IBM VSE/ESA
VM Guest Performance Considerations

Wolfgang Kraemer

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

Update 2001-07-15

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 document

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

This document is unclassified and intended for VSE customers.

This document is also available from the INTERNET via the VSE/ESA home page

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

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

Starting with VSE/ESA 2.4, the documents are also available on the VSE/ESA CD-ROM kit SK-2T0060 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 Performance Considerations'
'IBM VSE/ESA VM Guest Performance Considerations' (this document)
'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 2.5 Performance Considerations'
'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.

Acknowledgements

Thanks to all who contributed directly or indirectly, be it by measurements, suggestions or in other ways.
Specific thanks is due to Bill Bitner, VM/ESA Development, Endicott.

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

Please, as in the past, contact me 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

IBM VSE Development, Boeblingen Lab, Germany

Notes ...

Base Documents

This document essentially deals with VSE/ESA Guest performance aspects. It applies to all VSE/ESA releases, especially in ESA-mode under VM/ESA.

It comprises those charts formerly in the VM/VSE part of the VSE/ESA 1.3/1.4 Performance Considerations document.

The VSE/ESA performance documents (see a previous foil) are o available via the VSE/ESA home page and on the VSE/ESA CD-ROM kit.

VSE/ESA V2 Turbo Dispatcher under VM/ESA

VM/VSE performance considerations specific to the VSE/ESA V2 Turbo Dispatcher (on n-ways) are contained in the Turbo Dispatcher document.

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
Nway	CICS	zSeries	z/VM	ESS

Trademarks of other companies:

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

References

Some General VM/VSE References

The following are some references for further information in the context of VM/VSE guest setup and performance.

VM/ESA, 'CP Command and Utility Reference', SC24-5773

VM/ESA, Running Guest Operating Systems
Release 1.2.0, SC24-5522-02
Release 2.1.0, SC24-5755-00

VM/ESA, Planning and Administration
Release 1.2.1, SC24-5521-03
Release 2.2.0, SC24-5750-01

VSE/ESA under VM/ESA on an ES/9000
09/94 by Steve Lampasona, VSE Techn. Conf. Philadelphia

Exploiting VM/ESA Facilities for VSE/ESA
ITSO Boeblingen Red Book, 12/95, 77 pages, SG24-4678-00

VM/ESA Performance Monitoring Tools
ITSO Poughkeepsie, GG24-4152, 12/93, 281 pages
(VMPRF, RTM/ESA, VMPAF, FCON/ESA, EXPLORE/VM (tm), XAMAP (tm))

VM/ESA Storage Management with Tuning
GG24-3934-00, ITSO Red Book

The Value of VM for the VSE Enterprise, 09/98,
by James M. Savoie, IBM, WAVV 98, Albany, NY

You may also check the following URL
http://www.vm.ibm.com/devpages/bitner/presentations/vmvseprf.html
http://www.vm.ibm.com/perf/

VM/VSE Performance References

VM/ESA, Performance
Release 1.2.1, SC24-5642-01,
Release 2.1.0-2.3.0, SC24-5782-0x

'VM/ESA 1.1 CCW Translation Performance Improvements'
Washington System Center Flash 9220

References ...

VM/VSE Performance References (cont'd)

VM/VSE Performance Hints & Tips
ITSC Boeblingen, GG24-4260, 03/94, 110 pages

Evaluating 'EXPLORE/VSE' in a VM/VSE Environment
ITSC Boeblingen, GG24-4261-00, 09/94, 89 pages

VM/ESA Performance Reports:

VM/ESA Release 1.1.1 Performance Report, GG66-3236
VM/ESA Release 1.2.0 Performance Report, GG66-3245
VM/ESA Release 1.2.1 Performance Report, GC24-5673-00

VM/ESA Release 1.2.2 Performance Report,
VM Performance Group Endicott, 224 pages.

As VM22PERF PACKAGE on MKTTOOLS disk, latest update 94-07-15,
equivalent to GC24-5673-01

VM/ESA 2.1.0 Performance Report, GC24-5801,
VM Performance Group Endicott, 159 pages.

As VM210PRF PACKAGE on MKTTOOLS disk, latest update 95-09-15,
equivalent to GC24-5673-01

VM/ESA 2.2.0 Performance Report, GC24- ,
VM Performance Group Endicott, 96 pages.

As VM220PRF PACKAGE on MKTTOOLS disk, latest update 96-12-11,
equivalent to GC24-5673-01

VM/VSE Tuning and Performance
09/94 by Dan Janda and Richard Lewis, VSE Techn. Conf. Phila.
As TC94VMVS package on IBMVSE tools disk

'Is VM Good for VSE Performance?' by Bill Bitner,
VM/ESA and VSE/ESA Techn. Conf., Orlando 06/96, Rome 10/96;
and VSE/ESA Software Newsletter 3rd/4th Quarter 1998, p68 ff

VM/VSE From Both Sides Now - VSE Native, VSE/VM or LPAR,
Dan Janda and Richard Lewis IBM Endicott/Gaithersburg,
WAVV 09/98 Albany, NY.

VM for VSE Guest Performance, by Bill Bitner
VM/VSE Tech Conf, Orlando 05/99, Mainz 06/99.
WAVV 10/99 Cincinatti/Ft Mitchell, KY
VM/VSE Tech Conf, Orlando 06/00

Glossary

VM/VSE Performance References (cont'd)

'VM/ESA 2.3.0 Performance Update' by Bill Bitner, VM/VSE Technical Conference, Frankfurt, Germany, 03/98
VM/VSE Technical Conference, Reno, Nevada, 05/98, Session 26A

VM/ESA Running Guest Operating Systems: Version2 Release3, SC24-5755

Glossary

CFP	CCW Fast Path or Fast VM CCW translation A shorter path that saves CP CPU-time
ECKD	Extended Count Key Data An architecture for I/O commands (CCWs)
DIM	Data in Memory A concept to store as much data as possible/reasonable in processor storage
FPM	Full Pack Minidisk A method to define/use logical drives under VM
PPM	Partial Pack Minidisk A method to define/use logical drives under VM
MDC	Minidisk Caching Full track caching of minidisks in VM/ESA
ITR	Internal Throughput Rate A measure for processor and/or S/W effectivity: #transactions or batch jobs per CPU-second
MPG	Multiple Preferred Guest A function on ES/9000 processors, providing improved VM/ESA V=R/F guest support via PR/SM
LPAR	Logical Partition PR/SM partitions system into LPARs
PR/SM	Processor Resource Systems Manager An ES/9000 standard feature for logical partitioning
SIE	Start Interpretive Execution A S/390 instruction used for VM guests
VSCR	Virtual Storage Constraint Relief All that provides effectively more space below the 16 MB line

Introduction and Overview

PART A.

Introduction and Overview

Performance Value-Add of VM

Reasons to run VSE under VM (Performance View)

„ **VSE capacity reasons**

Capacity of a single VSE does not suffice

Run several VSE systems on the same processor complex under VM, w/o using PR/SM LPARs.

Applicable to

VSE/SP	mostly
VSE/ESA 1.1/1.2	often
VSE/ESA 1.3/2.1	sometimes

„ **Workload balancing reasons**

VSE dispatching not as sophisticated as in VM

VSE does not provide

- weighting factors for partition dispatch (except with VSE/ESA 2.2 Turbo Dispatcher in the PB group)
- capping of CPU usage
- flexible and controllable real storage allocation

„ **Resource sharing reasons**

PR/SM does not provide such a flexible resource sharing

- Real storage is shared for V=V guests
- Channels are shared even w/o EMIF
- DASD devices can be split up into minidisks

Performance Value-Add of VM ...

