



IBM eServer™ zSeries

# Best Practices for Oracle on Linux for zSeries

Configuring and Tuning Linux and z/VM for Oracle

Denny Dutcavich  
[dutch@us.ibm.com](mailto:dutch@us.ibm.com)

845.689.2226

© 2003 IBM Corporation

# Trademarks

**The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.**

AIX*	FICON Express	MQSeries*	RS/6000*
AIX 5L*	FlashCopy*	MVS	pSeries
CICS*	GDPS	Netfinity*	S/390*
DB2*	Geographically Dispersed Parallel Sysplex	NetView*	SP*
DB2 Universal Database	HiperSockets	OS/390*	Tivoli*
Domino	IBM*	OS/400*	WebSphere*
e-business logo*	IBM logo*	Parallel Sysplex*	xSeries
Enterprise Storage Server	IBM @server	PR/SM	z/Architecture
FICON	IMS	pSeries	z/OS
	Lotus*	RACF	z/VM
			zSeries

\* Registered trademarks of IBM Corporation

**The following are trademarks or registered trademarks of other companies.**

Linux is a registered trademark of Linus Torvalds

Penguin (Tux) compliments of Larry Ewing

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries

UNIX is a registered trademark of The Open Group in the United States and other countries.

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

\* All other products may be trademarks or registered trademarks of their respective companies.

**Notes:**

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

# Agenda

- Introduction
- Lessons Learned
  - CPU
  - Memory
  - Paging and Swap Space
  - VM and Guest Setup
  - Oracle parameters that affect I/O and throughput
  - Disk Considerations
  - Monitoring Linux and z/VM
- Key Success Factors

## Introduction

- The objective of this presentation is to provide the tips and techniques we learned to implement Oracle in a Linux for zSeries Environment. The inputs came from;
  - The porting and testing work performed at Oracle
  - Early Adopter customers
  - Tests run at IBM sites
- This information can be found in the Redbooks;
  - Experiences with Oracle9i on Linux for S/390 SG24-6552
  - Experiences with Oracle Database 10<sub>g</sub> on zSeries SG24-6482 (June)
  - Linux on IBM eServer zSeries Performance Mgmt and Tuning SG24-6926
- And information on how to setup a Linux guest in z/VM can be found in Redbook;
  - -z/VM and Linux on zSeries; From LPAR to Virtual Servers in Two Days SG24-6659
- Or in the Redpaper
  - Linux on IBM zSeries and S/390: Building SuSE SLES8 Systems Under z/VM to install a Linux guest



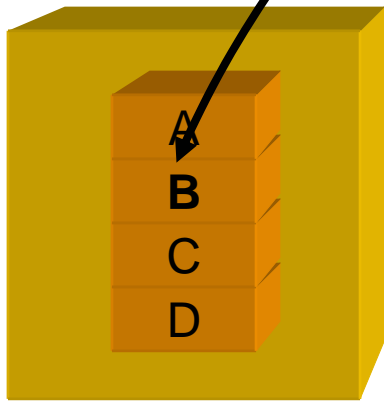
# CPU

- Performance on zSeries CPUs comparable to CPUs on other platforms of similar speed.
  - CPU speed is not entire story
    - zSeries has definite advantage with applications with mixed CPU and I/O
  - z/VM provides unique abilities to virtualize resources and simplify management of guests
  - Good planning is a must. IBM can do sizings and assist with planning and initial installation needs.
- On benchmarks
  - Benchmarks against other platforms are not a good thing
    - This will only test processor speeds
  - Best to test workloads selected for Linux on zSeries
- Allocation virtual CPUs
  - Virtual CPUs assigned to a guest should not exceed real CPUs
  - Assign all the real CPUs to a guest that is needed to obtained the necessary performance

