IBM Systems & Technology Group

z/VM Performance Update, z/VM 6.3

Revision 2014-04-14.1, BKW

Brian K. Wade, Ph.D.
bkw@us.ibm.com

Trademarks

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, ESCO, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/30, 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 – Your z/VM Performance Team

- **Dean DiTommaso**
- **Bill Guzior**
- **Steve Jones**
- **Virg Meredith – happy 50th, Virg ☺**
- **Patty Rando**
- **Dave Spencer**
- **Susan Timashenka – department manager**
- **Xenia Tkatschow**
- **Brian Wade**

Agenda

- **z/VM 6.3 thoughts**
  - A little about Large Memory
  - A little about HiperDispatch
  - Various other line items or small changes
  - Monitor record changes
  - Performance-related service
  - z/VM Performance Toolkit
- **Other thoughts**
  - Reminder about CPU MF
  - Continued evolution of System z CPC Performance workloads

z/VM 6.3 Highlights: A Performance View

- **Regression performance**
- **Large Memory considerations**
- **HiperDispatch considerations**
- **Large Dump considerations**
- **Some smaller changes**
- **Monitor record changes**
- **Performance-related service**
- **z/VM Performance Toolkit changes**

Regression Performance

- **Ran our usual library of workloads**
  - CMS interactive, various Apache configurations
- **Results are within usual 5% regression criteria**
- **Some workloads will see improvements:**
  - Constrained by reorder or demand scan
  - A few heavy guests along with small VCPU:LCPU ratio

Thoughts on Large Memory

Large Memory: Highlights

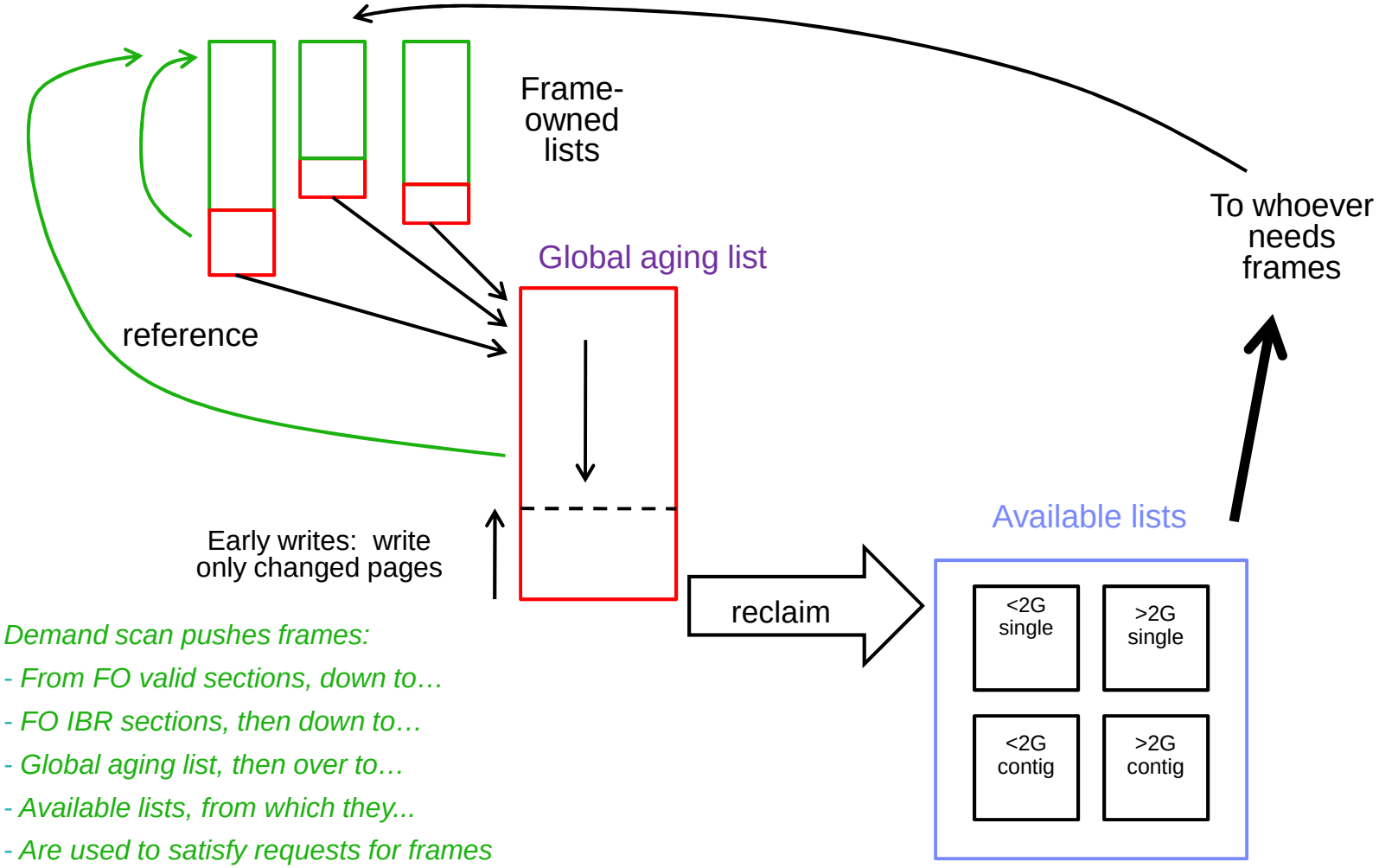
- **Exploit a 1 TB central memory**
  - Exploit larger real
  - Allow larger total virtual
  - XSTORE no longer required; best practice is not to use it
- **Remove algorithms and techniques known not to scale**
- **Improve SET RESERVED in a couple key ways**
  - Make it work
  - Extend it to NSS and DCSS
- **Overhaul CP Monitor records appropriately**
- **New or changed z/VM Performance Toolkit screens**

- **New planning heuristics: See Large Memory Deep Dive or z/VM Planning and Admin for guidance**

Large Memory, Scaling Problems Removed

- **Reorder is gone**
- **Demand scan no longer searches frame lists**
- **No more use of RRBE instruction (long CPI)**
- **DASD channel program now does scatter-scatter I/O**
- **Unchanged pages usually not rewritten**

New Approach: The Big State Diagram



Demand scan pushes frames:

- From FO valid sections, down to…

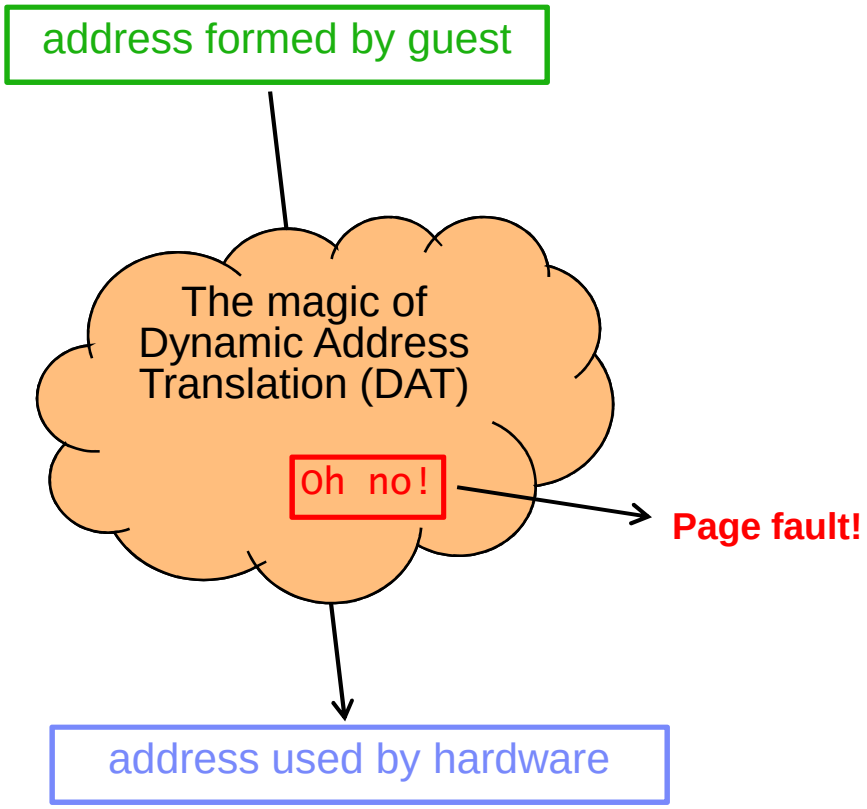
- FO IBR sections, then down to…

- Global aging list, then over to…

- Available lists, from which they...

