Managing z/VM and Linux Performance Best Practices

Jm Newell
IBM
jnewell@us.ibm.com

Session 7362

Thursday August 6th 2010

Agenda

- **Abstract**
- **Challenges**
- **Best Practices**
- **OMEGAMON XE for z/VM and Linux**
  - Integrated Monitoring Approach
  - Use Case Scenarios
- **Conclusion**

Abstract

- In today's virtualized environments it's important that we adhere to a set of best practices when it comes to managing the environment. Even though our applications may all run within the same physical environment many of the challenges faced managing an application stack spread across multiple servers still exist.

- Furthermore, there are unique challenges associated with z/VM and Linux environments for less experienced users.

- This presentation highlights the Performance and Availability management best practices for z/VM and Linux on System z while showing how OMEGAMON XE on z/VM and Linux can be used to measure for deviations from those best practices.

Virtual Linux servers have unique challenges versus running on physical machines.

- z/VM System Programmers and Linux Administrators may not be in the organization.

- We find that it is easy to over allocate resources; therefore, our monitoring examines resource usage of hardware, hypervisor, as well as the virtual machine. Real-time and historical metrics demonstrate peaks periods as well as average runtimes.

"Best Practices"

- z/VM
  - System Scope items
    - Maintenance, Memory, Paging, DASD, VDISK, Processors/ LPAR, System Utilization, DASD I/O, CP-Owned Allocation
    - Workloads: Virtual Processors, Paging

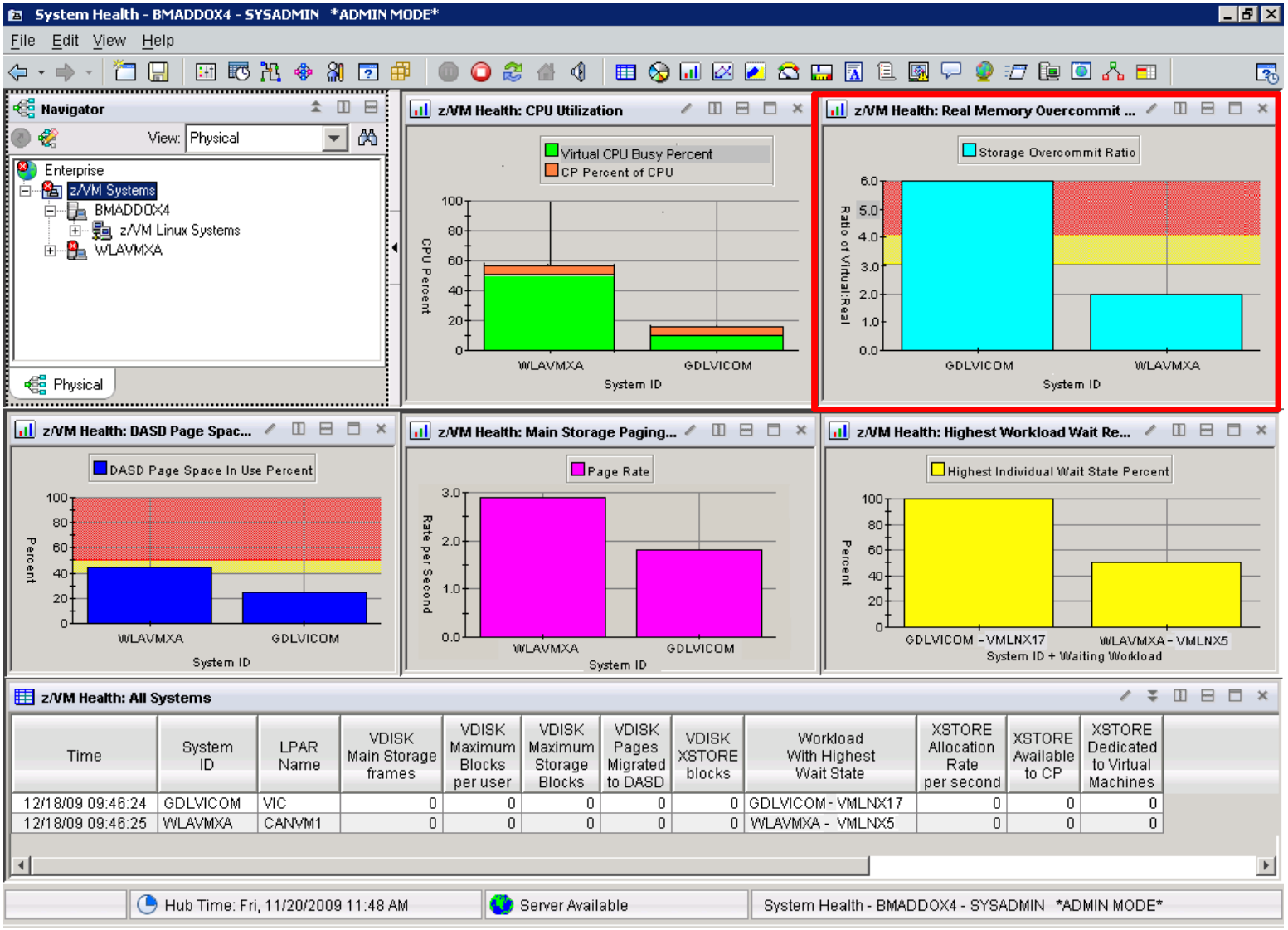
Maintenance Levels

- **Recommend maintaining current service levels.**
- **Apply latest Recommended Service Upgrade (RSU):**
  - z/VM Family
    - Released every 3-6 months
    - Contains cumulative service including all pre- and co-requisites in a pre-built format.
  - Includes service for all integrated components and the following pre-installed program products:
    - DirMaint, VM/RACF, Performance ToolKit
      - Available on tape, DVD, or electronically.
  - Separate Maintenance Stream for OMEGAMON
    - Available electronically, or via platform appropriate media.

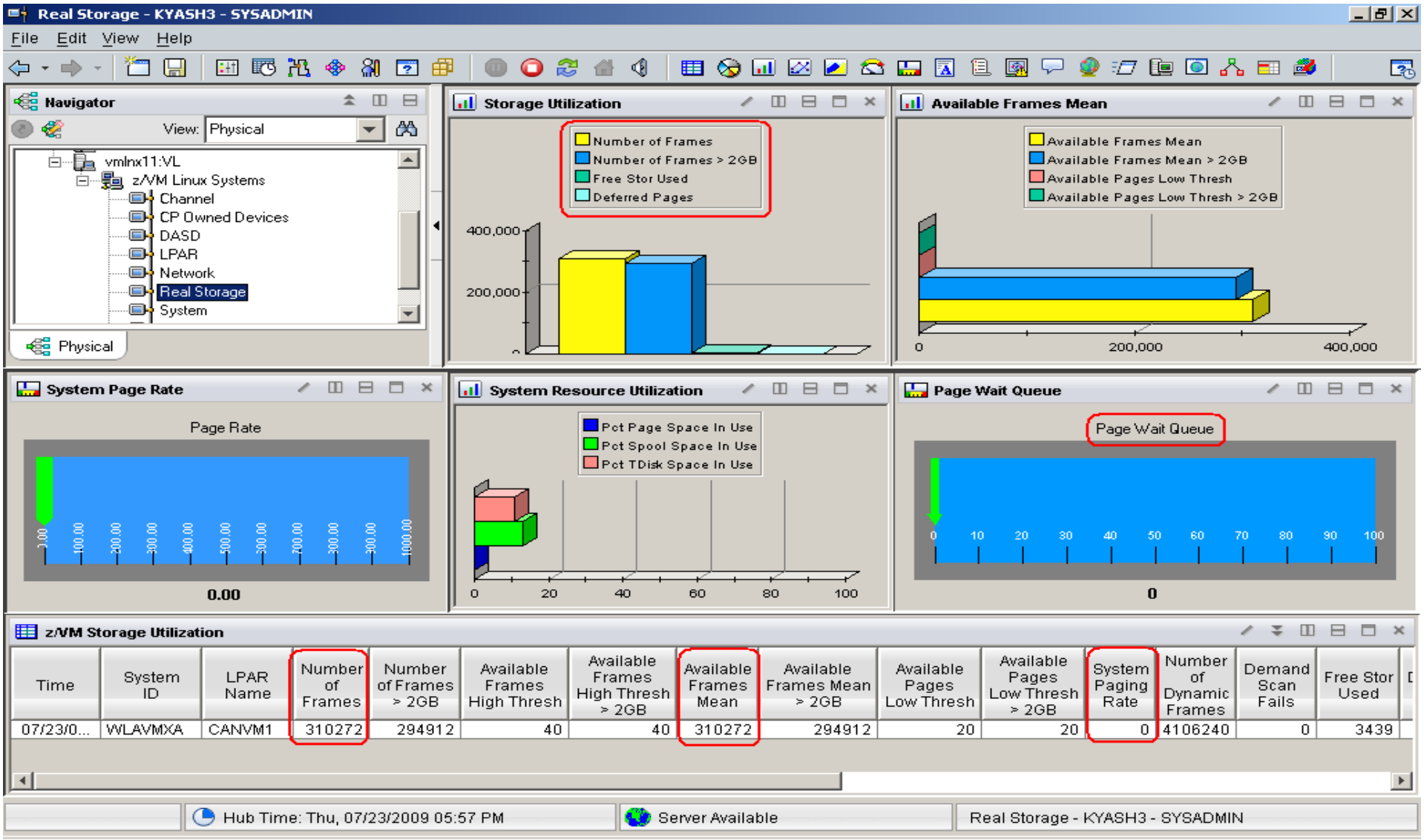
Memory Configuration

- **Plan on a virtual to real (V:R) memory ratio in the range of 1.5:1 to 3:1.**
- **Recommend configuring some processor memory as expanded storage:**
  - Serves as high speed cache.
  - Increases consistency of response time.
  - See http://www.vm.ibm.com/perf/tips/storconf.html for the gory details.
- **Rule of Thumb - start with 25% of memory configured as expanded:**
  - Typically 2–4GB of expanded storage is sufficient, 1GB minimum.
  - The lower the paging rate, the lower the amount of expanded storage required.
  - The greater the number of page frames available in central storage above 2GB, the higher the amount of expanded storage required.

