



Washington System Center

Linux Scalability VM and Linux Management Tools

Richard F. Lewis
IBM Washington System Center
rflewis@us.ibm.com

March, 2004

© 2003 IBM Corporation

IBM Washington System Center



Trademarks

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

e-business logo
IBM*
IBM eServer
IBM logo*
Performance Toolkit for VM
RMF
z/VM

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Intel is a trademark of Intel Corporation in the United States and other countries.
Java and all Java-related trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc., in the United States and other countries.
Microsoft, Windows and Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
UNIX is a registered trademark of The Open Group in the United States and other countries.

* 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.

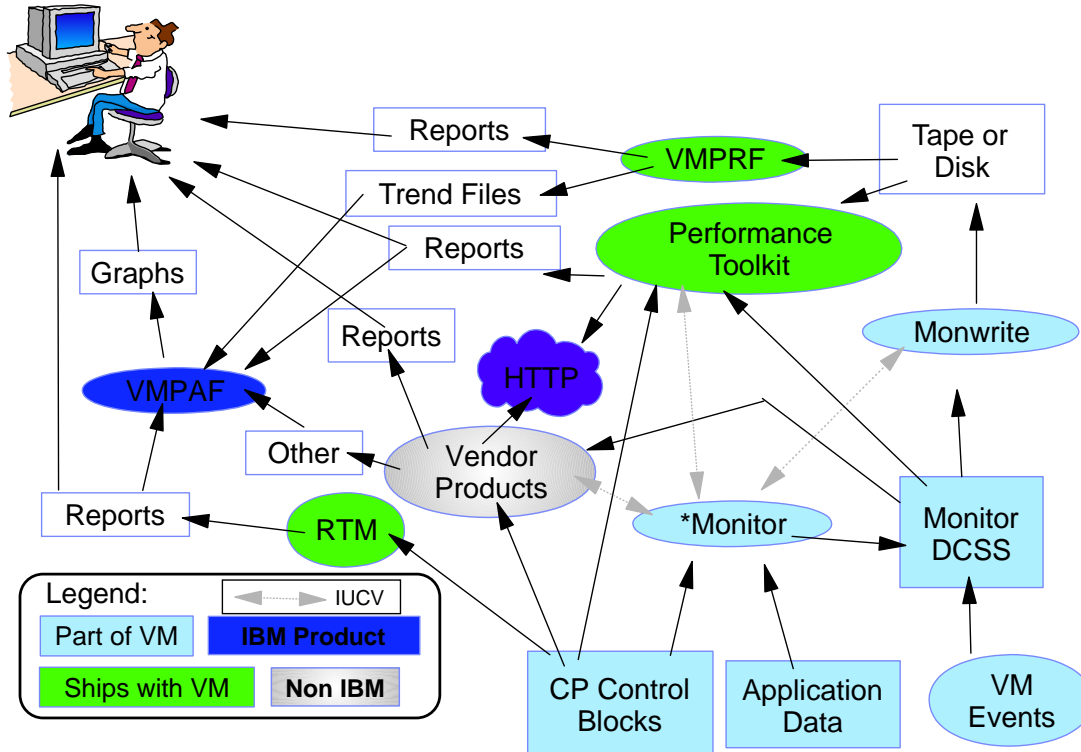
This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Agenda

- **Introduction**
- **Using Performance Toolkit**
- **Linux on zSeries Scaling**

Introduction

Performance Data Food Chain



Performance Product Strategy

- **Intention to phase out VMPRF and RTM**
 - high development costs
- **Intention to phase in FCON/ESA as Performance Toolkit for VM**
 - adds significant new function
- **Continue to encourage vendor activity**
 - competition breeds excellence
 - greater percentage of customer needs met

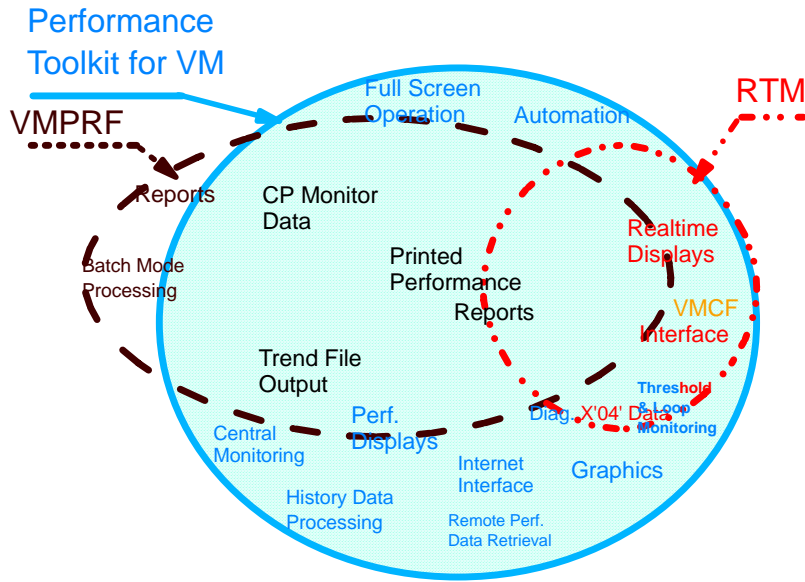
Performance Toolkit Naming

- **FCON = Full Screen Operator Console**
 - FCON/XA, FCON/ESA
- **FCX = 3 letter module prefix**
 - used in messages, displays, etc.
- **Performance Toolkit for VM = full name**
- **PERFKIT = module that invokes it**
- **PERFSVM = default userid it runs in**