„ **Paging performance reasons**

VM CP provides 'better' paging algorithms

An argument used by VSE customers.
Also, VM Page Data Set is more flexible than VSE's 1:1 mapping of pages to disk space

„ **H/W exploitation reasons**

VSE does not support certain hardware

- Multiple processors (before VSE/ESA 2.1)
- Expanded storage (if configured as such)
- Pinned data for DASD Fast Write (before 2Q95)
- Real CTCA usage by POWER
- ESCON Manager

„ **Performance enhancement reasons**

VM CP provides additional 'performance functions'

Usage of
- VM Fulltrack Minidisk Caching (MDC)
- VM Virtual Disk for the VSE Lockfile
- SQL/DS Guest Sharing

„ **Some further (non-performance) reasons**

CMS applications, VM specific functions

Separation of production and test

Migration vehicle

Flexibility of assigning hardware resources (device addresses, configuration, IOCDS changes)

Sharing of FBA volumes between VSEs

VSE DASDs under VM

VSE DASDs under VM

VSE DASDs under VM can be defined as

- **VM Minidisks (partial pack) ('PPMs')**
- **VM Full Pack Minidisks ('FPMs')**
- **DEDICATEd devices ('DEDs')**

	PPMs	FPMs	DEDs
Directory Statements	MDISK +MINIOPT	MDISK +DASDOPT	DEDICATE +DASDOPT (or ATTACH)
Eligible for MDC	yes	yes, via MDISK volid	no
Eligible for IOASSIST	no	no	yes
Eligible for VM CCW-X-lat. Bypass	no	V=R only	yes
A guest can change - device level fctns - subsystem lvl fctns	 no no	 yes(DEVCTL) yes(SYSCTL)	 yes(DEVCTL) yes(SYSCTL)
- MINIOPT for CACHE and MDC specification - DASDOPT for DEVCTL/SYSCTL specification			

1. VM Minidisks (partial pack) ('PPMs')

Such VM minidisks are being defined in the user directory in the MDISK/MINIOPT statement(s) and are LINKed to the user.

For any minidisk under VM you have to specify the extents (i.e. the real begin track and the number of tracks) in the MDISK statement. On virtual track 0 the minidisk has the VOL1 label.

Only when a minidisk starts at real cylinder 0 it is IPLable by VSE in a native environment (IPL record, VOL1 label and VTOC pointer).

> Make sure that NOCACHE is not set erroneously in the MINIOPT directory control statement, since VM would change all DEFINE EXTENT CCWs to Bypass Cache

VSE DASDs under VM ...

VSE DASDs under VM (cont'd)

2. VM Full Pack Minidisks ('FPMs')

VM full pack minidisk are defined similarly as normal VM minidisks with

- MDISK ... volid from cyl 0 to end
or - MDISK ... DEVNO but then not eligible for MDC

In addition, the DASDOPT directory statement can be used to specify the levels of control (DEVCTL, SYSCTL), mostly for DASD caching.

In contrast to DEDicated devices, several VM guests can be linked to the same full pack minidisk.

A full pack minidisk can have been created also with native VSE.

3. DEDICATEd devices ('DEDs')

DEDICATEd devices are in the user directory or are being ATTACHed by the operator to the user.

They
- usually benefit from full I/O assist
- are not shareable between VM guests
- are not eligible for VM MDC

VM/VSE Guest Setup (Summary)

VM/VSE Guest Setup (Summary)

„ **Optimally select guest definitions/setup**

Determine how much real storage can be dedicated

Run preferred guest(s) (V=R or V=F), especially with DEDicated devices

Run V=V only if required, or with many partial pack minidisks

Select guest performance parameters deliberately

Refer to next foil

„ **Optimally select DASD definitions**

Refer to Part B 'VM/VSE Guest DASD I/O Setup'

Dedicate as much DASDs as possible

This is beneficial for preferred guests

Avoid minidisks, but avoid at least shared DASD, where possible

But VM Full Track caching (MDC) needs minidisks

VM/VSE Guest Setup (Summary) ...

VM/VSE Guest Setup Performance Settings

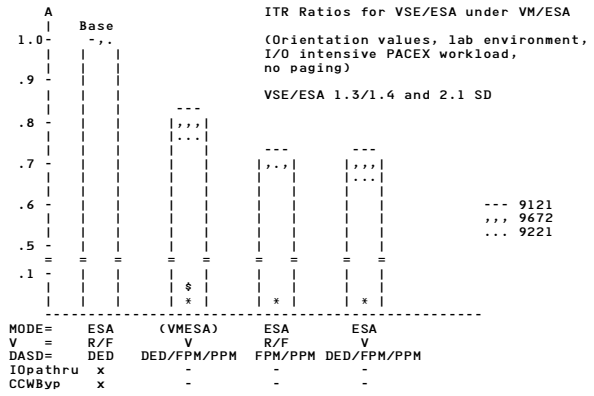
„ **Optimally set guest performance parameters in CP**

CP Command	Performance function for guest(s)	Note	Guest
SET IOASSIST ON	Enables I/O passthru (SIE assist)	a C	R F -
SET CCWTRAN OFF	Disable VM CCW translation	a C	R - -
SET PAGEX ON	Enable pseudo-page-fault facility	a DS	- - V
SET QUICKDSP	Bypass eligible list for key server machines	a D	R F V
SET SHARE	Sets priority weights (ABSolute, RELative, LIMITSoft, LIMITHard)	b D	R F V
SET RESERVED	Reserves real page frames for key V=V QUICKDSP machines	b S	- - V
LOCK	Fixes specified guest pages (Better: use RESERVE command)	b S	- - V
SET SRM STORBUF	Defines usage (%) of page pool DPA	b S	- - V
LDUBUF	Defines usage (%) of paging devices	b	- - V
DSPBUF	Limits #guests in dispatch list (Do not use in general)	b D	R F V
DSPSLICE	Size of dispatch time slice (Do not use in general)	b D	R F V
MAXWSS	(")	b S	- - V
DEDICATE	Dedicate a real processor to a virtual processor of a guest	b CD	R F V

Notes
- a Option may be beneficial in any case and for any number of VSE guests
- b Option only allows to prefer a specific VSE guest at cost of other VM tasks (VSE guests, CMS users)

The following are primary performance effects:
- C Impact on CP CPU-time (VM overhead)
- D Impact on dispatching/balancing
- S Impact on storage/paging

VM/VSE Guest ITR Comparison



- DED= Dedicated devices, FPM/PPM full or partial pack minidisks

* These cases assume CCW Fast Path (CFP), not available for VM/ESA 1.1.0, optional for VM/ESA 1.1.1, standard for >1.1.1, including VM 57265 for ECKD.

$ MODE=VMESA values for VSE/ESA <1.3 are .87/.79, (with VSE CCW translation for VSE/ESA 1.3 VMESA). VMESA n/a for VSE/ESA V2

-> V=R with dedicated devices gives best guest ITR
(only minor deltas between V=R and V=F)

-> V=R/F in general beneficial only with dedicated devices
(dedicated devices in general beneficial only with V=R)

-> For V=V the type of DASD (dedicated/minidisks) is unimportant

-> For average I/O intensive VSE workloads,
ITR ratios are closer to 1.0 and deltas smaller.

VM/VSE Guest Comparison

VM/VSE Guest Comparison

Ù **Mode Comparison for VSE Guests**

	V=R/F 'Preferred Guests'	V=V
When recommended?	For best ITR, use with DEDicated devices	If storage cannot be reserved, or V=R/F no more available
Paging done by	VSE only (Nobody, if VSE V2 NOPDS option used)	VM + VSE *) (VM Only, if DEF STOR >VSIZE, consider the VSE V2 NOPDS option) **)
Page Fault Reflection	Not applicable	Yes, if PAGEX ON
VSE 'Real' Size	Size of V=R/F area (Real real storage, not shared)	VM DEF STOR (Simul.real stor., 'shared')
VM CCW Translation I/O Passthru	Next charts "	Next charts "

*) Refer to charts on Double Paging
**) NOPDS in VSE does NOT save any CPU-time in VSE/ESA, just the need for the VSE Page Data Set. Refer to the VSE/ESA V2 Performance doc.