OMEGAMON Memory Configuration



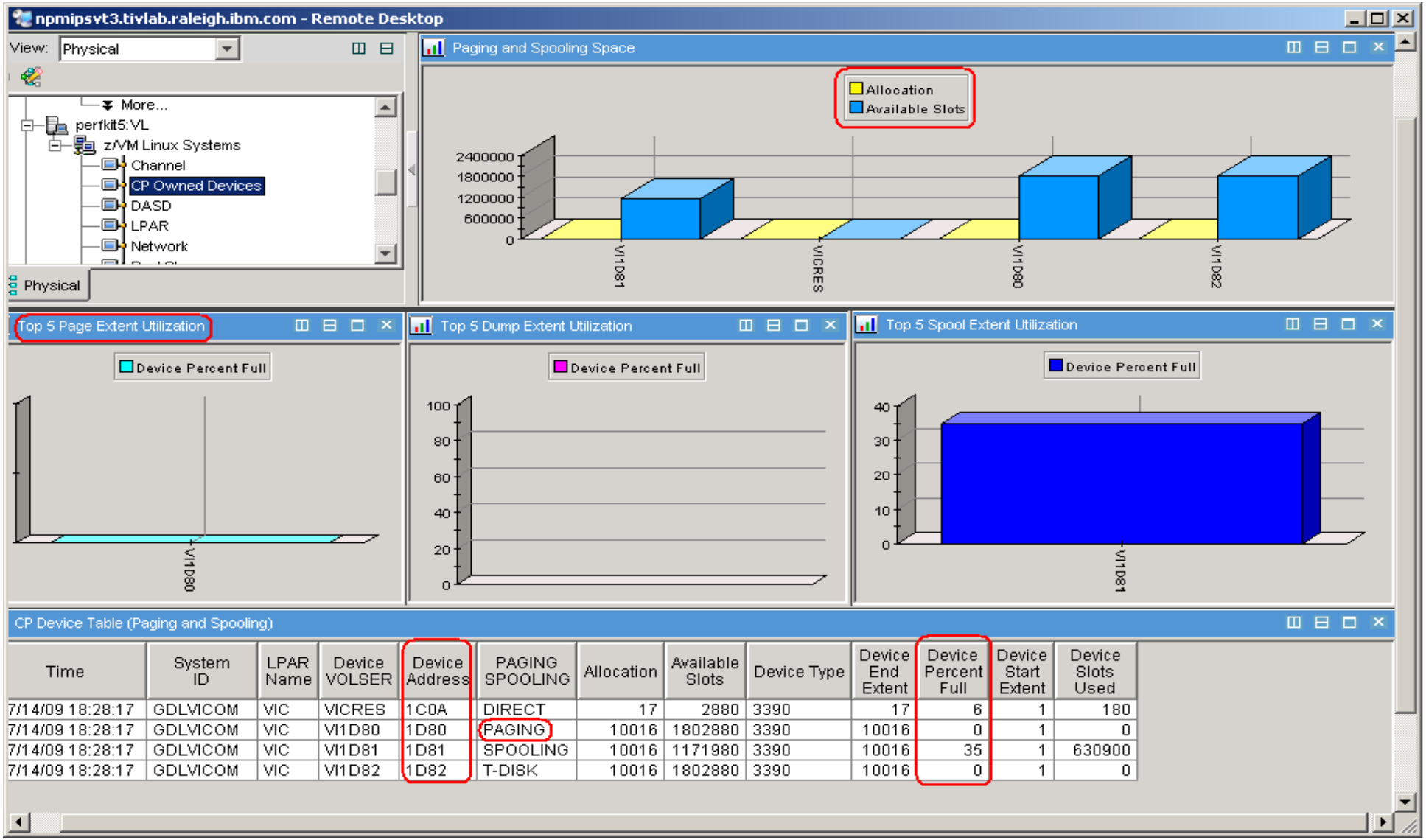
OMEGAMON Memory Configuration



Paging Subsystem

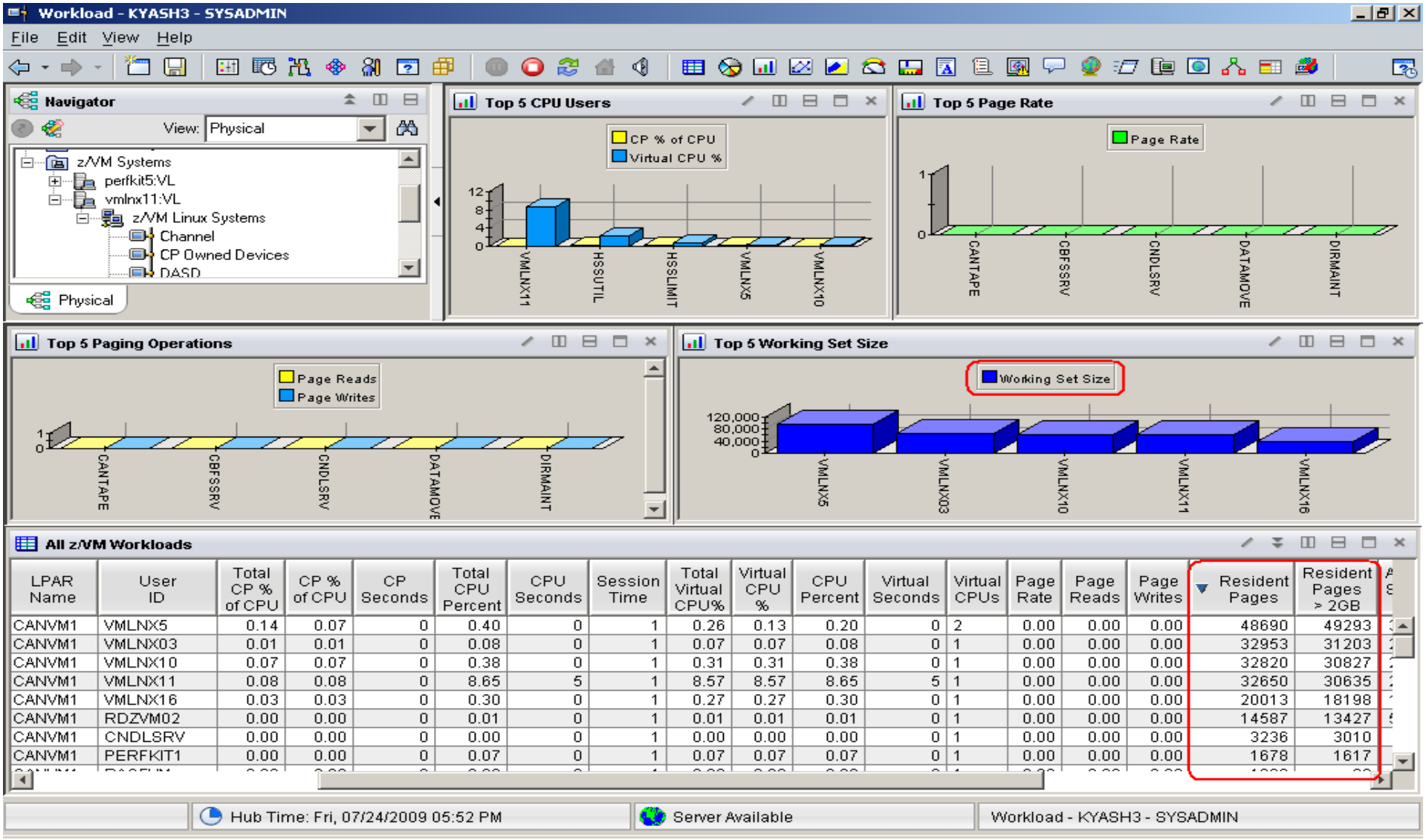
- **Plan for DASD page space utilization < 50% for performance reasons:**
  - Page space tends to get fragmented over time.
  - Large contiguous free space allows for greater paging efficiency.
  - Monitor usage with OMEGAMON XE or Q ALLOC PAGE command.
  - Block page size is the key performance indicator:
    - Aim for double digits – 10 or more pages per block set.
    - Performance Toolkit report DEV CPOWN (FCX109) "Block Page Size" Use multiple channels to spread out I/O to paging devices.
- **When Paging and Spool space fills, z/VM abends.**
  - VM warning message occurs around 90%
  - By the it's typically too late
- **Do not mix page space with any other space on a volume.**
- **Recommend using devices of the same size/geometry and performance characteristics.**
- **Calculation guidelines are located in the CP Planning and Administration Manual.**

OMEGAMON CP Owned Devices – Paging Subsystem



Time	System ID	LPAR Name	Device VOLSER	Device Address	PAGING SPOOLING	Allocation	Available Slots	Device Type	Device End Extent	Device Percent Full	Device Start Extent	Device Slots Used
7/14/09 18:28:17	GDLVICOM	VIC	VICRES	1C0A	DIRECT	17	2880	3390	17	6	1	180
7/14/09 18:28:17	GDLVICOM	VIC	VI1D80	1D80	PAGING	10016	1802880	3390	10016	0	1	0
7/14/09 18:28:17	GDLVICOM	VIC	VI1D81	1D81	SPOOLING	10016	1171980	3390	10016	35	1	630900
7/14/09 18:28:17	GDLVICOM	VIC	VI1D82	1D82	T-DISK	10016	1802880	3390	10016	0	1	0

Paging – Workload Workspace



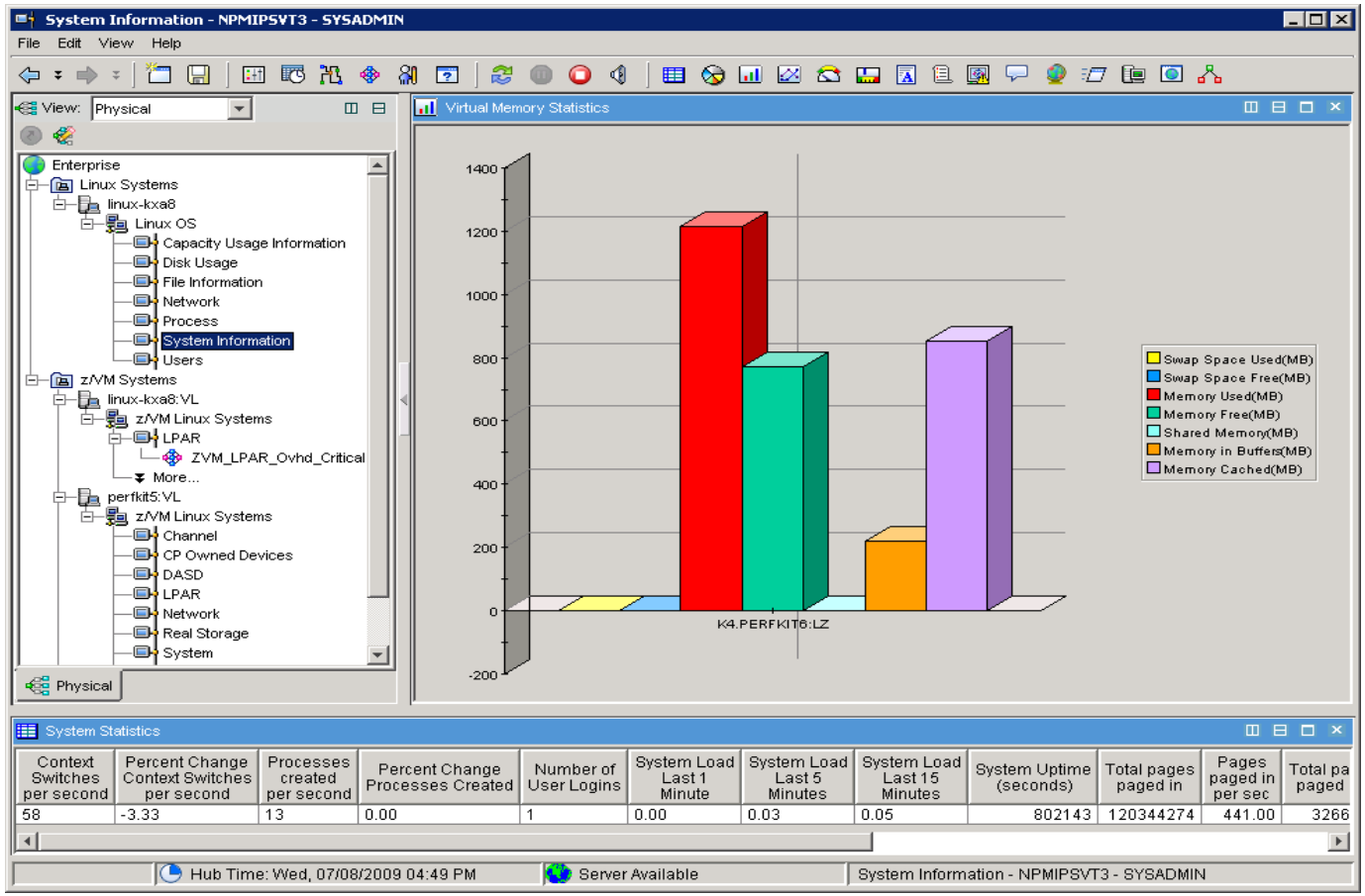
LPAR Name	User ID	Total CP % of CPU	CP % of CPU	CP Seconds	Total CPU Percent	CPU Seconds	Session Time	Total Virtual CPU%	Virtual CPU %	CPU Percent	Virtual Seconds	Virtual CPUs	Page Rate	Page Reads	Page Writes	Resident Pages	Resident Pages > 2GB
CANVM1	VMLNX5	0.14	0.07	0	0.40	0	1	0.26	0.13	0.20	0	2	0.00	0.00	0.00	48690	49293
CANVM1	VMLNX03	0.01	0.01	0	0.08	0	1	0.07	0.07	0.08	0	1	0.00	0.00	0.00	32953	31203
CANVM1	VMLNX10	0.07	0.07	0	0.38	0	1	0.31	0.31	0.38	0	1	0.00	0.00	0.00	32820	30827
CANVM1	VMLNX11	0.08	0.08	0	8.65	5	1	8.57	8.57	8.65	5	1	0.00	0.00	0.00	32650	30635
CANVM1	VMLNX16	0.03	0.03	0	0.30	0	1	0.27	0.27	0.30	0	1	0.00	0.00	0.00	20013	18198
CANVM1	RDZVM02	0.00	0.00	0	0.01	0	1	0.01	0.01	0.01	0	1	0.00	0.00	0.00	14587	13427
CANVM1	CNDLSRV	0.00	0.00	0	0.00	0	1	0.00	0.00	0.00	0	1	0.00	0.00	0.00	3236	3010
CANVM1	PERFKIT1	0.00	0.00	0	0.07	0	1	0.07	0.07	0.07	0	1	0.00	0.00	0.00	1678	1617

Is my Linux guest sized correctly?

- **More memory is not always better**
- **Excessive virtual machine sizes negatively impact performance.**
- **Linux will use all available memory**
  - Any space it doesn't need will be used as file buffer cache. Notice the large amount of cache used in example—indicates that guest may be sized too large
  - Larger Linux guests means that z/VM has to page out larger virtual machines when running other guests
- **One method—use monitor to watch for swapping. Shrink guest size until it starts swapping.**
- **Another method. Look at the Working Set Size for the Virtual Machine. This shows what z/VM is using for the guest.**
- **To handle some swapping, define a VDISK. This is much faster than swapping to a real minidisk**