- Are used to satisfy requests for frames

New Approach: Trial Invalidation



- **Page table entry (PTE) contains an "invalid" bit**
- **What if we:**
  - Keep the PTE intact but set the "invalid" bit
  - Leave the frame contents intact
  - Wait for the guest to touch the page
- **A touch will cause a page fault, but…**
- **On a fault, there is nothing really to do except:**
  - Clear the "invalid" bit
  - Move the frame to the front of the frame list to show that it was recently referenced
- **We call this trial invalidation.**

New Approach: Visit Everybody

- **Old way:**
  - Was focused on the dispatch list
  - Did not remember “where it left off”
- **New way:**
  - Focused on the list of logged-on users
  - Remembers “where it left off”
- **Objective: try to visit more equitably**

Workload: "Sweet Spot"

Our synthetic workload called *Sweet Spot* imitates behaviors we have seen in customer-supplied MONWRITE data.

Run ID	STWG522W	STXG3258	Delta	Pct
CP Level (p)	6.2.0	6.3.0		
SYSGEN Storage (P)	262144	393216	131072	50.0
Total CP Xstor (p)	131072	0	-131072	-100.0
Number of CPUs (p)	4	4	0	0.0
Aggregate ETR	0.0746	0.0968	0.0222	29.8
User Group 1 ETR	0.0065	0.0128	0.0063	96.9
User Group 2 ETR	0.0138	0.0236	0.0098	71.0
User Group 3 ETR	0.0268	0.0264	-0.0004	-1.5
User Group 4 ETR	0.0275	0.0341	0.0066	24.0
ITR (h)	77.77	105.60	27.83	35.8
System Util/Proc (p)	31.4	4.7	-26.7	-85.0
T/V Ratio (p)	1.51	1.08	-0.43	-28.5

By getting rid of both reorders and spin lock contention, we achieved huge drops in %CPU and T/V.

Workload: Apache Paging

Our Linux-based workload called *Apache Paging* is built to page heavily to DASD almost no matter how much central or XSTORE we give it.

Run ID	A38W952A	A38XM820
CP Level (p)	6.2.0	6.3.0
SYSGEN Storage (P)	262144	393216
Total CP Xstor (p)	131072	0
Number of CPUs	8	8
Tx/sec (c)	1.000	1.024
ITR (h)	1.000	1.017
XSTORE paging /sec	82489	0
DASD paging /sec	33574	31376

This is an example of a workload where the limit comes from something large memory will not fix.

Planning for Large Memory

- **Change XSTORE to central**
- **Plan enough DASD space (see Planning and Admin)**
- **Plan robust DASD configuration**
- **Check or add SET RESERVED settings**
- **Plan enough dump space**
  - http://www.vm.ibm.com/service/zvmpladm.pdf

Large Memory CP Monitor Changes

Domain	Record	Name	Type	Title	Fields, N / D / C
D0	R3	MRSYTRSG	sample	Real Storage Data (Global)	D C
D0	R4	MRSYTRSP	sample	Real Storage Data (Per Processor)	D
D0	R6	MRSYTASG	sample	Auxiliary Storage (Global)	N C
D0	R7	MRSYTSHS	sample	Shared Storage Data	D
D0	R23	MRSYTLCK	sample	Formal Spin Lock Data	N C
D1	R7	MRMTRMEM	config	Memory Configuration Data	N
D1	R15	MRMTRUSR	config	Logged on User	C
D2	R4	MRSCLADL	event	Add User to Dispatch List	D C
D2	R5	MRSCLDDL	event	Drop User from Dispatch List	D C
D2	R6	MRSCLAEL	event	Add User to Eligible List	C
D2	R8	MRSCLSTP	event	System Timer Pop	D
D3	R1	MRSTORSG	sample	Real Storage Management (Global)	N D C
D3	R2	MRSTORSP	sample	Real Storage Activity (Per Processor)	D
D3	R3	MRSTOSHR	sample	Shared Storage Management	N C
D3	R14	MRSTOASI	sample	Address Space Information Record	N C
D3	R15	MRSTOSHL	event	NSS/DCSS/SSP Loaded into Storage	N
D3	R16	MRSTOSHD	event	NSS/DCSS/SSP Removed From Storage	N C
D4	R2	MRUSELOF	event	User Logoff Data	N D C
D4	R3	MRUSEACT	sample	User Activity Data	N D C
D4	R9	MRUSEATE	event	User Activity Data at Transaction End	D C

z/VM Performance Toolkit: Highlights

- **Changed screens:**
  - FCX102 SYSTEM, Some Internal System Counters
  - FCX103 STORAGE, General Storage Utilization
  - FCX133 NSS, NSS and DCSS Utilization and Paging Activity
  - FCX146 AUXLOG, Auxiliary Storage Utilization, by Time
  - FCX147 VDISKS, Virtual Disks in Storage
  - FCX265 LOCKLOG, Spin Lock Log, by Time
- **Deleted screens:**
  - FCX254 AVAILLOG, Available List Management, by Time
  - FCX259 DEMNDLOG, Demand Scan Details, by Time
- **New screens:**
  - FCX290 UPGACT, User Page Activity — *page state transition rates*
  - FCX291 UPGACTLG, User Page Activity (benchmarks a user)
  - FCX292 UPGUTL, User Page Utilization Data — *page residency counts*
  - FCX293 UPGUTLLG, User Page Utilization Data (benchmarks a user)
  - FCX294 AVLB2GLG, Available List Data Below 2G, by Time — *available list counts*
  - FCX295 AVLA2GLG, Available List Data Above 2G, by Time
  - FCX296 STEALLOG, Steal Statistics, by Time — *steal algorithm activity*
  - FCX297 AGELLOG, Age List Log, by Time — *global aging list activity*

z/VM Performance Toolkit: New Columns and Concepts

Caption or Column Heading	What this means
Inst	*Instantiations:* the rate at which valid memory is being created *Instantiated:* the amount of valid memory
Relse	*Releases:* the rate at which memory is being released
Inval	*Invalidations:* the rate at which demand scan is marking memory invalid as a way to determine whether it is being touched
Reval	*Revalidations*: the rate at which invalid pages are being made valid because somebody touched them
Ready	*Ready reclaims* or *ready steals:* the frame was found and selected for reclaim and had already been prewritten to auxiliary storage
Not Ready	*Notready reclaims* or *notready steals:* the frame was selected for reclaim but we had to wait for the aux write to finish before we could take it
PNR	*Private, not referenced*: the page was read from aux as part of a block read, but it is still marked invalid because nobody has touched it yet
x<2G or *x*>2G	*Below 2 GB* or *Above 2 GB*: tells where the real backing frames are in real central
Sing	*Singles:* free frames surrounded by in-use frames (cannot coalesce)
Cont	*Contigs*: free frames in strings of two or more
Prot	*Protect threshold:* number of frames a singles-obtain must leave on a contigs-list

