ESX Workload Analysis: Lessons Learned

"Adjusting the Fit"

John Paul

Session ADC9398

VMWORLD 2006

Acknowledgements

- With special thanks and contributions from:
  - **Greg McKnight**, IBM Distinguished Engineer
    IBM Systems and Technology Group
  - **Victor Barra, Christopher Hayes, Doug Tapscott**
    Siemens Medical Solutions Health Services
  - And a few others who have asked to remain anonymous but have been invaluable by providing input to and reviewing this presentation

Presentation Focus

- Initial Lessons Learned
- The Five Contexts of Virtualization
- Basic Workload Analysis Approach
- Basic Performance Analysis Approach
- The Connection Between Workload and Performance Analysis
- Top Performance Counters per Context and Where to Find Them
- Server Processing Tiers
- Reasonability Checks for Each Tier
- How to Drill Down and Up
- Core Four Overviews, Key Counters, Bottlenecks
- Case Study
- Final Lessons Learned

Hold on to your chairs, we are have a lot to cover!

Initial Lessons Learned…

- Almost all workloads can be virtualized
- There will ALWAYS be host-busting workloads
- Most of your efforts will be geared towards reactive performance analysis
- You won't be able to find and resolve all performance problems
- Workload/performance analysis and capacity planning are more arts than sciences
- Most virtualization naysayers eventually turn into virtualization advocates
- The I/O and network subsystems need to be designed and baselined BEFORE loading
- Growth will happen faster then you think
- Plan for the norm, react to the exceptions
- Virtualization is all about resource consumption
- There will be workloads that outgrow your current capacity in your complex
- ALL of us are in the learning curve

Five Contexts of Virtualization

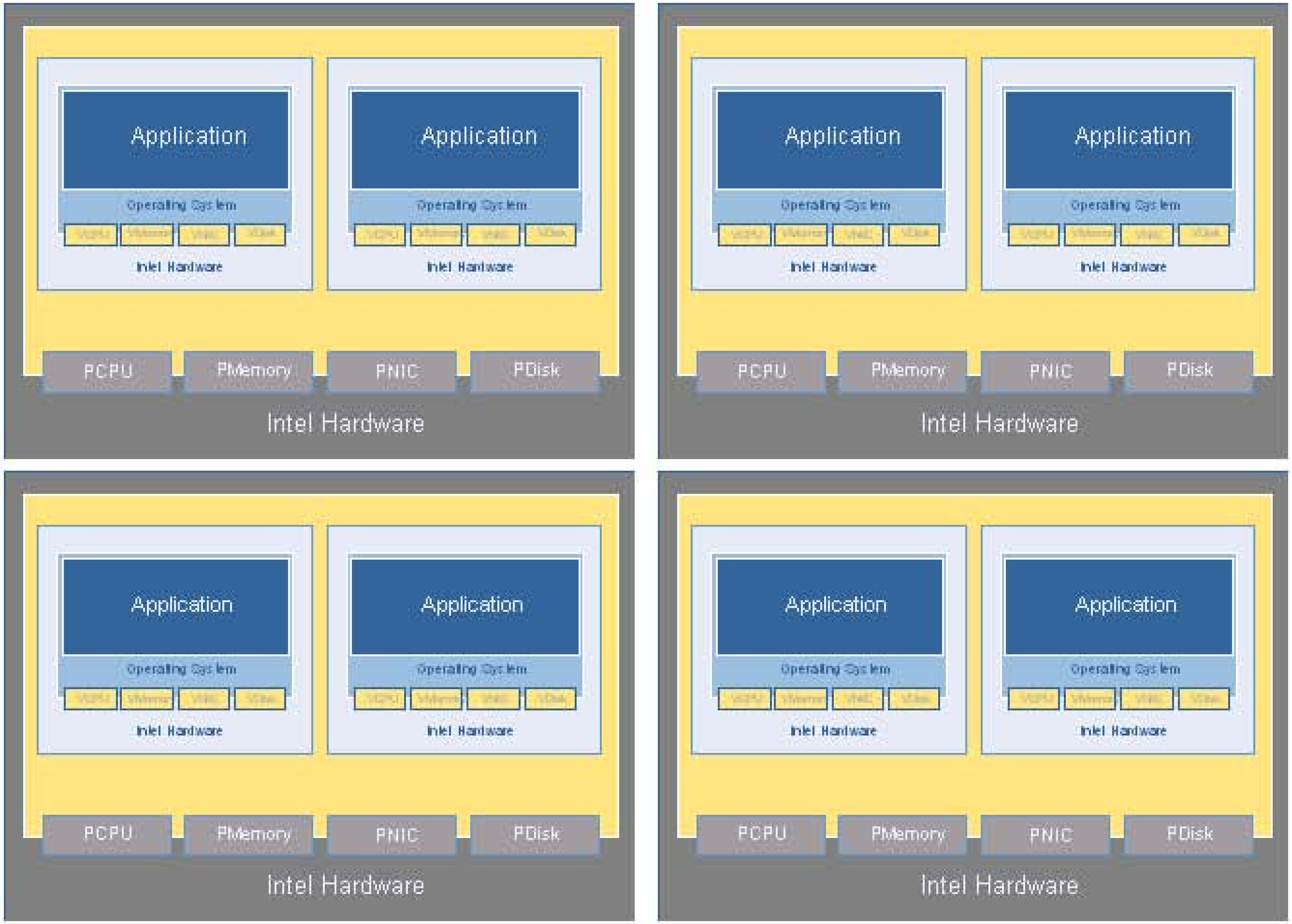
Physical Machine

Virtual Machine

ESX Host Machine

ESX Host Farm/Cluster

ESX Host Complex



Remember the virtual context

Reviewing the Basic Workload Analysis Approach

- **Categorize** the workloads from architectural, performance, and availability perspectives
- **Measure/calculate** your current capacity for the five contexts of virtualization
- **Identify** and **correct** current problems – don't virtualize known problems!
- **Determine** what are and are not ***currently*** good candidates for virtualization at the previously established financially viable target ratio
- **Design** a tiered VMware ESX infrastructure, tiers will evolve if not designed first
- **Apply** reasonability checks to your design before implementation
- **Build** your tiered VMware ESX infrastructure and expect **rapid** expansion
- **Load** your virtual containers aiming initially for a low virtual machine/server ratio
  - From existing servers using new installs, whenever possible
    - Initial resource allocation based upon physical consumption
  - With new workloads
    - Initial resource allocation based upon categorization of server
- **Measure** performance of ESX workloads, ***including*** a trend analysis

This approach has proven to be very effective!

Establishing the Basic Performance Analysis Approach

- **Identify** the virtual context of the reported performance problem
- **Monitor** the performance within that virtual context for an **overview**
  - Start with the overall health of the farm/complex, looking for atypical resource consumers (individual virtual machines)
  - Analyze those virtual machines
  - Identify processes using the largest amount of the Core Four resources
  - Apply a reasonability check on the resources consumed – "Is the amount of resources consumed characteristic of this particular application or task for the server processing tier?"
  - Look for repeat offenders! This happens often.
- **Expand** the performance monitoring to each virtual context as needed
  - Are other workloads influencing the virtual context of this particular application and causing a shortage of a particular resource?
- **Drill down or up** if the higher level diagnostics cannot identify the problem
- **Remedy** the problem
  - Correct the application configuration
  - Adjust the resources assigned to the virtual context
  - Remove the infrastructure problem which is degrading this virtual context

Use "rules of thumb" instead of absolutes

The connection between Workload Analysis and Performance Analysis….

- Workload Analysis…
  - In the virtual machine context is focused on the applications' and resultant virtual machine's resource consumption
  - In the ESX host context is primarily focused on the cumulative amount of resources consumed by all of the virtual machines on a single ESX host
  - In the farm and complex context is usually about infrastructure (SAN and network) resource consumption
- Performance Analysis…
  - In the virtual machine context is usually about virtual machine application configuration and assigning sufficient resources
  - In the ESX host context usually about having enough resources in the physical system to handle that particular workload
  - In the farm and complex context is usually about infrastructure and total capacity

Top Performance Counters per Context

Physical/Virtual Machine

CPU

- **Average physical CPU utilization****
- **Peak physical CPU utilization****
- CPU Time**
- **Processor Queue Length**

Memory

- **Average Memory Usage**
- **Peak Memory Usage**
- Page Faults
- Page Fault Delta

Disk

- I/O Reads
- I/O Writes
- I/O Read Bytes
- I/O Write Bytes
- **Split IO/Sec**
- **Disk Read Queue Length**
- **Disk Write Queue Length**
- **Average Disk Sector Transfer Time**

Network

- **Bytes Total/second**
- **Total Packets/second**
- Bytes Received/second
- Bytes Sent/Second
- Output queue length

