4260-1

-FBA to ECKD Migration Aid, Internal Disk for the Multiprise 2000-
ITSO Boeblingen brochure, 48 pages, S/390 White Paper 07/97
As SG242000 on MKTTOOLS, available to your IBM representative

VSE CACHE Command for Internal Disk

VSE CACHE Command for Internal Disk

Applies to MP 2000 and MP 3000 Internal Disk

Use the VSE CACHE command to display cache info

„ **No cache settings possible/required**

CACHE ... ON|OFF not accepted

„ **Check for real usage of IDFW/IDRFW**

Check that IDFW is really used (VSE/ESA V2) via

```
CACHE UNIT=cuu,STATUS  and  CACHE SUBSYS=cuu,STATUS
```

í **DASD Fast Write must be ACTIVE**

'NVS available' does not assure that Fast Write is really used.

„ **REPORT statistics**

```
CACHE UNIT=cuu,REPORT    CACHE SUBSYS=cuu,REPORT
```

í **ID 'WRITE hits' do NOT mean that IDFW is active or installed**

Cache-WRITE 'hits' are shown and NOT an indication for IDFW active, just for a WRITE, where the track was already in cache. The completion still may be signalled to S/W only when data are on the HDD, i.e. w/o Fast Write being active.

The Internal Disk cache management itself has only limited information regarding the 'Internal Disk WRITE hits'. Please consider that when interpreting detailed IDFW hit ratios.

- For more info on the CACHE command, refer to
  - Foils 'VSE/ESA DASD Cache Statistics' in this document
  - VSE/ESA System Control Statements, SC33-6613

Internal Disk SMF Measurement Data

Internal Disk SMF Measurement Data

Applies to MP 2000 and MP 3000 Internal Disk

- **Device timings are accumulated by the channel emulation function**

  CONNect and DISConnect times have to be interpreted correctly

- **PENDing time**

  Usually the time until the control unit executes the first CCW.

  - The control unit function is performed in the SAP, also, there is no physical channel path.

  > No Device busy, CU busy or director port busy exists. Just the load of the SAP will contribute to PEND time

- **DISConnect time**

  This is the time the SAP is disconnected from the device, since
  - a miss occurred for READ or
  - a WRITE has to be performed

- **CONNect time**

  This is the time the SAP spends executing a channel program for a logical volume, without the time being disconnected at misses (i.e. when data have to be moved acrosss the SCSI or SSA busses).

Multiprise 3000 Internal Disk (ID)

Multiprise 3000 Internal Disk (Summary)

Announced 09/99 with the Multiprise 3000 servers

- **An enhanced implementation of Internal Disk, compared to the Multiprise 2000 ID**

  Fast 18G HDDs (10000 RPM)
  SSA
  RAID-5
  Better DASD Fast Write
  IDFW replaced by IDRFW (Internal Disk RAID Fast Write)

- **Disk Cache also resides in S/390 memory**

  Up to 2 GB, in increments of 32 MB

- **'Usable' Capacity 72 up to 792 GB.**

  3390 formatted capacity is up to 757 GB.

  Up to 216 GB in Base CEC Unit, plus up to 288 GB each in Expansion Frame A and B
  Base CEC Unit must be fully populated, before Expansion Unit A is used, etc

- **Supported by VSE/ESA 2.2 with PTFs (and up)**

  VSE/ESA does not use/exploit those new CCWs which were introduced by the ESS I/O Subsystem in 07/99 (PFX, RTD, WFT etc)

- **Available only as option for Multiprise 3000 processors**

  Can be used concurrently to channel attached I/O subsystems. Emulated I/O via OS/2 only for migration purposes.

MP 3000 ID Config. Scheme

Multiprise 3000 ID Scheme

```
                                      System Assist    ----------------
                                        Processor     |                |
    ---------        ---------        ---------       |Expanded Storage|
   |         |      |         |      |         |      |                |
   | S/390   |      | S/390   |      | S/390   |      |----------------|
   |    CP   |      |    CP   |      |    SAP  |      |                |
   |         |      | (2-way) |      |         |      |Central Storage |
    ---------        ---------        ---------       |                |
       |                |                |            |----------------|
        ----------------------------------------------|                |
     Main memory bus         |                        |ID Disk Cache   |
                          -------                     |                |
                 *) <--..| 'MBA' |                    |----------------|
                          -------                     |                |
                         |Bridge |                    | HSA            |
                         |STI-PCI|                     ----------------
                          -------                      Processor Memory
                             |
    PCI   ------------------------------------------
                |                       |          |
             ------------          -----------     V
            |   RAID-5   |        | RAID-5    |  **) Emulated I/O, OS/2
            | adapter 1  |        | adapter 2 |      (Migration only,
  Port:     | B        A |        |  B        |       14 GB)
             ------------          -----------
  SSA loop:  |a|       |c|          |b|
             | |       | |          | |           *) DASD Channels
          .......... ......... ..........              (Parallel/ESCON)
  Box:    .BASE CPC. . EXP-B . . EXP-A   .
          .        . .       . .         .
  #HDDs:  .   16   . .  20   . .   20    .
          .......... ......... ..........
```

- **DASD Subsystem is managed by advanced RAID-5 adapters**
  - PCI attached
  - 64 MB of DRAM, 32 MB of battery powered NVRAM (NVS), each
  - 20 or 40 MB/sec per port

- **Disk drives (HDDs) are cabled in SSA loops**
  - 1 SSA loop per box
  - Up to 3 RAID-5 arrays per loop
  - 2 data paths to each HDD, each full duplex (statically assigned)
  - Up to 4 concurrent data transfers per loop
  - Multiple commands travel around the loop simultaneously
  - Component failures are recognized and re-routed

MP 3000 ID Config. Scheme ...

ID Configurations (of a box or SSA loop)

Configur. name	Array Configur. 5-way (4+P)	Array Configur. 7-way (6+P)	Total #HDDs (incl. 1 spare)	Gross 'usable' Capacity
Configurations in the Base CPC Unit				
B1	1	-	6	72 GB
B2	2	-	11	144
B3	3	-	16	216
Configurations in a Expansion Frame				
E1	1	-	6	72 GB
E2	1	1	13	180
E3	1	2	20	288

- 1 spare HDD is shared between all (1...3) arrays of 1 SSA loop (or 1 box)
- Total Capacity 216 + 288 + 288 = 792 GB
- Maximum Capacity '3390-formatted KB':
  207174 in Base CPC Unit
  275286 in each Expansion Frame
  757746 in total
- 1 SSA loop corresponds to 1 logical 3990-2 in IOCP

- **ID capacity increment is 1 full RAID-5 array**

ID RAID-5 Arrays

Array Type	#HDDs	Regions available	Max #addr.	3390 mapping example	'3390-MB'
5-way	5 =4+P	73	64	24 x 3390-3 + 1 x 3390-1	69058
7-way	7 =6+P	109	96	36 x 3390-3 + 1 x 3390-1	103114

- Naturally, parity data is wrapped around all HDDs of an array
- 'Regions' are ID internal allocation units
- Fully configured, 5x64 +4x96 = 704 addresses can be used (within the limits of total capacity)

MP 3000 ID Config. Scheme ...

Logical Volume Types

Logical Volume Type	Regions used	#Cylinders	'Logical Capacity'
3380J 3380E 3380K	1 2 3	885 1770 2655	630 MB 1260 1890
3390-1 3390-2 3390-3 3390-9	1 2 3 9	1113 2226 3339 10017	946 MB 1892 2838 8514
- 'Regions' are ID internal allocation units - Bytes per track: 3380 47476 3390 56664			

- **Higher exploitable capacity with 3390 volumes**
- **Different logical volumes now can coexist within a RAID array**

HDD Characteristics

Ultrastar 18ZX 3.5" Disks & Interposer

(HDD Carrier Assembly)

	MP 2000 ID	MP 3000 ID
Ultrastar	18XP *1 SCSI	18ZX SSA
Capacity (512 byte)	18.2 GB	18.2 GB
Rotation Avg SEEK Time Latency Data Rate(Inst) MB/sec (Sust) MB/sec SSA Feature MB/sec Buffer Size (used)	7200 RPM 7.5 msec 4.17msec 11.5 - 22.4 - .5 M	10020 RPM 6.5 msec 2.99msec 23.4 - 30.4 17.4 - 23.4 40 1M
*1 MP 2000 ID also had 9.1 GB HDDs		

MP 3000 ID Performance Aspects

HDD Failure

- **If an HDD fails, the RAID-5 adapter automatically rebuilds the data on the hot spare**

  - All data from all other HDDs in the array have to be read
  - I/O accesses from processor (production) continues
  - To limit the performance impact on production ...

    Only part of the HDD access capability (MB/sec) is being 'grabbed' for that rebuild

  -> HDD rebuild does not impact production too much, and does not take too long (appr. 1 hour)

Performance Aspects

- **Cache size in S/390 memory can be flexibly configured**

  32 MB in increments of 32 MB up to 2 GB.

  Use same rule of thumb as before (just more generously):

  - **Use about 1 to 3 MB cache for 1 GB of data**

    This is a 'Cache to Backstore Ratio' of 0.1% to 0.3%.