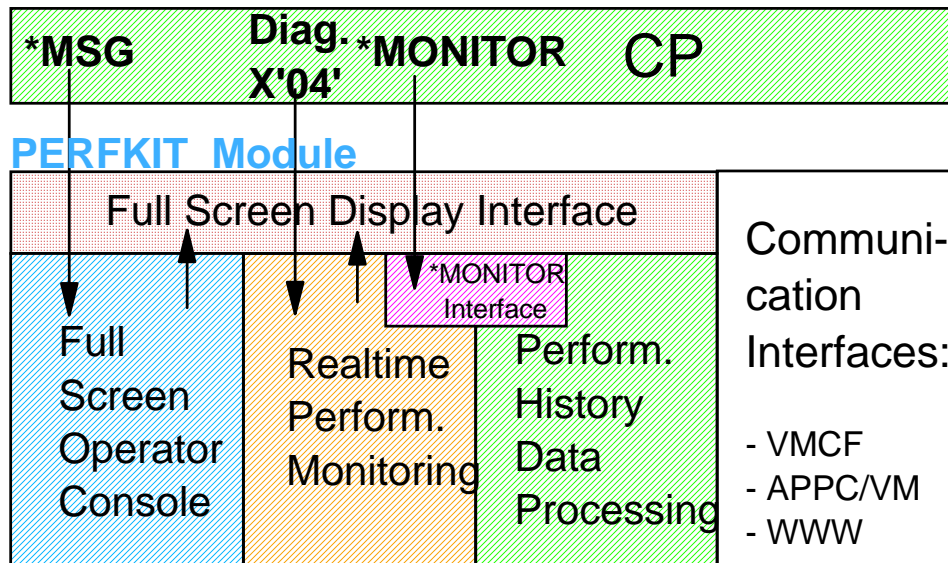
Program Functions

- **System Operation in Full-Screen Mode**
(Full Screen Operator **CON**sole)
- **Realtime Performance Monitoring**
 - Central monitoring facility for multiple systems
 - Multiple (remote, WWW) access to realtime perf. data
- **Performance History Data Processing**

Comparison with VMPRF and RTM



The PERFKIT Module

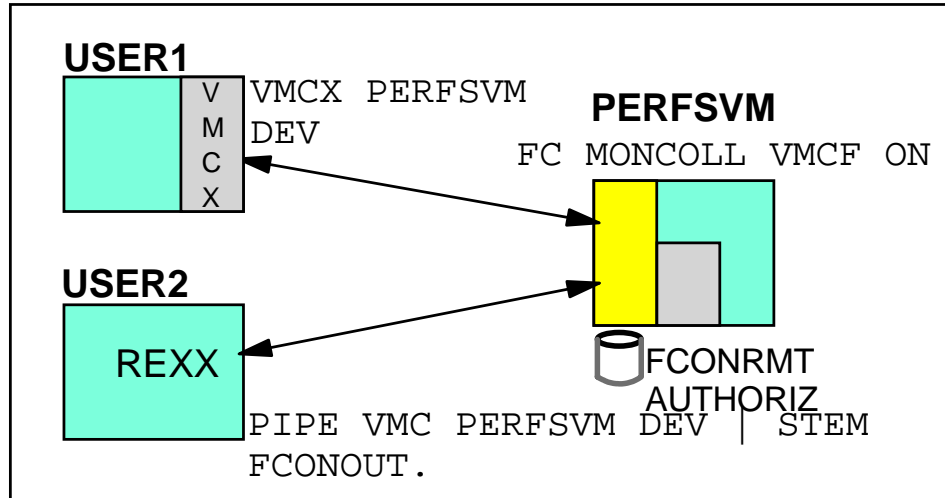


Using Performance Toolkit

Multiple (Remote) Access to Performance Data

- Local VMCF Interface
- APPC/VM Local & Remote
- WWW Interface for Standard Web Browsers