ESX Host

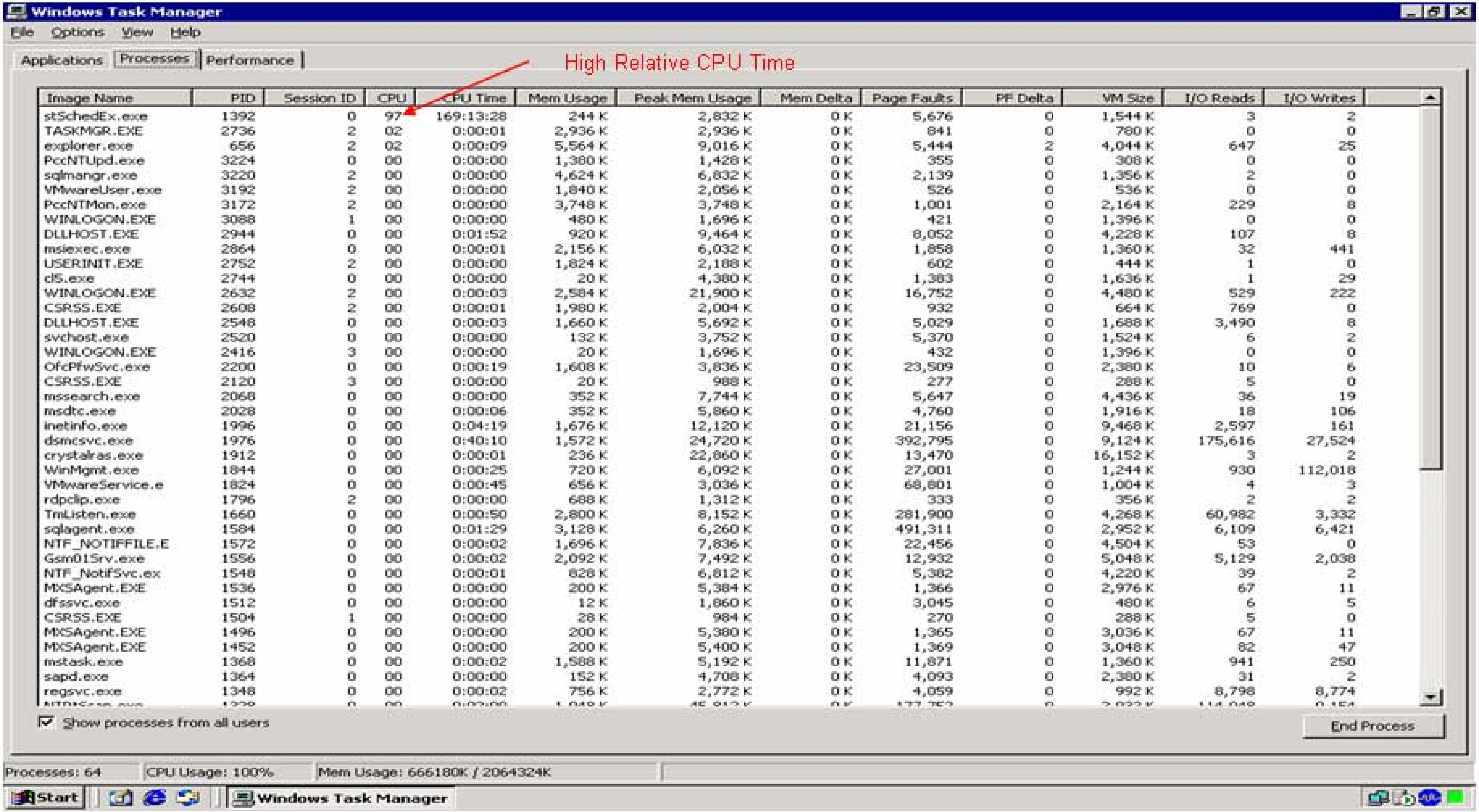
- Average physical CPU utilization
- Peak physical CPU utilization
- **Physical CPU load average**
- Logical CPU utilization
- CPU Effective Use
- Memory Usage
- Disk Reads/second
- Disk Writes/second
- NIC MB transmit/second
- NIC MB write/second
- % Used CPU (high consuming VMs)
- **%Ready to Run**
- %System (< 5% total)
- %Wait
- Allocated VM memory
- Active VM memory