For MODE=ESA supervisor (the only VSE mode since VSE/ESA 2.1):

- Several VSE address spaces possible, and data spaces
- VSE runs with DAT ON (even for NOPDS)
  VSE CCW translation is always required.
- VSE virtual size = VSE VSIZE

VM/VSE Guest DASD I/O Setup

PART B.

VM/VSE Guest DASD I/O Setup

VM/VSE I/O Related Terms

Some Performance Related Terms for VM/VSE

„ **I/O Passthru**

I/O Passthru (or 'I/O Assist' or 'SIE Assist')
is the ability that under certain conditions, I/O interrupts for guests do not cause CP interrupts and thus avoid VM overhead in such cases. On ES/9000 processors this is enabled by microcode.

Only dedicated devices (DASDs) of V=R or V=F guests are eligible for I/O passthru.

For additional info, not covered here, refer e.g. to
'VM/ESA Performance, SC24-5782-00
Chapter 3, under 'I/O Interpretation Facilities'.

Contains requirements for I/O Assist, reasons for dropping out of I/O Assist, e.g.
- SET IOASSIST OFF issued
- TRACE command active
- SET CCWTRAN ON issued
- SET RUN OFF in effect
- one or more virtual CPUs are in stopped state

and information on available data.

„ **VM CCW Translation Bypass**

VM CCW Translation Bypass (formerly often 'I/O Fast Path')
allows that under certain conditions the CCW translation by VM can be bypassed.

This can be done for V=R guests for dedicated devices and, in limited cases, also for full-pack minidisks. It can be controlled by the SET CCWTRAN command.
For a V=R machine the default for CCWTRAN is OFF. For a V=F machine CCWTRAN can not be set OFF. If you SET CCWTRAN ON for a V=R guest, then the benefit of I/O Assist is lost.

„ **Fast VM CCW Translation (CFP)**

Fast CCW Translation (or CCW Fast Path CFP)
is a CP feature where CP will use a more efficient path for the translation of virtual to real addresses associated with CCWs and for untranslation (real to virtual). It is only available for DASD I/O. It should not be confused with CCW Translation Bypass.

VM CCW Transl. Bypass and I/O Passthru

VM CCW Translation Bypass & I/O Passthru

DASD device	V=R Guest: No VM CCW translation	V=R Guest: I/O passthru	V=F Guest(s): No VM CCW translation	V=F Guest(s): I/O passthru
DED (dedic.)	X	X *2 *3	X	X *2 *3
FPM (full pack)	X *1	-	-	-
PPM (part.pack)	-	-	-	-

X means full benefit, - means no benefit/not applicable

*1 If additional conditions are fulfilled,
e.g. channel program starts with a 'valid' CCW sequence
SEEK, SET FILE MASK (CKD)
DEF EXT, LOC REC (ECKD)
DEF EXT, LOCATE (FBA)

*2 Virtual and real subchannel number must be equal
*3 PTF VM56244 avoids I/O assists off after ATTACH/DETACH

V=V guests: VM CCW translation always required
No I/O passthru (SIE assist)

- General assumptions:
  - ES/9000 with PR/SM for I/O passthru (I/O Interpretation Facilities)
  - VM/ESA ESA, only in first level (not in LPAR, not under VM)
- Some assumptions regarding VM CCW translation bypass:
  - No CP I/O or CCW trace for device was activated
  - CP has not a pending Sense for the device
  - No SET CCWTRAN ON (SET NOTRANS OFF) was issued for V=R (CCWTRAN OFF is default for V=R, unless not first level). SET CCWTRAN ON is not accepted for V=F, since option is always in effect.
- Some assumptions regarding I/O passthru:
  - No SET CCWTRAN ON was issued for V=R guests (which also causes V=R I/O passthru to be lost)
  - No SET IOASSIST OFF was issued (IOASSIST ON is default for all V=R/V=F guests)

For more info refer to SC24-5642-01 (VM/ESA Performance, Rel. 1.2.1)

VM/VSE I/O Related Hints

Benefit optimally from I/O Passthru

Ù **Ensure, DEDicated DASDs are/stay within I/O Assist**

„ **Without a VM monitor**

Query IOAssist

- ALL USERS SET must be ON
- IOASSIST SETTING must be ON for all V=R and V=F guests i.e. user must be eligible
- IOASSIST ACTIVE YES means that IOASSIST is used (a snapshot)

Query Virtual DAsd DETAILS

- Each vdev for a real device must be IOASSIST ELIGIBLE

The IND USER display for the number of total I/Os only show non-I/O-assisted I/Os (e.g. to console).
If this counter is increasing too fast, it may be an indication that some DASDs are not in IOASSIST

„ **With a VM monitor**

Take a look into

- RTM Device Display device list: 'IN I/O PASSTHRU'
- VMPRF report for IOASSIST: PRF098
- FCON/ESA monitor

Also, if I/O assisted, all DASD I/O counters for non-cached devices are reported as 0. (I/O counters for cached devices are retrieved from the control unit).

VM/VSE I/O Related Hints ...

Avoid/reduce CP Overhead for CCW Translation

CCW translation for guest virtual machines costs CP CPU-time.

VM provides means

- to avoid completely this overhead
or
- to reduce the CP overhead considerably.
(VM/ESA 1.2.1 and to a certain extent previous version/releases)

„ **Avoid VM CCW Translation**

1. I/O to dedicated DASD devices of V=R or V=F guests are handled by SIE assist (if available on the hardware). In this case CP does not even see the I/O (I/O Passthru), since u-code gets I/O interrupt.
2. I/O to dedicated DASD devices of a V=R guest are not subject to VM CCW translation, if CCWTRAN OFF (or NOTRANS ON) is set or in effect. This is default for V=R (a V=F system has always CCWTRAN ON!).

„ **Enhance VM CCW Translation (via CFP usage)**

CFP applies only to DASD-I/Os, not tape or CTC or other.

CFP is used for minidisks (full and partial pack) and for dedicated devices if SIE assist is not active for that device and CCWTRANs ON is set for the virtual machine.

CFP requires CCWTRAN ON in a virtual machine.

For V=V and V=F virtual machines this is always in effect. For V=R machines you can change the default CCWTRAN OFF to CCWTRAN ON, but then SIE assist is switched off for all (!) devices of that virtual machine.

VM (Fast) CCW Translation

VM (Fast) CCW Translation

DASD device	V=R	V=F	V=V Guest
DED (dedic.)	n/r	n/r	X CFP
FPM (full pack)	X*1 CFP*2	X CFP	X CFP
PPM (part.pack)	X CFP	X CFP	X CFP

X VM CCW translation required

CFP VM CCW translation with Fast Path attempted

n/r no VM CCW translation required (CCWTRAN OFF)

*1 If channel program does NOT start with a 'valid' CCW sequence (see previous chart)

*2 CFP usage is limited before VM/ESA 1.2.1

- To exploit CFP, it is required (e.g.):
  - Device not secondary of a duplexed pair.

  V=R: SET CCWTRAN ON (But this sets IOASSIST OFF)

  V=F: CCWTRAN always in effect, incl. I/O assists

  -> If you have both DEDicated devices and Minidisks, V=F is an alternative to V=R

„ **CFP PTFs**

	VM51012 CKD	VM57265 ECKD
VM/ESA 1.1.0	no	no
VM/ESA 1.1.1	opt	opt
VM/ESA 1.2.0	'std'	opt, called VM57443
VM/ESA 1.2.1	'std'	opt, " "
VM/ESA 1.2.2 (and newer)	'std'	'std'

Under VM/ESA 1.2.2, make sure that you have VM PTF UM27166 applied (APAR VM59317). This PTF avoids that when using record access and regular data format bit settings in a VSE guest, CFP is aborted.