Local VMCF Interface



VMCX Example

```

FCX100      CPU 2084  SER C3A6A  Interval 11:41:23 - 11:42:23  Remote Data

CPU Load
PROC %CPU %CP %EMU %WT %SYS %SP %SIC %LOGLD  Vector Facility  Status or
P00  5  2  3  95  1  0  95  5  not installed  Master
P01  5  1  3  95  1  0  95  5  not installed  Alternate

Total SSCH/RSCH  57/s  Page rate  .0/s  Priv. instruct.  242/s
Virtual I/O rate  8/s  XSTORE paging  .0/s  Diagnose instr.  172/s
Total rel. SHARE  4300  Tot. abs SHARE  0%

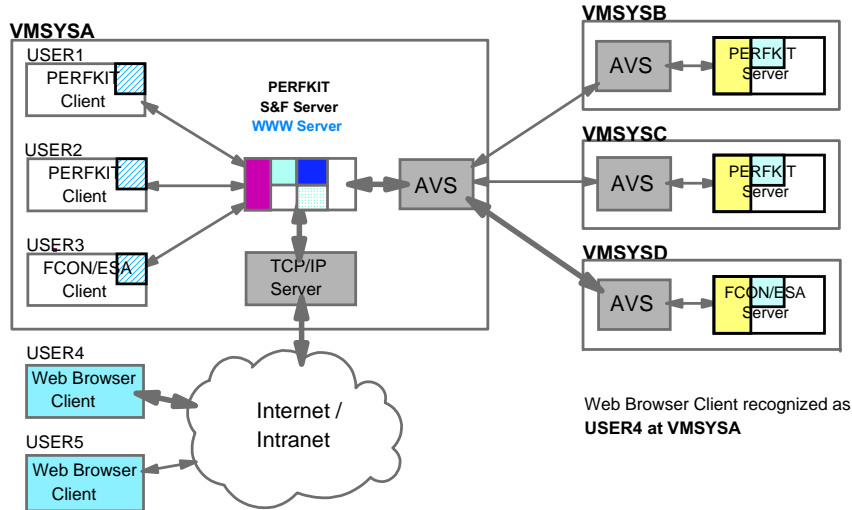
Queue Statistics:  Q0  Q1  Q2  Q3  User Status:
VMDBKs in queue  21  0  0  9  # of logged on users  31
VMDBKs loading  0  0  0  0  # of dialled users  0
Eligible VMDBKs  0  0  0  0  # of active users  20
El. VMDBKs loading  0  0  0  0  # of in-queue users  30
Tot. WS (pages)  2123k  0  0  247698  % in-Q users in PCWAIT  0
Expansion factor  2  1  2  % in-Q users in IOWAIT  65
85% elapsed time  2.688  .336  2.688  16.13  % elig. (resource wait)  0

Transactions  Q-Disp  trivial  non-trv  User Extremes:
Average users  .9  .0  .0  Max. CPU %  SEC1  2.4
Trans. per sec.  .2  .0  .0  Max. VECT %  .....  .....
Av. time (sec)  4.506  .000  .000  Max. IO/sec  SEC2  2.0
UP trans. time  .000  .000  .000  Max. PGS/s  .....  .....
MP trans. time  .000  .000  .000  Max. RESPG  RFL64ES8  432561
System ITR (trans. per sec. tot. CPU)  3.1  Max. MDCIO  .....  .....
Emul. ITR (trans. per sec. emul. CPU)  .0  Max. XSTORE  SEC2  60459

>>>>>>>> Enter next command, or 'QUIT' to exit <<<<<<<<<<<<

VM READ  OOSP1C
    
```


WWW Server Interface



Initial Perf. Data Selection Menu

Performance Toolkit for VM
 FL 440

Initial Performance Data Selection Menu (ZVMV4R40)
 Select performance screen

Command Refresh Systems Forw Help Auto-Refresh

General System Data 1. CPU load and trans. 2. Storage utilization 3. Storage subpools 4. Priv. operations 5. System counters 6. CP IUCV services 7. SPPOOL file display* 8. LPAR data 9. Shared segments A. Shared data spaces B. Virt. disks in stor. C. Transact. statistics D. Monitor data E. Monitor settings F. System settings G. System configuration H. VM Resource Manager I. Exceptions K. User defined data*	I/O Data 11. Channel load 12. Control units 13. I/O device load* 14. CP owned disks* 15. Cache extend. func.* 16. DASD I/O assist 17. DASD seek distance* 18. I/O prior. queueing* 19. I/O configuration 1A. I/O config. changes User Data 21. User resource usage* 22. User paging load* 23. User wait states* 24. User response time* 25. Resources/transact.* 26. User communication* 27. Multitasking users* 28. User configuration* 29. Linux systems*	History Data (by Time) 31. Graphics selection 32. History data files* 33. Benchmark displays* 34. Correlation coeff. 35. System summary* 36. Auxiliary storage 37. CP communications* 38. DASD load 39. Minidisk cache* 3A. Paging activity 3B. Proc. load & config* 3C. Logical part. load 3D. Response time (all)* 3E. RSK data menu* 3F. Scheduler queues 3G. Scheduler data 3H. SFS/BFS logs menu* 3I. System log 3K. TCP/IP data menu* 3L. User communication 3M. User wait states
---	---	--