Farm/Cluster/Complex

- % CPU utilization
- % Memory Utilization

** Remember that certain counters are "soft" and should be used as general guides only.

Windows Task Manager – Virtual Machine Context



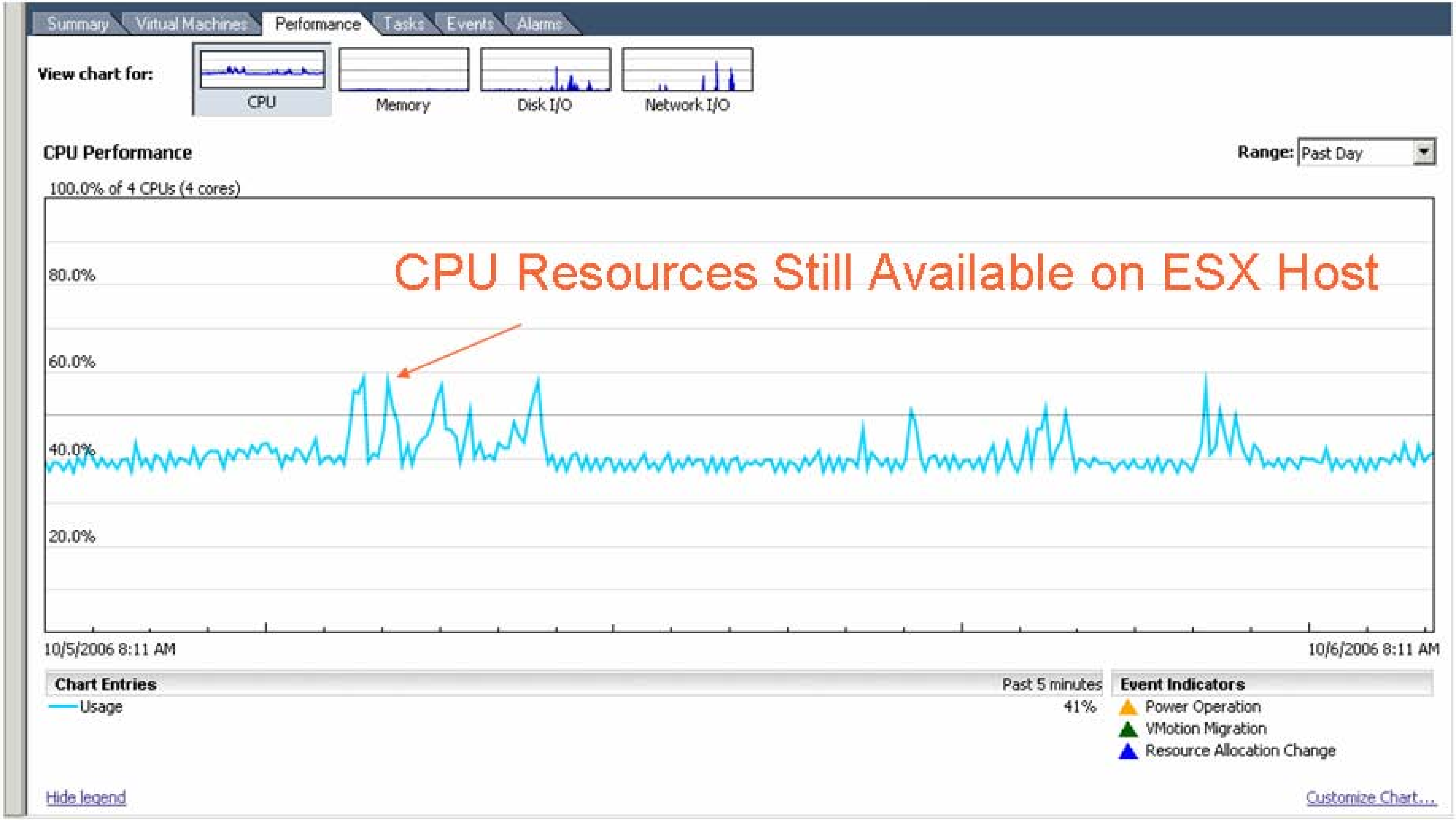
High Relative CPU Time

Image Name	PID	Session ID	CPU	CPU Time	Mem Usage	Peak Mem Usage	Mem Delta	Page Faults	PF Delta	VM Size	I/O Reads	I/O Writes
stSchedEx.exe	1392	0	97	169:13:28	244 K	2,832 K	0 K	5,676	0	1,544 K	3	2
TASKMGR.EXE	2736	2	02	0:00:01	2,936 K	2,936 K	0 K	841	0	780 K	0	0
explorer.exe	656	2	02	0:00:09	5,564 K	9,016 K	0 K	5,444	2	4,044 K	647	25
PccNTUpd.exe	3224	0	00	0:00:00	1,380 K	1,428 K	0 K	355	0	308 K	0	0
sqlmangr.exe	3220	2	00	0:00:00	4,624 K	6,832 K	0 K	2,139	0	1,356 K	2	0
VMwareUser.exe	3192	2	00	0:00:00	1,840 K	2,056 K	0 K	526	0	536 K	0	0
PccNTMon.exe	3172	2	00	0:00:00	3,748 K	3,748 K	0 K	1,001	0	2,164 K	229	8
WINLOGON.EXE	3088	1	00	0:00:00	480 K	1,696 K	0 K	421	0	1,396 K	0	0
DLLHOST.EXE	2944	0	00	0:01:52	920 K	9,464 K	0 K	8,052	0	4,228 K	107	8
msiexec.exe	2864	0	00	0:00:01	2,156 K	6,032 K	0 K	1,858	0	1,360 K	32	441
USERINIT.EXE	2752	2	00	0:00:00	1,824 K	2,188 K	0 K	602	0	444 K	1	0
cl5.exe	2744	0	00	0:00:00	20 K	4,380 K	0 K	1,383	0	1,636 K	1	29
WINLOGON.EXE	2632	2	00	0:00:03	2,584 K	21,900 K	0 K	16,752	0	4,480 K	529	222
CSRSS.EXE	2608	2	00	0:00:01	1,980 K	2,004 K	0 K	932	0	664 K	769	0
DLLHOST.EXE	2548	0	00	0:00:03	1,660 K	5,692 K	0 K	5,029	0	1,688 K	3,490	8
svchost.exe	2520	0	00	0:00:00	132 K	3,752 K	0 K	5,370	0	1,524 K	6	2
WINLOGON.EXE	2416	3	00	0:00:00	20 K	1,696 K	0 K	432	0	1,396 K	0	0
OfcPfwSvc.exe	2200	0	00	0:00:19	1,608 K	3,836 K	0 K	23,509	0	2,380 K	10	6
CSRSS.EXE	2120	3	00	0:00:00	20 K	988 K	0 K	277	0	288 K	5	0
mssearch.exe	2068	0	00	0:00:00	352 K	7,744 K	0 K	5,647	0	4,436 K	36	19
msdtc.exe	2028	0	00	0:00:06	352 K	5,860 K	0 K	4,760	0	1,916 K	18	106
inetinfo.exe	1996	0	00	0:04:19	1,676 K	12,120 K	0 K	21,156	0	9,468 K	2,597	161
dsmcsvc.exe	1976	0	00	0:40:10	1,572 K	24,720 K	0 K	392,795	0	9,124 K	175,616	27,524
crystalras.exe	1912	0	00	0:00:01	236 K	22,860 K	0 K	13,470	0	16,152 K	3	2
WinMgmt.exe	1844	0	00	0:00:25	720 K	6,092 K	0 K	27,001	0	1,244 K	930	112,018
VMwareService.e	1824	0	00	0:00:45	656 K	3,036 K	0 K	68,801	0	1,004 K	4	3
rdpclip.exe	1796	2	00	0:00:00	688 K	1,312 K	0 K	333	0	356 K	2	2
TmListen.exe	1660	0	00	0:00:50	2,800 K	8,152 K	0 K	281,900	0	4,268 K	60,982	3,332
sqlagent.exe	1584	0	00	0:01:29	3,128 K	6,260 K	0 K	491,311	0	2,952 K	6,109	6,421
NTF_NOTIFFILE.E	1572	0	00	0:00:02	1,696 K	7,836 K	0 K	22,456	0	4,504 K	53	0
Gsm01Srv.exe	1556	0	00	0:00:02	2,092 K	7,492 K	0 K	12,932	0	5,048 K	5,129	2,038
NTF_NotifSvc.ex	1548	0	00	0:00:01	828 K	6,812 K	0 K	5,382	0	4,220 K	39	2
MXSAgent.EXE	1536	0	00	0:00:00	200 K	5,384 K	0 K	1,366	0	2,976 K	67	11
dfssvc.exe	1512	0	00	0:00:00	12 K	1,860 K	0 K	3,045	0	480 K	6	5
CSRSS.EXE	1504	1	00	0:00:00	28 K	984 K	0 K	270	0	288 K	5	0
MXSAgent.EXE	1496	0	00	0:00:00	200 K	5,380 K	0 K	1,365	0	3,036 K	67	11
MXSAgent.EXE	1452	0	00	0:00:00	200 K	5,400 K	0 K	1,369	0	3,048 K	82	47
mstask.exe	1368	0	00	0:00:02	1,588 K	5,192 K	0 K	11,871	0	1,360 K	941	250
sapd.exe	1364	0	00	0:00:00	152 K	4,708 K	0 K	4,093	0	2,380 K	31	2
regsvc.exe	1348	0	00	0:00:02	756 K	2,772 K	0 K	4,059	0	992 K	8,798	8,774

☑ Show processes from all users

Processes: 64 CPU Usage: 100% Mem Usage: 666180K / 2064324K

Virtual Center – ESX Host Context



View chart for: CPU

CPU Performance

Range: Past Day

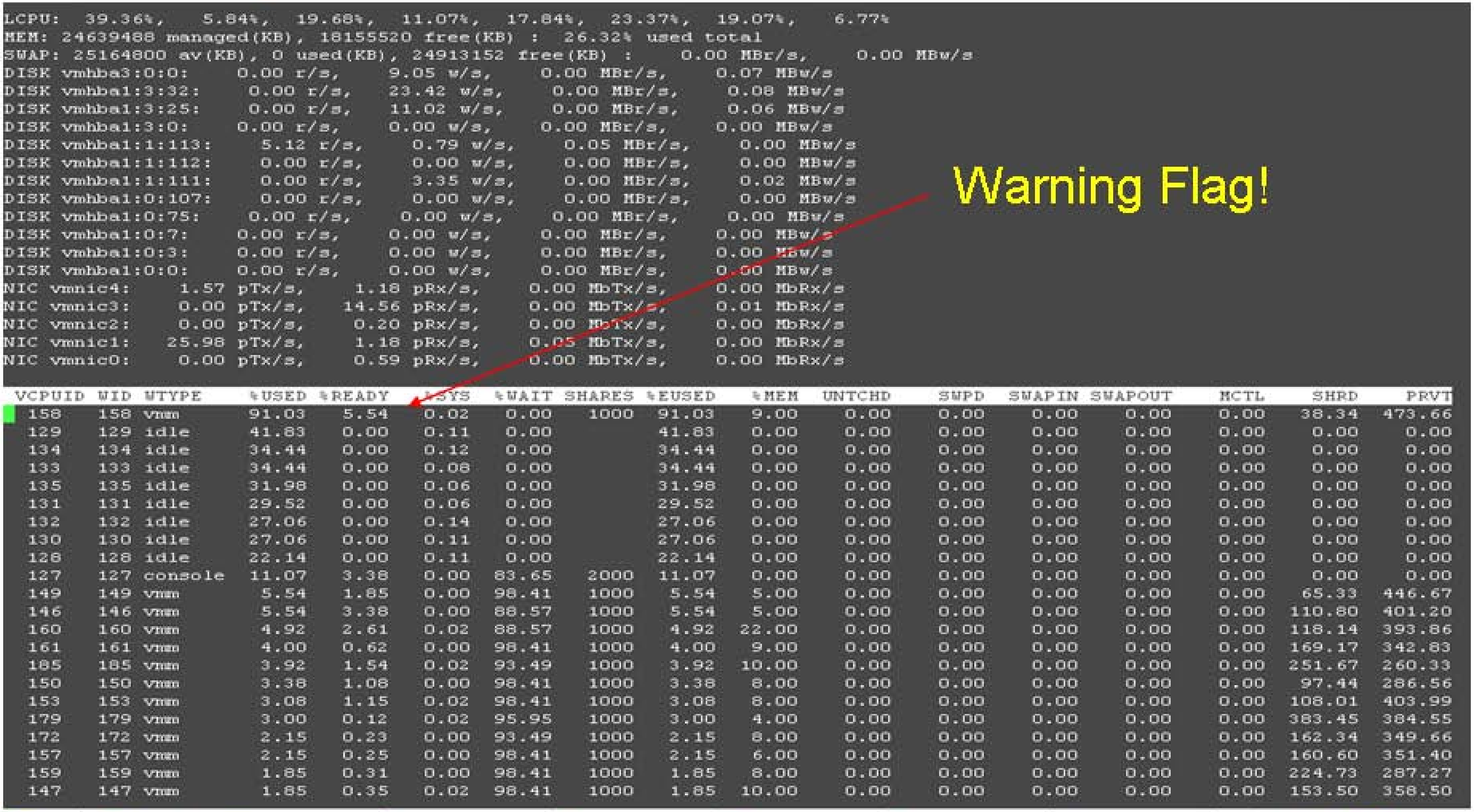
CPU Resources Still Available on ESX Host