VM (Fast) CCW Translation ...

VM Fast CCW Translation (cont'd)

„ **Fast CCW Translation counters**

- RTM/ESA Display System screen:

  FTR_ABOR, FTR_DONE, FTR_NE, and FTR_TOTL

- VMPRF 1.2 (APAR VM59889):

  System_Facilities_By_Time report (PRF104)

  CFP counters in VM monitor records added in VM/ESA 2.1.0

  Refer e.g. to
  'What's new in VMPRF' by Bruce Dailey
  Technical Report TR01.G020, October 1995

- FCON/ESA:

  Screen 38: Minidisk Cache

- By CP LOCATE HCPFTRRW, pointing to 3 fullword running counters:

  - # of fast CCW translations done
  - # of fast CCW translations started and then aborted
  - # of CCW translations, not eligible for CFP

Info on VM Monitors

VM/ESA Performance Monitoring Tools
ITSO Poughkeepsie, GG24-4152, 12/93, 281 pages
(VMPRF, RTM/ESA, VMPAF, FCON/ESA, EXPLORE/VM (tm), XAMAP (tm))

Realtime Monitor VM/ESA, Program Descriptions/Operations Manual,
Rel 5.2, SH26-7000-07, 06/94

VM Performance Reporting Facility, Users Guide and Reference,
Vers. 1.2, APAR VM55672, 09/93

VM (Fast) CCW Translation ...

VM/ESA CCW Fast Path Issue with ECKD (< 1.2.2)

„ **Problem**

By not exploiting VM CCW Fast Path (CFP) for ECKD

High total CPU-time by increased CP-time

Applies to:

all type of device definitions
(dedicated devices, full and partial pack minidisks).

all cases where VM CCW translation was required
(i.e not for dedicated devices of V=R/F guests with properly set options)

„ **Symptoms**

Higher T/V ratio (e.g. 1.7 instead of 1.3)

High 'virtual' DASD response times (seen from VSE)

Actual device response times are OK, shown by a VM monitor.

When measured within VSE (e.g. SDAID, CICS monitor..),
I/O response times are too high,
since markable CPU-time for std VM CCW translation included.

Í **Install the VM/ESA ECKD CCW translation fix**

VM57265 for VM/ESA 1.1.1
VM57443 for VM/ESA 1.2.0, 1.2.1

For more details refer to the chart on VM CCW translation.

VM/VSE Guest Comparison

V=V, V=F and V=R dependencies

Ù **V=V virtual machines**

CFP is used for all types of DASDs (minidisks and dedicated).
CCWTRAN ON is always active.

Usage of the VSE/ESA 2.1 NOPDS option possible,
if DEF STOR is big enough (no VSE page data set required)

> Only VM is paging

Ù **V=F virtual machines**

CCWTRAN ON is always active, but in contrast to a V=R machine, SIE assist can work at the same time.

1. If SIE assist is available,
   CP will not see I/Os to dedicated devices,
   CFP is used for minidisks.

2. If SIE assist is not available,
   CFP is used for dedicated devices and for minidisks.

Ù **V=R virtual machines**

VM Guest Recovery is only available for V=R guests.

Due to the fact that CCWTRAN could be on OR off, we have 4 different cases to consider

1. If SIE assist is available and CCWTRAN OFF is set (default),
   CP does not see I/Os to dedicated devices.

2. If SIE assist is not available and CCWTRAN is set OFF,
   CP will check all I/Os to dedicated devices, but there will be no CCW translation.

3. If CCWTRAN OFF and minidisks are used,
   CP will translate CCWs starting with CFP.

4. If CCWTRAN ON and minidisks are used,
   CP will use CFP.

> If you use dedicated devices AND minidisks in a V=R guest, you can either use SIE assist for the dedicated devices OR CFP for the minidisks. But not both together.

If you want to take advantage of both, run with V=F.

Overview on DASD and Guest Types

Overview on DASD and Guest Types

This summary shows for VSE guests, which type of I/O handling is in general possible for each individual type of DASD

Machine Type	SET IOASSIST	SET CCWTRAN	Device Type DED	FPM	PPM	MDCWrite FPM/PPM
V=R	ON	OFF	IOASS	NONE	NORMAL	NONE/NORMAL
V=R	ON	ON	FAST	FAST	FAST	FAST
V=R	OFF	ON	FAST	FAST	FAST	FAST
V=R	OFF	OFF	NONE	NONE	NORMAL	NONE/NORMAL
V=F	ON	-	IOASS	FAST		FAST
V=F	OFF	-	NONE	FAST		FAST
V=V	-	-	FAST	FAST		FAST

MDC VM Minidisk Caching for guests (WRITE-through)

IOASS I/O assist is active for that DEDicated device, no CP interrupt occurs and no VM CCW translation is needed. If I/O assist is not present on the processor, a CP interrupt is generated.

NONE VM CCW translation is not required

FAST VM CCW translation can try the Fast Path (CFP)

NORMAL VM CCW translation is done via the normal path

- IOASSIST OFF is the only case for all processors NOT having VM/ESA I/O assist (e.g. all ES/9000s do)
- I/O Assist is lost when CCWTRAN is set ON for a V=R guest
- For V=F machines CCWTRAN can not be set off, so that switch doesn't apply
- Fast CCW Translation may provide a benefit if the device happens to be out of I/O Assist at the time of the SIOF/SSCH
- For V=V guests CCWTRAN ON is always effective, and IOASSIST OFF
- Fast CCW Translation is slightly slower for V=R (vs V=F) due to requirements for V=R recoverability
- Be aware of all the specific cases and all additional conditions for I/O Assist and Fast CCW Translation

Checks for VM/VSE I/O Time Problems

Checks for VM/VSE I/O Time Problems

Symptom: Too long msec per DASD I/O

Important for cached and/or ECKD requiring devices,
like ESCON attached DASDs and/or 3390s, 9345s, RAMACs

„ **Make sure, 'problem' is not caused by**
device contention
big amounts of data transferred per I/O
wrong sector settings in channel programs
May cause 1 lost revolution per CCW

„ **Is the high msec/IO time**
seen by VSE only?
May be the delays within CP are caused by not using/aborting from VM Fast Path CCW Translation (CFP)
seen also in VM?

„ **Check cache settings for this device/subsystem**
- in VM
- in VSE

„ **Make sure, ECKD channel programs are used**
- in VM
- in VSE

„ **Make sure, the ECKD channel programs have the correct cache bit settings**
- in VM
- in VSE

Í **It is often NOT sufficient to trace a problem within VSE guest only**

For problems with high msec/IO, refer to
'VSE/ESA I/O Subsystem Performance Considerations'

Sharing Dedicated Devices?

Dedicate a device to >1 virtual machine?

- To make use of SIE assist for devices used by more than one virtual machine (minidisks are not SIE assisted), some installations defined a device more than once in the IOCDS (More than one path to the device is required in this case).

  This method is not supported, worked in the past, but there is no guarantee that it will work in the future.

More VM/VSE Guest Type Considerations

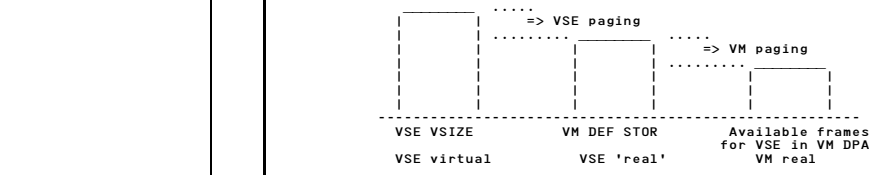
PART C.