Example for Performance Data Display

Performance Toolkit for VM
General I/O Device Load and Performance (ZVMV4R40)
Select a device for I/O device details

FL 440

 Auto-Refresh

Interval 19:23:03-19:24:03, on 2004/03/23 (Select average for mean data)

Addr	Type	Label/ID	Mdisk	Pa-	<Rate/s->	<----->	Time (msec)	----->	Req.	<Percent>	SEEK	Recov	<-Throt				
			Links	ths	I/O Avoid	Pend	Disc	Conn	Serv	Resp	CUWt	Qued	Busy	READ	Cyla	SSCH	Set/s
>>	All	DASD	<<														
0124	3380	OS39H7	0	1	.0	.000	0
0125	3380	HFSUS1	0	1	.0	.000	0
0A80	3390-3	OS39R7	0	1	.0	.000	0
0A82	3390-2	OS3R7A	0	1	.0	.000	0
0A85	3390-1	M3KPLX	0	1	.0	.000	0
4340	3390-3	240RES	0	2	.0	.000	0
4341	3390-3	LNXL X	0	2	.0	.000	0
4342	3390-3	SUS8	0	2	.0	.000	0
4343	3390-3	VMILN2	0	2	.0	.000	0
4344	3390-3	DEBIAN	0	2	.0	.000	0
A003	3390-3	440RES CP	123	1	.0	.000	0
A004	3390-3	440W01 CP	40	1	.0	.000	0
A005	3390-3	440W02 CP	8	1	.1	.000	9
A009	3390-3	PAC001 CP	0	1	.0	.000	0
A00A	3390-3	BASRES	0	1	.0	.000	0
A00B	3390-3	W0J001	0	1	.0	.000	0
A00C	3390-3	W0J002	0	1	.0	.000	0
A00D	3390-3	440U01	2	1	.0	.000	0
A00E	3390-3	BASW01	0	1	.0	.000	0
A00F	3390-3	M2K353	0	1	.0	.000	0

Hyperlink selection of:

Sort sequence

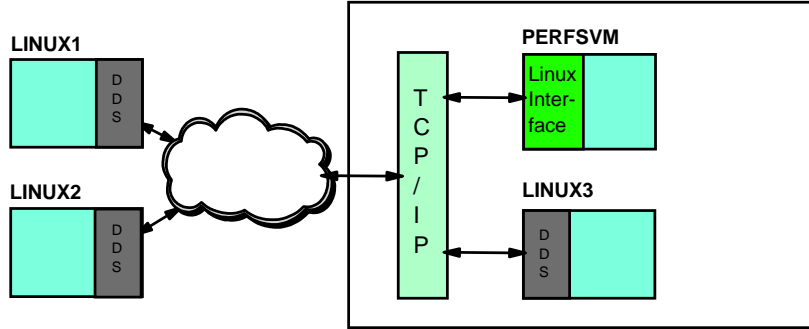
Context help

Device details

Linux Performance Data

- Retrieval Based on RMF DDS Interface
 - Originally developed for use with RMF PM
- Permanent Data Collection in Linux
- History Data Saved in Linux
- Selective 'ad hoc' Retrieval via TCP/IP
 - XML data retrieval requests
 - Linux systems not necessarily under same VM
 - Only data for requested report are retrieved

Accessing Linux Performance Data - Concepts



FC MONCOLL LINUXUSR ON

Accessing Linux Performance Data

File FCONX LINUXUSR

```

*****
** Initialization file with IP address definitions **
** for Linux systems that may have to be monitored. **
*****
*
LINUX1 1.111.111.111:8803
LINUX2 2.222.222.222:8803
LINUX3 3.333.333.333:8803
...
...
    
```

Set FC MONCOLL LINUXUSR ON - in FCONX \$PROFILE

- ➔ Defines IP addresses of Linux systems from which performance data may have to be retrieved.
- Systems to be monitored must be defined in this file!**