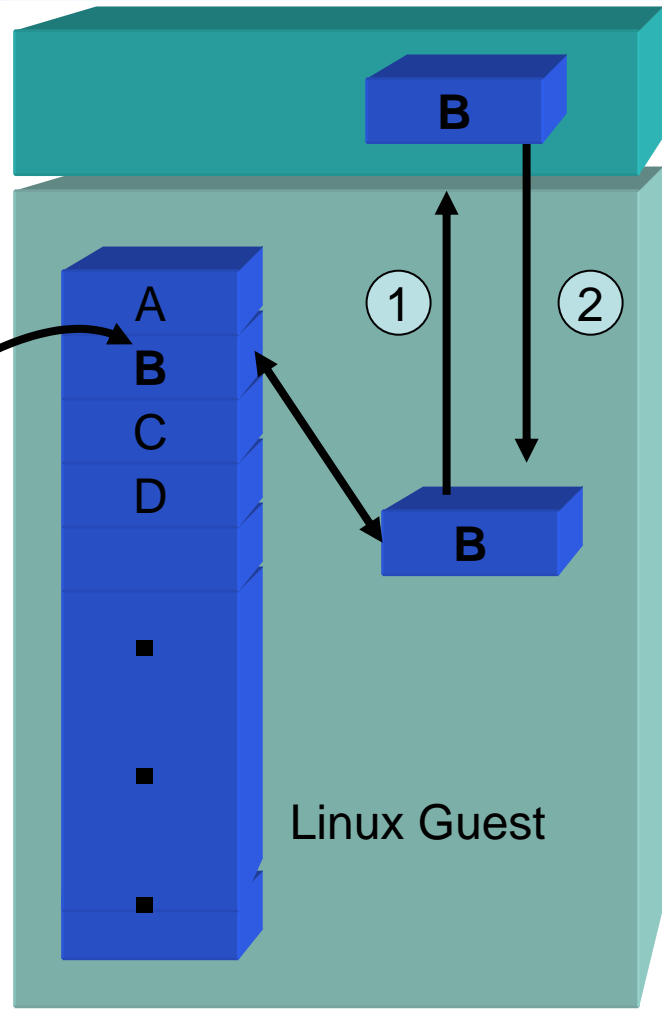


# Paging

- ③ Linux completes P.O.
  - Moves page to swap device (VDISK)



Linux Paging



z/VM Main Memory

## z/VM Expanded Storage

- ① Page B paged out by z/VM
  - Available for P.I.
  - Linux clueless
  - Paged to eStore
- ② Page B paged in by z/VM
  - Linux needs to page
  - Page fault accessing Page B

**Paging can be OK. Make sure first level of paging is to memory!!!**



# Memory

- Sizing memory is critical
  - A Linux guest running Oracle should have memory for
    - Oracle SGA and PGA (if needed)
    - Linux
  - Oracle Systems Global Area (SGA)
    - 750 MBytes is max for Oracle9*i* on either 31bit or 64bit Linux (can be larger if necessary)
    - Databases with very large SGAs in Oracle 10*g* may perform better in a LPAR rather than under z/VM
- Not enough memory
  - KILLMEM process shuts down process if there is not enough memory for the Linux kernel to run
  - Linux will continue to shut down processes in an effort to stay alive, it won't commit suicide
- Paging happens
  - But paging to memory helps

# Paging

- VM Paging Space is critical
  - Configure expanded storage as the first paging device for z/VM
    - Configure at least 25% of physical memory for xStore
  - Paging Space
    - Provide enough disk to keep 50% free
    - Separate swap space from
      - guest data
      - sysres
- Linux Swap
  - Linux swaps (pages) to manage memory
    - Insure enough swap space configured
  - Configure VDISKS for Linux Swap Space
    - Swaps are done to memory with VDISK
    - No memory utilized until swap occurs
    - Use judiciously

## I/O Contention Caused by the 2GB Line

- z/VM supports 64bit guests but, I/O is still 31bit
  - Pages that require certain CP processing must reside below 2GB in z/VM's central storage (host real memory). This includes things such as I/O channel programs and data (both traditional SSCH and the newer QDIO), simulation of instructions, and locked pages (e.g. QDIO structures for real devices).
  - See <http://www.vm.ibm.com/perf/tips/2GSTORAG.HTML>
- Some configuration hints and tips
  - Virtual Machines that do not drop from the queue can hold pages below the 2GB line.
  - Use Guest Lan or VSwitch
  - MDC vs SSCH
  - Use VDISKS
  - Insure guest memory sized correctly
  - Idle guests should be idle
  - Consider multiple LPARs



## 2GB Contention as It Relates to Oracle in SLES8

- You may experience bottlenecks if the cumulative size of the Oracle `DB_CACHE_SIZE` and `LOG_BUFFER` for all virtual machines exceeds about 2GB and there is substantial I/O
- Alternatives for Oracle implementations for SLES8
  - Use `filesystemio_options=setall (asyncio and directio)` parameter in the `init.ora` (10g only and SLES9)
  - Use multiple z/VM LPARs
  - Run Linux in a native LPAR
- Also Turn off Timer (`hz_timer`)
  - `sysctl -w kernel.hz_timer=0`
  - Timer pops cause guest to stay on queue and hold pages
- SRM Tuning for constrained systems (or constrained time periods) to over commit resources
- Use `QUICKDISP(atch)` judiciously!!!
  - Immediate dispatching from eligible to dispatch queue
  - Turning it on for all guests will not help