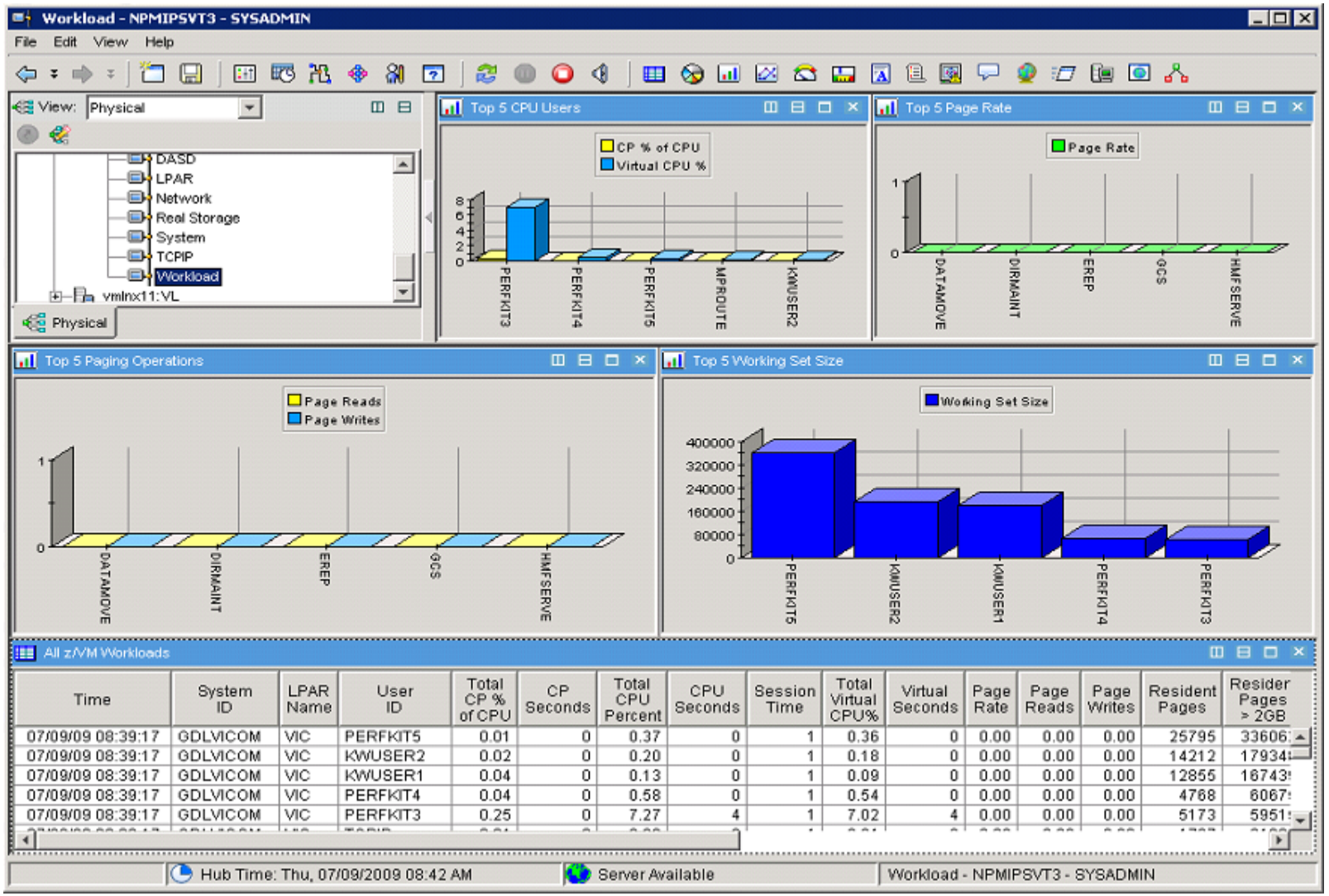
Sizing Linux Guests

Memory usage of a particular Linux virtual machine



Sizing Linux Guests

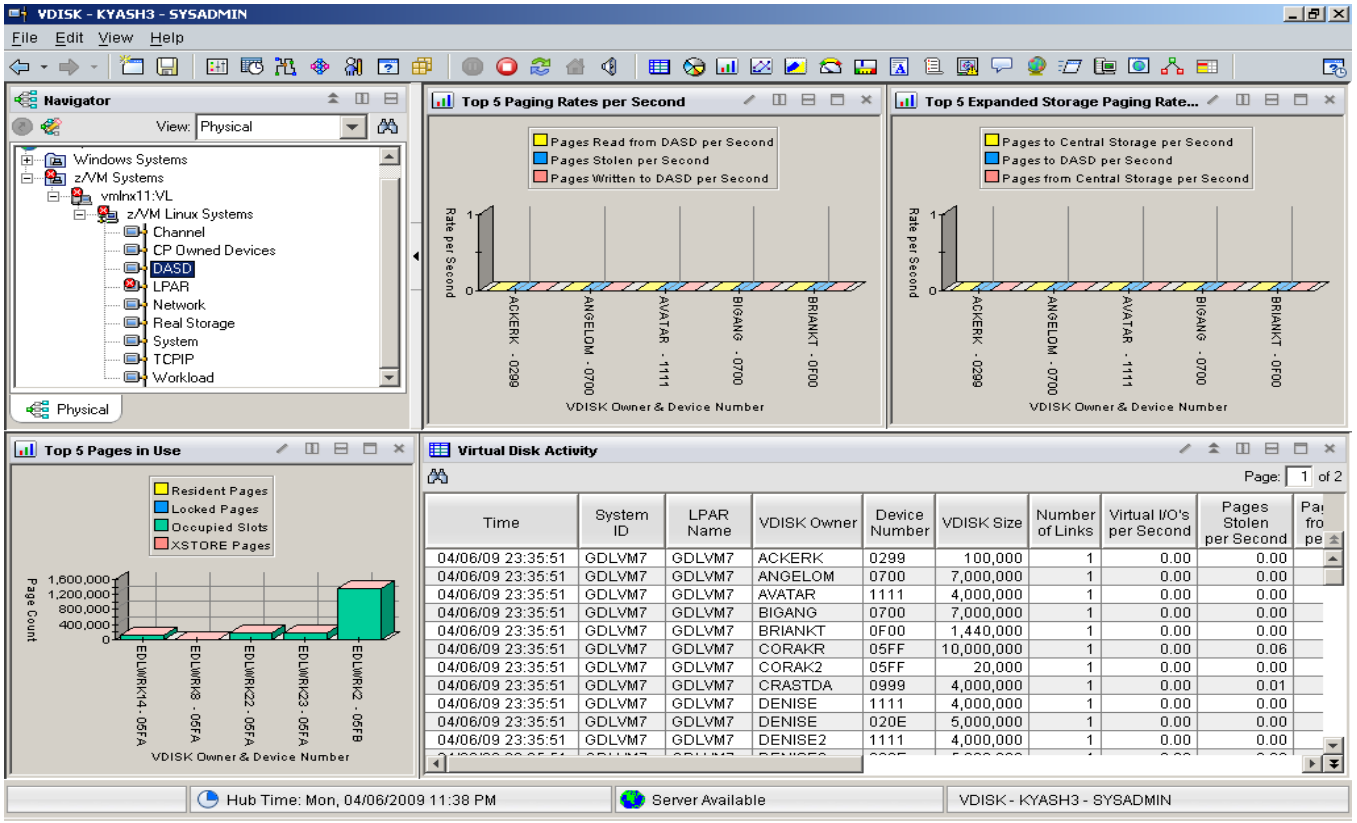
Working Set Size can be found on the Workload workspace of the z/VM agent



VDISK

- **What is it?**
  - FBA (Fixed Block Architecture disk) device emulated in-memory
    - Translation: Very fast “device”.
  - High performance paging device for Linux on z.
  - Memory is allocated by CP from the Dynamic Paging Area
  - Allocated only when referenced
    - Allocating a 10 MB device does NOT instantly consume 10 MB of pages.
    - Pages are allocated when needed.
  - Need to factor VDISK in the overall memory planning for system.

VDISK Workspace



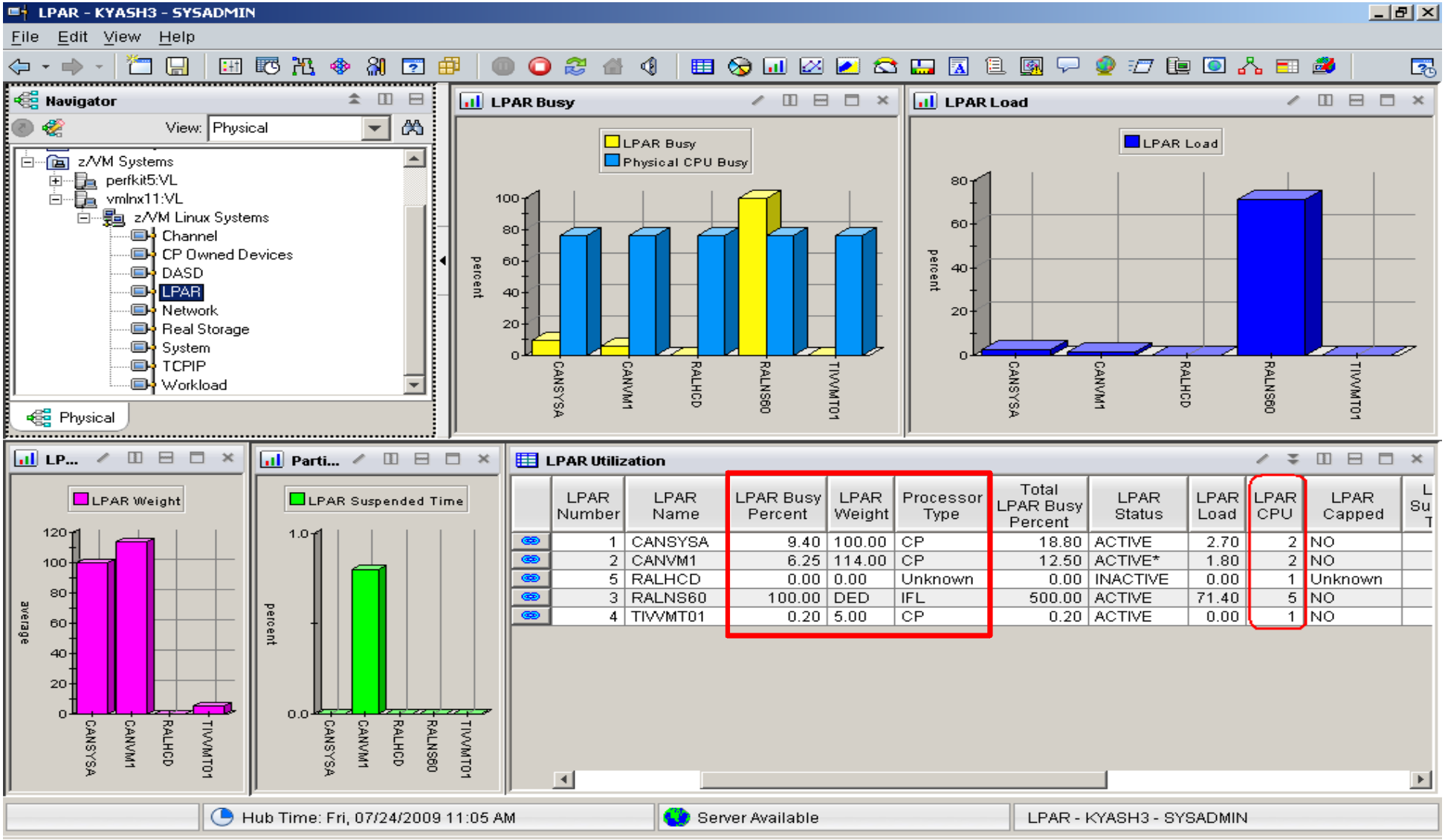
Do not ignore the hardware!

- **Just because Linux resources are virtual, do not ignore the hardware!**
  - Hardware is another potential layer of shared resources.
  - LPAR weight, CPU sharing, LPAR load, and other attributes need to be monitored for overall system performance.
  - The measurement should include the entire CEC and not just the LPAR hosting z/VM.

Processors

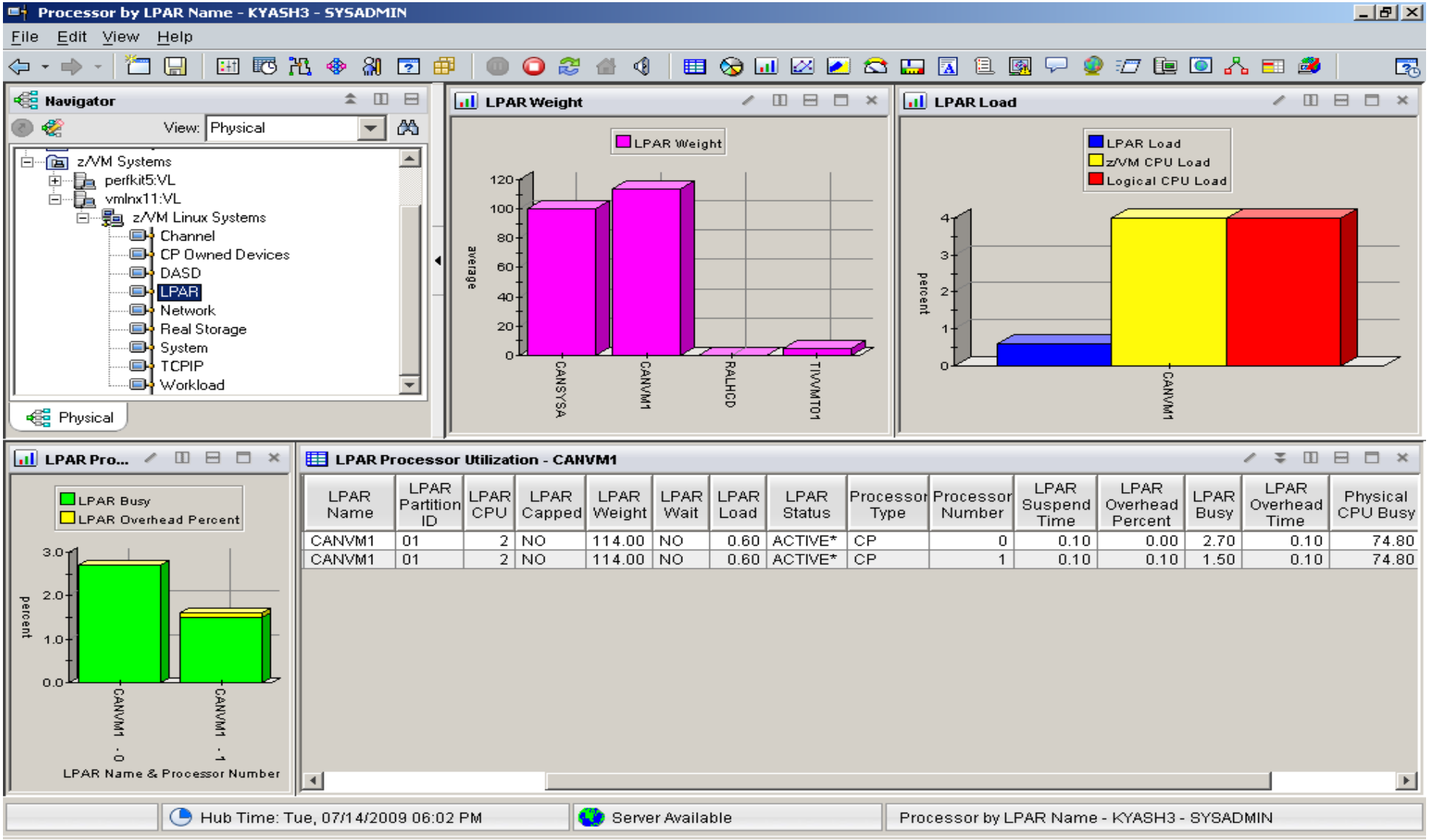
- **Logical Processors**
  - LPAR recommendation – no greater than a 4:1 logical to real ratio.
  - z/VM 5.1 and z/VM 5.2 support up to 24 processors.
  - z/VM 5.3 and z/VM 5.4 support up to 32 processors.
- **Virtual Processors**
  - Various guest systems and workloads scale differently.
  - No rule-of-thumb.
  - Virtual Machine recommendation:
    - Configure the number of virtual processors per guest for peak workload, but no more.
    - Do not define more virtual processors to a guest than logical processors defined to a z/VM LPAR.
  - High diagnose x'44' rates may be an indication of too many virtual processors.
    - Performance Toolkit reports CPU (FCX100) or PRIVOP (FCX104) can be used to monitor diagnose rates.

