IBM Systems & Technology Group

z/VM System Limits

Revision 2011-07-22 (BKW)

IBM z/VM Performance Evaluation
Bill Bitner bitnerb@us.ibm.com
Brian Wade bkw@us.ibm.com

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DBE, e-business logo, ESCON, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/390, System z, VM/ESA, VSE/ESA, Websphere, xSeries, z/OS, zSeries, z/VM

The following are trademarks or registered trademarks of other companies

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
LINUX is a registered trademark of Linus Torvalds
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.
Intel is a registered trademark of Intel Corporation
* 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.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

Acknowledgements

- **Contributors to this material**
  - Bill Bitner
  - Wes Ernsberger (now retired - ☹ - we miss you Wes)
  - Bill Holder
  - Virg Meredith
  - Brian Wade

Agenda

- **Describe various limits**
  - Architected
  - Consumption
  - Latent
- **Show how to keep tabs on consumables**
- **Discuss limits that may be hit first**

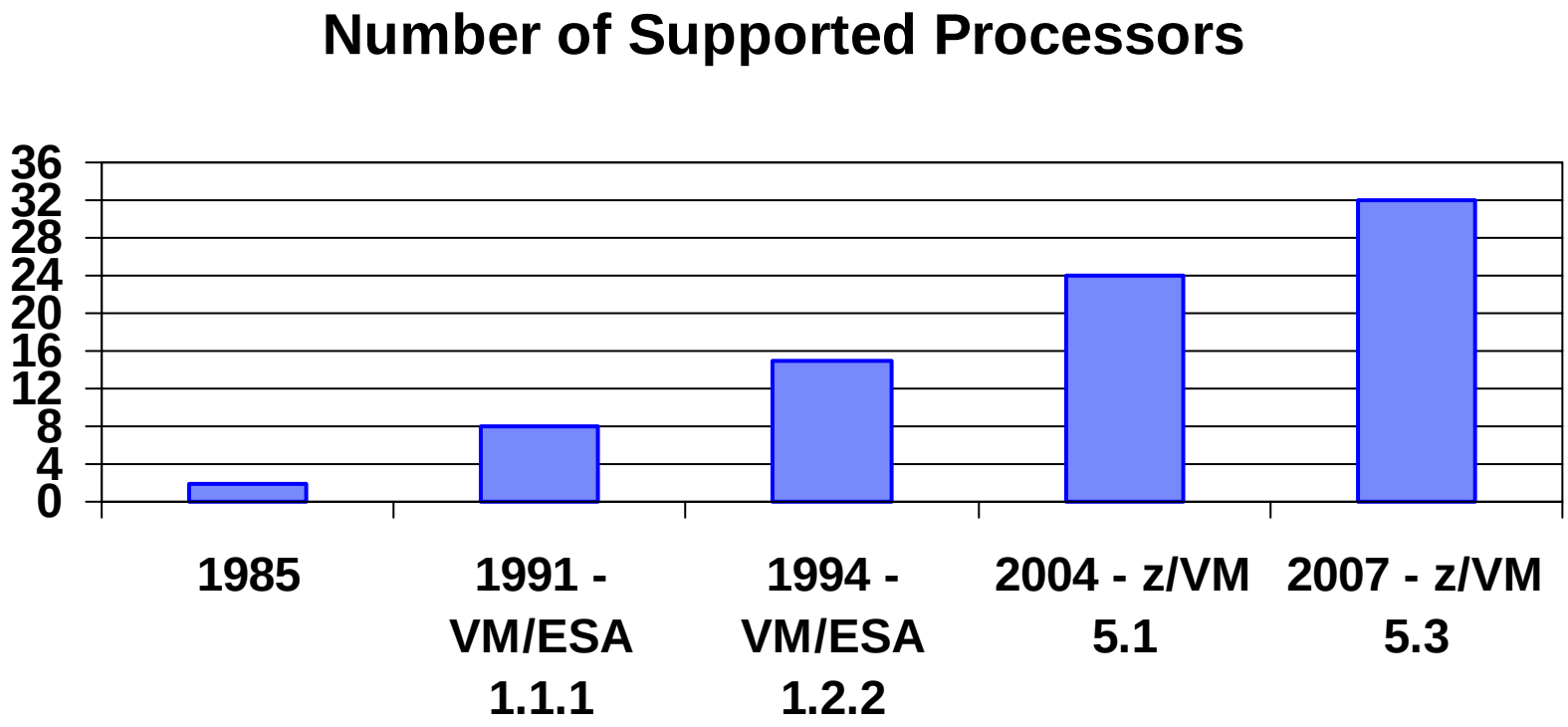
Limits

- **Processors**
- **Memory**
- **I/O**
- **Others**
- **Latent limits**

Processors

- **Processors (architected): 64**
  - Includes all engine types (CP, zAAP, zIIP, IFL…)
- **Processors (hardware): 80 (z196), 64 (z10 EC), 54 (z9 EC)**
- **Logical processors in a partition (hardware): 80 (z196), 64 (z10), 54 (z9)**
- **Logical processors in a z/VM partition (unsupported): 64 (z196, z10), 54 (z9)**
- **Logical processors in a z/VM partition (support statement): 32**
- **Master processor (architected): 1**
  - 100%-utilized master is the issue
  - z/VM will elect a new master if master fails
- **Virtual processors in single virtual machine (architected): 64**
  - But $N_{Virtual} > N_{Logical}$ is not usually practical
- **Number of partitions: 60 (z196, z10, z9)**