Chart Entries	Past 5 minutes
Usage	41%

Event Indicators

- ▲ Power Operation
- ▲ VMotion Migration
- ▲ Resource Allocation Change

ESX Top – ESX Host Context

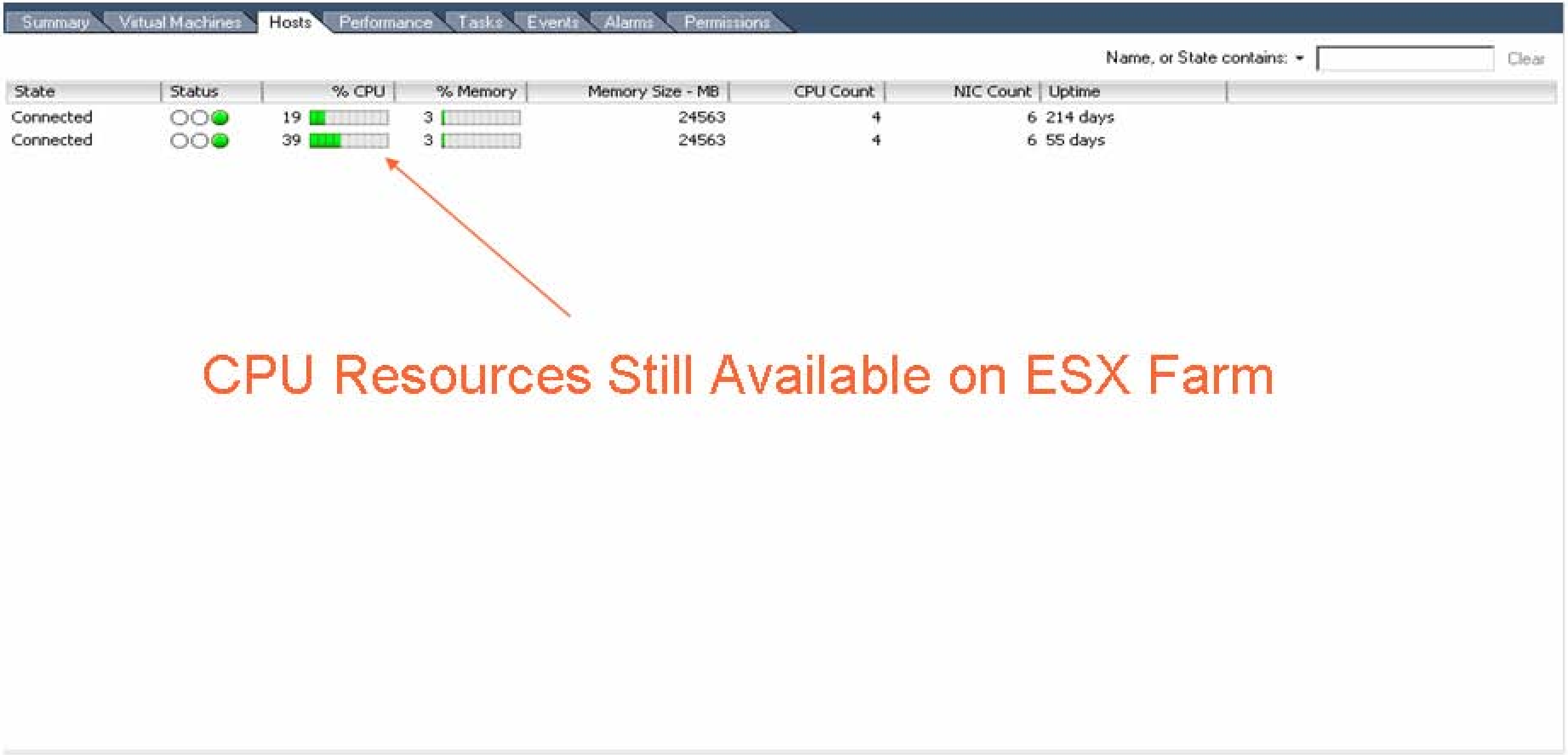


```
LCPU:  39.36%,   5.84%,  19.68%,  11.07%,  17.84%,  23.37%,  19.07%,   6.77%
MEM: 24639488 managed(KB), 18155520 free(KB) :  26.32% used total
SWAP: 25164800 av(KB), 0 used(KB), 24913152 free(KB) :    0.00 MBr/s,    0.00 MBw/s
DISK vmhba3:0:0:    0.00 r/s,    9.05 w/s,    0.00 MBr/s,    0.07 MBw/s
DISK vmhba1:3:32:    0.00 r/s,   23.42 w/s,    0.00 MBr/s,    0.08 MBw/s
DISK vmhba1:3:25:    0.00 r/s,   11.02 w/s,    0.00 MBr/s,    0.06 MBw/s
DISK vmhba1:3:0:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
DISK vmhba1:1:113:    5.12 r/s,    0.79 w/s,    0.05 MBr/s,    0.00 MBw/s
DISK vmhba1:1:112:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
DISK vmhba1:1:111:    0.00 r/s,    3.35 w/s,    0.00 MBr/s,    0.02 MBw/s
DISK vmhba1:0:107:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
DISK vmhba1:0:75:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
DISK vmhba1:0:7:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
DISK vmhba1:0:3:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
DISK vmhba1:0:0:    0.00 r/s,    0.00 w/s,    0.00 MBr/s,    0.00 MBw/s
NIC vmnic4:     1.57 pTx/s,     1.18 pRx/s,    0.00 MbTx/s,    0.00 MbRx/s
NIC vmnic3:     0.00 pTx/s,    14.56 pRx/s,    0.00 MbTx/s,    0.01 MbRx/s
NIC vmnic2:     0.00 pTx/s,     0.20 pRx/s,    0.00 MbTx/s,    0.00 MbRx/s
NIC vmnic1:    25.98 pTx/s,     1.18 pRx/s,    0.05 MbTx/s,    0.00 MbRx/s
NIC vmnic0:     0.00 pTx/s,     0.59 pRx/s,    0.00 MbTx/s,    0.00 MbRx/s

 VCPUID  WID  WTYPE     %USED %READY   %SYS  %WAIT SHARES %EUSED    %MEM  UNTCHD     SWPD  SWAPIN SWAPOUT     MCTL     SHRD     PRVT
   158   158 vmm        91.03   5.54   0.02   0.00   1000  91.03    9.00    0.00     0.00    0.00    0.00     0.00    38.34   473.66
   129   129 idle       41.83   0.00   0.11   0.00         41.83    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   134   134 idle       34.44   0.00   0.12   0.00         34.44    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   133   133 idle       34.44   0.00   0.08   0.00         34.44    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   135   135 idle       31.98   0.00   0.06   0.00         31.98    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   131   131 idle       29.52   0.00   0.06   0.00         29.52    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   132   132 idle       27.06   0.00   0.14   0.00         27.06    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   130   130 idle       27.06   0.00   0.11   0.00         27.06    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   128   128 idle       22.14   0.00   0.11   0.00         22.14    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   127   127 console    11.07   3.38   0.00  83.65   2000  11.07    0.00    0.00     0.00    0.00    0.00     0.00     0.00     0.00
   149   149 vmm         5.54   1.85   0.00  98.41   1000   5.54    5.00    0.00     0.00    0.00    0.00     0.00    65.33   446.67
   146   146 vmm         5.54   3.38   0.00  88.57   1000   5.54    5.00    0.00     0.00    0.00    0.00     0.00   110.80   401.20
   160   160 vmm         4.92   2.61   0.02  88.57   1000   4.92   22.00    0.00     0.00    0.00    0.00     0.00   118.14   393.86
   161   161 vmm         4.00   0.62   0.00  98.41   1000   4.00    9.00    0.00     0.00    0.00    0.00     0.00   169.17   342.83
   185   185 vmm         3.92   1.54   0.02  93.49   1000   3.92   10.00    0.00     0.00    0.00    0.00     0.00   251.67   260.33
   150   150 vmm         3.38   1.08   0.00  98.41   1000   3.38    8.00    0.00     0.00    0.00    0.00     0.00    97.44   286.56
   153   153 vmm         3.08   1.15   0.02  98.41   1000   3.08    8.00    0.00     0.00    0.00    0.00     0.00   108.01   403.99
   179   179 vmm         3.00   0.12   0.02  95.95   1000   3.00    4.00    0.00     0.00    0.00    0.00     0.00   383.45   384.55
   172   172 vmm         2.15   0.23   0.00  93.49   1000   2.15    8.00    0.00     0.00    0.00    0.00     0.00   162.34   349.66
   157   157 vmm         2.15   0.25   0.00  98.41   1000   2.15    6.00    0.00     0.00    0.00    0.00     0.00   160.60   351.40
   159   159 vmm         1.85   0.31   0.00  98.41   1000   1.85    8.00    0.00     0.00    0.00    0.00     0.00   224.73   287.27
   147   147 vmm         1.85   0.35   0.02  98.41   1000   1.85   10.00    0.00     0.00    0.00    0.00     0.00   153.50   358.50
```

Warning Flag!