More VM/VSE Guest Type Considerations

VM/VSE Guest Type Considerations

'Double Paging' for V=V (MODE=ESA) VSE Guests

Ù **Relevant Storage Sizes**



- VSE paging is the higher, the bigger VSE VSIZE vs the DEF STOR size is.
  No VSE page-I/Os at all if DEF STOR > VSIZE (roughly)

- VM paging is the higher, the bigger DEF STOR size vs the available page frames for VSE in the DPA is.
  No VM page-I/O if a VSE page is in VM DPA

Ù **Paging Remarks/Comparison**

	VSE Guest	VM CP
Page Data Set size	VSIZE (+VIO) Fixed assignment	up to DEF STOR Dynamic assignment
Page I/O operations	Normal CP overhead No VSE CCW transl. (SYSIO) Mostly single, few block paging	More block paging, segment oriented
Replacement strategy	VSE LRU	CP LRU

- VSE page-I/Os have same CP overhead as normal file I/Os

- VM reflects page-faults for VSE pages via 'pseudo-page-fault' interrupts, allowing VSE guest to continue with another task

- LRU = Least Recently Used
  Apart from customer statements on 'better' VM paging no further direct indications available on that

- To avoid (unnecessary) page-ins of a VSE page by VM with invalid (to be cleared) content,
  VSE uses CP Diag10 at all occasions where in native case a VSE page is cleared (e.g. INVPAGE etc.)

VM/VSE Guest Type Considerations ...

'Double Paging' for V=V (MODE=ESA) Guests (cont'd)

- **Disadvantages of 'Double Paging'**
  - **Paging CPU-time**

    VSE paging for V=V costs more total CPU-time (VSE + CP overhead) than VM paging

  - **'Unnecessary' VSE page I/Os**

    Even if there would be a page frame available in the VM DPA, VSE may have to do a page-I/O.

  - **'Double' page-I/Os**

    It may occur that if a single page is not in real storage, both a VM and a VSE page-I/O must be done:

    If VSE has selected a frame only in VSE 'real' storage, but not in VM real storage, a VSE page-I/O may cause also a VM page-I/O, to first bring the VSE 'frame' into VM real storage

- **Don't care on 'double paging' if VM+VSE paging is low**

Reasons to run V=V Guests

Reasons to run VSE (MODE=ESA) guests with V=V

Here follows a list of reasons why customers do not run V=R/F with its associated performance benefits such as:

- VM CCW translation bypass (Dedicated, or V=R Full Pack Minis)
- I/O passthru (Dedicated, on ES/9000s)
- Avoid double paging (refer to separate chart)

Overview

- **Non ES/9000 Processor**
- **Heavy Load Variation(s) for VSE guest(s)**
- **Too Small Real Storage installed**
- **Shared Minidisks**
- **Big DEF STOR Size avoids VSE Paging for V=V**
- **Exploiting Expanded Storage for VSE Guest Paging**
- **Functional Requirement for V=V**
- **Historical Reasons**

The disadvantage of V=V vs V=R/F is smaller if

- mostly minidisks are used
- VM Fast CCW Translation is used
- SET RESERVED is used to fix pages

If you run V=V for other reasons than shown below,

- check the reason why you still run V=V
- check the potential for you to improve guest performance

Reasons to run V=V Guests ...

Reasons to run VM/VSE guests V=V (Details)

- **Non ES/9000 Processor**

  Running on elder processors without PR/SM support for Multiple Preferred Guests (MPG) does only allow a V=R guest.

  The number of V=R/F guests on an ES/9000 supported by VM/ESA is 1xV=R + 5xV=F, or 6xV=F, sufficient for most installations.

- **Heavy Load Variation(s) for VSE guest(s)**

  Demand for real storage by individual VSE guests varies so heavily (across a day, week,..) that optimal V=R/F sizes are hard to define, except abundant real storage is available.

  If only the minimum permanently required V=R/F sizes are defined for the guest(s), no other real storage can be used, since only VSE is paging for V=R/F.
  Especially in case of CICS production a higher paging rate is not acceptable.

- **Too small Real Storage installed**

  For some strange reason, the customer has too small real storage and wants no reservation for even constant loads.
  Even the 128 MB limitation for integrated adapters on 9221s should allow V=R/F guests.

  > It is your decision

Reasons to run V=V Guests ...

Reasons to run VM/VSE guests V=V (cont'd)

- **Shared DASDs**

  VM/VSE DASDs have to be set up as VM minidisks (FPM or PPM) e.g. to exploit VM MDC for I/O caching of a guest.

  If most of the DASDs of a VSE guest are to be defined as minidisks, the CPU-time benefit of V=R/F vs V=V is smaller and thus V=V may still be used.

  In any case, check whether you can at least partly run with dedicated devices if DASD volumes are non-shared.

- **Exploiting Expanded Storage for VM MDC**

  For VM Minidisk Caching, also expanded storage can be used (though, using real storage is performance-wise superior).
  So this would be an exceptional case

- **Big DEF STOR Size avoids VSE Paging for V=V**

  VSE in such a case will never do any page-I/O, since VSIZE is smaller than the simulated real storage by VM. Nevertheless, the VSE Page Data Set is required (before VSE/ESA 2.1).

  Use the VSE/ESA 2.1 NOPDS option to get rid of the VSE Page Data Set in such a case.

  Performance aspects:

  - There is no CPU-time benefit as compared to the same V=V guest with a smaller DEF STOR, if VSE would not actually page.
  - No VSE page-I/Os on top of VM paging

  Normal case (DEF STOR < VSIZE):

  VSE paging I/Os (if present) for V=V guests require high CP overhead (no CCW bypass, no IO passthru), therefore a too small DEF STOR for V=V guests is more harmful regarding CPU-time than for V=R/F guests.

  It is possible for V=V guests to reserve real storage via LOCK or, better, SET RESERVED commands.

Reasons to run V=V Guests ...

Reasons to run VM/VSE guests V=V (cont'd)

Ù **Exploiting Expanded Storage for VSE Guest Paging**

It may be beneficial to do paging for V=V guest(s) via expanded storage (high speed paging area), if

- on an ES/9000 only part of the central storage (128M) can be used as main storage if integrated adapters are used (9221)
- on an ES/9000 part of the total processor storage is available and only usable when configured as exp. storage (9121, 9672)
- unused 'real' expanded storage is available (9021)

Since VM/ESA 1.2.2 Minidisk caching is available, expanded storage can also be used for that purpose, though, real storage is even better.

> Specific cases

Ù **Functional Requirement for V=V (e.g. old GCS)**

RIO370 is an optional VM real storage area, required when authorized V=V machines (e.g. GCS) want to do Real Channel Program Execution (This applies to the elder 24-bit DIAG98 only).
If present, it must completely reside below the 16M real line, together with all V=R/F reserved areas.

> V=R/F guests are hardly possible when RIO370 is required.

Ù **Historical Reasons**

Customer migrated from the old MODE=VM supervisor and just kept V=V.
If reserved real storage for the VSE guest was required, it was done via SET LOCK or better SET RESERVED commands.

> Exploit V=R/F if reasonable

Virtual Disks for VM/VSE

PART D.