Processor Scaling



Processors: FCX100 CPU

```
FCX100  Run 2007/09/06 14:00:28          CPU
                                         General CPU Load and User Transactions
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                    CPU 2094-700
For   3180 Secs 00:53:00                                    z/VM   V.5.3.0 SLU 0701
```

```
 CPU Load                                                    Vector Facility   Status or
 PROC TYPE %CPU  %CP %EMU %WT %SYS %SP %SIC %LOGLD  %VTOT %VEMU  REST ded. User
 P00  IFL    16   2   14  84   2   0   84     16      ..     ..    ... ........
 P15  IFL    18   2   16  82   1   0   80     18      ..     ..    ... ........
 P14  IFL    18   2   16  82   1   0   80     18      ..     ..    ... ........
 P13  IFL    18   2   16  82   1   0   80     18      ..     ..    ... ........
 P12  IFL    18   2   16  82   1   0   81     18      ..     ..    ... ........
 P11  IFL    18   2   17  82   1   0   80     19      ..     ..    ... ........
  ... truncated …
```

1. T/V ~ 18/16 = 1.13 a little CP overhead here
2. Master does not seem unduly burdened

Processors: FCX114 USTAT

```
FCX114  Run 2007/09/06 14:00:28          USTAT                                          Page  186
                                         Wait State Analysis by User
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                                          CPU 2094-700
For   3180 Secs 00:53:00                                                  z/VM   V.5.3.0 SLU 0701
```

										<-SVM and->							<--%Time spent in-->				Nr of
Userid	%ACT	%RUN	%CPU	%LDG	%PGW	%IOW	%SIM	%TIW	%CFW	%TI	%EL	%DM	%IOA	%PGA	%LIM	%OTH	Q0	Q1	Q2	Q3	E0-3	Users
>System<	64	1	0	1	0	0	0	83	0	0	0	3	0	0	0	10	1	29	10	57	0	211
TCPIP	100	0	0	0	0	0	0	0	0	3	0	97	0	0	0	0	3	0	0	0	0	
RSCSDNS1	100	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	
SNMPD	100	0	0	0	0	0	0	0	0	2	0	98	0	0	0	0	2	0	0	0	0	
SZVAS001	100	2	0	0	0	0	0	97	0	0	0	0	0	0	0	1	0	3	12	85	0	

1. %CPU wait is very low – nobody is starved for engine
2. %TIW is “test idle wait” – we are waiting to see if queue drop happens

Memory (1 of 3)

- **Central storage**
  - CEC limit for a partition: 512 GB-HSA (z9), 1 TB (z10, z196)
  - Supported for a z/VM partition: 256 GB
  - Unsupported for a z/VM partition: 512 GB
- **Expanded storage**
  - CEC limit for a partition: 16 TB
  - Supported for a z/VM partition: 128 GB
  - Unsupported for a z/VM partition: 600 – 700 GB
  - See http://www.vm.ibm.com/perf/tips/storconf.html
- **Virtual machine size (hardware):**
  - z196, 16 TB; z10, 8 TB; z9, 1 TB
  - On z990 and z900, 256 GB

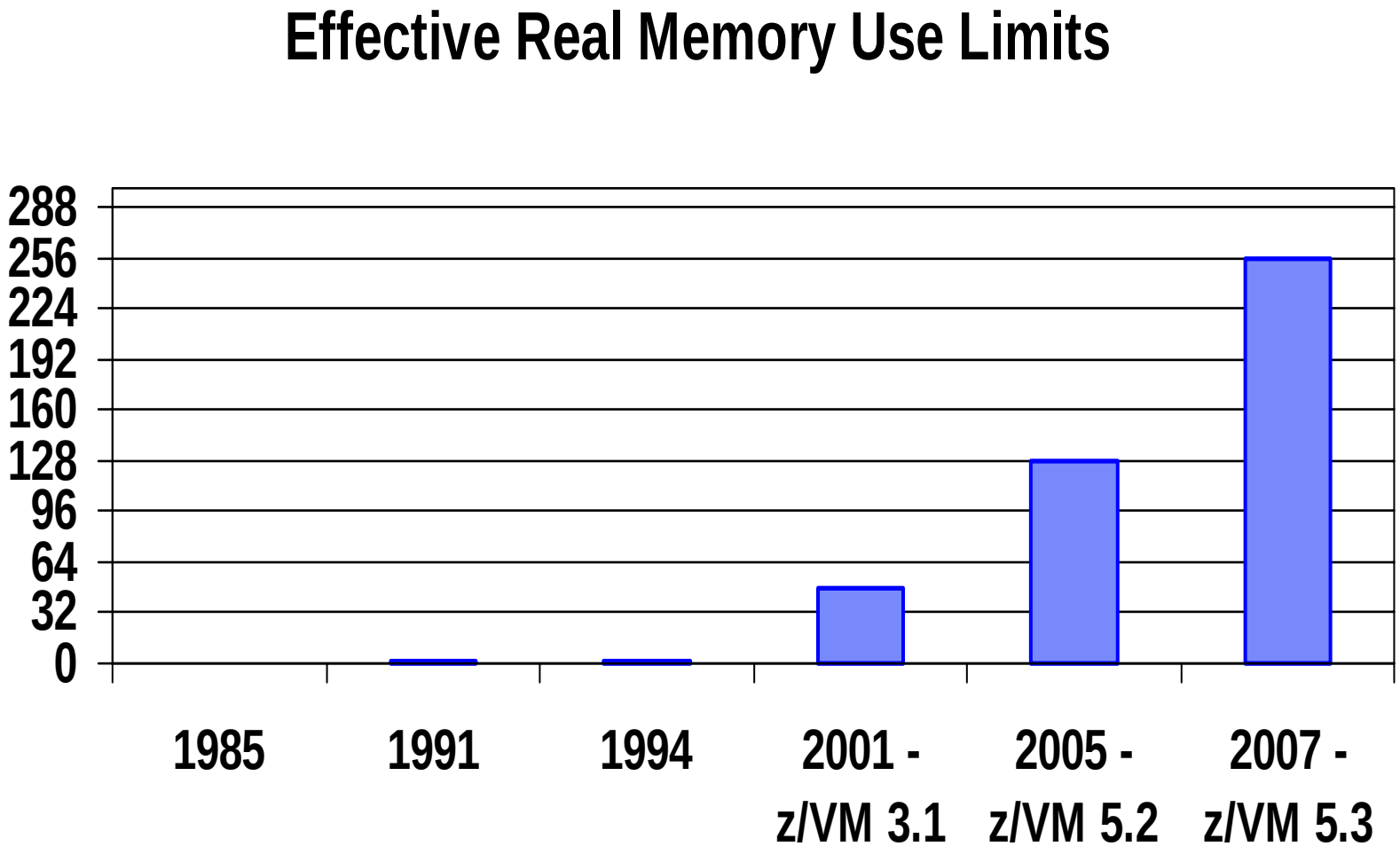
Memory (2 of 3)