LPAR Utilization Workspace



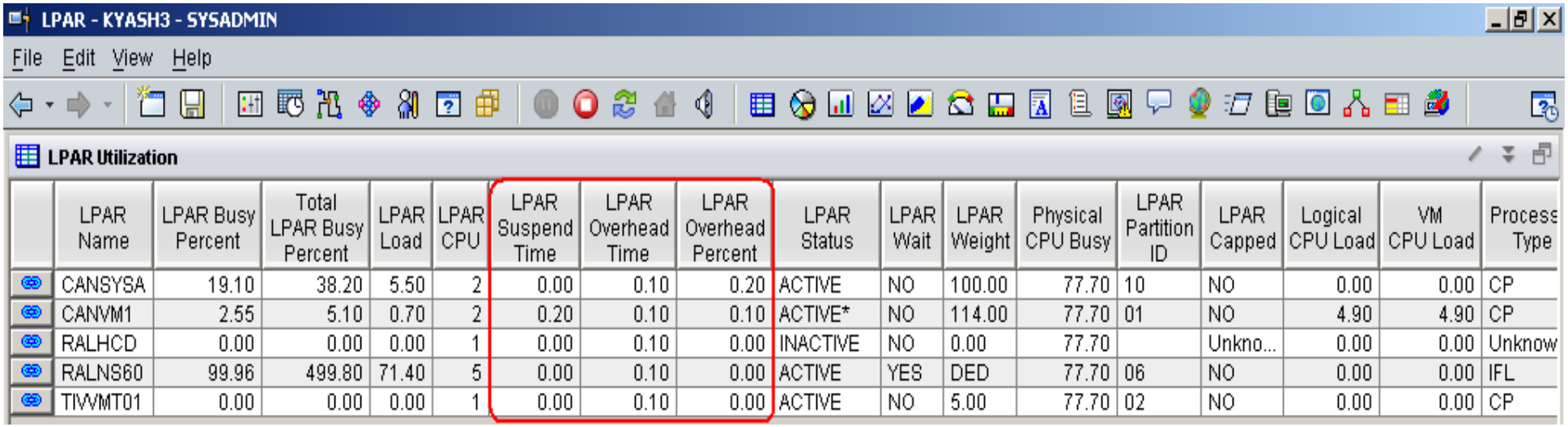
	LPAR Number	LPAR Name	LPAR Busy Percent	LPAR Weight	Processor Type	Total LPAR Busy Percent	LPAR Status	LPAR Load	LPAR CPU	LPAR Capped
	1	CANSYSA	9.40	100.00	CP	18.80	ACTIVE	2.70	2	NO
	2	CANVM1	6.25	114.00	CP	12.50	ACTIVE*	1.80	2	NO
	5	RALHCD	0.00	0.00	Unknown	0.00	INACTIVE	0.00	1	Unknown
	3	RALNS60	100.00	DED	IFL	500.00	ACTIVE	71.40	5	NO
	4	TIVVMT01	0.20	5.00	CP	0.20	ACTIVE	0.00	1	NO

Processors – LPAR Processor Workspace



LPAR Name	LPAR Partition ID	LPAR CPU	LPAR Capped	LPAR Weight	LPAR Wait	LPAR Load	LPAR Status	Processor Type	Processor Number	LPAR Suspend Time	LPAR Overhead Percent	LPAR Busy	LPAR Overhead Time	Physical CPU Busy
CANVM1	01	2	NO	114.00	NO	0.60	ACTIVE*	CP	0	0.10	0.00	2.70	0.10	74.80
CANVM1	01	2	NO	114.00	NO	0.60	ACTIVE*	CP	1	0.10	0.10	1.50	0.10	74.80

LPAR Utilization Workspace – Tabular View



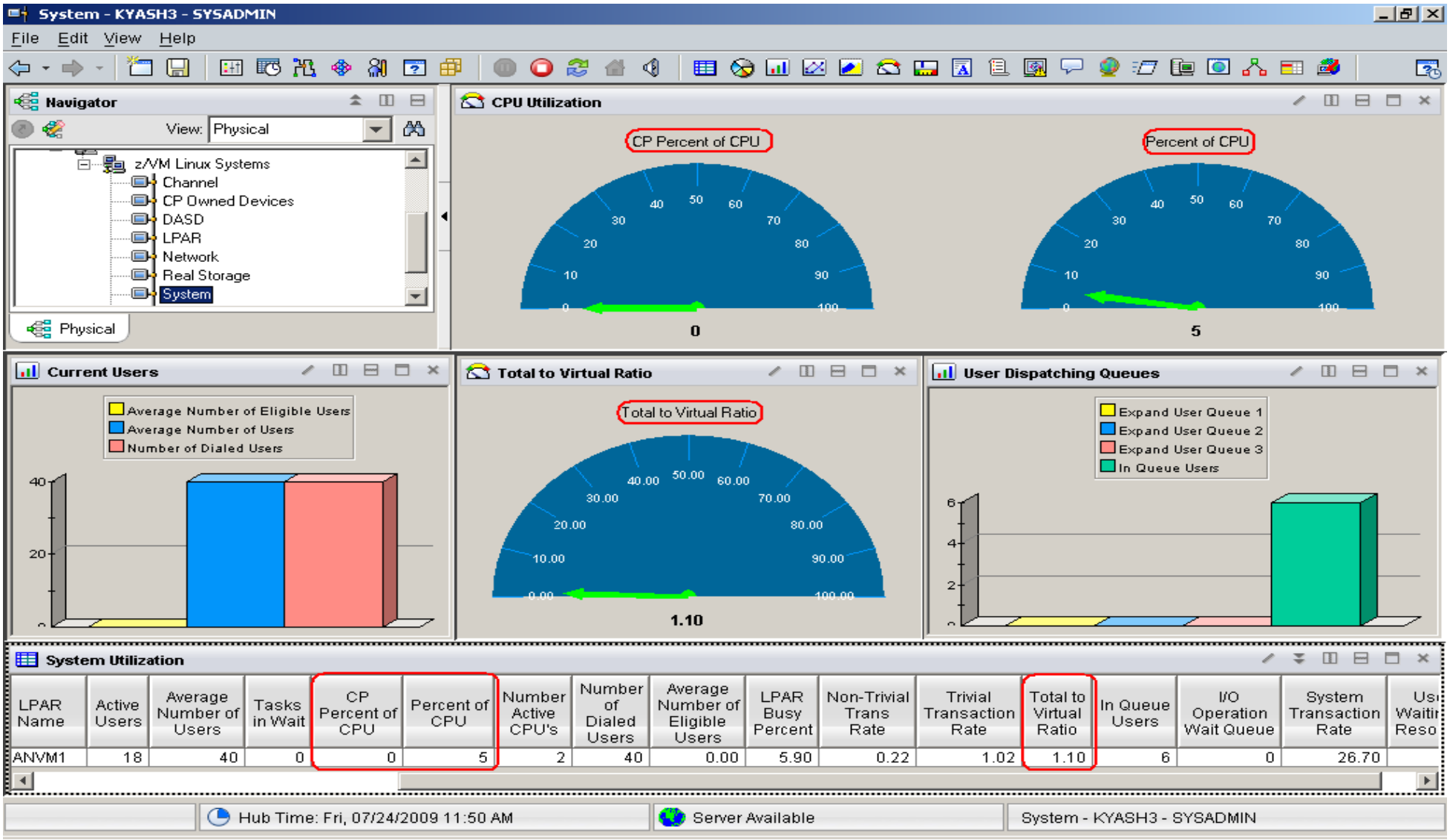
	LPAR Name	LPAR Busy Percent	Total LPAR Busy Percent	LPAR Load	LPAR CPU	LPAR Suspend Time	LPAR Overhead Time	LPAR Overhead Percent	LPAR Status	LPAR Wait	LPAR Weight	Physical CPU Busy	LPAR Partition ID	LPAR Capped	Logical CPU Load	VM CPU Load	Process Type
	CANSYSA	19.10	38.20	5.50	2	0.00	0.10	0.20	ACTIVE	NO	100.00	77.70	10	NO	0.00	0.00	CP
	CANVM1	2.55	5.10	0.70	2	0.20	0.10	0.10	ACTIVE*	NO	114.00	77.70	01	NO	4.90	4.90	CP
	RALHCD	0.00	0.00	0.00	1	0.00	0.10	0.00	INACTIVE	NO	0.00	77.70		Unkno...	0.00	0.00	Unknow
	RALNS60	99.96	499.80	71.40	5	0.00	0.10	0.00	ACTIVE	YES	DED	77.70	06	NO	0.00	0.00	IFL
	TIVVMT01	0.00	0.00	0.00	1	0.00	0.10	0.00	ACTIVE	NO	5.00	77.70	02	NO	0.00	0.00	CP

- LPAR Suspend Time: RoT: 5% Suspend time is yellow line, 10% is red line for concern.
- LPAR Overhead: This should generally be less than 5% of the Physical IFLs (CEC in an all-IFL configuration) for general LPAR management overhead, and then less than 5% of the z/VM partition IFLs.

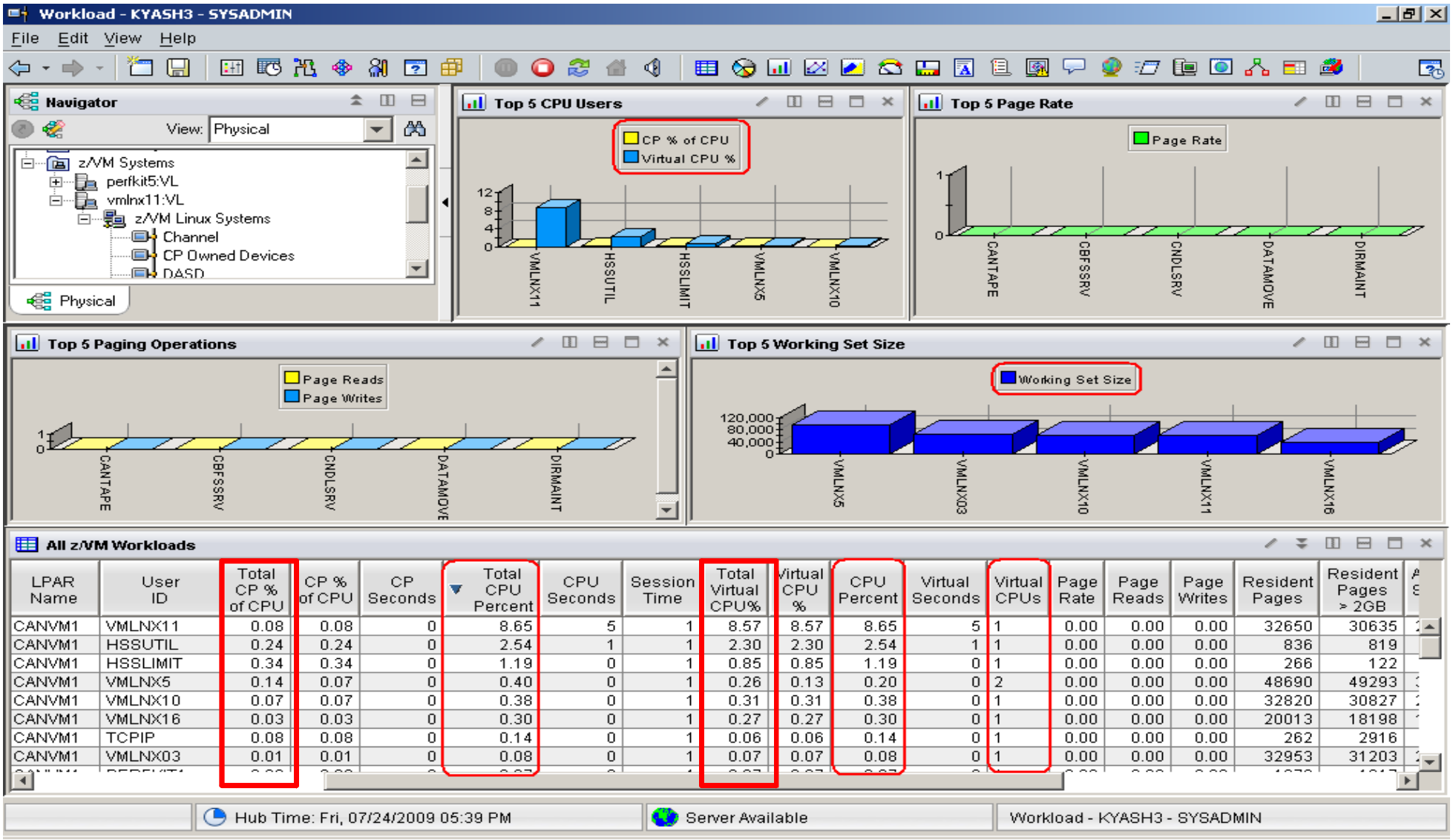
Processor Utilization

- Total Processor Utilization **This is the processor utilization from the VM perspective and includes CP, VM System, and Virtual CPU time.**
- System Time: **This is the processor time used by the VM control program for system functions that are not directly related to any one virtual machine. This should be less than 10% of the total.**
- CP Processor Time: **This is the processor time used by the VM control program in support of individual virtual machines.**
- Virtual Processor Time: (Emulation Time): **This is processor time consumed by the virtual machine and the applications within it.**
- Total to Virtual Ratio **The ratio of total processor time to virtual processor time is often used as an indicator of z/VM efficiency or overhead. The closer to 1.0, the better the z/VM efficiency. RoT: Should explore causes of a ratio over 1.30.**