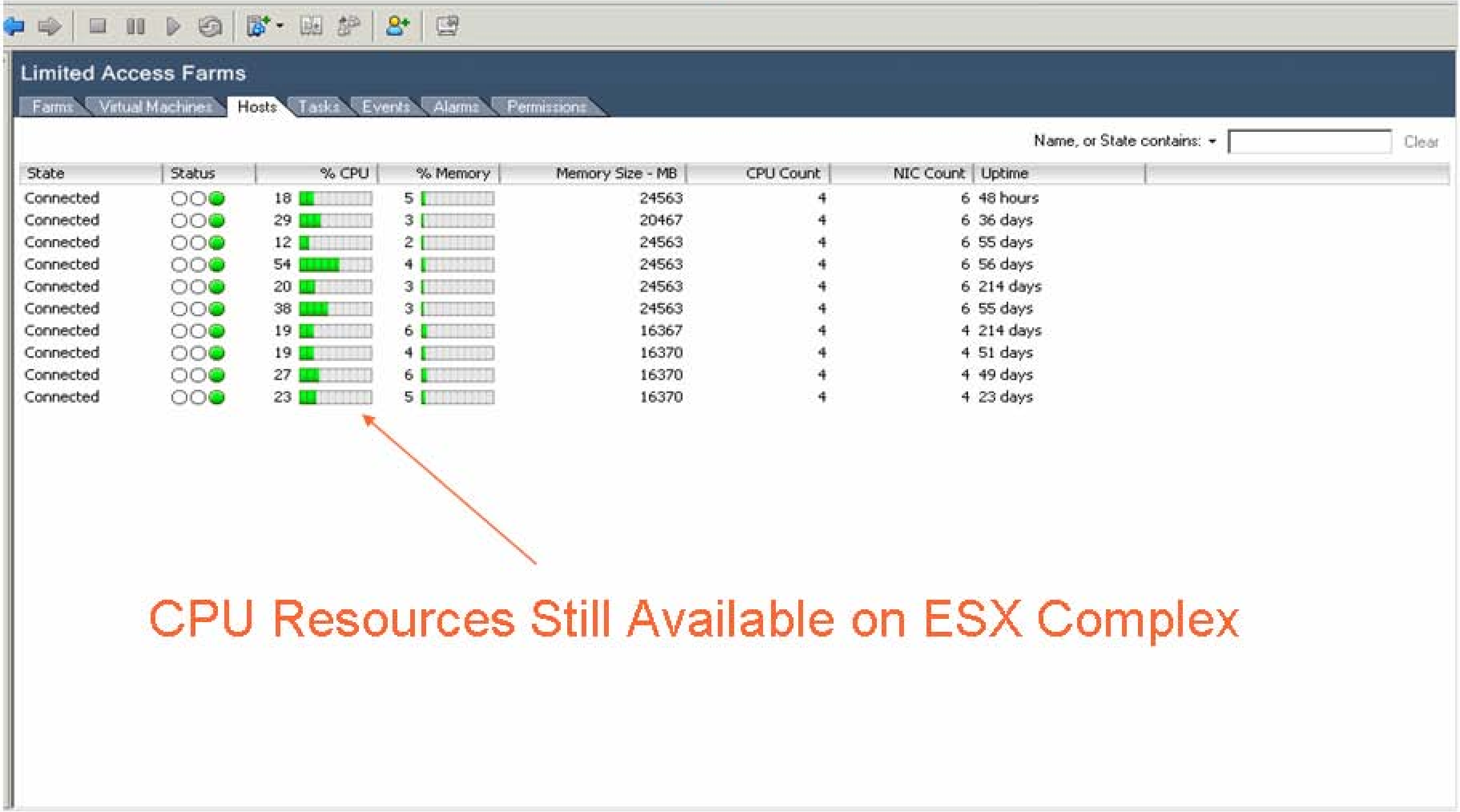
Virtual Center – ESX Farm Context



State	Status	% CPU	% Memory	Memory Size - MB	CPU Count	NIC Count	Uptime
Connected	○○●	19	3	24563	4	6	214 days
Connected	○○●	39	3	24563	4	6	55 days

CPU Resources Still Available on ESX Farm

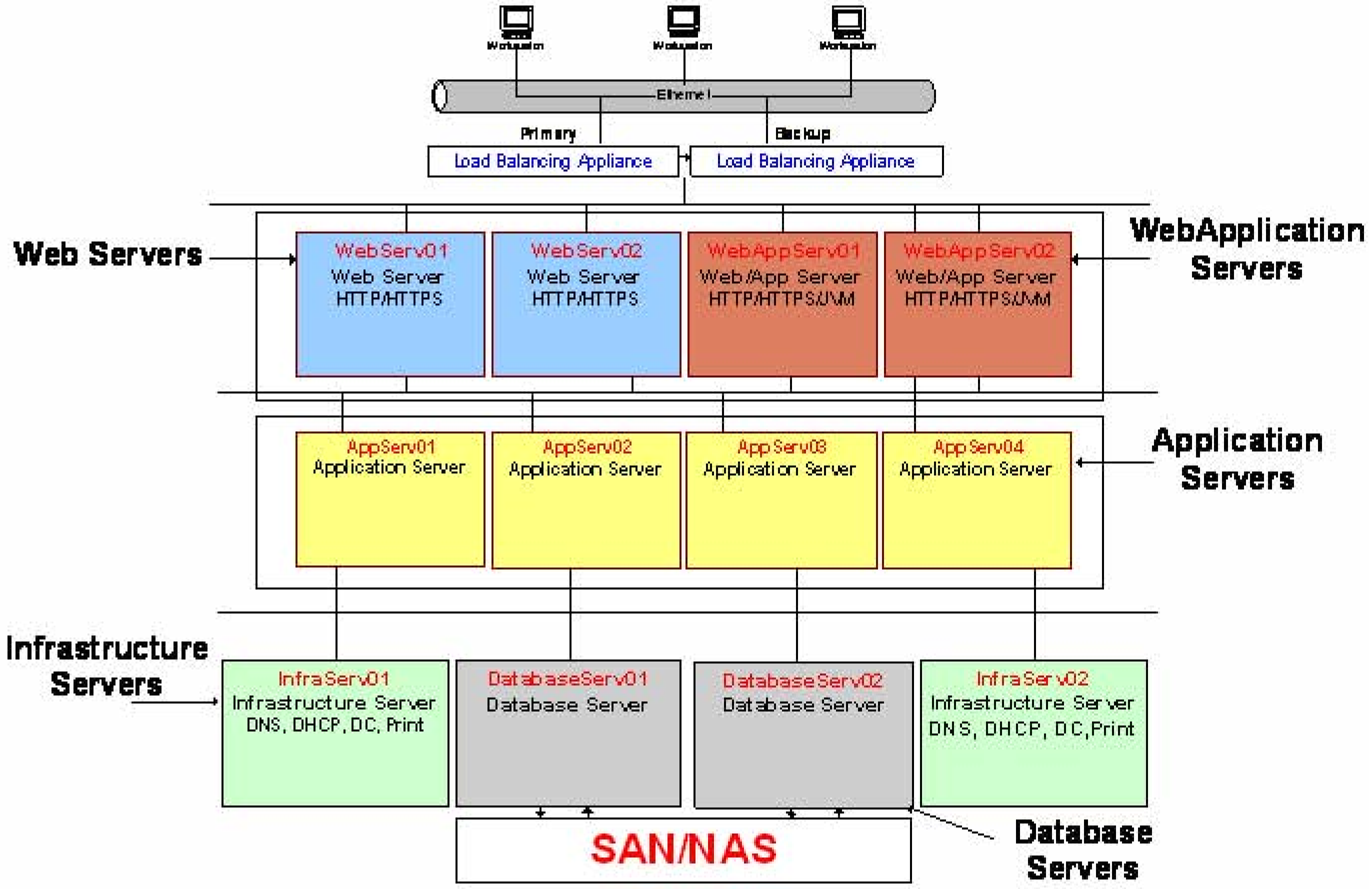
Virtual Center – ESX Complex Context



State	Status	% CPU	% Memory	Memory Size - MB	CPU Count	NIC Count	Uptime
Connected	○○●	18	5	24563	4	6	48 hours
Connected	○○●	29	3	20467	4	6	36 days
Connected	○○●	12	2	24563	4	6	55 days
Connected	○○●	54	4	24563	4	6	56 days
Connected	○○●	20	3	24563	4	6	214 days
Connected	○○●	38	3	24563	4	6	55 days
Connected	○○●	19	6	16367	4	4	214 days
Connected	○○●	19	4	16370	4	4	51 days
Connected	○○●	27	6	16370	4	4	49 days
Connected	○○●	23	5	16370	4	4	23 days

CPU Resources Still Available on ESX Complex

Server Processing Tiers



Reasonability Checks – Web Servers

- Memory and CPU resource consumption should be relatively light unless TCP sockets are constantly created and released
- Secure sockets could increase CPU consumption
- Network I/O is often the top resource consumer
- Most files are usually static except for log files
- High amount of memory sharing
- Target ratio** of 3-4 virtual machines per CPU/socket
- Single CPU allocation
- Average memory allocation of 384-500MB of RAM

**The target ratios in the Reasonability Checks slides are dependent upon many factors in each environment. Some locations have been found to have higher ratios. The ratios listed in these slides have come from real world experience in a large ESX deployment.