- **Instantiated guest real limit imposed by PTRM space limits (architected): 8 TB**
  - 16 4-GB PTRM spaces; each PTRM space can map 512 GB of guest real
- **Virtual to real ratio (practical): about 2:1 to 3:1**
  - Assumes guests that tend to use all of their memory
  - Some performance-sensitive production workloads will require 1:1
  - If you really, really do your homework on your paging subsystem you can push this up somewhat
  - VMRM-CMM can help with this too (it encourages Linux to Diag x'10' its guest real)
  - Many factors come into play here:
    - Relative mix of active and idle guests
    - Workload's or SLA's sensitivity to delays
    - Exploitation of shared memory
  - For guidance, see http://www.vm.ibm.com/perf/tips/memory.html

Memory (3 of 3)

- **Paging space (architected) (optimal when <= 50% allocated):**
  - 11.2 TB for ECKD
  - 15.9 TB for Emulated FBA on FCP SCSI
- **Paging volumes: 255**
- **Concurrent paging I/Os per paging volume: 1 for ECKD, >1 for EDEV (have observed 1.6)**
- **System Execution Space (SXS) (architected): 2 GB**
  - For practical purposes it is 2 GB, but there are structures in the space placed above 2 GB
- **DCSS size (architected):**
  - Each segment can be up to 2047 MB
  - Segments can map into > 2 GB, starting in z/VM 5.4
- **Minidisk Cache (architected): 8 GB**
  - Practically somewhat less, ~ 2 GB

Memory Scaling



Page Slots: FCX146 AUXLOG

```
FCX146  Run 2007/09/06 14:00:28          AUXLOG
                                         Auxiliary Storage Utilization, by Time
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00
For   3180 Secs 00:53:00
```

	<Page Slots>		<Spool Slots>		<Dump Slots>		<----- Spool Files ----->				<Average MLOAD>	
Interval	Total	Used	Total	Used	Total	Used	<-Created-->		<--Purged-->		Paging	Spooling
End Time	Slots	%	Slots	%	Slots	%	Total	/s	Total	/s	msec	msec
>>Mean>>	87146k	44	5409096	52	0	..	54	.02	54	.02	2.8	.8
09:08:00	87146k	44	5409096	52	0	..	1	.02	1	.02	2.3	.8
09:09:00	87146k	44	5409096	52	0	..	1	.02	1	.02	3.9	.8
09:10:00	87146k	44	5409096	52	0	..	1	.02	1	.02	3.6	.8
09:11:00	87146k	44	5409096	52	0	..	1	.02	1	.02	2.8	.8
09:12:00	87146k	44	5409096	52	0	..	1	.02	1	.02	2.9	.8

1. This system is using 44% of its page slots.
2. 87146k slots / (256 slots/MB) = 332 GB of paging space.
3. 332 GB * 44% = 146 GB in use.

DASD I/O: FCX109 DEVICE CPOWNED

```
FCX109  Run 2007/09/06 14:00:28        DEVICE   CPOWNED                                          Page  152
                                       Load and Performance of CP Owned Disks
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                                    CPU 2094-700
For   3180 Secs 00:53:00                                                     z/VM   V.5.3.0 SLU 0701
```

```
 Page / SPOOL Allocation Summary
 PAGE slots available       87146k          SPOOL slots available      5409096
 PAGE slot utilization          44%         SPOOL slot utilization          52%
 T-Disk cylinders avail.   .......          DUMP slots available             0
 T-Disk space utilization      ...%         DUMP slot utilization           ..%
```

< Device Descr. ->					<---------- Rate/s ---------->						User		Serv	MLOAD	Block	%Used
Addr	Devtyp	Volume Serial	Area Type	Area Extent	Used %	<--Page---> P-Rds	P-Wrt	<--Spool--> S-Rds	S-Wrt	Total	SSCH +RSCH	Inter feres	Queue Lngth	Time /Page	Time Resp	Page Size	for Alloc
F08B	3390	VS2P49	PAGE	0-3338	45	2.6	1.7	...	...	4.4	1.6	1	.02	2.4	2.4	7	89
F090	3390	VS2P69	PAGE	0-3338	45	2.7	1.6	...	...	4.3	1.6	1	0	2.7	2.7	7	84

1. Interesting fields: slot utilization, MLOAD, Queue Lngth.
2. See a wait queue? See also %PGW and %LDG in FCX114 USTAT.