# Example of I/O contention

- This example demonstrates the problem
  - Note the high paging rates to xStor
  - Varied memory sizes and compared results
- Configuration
  - 8 GB Guest
  - 1 GB SGA
  - I/O stress test
- ESAMON used to display indications

Time	<---Users---> <-avg number-> On Actv In Q			Transactions Per Avg. Minute Resp		<Processor> Utilization Total Virt.		Storage (MB) Fixed Active User Resid.		<-Paging--> <pages/sec> XStore DASD		<-----I/O-----> <-DASD--> Other <-Cache--> Rate Resp Rate Rate %Hit			<MiniDisk> Spool Page Rate <-per second->		Communications IUCV VMCF		Captur Ratio (pct)	
08/10/04																				
13:16:00	35	11	6.0	71.0	1.756	37	35	56.9	3662.8	0	0	325	1.9	0	446.0	72.6	0	2	0	100.00
13:17:00	35	11	7.0	68.0	0.221	98	96	56.2	4130.3	0	0	287	1.1	0	111.1	88.5	0	3	0	100.01
13:18:00	35	11	7.0	68.0	1.404	101	98	57.4	4541.5	0	0	195	0.5	0	60.3	90.0	0	4	0	100.00
13:19:00	35	11	8.0	56.0	1.597	100	76	56.1	4595.3	1208	0	606	0.7	0	31.7	87.9	0	5	0	100.00
13:20:00	35	11	7.0	57.0	1.496	96	59	55.9	4575.7	1405	0	659	0.7	0	11.4	72.3	0	6	0	99.99

# Reconfigured Memory

- This example demonstrates better Performance
  - No paging to xStore
  - Same tests as previous chart
- Configuration
  - 1 GB Guest
  - 768 SGA
- ESAMON used to display indications

Time	<---Users---> <-avg number-> On Actv In Q			Transactions Per Avg. Minute Resp		<Processor> Utilization Total Virt.		Storage (MB) Fixed Active User Resid.		<-Paging--> <pages/sec> XStore DASD		<-----I/O-----> Rate Resp Rate			<MiniDisk> <-Cache--> Rate %Hit		Spool Page Rate	Communications <-per second-> IUCV	VMCF	Captur Ratio (pct)	
08/10/04																					
12:06:00	35	11	6.0	84.0	0.517	2	2	57.4	2634.1	0	0	132	0.5	0	2.0	59.5	0	1	0	100.00	
12:07:00	35	11	6.0	101.0	0.368	2	1	56.0	2634.1	0	0	133	0.4	0	0.2	90.9	0	1	0	100.00	
12:08:00	35	11	6.0	107.0	2.676	1	1	56.8	2634.1	0	0	130	0.5	0	0.4	54.5	0	2	0	100.00	
12:09:00	35	11	6.0	109.0	1.596	1	1	57.5	2634.2	0	0	127	0.4	0	0.2	81.8	0	1	0	100.00	
12:10:00	35	11	4.0	98.0	0.848	2	1	56.2	2635.4	0	0	129	0.4	0	1.3	38.3	0	1	0	100.00	
12:11:00	35	11	6.0	113.0	1.306	1	1	56.1	2635.4	0	0	128	0.4	0	0.2	100	0	2	0	100.00	
12:12:00	35	11	7.0	100.0	0.432	19	18	56.4	3046.2	0	0	185	1.1	0	122.3	59.1	0	1	0	100.01	
12:13:00	35	11	6.0	65.0	1.757	87	84	56.0	4011.4	0	0	422	1.4	0	471.0	76.2	0	2	0	100.01	
12:14:00	35	11	7.0	72.0	2.586	100	98	56.3	4412.3	0	0	398	0.9	0	65.2	89.3	0	4	0	99.99	
12:15:00	35	11	7.0	64.0	0.055	96	94	56.5	4808.9	0	0	420	0.9	0	41.8	89.5	0	5	0	100.00	

## Oracle init.ora options

- No async I/O (aio) for Oracle9i
- The following init.ora parameters can affect I/O performance with Oracle 10g
  - filesystemio\_options =
    - directio (dio)
    - asyncio (aio)
    - setall