  Another Rule-of-Thumb (ROT) for DASD cache sizes is

  - **Use about 1 to 3 MB cache per 1 IO/sec**

    Both ROTs coincide for an average 'Access Density' of 1 IO/sec per installed GB DASD (refer to Chart D16).

    So, it may be benefial to look at both ROTs.

  - **Leave enough main storage for VSE/ESA**

    Very rough rule for VSE: up to 6 MB per MIPS consumed ...

    - to allow good exploitation of Data In Memory (DIM)
    - to just be on the safe side and do not page

MP 3000 ID Performance Aspects ...

Performance Aspects (cont'd)

- **DFW may be reset by ID, if NVS not fully functional**

  IBF is optional, BUT no more needed for ID DFW

- **Check status via CACHE SUBSYS=cuu,STATUS**

  DFW must be ACTIVE, NVS must be ON

- **Fast Write misses for Update WITEs are only obtained**

  - at first reference of a track (if format is 'unpredictable')
  - at a subsequent reference, if track has been discarded meanwhile from cache and if track format is 'unpredictable'.

  Format WRITEs usually are hits

  (since formatting channel pgms usually start at begin-of-track)

- **Seq. performance is better than on MP 2000 ID**

  Holds for READs and WRITEs, especially for single stream:

  - Faster HDD
  - Overlapped operations to HDDs of RAID-5 array
  - High SSA loop capacity (mult. streams)

- **No Sequential Detect function in ID**

  Thus, it is still important for the S/W to use SEQ caching indication in ECKD channel programs in order to initiate pre-staging in a cached I/O subsystem.

  (Each HDD does seq. pre-stage anyhow, until interrupted, so the benefit may be no more as huge as with former HDDs)

MP 3000 ID Performance Aspects ...

Performance Aspects (cont'd)

- **Record Caching**

  If a track is predictable format, only the pertinent record(s) are being read from DASD, otherwise all track is being read

- **Unit of Staging**

  In case of a cache miss, the following are the units staged from HDDs to cache:

	' Predictable' Tracks	'Unpredictable' Tracks
Normal caching	Rest-of-Track	Total Track
Record caching	Record(s) only	Total Track

- **No Adaptive Caching is implemented in ID**

  I/O subsystem does not adaptively change between Normal and Record Caching.

  > Performance may change if Record Caching is used by the underlying S/W
  > - cache friendly
  > - cache unfriendly

MP 3000 ID Performance Hints

Performance Hints

„ **I/O Distribution within MP 3000 ID Subsystem**

í **RAID-5 automatically spreads hot spots and high I/O rate of a volume across all HDDs of an array**

í **Still avoid a too high I/O rate to a single logical volume**

A high IOSQ time in the operating system would be the result, if I/Os are from CICS or from multiple batch partitions

í **Try to roughly balance I/O activity across SSA loops**

í **Still, all volumes must be ADDed in VSE as 'ECKD'**

Not only a functional requirement, also optimal ECKD channel programs are benefial.

í **For better exploitation of DASD capacity ... use 3390 track format**

No performance impact expected for 3380 vs 3390

MP3000 ID Performance Results

MP3000 ID Performance Results for VSE/ESA

„ **Measurement Environment**

Hardware

- MP 3000 Model H50 (Uni-processor)
- Internal Disk microcode as of 99-11-03
- Fully populated BASE CPC, with 216 GB Internal Disk
- 256 MB Internal Disk cache, as part of processor memory

Software

- VSE/ESA 2.3.2 with Turbo Dispatcher (status DY44820)

Workload

- PACEX batch workload (1, 4, 8, and 16 concurrently active VSE partitions)
- VSE System volumes plus, roughly 1 volume per partition, using mostly 1 RAID-5 array (2 arrays for PACEX16)
- I/O loads are characterized by a R/W ratio of 1.52, i.e. 39.7% of all I/Os are WRITEs (a relatively high WRITE content for total workloads)

„ **Measurement Goals**

- Drive the MP3000 Internal Disk with varying I/O rates
- Compare performance values (as far as meaningful/possible) to formerly obtained results for MP2000 ID.

MP3000 ID Performance Results ...

Measurement Results

MP3000 Internal Disk Performance

Case	#VSE vols	#arrays	msec/IO	IO/sec	SAP util	CPU util
PACEX1	3+1	1	1.4 ms	642	23%e	8%
PACEX4	3+4	1	2.3 ms	1470	52%e	20%
PACEX8	4+8	1	3.2 ms	1790	62%e	25%
PACEX16	6+16	2	7.8 ms	1961	69%	31%

- PACEXn means n times 7 jobs in 1 partition each
- VSE/ESA 2.3.2 with Turbo Dispatcher
- R/W ratio =1.52 (39.7% WRITEs)
- 5 HDDs per array (Base CPC only)

- READ hit ratio varied from 0.90 to 0.95
- WRITE hit ratio varied from 0.91 to 0.97
- Most of msec/IO was DISCONNECT time (as expected)

í **Very very good I/O response times**

even at higher I/O load and 1 array only (PACEX8)

The high SAP utilization (vs the CPU utilization) is caused by the very I/O and WRITE intensive PACEX workload

MP3000 ID Performance Results ...

Comparison to MP 2000 (08/97) ID Results

	Throughput (IO/sec)		I/O response time	
Case	MP2000->MP3000 ID	Ratio	MP2000->MP3000 ID	Impr.
PACEX1	289 -> 642	2.22	3.3 -> 1.4 ms	2.39
PACEX4	880 -> 1470	1.67	4.2 -> 2.3 ms	1.82
PACEX8	1346 -> 1790	1.33	4.6 -> 3.2 ms	1.44
PACEX16	na -> 1961	na	na -> 7.8 ms	na

As base line, here the MP2000 results as of 08/97 were used, i.e. 256 MB cache at a 2003-116, with 9 GB HDDs.

í **Much better I/O response times, even at higher I/O rates**

PACEX(1) Scenario Consideration

Consideration of e.g. the PACEX1 scenario (no queueing):

1 I/O stream = 1 partition with 18000 I/Os to 1 user volume

```
                CPU-time        I/O time (msec/IO)
              |----------|==============================|

MP2000:         0.455 ms            3.34 ms

MP3000:         0.133 ms            1.39 ms
```

The resulting I/O rates calculated with these 2 values are very close to the measured ones:

MP2000: 1000 ms / (0.45+3.34) ms = 264 IO/sec
MP3000: 1000 ms / (0.13+1.39) ms = 658 IO/sec

If MP2000 would have had the same processing speed as the MP3000 model used:
1000 ms / (0.13+3.34) ms = 288 IO/sec

would have resulted, then.

Conclusion:
The significantly faster MP3000 speed only had a minor impact on the increased I/O throughput/speed.

MP 3000 ID PTFs

VSE native

VSE/ESA PTFs, required for running MP 3000 ID:

APAR	PTF	Component
DY45179	UD51135	VSE/AF 2.2 (IOCP)*)
	UD51136	VSE/AF 2.3 (IOCP)
	UD51138	VSE/AF 2.4 (IOCP)
DY45181	UD51077	EREP
PQ26800	UQ90021	ICKDSF
PQ29648	-	addt'l info to PQ26800
The new IOCP CHPIDs refer to new types of channels TYPE=DSD: 'Direct System Device channel' for ID TYPE=EIO: 'Emulated I/O channels' for EMIO		

VM/VSE

```
APAR VM62180 + VM62111 + VM62312 are required to
             - allow exploitation of new CCWs (from ESS)
               by VM/ESA guests (does not apply to VSE)
             - define the new IOCP CHPIDs
```

More Info

- Internal Disk White Paper. Update for MP 3000. Available to your IBM representative via MKTTOOLS
- IBM Ultrastar family specifications. Via http://www.storage.ibm.com
- Multiprise 3000 announcement and documentation: Via http://www.s390.ibm.com/multiprise
  - Multiprise 3000 Reference Guide, G326-3081-00
  - Multiprise 3000 Product Advisor (Web based capacity planning tool)
- Internal Disk Subsystem Reference Guide, SA22-1025 Via http://www.ibm.com/servers/resourcelink
- Internal Disk Subsystem User's Guide, SA22-1026 Via http://www.ibm.com/servers/resourcelink
- S/390 Multiprise 3000 Integrated LAN Adapter Feature. Performance Report, Nov 99. GF22-5136
- Multiprise 3000 Technical Introduction. IBM Redbook SG24-5633-00, 11/99. 133 pages
- Planning and Implementation for the Multiprise 3000 VM/VSE Technical Conference 06/2000, Orlando. By Dennis Ng

DIM and I/O Caching, Global View

PART K.

DIM and I/O Caching, Global View

For information on VM/ESA Minidisk Caching (MDC), refer e.g. to 'IBM VSE/ESA VM Guest Performance Considerations'

For all VSAM Shareoption related files,

thanks to Horst Sinram VSE/VSAM Development for assistance.