LXCPU userid - LINUX CPU Utilization Details

```

FCX230      CPU 2084  SER C3A6A  Interval 10:34:00 - 10:35:00  Perf. Monitor

Linux CPU Utilization for System RFLSLESS

<--- Percent CPU Utilization ---->  <-Accumulated (s)->
Processor      Total  User  Kernel  Nice  Idle  TotTm  UserTm  KernTm
>>Mean>>      0      0      0      0      100   ---    ---    ---
cpu0           0.01    0    0.01    0    99.98  ---    ---    ---
cpu1           0      0      0      0      100   ---    ---    ---

Process Name
atd.389        0      0      0      0    ---    0.01   0.01    0
bdflush.9     0      0      0      0    ---    0      0      0
init.1        0      0      0      0    ---    2.17   0.17    2
keventd.5     0      0      0      0    ---    0      0      0
kinoded.11    0      0      0      0    ---    0      0      0
klogd.274     0      0      0      0    ---    0.31   0.26    0.05
kmcheck.4     0      0      0      0    ---    0.06   0      0.06
ksoftirqd_CPU0.6 0      0      0      19   ---    195.8  0      195.8
kswapd.8      0      0      0      0    ---    14.49  0      14.49
kupdated.10   0      0      0      0    ---    7.31   0      7.31
lvm-mpd.47    0      0      0      ...  ---    0      0      0
master.380    0      0      0      0    ---    0.49   0.1     0.39
mdrecoveryd.12 0      0      0      0    ---    0      0      0
migration_CPU0.2 0      0      0      0    ---    0      0      0
migration_CPU1.3 0      0      0      0    ---    0      0      0
portmap.312   0      0      0      0    ---    ...    ...    ...
qethsoftd0004.213 0      0      0      0    ---    ...    0      0
qmgr.394      0      0      0      0    ---    0.54   0.15    0.39
    
```

LXMEM userid - LINUX Memory Details

```

FCX229      CPU 2084  SER C3A6A  Interval 10:34:00 - 10:35:00  Perf. Monitor

Linux Memory Util. & Activity Details for System RFLSLESS

Total memory size      502MB  Swap space size      99MB
Total memory used      498MB  % Swap space used    0%
  Used for buffer      23MB   Swap-in rate         0/s
  Used for shared      0MB    Swap-out rate        0/s
  Used for cache      320MB  Page-in rate         0/s
Total free memory      3MB    Page-out rate        1.866/s

<----- Size ----->  <----- Page Fault Rate/s ----->
(Bytes)      (kB)      Minor  Major  <-Incl.Children->
Process Name  VirtSize  ResidSet  MinPgFlt  MajPgFlt  MinPFltC  MajPFltC
gpmddsrv.5931 28667900  2024     0        0          0         0
gpmddsrv.5932 28667900  2024     0        0          0         0
gpmddsrv.5933 28667900  2024     0        0          0         0
gpmddsrv.5934 28667900  2024     0        0          0         0
gpmddsrv.5935 28667900  2024     0        0          0         0
qmgr.394      7454720  1840     0        0          0         0
bash.423     3002370  1832     0        0          0         0
sshd.318     4386820  1596     0        0          0         0
master.380   4759550  1548     0        0          0         0
pickup.5790  4726780  1464     0        0          0         0
klogd.274    2019330  1196     0        0          0         0
gengat.5918  2699260  1172     0        0          0         0
procat.5922  2695170  1148     1        0          0         0
login.418    2146300  1144     0        ....      0         0
    
```

LXFILSYS userid - LINUX FileSystem Details

```

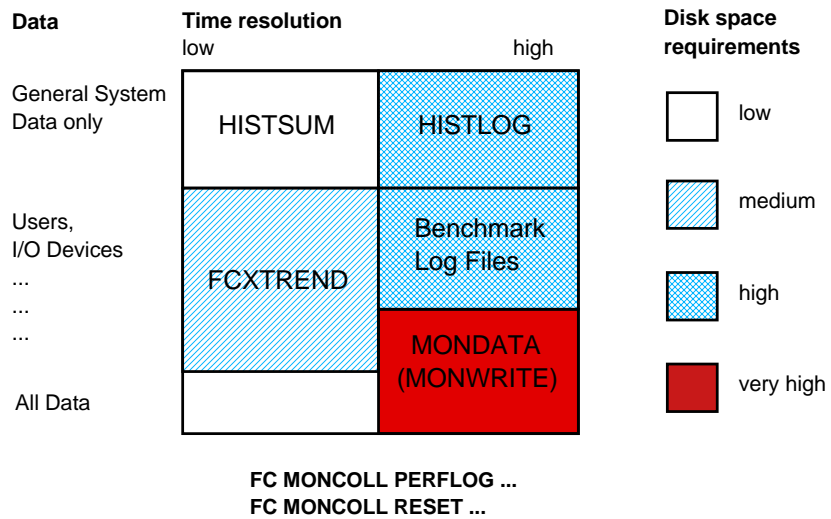
FCX228      CPU 2084  SER C3A6A  Interval 10:35:00 - 10:36:00  Perf. Monitor

Linux Filesystem Usage for System RPLSLES8

DASD I/O Activity
I/O request rate per second      0.43
I/O response time/request (msec) 10.05
I/O response time/sector (msec)  0.837

-----
Filesystem      <---- MBytes ---->  <-Percent->
Name            Size      Free    %Used %Free
>Total>         7228    4779   30.4  69.5
/dev/dasdal     2310     525    76.0  23.9
/dev/dasdbl     2310    1820   16.9  83.0
/dev/dasdc1    2310    2139    2.3  97.6
/dev/dasddl      47       44     0    100
shmfs           251     251     0    100
    
```