Virtual Disks for VM/VSE

Virtual Disks for VM/VSE

General

„ **Both VSE and VM Virtual Disk (VD) can significantly enhance Elapsed Times by avoiding physical I/Os**

Percentage of Elapsed Time savings depend on how I/O bound an activity is/was

„ **Both can be used even on ESA/370 processors**

No ESA/390 required

„ **Mapping concept (and thus pathlengths) are not too complex, since FBA logic used**

VM Virtual Disk Specifics

„ **CPU-time aspects**

- No VM CCW translation required
- More CPU-time savings in CP when interrupt is saved

VM/VSE Virtual Disk Sample Results

VM/VSE Virtual Disk Sample Results

„ **Extreme I/O-intensive VSAM KSDS job**

Load 100K records & randomly read 20K, update 10K, insert 10K records
(100 byte records, 7 byte keys)

Guest	Real Disk (Base)	VSE VD	VM VD
Elapsed time ratio			
V=R, dedicated (assisted I/O)	1.0 (469 sec)	0.27	0.24
Other (unassisted I/O)	1.0 (473 sec)	0.28	0.24
Total CPU-time ratio			
V=R, dedicated (assisted I/O)	1.0 (39.8 sec)	1.02	1.63
Other (unassisted I/O)	1.0 (62.7 sec)	0.94	1.09
- Consider that all 1.0 bases (...) are different - The smaller the values, the better			

Measurement Environment:

- 9121-260
- VSE/ESA 1.3.2 under VM/ESA 1.2.1
- 3390 real DASDs, VDs as FBA
- VSE DASDs defined as dedicated devices
- IOASSIST ON, and CCWTRAN OFF, where possible
- VM/ESA ECKD fix for CFP
- Only test job was active, i.e. single batch environment

VM/VSE Virtual Disk Sample Results ...

„ **COBOL Compile & Link ('avg' I/O-intensive) job**

Guest	Real Disk (Base)	VSE VD (under VM)	VM VD
Elapsed time ratio			
V=R, dedicated (assisted I/O)	1.0 (71 sec)	0.58 (41 sec)	0.59 (42 sec)
Other (unassisted I/O)	1.0 (77 sec)	0.54 (42 sec)	0.54 (42 sec)
CPU-times (sec)			
V=R, dedicated (assisted I/O)	TOT 5.69 VSE 4.29 CP 1.40	TOT 4.59 VSE 3.90 CP 0.69	TOT 5.03 VSE 3.74 CP 1.29
Other (unassisted I/O)	TOT 7.51 VSE 4.34 CP 3.17	TOT 5.35 VSE 3.93 CP 1.42	TOT 5.81 VSE 3.85 CP 1.96
Total CPU-time ratio			
V=R, dedicated (assisted I/O)	1.0 (5.69 sec)	0.80	0.88
Other (unassisted I/O)	1.0 (7.51 sec)	0.71	0.77

- Consider that all 1.0 bases (...) are different
- The smaller the values, the better

Measurement Environment:
- 9121-320 with 3380 DASDs
- VSE/ESA 1.3.2 under VM/ESA 1.2.2
- VSE DASDs defined as DEDicated devices
- IOASSIST ON, where possible
- CCWTRAN OFF, where possible
- VM/ESA ECKD fix is standard for VM/ESA 1.2.2
- 4 test jobs were active concurrently: multiple partition case
- Note that in this specific case, CPU-time was saved since the jobs contained also a Link-Edit step, which profited from the CKD to FBA transition (better blocking).
- The same Elapsed time for VSE and VM VD is due to same remaining number of physical I/Os
- Be careful when interpreting VSE I/O numbers to DEDicated devices if measured via VM
  (only seen if 3990 cached and properly setup)

VM/VSE Virtual Disk Sample Results ...

Conclusions

If a Virtual Disk is function-wise applicable...

Í **Use a VSE VD**

for a V=R guest if real file on a dedicated real disk

(i.e. assisted real I/O)
at some costs in VSE

in all other cases

(base file with unassisted I/O)
at some costs in VSE VTIME, but savings in CP overhead

Í **Use a VM VD**

appearing to VSE as a 9336-20

if functionwise needed

- any files shared across multiple VSE guests
  (e.g. use a VM Virtual Disk for the VSE LOCK file alone)

- any virtual disk that must survive VSE IPLs
  (VM VDs exist as long as at least 1 LINK exists to it)

if this would be the only way to exploit available expanded storage

(via VM paging for V=V guests, but,
VM MDC is to be preferred if expanded storage left)

For more info refer to:

'VM/VSE Performance Hints and Tips', GG24-4260, ITSC BOE, 03/94

VM/ESA Fulltrack Minidisk Caching

PART E.

VM/ESA Fulltrack Minidisk Caching

VM/ESA Fulltrack Minidisk Caching

VM/ESA 1.2.2 Fulltrack Minidisk Caching (MDC)

„ **is now available to any VM guest**

Not only CMS, not only for 4K blocksize.
Record level minidisk caching is not applicable for VSE guests

„ **has been enhanced**

Caching Characteristics

„ **READ caching, on a 'per track' base**

(1 FBA 'track' = 32K)

„ **Synchronous WRITEs to DASD ('write thru')**

MDC cache in storage is updated in case of a hit,
but only if track was in MDC-cache.

Í **no performance benefit for any WRITEs or if data READ only once**

> no integrity exposure

„ **MDC-Staging is done on Full Track base (1 track/IO) at a READ miss**

- No pre-staging across tracks is being done.
- WRITE misses do NOT cause staging.

„ **Caching is done in main or expanded storage, or both**

> best performance obtained if main storage used

• Cache management is done on a 4K (page) base for 'standard' tracks

- CKD/ECKD tracks with fixed record lengths of specific sizes:
  256, 512 bytes, 1K, 2K, 4K
- FBA tracks

(Of importance for CMS 4K blocks, and for data moves to/from VM expanded storage)

VM/ESA Fulltrack Minidisk Caching ...

Scope of Functional Applicability

„ **Any disk functionwise eligible, independent of the R/W ratio**

But R/W ratio affects performance vs non-MDC disks,
as does the READ access pattern and blocking (KB/IO)

„ **MDC available for 3380s, 3390s (incl. RAMACs), 9345s, FBAs**

FBA-minidisks (Refer to VM/ESA Planning and Administration):

- Definition must start and end on page boundary (8 blocks, page aligned).
- For optimal FBA minidisk performance, the minidisk should be defined on pseudo track boundaries (64K blocks).

„ **MDC is possible on all ESA/370 processors:**

ESA/370 address spaces and access registers used

(Provided the processor is supported by VM/ESA 1.2.2).
But, ITR performance benefits smaller if ESA access registers not optimally implemented (e.g. 4381-9xE).

„ **MDC disks can be shared between VSE guests, if under the same VM**

All cases apply where sharing so far was possible w/o MDC
(there is one global MDC-cache maintained by CP)

„ **MDC disks cannot be shared across VM systems**

Minidisk in fact must not reside on a DASD physically shared across VM systems, except only 1 guest has WRITE access to it

„ **Tracks with records >32K are NOT MDC-cached**

MDC processing is being exited, I/O is passed to I/O subsystem.
Just a performance issue for the affected tracks

VM/ESA Fulltrack Minidisk Caching ...

VM/ESA Fulltrack Minidisk Caching (MDC) (cont'd)

„ **Tracks for which Format WRITEs are done are discarded from MDC-cache**

„ **Disk must be (re-)defined as Full/Partial Pack minidisk,**

if previously it was defined as DEDicated device

Directory statements:

- MDISK ... DEVNO ... not MDC-cacheable (as of 11/95)
- MDISK ... VOLSER... MDC-cacheable

Do not forget to prefer VSE productions with MDC via the directory option

- OPTION NOMDCFS
  (guest I/O rate not limited by fair share limit)

„ **CPU-time aspects**

Remaining physical I/Os to the minidisk

- do not have I/O passthru
  as existed for DEDicated devices of V=R/V=F guests
- do not have VM CCW translation bypass
  as existed for DEDicated devices of V=R/V=F guests
  as existed for FPMs of V=R guests

> Highest savings in CPU-time for

- V=V guests (any type of DASD: DED, FPM, PPM)
- V=F guests (any type of VM minidisk)
- V=R guests (PPMs)

• For more details refer e.g. to:

'VM/ESA 1.2.2 Performance Report', VM22PERF PACKAGE on MKTTOOLS
(includes many details and performance results)

The sample results on a later chart just show a rough overview.