DIM/Caching Hierarchy

DIM/Caching Hierarchy

```
  Full Key   VSAM   Non-VSAM  Non-volatile  Volatile (work)
  VSAM LSR   READ     READ     WRITE          READ/WRITE
    READ                                    Shared  NonShared
                                          (VSE ext.)(VSE int.)
                                                               Level
|          |       |         |           |         |         |
|----------|       |         |           |         |         |
| CICS Data|       |         |           |         |         |
|  Tables  |       |         |           |         |         |
|          |       |         |           |         |         | Part.
|----------|-------|         |...........|.........|.........| Cache
|                  |         |           |         |         | 'PC'
|  VSAM LSR (NSR)  |         |           |         |         |
|                  |         |           |         |         |
|------------------|---------|...........|---------|---------|  ...
|                            |           | also    |         |
|   VSE-global S/W Caching   |           | VSE S/W |Virt.Disk| Global
|  (CACHE/VSE, OPTI-CACHE,   |           | Caching |(VSE VD, | Cache
|   BIM VIO non-volat. ...)  |           |         | BIM VIO)| 'GC'
|                            |           |         |         |
|----------------------------|...........|---------| --------| ...
|       VM MDC Caching       |           |         |           VM
|       (CACHE-MAGIC ...)    |           |  VM VD  |           Cache
|                            |           |         |           'VC'
|----------------------------------------|---------|           ...
|           Multiprise Int. Disk         |         |
|                                        |
|       H/W (Subsystem) Cache w/ NVS     |                      HW
|           (READ/WRITE Caching)         |                      Cache
|                                        |                      'HC'
|----------------------------------------|
```

- Very schematic view, showing each possible layer (VM cache layer is available only for VM/VSE)
- Categories at top designate performance benefit areas, not file eligibilities
- Naturally, it is not reasonable to implement each layer in a single environment
- VSE-global S/W caching and VM MDC Caching are similar, except when data are shared between VSEs
- Multiprise Internal Disk caching cannot be shared across processors

Overall Statements

Overall Statements

- **Caching with NVS in the I/O Subsystem**

  is the only means for WRITE caching w/o any risk

  (also across processors)

  no impact on CPU-time

- **All S/W DIM means or S/W caching products require additional real memory**

  The more versatile (files eligible) a product is

  and/or

  the more global (across VSE partitions) a product works

  and/or

  the better it can react to workload/access pattern changes ...

  - **the more effectively the available real memory is exploited**
  - **the bigger are the overall performance gains**

- **Effective READ caching with full READ integrity**

  for all participants can only be done

Data shared	Caching types allowed
Within a single VSE	VSE Global Cache GC +VC +HC
Across VSEs under same VM	VM Cache VC +HC
Across processors	H/W Cache HC

VSAM Share Options and I/O Caching

VSAM SHROPT (2)

Ù **Any number of READs, plus 1 WRITE**

```
READ means here 'INPUT OPEN', WRITE means 'OUTPUT OPEN',
also READ/WRITE is used for VSAM GET/UPDATE requests

      <..........VSE1............>| <.........VSE2..........
                                  |
         Part1         Part2      |      ....
                                  |
         R / W         R          |     R
                                  |
         A   |         A          |     A
         |   V         |          |     |
         |   |         |          |     |
         PC  -         PC*        |     PC*
         |   |         |          |     |
         |   |         |          |     |
         A   |         A          |     A
         |...|...GC....|          |     |..GC*..
         |   |         |                |..VC*..
         |...|.........|........VC......|.......
         |   |         |                |
         |   |         |      ------------
         |   |          --<---|            |
         |   |                | VSAM SHR 2 |
          -------<------------|    file    |
             |                |            |
             |                 ------------
             V                     A
             |                     | WRITEs
              ---------------------

    PC   Partition Cache for READ allowed,
         but only the single updating partition has full READ
integrity

    GC   Global Cache possible, since all WRITEs done from 1 VSE.
         All partitions of VSE1 with full READ integrity

    VC   VM Caching allowed, even if VSE2 is outside VM/VSE1

    *    Caching w/o full READ integrity
    -    No locks at all needed for READ and WRITE
```

í **Any type of READ caching allowed (at any level)**

```
Full READ integrity not in all cases
```


VSAM Share Options and I/O Caching ...

VSAM SHROPT (4 3)

Ù **Any number of READs, plus (any number of WRITEs, but from 1 VSE only)**

```
      <..........VSE1............>| <.........VSE2..........
                                  |
         Part1         Part2      |      ....
                                  |
         R / W         R / W      |     R
                                  |
         A   |         A   |      |     A
         |   V         |   V      |     |
         |   |         |   |      |     |
        NPC  -        NPC  -      |    NPC
         |   |         |   |      |     |
         A   |         A   |      |     A
         |   |         |   |      |     |
         |...|...GC....|...|      |    NGC
         |   |         |   |            |
         |...|.........|...|....VC......|  NVC
         |   |         |   |            |
         |   |         |   |      ------------
         |   |          --<---|            |
         |   |             |  |VSAM SHR4 3 |
          -------<------------|    file    |
             |             |  |            |
             |             |   ------------
             V             V       A
             |             |  ...  | WRITEs
             |             ---. .---
              ----------------. .
                              ... VSE internal LOCK (all WRITEs)

    NPC  No Partition Cache for READ allowed, since updates from
other
         partitions not propagated (VSAM reads always 'from DASD')

    NGC  No Global Cache for READ allowed (VSE-global S/W Cache),
         since updates from other VSEs are not propagated

     GC  Global Cache possible, since all WRITEs done from 1 VSE

     VC  VM Caching allowed, if VSE2 under same VM

    NVC  No VM Caching allowed, if VSE2 under separate VM
```

í **For 'READ-Only VSEs' only VC under same VM allowed**

í **For 'WRITE VSE' only GC and VC is allowed**

VSAM Share Options and I/O Caching ...

VSAM SHROPT (4 4)

Ù **Any number of READs, plus any number of WRITEs**

```
      <..........VSE1............>| <.........VSE2..........
                                  |
         Part1         Part2      |      ....
                                  |
         R / W         R / W      |     R  /  W
                                  |
         A   |         A   |      |     A     |
         |   V         |   V      |     |     V
         |   |         |   |      |     |     |
        NPC  -        NPC  -      |    NPC    -
         |   |         |   |      |     |     |
         A   |         A   |      |     A     |
         |   |         |   |      |     |     |
         |   |  NGC    |   |      |     |     | NGC
         |   |  NVC    |   |            |     | NVC
         |...|.........|...|....VC......|.....|
         |   |         |   |            |     |
         |   |         |   |  ------------    |
         |   |          --<---|            |  |
         |   |             |  |VSAM SHR4 4 |  |
          -------<------------|    file    |  |
             |             |  |            |  |
             |   ----------    ------------   |
             |  |   --------        A         |
             |  -->|  LOCK  |       | WRITE   V
              ---->|  file  |-------          |
                -->|        |                 |
                |   --------                  |
                 ---------------<-------------

    NPC  No Partition Cache for READ allowed, since otherwise
updates
         from other partitions not propagated. Each VSAM READ must
be
         'from DASD' (NSR/LSR), except GET SEQ NOUPDATE

    NGC  No Global Cache for READ allowed (VSE-global S/W Cache),
         since updates from other VSEs are not propagated

     VC  VM Cache allowed only if all VSEs under same VM

     -   All WRITEs first require a LOCKfile access.
         Since a 2nd VSE is also allowed to WRITE,
         VSAM CAs are locked for owning tasks.
         (If all VSEs are under same VM, LOCKfile can be cached in
VM)
```

í **PC and GC READ caching NOT allowed**

í **VC is ONLY allowed if ALL VSEs under same VM**

VSAM Share Options and I/O Caching ...

VSAM Share Option 4 and I/O Caching

Ù **READ caching for VSAM SHROPT 4 is NOT allowed**

VSE Global Caching (GC)	
VSAM SHROPT (4 3)	for 2nd (READ Only) VSE
(4 4)	for all partitions in all VSEs
VM Caching (VC)	
VSAM SHROPT (4 3)	for 2nd (READ Only) VSE, except under same VM
(4 4)	except all VSEs under same VM

A good caching product should at least tell this to the user.

```
If possible, this should also be directly implemented in the product
itself and not only documented somewhere
```

More VSAM Data Sharing

```
Refer e.g. to

  - 'VSAM Data Sharing Across IBM S/390 Systems'
    by Horst Sinram, IBM Boeblingen
    VM and VSE Tech Conf, Mainz, Germany 06/97, Session 50F
```

VSE DIM Means and I/O Caching

Individual Statements

(Performance comparisons can only be done if function-wise applicable)

- **CICS Data Tables have unbeatable benefits**

  but only for full-key KSDS READs

  Should be used if applicable

- **VSAM LSR is THE means for 'usual DIM' (most used files)**

  but subpools should not be too big

  A very intelligently implemented S/W cache vendor product may show CPU-time benefits vs 'big' LSR DIM

  No measurement results available

- **VSE-global S/W caching vendor products (caching in VSE storage)**

  may save some VSE CPU-time, but at least save CP time if under VM
  Depends on VM guest and DASD setup

  can for SHROPT 4 only be used for the single WRITE-VSE of SHROPT 4 3

  (READ caching advantages, similar to e.g. VSAM GSR in MVS)

VSE DIM Means and I/O Caching ...