## Init.ora Parameter dbwr\_io\_slaves

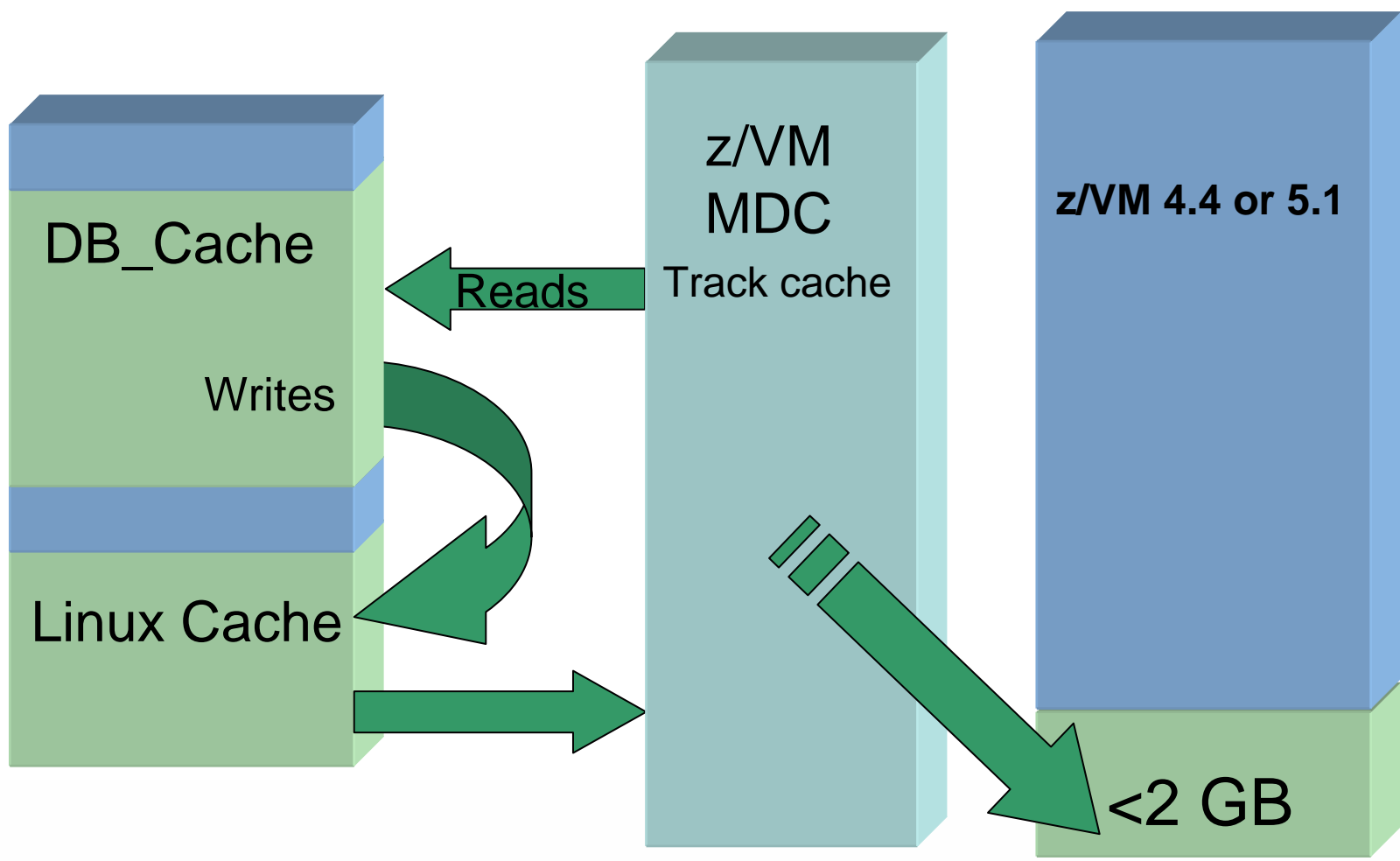
- Oracle9i for s/390 was not built with support for Asynchronous I/O
  - Use dbwr\_io\_slaves=4 in the init.ora
  - Use more slaves as needed to handle I/O
- Changes for Oracle Database 10g
  - SuSE SLES8 is built with async I/O
  - Oracle database server kernel built with Async I/O
    - Run make after install to enable Async I/O
    - Set options in init.ora