System Processor Utilization Workspace



Processors – Workload Workspace

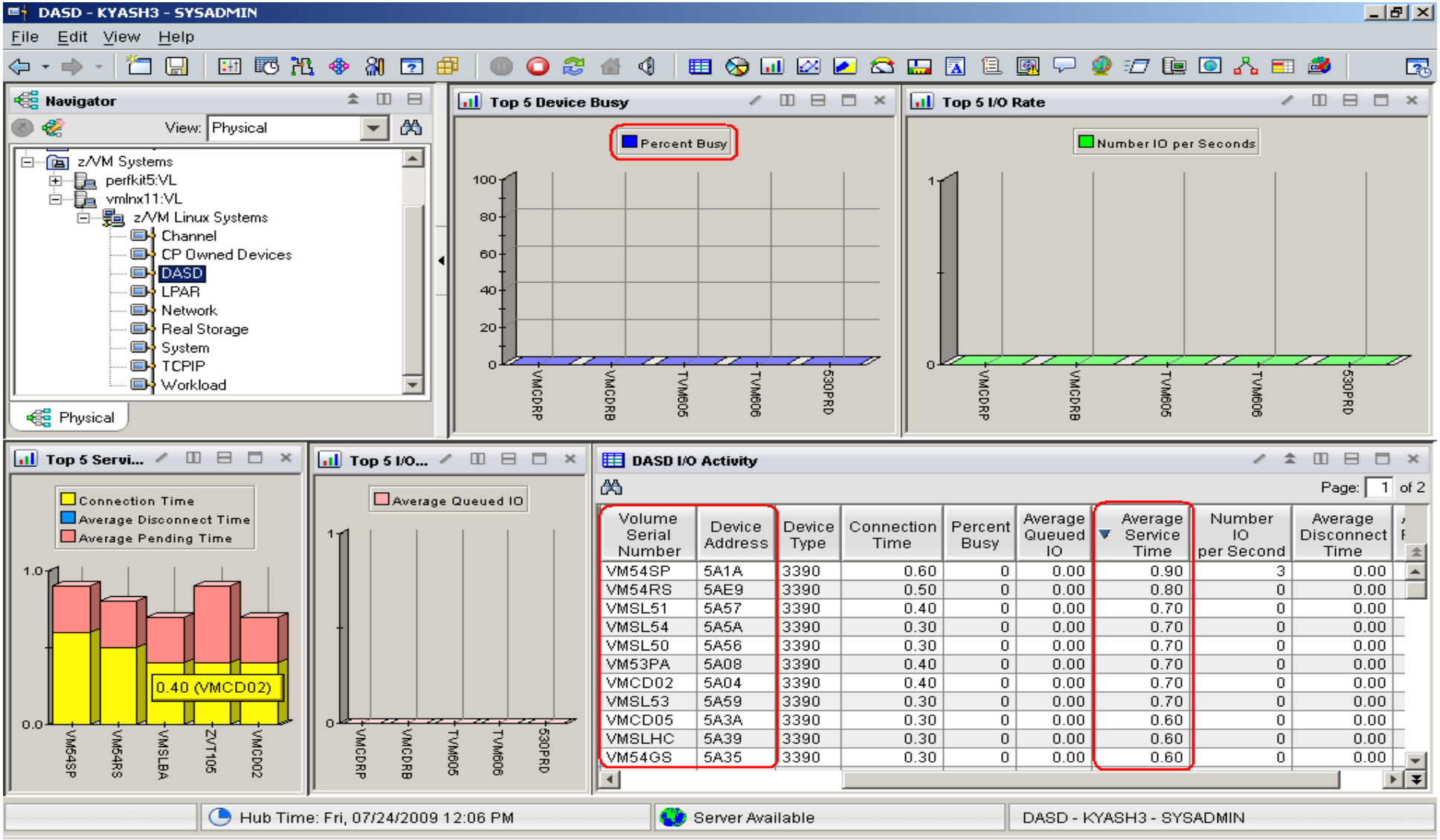


LPAR Name	User ID	Total CP % of CPU	CP % of CPU	CP Seconds	Total CPU Percent	CPU Seconds	Session Time	Total Virtual CPU%	Virtual CPU %	CPU Percent	Virtual Seconds	Virtual CPUs	Page Rate	Page Reads	Page Writes	Resident Pages	Resident Pages > 2GB
CANVM1	VMLNX11	0.08	0.08	0	8.65	5	1	8.57	8.57	8.65	5	1	0.00	0.00	0.00	32650	30635
CANVM1	HSSUTIL	0.24	0.24	0	2.54	1	1	2.30	2.30	2.54	1	1	0.00	0.00	0.00	836	819
CANVM1	HSSLIMIT	0.34	0.34	0	1.19	0	1	0.85	0.85	1.19	0	1	0.00	0.00	0.00	266	122
CANVM1	VMLNX5	0.14	0.07	0	0.40	0	1	0.26	0.13	0.20	0	2	0.00	0.00	0.00	48690	49293
CANVM1	VMLNX10	0.07	0.07	0	0.38	0	1	0.31	0.31	0.38	0	1	0.00	0.00	0.00	32820	30827
CANVM1	VMLNX16	0.03	0.03	0	0.30	0	1	0.27	0.27	0.30	0	1	0.00	0.00	0.00	20013	18198
CANVM1	TCPIP	0.08	0.08	0	0.14	0	1	0.06	0.06	0.14	0	1	0.00	0.00	0.00	262	2916
CANVM1	VMLNX03	0.01	0.01	0	0.08	0	1	0.07	0.07	0.08	0	1	0.00	0.00	0.00	32953	31203

Direct Access Storage Devices (DASD)

- Avg Pending Time for DASD
  - Average pending time for real DASD I/Os. RoT: Should be less than 1 millisecond.
- **Items worth keeping an eye on:**
  - **Number of I/O's per Second, Percent Busy**
  - **Avg Service Time** Average service time for real DASD devices (sum of the pending, connect, and disconnect times).
  - **DASD I/O Rate** Rate of traditional real I/Os per second to real DASD devices. Worth monitoring.

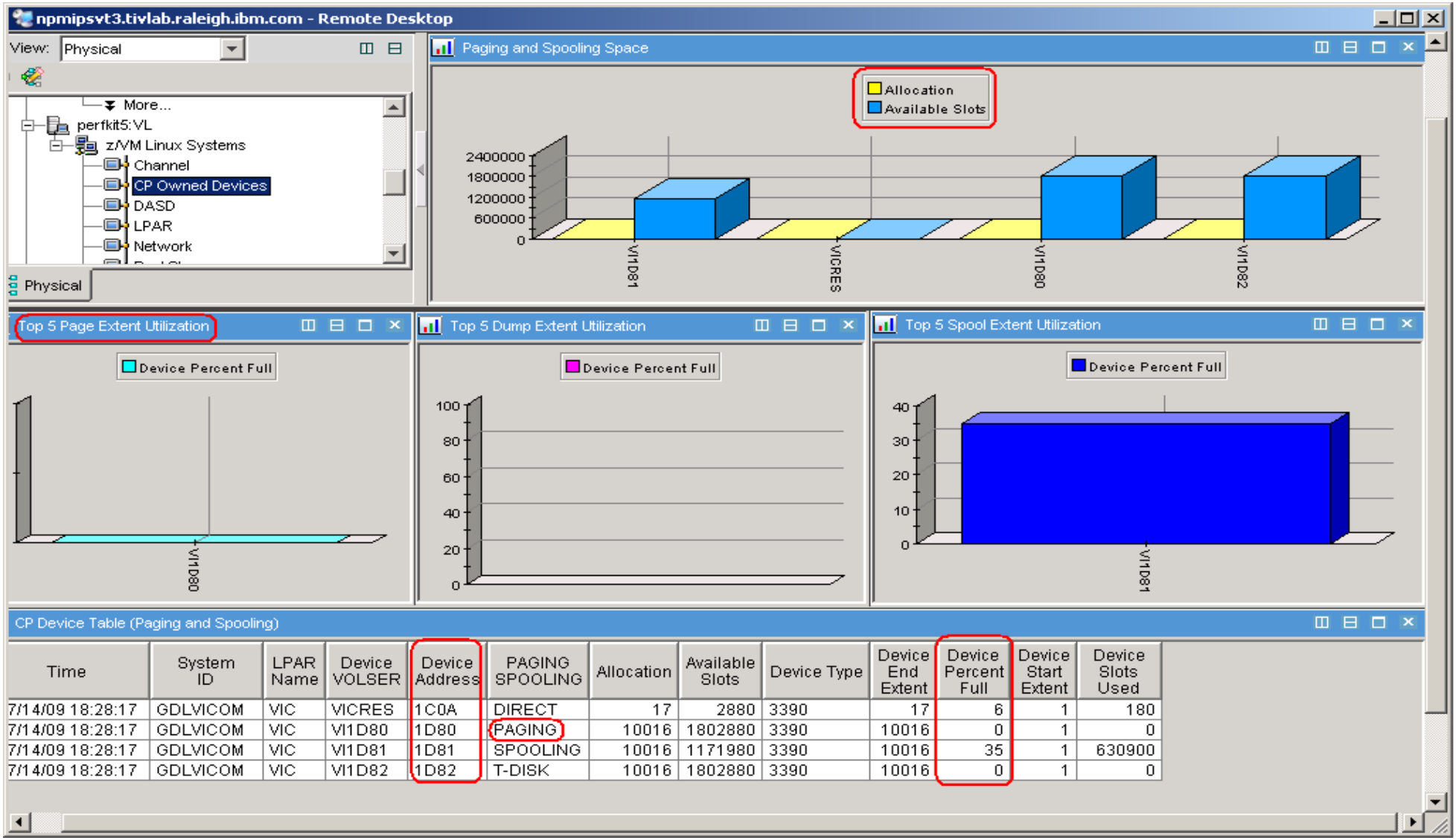
DASD I/O Workspace



System Dump & Spool Space

- **Dump Space**
  - Ensure there is sufficient dump space defined to the system.
  - Dump space requirements vary according to memory usage.
    - Q DUMP – identifies allocated dump space.
    - Calculation guidelines are located in CP Planning and Administration Manual.
- **Spool Space**
  - Various uses:
    - User printer, punch, reader files (console logs)
    - DCSS, NSS
    - System files
    - Page space overflow
  - Spool Management:
    - Monitor with Q ALLOC SPOOL command, OMEGAMON XE or Operations Manager for z/VM
    - command.
    - SFPURGER utility:
      - Rule based tool to clean up spool space.
      - Included in the no charge CMS Utilities Feature (CUF).

System Dump & Spool Space

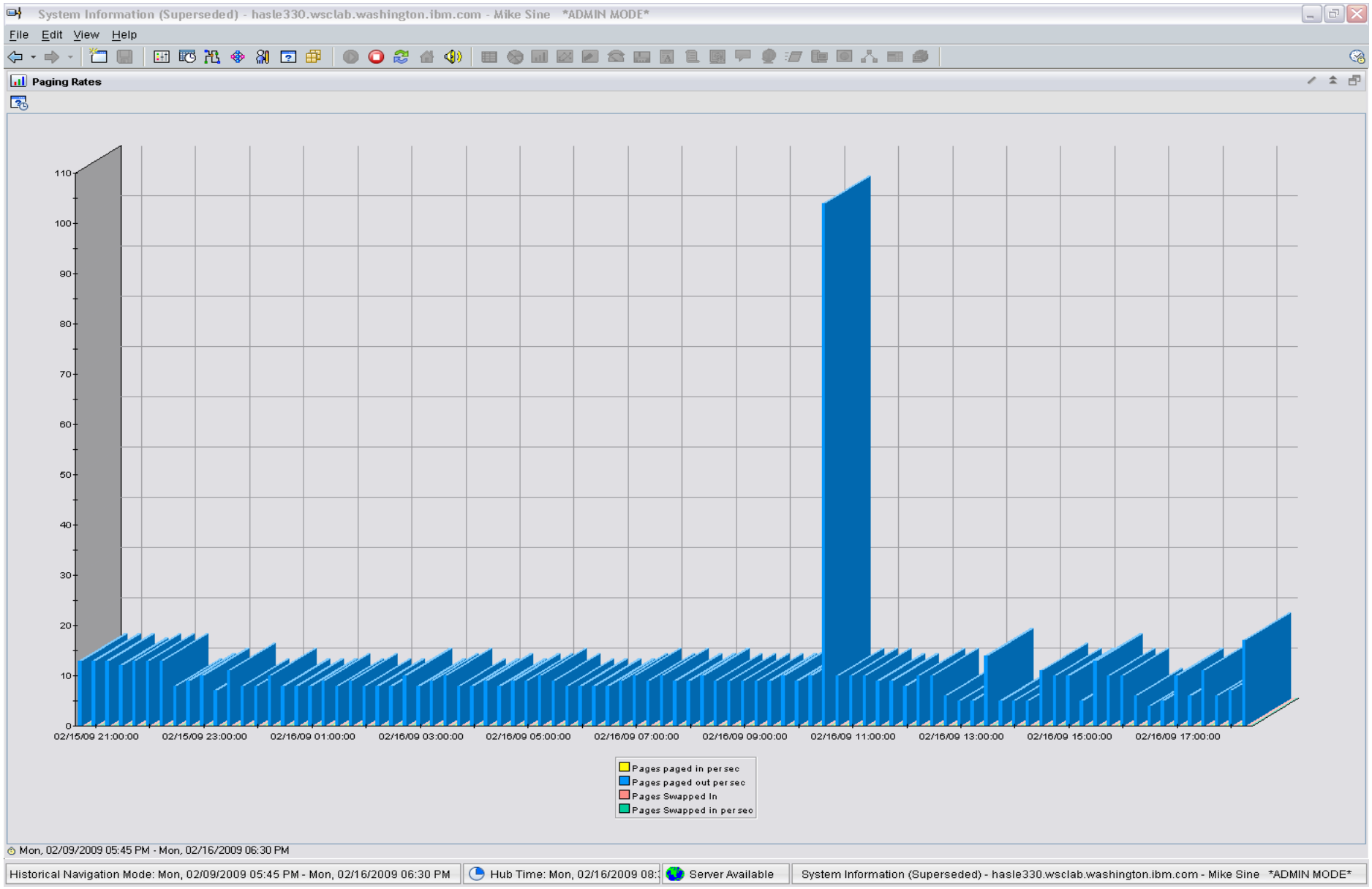


Time	System ID	LPAR Name	Device VOLSER	Device Address	PAGING SPOOLING	Allocation	Available Slots	Device Type	Device End Extent	Device Percent Full	Device Start Extent	Device Slots Used
7/14/09 18:28:17	GDLVICOM	VIC	VICRES	1C0A	DIRECT	17	2880	3390	17	6	1	180
7/14/09 18:28:17	GDLVICOM	VIC	VI1D80	1D80	PAGING	10016	1802880	3390	10016	0	1	0
7/14/09 18:28:17	GDLVICOM	VIC	VI1D81	1D81	SPOOLING	10016	1171980	3390	10016	35	1	630900
7/14/09 18:28:17	GDLVICOM	VIC	VI1D82	1D82	T-DISK	10016	1802880	3390	10016	0	1	0