History Data Files



Graphics

- Simple Plots with Commands **PLOT...**
 - No additional graphics SW required
- GDDM Line Graphics with Commands **GRAPH...**
 - Requires GDDM on the system where graphics are to be shown
- Line Graphics with Java Applet via WWW Interface
 - Based on graphics capability of WS and Web Browser's Java support
 - No additional graphics SW required

GRAPHICS Selection Menu

Performance Toolkit for VM Graphics Selection Menu (ZVMV4R40)
FL 440

Command Refresh Systems Menu Return Help Auto-Refresh

General Specifications

Output format:

Data origin:

Graphics type:

Selected period:

Selected days:

Selected hours:

Variable Selection

X-Variable: SSCH/RSCH rate

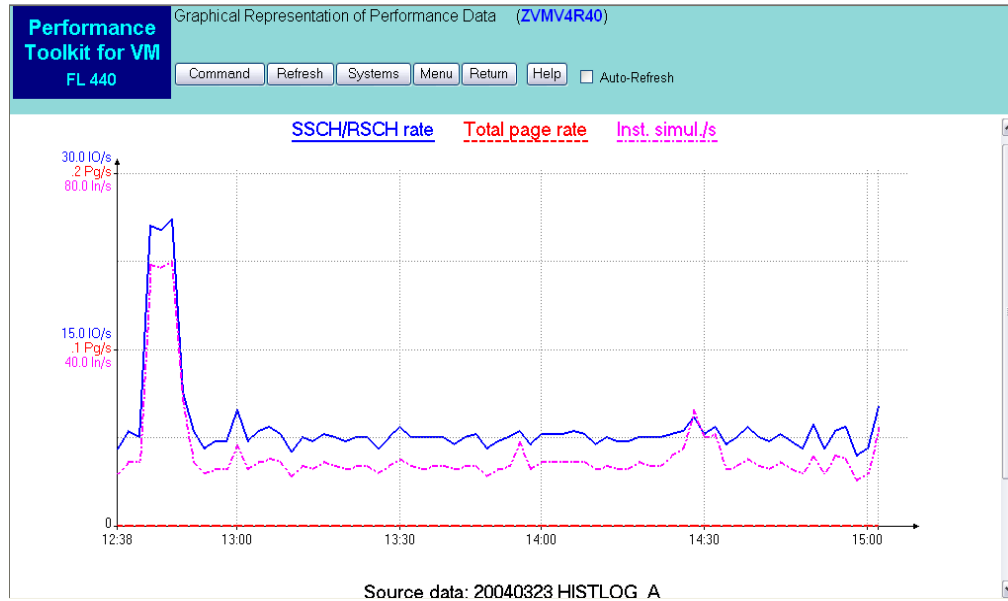
Truncate at:

Y-Variables:

1	<input type="text" value="IO/S"/>	SSCH/RSCH rate
2	<input type="text" value="PG/S"/>	Total page rate
3	<input type="text" value="priv"/>	
4	<input type="text"/>	

Cumulative

GRAPHSUM Example



IDLEUSER (RTM Compatibility)

FCX238		CPU 2064		SER 51524		Status		06:58:11		Perf. Monitor	
Userid	Min Idle	Userid	Min Idle	Userid	Min Idle	Userid	Min Idle	Userid	Min Idle	Userid	Min Idle
AVS	30	BUCKETS	189	DATAMOVE	2	DGTSRV01	189	DGTSRV02	189		
DGTSRV03	189	DIRMAINT	1	FTPSEVB	189	GCS	189	HAMILTJL	189		
HARDWARE	189	HOLDER	189	IMAPAUTH	189	IMAP3SRV	8	ISPVMS	189		
JSWITZER	189	K4SERV	8	LININS	189	LNXMSTR	189	ML1	189		
MULTISRV	8	OPERACCT	6	OPERSYMP	189	OSADMIN1	189	PERFSVM	189		
PERFTEST	189	PORTMAP	8	RAICHER	189	RECOVERY	189	RMSMASTR	5		
RTMTST4	189	RXAGENT1	8	SFCM1	189	SFSESA	189	SMSMASTR	189		
SMSSRV01	189	SNMPD	8	SNMPQE	1	SQLMACH	2	TCPMAINB	12		
TCPMAINC	189	TFTPD	8	TOOLS	189	TPOPER	189	VMNFS	3		
...											
...											