MDC Performance Monitoring

MDC Performance Monitoring

VM monitors display MDC related data

e.g.
- total virtual #READs and #WRITEs
- #I/Os avoided
- virtual R/W ratio
- MDC read hit ratio
- storage used for MDC

RTM/ESA

MDCACHE and MDCLOG screens:

DISPLAY MDCAChe
DISPLAY MDCLog

VMPRF

MINIDISK_CACHE_USAGE_BY_TIME (PRF103) screen.
Also look at the TREND and SUMMARY records

FCON/ESA

Detailed I/O Display screen

MDC-cache unfriendliness can be assumed, if a low hit ratio is obtained, though all of the following conditions are fulfilled:

- logical device eligible for MDC
- MDC actually is active
- not too many WRITEs vs READs

VM MDC PTF

When VSE/VSAM APAR DY43335 is installed (dated 10/95), you need for VM MDC-caching the PTF for VM APAR VM60996:

VM/ESA 1.2.2 UM28334
VM/ESA 2.1.0 UM28333
VM/ESA 2.2.0 UM28335

When MDC Cannot Give Benefits

When MDC Cannot Give Benefits

In all these cases, MDC cannot bring benefits in general and for specific jobs.

It may be advisable

- to NOT MDC-cache such a volume, or ...
- to move files around to better separate them regarding MDC-Caching

„ **All cases with 'WRITE only' or 'WRITE mostly'**

WRITEs are not cached; but very few instructions are required to check whether a track is already in MDC-cache (and would need to be updated).

Check if such files can reside on a DEDicated disk in a preferred guest.

„ **All cases where data is read only once**

The real benefit of minidisk caching is for multiple reference of the same data.
Writing data to disk and reading them in later ('workfiles'), always means that those tracks are NOT in MDC-cache.

- SORT workfiles (e.g.)

„ **All cases with 'Format Writes'**

Format Writes cause a purge of the track in MDC-cache and a temporary bypass of the MDC-cache for the pertinent track(s)

„ **All 'applications' which are 'cache unfriendly'**

e.g. SQL/DS or DL/1 data bases

Here a 3990-6 cache is optimal, since VSAM uses both 'Record Caching' and 'Regular Data Format'

- to cope for cache unfriendly access patterns
- to get write hits also for update writes

(VM minidisk caching on record level is only available for CMS, APAR VM61045)

„ **MDC benefits are smaller if DIM already applied**

(e.g. in CICS/VSAM, via LSR pools or Data Tables)

When MDC Should Not be Used

When VM Minidisk Caching should NOT be used

„ **All cases with (mostly) seq. READs,**
- if READ blocking (#bytes/IO) is > 1 track
- if SEQ bits are used in ECKD channel pgms
(even if only 1 track/IO)

- FBA: > 32K/IO
- CKD/ECKD: > tracksize/IO

Note, if a utility is optimally designed, MDC cannot read data faster from DASD. This holds e.g. for

- VSE FASTCOPY
- VSAM B/R

This is especially true, if the guest utility or program uses

- multiple track READs, or
- sequential caching bits on top of 1 track/IO
(and devices are cached and accessed as ECKD)

VM MDC does not do multiple track/IO:
Pure sequential I/Os are no good MDC-cache candidates, except the utility or program is not optimal for this purpose.

„ **All cases where tracks contain physical record sizes >32K**
This causes a purge of the track in MDC-cache and a temporary bypass of the MDC-cache for the pertinent track(s)

„ **VSAM files pre-dominantly used for BROWSE**

When it definitly must be avoided that a high VSAM I/O intensive CICS transaction (Browse) dominates other txns in the same CICS (VSAM NSR, for LSR no read ahead is done)

„ **VSE/ESA Lock File**

Even if only a VSE Lock File would be on such a minidisk.

- Any minidisk accessed with RESERVE/RELEASE CCWs (as used for the VSE Lock File) falls out of MDC-caching!

- Lock File is WRITE intensive

MDC Usage Recommendations

MDC Usage Recommendations

Í **Do NOT expect MDC-caching benefits for all types of use/files**

Í **Switch MDC off for fast Guest BACKUP/RESTORE**

Í **Put files with phys. recordsize >32K on non-MDC-cached minidisks**

Í **Do not MDC-cache SORTWK files**

They are READ once, and may use >32K records

Í **Check MDC Usage/Status for SEQuentially used files**

Í **Put VSE/ESA Lockfile on VM Virtual Disk**

MDC Controls and Warning

MDC Controls

„ **Flexible MDC controls available**

- Real and/or expanded storage for MDC can be specified as fixed or even variable sizes

```
SET MDCache STorage maxM          Min and Max main storage
            STorage minM maxM     for MDC
```

Carefully observe VM paging when maxM is not set or too high

- Caching can be even switched ON|OFF on a temporary base in a hierarchical and/or selective manner:

```
SET MDCache SYSTEM ON|OFF         All disks/guests
            RDEV   OFF rdev       Real disk
            MDIsk  ON|OFF vdev    Minidisk
            INSert ON|OFF uid     Specific guest
```

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

Warning/To be observed

Like any misuse of DIM, too much data in a limited real storage may cause disastrous paging.

This is especially important for VM/ESA V2, where MDC is the default for ALL minidisks!!

Í **Check VM paging carefully**

- Make sure that any V=V guest still has enough real storage available

- Slight changes in SET MDC STOR maxM may significantly change VM paging rates (and V=V/V=R guest performance)

More info

• For more info on Minidisk Cache, refer e.g. to

- Minidisk Cache, by Bill Bitner
http://www.vm.ibm.com/perf/tips/prgmdcar.html

Some MDC Performance Results

„ **Very I/O intensive PACEX8 Batch workload**

V=R Guest runs			
	Uncached DEDicated (Base1)	Uncached FPMinis (Base2)	MDC Cached FPMinis 18 MB, fixed
Elapsed Time	566 sec '1.0'	562 sec 0.99 '1.0'	361 sec 0.64 0.64
CPU-time	131.3 sec '1.0'	174.9 sec 1.33 '1.0'	174.6 sec 1.33 1.00
DASD I/Os	167.7K '1.0'	171.3K 1.02 '1.0'	94.8K 0.57 0.55
V=V Guest runs			
	Uncached FPMinis (Base3)	MDC Cached FPMinis 16 MB fixed	MDC Cached FPMinis 48 MB fixed
Elapsed Time	569 sec '1.0'	389 sec 0.68	278 sec 0.49
CPU-time	184.8 sec '1.0'	180.4 sec 0.98	164.8 sec 0.89
DASD I/Os	172.7K '1.0'	111.3K 0.64	83.0 K 0.48

- Consider that all '1.0' bases are different
- The smaller the values, the better

- 9121-320, no paging, VSE/ESA 1.3.2, uncached 3380 DASDs, MDC in real storage, controlled lab environment

„ **Up to 50% Elapsed Time reductions,**
highly dependent on workload and environment

„ **CPU-time ratio depends also on source environment**

Source I/Os:	CPU-time:	vs:
fully assisted	some increase	V=R DEDs
partly assisted	about same	V=R FPMs
not assisted	some reduction	any V=V disks

Some MDC Performance Results ...

An Experiment: MDC vs 3990 Caching

„ **Workload**

A CICS Assembler transaction, using VSAM KSDS for READs.
Here the same record was read twice, and always was a hit,
in MDC or in 3990 cache.

„ **Environment**

9221-421 processor, 3380s at a (old) 3990-3 (parallel) channel.

VSE/ESA 1.3.6 and VM/ESA 1.2.2.