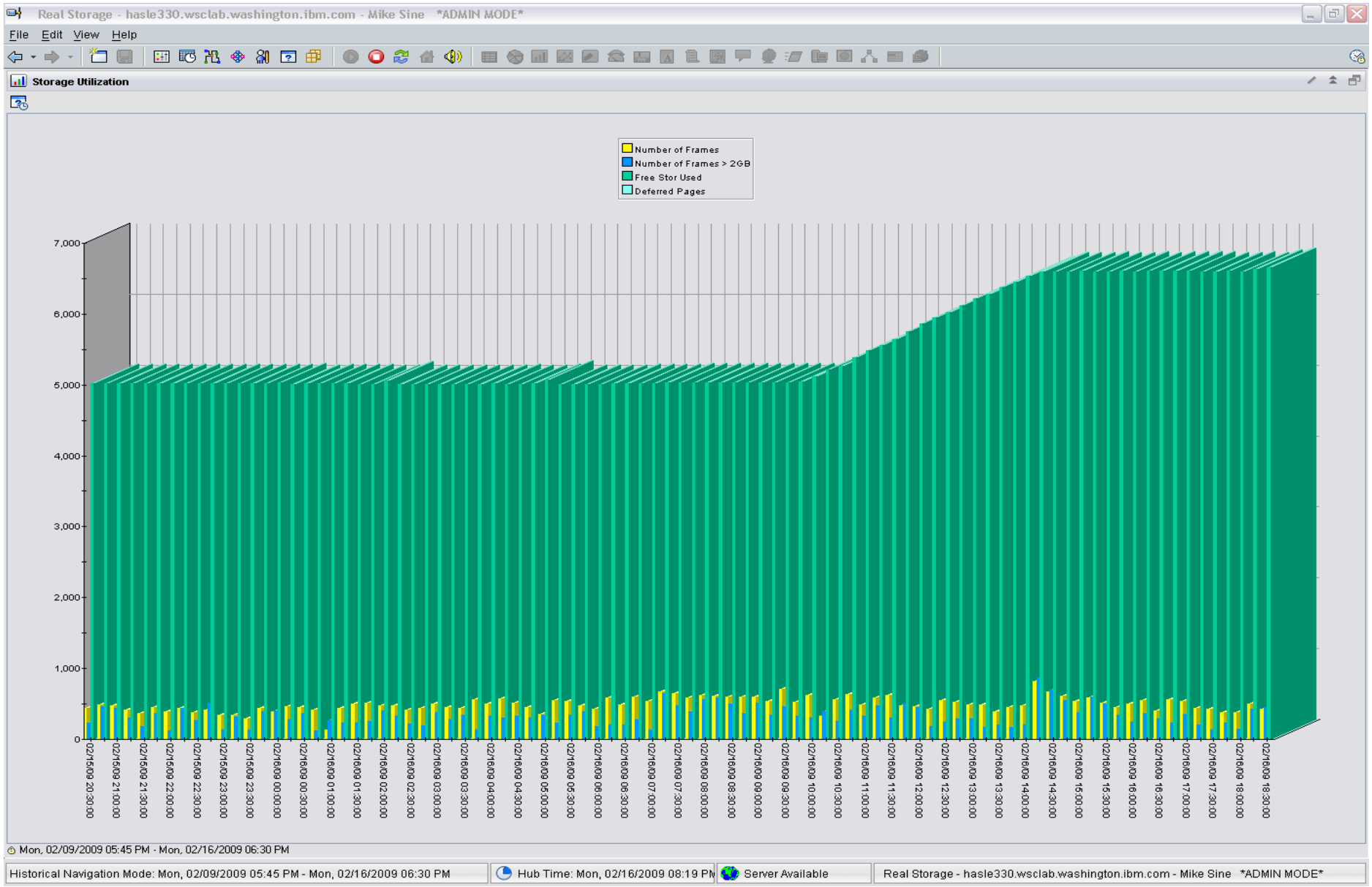
Persistent Historical Views

- This makes it easier to see anomalies, or match spikes.
- Capturing performance data as a base line is a must:
  - General history data – business as usual.
  - Detailed raw monitor data prior to and following any major changes.
  - Ability to review attributes of a past incident.

Persistent Historical Views



Persistent Historical Views



OMEGAMON XE on z/VM and Linux
An Integrated Monitoring Approach

- **Provides performance monitoring for z/VM and Linux guests**
- **Linux agents gather performance data from Linux guests**
- **z/VM agent gathers performance data from z/VM**
  - Including z/VM view of guests
  - Uses IBM Performance Toolkit for VM as its data source
  - Linux provides APPLDATA to CP monitor
- **Executes automated actions in response to defined events or situations**
- **Part of the Tivoli Management Services infrastructure and OMEGAMON family of products**
  - Specifically focused on z/VM and Linux guests

Available Performance Metrics

z/VM

- z/VM Linux Default Workspace
- PAGING and SPOOLING Utilization
- DASD
- LPAR Utilization
- Processors
- NETWORK Utilization (Hipersocket and Virtual Switch)
- REAL STORAGE Utilization
- TCPIP Utilization – Server
- TCPIP Utilization - Users
- SYSTEM Utilization
- System Terminal Workspace
- Workload (z/VM User ID) Activity
- Linux Workload Workspace
- ApplData Workspace
- Channels
- Minidisk Cache
- CCW Translation
- DASD Cache
- Control Unit Cache
- Spin Locks
- Virtual Disks
- Resource Constraint Analysys

Linux

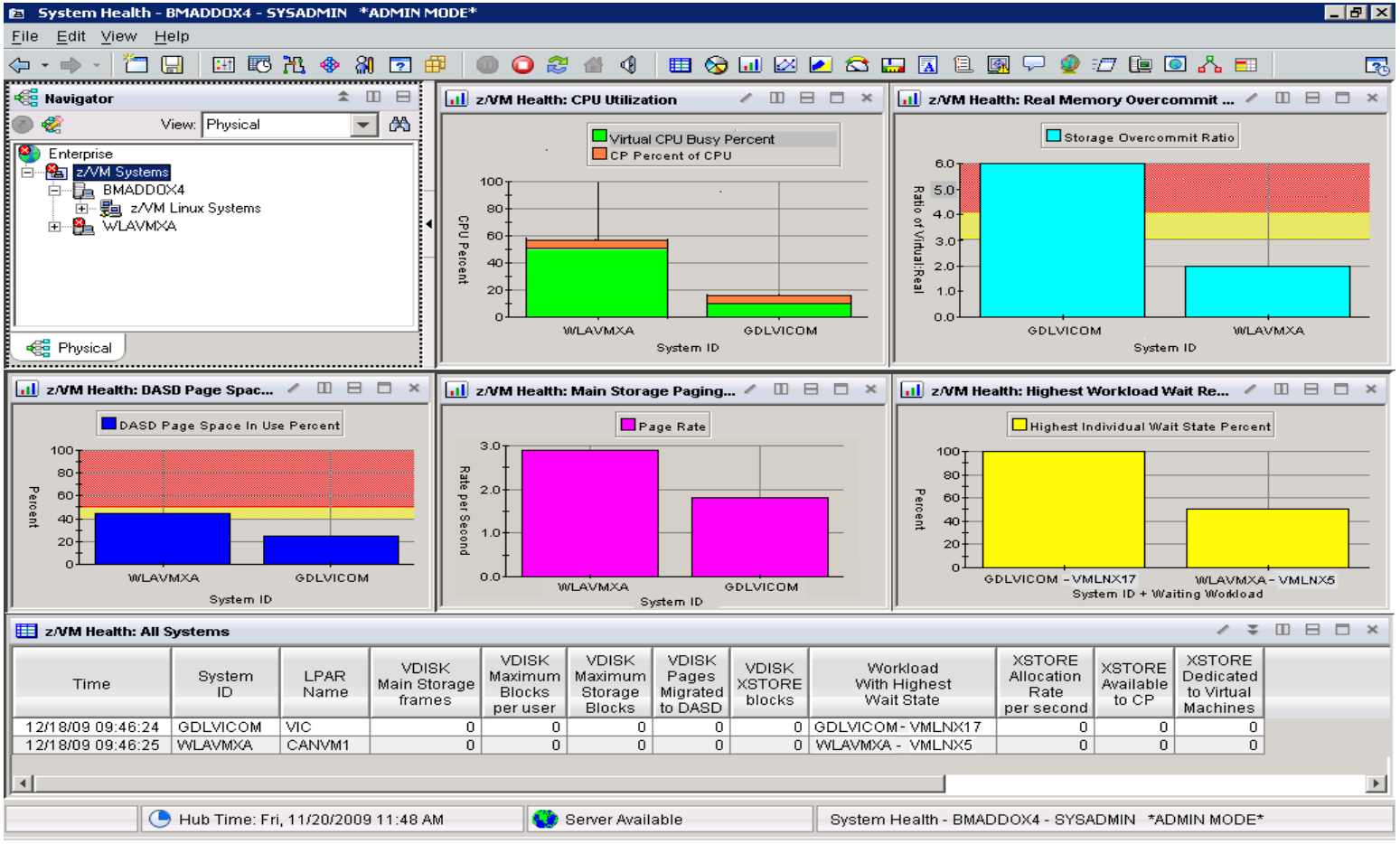
- Linux OS
- Capacity Usage
- Disk Usage
- File Information
- Network
- Process
- System Information
- Users

Use Case Scenarios

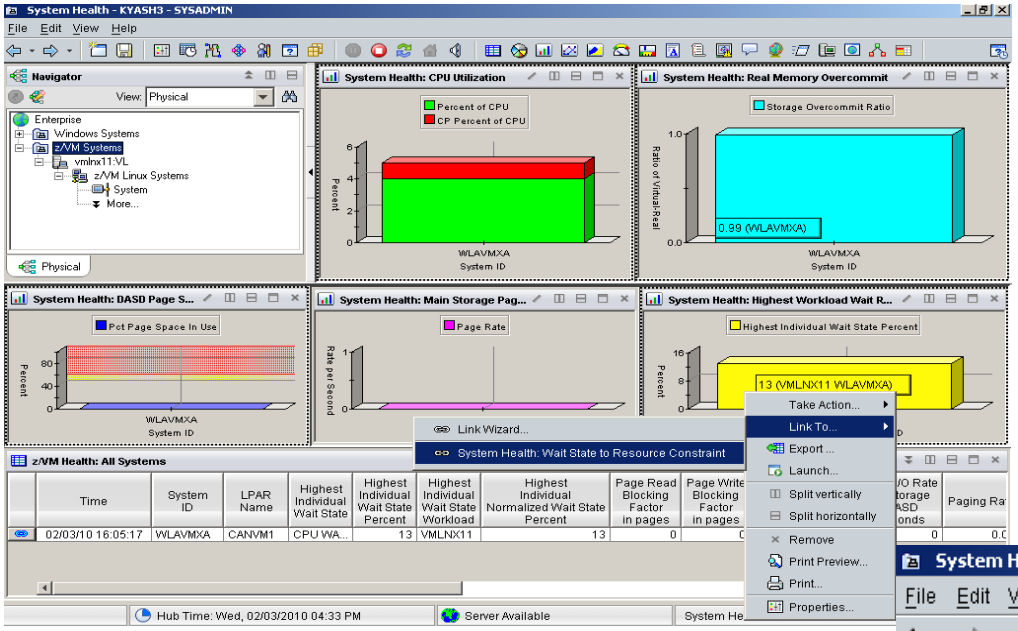
- **Overall health of your z/VM systems**
- **Adding Additional Linux Servers**
- **System running slowly**

Scenario 1— Overall Health of Your System

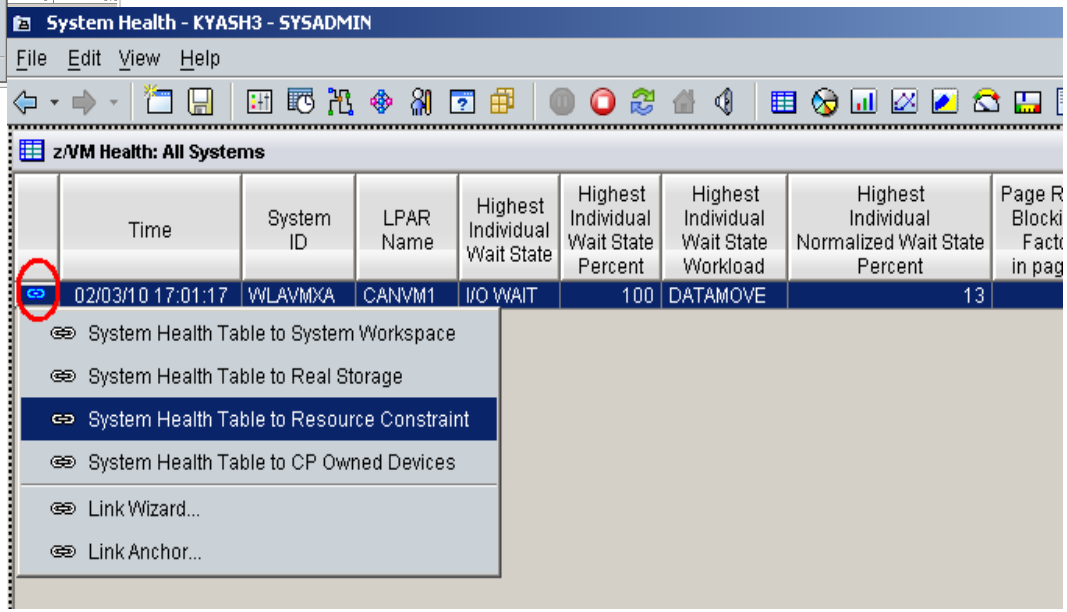
At a quick glance you can see the % CPU usage, ratio of real to virtual memory ratio, paging space, paging rates highest wait state, and VDISK usage for all your z/VM systems



Scenario 1— Overall Health of Your System

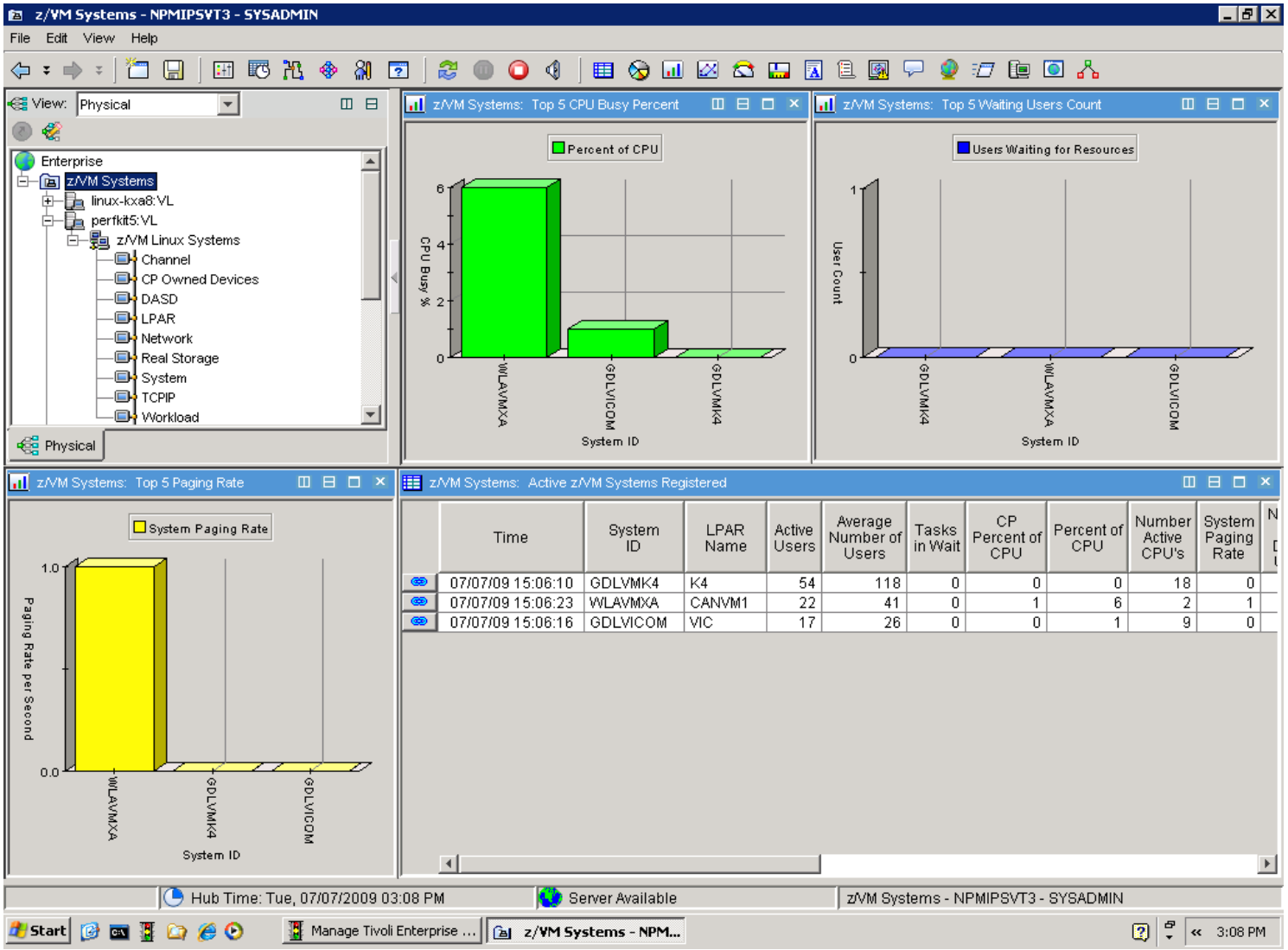


Links are available to drill down for further detailed data



Scenario 1 — Overall Health of Your System

By following the link to the System workspace, you can see at a quick glance the %CPU usage, number of users in a wait state, and paging rates of all your z/VM systems