V:R Ratio and Segment Tables: FCX113 UPAGE

```
FCX113  Run 2007/09/06 14:00:28        UPAGE                                                    Page  173
                                       User Paging Activity and Storage Utilization
From 2007/09/04 09:07:00                                                                 VS2
To   2007/09/04 10:00:00                                                                 CPU 2094-700   SN 2BFBD
For   3180 Secs 00:53:00                                                                 z/VM   V.5.3.0 SLU 0701
```

	Data	<--- Paging Activity/s --->			<--- Number of Pages --->											
	Spaces	<Page Rate>		Page	<--Page Migration-->						<-Resident->		<--Locked-->			Stor	Nr of
Userid	Owned	Reads	Write	Steals	>2GB>	X>MS	MS>X	X>DS	WSS	Resrvd	R<2GB	R>2GB	L<2GB	L>2GB	XSTOR	DASD	Size	Users
>System<	.0	1.7	1.1	4.1	.0	2.4	3.7	1.4	122050	0	2347	106962	6	24	12240	179131	1310M	212
ABCDEFGH	.0	.0	.0	.0	.0	.0	.1	.0	13	0	0	0	0	0	483	254	32M	
DATAMOVA	.0	.0	.0	.0	.0	.5	.5	.0	147	0	0	0	0	0	220	368	32M	
DATAMOVB	.0	.0	.0	.0	.0	.6	.6	.0	192	0	0	0	0	0	220	366	32M	
DATAMOVC	.0	.0	.0	.0	.0	.6	.6	.0	191	0	0	0	0	0	220	369	32M	
DATAMOVD	.0	.0	.0	.0	.0	.6	.6	.0	189	0	0	0	0	0	220	362	32M	

1. Resident guest pages = (2347 + 106962) * 212 = 88.3 GB
2. V:R = (1310 MB * 212) / 91 GB = 2.98 (FCX103 shows 91 GB central)
3. Segment table pages: hard to say. Conservatively:
   212 guests * (4 ST/guest * 4 pg/ST) = 13 MB

PTRM Space: FCX134 DSPACESH

```
FCX134  Run 2007/09/06 14:00:28          DSPACESH
                                 Shared Data Spaces Paging Activity

From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                                    CPU 2094-700
For   3180 Secs 00:53:00                                                    z/VM   V.5.3.0 SLU
   0701
```

```
   _
```

```
 _____                          .     .     .     .     .     .     .     .     .     .     .     .     .     .     .
                                <--------- Rate per Sec. ---------> <------------------Number of Pages------------------>
 Owning                                                                    <--Resid--> <-Locked--> <-Aliases->
 Userid     Data Space Name      Pgstl Pgrds Pgwrt X-rds X-wrt X-mig Total Resid R<2GB  Lock L<2GB Count Lockd XSTOR  DASD
 >System<   --------              .026  .016  .001  .015  .026  .000  103k  1208    51     0     0     0     0    34  4981
 SYSTEM     FULL$TRACK$CACHE$1    .000  .000  .000  .000  .000  .000  524k     0     0     0     0     0     0     0     0
 SYSTEM     ISFCDATASPACE         .000  .000  .000  .000  .000  .000  524k   113     8     8     8   113   100     0    27
 SYSTEM     PTRM0000             4.257  .492  .442 3.957 4.036  .000 1049k  386k 15885     0     0     0     0  5195  683k
 SYSTEM     REAL                  .000  .000  .000  .000  .000  .000   24M     0     0     0     0     0     0     0     0
 SYSTEM     SYSTEM                .080  .001  .034  .079  .080  .000  524k    45    10     0     0    44     0    47  510k
```

1. PTRM space = 386,000 pages = 1.47 GB of PGMBKs.
2. This maps 128 * 1.47 GB = 188.5 GB of guest storage.
3. z/VM 5.3 and later can have >2 GB of PGMBKs.

Real Memory: FCX254 AVAILLOG

```
FCX254  Run 2007/09/06 14:00:28        AVAILLOG                                                    Page  190
                                       Available List Management, by Time
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                                          CPU 2094-700
For   3180 Secs 00:53:00                                                          z/VM   V.5.3.0 SLU 0701
```

	<---------- Available List Management ---------->																
	<---- Thresholds ----->				<---------- Page Frames ---------->						<-Times->		<----------- Replenishment ------------>				Perct
Interval	<---Low--->		<--High--->		<Available>		<Obtains/s>		<Returns/s>		<-Empty->		<--Scan1-->		<--Scan2-->		<-Em-Scan->		Scan	Emerg
End Time	<2GB	>2GB	<2GB	>2GB	<2GB	>2GB	<2GB	>2GB	<2GB	>2GB	<2GB	>2GB	Compl	Pages	Compl	Pages	Compl	Pages	Fail	Scan
>>Mean>>	20	7588	5820	13388	5130	7678	323.3	857.4	311.5	844.8	0	0	27	1381k	63	1380k	58	84490	82	88
09:08:00	20	7680	5820	13480	6665	15122	353.3	838.5	353.2	1007	0	0	0	43091	3	26491	0	0	3	100
09:09:00	20	7680	5820	13480	3986	5496	163.1	640.2	108.9	442.7	0	0	1	14528	0	0	0	0	0	0
09:10:00	20	7681	5820	13481	6622	9542	222.4	556.1	257.0	598.3	0	0	0	30103	2	8868	0	0	1	100
09:11:00	20	7681	5820	13481	4982	6710	292.1	615.2	248.8	533.6	0	0	0	21246	0	8547	1	3989	1	100
09:12:00	20	7681	5820	13481	4769	1560	284.9	946.9	254.4	830.0	0	0	0	18253	0	22438	2	656	1	100

1. Pct ES = 88% generally this system is tight on storage (might just be intense use of VDISKs – where would you look?)
2. Scan fail >0 generally this system is tight on storage
3. Times Empty = 0 this indicates it isn’t critical yet

SXS Space: FCX261 SXSAVAIL

```
FCX261  Run 2007/09/06 14:00:28          SXSAVAIL                                                    Page  261
                                         System Execution Space Page Queues Management
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                                              CPU 2094-700
For   3180 Secs 00:53:00                                                              z/VM   V.5.3.0 SLU 0701
```