KSDS file, using VSAM LSR (4K data, 2K index-CI) and SHROPT 4
(SHROPT 4 deliberately was used to force 2 VSE I/Os for each EXEC
CICS READ, in order to read sequence set index-CI and the data-CI)

„ **Results**

Elapsed and CPU-timings from CICS Auxtrace		
	No MDC used	With VM MDC
	2 log. VSE SSCHs = 2 phys. SSCHs (2 3990-3 hits)	2 logical VSE SSCHs no phys. SSCHs (2 MDC hits)
Elapsed Time CPU Time -> I/O Time	6.2 msec 2.0 msec 2x 2.1 msec/IO	2.6 msec 2.0 msec 2x 0.3 msec/IO
I/O time is the time for a logical I/O seen from VSE		

„ **Conclusions**

> MDC hits were much faster than DASD cache hits.

(MDC hits even were so fast that the EXCPAD VSAM exit used by
CICS avoided the processing of an SVC7 for that transaction)

Some PR/SM LPAR Aspects

PART F.

Some PR/SM LPAR Aspects

This part is only a rough performance related overview.

For details and for latest updates of PR/SM functions refer e.g. to

- PR/SM Planning Guide, GA22-7236-00 (09/96) or later

Partitioning

Principal Partitioning Possibilities

Physical Partitioning

Subdivision of multiprocessors into 2 separate processor complexes

Logical Partitioning

This is done via PR/SM (Processor Resource/Systems Manager),
with LPARs (Logical PARtitions).

PR/SM is a standard H/W feature on ES/9000 processors and follow
ons.

Much more flexible than physical partitioning.

Up to 15 LPARs per system.

Software Partitioning

Subdivision of resources by VM for use by VM guests

Combinations possible

PR/SM Overview

Purpose

Coexistence of multiple SCPs

SCP means 'System Control Program'

Provision of several Logical Partitions (LPARs) with independent operation

... even on a uni-processor complex

Exploitation of larger machines with 'smaller' SCPs

Formerly e.g. S/370 mode SCPs on ESA, VSE/SP on multiprocessors...

... even without VM

... with performance approaching native performance in specific cases

This applies primarily to a DEDicated LPAR.
Usually, the performance of a guest in a SHARED LPAR is close to
the performance of a preferred VM guest with all DASDs dedicated.

(up to here: similar reasons as for using VM guests)

PR/SM is the prerequsite for VM/ESA MPG support

Multiple Preferred Guests, means SIE Assist, which includes I/O
Assist and SIGP Assist

LPAR Overview

Logical Partition (LPAR)

A collection of processor complex resources that can be run by an operating system

Up to 10 LPARs (20 if processor complex is additionally physically partitioned), depending on processor type

Characterized by:

- **Mode (S/370, ESA/390)**

  9672-Rx5 (G4) CMOS processors no more allow S/370 LPARs, nor S/370 VM guests

- **# logical processors**

  Any physical processor (CP or engine) may be SHARED among LPARs, or may be DEDICATED to a single LPAR for exclusive use.

  Any LPAR can only have shared or dedicated processors, not both.

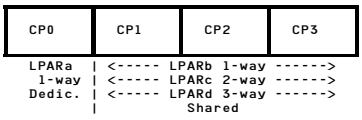
  - **'Dedicated' or 'Shared' LPARs**

    Both types can be active concurrently, but on different subsets of engines

  The total #logical processors for all shared LPARs may be larger than the number of physical processors serving the shared LPARs.

  But 1 LPAR can at most have as many logical processors as engines are available on the processor.

  Example: 1 Dedicated and 3 Shared LPARs on a 4-way



LPAR Overview ...

Logical Partition (cont'd)

- **Central (main) storage**

  Sharing of allocated central storage is not allowed between LPARs

- **Expanded storage (optional)**

  Contiguous areas with at least 1 MB granularity, not shareable. Not required/directly useable for VSE native

- **Channel paths, subchannels**

  On EMIF capable processors, ESCON channels may be shared between LPARs.

  Channel paths may be reconfigurable, but can be used only by one LPAR at any point in time.

  Devices can be shared as long as attached via different channel paths:
  - DASDs via control unit
  - Printers only via switchbox
  - Terminals not in general

  Each VSE needs a separate console attached to a separate local non-SNA control unit.

LPAR Performance Dependencies

LPAR Performance Dependencies

(Absolute performance or performance vs native, refer to GA22-7123)

- **Number of processors**

  It is recommended to define only the required number to handle the workload

- **LPAR mode**

  If a S/370 guest is allowed, SIO or SIOF must be translated to SSCH

- **SCP (e.g. VSE)**

- **Workload**

  Instruction mix, incl. frequency of SIE invocation, ...

- **LPAR Status (Dedicated or Shared)**

  If possible, use dedicated LPARs (with own physical processors). This reduces LPAR overhead.

- **Processing Weights and Capping Status**

  Processing weights (shared LPARs only) range from 1 to 999.

  Capping status can be YES or NO for an LPAR.

  A capped shared LPAR cannot use more than its share based on its processing weight, even if other shared LPARs would not need it: 'hard limit'.

  -> Do not cap an LPAR without real need

- **'Wait Completion' definition (Shared LPARs)**

  No = default = event driven = recommended for shared engines

LPAR Performance Dependencies ...

LPAR Performance Dependencies (cont'd)

- **Processor running time interval**

  This is a CPU-time-slice, either system or user determined

- **# LPARs**

- **# real and # logical processors being online**

  Keep the ratio logical-to-physical processors as low as possible

- **Processor utilization(s)**

  'Low Utilization' effects for event driven scheduling

- **Central storage available in LPAR**

- **Amount of activities in concurrent LPARs**

- **Note of caution for co-existence with OS/390**

  In case, OS/390 Sysplex runs in other PR/SM LPARs, using External Time Reference (ETR) interrupts ...

  - ETR interrupts are ignored by VSE/ESA (and others)

  - Pending ETR interrupts may cause high PR/SM overhead

  Install the PTFs for

VSE/ESA 2.x	APAR DY45481	UD51541(2.1/2.2) UD51543(2.3) UD51548(2.4)
VSE/ESA 2.5	APAR DY45502	UD51527

  They simply catch these ETR interrupts and throw them away on any type of processor.

VM/VSE Guests vs PR/SM LPARs

VM/VSE Guests vs PR/SM LPARs

Ù **Some Aspects on Shareable H/W Resources**

H/W resource	Shareable between LPARs	Shareable between VM guests
Processor Power	YES, except DEDICATEd LPAR	YES, except phys. processor is DEDICATEd
Processor Storage	NO	Only for V=V guests
Channels	Only with EMIF	YES
Control Units	YES, if channels separate, or EMIF	YES
Disks	YES, if control unit shared	YES, except DEDICATEd devices

- EMIF means ESCON channels shared between LPARs (ESCON Multiple Image Feature)

Ù **More Info**

refer e.g. to

- LPAR vs MPG, by Romney White, IBM
  VM/VSE Tech Conf Orlando, 05/1999
  VM/VSE Tech Conf Orlando, 06/2000

VSE/ESA under z/VM on eServer zSeries

PART G.

VSE/ESA under z/VM on eServer zSeries

This part is a preliminary outlook.

For details and for latest information, refer e.g. to

- z/VM Version 3 Release 1,
  WAVV 2000 Colorado Springs, 10/2000
- the IBM Announcements of 2000-10-03, 2001-02-20

z/VM is generally available since 2001-02-23

z/VM Performance Benefits for VSE/ESA

z/VM Performance Benefits for VSE/ESA Guests

Besides support of 31-bit VM guests (as VM/ESA)...

Ù **Exploitation of 64-bit real (>2 GB real memory) on IBM eServer zSeries 900 processors**

„ **for V=V 31-bit guests: VM page pool**

Í **Higher total VM capacity for more guests (or more guest virtual storage)**

„ **for any 31-bit guest: VM Minidisk Caching etc**

Í **More Data In Memory possible, outside of guest virtual storage**

CP nucleus and V=R/F areas must stay below the 2 GB line

Ù **Native FlashCopy support for ESS, e.g. for asynchronous backups or test environments**

EOD End Of Document
HAND Have A Nice Day