Scenario 1— Overall Health of Your System

- **Things to look for**
  - CPU usage
    - Is any one system using more CPU than expected
    - Is any one system using less CPU than expected—you may have an underutilized processor and be wasting capacity
    - Remember, a DEDICATED processor will show 100%
  - Users waiting for resources
    - Number of users at the end of the monitoring interval who are either in:
      - Eligible list—waiting to enter the dispatch list
      - Nondispatchable
        - Waiting for paging
        - Waiting for I/O completion
      - Dispatchable
        - Waiting for a processor

Scenario 1— Overall Health of Your System

- **Things to look for**
  - System paging rate
    - Number of page reads per second
    - Not a complete indicator of your paging effectiveness, but a good first glance
      - If the rate is low, and you don't have many users waiting or paging to complete (dispatch list), then you don't have a problem
      - If rate is low and you DO have many users in dispatch list, it may be an indication of a paging problem.
        - High dispatch list number could be for other reasons such as I/O contention. You need to check.
    - If the rate is high, then you may need to tune your paging subsystem.

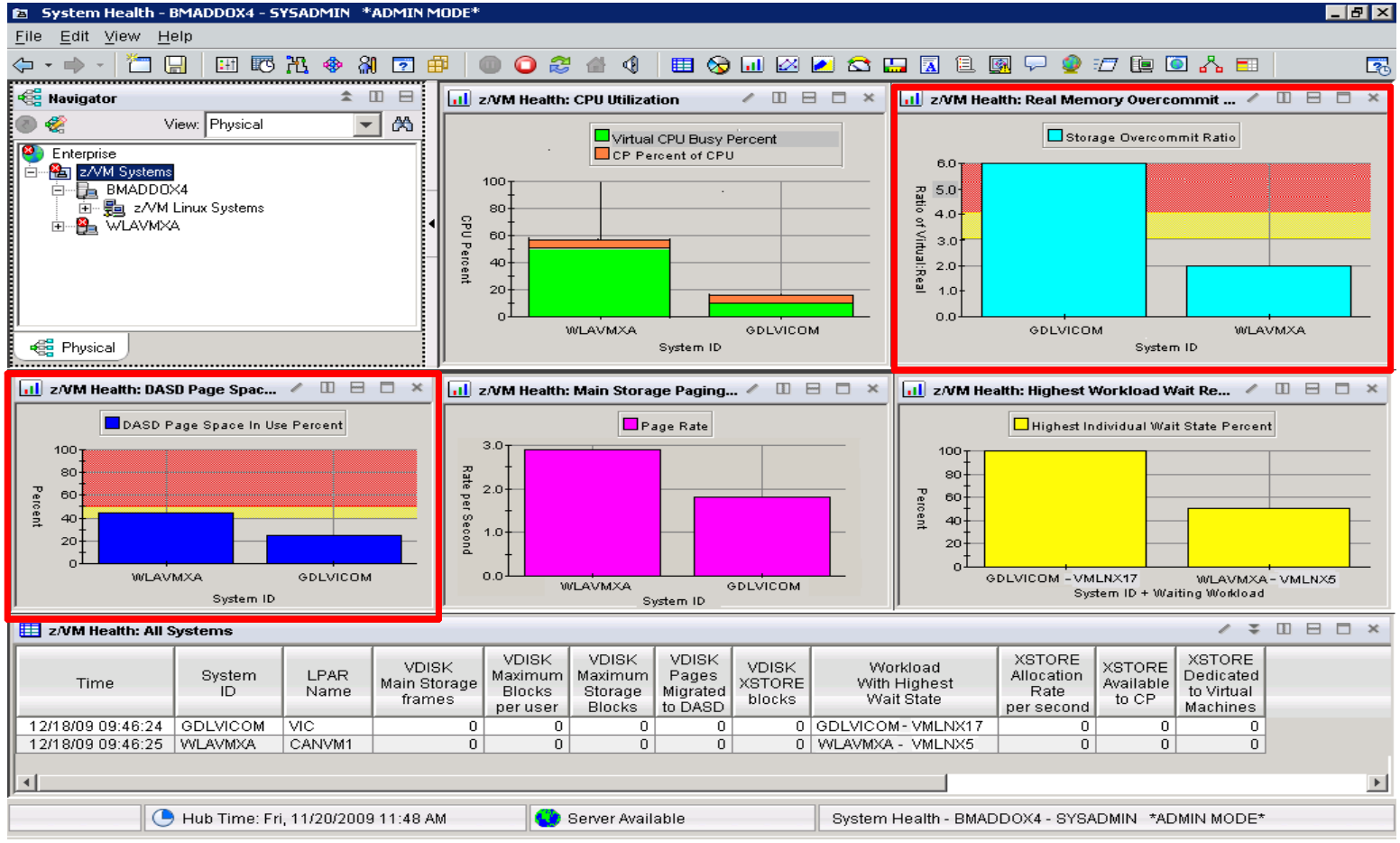
Scenario 2— Adding Additional Linux Servers

Again by using the System Health Workspace at a quick glance you can see ratio of real to virtual memory ratio.

As a rule of thumb you do not want to overcomitt memory greater than 3:1

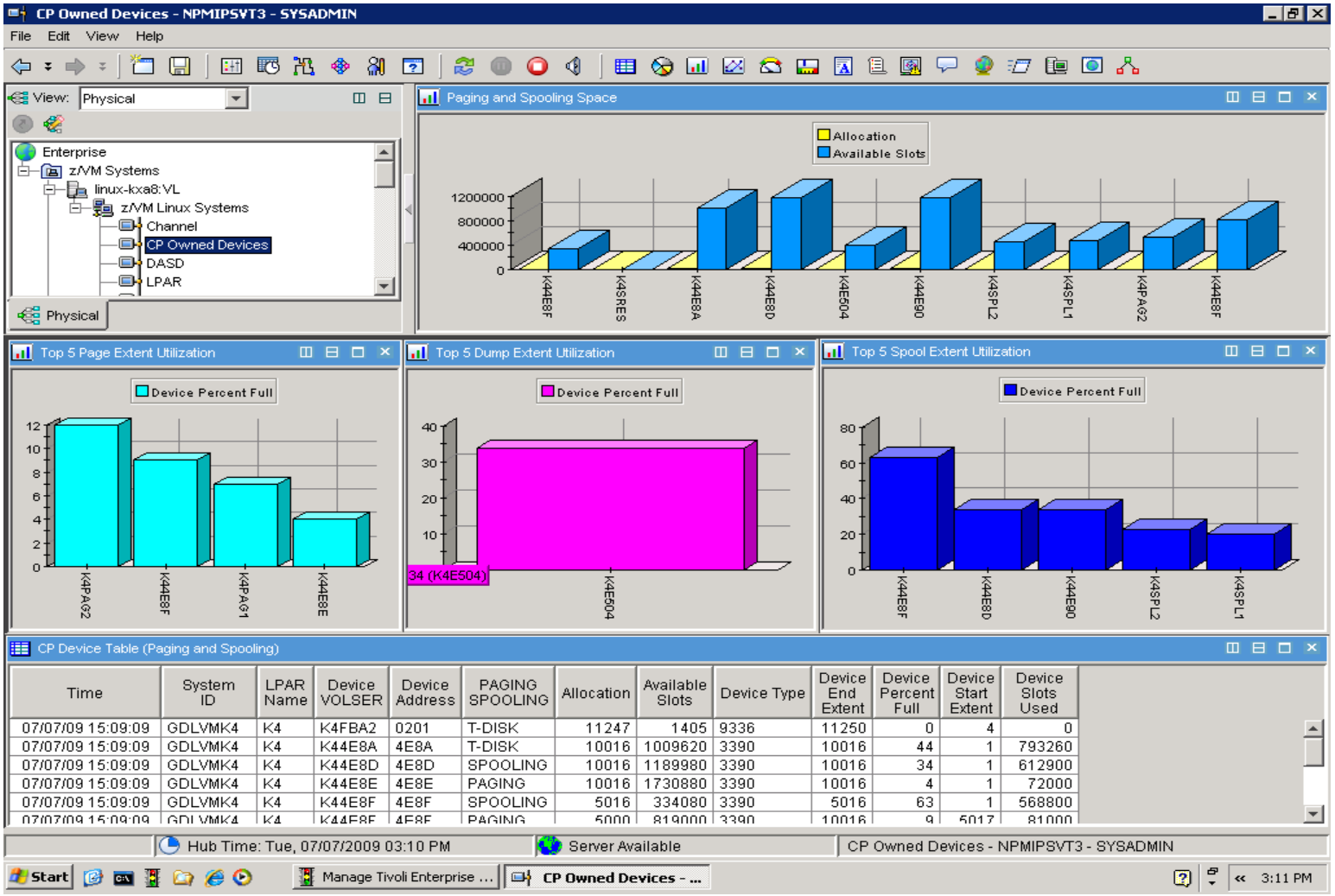
Additional page space is also needed to be added before more workload is added

To better understand the overall Paging Utilization Data, follow the link from the DASD Page Space Utilization view to get additional details on the paging configuration



Scenario 2 — Adding Additional Linux Servers

Using the information in the CP Owned Volumes workspace, one can determine available paging slots, the allocation of existing free space and whether the paging subsystem can handle additional large guests

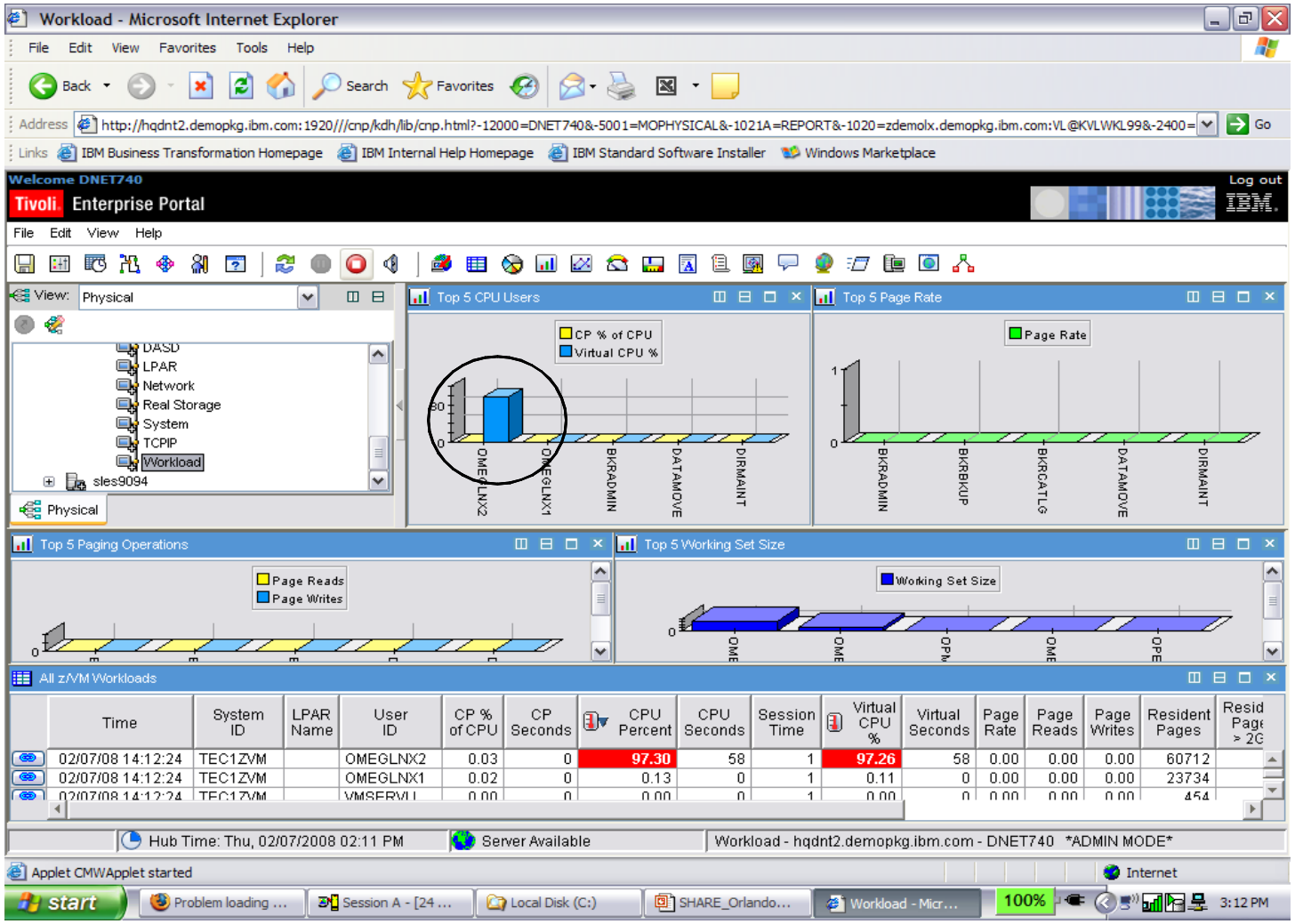


Scenario 2 — Adding Additional Linux Servers

- **General tips**
  - Page space utilization should always be < 50%
  - Never put Paging and Spool space on the same volume
  - Allocate Spool and Page volumes to try and reduce I/O contention by separating them as much as possible (control unit, channel, etc)
  - Dedicated paging devices reduce contention for paging
  - Try to avoid putting highly used files on the same volume as paging and spool space, such as the CMS system disk
  - Use your fastest devices for Paging
  - Multiple Paging devices allow more overlap of paging operations
  - Expanded storage can be used for paging
  - Directory space is not heavily used, can be placed anywhere