Individual Statements (cont'd)

- **Virtual Disks**

  are only for work data,
  except when enhanced by a non-volatile option (BIM-VIO)
  are no 'real caching' products, since staging/de-staging done by paging

  Use VSE VD instead of VM VD, if applicable

- **VM MDC for guests**

  is very versatile
  is the only means to cache production data shared between VSEs under same VM with multiple WRITEs (SHROPT 4 3 or 4 4)

  BUT ...
  does not apply to native VSE
  does not save VSE CPU-time
  can only be used across multiple VSEs if ALL WRITE-VSEs are under this VM

- **Always use H/W Caching (esp. WRITEs with NVS)**

  Is well suited to complement usual DIM or any type of S/W caching

  Is required/provided anyhow for RAID-5 (RAMAC)

Misc VSE I/O Aspects

PART L.

Misc VSE I/O Aspects

File Placement on Disk(s)

File Placement Within a Disk

- **For traditional (non-simulated) S/390 disks put files about in the middle of the pack**

  (if possible also VTOC, except for VSAM-Only disks)

  -> Reduce overall SEEK times

  Is less important if physical S/390 device is cached.
  Is unimportant if S/390 logical device is RAID5 on 3.5" HDDs

File Placement and Sharing Across S/390 Disks

- **In any case, S/390 DASD utilization should be about balanced**

  About overall balanced S/390 DASD utilizations is beneficial not only for physical devices, but to some extend also for simulated devices e.g. via RAID-5.
  Per (logical) S/390 device only 1 SSCH can be active at 1 point in time (seen from VSE)

- **Put non-shared data on non-shared S/390 disks**

  This is a general rule which does not only bring performance benefits, but also is reasonable for non-performance reasons

- **Avoid, whenever possible, to ADD S/390 disks as cuu,SHR**

  Reduces sharing overhead, especially for non-shared files

- **For more hints on DASD Sharing, refer to the VSE/ESA V2 base document**

VTOC Performance

Basics

- **VTOC accesses are done by the Common VTOC Handler (CVH)**

  - at each BAM label read
  - at each BAM or VSAM space allocation
  - for safety reasons at each VSAM file OPEN and each first usage of a new VSAM extent by a file to check coincidence between catalog and VTOC info ($$BOVS01 just reads the VTOC file label (Format-4))

  Search of the whole VTOC is required
  - to look for 'equal ids'
  - to check for overlapping extents

- **VTOC position/layout is less important for**

  - fully VSAM owned volumes
  - cached volumes

Optimized VTOC layout for VSE/ESA V2 system volumes

Used in the Base Install process if 'Automatic VTOC Initialization' was selected

- **Increased FBA VTOC CISIZE from 1K to 8K**
- **Reduced CKD/ECKD VTOC size to 4 tracks**

  Tracks 11-15

General Recommendation

- **Place VTOC in about the middle of a disk**

  This may show slight overall seek improvements for multi-thread environments except in cases with VM Partial Pack Minidisks (in spite of reading the volume label by the CVH)

General CKD/ECKD VTOC Hints

General CKD/ECKD VTOC Hints

- **Place VTOCs on the last tracks of a cylinder**

  Label read (READ Format-1 Label by name) uses multitrack search. They search, if required, until end-of-cylinder, even if extent (here VTOC) ends at an earlier track.

  When a new BAM file or VSAM space is defined, it is necessary to read all F1 labels, in order to avoid 'overlap on unexpired file'.

  Such VTOC CKD/ECKD I/Os read 1 entry each

- **Use only as many tracks as required**

  A mostly VSAM owned disk volume may need only 1 track

Í **Reduces time for VTOC accesses**

- **VTOC CKD/ECKD Capacities**

#VTOC entries per track	
3375 3380 3390 9345	51 53 50 45
- 4 tracks used for VSE system volumes	

- **ICKDSF 16 for VTOC Extension/Relocation**

  Use REFORMAT with EXTVTOC to extend an existing CKD/ECKD VTOC, use REFORMAT with NEWVTOC to move and extend an existing VTOC

General FBA VTOC Hints

General FBA VTOC Hints

- **Use 8K CISIZE**

  44 byte key + 96 byte data = 140 byte record

- **Per FBA-VTOC-I/O 1 CI is read**

  > Bigger CIs reduce the number of VTOC I/Os

- **VTOC FBA Capacities**

#VTOC entries per CI (0671, 3370, 9332, 9335, 9336)	
1K-CI 8K-CI	7 57
- 4 8K-CIs used for VSE system volumes	

BUFSIZE Consideration

BUFSIZE Consideration

- **Background info**

  The CCW translation for I/O operations uses translation buffers (72 byte each) to store copy blocks

  The number of required translation buffers directly depends on the complexity of a channel program

  Any copy block is kept at least until the corresponding I/O operation has been completed:

  - For NOFASTTR, buffers are freed after I/O interrupt handling ('re-translation')
  - For FASTTR, buffers are kept for reuse, but at most 1 sec

  All copy blocks are handled partition individually (e.g search for FASTTR duplicates), but total bufferspace (BUFSIZE) is common for a VSE system.

  For FASTTR vs NOFASTTR refer to the VSE/ESA 1.3/1.4 performance document

- **Number of translation buffers required**

  The BUFSIZE requirement (check via SIR) increases with

  - the complexity of the channel programs used
  - the total I/O rate
  - the average msec per I/O
  - the number of active partitions
  - the number of partitions and the I/O rate for FASTTR

BUFSIZE Consideration ...

BUFSIZE (cont'd)

- **No VSE msg when a task is waiting for copy blocks**

  STATUS part-id snapshots may show you this situation.

  In seldom, serious cases, message

  0V06I NOT ENOUGH BUFFERS FOR CHANNEL PROGRAM TRANSLATION

  may occur.

- **High water mark for used copy blocks**

  For problem analysis purposes, the high water mark of used copy blocks can be displayed by the SIR command:

  COPY-BLKS = 00195 HIGH-MARK = 001690 MAX = 3000

  With SIR RESET, the HIGH-MARK can be reset.

  Note that with FASTTR, HIGH-MARK usually is close to any reasonably specified MAX (=BUFSIZE) value i.e. the HIGH-MARK value is less informative.

- **For NOFASTTR, BUFSIZE=2000 is mostly sufficient.**

  For FASTTR or higher I/O rates, use up to 3000
  Via 4K rounding, actual BUFSIZE is larger than specified

- **In VSE/ESA 2.4 FASTTR is no more available**

  - FASTTR was not easy to handle with
  - its benefits were very limited
  - was not so suited for Turbo Dispatcher

VSE/ESA Missing Interrupt Handler

VSE/ESA Missing Interrupt Handler (MIH)

- **Functional purpose**

  Detect any interrupt which (for whatsoever reason) was lost, not to react to slow I/Os

  After MIH seconds (VSE default is 180 sec), the VSE I/O supervisor displays an emergency message 0Exx to the console.
  Depending on the situation and user response, this may initiate recovery to the target device.

- **Performance aspect**

  MIH never should be set smaller than the longest possible I/O operation, initiated by VSE
  (i.e. not of any subsystem initated long destaging activity).

  This is expected to be full cylinder operations or similar long running DASD I/Os.

  But, usually the MIH value is determined by long tape operations, such as REWIND.

  If MIH is smaller, unnecessary messages would be issued costing also CPU time overhead

- **Recommendation**

  Leave the MIH times at the default (MIH=180)
  Note that in VSE the MIH value applies to all types of I/Os.
  We are not aware of any requirement to have MIH set > 180 sec.

  Please contact us if you would need to increase this value, e.g. via the SIR MIH command in VSE/ESA V2, as documented in the 'Hints and Tips for VSE/ESA' brochure:

  ```
  SIR MIH       displays the current MIH setting
  SIR MIH=nnnn  sets MIH value to nnnn sec
  ```

For general info on MIH, your IBM representative may refer to WSC Flash 9508 'RAMAC MIH Considerations'

I/O Performance PTFs

PART M.

I/O Performance PTFs

VSE/ESA 1.3/1.4 Performance APARs/PTFs

Some 1.3/1.4 APARs/PTFs for I/O performance

The next 2 PTFs became available 03/94 and refer to DASD caching with VSAM (thus the PTFs are standard since VSE/ESA 1.3.5):

* DY43072 UD90363 VSAM support for 3990-6 Enhancements

  This PTF provides the VSAM support for 'regular data format' and for 'record cache mode' of the 3990-6 enhancements.
  Also, seq. bits are set for better cache control during VSAM SPEED load mode.
  This PTF installed (or by default included since 1.3.5) requires a 9340 u-code patch (E6392AC)

* DY43138 UD49025 VSAM B/R cache bit settings for ECKD

  This PTF uses the sequential caching bits instead of bypass cache in order to speed up Backup(!) to a target disk.
  It also applies to 9345 Cache, which in its latest EC 486392 adequately exploits the sequential setting.