z/VM Performance Toolkit: New Report FCX292 UPGUTL

```
FCX292  Run 2013/04/10 07:38:36         UPGUTL                                                          Page  103
                                        User Page Utilization Data
From 2013/04/09 16:02:10                                                                       SYSTEMID
To   2013/04/09 16:13:10                                                                       CPU 2817-744    SN A6D85
For    660 Secs 00:11:00                "This is a performance report for SYSTEM XYZ"          z/VM    V.6.3.0 SLU 0000
_____________________________________________________________________________________________________________________

 ______        .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .      .
                <---------------------------------------------- Storage --------------------------------------------->
                                  <------------------------- Resident --------------------------->
            Data                                                <----- Invalid But Resident ------>           Base
          Spaces                  <---- Total ----> <-Locked--> <-- UFO --> <-- PNR --> <-AgeList->           Space Nr of
 Userid    Owned   WSS  Inst Resvd T_All  T<2G  T>2G  L<2G  L>2G  U<2G  U>2G  P<2G  P>2G  A<2G  A>2G XSTOR   AUX  Size Users
 >>Mean>>     .0 5284M 6765M  5611 5286M   27M 5259M  1010  232K  6565 2238K 59588   26M 53080  107M    .0 1815M 7108M    73
 User Class Data:
 CMS1_USE     .0 3320K   19M    .0  484K    .0  484K    .0  4096    .0 69632    .0  244K    .0  344K    .0   19M 2047M     1
 LCC_CLIE     .0  364M  485M    .0  365M 11264  365M    .0  208K    .0  325K    .0 2686K    .0 8177K    .0  164M 1024M     8
 LXA_SERV     .0 7974M   10G    .0 7978M   41M 7937M    .0  206K  9984 3327K 90624   39M 80725  161M    .0 2719M 10240M   48

 User Data:
 DISKACNT     .0 4976K 5156K     0    4K     0    4K     0     0     0    4K     0     0     0     0     0 5152K   32M
 DTCVSW1      .0  184K   11M     0  196K    8K  188K    8K    4K     0    4K     0     0     0  168K     0   11M   32M
 DTCVSW2      .0  180K   11M     0  184K     0  184K     0    4K     0    4K     0     0     0  164K     0   10M   32M
 EREP         .0 4912K 4944K     0    4K     0    4K     0     0     0    4K     0     0     0     0     0 4940K   32M
 FTPSERVE     .0   84K 5764K     0   88K     0   88K     0    4K     0    4K     0     0     0   76K     0 5760K   32M
 GCSXA        .0  204K  208K     0    8K     0    8K     0    4K     0    4K     0     0     0     0     0  200K   16M
 LCC00001     .0  364M  488M     0  365M     0  365M     0  204K     0  228K     0 2884K     0 8660K     0  192M 1024M
 LCC00002     .0  369M  492M     0  371M   20K  371M     0  204K     0  224K     0 2312K     0 7736K     0  159M 1024M
 LCC00003     .0  363M  484M     0  364M     0  364M     0  204K     0  252K     0 2852K     0 8372K     0  215M 1024M
 LCC00004     .0  363M  483M     0  363M   16K  363M     0  204K     0  228K     0 2724K     0 8512K     0  185M 1024M
```

Look for the new concepts: Inst   IBR   UFO   PNR   AgeList

Amounts are in bytes, suffixed. Not page counts!

FCX113 UPAGE is still produced.

z/VM Performance Toolkit: New Report FCX290 UPGACT

```
FCX290  Run 2013/04/10 07:38:36          UPGACT                                                    Page  102
                                         User Page Activity
From 2013/04/09 16:02:10                                                                  SYSTEMID
To   2013/04/09 16:13:10                                                                  CPU 2817-744   SN A6D85
For    660 Secs 00:11:00                 "This is a performance report for SYSTEM XYZ"    z/VM   V.6.3.0 SLU 0000
```

		<---------- Storage ---------->						<------- Movement/s -------->				
	Stl	<--- Transition/s ---->			<-Steal/s->						<Migrate/s>		Nr of
Userid	Wt	Inst	Relse	Inval	Reval	Ready	NoRdy	PGIN	PGOUT	Reads	Write	MWrit	Xrel	Users
>>Mean>>	1.0	143K	5142	849K	718K	999K	.0	.0	.0	958K	761K	.0	.0	73
User Class Data:														
CMS1_USE	1.0	15515	15801	2377	1632	5145	.0	.0	.0	.0	1980	.0	.0	1
LCC_CLIE	1.0	658K	20875	488K	486K	60875	.0	.0	.0	54212	22869	.0	.0	8
LXA_SERV	1.0	108K	1095	1191K	994K	1506K	.0	.0	.0	1447K	1153K	.0	.0	48
User Data:														
DISKACNT	1.0	0	0	0	0	0	0	0	0	0	0	0	0	
DTCVSW1	1.0	0	0	3072	2855	0	0	0	0	0	0	0	0	
DTCVSW2	1.0	0	0	3004	2780	0	0	0	0	0	0	0	0	
EREP	1.0	0	0	0	0	0	0	0	0	0	0	0	0	
FTPSERVE	1.0	0	0	1434	1434	0	0	0	0	0	0	0	0	
GCSXA	1.0	0	0	0	0	0	0	0	0	0	0	0	0	
LCC00001	1.0	601K	18686	501K	498K	65139	0	0	0	49866	23670	0	0	
LCC00002	1.0	657K	24955	487K	486K	54725	0	0	0	44522	18991	0	0	
LCC00003	1.0	565K	23012	485K	481K	64065	0	0	0	44783	19859	0	0	
LCC00004	1.0	602K	24104	499K	495K	63178	0	0	0	48811	24588	0	0	
LCC00005	1.0	717K	25675	500K	499K	65865	0	0	0	66002	28753	0	0	

Look for the new concepts: Inst Relse Inval Reval Ready NoRdy

z/VM Performance Toolkit: New Report FCX295 AVLA2GLG

```
FCX295  Run 2013/04/10 07:38:36          AVLA2GLG                                                    Page    25
                                         Available List Data Above 2G, by Time
From 2013/04/09 16:02:10                                                                  SYSTEMID
To   2013/04/09 16:13:10                                                                  CPU 2817-744    SN A6D85
For    660 Secs 00:11:00                 "This is a performance report for SYSTEM XYZ"    z/VM    V.6.3.0 SLU 0000
```

	<---------- Storage ---------->						<--Times-->		<-Frame Thresh-->		
Interval	<Available>		<Requests/s>		<Returns/s>		<-Empty/s->		Sing	<-Contigs->	
End Time	Sing	Cont	Sing	Cont	Sing	Cont	Sing	Cont	Low	Low	Prot
>>Mean>>	23M	267M	47M	59M	47M	51M	.0	.0	1310	15	15
16:02:40	0	938M	32M	126M	502K	30310	.0	.0	1332	15	15
16:03:10	152K	4556K	50M	89M	49M	59M	.0	.0	1168	15	15
16:03:40	400K	4824K	68M	82M	71M	79M	.0	.0	1321	15	15
16:04:10	0	5896K	49M	72M	52M	70M	.0	.0	2409	15	15
16:04:40	0	2124K	40M	60M	41M	59M	.0	.0	1308	15	15
16:05:10	876K	3488K	54M	52M	55M	51M	.0	.0	1118	15	15
16:05:40	0	3624K	53M	58M	54M	57M	.0	.0	1409	15	15
16:06:10	2016K	4464K	49M	57M	51M	56M	.0	.0	1273	15	15

Look for the new concepts: Singles Contigs Prot

Amounts are in bytes, suffixed. Not page counts!

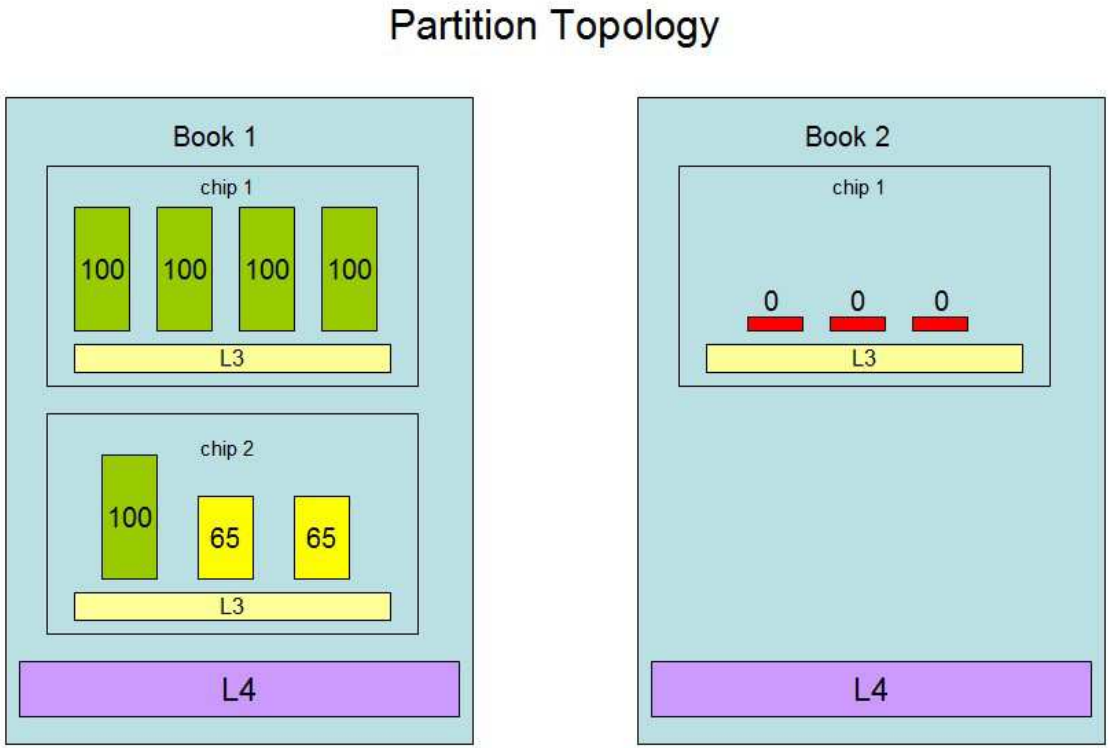
FCX254 AVAILLOG is no longer produced.