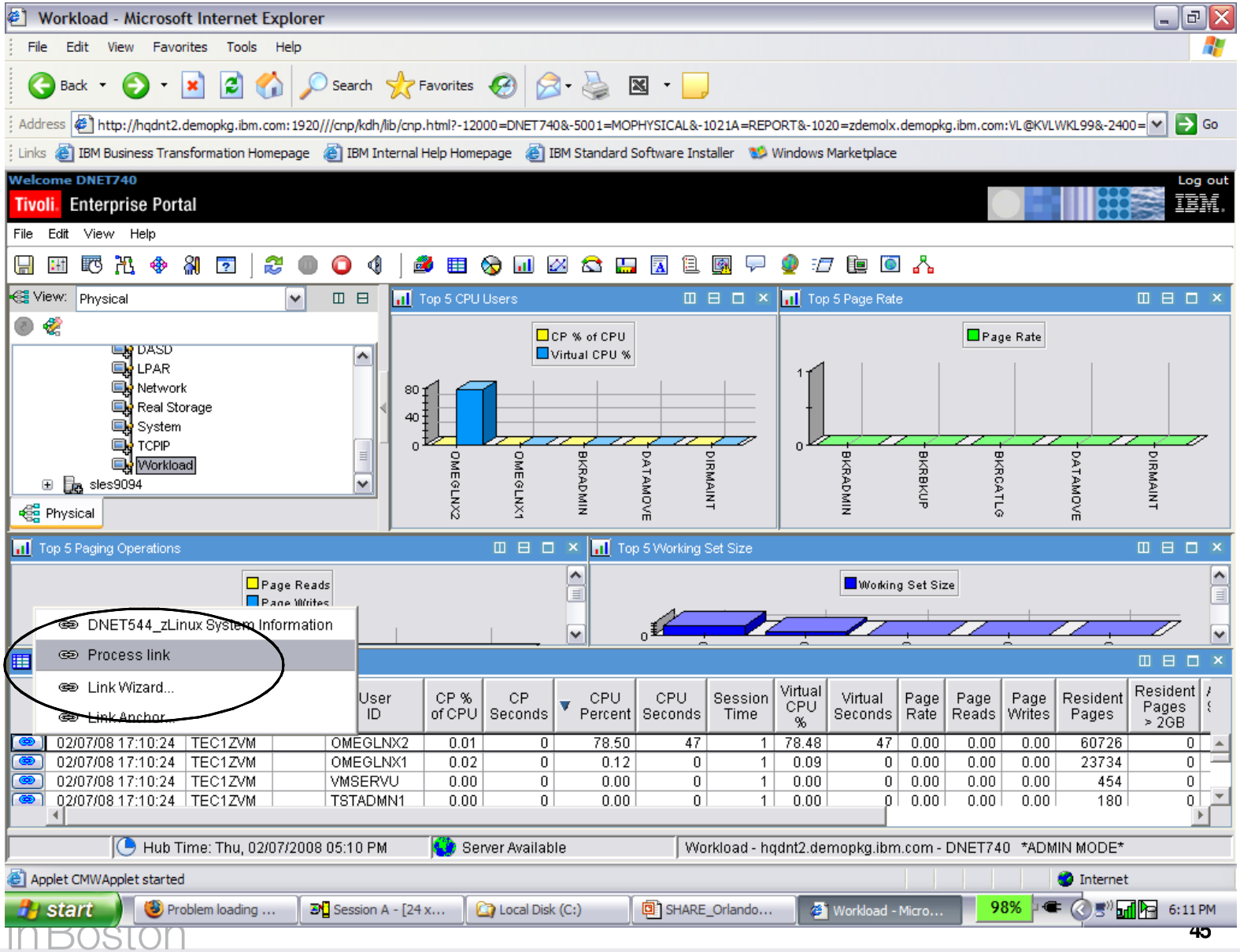
Scenario 3 — System Running Slowly

System is running slowly. Check Workload workspace to see if any particular user is hogging the CPU.



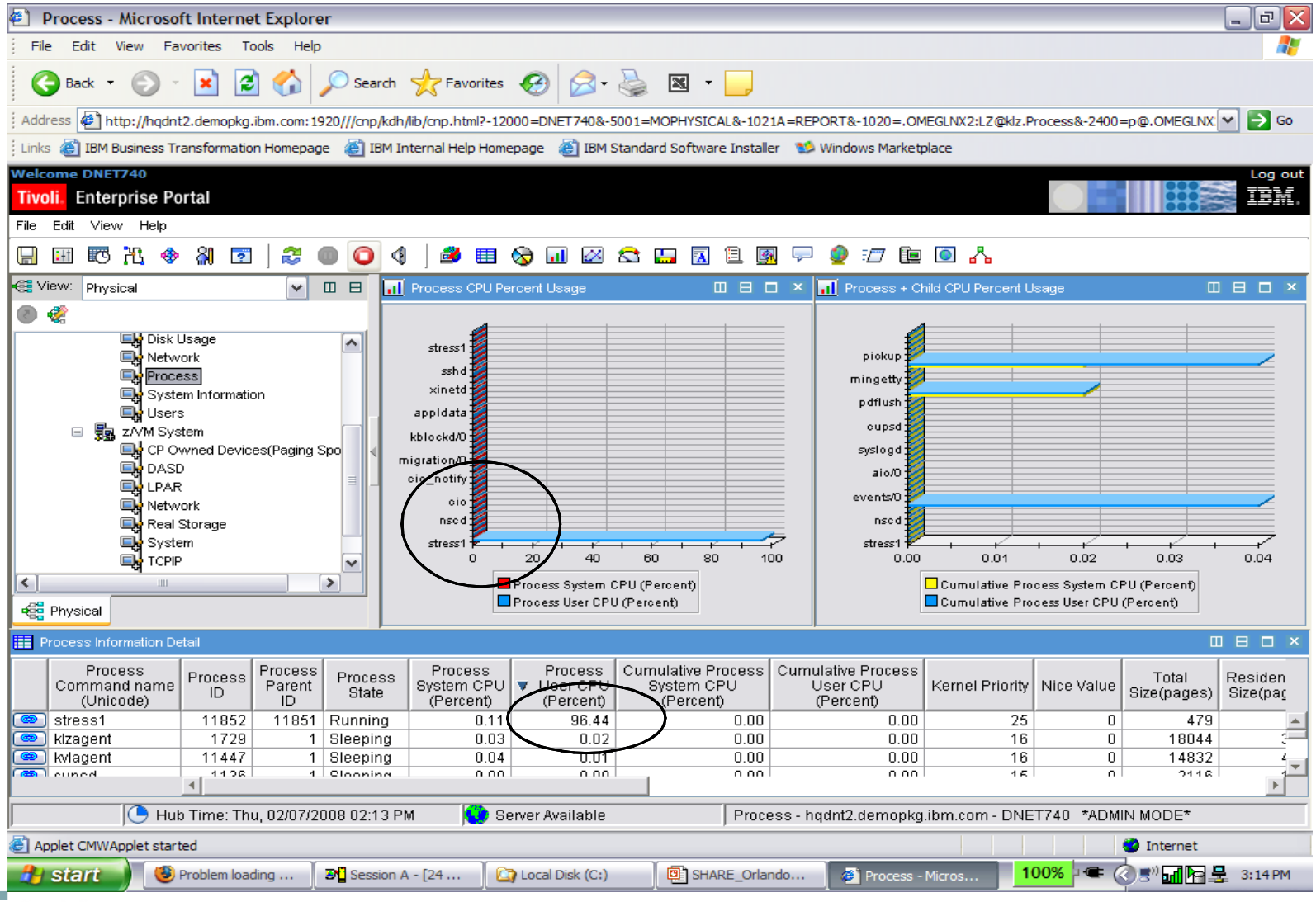
Scenario 3 — System Running Slowly (cont)

Predefined Link to take You directly To the Process workspace



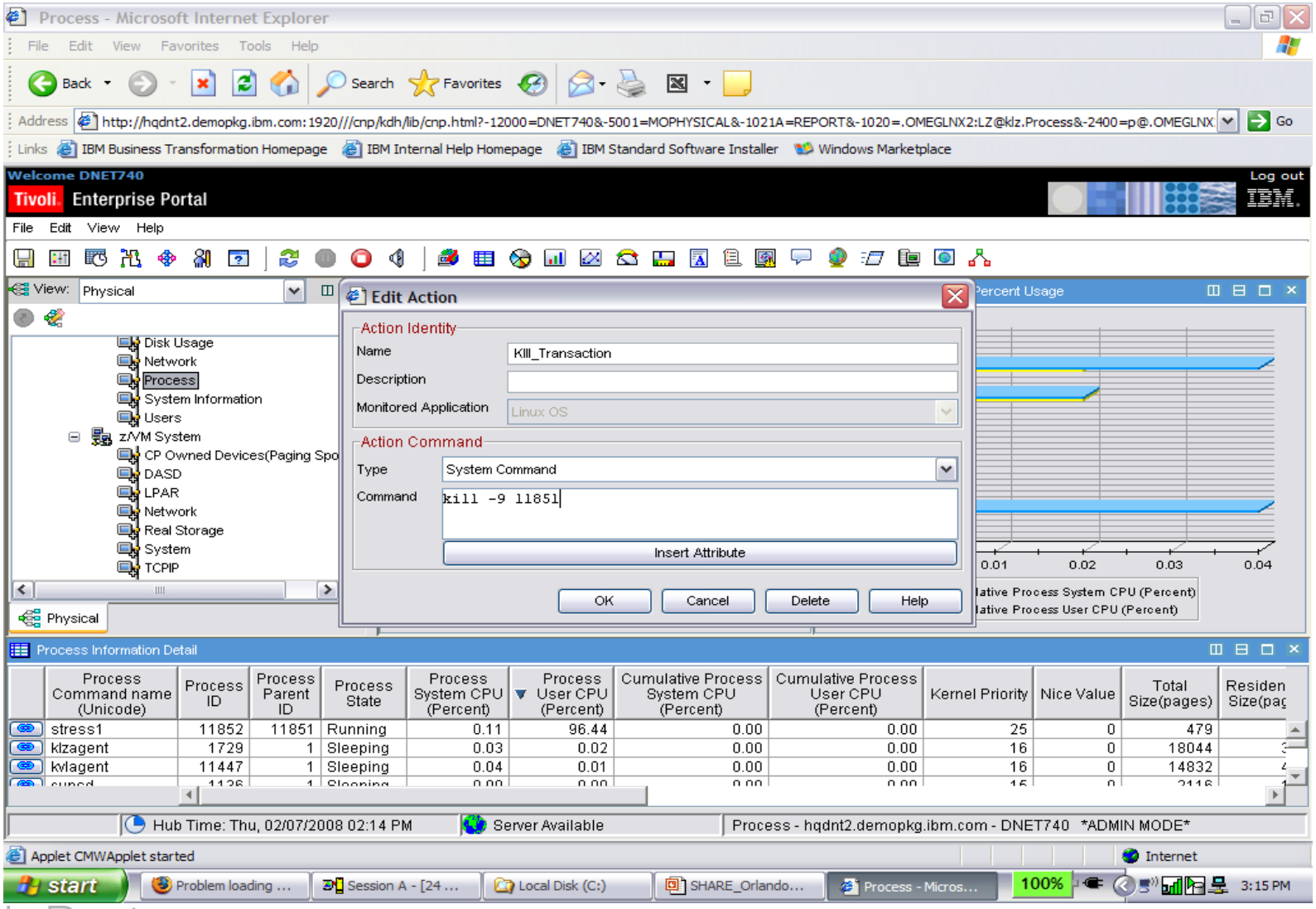
Scenario 3 — System Running Slowly (cont)

See if there is a process which is using too much CPU



Scenario 3 — System Running Slowly (cont)

You can issue a Take Action command to stop the offending process

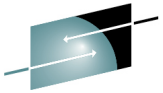


Conclusion

- This presentation has highlighted the best practices for performance and availability management in managing z/VM and Linux on System z. To maximize the benefits of your shared environment, you must also consider the following factors:
  - Security (IBM RACF®, IBM Tivoli zSecure for RACF z/VM)
  - Directory Maintenance (DIRMAINT)
  - Backup and Recovery (IBM Backup and Restore Manager)
  - Automation (z/VM Operations Manager, System Automation for Multiplatform, System Automation Application Manager)
  - Accounting and Chargeback (Tivoli Usage and Accounting Manager)
  - Real resource management (Tape Manager, OSA/SF)
  - Virtual machine provisioning and management (IBM Systems Director, IBM Tivoli Provisioning Manager, IBM Tivoli Service Automation Manager)

Additional Information

- IBM zAdvisor Article - Performance and Availability Best Practices for Managing z/VM and Linux on z
  - http://www-01.ibm.com/software/tivoli/systemz-advisor/2010-06/performance-and-availability.html
- Performance Considerations for Linux Guests
  - http://www.vm.ibm.com/perf/tips/linuxper.html
- OMEGAMON XE for z/VM and Linux Homepage
  - http://www-01.ibm.com/software/tivoli/products/omegamon-xe-zvm-linux/



धन्यवाद

Hindi

多謝

Traditional Chinese

감사합니다

Korean

Спасибо

Russian

Gracias

Spanish

Thank You

English

Obrigado

Brazilian Portuguese

شكرا

Arabic

Danke

German

多谢

Simplified Chinese

Grazie

Italian

Merci

French

நன்றி

Tamil

ありがとうございました

Japanese

ขอบคุณ

Thai