## Use of filesystemio\_options=setall

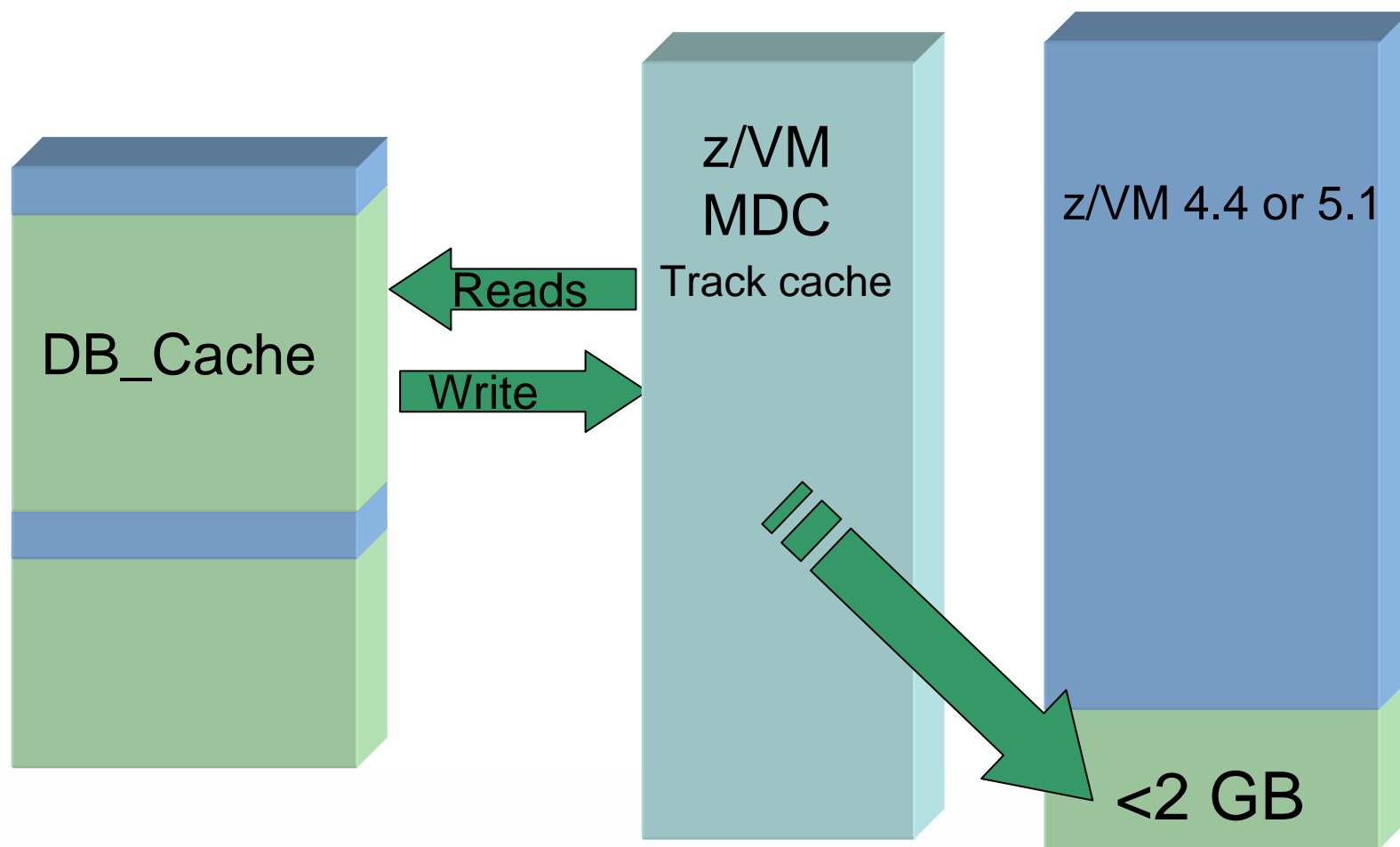
- filesystemio\_options=setall
  - Enables asyncio and directio for filesystems
  - These are unrelated options, but set through one parameter
    - setall
    - asyncio
    - directio
- Our experiences
  - SLES9 initial testing shows that setall works well
  - SLES8 use directio alone for best results