Thoughts on HiperDispatch

HiperDispatch: Highlights

- **Use of vertical mode partitions**
- **Running widely if power is there**
- **Automatic reduction of MP level**
- **Topology-aware dispatching**
- **New or changed CP Monitor records**
- **New or changed z/VM Performance Toolkit screens**
- **How to plan for HiperDispatch**

Vertical-Mode Partitions



Features:

Concentrated entitlement

Durable placement

Advantages:

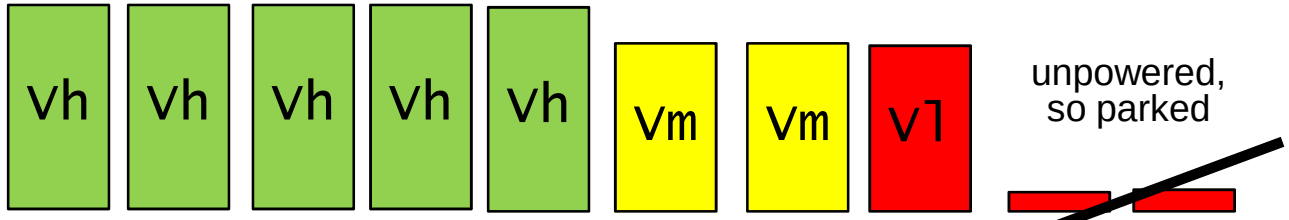
Quiet place to run

Opportunity to reduce MP level

Running Widely When The Power Is There

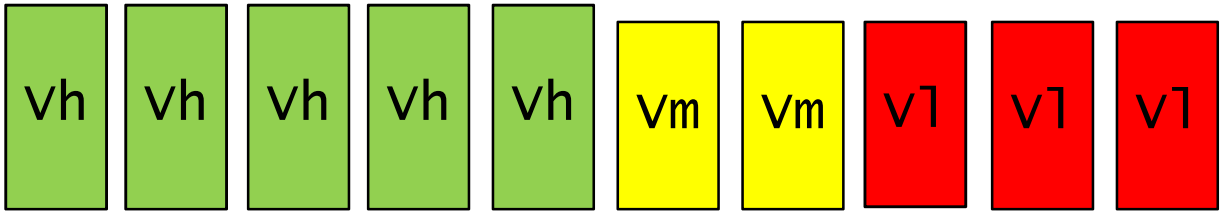
Entitlement: 630% via 5 Vh @ 100, 2 Vm @ 65, 3 Vl @ 0

Prediction:
- Power avbl 750%:



750 = 100 100 100 100 100 83.3 83.3 83.3 0 0

Prediction:
- Power avbl 950%:



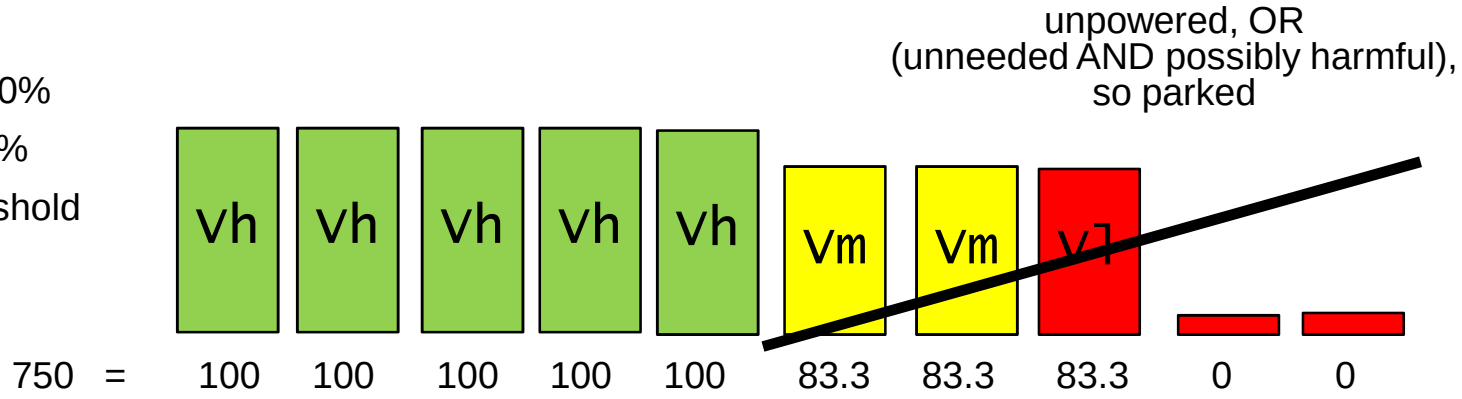
950 = 100 100 100 100 100 90 90 90 90 90

Parking to Try To Reduce z/VM Overhead

Entitlement: 630% via 5 Vh @ 100, 2 Vm @ 65, 3 Vl @ 0

Prediction:
- Power avbl 750%
- Utilization 350%
- T/V > low threshold

CPUPAD 100%



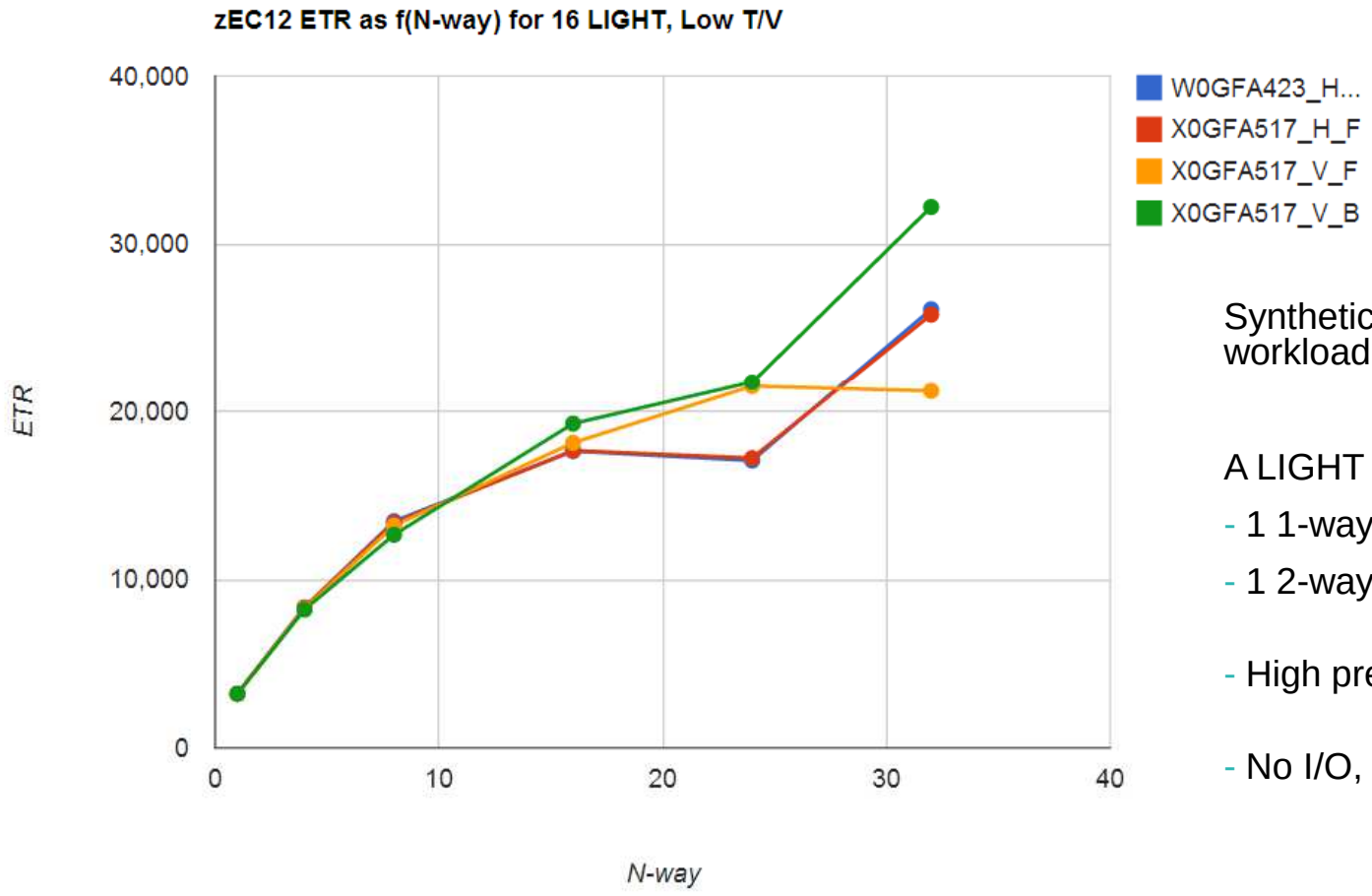
z/VM parks apparently unneeded processors, but only if T/V is projected high and load is projected below capacity. Safety margin is controlled via CP SET SRM CPUPAD.