The next PTF was closed 11/14/94 and is contained in 1.3.6:

* DY43312 UD49234 UD49237 PTFs retrofitted from VSE/ESA 2.1

  This PTF contains also an enhancement of the CKD/ECKD conversion routine, beneficial for WRITEs with specific CKD channel programs (e.g. CICS journal)

The next PTF was closed 02/10/95 and is not included in 1.3.6.

* DY43414 UD49333 VSAM B/R restore performance for 3990-3

  This PTF sets the beginning of the extent address in the DEFINE EXTENT CCW for VSAM B/R to the begin of the current extent, in order to allow an optimal sequential de-staging for 3990 type of cached control units during RESTORE

* DY43335 UD49325 UD49332 RAMAC Array DASD and Format Writes

  This PTF corrects a problem in the RAMAC Array DASD, which loses a revolution when a standard R0-record is written and a specific bit is not set.

VSE/ESA 1.3/1.4 Performance APARs/PTFs ...

Some 1.3/1.4 APARs/PTFs for I/O performance (cont'd)

```
* DY42800    UD48965    VSAM Load mode performance for CI-mode files
             UD48966

   This PTF allows that multiple CIs are chained in the same VSAM
   channel program when loading or pre-formatting a CNV opened file.
   It was retrofitted from VSE/ESA 2.1 to VSE/ESA 1.3


* DY43207    UD49163     IPL accepts ADD cuu,ECKD for 3380 devices
             UD49164     if attached to an ECKD capable synchronous
                         control unit.

   Further functional enhancements are included in this fix.


* DY43836    UD49763     VSAM I/O performance for ECKD format writes

   This PTF corrects a VSAM sector value when doing format WRITEs
   to ECKD attached devices.
   It applies to all ECKD DASD attachments and especially to RAMAC
   Array Subsystem. VSAM REPRO is affected and formatting of new
   extents, no benefit for VSAM B/R Restore


* DY43416  UD49348     VSAM performance improvement for CNV load mode

   This PTF allows chaining of several CIs when loading a VSAM file
   with MACRF=CNV (CI-processing) and VSAM buffering (MACRF=NUB).

   It applies especially to ADSM/VSE if disk space is acquired via
   DEFINE VOLUME.


* DY44358  UD50212/15  Misc. plus RESET of SIR dynamic counters

   This PTF for VSE/ESA V1.3/1.4 allows to RESET SIR counters,
   so far incremented always since IPL time.


This list of APARs

   - is provided to give fast hints to resolved performance problems.
     PTF numbers may have changed, so always refer to APARs when
     ordering fixes.

   - is also contained in the base documents
```


VSE/ESA 2.1/2.2 Performance APARs/PTFs

Some 2.1/2.2 APARs/PTFs for I/O performance

```
* DY43697/8 UD49662    Some functional and performance enhancements:
            UD49664    Turbo Dispatcher improvements,
            UD49671-3  CKD/ECKD conversion routine enhancements,
                       CACHE SUBSYS=cuu,REPORT provides summary data

   With this PTF, e.g. the CKD/ECKD conversion routine is smarter
   to CKD programs with format writes if no sector value is given.
   Also, for native VSE, all data of all devices at a subsystem
   are now accumulated to directly provide the overall hit ratio.


* DY43844    UD49764     VSAM I/O performance for ECKD format writes

   This PTF corrects a VSAM sector value when doing format WRITEs
   to ECKD attached devices.
   It applies to all ECKD DASD attachments and especially to RAMAC
   Array Subsystem. VSAM REPRO is affected and formatting of new
   extents, no benefit for VSAM B/R Restore.


* DY44070  UD49933     VSAM catalog mgmnt, VSAM managed files on ECKD

   This PTF corrects some VSAM catalog management problems and
   provides channel program enhancements for VSAM managed SAM files
   on all types of ECKD devices (3380, 3390, 9345, RAMAC)


* DY43585  UD49565/66  Misc. problems plus CKD/ECKD conversion

   This PTF corrects also a performance problem created by non-
   optimal CKD/ECKD conversion (avoids protection checks for
   programs with multiple SEEKs)


* DY44277  UD50216/17  Misc. plus RESET of SIR dynamic counters

   This PTF for VSE/ESA V2.1/2.2 allows to RESET SIR counters,
   so far incremented always since IPL time.
   Check the PTF numbers, which may be obsolete meanwhile.


* DY44442  UD50251/52  Misc. plus SIR SMF,cuu command

   This PTF also includes a supervisor PTF for parallel POWER
   and an enhanced GETVIS SVA,DETAIL display
```


Appendix A: Tape Subsystems

PART N.

Appendix A: Tape Subsystems

3490 Performance Features

IBM 3490 and 3490E Tape Drives

„ **Survey**

	3490		3490E		
Models	D31/D32	A01/A02 B02/B04	D41/D42	C10/C11 /C22	A10/A20 B20/B40
# tracks	18	18	36	36	36
IDRC	opt.	std.	std.	std.	std.
ACL	opt.	std.	opt.	-/std. /std.	std.
Max.#channels tot	2	4/8	2	2	4/8
ESCON	1	2/4	2RPQ	2	4/8
# drives	1/2	2-16	1/2	1/1/2	2-16
Buffersize	2M	2M	2M	opt.8M	8M
Rewind speed	4m/sec	4m/sec	5m/sec	5m/sec	5m/sec
Perf.Enhancem.	-	-	opt.	feature	std.
Autoblocking even w/o IDRC	no	no	yes	yes	yes
# tape strings (controllers)	1	1/2	1	1	1/2
ADDed as	3490@	3490	3490E	3490E	3490E

@ ADDed in VSE as 3480 if w/o IDRC
- Perf. Enhancement includes a faster compactor and a larger auto-block size (128K)

```
Notes:
 - Uncompacted maximum (instantaneous) data rate:
     3MB/sec (independent of channel type and attachment)
     (2m/sec tape speed at appr. 1500 bytes/mm gives
     appr. 3000000 bytes/sec = 2929 KB/sec (K=1024))
   Value also applies for the aggregate tape-string data rate
   (1 controller).
 - 1 drive may be connected to >1 string (controller),
   but not be concurrently used (flexibility, no capacity benefit)
 - 3490-C1A and C2A drives in 3494 only
 - Cartridge capacity (without compaction):
   18 track: 200 MB, 36 track: 400 MB, 800 MB (enhanced)
```

Determining Factors for Tape Performance

Determining Factors for Actual Tape Throughput

Ù Achievable Effective Tape Data Rate

„ Tape attachment

Tape model and mode
Number of concurrent drives used
Number of tape strings
Number and speed of tape channels

„ Speed of supplied tape data

Application characteristics
Processor speed
Type of DASDs
Number of DASDs
Number and speed of DASD channels

NOTE:

The 4 charts here on 3490E performance are an excerpt from 09/93 (IBM INTERNAL USE) charts. Your IBM representative may explain additional results, if you wish.

Overall 3490E Summary

3490E vs 3490

Compared at same tape channel speed

- higher cartridge capacity saves change time
- bi-directional recording saves rewind time

„ Performance benefits essentially only
- if no IDRC is used
- at smaller blocksizes

caused by autoblocking and 'performance enhancement'

„ Same basic drive speed and rate as 3490 std.

ESCON vs parallel tape channel

„ Performance benefits only when tape channel(s) become a bottleneck

i.e. channel speed for single drive operation,
channel capacity for multiple drive operation

Tape channels become the more a bottleneck, the

- more drives are used concurrently
(preferrably at separate tape strings)
- higher the IDRC compaction factor is
(already at single drive operation)
- fewer tape channels are used
- lower job(s) are bound on the DASD side
(multiple DASD channels required)

FAST COPY Results for 3490E and 3390s

Environment

„ VSE/ESA 1.3.2 in 9121-190 ESA LPAR
1 to 3 1.5 MB batch partitions of equal priority, 34/64 MB real

„ FAST COPY DUMP VOLUME, OPTIMIZE=4
Backup of DOSRES (1 drive) and SYSWK1 (2 drives) and user volume (3 drives)

„ 3390-01 DASDs at 2 DASD channels
Same results would be obtained for 3390-02/03

„ 3490E B40 drives at same string, 2 tape channels

„ 4.5 MB parallel and 9.0 MB/sec ESCON channels for DASD and tape

FAST COPY Results

	Channel type DASD,tape	Options	Revolutions /track read	EDR MB/sec
Single drive	ESCON,	IDRC 2.87:1	1.27	2.27
"	"	no IDRC	1.27	2.27
"	"	IDRC, OPT=1	2.00	1.45
"	Parallel	IDRC 4.30:1	1.27	2.46
"	"	no IDRC	1.29	2.42
Two drives	ESCON,	IDRC -	-	4.43
"	Parallel	IDRC -	-	3.41
Three drives	ESCON,	IDRC -	-	5.14
"	Parallel	IDRC -	-	3.56