	<-- Backed <2GB Page Queue -->					<-- Backed >2GB Page Queue -->					<---------- Unbacked Page Queue ----------								
Interval	Avail	<-Pages/s-->		<Preferred>		Avail	<-Pages/s-->		<Preferred>		Avail	<-Pages/s-->		<Preferred>		<---- Replenishment ---->			
End Time	Pages	Taken	Return	Used	Empty	Pages	Taken	Return	Used	Empty	Pages	Taken	Return	Used	Empty	Thres	Att/s	Stolen	MinPgs
>>Mean>>	26	.513	.509	.513	.000	3	1.798	1.804	1.798	4.114	466946	130.3	130.1	126.2	.000	128	.000	128	...
09:08:00	26	.483	.383	.483	.000	0	1.650	1.650	1.650	3.667	467829	128.2	127.3	124.5	.000	128	.000	128	...
09:09:00	26	.500	.500	.500	.000	0	.583	.583	.583	3.067	465679	120.8	84.98	117.8	.000	128	.000	128	...
09:10:00	27	.517	.533	.517	.000	0	1.183	1.183	1.183	4.000	467657	109.1	142.1	105.1	.000	128	.000	128	...
09:11:00	27	.517	.517	.517	.000	0	1.633	1.633	1.633	2.917	467632	137.2	136.8	134.3	.000	128	.000	128	...
09:12:00	29	.450	.483	.450	.000	0	2.000	2.000	2.000	3.383	467654	129.9	130.2	126.5	.000	128	.000	128	...
09:13:00	27	.517	.483	.517	.000	0	2.483	2.483	2.483	3.550	467698	139.3	140.0	135.7	.000	128	.000	128	...
09:14:00	25	.550	.517	.550	.000	0	2.000	2.000	2.000	2.750	465651	119.0	84.92	116.3	.000	128	.000	128	...

1. How we touch guest pages: (1) 64-bit; (2) AR mode; (3) SXS.
2. There are 524,288 pages in the SXS.
3. This system has 466,000 SXS pages available on average.

MDC: FCX178 MDCSTOR

```
FCX178  Run 2008/04/15 10:00:22          MDCSTOR           Page   76
                                         Minidisk Cache Storage Usage, by Time
From 2008/04/15 09:47:11
To   2008/04/15 10:00:11                      CPU 2084-320   SN 17F2A
For    780 Secs 00:13:00                      z/VM   V.5.3.0 SLU 0000

            <---------------- Main Storage Frames ----------------->
 Interval         <--Actual--->    Min    Max  Page    Steal
 End Time   Ideal   <2GB   >2GB    Set    Set Del/s Invokd/s   Bias
 >>Mean>>   5839k  82738  1354k      0  7864k     0     .000   1.00
 09:57:41   5838k 119813  1932k      0  7864k     0     .000   1.00
 09:58:11   5838k 119813  1932k      0  7864k     0     .000   1.00
 09:58:41   5838k 119825  1932k      0  7864k     0     .000   1.00
 09:59:11   5838k 119825  1932k      0  7864k     0     .000   1.00
 09:59:41   5838k 119825  1932k      0  7864k     0     .000   1.00
 10:00:11   5838k 119837  1932k      0  7864k     0     .000   1.00
```

- Xstore not used for this configuration so edited out from report.
- Add up the pages in main storage for this run and you get about 8 GB in use for MDC.

MDC Spaces: FCX134 DSPACESH

```
FCX134  Run 2008/04/15 10:00:22          DSPACESH
                                         Shared Data Spaces Paging Activity
From 2008/04/15 09:47:11
To   2008/04/15 10:00:11
For    780 Secs 00:13:00                 This is a performance report for system XYZ
```

```
 ______                                  .      .     .     .     .     .     .     .     .     .
                                            <------------------Number of Pages------------------>
 Owning                            Users         <--Resid--> <-Locked--> <-Aliases->
 Userid    Data Space Name         Permt  Total Resid R<2GB  Lock L<2GB Count Lockd XSTOR  DASD
 >System<  --------                    0  1507k  5665   101     0     0   100     0     0     0
 SYSTEM    FULL$TRACK$CACHE$1          0   524k     0     0     0     0     0     0     0     0
 SYSTEM    FULL$TRACK$CACHE$2          0   524k     0     0     0     0     0     0     0     0
 SYSTEM    FULL$TRACK$CACHE$3          0   524k     0     0     0     0     0     0     0     0
 SYSTEM    FULL$TRACK$CACHE$4          0   524k     0     0     0     0     0     0     0     0
 SYSTEM    ISFCDATASPACE               0   524k     0     0     0     0     0     0     0     0
 SYSTEM    PTRM0000                    0  1049k 44489     0     0     0     0     0     0     0
 SYSTEM    REAL                        0  7864k     0     0     0     0     0     0     0     0
 SYSTEM    SYSTEM                      0   524k   805   787     0     0   800     0     0     0
 SYSTEM    VIRTUAL$FREE$STORAGE        0   524k    23    23     0     0     0     0     0     0
```

- You'll see the address spaces used for MDC (track cache)
- Values here are zero for page counts, ignore.