# I/O Without directIO Option

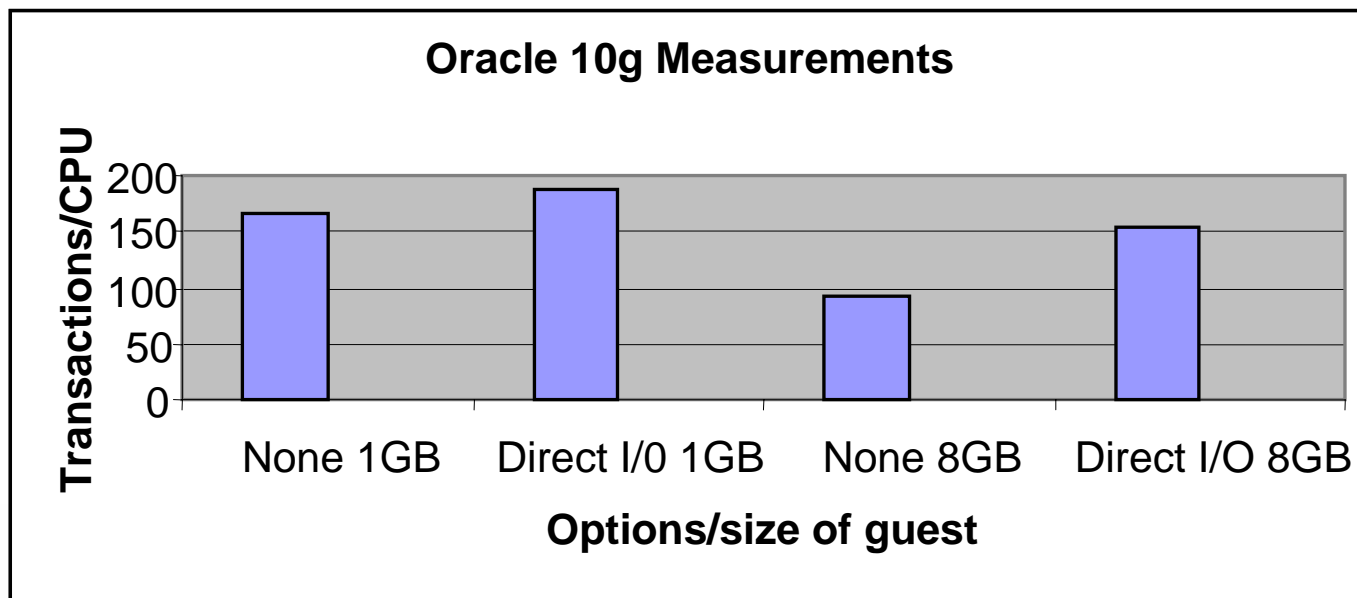


## I/O with directIO Option



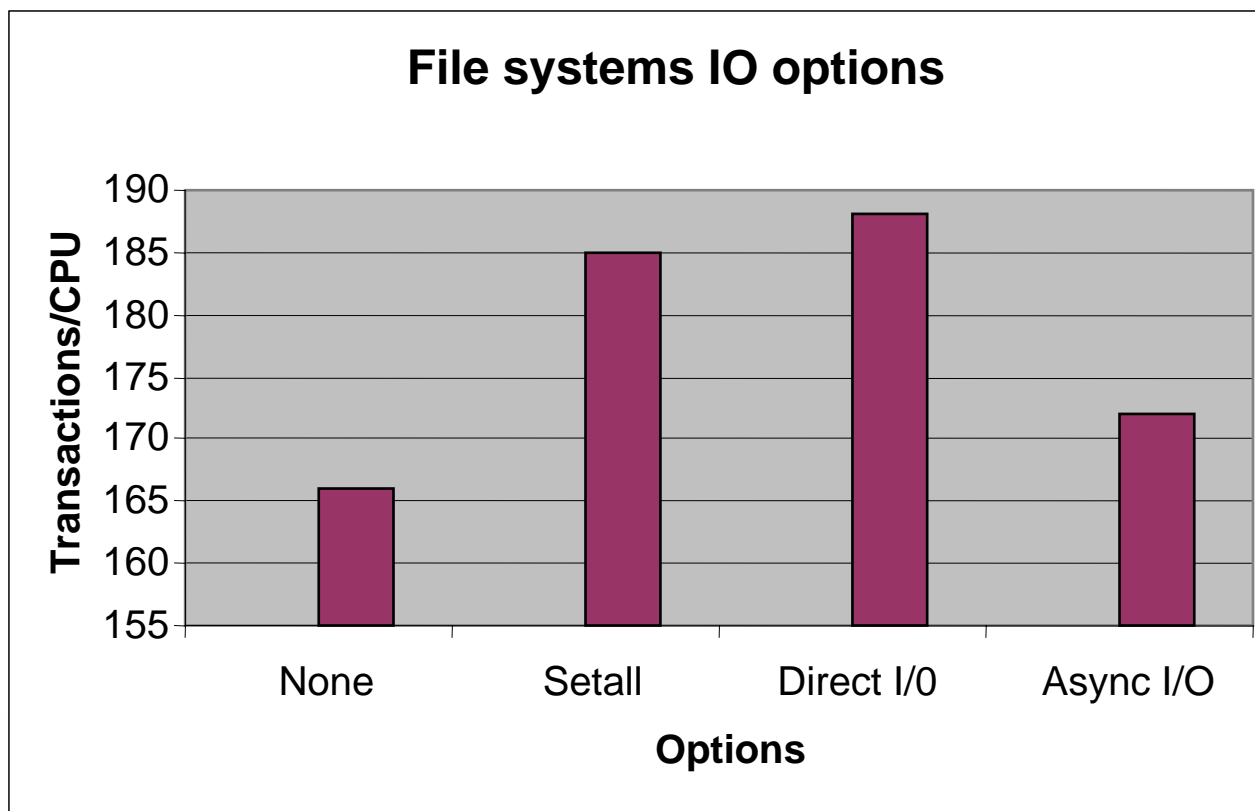
# Measurements with directIO

- Oracle Database10g
- SuSE SLES8



# Oracle init.ora options

- Oracle Database 10g
- SuSE SLES8

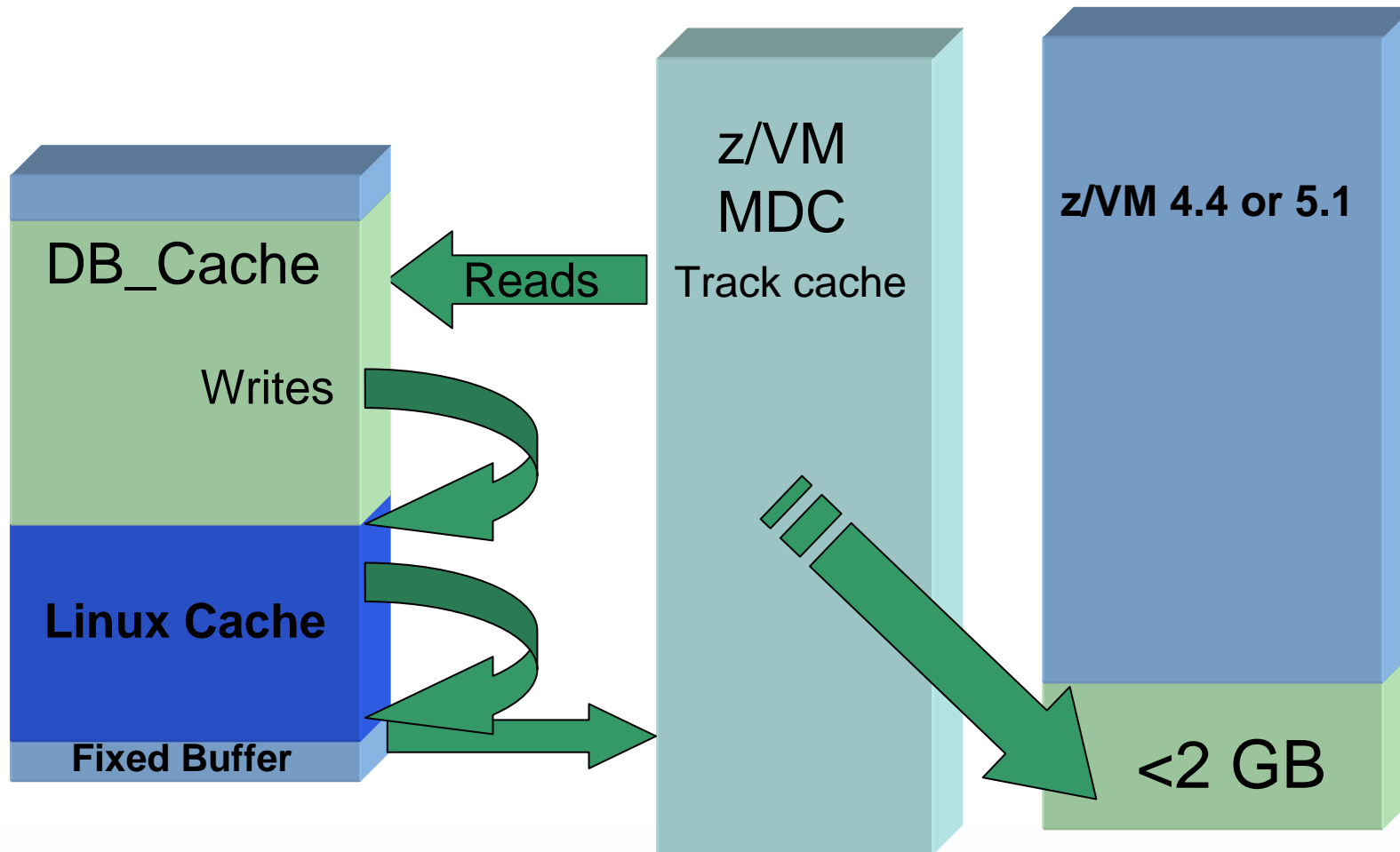


## Other Considerations

- Please be aware if you are using SLES8
  - And you run 10g, you may not experience problems.
  - Or if you run 9i, even though it is 31bit, you may still exhibit symptoms.
- Fixed Buffer Patch
  - Applies to ECKD DASD
    - Tunable
  - Available in SuSE SLES9 SP1 and RHEL4
    - Oracle Database certification for these distro's projected
- z/VM
  - New releases on the horizon