Conclusions

„ Higher overall throughput with ESCON, but only if channels become a bottleneck in multiple thread

Other ESCON benefits NOT considered here

VSE/FAST COPY OPTIMIZE Consideration

OPTIMIZE 4 vs OPTIMIZE 1 Performance

„ Environment

VSE/ESA 1.3.2
9121-190 processor, ESA LPAR
3390-01 DASDs (uncached) at ESCON channel
3490E B40, Mode='08' (IDRC), ESCON

8119 DASD tracks dumped (DOSRES), 334 MB total

„ FAST COPY OPTIMIZE Results

	OPTIMIZE 1	OPTIMIZE 4	Delta / Factor	
Elapsed time	230 sec	147 sec	-36%	1.56
CPU-time	n/a	n/a	-62%	2.65
DASD I/Os	8205	1711	-79%	4.77
Tape I/Os	8164	1675	-79%	4.77
Revol./track	2.00	1.27	-36%	1.56
Eff.Data Rate	1.45 MB/sec	2.27 MB/sec	+56%	1.56

Single FAST COPY is DASD revolution bound
(as expected)

OPTIMIZE 4 much more efficient
Elapsed and CPU-time

„ General OPTIMIZE Hints

- OPTIMIZE=1/2/3/4 (for 1/2/3/5 tracks per DASD and tape I/O) is specifiable for DUMP (ALL/VOLUME/FILE) only, i.e. not for COPY or RESTORE
- For RESTORE, the OPTIMIZE values from DUMP are used implicitly

Autom. Cartridge Loader Improvements

Automatic Cartridge Loader Improvements

- **Improved ACL selection in VSE/ESA 2.1:**

  'exhaustive' complemented by asynchronous 'alternate'

  via ACL=YES/NO parameter

- **Improved elapsed times for tape activities (ACL=NO)**

  - for non-3490E cartridges

  Rewind always required

  - if 3490E cartridge not full

  (Partial) Rewind in spite of bi-directional recording

3590 High Performance Tape Subsystem

3590 High Performance Tape Subsystem -Summary-

Supported with VSE/ESA 2.2.0 and up.
To be added in VSE with ADD cuu,TPA

- **High Reliability/Capacity/Performance**
- **New tape cartridge storage technology**

  - 8x16=128 longitudinal serpentine tracks (4x forward+backward, writing 16 tracks each time, addt'l servo tracks)

    -> Reduced avg REWIND times (bi-directional recording)

  Cartridge capacity of 10/up to 30 GB (uncompacted/compacted)
- **High tape speed and data rates**
  - 2 meter/sec
  - 9 MB/sec device peak (instantaneous) data rate (READ and WRITE, 3 times as much as 3490E)
  - Search speed 15x3490E = 166 MB/sec
- **On S/390: ESCON attachment only**
- **3590-A00 and -A50 controller**
  - 1 or 2 ESCON channel paths, up to 43 km distance
  - up to 4 drives (B11s and B1As)
  - 2x64 logical channels

  For 2 ESCON channels (or 3490E-mode), DY44364 is required
- **3590-B11 and -B1A drives**

  B11: 10-cartridge Automatic Cartridge Facility (ACF) (Random Mode not supported with ESCON)
  B1A: Use in 3494 Tape Library Server, no ACF

3590 High Performance Tape Subsystem ...

3590 Tape Subsystem -More Details-

- **Revised compaction technology (LZ1)**
- **Resources per 3590 drive and peak data rates**

```
                   A00
                 --A50->    --B11->
                |           |
                |           -----    ----    -------
 -----------    | SCSI bus |SCSI |  |com-|  | 4 MB  |   ---------
  ESCON <==>|-| ===<==>==|cont-|--|pac-|--| buffer|--|cartridge|
 -----------    |          |roll.|  |tion|  |       |   ---------
                |           -----    ----    -------
  17 MB/sec        20 MB/sec                              9 MB/sec
                                                        (physical)
```

(3590 control unit appears to drive as a SCSI host)

- **Up to 2 concurrent I/O operations per A00 controller**
  - 1 or 2 ESCON channels per A00 controller
  - 2 internal SCSI buses per A00 controller, connected to each drive
  - Up to 4 B11 drives per A00 controller
- **Automatic reblocking to 384 KB blocks**

  This reblocking, naturally, is transparent to any S/390 S/W.

  In order to save S/390 I/O pathlength and thus CPU-time, it is still important to specify adequate blocksizes e.g. in utilities.

  Also it is still important to use enough I/O buffers

3590 High Performance Tape Subsystem ...

3590 Tape Modes

TPA architecture as such allows many mode settings:

'00' to '0F' (buffered)
'20' to '2F' (un-buffered)

(those from 3480/3490, plus more by new WRITE formats 0 to 7).

For the 3590, only the WRITE formats
0 Use device default
1 3590 cartridge format
7 Use media default
are valid, all resulting in the same effective (3590) mode.

Using WRITE format 0, the following modes apply for 3590s

	Uncompacted	Compacted
Buffered write	00	08 (default)
Tape-write-immediate	20	28

Mode 08 is the default mode, valid if both
- the 3590 drives have been ADDed w/o any mode and
- the ASSGN is done w/o specifying a mode value

For performance reasons, mode 08 should be used

Performance Remarks

- Utmost achievable effective (sustained) data rates (no DASD involved, 3:1 compression ratio, single drive)

Blocksize	max EDR
32K	8 MB/sec
64K	11 MB/sec

Actual achievable data rates (with DASDs involved) are much lower in practice, see Reminder on next foil

- Actual Total Elapsed Times may include also REWIND times
  - not included in 3490(E) S/W times
  - included in 3590 S/W times (since S/W waits for CE, not DE)

  Consider REWIND times separate from any data rates

3590 High Performance Tape Subsystem ...

Ù **3590 Microcode Performance Patch**

A performance enhancement for chained READs and WRITEs has been developed and was integrated into any u-code (EC-) level 01/97.

Make sure you have that level for performance reasons

Ù **Reminder**

The following, taken from the announcement, again must be understood:

```
The actual throughput a customer may achieve
is a function of many components, such as

     - system processor
     - ESCON tape controller
     - associated drive configuration
     - data block size
     - data compressibility

and dependencies on other I/Os, such as
     - DASD   and the
     - system or application S/W used.

Usually, for single drive/single DASD,
the DASD represents the bottleneck
```

Ù **More Information**

- Magstar and IBM 3590 High Performance Tape Subsystem ITSO Red Book, GG24-2506-00, 04/95, 154 pages (Chapter 6, page 127-141 contains performance considerations)
- IBM 3590 High Performance Tape Subsystem Introduction and Planning Guide, GA32-0329-00 05/95, 82 pages
- IBM 3590 Tape Subsystem, Presentation Guide, G325-3306-01 (09/96), as G3253306 package on MKTTOOLA
- VSE/ESA Enhancements, Version 2.2, SC33-6629-00, 12/96

3591 High Performance Tape Control Unit

3591 High Performance Tape Control Unit

Announced 03/96, available since 05/96

„ **Rackmounted control unit 3591-A01**

„ **ESCON attached to S/390 processors**

1 ESCON channel per 3591

„ **Attaches 1 to 4 3590-B11 tape drives**

„ **Appears to S/W as '3490E'**

Í **Allows to use 3590 technology w/o new S/W**

Add in VSE/ESA as 'ADD cuu,3490E'

For more info refer to:

- IBM 3591 Introduction, Planning, and User's Guide, GA32-0558

3494 Tape Library Dataserver

3494 Tape Library Dataserver

Eliminates manual tape handling

Ù **Tape Library with modular coexistence of 3490E and 3590 tape drives**

```
3490E   C1A-, C2A-models  (1 to 16 drives)
3590    B1A       -models  (1 to 46 drives)
```

Ù **Cartridge accessor on a rail system**

Ù **From 8 up to 16 frames for flexible configuration/capacity**

- Up to 187 TB (compacted 3590)
- Up to 6240 cartridges

Ù **Cross platform support**

Ù **Common architecture with 3495**

Ù **1 Magstar Virtual Tape Server can be included**

Refer to next foil

3494 Virtual Tape Server

Magstar 3494 Virtual Tape Server (VTS)

Expands IBM tape automation

Ù **Consists of Virtual Tape Server Frame B16,**

3494 VTS Control Unit (RISC based) and
72 GB (formerly also 36 GB) of RAID SSA disk as Tape Volume Cache

adjacent to D12 Drive Frame

3 or 6 3590-B1A tape drives

Ù **Simulates 32 virtual 3490E drives**

2 3490E-A20 strings with 16 drives each

No physical tape drive needs to be allocated if underutilized, no waiting for a free drive

Ù **Nearly 100% 3590 cartridge capacity exploitation**

via volume stacking of logical volumes

Ù **Provides up to 50,000 logical volumes**

(400 (CST) or 800 MB (ECCST) each)