Topology-Aware Dispatching

- **If can't send home, send close to home**
  - Same chip? Same book?
- **Try to keep a virtual MP's VCPUs together**
- **Try not to do long-drag steals**
  - Cross-book, cross-chip
- **Be smart about which real CPU we wake up**
  - Same chip as stacked work? Same book?

- **Rebalance: only certain workloads are suitable**
  - A few heavy users, low VCPU:LCPU ratio, clearly distinguishable %CPU

Memory-Touching Workload, Light Edition



Synthetic, memory-touching workload

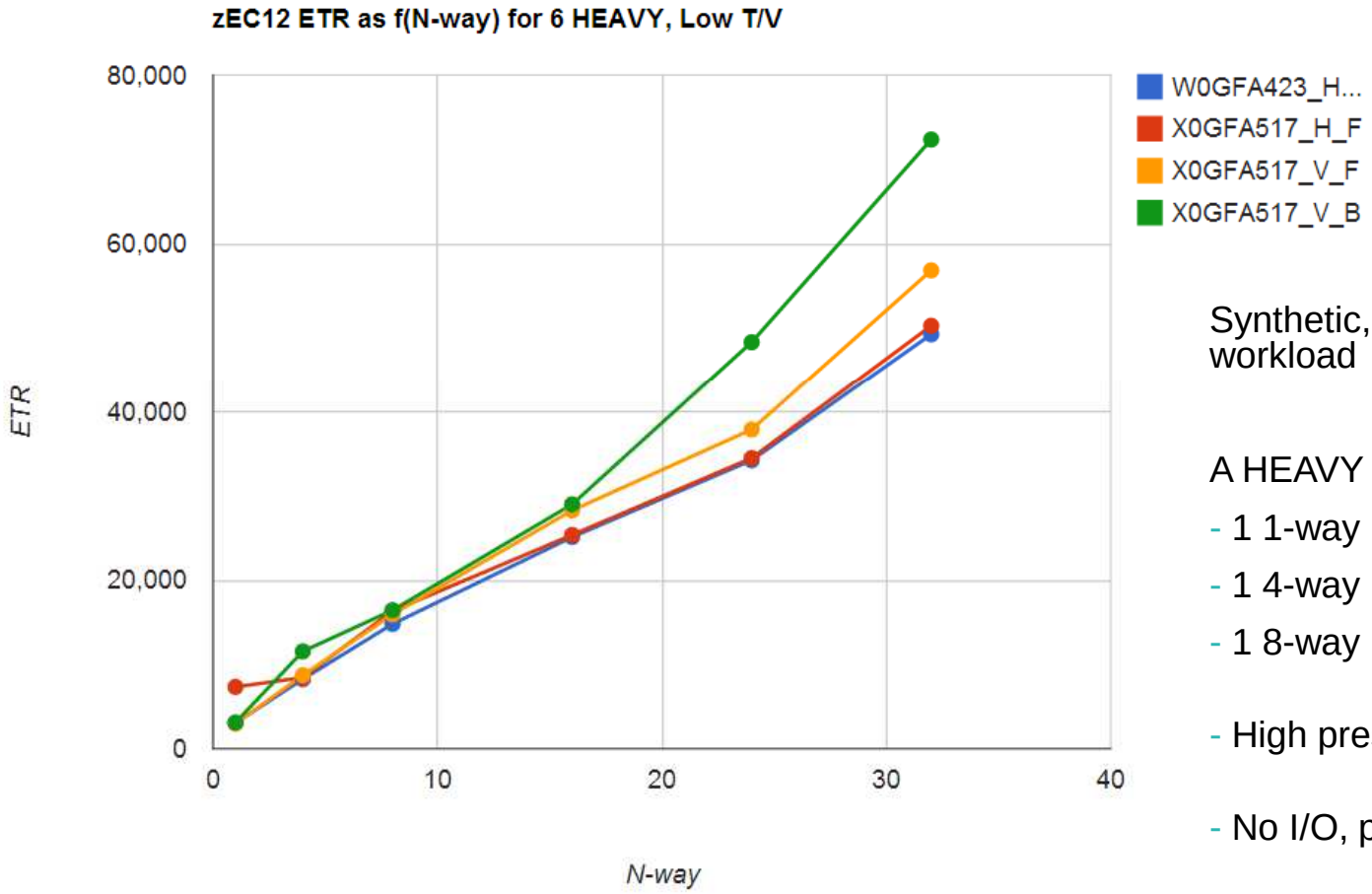
A LIGHT tile is 81% busy:

- 1 1-way @ 15% each
- 1 2-way @ 33% each

- High pressure on nest

- No I/O, paging, etc.

Memory-Touching Workload, Heavy Edition



Synthetic, memory-touching workload

A HEAVY tile is 540% busy:

- 1 1-way @ 15% each
- 1 4-way @ 31% each
- 1 8-way @ 50% each

- High pressure on nest

- No I/O, paging, etc.

Planning for HiperDispatch

- **Before:**
  - Decide what “success” looks like: metrics and values
  - Measure: transaction rates, MONWRITE data
- **Turn on Global Performance Data Control (activation profile)**
- **After:**
  - We think vertical and reshuffle are probably the right choice for you
  - Do same measurements
  - Compare
- **The way out:**
  - CP SET SRM POLARIZATION HORIZONTAL
  - SRM statement in system configuration file

Comments on Workloads

- **Amenable workloads for z/VM HiperDispatch:**
  - High-CPU, CPU-constrained workloads
    - Improving cache behavior stands to improve performance
  - Active VCPU:LCPU ratio isn't too large
    - High ratio has too much context switching to feel much effect
  - Runs in a partition having multiple topology containers
    - Gives z/VM an opportunity to separate guests from one another
- **Compare those statements to IBM's statements about PR/SM and partitions**
- **Indifferent workloads for z/VM HiperDispatch**
  - Constrained by something else, such as I/O
  - Memory-overcommitted
  - High VCPU:LCPU ratio with every virtual CPU active just a little bit
  - Workloads with bad memory access habits
- **Remember that vertical mode also keeps your partition away from the other partitions**