# Fixed Buffer Patch



**Applies to ECKD DASD only**

# DASD Considerations

- Distribute data across arrays
  - Not doing this can cause BIG performance issues
- LVM Striping
  - ESCON one stripe per channel
  - FICON one stripe per volume
  - FCP one stripe per volume or LUN
  - Optimal strip sizes
    - 16K or 32K (but may vary depending on your workload)
- Consider use of Parallel Access Volumes (PAV) for concurrent I/O to same volume
  - O/S still issues single I/Os to devices
  - SCU can do I/Os in parallel for reads and writes to different domains
  - Only valid with LVM and ESS

## Recommendations for Running 10g

- z/VM
  - Make sure all current patches applied
  - Use Best practices for using functions such as VSwitch in z/VM
- zSeries
  - SCSI/FCP is a good choice but
    - Fixed Buffer patch is only ECKD DASD
- Linux
  - Use the 251 (or higher) kernel on SLES8 (Oracle developed on 112)
- Oracle
  - Insure that async I/O is enabled (run make after installing)
  - Use
    - filesystemio\_options=setall on SLES9 (when certified)
    - filesystemio\_options=directio on SLES8

# What Do I Run in a Linux Guest?

- Production - Recommendation
  - Only a database
  - Only one database
  - Put app servers in separate guest
  - If you chose to configure other, monitor paging
- Test/Dev/etc
  - Single database is better
  - Multiple DBs can be OK
  - You may experience performance issues with multiple DBs
    - If all are doing queries you should be OK
    - If one starts doing loads or imports – problem!
- Why
  - Linux does not differentiate between instances
  - If Linux needs memory (i.e. import or load)
    - It will page other SGAs
- Is there a fix?
  - Init.ora parameter lock\_sga in Oracle9i – but Oracle needs to run as root
  - Investigating Oracle Database 10g

## Monitoring Performance

- z/VM has the only accurate view of what resources each guest is consuming. Any investigation needs to start in z/VM.
- The goal of z/VM is convince each guest that it is running alone on the machine and has all of its own resources.
  - Therefore, no guest can possibly tell what is really happening on the system.
  - Don't under estimate the value of sar or vmstat
- Consider either
  - z/VM Toolkit
  - EASMON from Velocity Software
    - <http://www.velocity-software.com/>
  - You can also use CP commands when you suspect problems
- Use Oracle Statspack to monitor your instances



# Key Success Factors

- Memory is critical
  - Small virtual guest size – less is better
  - Let z/VM manage memory not Linux
    - Use VDISK for Linux swap
- Monitor resource
  - Understand limits
  - Make changes (i.e.tuning) that may be necessary
- Paging and swap space necessary
  - Both should use memory devices
  - Use Best Practices for setting up paging space
- Avoid I/O bottlenecks
  - Distribute data in the ESS across arrays
  - Consider striping with LVM
  - Faster is better



# Information Sources

- <http://www.ibm.com/redbooks>
  - SG24-6552 Experiences with Oracle9i for Linux on zSeries
  - REDP-3859 Experience Installing Oracle Database 10g on Linux for zSeries
  - SG24-7023 Linux on IBM eServer zAeries and S/390; Best Security Practices
- <http://www.oracle.com/ibm>
  - IBM platform information
- <http://otn.oracle.com>
  - (Select “download code”)
- <http://www.vm.ibm.com/perf/tips>
  - General z/VM Tuning Tips
- <http://www.vm.ibm.com/perf/tips/2GSTORAG.HTML>
  - 2GB I/O information
- <http://www-124.ibm.com/developerworks/oss/linux390/index.shtml>
  - Lot’s of information on Linux for zSeries
- [http://awlinux1.alphaworks.ibm.com/developerworks/linux390/perf/tuning\\_rec\\_dasd.shtml](http://awlinux1.alphaworks.ibm.com/developerworks/linux390/perf/tuning_rec_dasd.shtml)
  - Hints and Tips for Selecting and Tuning I/O options

## Acknowledgements

- Many thanks to the following contributors
  - Oracle – Barry Perkins, Mike Morgan, Betsie Spann
  - IBM BOE – Martin Kammerer, Ulrich Weigand
  - IBM Storage – Pat Blaney
  - IBM team – Bruce Frank, Tom Russell, Kathryn Arrell, Denny Dutcavich, Laurent Dupin
  - Velocity Software – Barton Robinson
  - z/VM Performance– Bill Bitner, Chuck Morse, Rich Lewis, Ann Jackson