Ù **Total capacity of up to 40 TB per VTS**

Ù **Tape data are cached on disk**

- tape data are LRU cached, beneficial if re-used, fast random access with tape motion commands
- data movement to real tape is done later (after demount)
- used for all accesses to the virtual volumes
- 3590 physical cartridges are managed, using thresholds to determine when to consolidate partially full volumes

3494 and VTS

VSE Support

	3494		3495	
	w/o VTS	w/ VTS	w/o VTS	w/ VTS
VSE/ESA native	1.3.5+PTFa)	no b)	no	no
VM/VSE	1.3.5+PTF	1.3.5+PTF	no c)	no

- VSE tape mgmnt system highly recommended for usage reasons (EPIC/VSE, DYNAM/T, from CA and BVS ESA, from infosoft (VM/VSE))
- PTF is UD90367/90368 (APAR DY43306)

a) Native support provided by a LAN attachment to the 3494 Library Manager (Library Conytrol Path)
b) 3494-B16 has no LAN attachment, ESCON only
c) might work, but not supported

More info on 3494/VTS/VSE

Refer to

- IBM 3494 Tape Library Dataserver and VSE/ESA, 08/96, consists of 3 documents. As VSE3494 PACKAGE on IBMVSE tools disk
- IBM Magstar 3494 Tape Library, G325-3300-05, 09/96 Presentation Guide, as G3253300 package on MKTTOOLS
- IBM Magstar Virtual Tape Server, G325-3322-00, 09/96 Presentation Guide, as G3253322 package on MKTTOOLS

Appendix B: IOCP and Performance

PART O.

Appendix B: IOCP and Performance

For general information on IOCP, refer to

'Input/Output Configuration Program User's Guide', GC38-0401-07, 04/98 (includes Multiprise 2000 Internal Disk)

Refer also to APAR DY44630 (PTF UD50566) for VSE IOCP.

IOCP Introduction

IOCP Introduction

IOCP = Input/Output Configuration Program

- **Configures the Channel Subsystem of XA/370, ESA/370 or ESA/390 capable processors**
  E.g. 4381s, ES/9000s or 9672 CMOS servers
- **Uses for this task a source input deck, the**
  IOCDS = Input/Output Configuration Data Set
- **The IOCDS can be**
  - **modified with any text editor**
  - **generated (Build) by the IOCP**
  - **stored on the hard-disk of the service processor**
- **New IOCDS is active, after processor has been IMLed**
  (also known as POR = Power-On-Reset)
- **Several IOCDSs may be defined/saved, but only 1 IOCDS is active at any point in time**

IOCP Modifications

IOCP Modifications

To modify the IOCP (with any text editor), there are two possibilities:

- **Stand-Alone**
  - **included in processor microcode**
  - **to be edited on the Service Processor (SVP) of the ES/9000 or S/390 9672 processors**
  - **serviced via MES**

  To use the latest IOCP version, install the latest microcode level on the processor

- **Operating System based**
  - **included in MVS, VM or VSE**
  - **to be edited with e.g. VSE/ICCF, VSE/DWF**
  - **serviced via PTFs**

  Current IOCP version is 1.4 (1.5 for Multiprise)
  Introduced with DY43581 (VSE/ESA V2) or DY43491 (VSE/ESA V1)

IOCP Versions

IOCP Versions

The IOCP is available in different versions
(contained in VSE/ESA base since 1.3):

- **IXP IOCP**

  Required for ES/9221 integrated adapters (e.g. ICA)

  Refer to GC38-0097

- **IZP IOCP**

  Latest version, required for 9672 CMOS servers,
  for ESCON, EMIF,
  not usable for ES/9221 integrated adapters

  Refer to GC38-0401

- **Which version to be used?**

  Stand-Alone:
  - correct version is used since in microcode
  Must be used on a newly installed machine

  VSE based:
  - use IXPIOCP or IZPIOCP in EXEC statement

  Error messages show which version is used :
  Error msg IXPxxxx - IXPIOCP used
  Error msg IZPxxxx - IZPIOCP used

Hardware Positioning of the IOCDS

Hardware Positioning of the IOCDS

Assignment of ... to ... | Change in Assignm. requires...

VSE ADD | cuu > VSE | IPL

VM | cuu > Guest (ATTACH, DEDICATE) | Guest IPL/SET RDEV

LPAR image (optional) | Storage, Mode CHPID > 370 channel (S/370 mode only) | LPAR Activation

IOCDS | cuu > CHPID | IML (POR)

A
Customer
IBM Hardware SE
V

H/W Configuration | CHPID > Hardware | Power Off/On

IOCDS Tailoring and System Down Time

IOCDS Tailoring and System Down Time

System down time
Stand Alone VSE based
* *

IPL
A

LPAR Activation Optional
A A A

IML
A *
no errors
\>
errors

IOCP Generation (Build) at SVP EXEC IZPIOCP
*
A

Edit IOCDS SVP Editor ICCF Editor
A
<

Multiple Channel Paths to 1 Device

Multiple Channel Paths to 1 Device

- **ESA Channel I/O Subsystem provides DPS and DPR between up to 4/8 channel paths**

- **In S/370-mode, S/W must manage 'Alternate Pathing'**

  Between 2 'subsequent' channels, specified via
  ADD cuu (S),type

- **S/370 LPARs convert SIOF requests to SSCHs, thus providing alternate pathing as for ESA-mode**

VSE-mode	S/370-mode		ESA-mode
Processor	w/o IOCDS	w/ IOCDS	w/ IOCDS
Example	9370	4381-9xE, 9672 LPAR	9x21, ES/9000, 9672
VSE I/O	SIOF	SIOF (*)	SSCH
Max #paths	2, via S/W	4/8 via IOCDS	4/8 via IOCDS
VSE ADD	ADD cuu(S),3380	ADD cuu,3380	ADD cuu,3380

(*) S/370 LPAR converts SIOF to SSCH
- ADD 3380 if possible as ECKD

IOCP Statements

IOCP Statements (Macroinstructions)

- **ID**

  Optional heading for output listings

- **CHPID**

  Describes (physical) channels/channel paths

```
- Characteristics  (Byte/Block/ESCON-channel,
                    ESCON CTC-connection, Int.Disk SCSI bus)
- Relates channel paths to channel numbers/channel sets
```

- **CNTLUNIT**

  Describes control unit images associated to the channel paths

```
- Characteristics of the control unit image
- Channel paths that can be used to reach the CU image
- Unit addresses that the control unit image recognizes
- Channel protocol used (DCI/Streaming 3MB/Streaming 4.5MB/...)
```

- **IODEVICE**

  Describes (logical) devices at the control units

```
- Device characteristics
- Control units to which the device is attached
- Device address number (must be in range 000 - FFF for VSE)
```

If the devices seen by the S/390 S/W are not simulated (e.g. RAMAC), the logical devices are also physical devices.

A similar consideration applies to control unit images, which may be e.g. a real 3990 Storage Cluster or a simulated one (RAMAC e.g.).

(NOTE that in IOCP terms a 'logical control unit' is different, i.e. a set of all control unit images that physically or logically attach I/O devices in common)

Example 1: Multiple Paths to a Device

Example 1: Multiple Paths to a Device

```
                                         3990          3390s
(Uni  or      CHPID 20                   CL 0           DASDs
N way)
Processor     CHPID 21                                  100
(ESA mode)                                               11F
              CHIPD 22                   CL 1

              CHIPD 23
```

- **IOCDS**

a) 4.5 MB Parallel Channels

```
CHPID     PATH=(20,21,22,23),TYPE=BL
*
CNTLUNIT  CUNUMBR=001,PATH=(20,21),UNITADD=((00,32)),    X
          SHARED=N,PROTOCL=S4,UNIT=3990
CNTLUNIT  CUNUMBR=002,PATH=(22,23),UNITADD=((00,32)),    X
          SHARED=N,PROTOCL=S4,UNIT=3990
*
IODEVICE  ADDRESS=(100,32),CUNUMBR=(001,002),UNIT=3390
```

b) ESCON Channels

```
CHPID     PATH=(20,21,22,23),TYPE=CNC
*
CNTLUNIT  CUNUMBR=001,PATH=(20,21),UNITADD=((00,32)),    X
          UNIT=3990
CNTLUNIT  CUNUMBR=002,PATH=(22,23),UNITADD=((00,32)),    X
          UNIT=3990
*
IODEVICE  ADDRESS=(100,32),CUNUMBR=(001,002),UNIT=3390
```

- **ADD statements in VSE/ESA**

```
ADD 100:11F,ECKD
```


Example 2: 3490 Tape Attachment

Example 2: 3490 Tape Attachment

3490 with 2 Paths and 4 Drives, plus Preferred Paths

```
                                         3490 Subsystem
(Uni or      CHPID 20           A                   130
N way)
Processor                                           131
(ESA mode)
             CHIPD 21           B                   132
```