Selectable with IDLEUSER command

PROCSUM Log (VMPRF Compatibility)

```

FCX239      CPU 2064  SER 51524  Interval 03:49:11 - 07:14:11  Perf. Monitor

      <----- CPU -----> <----- Spin Lock Activity ----->
      <--Ratio-->           <----- Total -----> <--- Scheduler ---> <--
Interval    Pct      Cap- On-  Locks Average  Pct  Locks Average  Pct  <--
End Time    Busy    T/V  ture line /sec   usec Spin  /sec   usec Spin st
>>Mean>>   1.0  1.66 .7789  9.0  74.8  1.825 .002  58.4  2.029 .001
05:30:11   1.0  1.71 .7760  9.0  65.9  1.941 .001  50.5  2.208 .001
05:35:11   1.0  1.71 .7672  9.0  69.1  1.881 .001  52.7  2.151 .001
05:40:11   1.0  1.69 .7717  9.0  71.7  1.611 .001  55.2  1.798 .001
05:45:11   1.0  1.69 .7794  9.0  72.5  1.646 .001  56.0  1.834 .001
05:50:11   1.0  1.67 .7748  9.0  71.2  1.578 .001  55.1  1.762 .001
05:55:11   1.0  1.69 .7708  9.0  72.4  1.642 .001  56.9  1.801 .001
06:00:11   1.0  1.68 .7804  9.0  70.0  1.558 .001  55.0  1.714 .001
...
...
    
```

Similar to SYSTEM_SUMMARY2_BY_TIME
 Selectable with PROCSUM command

VSWITCH (Virtual Switches Data)

```

FCX240      Data for 2002/11/13  Interval 10:58:41 - 10:59:41  Perf. Monitor

----- .
      Q Time <--- Outbound/s ---> <--- Inbound/s ----> <--- Signal
      S Out  Bytes <--Packets-->  Bytes <--Packets--> <-- issued/
Addr  Userid  V  Sec  T_Byte T_Pack T_Disc R_Byte R_Pack R_Disc Write Read
>> System <<  8 300  0 .0 .0  0 .0 .0 .0 .0 .0
0113 TCPIP1  8 300  0 .0 .0  0 .0 .0 .0 .0 .0
0116 TCPIP1  8 300  0 .0 .0  0 .0 .0 .0 .0 .0
0119 TCPIP1  8 300  0 .0 .0  0 .0 .0 .0 .0 .0
...
...
    
```

Selectable with VSWITCH command

VMRM (VM Resource Manager Data)

```
FCX241      Data for 2003/04/07  Interval 10:48:14 - 10:49:14  Perf. Monitor
.
.
.
VM Resource Manager      Impor <-- DASD --> <-- CPU ---> Active
Server  Workload         tance  D-Goal D-Act   C-Goal C-Act  Samples
IRDSVM  WORK1                10    100  ...    100  ...    1
IRDSVM  WORK2                 5     50  ...    50   ...    1
IRDSVM  WORK3                 1     1   ...    1   ...    1
IRDSVM  WORK4                10    100  100    100  72    1
IRDSVM  WORK5                 5     50  100    50   77    1
IRDSVM  WORK6                 1     1   100    1    63    1
IRDSVM  WORK7                10    100  100    100  63    1
IRDSVM  WORK8                 5     50  100    50    3    1
IRDSVM  WORK9                 1     1   ...    1   ...    1
...
...
```

Selectable with VMRM command

Quick Setup Summary...

- **Enable product (using VMFINS ENABLE)**
- **Uncomment CP MONITOR commands in PERFSVM PROFILE EXEC**
- **Customize FCONX \$PROFILE**
 - Update FC MONCOLL RESET for production of trend records - add T to time specifications
 - E.g. FC MONCOLL RESET 06:00T ...
 - To gather Linux data
 - FC MONCOLL LINUXUSR ON
 - Create FCONX LINUXUSR file listing Linux systems
 - To allow vmcx and appc data display
 - FC MONCOLL VMCF ON
 - Create FCONRMT AUTHORIZ file
 - Userids authorized to display data or execute commands

Quick Setup Summary...

■ Customize FCONX \$PROFILE (cont.)

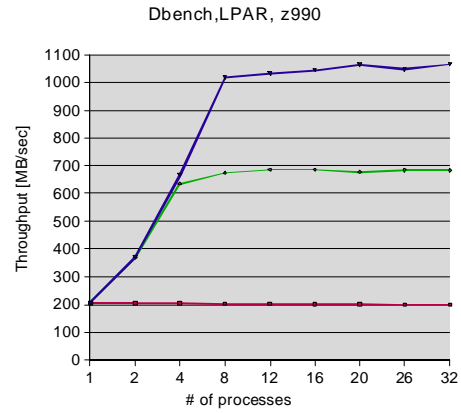
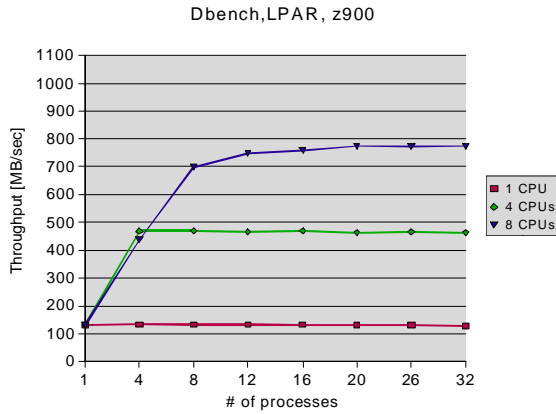
- To allow web based data display
 - FC MONCOLL WEBSRV ON TCPIP TCPIP 81
 - Create FCONRMT SYSTEMS file
 - nodename PERFSVM ESA N FCXRES00
 - Add to FCONRMT AUTHORIZ
 - nodename PERFSVM S&FSERV
 - Update PROFILE TCPIP to authorize PERFSVM to PORT 81