HiperDispatch CP Monitor Changes

Domain	Record	Name	Type	Description of Change
D0	R2	MRSYTPRP	sample	Added polarity, entitlement, and park-time accumulator
D0	R16	MRSYTCUP	sample	Added partition current weight
D0	R23	MRSYTLCK	sample	Added the HCPDSVTL topology lock
D1	R4	MRMTRSYS	config	Added bit indicating whether system is horizontal or vertical
D1	R5	MRMTRPRP	config	Added park state, polarization, entitlement, and topological location
D1	R16	MRMTRSCH	config	Added h/v bit, CPUPAD settings, and EXCESSUSE settings
D2	R7	MRSCLSRM	event	Added h/v bit, CPUPAD settings, and EXCESSUSE settings
D4	R2	MRUSELOF	event	Added rebalance results and steal results
D4	R3	MRUSEACT	sample	Added rebalance results and steal results
D5	R2	MRPRCVOF	event	Added park/unpark failure as reason varied off
D5	R3	MRPRCVON	event	Added parked as a state; use iff neither D5 R17 nor D5 R18 are seen
D5	R15 (new)	MRPRCDSV	event	Records assignment of processors to dispatch vectors
D5	R16 (new)	MRPRCPUP	event	Records park/unpark decision
D5	R17 (new)	MRPRCRCD	sample	Records processor's VMDBK steal behavior
D5	R18 (new)	MRPRCDHF	sample	Records PLDV population trends

z/VM Performance Toolkit

- **Themes in the changes in existing Perfkit screens**
  - CPU entitlement appears in sensible places, e.g. FCX100 CPU
  - Percent-parked appears in sensible places, e.g. FCX100 CPU
  - Parked time is correctly accounted for, e.g. FCX126 LPAR %Susp
  - SRM settings are reported where they ought to be, e.g. FCX154 SYSSET
  - Interesting events are reported in FCX180 SYSCONF as they should
  - Number of unparked CPUs appears in sensible places, e.g. FCX225 SYSSUMLG
  - Counts of new monitor records appear in FCX155 MONDATA as they should
  - Obsolete data is compatibly deleted in certain places, e.g. FCX144 PROCLOG
- **New reports sure to attract interest:**
  - FCX287 TOPOLOG shows a log of partition topology, container-major
  - FCX298 PUORGLOG shows a log of partition topology, CPU-major
  - FCX299 PUCFGLOG shows a log of the park/unpark state
  - FCX301 DSVBKACT replaces the PLDV emptiness columns on FCX144 PROCLOG
  - FCX302 PHYSLOG shows a physical CPU utilization log of the CEC by type pool
  - FCX303 DSVSLOG replaces the PLDV steal columns on FCX144 PROCLOG
  - FCX304 PRCLOG is where you should now look instead of FCX144 PROCLOG
  - FCX306 LSHARACT reports the partitions' entitlements vs. logical CPU counts
- **Obsolete reports**
  - FCX144 PROCLOG is still there for now, but start using FCX304 PRCLOG instead

New Report PUORGLOG

```
1FCX298  Run 2013/05/20 10:39:48          PUORGLOG
                                          Processor Unit organization log
 From 2013/05/19 03:39:31
 To   2013/05/19 03:41:31
 For    120 Secs 00:02:00                  Result of GF003855 Run
 ____________________________________________________________________________

  Logical PU organization for Partition PPRF1    (GDLBOFVM)

  Date  Time     CPU Type PPD Ent. Location
  05/19 03:39:31   0 CP   VhD  100 1:6
  05/19 03:39:31   1 CP   VhD  100 1:6
  05/19 03:39:31   2 CP   VhD  100 1:5
  05/19 03:39:31   3 CP   VhD  100 1:5
  05/19 03:39:31   4 CP   VhD  100 1:5
  05/19 03:39:31   5 CP   VhD  100 1:5
  05/19 03:39:31   6 CP   VhD  100 1:5
  05/19 03:39:31   7 CP   VhD  100 1:4
  05/19 03:39:31   8 CP   VhD  100 1:4
  05/19 03:39:31   9 CP   VhD  100 1:4
  05/19 03:39:31  10 CP   VhD  100 1:4
  05/19 03:39:31  11 CP   VhD  100 1:2
  05/19 03:39:31  12 CP   VhD  100 1:2
  05/19 03:39:31  13 CP   VhD  100 1:2
  05/19 03:39:31  14 CP   VhD  100 1:2
                                … truncated …
```

Notes:

Vh	vertical high
Vm	vertical medium
Vl	vertical low
VhD	vertical high, dedicated partition
Ent	entitlement (100% = 1 CPU's worth)
Location	book:chip (z10: book)

New Report LSHARACT

```
1FCX306  Run 2013/02/19 12:10:57          LSHARACT
                                          Logical Partition Share

 From 2013/02/19 11:49:58
 To   2013/02/19 11:56:10
 For    372 Secs 00:06:12                  Result of GFCM0107 Run
```

LPAR Data, Collected in Partition RPRF2

Physical PUs, Shared: CP- 40 ZAAP- 2 IFL- 16 ICF- 1 ZIIP- 3
Dedicated: CP- 4 ZAAP- 0 IFL- 0 ICF- 0 ZIIP- 0

Proc Type	Partition Name	LPU Num	LPAR Weight	Entlment	<LPU Total,%> Busy	Excess	LPU Conf
CP	RCPX4	10	10	59.3	3.0	.0	o
CP	RCTS1	5	10	59.3	311.9	252.6	o
CP	RCTS2	5	30	177.8	1.0	.0	o
CP	RCT1	20	30	177.8	111.3	.0	o
CP	RCT2	10	10	59.3	11.2	.0	o
CP	REXT1	5	10	59.3	.0	.0	o
CP	REXT2	4	10	59.3	.0	.0	o
CP	RINS	10	10	59.3	.0	.0	o
CP	RPRF1	4	DED	...	...	...	.
CP	RPRF2	24	335	1985.2	1548.4	.0	o
CP	RSPX1	6	40	237.0	481.3	244.3	o
CP	RSPX2	6	40	237.0	499.7	262.7	o
CP	RSPX5	6	40	237.0	126.5	.0	o
CP	RST1	10	10	59.3	16.2	.0	o
CP	RST1X	6	10	59.3	102.5	43.2	o
CP	RST2	6	50	296.3	.9	.0	o
CP	RST3	3	30	177.8	1.2	.0	o
ICF	RCTS2	1	10	25.0	.0	.0	-
ICF	RCT1	1	30	75.0	.0	.0	-
IFL	RCTS2	2	10	188.2	.0	.0	-
IFL	RCT1	2	30	564.7	.0	.0	u
IFL	RSTL1	16	45	847.1	449.2	.0	o
ZAAP	RCPX4	1	10	40.0	.1	.0	-
ZAAP	RCTS2	1	10	40.0	.0	.0	-
ZAAP	RCT1	1	30	120.0	.0	.0	u
ZIIP	RCPX4	1	10	60.0	.3	.0	-
ZIIP	RCTS2	1	10	60.0	.0	.0	-
ZIIP	RCT1	1	30	180.0	.0	.0	u

You now have an easy way to see the entitlements of your partitions.