Reasonability Checks – Web/Application Servers

- HTTP, HTTPS, java applet/servlet processing
  - Apache Tomcat, IBM Websphere, BEA Weblogic app servers
- Use the application server monitoring tools to measure how the allocated memory is actually being used
- Memory and CPU resource consumption can be VERY high
- There is a low amount of memory sharing within Java Virtual Machines since memory segments are typically unique
- Target ratio** of 1-2 virtual machines per CPU/socket, depending upon the load
- Symmetric multi-processor allocation is best, depending upon the ESX host
- Average memory allocation 1-2GB of RAM per virtual with maximum Java heap size of 75% of the memory allocation
- Tuning of JVM heap size and garbage eating cycle is key. Make sure that the application server returns unused memory to the OS
- Make sure that the JVM heap size settings are not set so high that they consume most of the allocated memory for the Virtual Machine, causing excessive paging at the Windows operating system level

Reasonability Checks – Application Servers

- Essentially batch processing so measure average and peaks. Know the personality of the workloads, especially during high usage time
- Resource consumption is based upon the application but can be artificially throttled with ESX share settings
- This layer can potentially consume all of the ESX host resources during peak periods, which is not necessarily a bad thing but should be watched
- Disk I/O is often the constraining factor
- Data drives typically largest for this layer, which requires a careful analysis of the LUN sizes (if SAN) and LUN performance
- Target ratio** of 2-3 virtual machines per CPU/socket, based on workload
- Symmetric multi-processor allocation is sometimes necessary
- The application layer should be initially sized close to the shared physical machine allocation

Reasonability Checks – Database Servers

- DBMS software (e.g., Microsoft SQL Server) is the primary consumer
- The database may reside on external SAN or local disk
- Disk and network I/O overhead are constraining factors
- A substantial virtualization degradation (up to 25%) can occur in this layer. It is best to start out very conservatively and then add load in your own environment
- Target ratio** of 1-2 virtual machines per CPU/Socket
- Memory allocation of at least 1GB per virtual machine
- Symmetric Multi-Processor allocation IS necessary if the host has > two CPUS
- Check the initialization parameters of the DBMS and ensure that the DBMS does not fully consume the allocated resources for the Virtual Machine
- Database layers can be very different so get your DBAs involved. The more information, the better

Reasonability Checks – Infrastructure Servers

- Print, security patch, anti-virus, “non-user” servers, file servers
  - Peak resource periods do not adversely affect functions
  - Target ratio of 3-4** virtual machines per CPU/Socket
  - Memory allocation of 384-512 MB per virtual machine
  - Single CPU allocation per virtual machine
- DHCP, DNS, Domain Controllers
  - Carefully analyze your network utilization and flow
  - Heavy network I/O in 2.5.x consumes high CPU, better in 3.0
  - Virtualize these in development, tread carefully in production
  - Base your decision on whether to virtualize these types of workloads on your overall network load as well as your expected load on these servers

Use this layer to “fill” farm space

Nothing Definite Yet? Time to Drill Down (and Up)

- How does it really work? One of the biggest lessons learned is the need to thoroughly understand the underlying infrastructure and its impact on the virtual solution
- What is being consumed? Performance analysis is mostly about what is being consumed and whether that consumption is reasonable, given what is known about the applications and infrastructure.
  - Do a quick upwards view to the farm and complex to see if there are other similar problems. If so, suspect the underlying infrastructure
- Where are the serious bottlenecks? Most serious bottlenecks are regularly seen so look for the repeat offenders keeping in mind that:
  - Each application causes different bottlenecks to occur
  - The same server will perform differently for different applications
  - Remember: There are always bottlenecks!

General Bottlenecks

- Memory and disk bottlenecks will have similar symptoms
  - Insufficient memory will require synchronous disk I/O for most network requests because the memory buffer will not be large enough to contain data for most requests
  - A slow disk will likely result in memory buffers filling with write data (or waiting for read data) which will delay all requests because free memory buffers are unavailable for write requests (or response is waiting for read data in disk queue)
  - Disk/controller/memory utilization will typically be very high
  - Most network transfers will happen only after disk I/O has completed
  - Very long response time, low network utilization
  - Processor utilization will usually be low
    - Since disk I/O can take a relatively long time, and disk queues will become full, the processor will be idle or have low utilization as it waits long periods of time before processing next request
- To have a network bottleneck, the network must not be waiting on any disk I/O so the server probably has sufficient memory

CPU Overview – Virtual Machine Context

- Look for any process that is sustaining a high CPU utilization for an extended period of time or is regularly peaking
  - If high CPU utilization is seen, check disk and network I/O rates to see if the root cause is actually I/O
  - Check memory allocation within the Virtual Machine since insufficient memory allocation can drive up CPU consumption
  - A %Ready to Run (ESXTop) greater then 5% may reveal insufficient CPU resources were allocated to the Virtual Machine
  - If SMP is turned on consider going to a uniprocessor configuration
  - If CPU utilization is 100% add CPU shares or allocation first **before** considering SMP
- Remember….
  - User-level application code runs directly at near-native speed
  - Remaining operating system code and virtualized instructions have varying overhead
  - The faster the CPUs (and underlying bus structures) the better the system can absorb the virtual overhead, particularly under stress

Memory Overview - Virtual Machine/ESX Host Contexts

- Remember to keep the context of the different memories clear
  - Virtual Machine “physical” memory
  - Virtual Machine virtual/paging memory
  - ESX Host physical memory
  - ESX Host virtual/paging memory
- Physical memory shortages can be a roadblock at different contexts
- Page Reads/sec, Page Writes/sec
  - Page Reads/sec are the number of disk reads done for paging
  - Page Writes/sec are the number of disk writes done for paging
  - Most other memory counters are for virtual memory
    - These cannot be used to diagnose when a server is running low on physical memory
    - Available MBytes useful for understanding physical memory usage

Storage Subsystem Overview

- The design needs to start with the expected amount of I/Os per second (IOPS) and the expected types of I/Os (sequential read/writes, random reads/writes) and ratios of each
  - Per LUN
  - Per individual paths to the storage subsystem
  - Remember to consider capacity AND performance AND availability/redundancy
    - For multi-threaded I/O intensive applications, more disks = more performance
    - Random read/write workloads usually require lots of disks to scale
    - For random write intensive environments:
      - RAID-10 about 50% greater throughput than RAID-5 at the disk level
- RAID Ratio of performance for comparing RAID strategies:
  - %Reads * (Physical Read Ops) + %Writes * (Physical Write Ops)
- RAID-10, RAID-1, RAID 0+1, RAID-1+0
  - Two physical disk writes per logical write request are required
  - I/O Performance = % Read * (1) + % Write * (2)
- RAID-5
  - Four physical disk I/O operations per logical random write request are required (two reads and two writes)
  - I/O Performance = % Read * (1) + % Write * (4)

When a Disk is not a Disk – SAN Considerations

- Unless you are booting from SAN you need to consider both local disk and network storage
- Local disk involve an on-board or plug-in SCSI controller with a small amount of read/write cache
  - Data drives compete with system drives and paging
- SAN solutions can include network fabric, network switches, network adaptors, host bus adaptors, frame adaptors, front-end processors, microcode, a variety of bus structures, and GBs of cache
  - SAN performance analysis starts with the host machine
    - Start with disk busy, average sector transfer time, IOPS
  - At the SAN level, start with the “back end” physical disks, using SAN management tools
    - The bigger the performance problem the more likely it is in the back end disk area
    - Don’t expect much more than 100 IOPS from a physical disk
  - Work your way upwards inside the SAN, working your way to the SAN fabric
  - Remember that the SAN has the similar challenges to ESX, which is competition for shared resources
    - Look for competition at the physical disk and LUN levels
  - Random reads, random writes, sequential reads, sequential writes may get homogenized in a SAN
  - I/O block sizes can be changed as the data is moved down the I/O path
  - Native SAN tools tend to measure at larger sampling intervals so results will be smoothed
  - Though individual components of a SAN or NAS have absolute throughput limits the aggregate SAN throughput limits are not the sum of its parts

Disk Overview – Virtual Machine/ESX Host Contexts

- Disk bottlenecks are the most common bottleneck
  - Most configure disks based upon capacity requirements, not performance
- Average random disk I/O is about 7 – 8 mSec
  - Disks run optimally with no more than 2 – 3 I/Os in queue
  - At 2 – 3 I/Os per disk average disk latency would be about 21mSec @7mSec latency
- Avg. Disk sec/Transfer
  - If greater than 20-30mSec then the disk is a bottleneck
  - Look at physical disk not logical disk counters
- Split I/Os/Sec
  - Should not be more than small percentage of I/O E.g. 1%
  - High Split I/Os mean the disk is fragmented or the array is not aligned
  - Also check I/O size by monitoring Avg. Disk Bytes/Transfer

Network Overview – All Contexts

- Network performance problems can be difficult to diagnose
- Suspect network problems if poor performance and no other obvious server performance problems
  - Disk latencies are low < 20mSec
  - Sufficient Memory - No/low hard disk paging
    - Page/Reads and Page/Writes/sec < 100
    - and Avg. Disk Sec/Transfer < 15 - 20mSec
  - CPU utilization can be high or low
    - CPU utilization is high if packet sizes are small
    - Many trips through TCP/IP stack
  - CPU utilization is low if packet sizes are large
    - Most time spent doing DMA by LAN adapter