Reorder Processing - Background

- ***Page reorder* is the process of managing user-frame-owned lists as input to demand scan processing.**
  - It includes resetting the HW reference bit.
  - Serializes the virtual machine (all virtual processors).
  - In all releases of z/VM
- **It is done periodically on a virtual machine basis.**
  - Even if the system is not paging.
- **The cost of reorder is proportional to the number of resident frames for the virtual machine.**
  - Roughly 130 ms/GB resident on z10
  - Delays of ~1 second for guest having 8 GB resident
  - This can vary for different reasons +/- 40%

Reorder Processing - Diagnosing

- **Performance Toolkit**
  - Check resident page fields (“R<2GB” & “R>2GB”) on FCX113 UPAGE report
    - Remember, reorder works against the resident pages, not total virtual machine size.
  - Check Console Function Mode Wait (“%CFW”) on FCX114 USTAT report
    - A virtual machine may be brought through console function mode to serialize reorder. There are other ways to serialize for reorder and there are other reasons for CFW, so this is not conclusive.
- **REORDMON**
  - Available from Bill Bitner or http://www.vm.ibm.com/download/packages/
  - Works against MONWRITE data for all monitored virtual machines
  - Works in real time for a specific virtual machine
  - Provides how often reorder processing occurs in each monitor interval

Reorder Processing - Mitigations

- **Try to keep the virtual machine as small as possible.**
- **Virtual machines with multiple applications may need to be split into multiple virtual machines with fewer applications.**
- **APAR VM64774 is now available**
  - Implements a flexible SET REORDER function
- **See http://www.vm.ibm.com/perf/tips/reorder.html for more details.**

I/O (1 of 3)

- **Number of subchannels in a partition (aka device numbers) (architected): 65,535**
- **CHPIDs per server:**
  - z9 EC or z10 EC: 1024 ESCON, 336 FICON Express 4, 96 OSA Express 3, 16 HiperSockets
  - z196: 240 ESCON, 336 FICON Express8, 336 FICON Express4, 96 OSA Express3, 32 HiperSockets
- **Device numbers per disk volume**
  - Without PAV, 1
  - With PAV or HyperPAV, 256 (base plus 255 aliases, but can use only 7 aliases for ESCON)
- **Virtual devices per virtual machine: 24576 (24K)**
- **Concurrent real I/Os per ECKD disk volume: 1 usually, but more with PAV or HyperPAV if of guest origin**
- **Concurrent real I/Os per chpid (aka “open exchange limit”)**
  - 1 for ESCON
  - 32 for FICON Express
  - 64 for FICON Express2 and later
- **I/O rates:**
  - Fastest FICON is an 8 Gb/sec link (translates to about 640 MB/sec)
  - About 1-2 msec per I/O are required for a nominal DASD I/O from a z9 to a 2107 (aka rates of 500-1000/sec/device)
- **Ref: http://www-03.ibm.com/systems/z/hardware/**

I/O (2 of 3)

- **ECKD volume sizes**
  - Largest ECKD minidisk that can contain a CMS file system (architected): 32768 cylinders (22.5 GB)
  - Largest ECKD volume, period: 65536 cylinders (43 GB)
- **EFBA volume sizes**
  - Largest EFBA minidisk that can contain a CMS file system (architected): 381 GB (tough to beat for archiving)
    - Practical limit is 22 GB due to CMS in-memory file system structures under 2 GB, unless very few, very large files
  - Largest EDEV CP can use: 1024 GB (but PAGE, SPOL, DRCT must be below 64 GB line on volume)
  - Largest EDEV, period: 2^{32} FB-512 blocks (2048 GB)
- **VDISK size (architected): 2 GB minus eight 512-byte blocks**
- **Total VDISK (architected): 2 TB**
- **Single VSWITCH OSAs: 8**

I/O (3 of 3)

- **Number of files on a user's accessed CMS disks or directories**
  - 262,144 (16 MB / 64 bytes per FST)
  - You cannot use all of your <16MB storage for FSTs anyway
  - Files residing in SFS DIRCONTROL directories *in a data space* do not charge to this limit because those FSTs aren't below 16 MB in CMS storage
- **SFS limits**
  - 32767 storage groups
  - 2^{32}-1 blocks (16 TB) per file pool, storage group, file space, file
  - 2 GB of data per DIRCONTROL-directory-in-data-space
  - No architected limit on numbers of files or users
  - See Appendix B in CMS File Pool Planning, Administration, and Operation

DASD I/O: FCX108 DEVICE

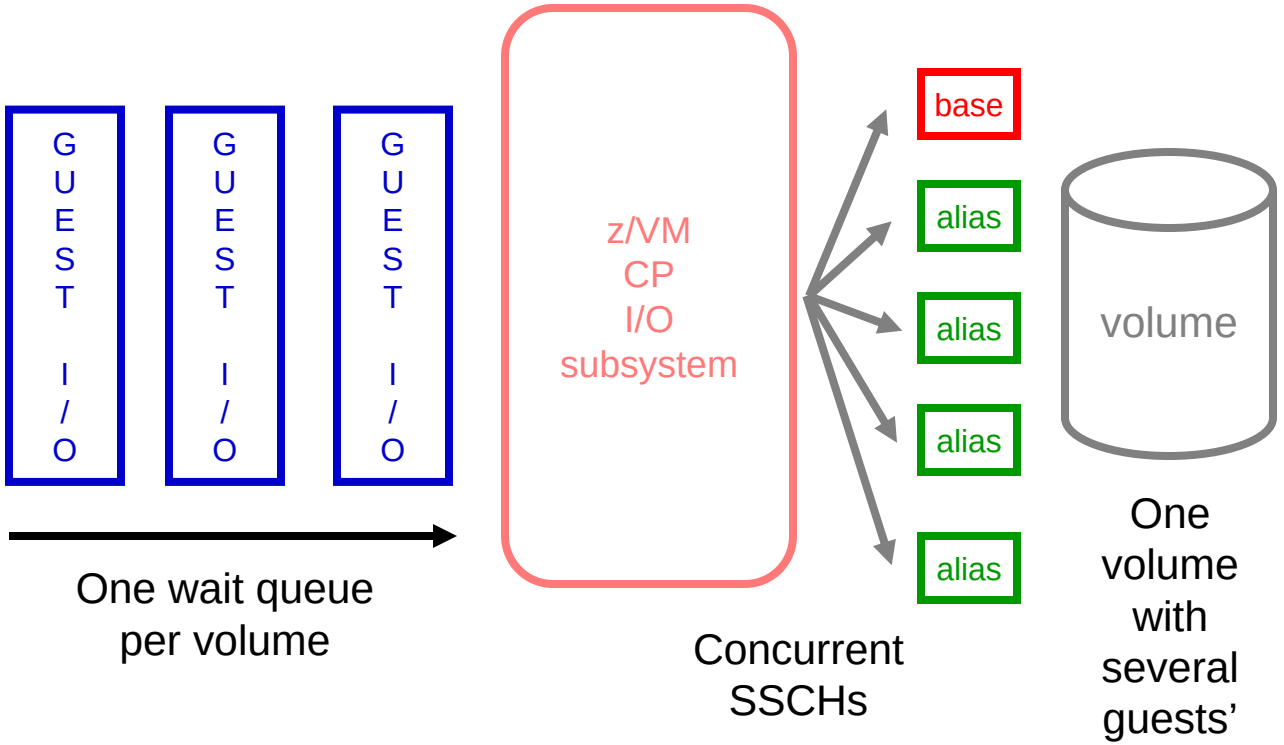
```
FCX108  Run 2007/09/06 14:00:28          DEVICE                                          Page  110
                                         General I/O Device Load and Performance
From 2007/09/04 09:07:00
To   2007/09/04 10:00:00                                                  CPU 2094-700   SN
For   3181 Secs 00:53:01                                                   z/VM   V.5.3.0 SLU 0701
```