Features:

- Reports by partition and CPU type
- Reports entitlement in percent
- Reports percent-busy of the partition's CPUs of that type
- Reports whether the partition is consuming beyond its entitlement ("Excess")
- Reports LPU configuration wrt entitlement:
  - "o" – overconfigured
  - "u" – underconfigured
  - "-" – apparently just right

New Report PUCFGLOG

```
1FCX299  Run 2013/06/24 09:36:54          PUCFGLOG                                                    Page    6
                                          Processor Unit Configuration log
 From 2013/02/19 11:49:52                                                                  GFCM0107
 To   2013/02/19 11:56:10                                                                  CPU 2817-744   SN B6D85
 For    378 Secs 00:06:18                  Result of GFCM0107 Run                         z/VM   V.6.3.0 SLU 0000
```

					Type		<--- Last --->				<--- Next --->						
Date	Time	Type	OnL	Entitl	Cap	CPUPAD	EX	Load	XP	XPF	T/V	LCei	XPF	T/V	N	NotVh	UpCap	LPU Unparked mask
02/19	11:49:54	CP	24	1985.2	...	100.0	70	2.2	1159.4	892.8	3.519	3.9	885.9	200.5	2	.0	200.0	00300000_00000000
02/19	11:49:56	CP	24	1985.2	...	100.0	70	.5	1153.3	888.1	256.0	1.7	883.4	201.3	2	.0	200.0	00300000_00000000
02/19	11:49:58	CP	24	1985.2	...	100.0	70	.5	1159.7	893.1	122.3	1.7	885.2	204.2	2	.0	200.0	00300000_00000000
02/19	11:50:00	CP	24	1985.2	...	100.0	70	.7	1136.7	875.4	53.45	1.7	857.7	172.5	2	.0	200.0	00300000_00000000
02/19	11:50:02	CP	24	1985.2	...	100.0	70	.9	1128.6	869.2	4.531	1.7	863.0	172.5	2	.0	200.0	00300000_00000000
02/19	11:50:04	CP	24	1985.2	...	100.0	70	1.3	1034.5	778.8	1.822	1.8	688.3	172.4	2	.0	200.0	00300000_00000000
02/19	11:50:06	CP	24	1985.2	...	100.0	70	.6	1157.1	891.1	38.57	1.8	856.4	168.5	2	.0	200.0	00300000_00000000
02/19	11:50:08	CP	24	1985.2	...	100.0	70	.5	1162.9	895.5	250.8	1.7	856.9	211.1	2	.0	200.0	00300000_00000000
02/19	11:50:10	CP	24	1985.2	...	100.0	70	44.8	1161.8	894.7	2.214	89.1	858.9	211.1	2	.0	200.0	00300000_00000000
02/19	11:50:12	* CPU Park/Unpark State changed																
02/19	11:50:12	CP	24	1985.2	...	100.0	70	199.7	1145.1	881.9	1.517	354.6	858.5	197.6	5	.0	500.0	00300000_00000000
02/19	11:50:14	* CPU Park/Unpark State changed																
02/19	11:50:14	CP	24	1985.2	...	100.0	70	501.6	1155.6	890.0	1.009	803.5	858.3	197.5	10	.0	1000.0	013C0000_00000000
02/19	11:50:16	* CPU Park/Unpark State changed																
02/19	11:50:16	CP	24	1985.2	...	100.0	70	999.6	1147.4	883.6	1.001	1497.6	857.9	146.5	16	.0	1600.0	0FFC0000_00000000
02/19	11:50:18	* CPU Park/Unpark State changed																
02/19	11:50:18	CP	24	1985.2	...	100.0	70	1599.3	1155.1	889.6	1.001	2199.1	857.7	130.3	23	100.0	2300.0	FFFF0000_00000000
02/19	11:50:20	* CPU Park/Unpark State changed																
02/19	11:50:20	CP	24	1985.2	...	100.0	70	2297.6	1179.7	908.5	1.001	2995.8	860.2	125.6	24	100.0	2400.0	FFFFFE00_00000000
02/19	11:50:22	* CPU Park/Unpark State changed																
02/19	11:50:22	CP	24	1985.2	...	100.0	70	2397.1	1144.5	881.4	1.005	2496.6	854.3	125.4	24	100.0	2400.0	FFFFFF00_00000000
02/19	11:50:24	CP	24	1985.2	...	100.0	70	2080.5	1181.8	910.1	1.002	2569.2	887.6	125.3	24	100.0	2400.0	FFFFFF00_00000000
02/19	11:50:26	CP	24	1985.2	...	100.0	70	1681.3	1140.0	878.0	1.002	2660.9	845.8	122.1	24	100.0	2400.0	FFFFFF00_00000000
02/19	11:50:28	CP	24	1985.2	...	100.0	70	1632.4	1169.6	900.7	1.002	2684.7	886.2	1.660	24	100.0	2400.0	FFFFFF00_00000000
02/19	11:50:30	CP	24	1985.2	...	100.0	70	1587.7	1149.4	885.2	1.002	2635.4	869.6	1.252	24	100.0	2400.0	FFFFFF00_00000000
02/19	11:50:32	CP	24	1985.2	...	100.0	70	1878.3	1129.6	869.9	1.011	2560.8	854.7	1.008	24	100.0	2400.0	FFFFFF00_00000000
02/19	11:50:34	CP	24	1985.2	...	100.0	70	1824.3	1176.2	905.8	1.002	2425.8	884.3	1.007	24	100.0	2400.0	FFFFFF00_00000000

Look for: effect of high T/V, workload ramp-up, U' and XPF' values; power of a non-Vh

New Report DSVSLOG

```
1FCX303  Run 2013/05/20 10:32:38          DSVSLOG
                                          DSVBK Steals per logical CPU Log, by Time
 From 2013/05/19 02:03:25
 To   2013/05/19 02:05:19
 For    114 Secs 00:01:54                  Result of GF003820 Run
```

Interval End Time	C P U	Type	PPD	Ent.	DVID	Pct Park Time	DSVBK Steal /s Lvl-00	Lvl-01	Lvl-02	Lvl-03	Lvl-04	Lvl-05
>>Mean>>	0	CP	Vh	100	0000	0	4.404	4.088	.000	....	....	....
>>Mean>>	1	CP	Vh	100	0001	0	2.456	2.561	.000	....	....	....
>>Mean>>	2	CP	Vh	100	0002	0	6.877	.921	.000	....	....	....
>>Mean>>	3	CP	Vh	100	0003	0	7.596	.930	.000	....	....	....
>>Mean>>	4	CP	Vh	100	0004	0	4.500	.482	.000	....	....	....
>>Mean>>	5	CP	Vh	100	0005	0	3.614	.228	.000	....	....	....
>>Mean>>	6	CP	Vh	100	0006	0	4.518	.482	.000	....	....	....
>>Mean>>	7	CP	Vh	100	0007	0	2.912	.386	.000	....	....	....
>>Mean>>	8	CP	Vh	100	0008	0	1.412	.421	.000	....	....	....
>>Mean>>	9	CP	Vh	100	0009	0	1.386	.184	.000	....	....	....
>>Mean>>	10	CP	Vh	100	000A	0	2.070	.544	.000	....	....	....
>>Mean>>	11	CP	Vh	100	000B	0	2.114	.149	.000	....	....	....
>>Mean>>	12	CP	Vh	100	000C	0	5.886	1.623	.000	....	....	....
>>Mean>>	13	CP	Vh	100	000D	0	3.772	.702	.000	....	....	....
>>Mean>>	14	CP	Vh	100	000E	0	3.026	.675	.000	....	....	....
>>Mean>>	15	CP	Vh	100	000F	0	2.658	.360	.000	....	....	....
>>Total>	16	CP	Vh	1600	MIX	0	59.202	14.737	.000	....	....	....

Reports VCPU steal behavior by the distance the steal dragged the VCPU.

- Lvl-00: you stole it from a CPU in your chip (z10: … in your book)
- Lvl-01: you stole it from a CPU in your book (z10: … in another book)
- Lvl-02: you stole it from a CPU on another book (z10: … not applicable)

New Report PHYSLOG

```
1FCX302  Run 2013/06/24 09:36:54          PHYSLOG
                                          Real CPU Utilization Log
 From 2013/02/19 11:49:58
 To   2013/02/19 11:56:10
 For    372 Secs 00:06:12                  Result of GFCM0107 Run
```