- **IOCDS**

a) 4.5 MB Parallel Channels:

```
*
CHPID     PATH=(20,21),TYPE=BL
*       1st Channel
CNTLUNIT  CUNUMBR=001,PATH=(20),PROTOCL=S4,SHARED=N,          X
          UNITADD=((30,16)),UNIT=3490
*       2nd Channel
CNTLUNIT  CUNUMBR=002,PATH=(21),PROTOCL=S4,SHARED=N,          X
          UNITADD=((30,16)),UNIT=3490
*
IODEVICE  .... as shown below
```

b) ESCON Channels:

```
*
CHPID     PATH=(20,21),TYPE=CNC
*       1st Channel
CNTLUNIT  CUNUMBR=001,PATH=(20),UNITADD=((30,16)),UNIT=3490
*       2nd Channel
CNTLUNIT  CUNUMBR=002,PATH=(21),UNITADD=((30,16)),UNIT=3490
*
*          3490   130 133 plus preferred paths
IODEVICE  ADDRESS=(130,1),CUNUMBR=(001),UNIT=3490,STADET=N,PATH=20
IODEVICE  ADDRESS=(131,1),CUNUMBR=(002),UNIT=3490,STADET=N,PATH=21
IODEVICE  ADDRESS=(132,1),CUNUMBR=(001),UNIT=3490,STADET=N,PATH=20
IODEVICE  ADDRESS=(133,1),CUNUMBR=(002),UNIT=3490,STADET=N,PATH=21
*
```

- **ADD statements in VSE/ESA**

```
ADD 130:133,3490
```


Example 3: Alternate Pathing

Example 3: Alternate Pathing

Multiple 4.5M Parallel Paths to a Device from 2 CPUs

```
(Uni or N way)                                   (Uni or N way)
     CPU 1                                            CPU 2
   (ESA mode)
                                                              LPAR1
                                                   (S/370 mode)

      CHPID's                                              CHPID's

21  22  23  24                                     26  27        LPAR
                                                                 image
                                                          S/370 channels

                                                              6     7



                                 3990                                       3990
                 Storage Cluster 1   Storage Cluster 2



                                 3390
                          (cuu : X20 X3F)
```

Example 3: Alternate Pathing ...

Example 3 - Definitions CPU1

„ **IOCDS**

```
*           CHPID's
CHPID     PATH=(21),TYPE=BL
CHPID     PATH=(22),TYPE=BL
CHPID     PATH=(23),TYPE=BL
CHPID     PATH=(24),TYPE=BL
*
*           3990
*               Storage Cluster 1
CNTLUNIT  CUNUMBR=001,PATH=(21,23),PROTOCL=S4,SHARED=N,         X
          UNITADD=((20,32)),UNIT=3990
*
*               Storage Cluster 2
CNTLUNIT  CUNUMBR=002,PATH=(22,24),PROTOCL=S4,SHARED=N,         X
          UNITADD=((20,32)),UNIT=3990
*
*           3390   X20 X3F
IODEVICE ADDRESS=(120,32),CUNUMBR=(001,002),UNIT=3390,STADET=N
```

„ **ADD statements in VSE/ESA (ESA mode)**

```
Note:
 In ESA mode the alternate pathing is done by the Channel Subsystem!

    ADD 120:13F,ECKD,SHR        Use SHR only if shared disks used !
```


Example 3: Alternate Pathing ...

Example 3 - Definitions CPU2

„ **IOCDS**

```
Note:

VSE/ESA can run on the ES/9000 (except early ES/9221 machines) and
9672 processors in S/370 mode only in an LPAR !

The LPAR image, defined at the service processor, has to reflect
the mapping of CHPIDs to S/370 channels !
```

```
*               CHPID's for LPAR1
CHPID     PATH=(26),TYPE=BL,PART=(LPAR1,REC)
CHPID     PATH=(27),TYPE=BL,PART=(LPAR1,REC)
*
*               1st Storage Cluster 3990
CNTLUNIT  CUNUMBR=001,PATH=(26),PROTOCL=S4,SHARED=N,            X
          UNITADD=((20,32)),UNIT=3990
*
*               2nd Storage Cluster 3990
CNTLUNIT  CUNUMBR=002,PATH=(27),PROTOCL=S4,SHARED=N,            X
          UNITADD=((20,32)),UNIT=3990
*
*           3390   X20 X3F
IODEVICE  ADDRESS=(620,32),CUNUMBR=(001,002),UNIT=3390,STADET=N
```

„ **ADD statements in VSE/ESA 1.3./1.4. (S/370 mode)**

```
Note:
    In S/370-mode alternate pathing is done by the operating system

 ADD 620:63F(S),ECKD  This definition requires that the LPAR is
                      defined to map CHPID 26 to S/370 channel 6
                      and            CHPID 27 to S/370 channel 7
```


IOCP and Performance -Rules-

IOCP and Performance

Note:
All comments and examples are valid for S/390 ES/9000 processors or 9672 servers.
They are also valid for ES/4381 processors like the 4381-P13 (XA/370) or the 4381-9xE (ESA/370) processors.

„ **Wrong IOCDS definitions can cause massive performance degradation**

to be observed in terms of
- **high I/O service times**
- **high Online response or Batch elapsed times**
- **more CPU-time**

„ **Also functional problems may occur**

General Rules for Correct IOCDS Definitions:

„ **Use the right VSE based version of IOCP**
Refer to chart on IOCP versions

Dependent of the mode of VSE/ESA (S/370 or ESA mode) ...

„ **Define the correct statements in the IOCDS and the VSE IPL procedure**
Refer to Example 3

Before changing an IOCDS ...

„ **Take a look into Appendix D of**

'I/O Configuration Program - User's Guide',
GC38-0097 (IXP version) or GC38-0401 (IZP version)

The rules listed there must always be followed

IOCP and Performance -Rules- ...

Specific Rules for CHPID

„ **Up to 8 chpids in each CHPID macro possible**
„ **For parallel channels,**
Specify always (if possible) CHPID TYPE=BL.

Use TYPE=BY only if really needed
(e.g. for RSCS connections)

Specific Rules for CNTLUNIT

„ **Specify exactly 1 CNTLUNIT statement for each 'control unit image' in a physical CU box**
Control unit images (having a unique CUNUMBR):

- 3880 Storage Director (SD)
- 9343 Storage Cluster (CL)
  (1 per 9343-CC2, 2 per 9343-CC4 or -DC4)
- 3990 Storage Cluster (CL)
  (2 per 3990-6 or 9390-001, 4 per 9390-002)
- RAMAC Array Subsystem Cluster (CL)
  (2 per 9394-001 or -002, 4 per 9394-002)
- 3490/3490E Channel Attachment (CA)

This recommendation assures automatically that successive connection attempts are to another control unit image ('ping-pong' between storage control clusters)

„ **For parallel channels,**
Specify always (if possible) SHARED=N.
SHARED=N allows multiple concurrent I/O requests

Use SHARED=Y only if really needed:
e.g. for - 3420 tape units
- 3x74 controllers
  (SHARED=N may result in CPU-time overhead)

IOCP and Performance -Rules- ...

Specific Rules for CNTLUNIT (cont'd)

- **For parallel channels,**
  Specify always (if possible) PROTOCL=S4.
  Use PROTOCL=S only if really needed:
  e.g. for - 4381
  (S for 3.0 MB/sec and S4 for 4.5 MB/sec parallel channels)
- **Some parameters in the CNTLUNIT statement are dependent on the settings in the control unit**
  e.g. PROTOCL, UNITADD
- **Specify via UNITADD the same number of devices as the H/W CE has set in the control unit**
  Occuring are powers of 2 (control unit dependent):
  2, 4, 8, 16, 32 or 64
  - > If controller is set for more devices than in IOCDS, lost interrupts require extra CPU-time for error recovery
- **IODEVICEs may be added as 'look-ahead' in the IOCDS**
  Avoids a separate IOCP build before a new device is attached.
  A separate VSE IPL is sufficient
- **Any IODEVICE which is NOT actually attached must NOT be ADDed in VSE**
  This assures that VSE never will issue any I/O request to that non-existing device

IOCP and Performance -Rules- ...

Specific Rules for IODEVICE

- **Each physical/logical I/O device must be represented by exactly 1 IODEVICE statement**
  A logical I/O device is e.g. a RAMAC simulated device.
  A single IODEVICE statement may represent several consecutive I/O devices
  Alternate Pathing in S/370-mode is an exception (see Example 3)
- **Specify in the ADDRESS parameter the same number of devices as the H/W CE has set in the control unit**
- **If 3490/3490E are defined, preferred paths should be defined via IODEVICE PATH=chpid**
  (see Example 2)

Enterprise Storage Server (ESS)

PART P.

Enterprise Storage Server (ESS)

First announced 99-07-27.

ESS performance information together with FlashCopy
was added in the new VSE/ESA V2.5 performance document.

Refer also e.g. to

- the ESS announcement letter, dated 99-07-27
- the ESS home page
  http://www.ibm.com/storage/ess
- IBM ESS Introduction and Planning Guide, GC26-7294
  available via the URL above
- IBM ESS Performance White Paper
  Version 1.0, 69 pages, by John Ponder et al.
  Via ESS home page

EOD (End of document)

HAND (Have a nice day)