<-- Device Descr. -->			Mdisk	Pa-	<-Rate/s->		<------- Time (msec) ------->					Req.	<Percent>		SEEK	Recov	<-Throttle->	
Addr	Type	Label/ID	Links	ths	I/O	Avoid	Pend	Disc	Conn	Serv	Resp	CUWt	Qued	Busy	READ	Cyls	SSCH	Set/s	Dly/s
>> All DASD <<			....		.5	.4	.2	.1	3.4	3.7	3.7	.0	.0	0	17	1173	0	...	.0
F024	3390	VS2426	1	4	12.9	147.0	.2	.7	.4	1.3	1.3	.0	.0	2	91	193	0	...	...
0C20	CTCA		...	1	12.6	...	.3	.2	.6	1.1	1.1	.0	.0	1	..	...	0	...	...
F685	3390	VS2W01	290	4	11.8	.3	.2	.0	.3	.5	.5	.0	.0	1	84	89	0	...	...
F411	3390	VS2613	1	4	10.6	.5	.2	.3	.4	.9	.9	.0	.0	1	1	1303	0	...	...

1. Interesting columns: Avoid, Serv, Req. Qued.
2. Req. Qued > 0 is a good trigger for looking at PAV or HyperPAV.
3. For queuing on page or spool, use FCX109 DEVICE CPOWNED.

How z/VM Exploits PAV For Its Guests

- Guests do minidisk I/O
- Old: we serialized the resultant real I/Os
- New: we use the volume's PAV aliases to drive the real I/Os concurrently
- One wait queue for the volume
- Many device numbers for the volume



Note: z/VM 5.2 plus VM63855 (UM31771) (May 2006)

FICON Open Exchange Limit

- **Parallel and ESCON: 1 I/O at a time on a chpid**
  - Pending time >0 could mean chpid contention
  - Controller disconnect was a good thing, and so they did
- **FICON: 64 (was 32) I/Os at a time on a chpid**
  - Pending time >0 probably now means slow IR
  - Little motive for controller to disconnect anymore
    - Controller cache miss is still a good reason
- **Calculating “open exchange level” is not easy**
- **Very seldom is this an issue anyway**

Other

- **Number of spool files (architected):**
  - 9999 per user
  - 1.6 million spool files per system
    - 1024 files per warm start block * (180 * 9) warm start blocks
- **Number of logged-on virtual machines (approximate): about 100,000 (per designers)**

Metrics for Formal Spin Locks

FCX265 CPU 2094 SER 19B9E Interval 02:31:51 - 12:34:01 GDLVM7

		<---------- Spin Lock Activity ---------->						
		<----- Total ----->			<--- Exclusive --->		<----- Shared ---->		
Interval		Locks	Average	Pct	Locks	Average	Pct	Locks	Average	Pct
End Time	LockName	/sec	usec	Spin	/sec	usec	Spin	/sec	usec	Spin
>>Mean>>	SRMATDLK	1.9	.539	.000	1.9	.539	.000	.0	.000	.000
>>Mean>>	RSAAVCLK	.0	2.015	.000	.0	2.015	.000	.0	.000	.000
>>Mean>>	FSDVMLK	.0	24.97	.000	.0	24.97	.000	.0	.000	.000
>>Mean>>	SRMALOCK	.0	.000	.000	.0	.000	.000	.0	.000	.000
>>Mean>>	HCPTRQLK	4.1	.195	.000	4.1	.195	.000	.0	.000	.000
>>Mean>>	SRMSLOCK	34.0	1.096	.001	32.7	1.037	.001	1.3	.001	.000