Interval End Time	Type	<PU Num> Conf	Ded	Total Weight	%LgclP	%Ovrhd	LpuT/L	%LPmgt	%Total	TypeT/L
>>Mean>>	CP	44	4	675	3387.1	27.947	1.008	31.870	3446.9	1.018
>>Mean>>	ZAAP	2	0	50	.093	.042	1.451	.424	.559	6.015
>>Mean>>	IFL	16	0	85	448.16	1.017	1.002	2.108	451.28	1.007
>>Mean>>	ICF	1	0	40	.004	.003	1.624	2.257	2.263	563.66
>>Mean>>	ZIIP	3	0	50	.193	.090	1.465	1.204	1.487	7.694
>>Mean>>	>Sum	66	4	900	3835.5	29.099	1.008	37.864	3902.5	1.017
11:50:04	CP	44	4	675	1963.9	33.262	1.017	36.226	2033.4	1.035
11:50:04	ZAAP	2	0	50	.004	.001	1.306	.037	.042	10.107
11:50:04	IFL	16	0	85	501.44	1.087	1.002	2.372	504.90	1.007
11:50:04	ICF	1	0	40	.007	.004	1.566	2.277	2.289	312.13
11:50:04	ZIIP	3	0	50	.005	.002	1.334	.093	.100	19.003
11:50:04	>Sum	66	4	900	2465.4	34.356	1.014	41.006	2540.7	1.031
11:50:10	CP	44	4	675	2074.2	25.632	1.012	28.117	2127.9	1.026
11:50:10	ZAAP	2	0	50	.004	.001	1.340	.003	.008	2.013
11:50:10	IFL	16	0	85	502.09	.993	1.002	2.130	505.21	1.006
11:50:10	ICF	1	0	40	.007	.004	1.568	2.165	2.176	322.32
11:50:10	ZIIP	3	0	50	.004	.001	1.354	.096	.102	24.829
11:50:10	>Sum	66	4	900	2576.3	26.632	1.010	32.511	2635.4	1.023
11:50:16	CP	44	4	675	2753.4	23.553	1.009	25.725	2802.7	1.018
11:50:16	ZAAP	2	0	50	.003	.001	1.352	.002	.007	2.015
11:50:16	IFL	16	0	85	502.84	.728	1.001	1.603	505.17	1.005
11:50:16	ICF	1	0	40	.006	.003	1.508	2.168	2.178	335.01
11:50:16	ZIIP	3	0	50	.004	.001	1.317	.093	.098	27.041
11:50:16	>Sum	66	4	900	3256.3	24.287	1.007	29.592	3310.1	1.017

You now have an easy way to see how busy your CEC is. (At last!)

Features:

- Tallied by CPU type (CP, IFL, …)
- One group of rows every sample interval
- Reports all three ways CPU gets used:
  - By logical CPUs
  - By PR/SM, chargeable
  - By PR/SM, unchargeable
- New concepts:
  - LPU T/L: like “guest T/V”
  - Type T/L: like “system T/V”

z/VM 6.3: More than Just Large Memory and HiperDispatch

Large Memory Dump: Highlights

- **Assures the system can produce a dump of a 1 TB system**
- **Includes changes to hard abend dump, SNAPDUMP, dump loader, and VM Dump Tool**
- **Changes to help improve the speed of hard abend dumps**
  - Speed of dump-to-ECKD has been improved
  - Speed of dump-to-SCSI has been improved
- **Can now stand-alone dump in hard-abend format to either ECKD or SCSI**
- **Emphasis on specifying DUMP operand on CP_OWNED statement**
- **Recovery of preallocated dump space after a SNAPDUMP**
- **SET DUMP can now list up to 30 DASD devices to receive the dump**
- **D1 R17 new fields to describe frames constituting 2 MB buffer reserved for dump**

Large Memory Dump: A Couple of Runs

ECKD

Dump Rel (O)		6.2	6.3		
Run ID		JW1E2561	JX1E2561	Delta	Pct
Dump rate (O)	rec/sec	3943	6186	2243	56.9
Dump Elapsed (O)	sec	257	164	-93	-36.2
Dump Records (O)	rec	1013483	1014549	1066	0.1

SCSI

Dump Rel (O)		6.2	6.3		
Run ID		JW1S2561	JX1S2561	Delta	Pct
Dump rate (O)	rec/sec	1136	3334	2198	193.5
Dump Elapsed (O)	sec	903	308	-595	-65.9
Dump Records (O)	rec	1025723	1026767	1044	0.1

z/VM 6.3: Other Performance Items

- **FCP Data Router**
  - Data movement assist for FCP card
  - Lets card move data from System z memory directly to its SCSI card
  - Available on z196 GA2 and later
- **Local TLB Clearing Facility**
  - Guests can now use IPTE or IDTE with the local-clearing-control (LC) bit
  - Ask your OS whether it does this
- **Access-Exception Fetch/Store Indication Facility**
  - More detail in Translation Exception Identifier when a storage access is denied
  - Ask your OS whether it can make use of this
- **VSWITCH Recovery Stall Prevention**
  - Properly handle missing interrupt from Clear Subchannel (CSCH) related to uplink port
- **CCW fast-trans and MDC both now allow Prefix-LRE CCW**
  - Important because some later Linux distros use Prefix-LRE
- **VM65156 missing path in guest mask no longer causes MDC bypass**

Monitor Record Changes

- **All the HiperDispatch changes**
- **All the Large Memory changes**
- **For FCP Data Router: D1 R19, D6 R25**
- **HiperSockets changes: D1 R19, D6 R25, D6 R26, D6 R27**
- **For Large Memory Dump: D1 R7, D3 R1**
- **VSWITCH Edge Port Aggregator: D6 R21, D6 R35**
- **VSWITCH Recovery Stall Prevention: D6 R22**
- **Additional debug: D0 R17, D0 R20, D3 R4, D3 R11, D5 R8, D5 R10, D6 R3, D6 R4, D6 R7, D6 R8, D6 R14, D6 R31, D9 R3**

z/VM Performance Toolkit

- **High Performance FICON changes**
  - SYSLOG, SYSTEM, DEVICE HPF, HPFLOG, SYSCONF, IOCHANGE, LCHANNEL all updated
- **VSWITCH HiperSockets Bridge changes**
  - GVNIC, VNIC, GVSWITCH, VSWITCH, QDIO, IOCHANGE all updated
- **LGR changes**
  - New reports LGRELOG and LGRDATA
- **Large Memory Changes**
  - 6 changed, 2 deleted, 8 new
- **HiperDispatch Changes**
  - 7 changed, 1 obsolete, 8 new

The CPU Measurement Facility

CPU Measurement Facility Counters

- **CPU MF counters are a System z hardware facility that records the performance of the CPU and nest**
  - Instructions, cycles, cache misses, … processor-ish stuff
- **Available on zEC12, zBC12, z196, z114, and z10 EC/BC**
- **The CPU MF counter values:**
  - Help IBM to understand how your workload stresses a CEC
  - Help IBM to map your workload into the LSPR curves
  - Help IBM to understand your system when there is a processor performance problem
- **z/VM 5.4 or later can all collect the CPU MF counters from the hardware**
  - z/VM 5.4 and 6.1: VM64961, UM33440 (5.4), UM33442 (6.1)
  - Counters come out in a new Monitor record, D5 R13 MRPRCMFC
- **We want volunteers to send us MONWRITE data!**
  - Your contributions will help us to understand customer workloads!

CPU MF Counters and CP Monitor, Details

- **Counter sample record is in the Processor domain**
- **MONITOR SAMPLE command manipulates counter collection**
- **QUERY MONITOR reveals whether counter collection is on**
- **z/VM writes the collected counters into the Monitor data stream**
  - D5 R13 MRPRCMFC, Processor domain, sample record
- **The D5 R13 records land in your MONWRITE data**
- **CPUMF package on www.vm.ibm.com/download/packages/ can reduce the counters**

IBM Wants Your CPU MF Counter Data

- **Your data will help IBM to build a library of customer workloads**
- **Collect an hour's worth of MONWRITE data…**
  - From a peak period, <- very important!
  - With CPU MF counters enabled,
  - With one-minute sample intervals
- **Contact Richard Lewis at rflewis at us.ibm.com**
- **Richard will send you instructions on how to transmit the data to IBM**
- **No deliverable will be returned to you**
- **We will be ever grateful for your contribution**

Other Thoughts

Evolution of System z CPC Performance Workloads

- **Now include a memory-constrained configuration**
  - Was traditionally all memory-rich
- **From 16-way to 32-way**
- **From workload-indexed to RNI-indexed**
  - We do want your CPU MF counter data
- **Our goal is to QA the CPC on workloads that represent z/VM customers' environments**

Summary

Summary

- **z/VM 6.3 is a performance release**
- **Large memory: we expect scaling to 1 TB**
- **HiperDispatch: we expect improvements for amenable workloads**
- **Large memory dump: necessary for large memory**
- **Lots of CP Monitor and z/VM Performance Toolkit changes**
- **Keep that CPU MF data coming – Richard wants to hear from you**

- **… and when you have a moment, send us some MONWRITE data ☺**