■ For use of vmcx and appc remote display from userid other than PERFSVM

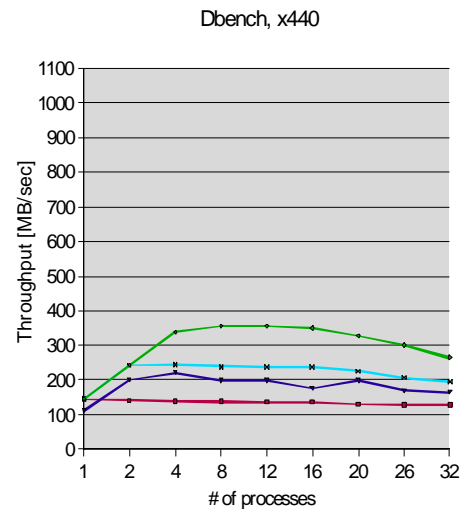
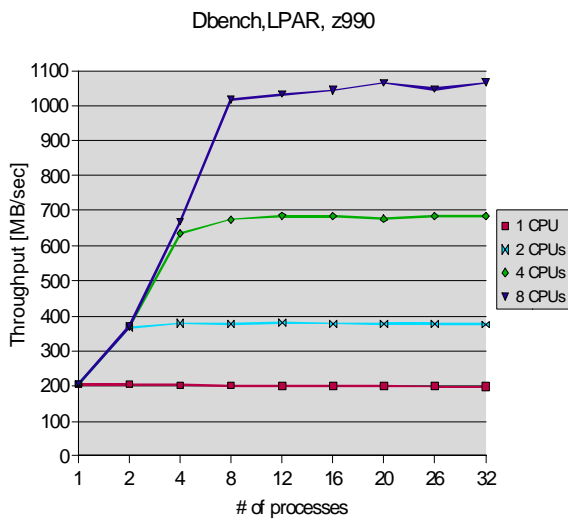
- Make contents of 4VMPTK40 200 mdisk available to performance data display virtual machines
 - Content of this minidisk includes perffit module, vmcx module, etc.

Linux Scaling

Scalability - z900 vs z990, ext2, 31 bit



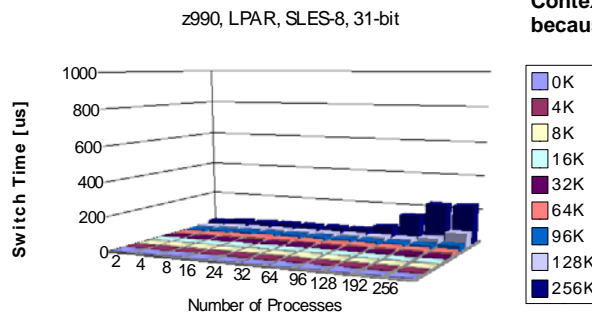
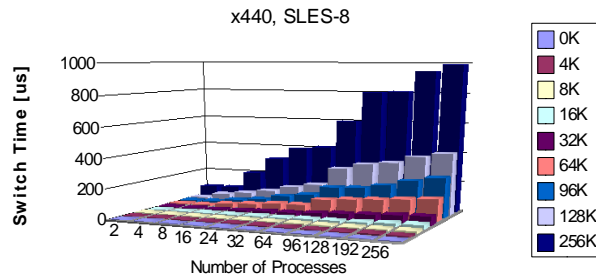
Scalability - z990 vs Intel, ext2, 31/32Bit



z990 shows good scaling behavior

x440 shows best throughput with 4 CPU, strong throughput degradation with more than 4 CPUs

Kernel - Context Switches



Context Switches much faster on zSeries because of large shared caches

Summary

- **z/VM Performance Toolkit**
 - Positioned to replace VMRTM and VMPRF
 - Currently provides most of the functions of VMRTM and VMPRF
 - Local and remote performance data access
 - Display of Linux performance data using rmfpm
 - History data collection and storage
 - Data graphing capability
- **Linux scalability**
 - z990 excellent context switching platform
 - Linux scales well to 8 processors and more depending on workload

References

- **General information**
 - <http://www.vm.ibm.com/related/perfkit/>
- **Performance Toolkit Book**
 - <http://publibz.boulder.ibm.com/epubs/pdf/hcs17a00.pdf>
- **Comparison to VMPRF**
 - <http://www.vm.ibm.com/related/perfkit/pkitprf.html>
- **Comparison to RTM**
 - <http://www.vm.ibm.com/related/perfkit/pkitrtm.html>