Network Overview – ESX Host Context

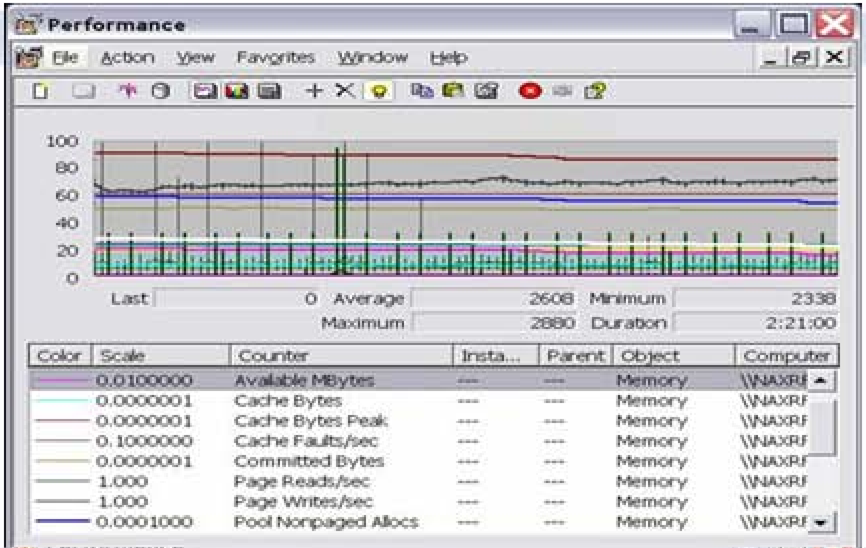
- Server Network Adapter Problems Can be Diagnosed
  - Bytes Total/sec, Packets/sec
  - Chunky Workloads: Bandwidth limited
    - Image servers, file/web/mail servers, video servers, database warehousing
    - Bytes Total/Sec should be well below 50-60% of wire speed
    - Wire speed on full duplex adapters is about 1.3 - 1.6x rated speed
    - Peak sustained rates
      - 1Gbit Enet = 130MB/Sec to 160MB/Sec
      - 100Mbit Enet = 13 - 16MB/Sec
  - Chatty Workloads: - High Packet Rates - Packets/Sec maximums:
    - Message/Chat server, database transaction processing, lots of small files
      - 1Gbit Enet adapter at 50% utilization will drive about 50K - 80K packets/sec
      - 100Mbit Adapter at 50% utilization is about 5K to 8K packets/sec
  - If Bytes/sec is well below wire speed and Packets/sec is low
    - A network bottleneck could still exist in the external network
      - This will require a LAN Analyzer to diagnose

Case Study

File Server Hangs for Several Seconds

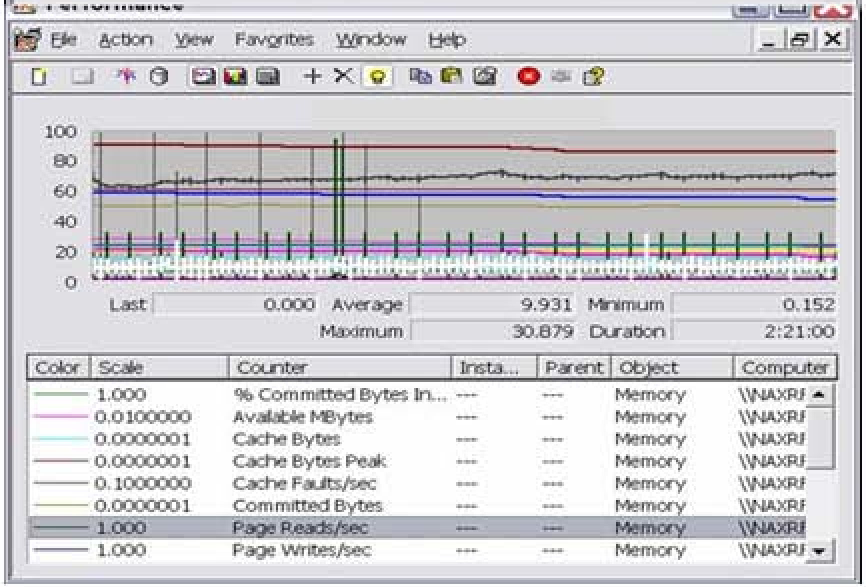
Memory Analysis

- About 2.6 – 3.0GB free Memory Available
  - > No significant paging
  - > Cache Bytes stable (150MB)
- Recommendations
  - > No memory configuration issues noticed
  - > The memory configuration is suitable for this workload



Last 0 Average 2608 Minimum 2338
Maximum 2880 Duration 2:21:00

Color	Scale	Counter	Insta...	Parent	Object	Computer
—	0.0100000	Available MBytes	---	---	Memory	\\WAXRF
—	0.0000001	Cache Bytes	---	---	Memory	\\WAXRF
—	0.0000001	Cache Bytes Peak	---	---	Memory	\\WAXRF
—	0.1000000	Cache Faults/sec	---	---	Memory	\\WAXRF
—	0.0000001	Committed Bytes	---	---	Memory	\\WAXRF
—	1.000	Page Reads/sec	---	---	Memory	\\WAXRF
—	1.000	Page Writes/sec	---	---	Memory	\\WAXRF
—	0.0001000	Pool Nonpaged Allocs	---	---	Memory	\\WAXRF

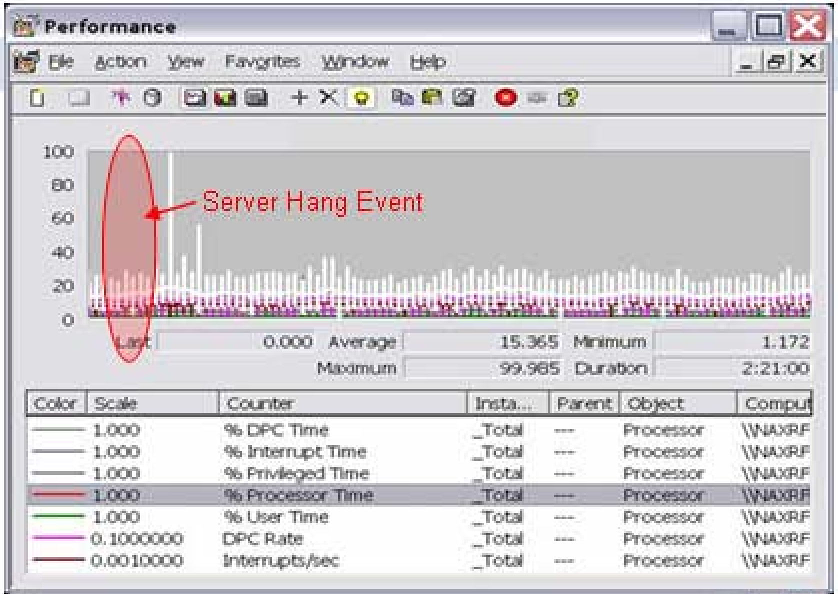


Last 0.000 Average 9.931 Minimum 0.152
Maximum 30.879 Duration 2:21:00

Color	Scale	Counter	Insta...	Parent	Object	Computer
—	1.000	% Committed Bytes In...	---	---	Memory	\\WAXRF
—	0.0100000	Available MBytes	---	---	Memory	\\WAXRF
—	0.0000001	Cache Bytes	---	---	Memory	\\WAXRF
—	0.0000001	Cache Bytes Peak	---	---	Memory	\\WAXRF
—	0.1000000	Cache Faults/sec	---	---	Memory	\\WAXRF
—	0.0000001	Committed Bytes	---	---	Memory	\\WAXRF
—	1.000	Page Reads/sec	---	---	Memory	\\WAXRF
—	1.000	Page Writes/sec	---	---	Memory	\\WAXRF

Processor Analysis

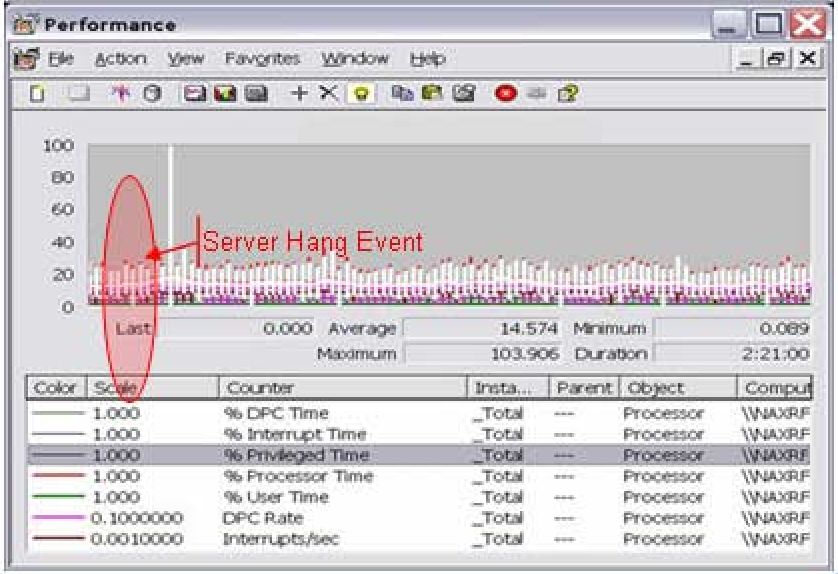
- Average processor utilization is only 15%
- A peak of 99.9% seen as white impulse line.
- Second chart shows that peak and MOST of the processor time is kernel time (OS)
  - Tip:
    - Use %Privilege Time to help identify driver or kernel problems
    - Use %User Time to help identify application problems
- Summary
  - Since the processor is only busy for a very brief time, but at only 15% for most of the day, faster processors should not be needed when the cause for the peak can be determined and remedied



Server Hang Event

Last 0.000 Average 15.365 Minimum 1.172
Maximum 99.985 Duration 2:21:00

Color	Scale	Counter	Insta...	Parent	Object	Comput
—	1.000	% DPC Time	_Total	---	Processor	\\WAXRF
—	1.000	% Interrupt Time	_Total	---	Processor	\\WAXRF
—	1.000	% Privileged Time	_Total	---	Processor	\\WAXRF
—	1.000	% Processor Time	_Total	---	Processor	\\WAXRF
—	1.000	% User Time	_Total	---	Processor	\\WAXRF
—	0.1000000	DPC Rate	_Total	---	Processor	\\WAXRF
—	0.0010000	Interrupts/sec	_Total	---	Processor	\\WAXRF