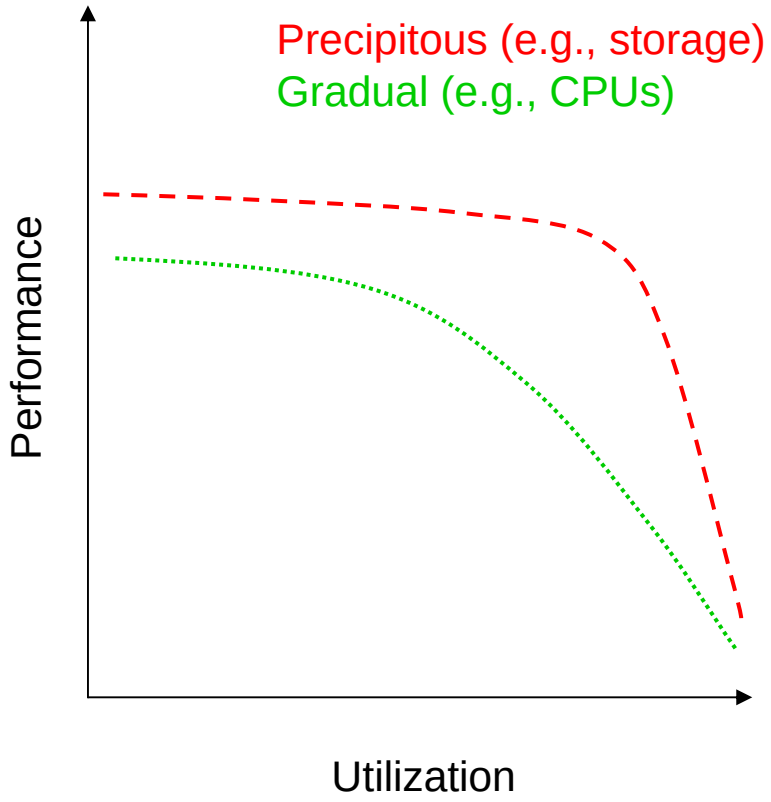
This is really for our use. Just look at T/V.

Latent Limits

- **Sometimes it's not an architected limit**
- **Sometimes it's just "your workload won't scale past here, because…"**
- **In our studies of z/VM 5.3, we found these kinds of latent limits:**
  - Searching for a below-2-GB frame in lists dominated by above-2-GB frames (storage balancing functions)
  - Contention for locks, usually the scheduler lock
- **These kinds of phenomena were the reasons we published the limits to be 256 GB and 32 engines**
  - We wanted to publish supported limits we felt would be safe in a very large variety of workloads and environments
  - Some of our measurement workloads scaled higher than this

Other Notes on z/VM Limits

- **Sheer hardware:**
  - z/VM 5.2: 24 engines, 128 GB real
  - z/VM 5.3: 32 engines, 256 GB real
  - System z: 65,000 I/O devices per partition
- **Mad-scientist stuff we've tried:**
  - 54 engines
  - 1 TB / 128 GB with 100 10 GB Linux Apaches
  - 440 GB / 20 GB with 8 1 TB thrasher guests
  - 256 GB of thrashers in 3 GB of central
- **Utilizations we routinely see in customer environments**
  - 85% to 95% CPU utilization without worry
  - Hundreds of thousands of XSTORE pages per second
  - Tens of thousands of DASD pages per second
- **Our limits tend to have two distinct shapes**
  - Slow rolloff: CPUs
  - Fast rolloff: storage



Keeping Tabs on Consumption Limits

- **Processor**
  - CPU utilization: FCX100 CPU, FCX126 LPAR, FCX144 PROCLOG, FCX114 USTAT
  - Good article on CPU at http://www.vm.ibm.com/perf/tips/lparinfo.html
- **Memory & Paging**
  - Page slots in use: FCX146 AUXLOG
  - Paging I/O: FCX109 DEVICE CPOWNED
  - V:R Memory ratio: FCX113 UPAGE
  - PTRM space consumed: FCX134 DSPACESH
  - Storage in use for segment tables: FCX113 UPAGE
  - Consumption of SXS space: FCX261 SXSAVAIL
  - MDC: FCX178 MDCSTOR, FCX134 DSPACESH
  - Consumption of real memory: FCX103 STORAGE, FCX254 AVAILLOG
  - Consumption of expanded storage: FCX103 STORAGE
  - Good article on paging at http://www.vm.ibm.com/perf/tips/prgpage.html
- **I/O**
  - Guest DASD I/O: FCX108 DEVICE
  - Concurrency on FICON chpids: FCX131 DEVCONF, FCX215 INTERIM FCHANNEL, FCX168 DEVLOG, FCX232 IOPROCLG

What Consumption Limits Will We Hit First?

- **Guest-storage-intensive workload:**
  - page slots on DASD… at 5-6 TB things start to get interesting… mitigate by paging to SCSI
  - paging I/O concurrency – only 255 at a time – mitigate by paging to SCSI
  - utilization on paging volumes and chpids -- watch for MLOAD elongation -- mitigate by spreading I/O
  - reorder processing – use more and smaller guests
  - mitigation by application tuning… perhaps smaller guests
  - segment table constraints: probably an issue at 50% (128 TB of logged-on guest real) … not anytime soon
- **Real-storage-intensive workload:**
  - Ability of the system to page will limit you: ensure adequate XSTORE and paging capacity
  - You can define > 256 GB of real storage, but we are aware that some workloads cannot scale that high
  - Mitigation by application tuning or by using CMM
- **CPU-intensive workload:**
  - FCX100 CPU, FCX126 LPAR, and FCX 114 USTAT will reveal CPU limitations
  - You can define > 32 engines, but we are aware that some workloads cannot scale that high
  - Mitigation by application tuning
- **I/O-intensive workload:**
  - Device queueing: consider whether PAV or HyperPAV might offer leverage
  - Chpid utilization: add more chpids per storage controller
- **Ultimately partitions can be split, but we would prefer you not have to do this (too complicated)**
- **Without trend data (repeated samples) for *your* workloads it is difficult to predict which of these limits *you* will hit first**

Summary

- **Knowing limits:**
  - Real resource consumption
  - Limits to managing the virtualization of real resources
- **Measuring limits:**
  - Knowing where to watch for these limits
  - Including these in capacity planning
- **Managing limits:**
  - Tuning and configuring
  - Planning for growth

IBM Systems & Technology Group

End of Presentation