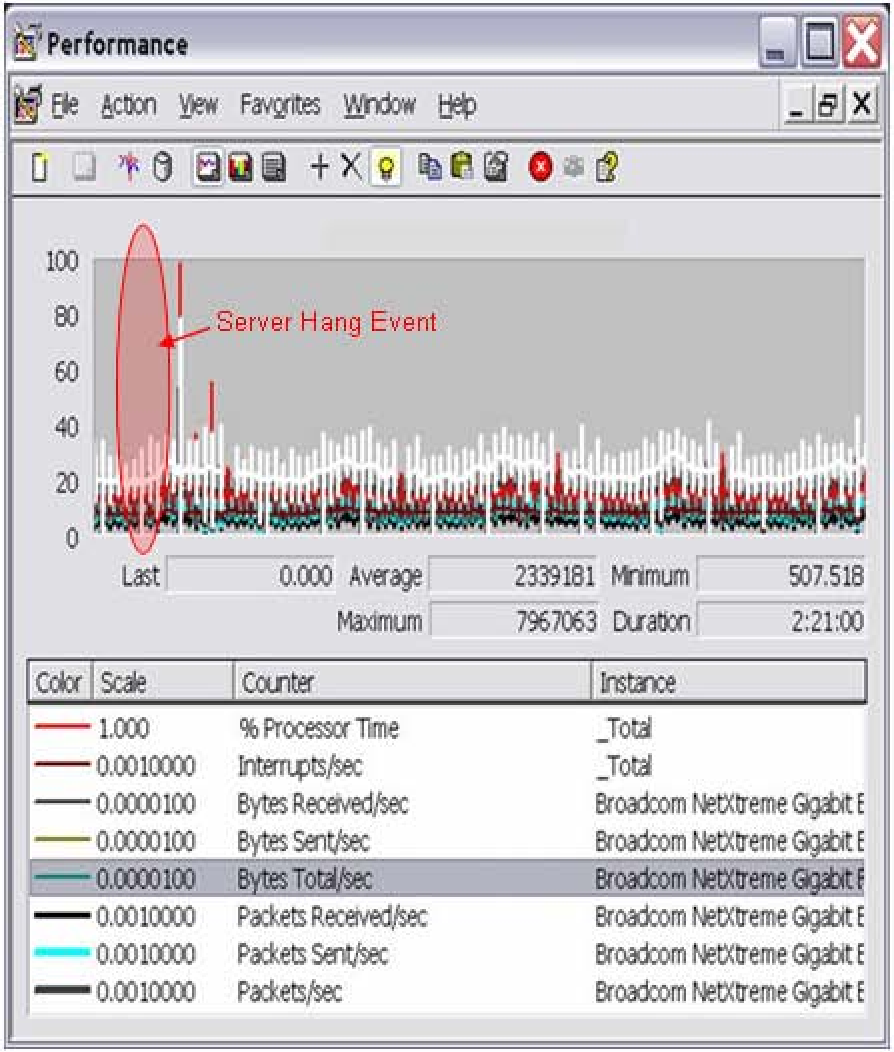
Server Hang Event

Last 0.000 Average 14.574 Minimum 0.089
Maximum 103.906 Duration 2:21:00

Color	Scale	Counter	Insta...	Parent	Object	Comput
—	1.000	% DPC Time	_Total	---	Processor	\\WAXRF
—	1.000	% Interrupt Time	_Total	---	Processor	\\WAXRF
—	1.000	% Privileged Time	_Total	---	Processor	\\WAXRF
—	1.000	% Processor Time	_Total	---	Processor	\\WAXRF
—	1.000	% User Time	_Total	---	Processor	\\WAXRF
—	0.1000000	DPC Rate	_Total	---	Processor	\\WAXRF
—	0.0010000	Interrupts/sec	_Total	---	Processor	\\WAXRF

Network Adapter Analysis

- Network subsystem is very underutilized and is not causing any bottleneck
  - Peak throughput is correlated to the spike in CPU utilization
  - No bottleneck as the LAN is only moving about 8MB/Sec
  - And packets/Sec are about 15,500/sec
- Recommendations
  - Subsystem healthy
  - LAN traffic correlated with CPU spike proves that the spike is in response to a greater request in file server load
  - High bytes/sec show server is responding to the increased load from application servers



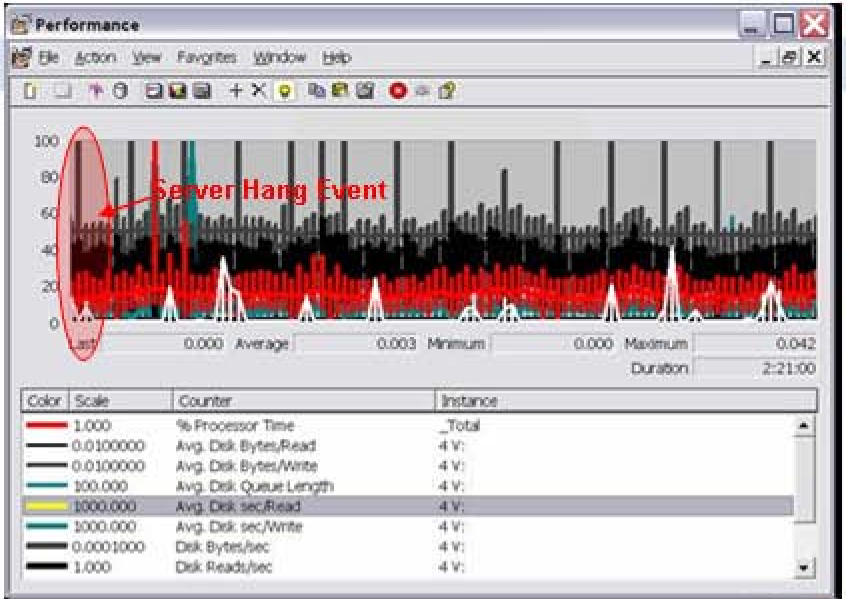
Last 0.000 Average 2339181 Minimum 507.518
Maximum 7967063 Duration 2:21:00

Color	Scale	Counter	Instance
—	1.000	% Processor Time	_Total
—	0.0010000	Interrupts/sec	_Total
—	0.0000100	Bytes Received/sec	Broadcom NetXtreme Gigabit E
—	0.0000100	Bytes Sent/sec	Broadcom NetXtreme Gigabit E
—	0.0000100	Bytes Total/sec	Broadcom NetXtreme Gigabit E
—	0.0010000	Packets Received/sec	Broadcom NetXtreme Gigabit E
—	0.0010000	Packets Sent/sec	Broadcom NetXtreme Gigabit E
—	0.0010000	Packets/sec	Broadcom NetXtreme Gigabit E

Disk Performance Analysis

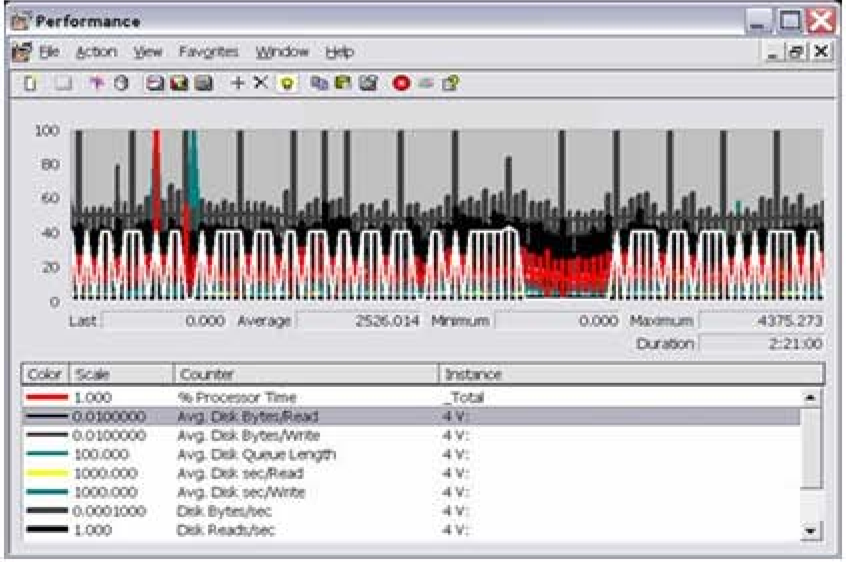
- Avg. Disk Sec/Read (white line) is a minor system bottleneck
  - Avg. Disk Sec/Read is only 3mSec (good)
  - White spikes on top chart are rarely above 20mSec indicating a very minor read bottleneck

- Avg. Bytes/Read (white line) indicate about 4KB maximum
  - Recommendations
    - Disk V: reads do not indicate a significant bottleneck



Last 0.000 Average 0.003 Minimum 0.000 Maximum 0.042
Duration 2:21:00

Color	Scale	Counter	Instance
🟥	1.000	% Processor Time	_Total
⬛	0.0100000	Avg. Disk Bytes/Read	4 V:
⬛	0.0100000	Avg. Disk Bytes/Write	4 V:
🟦	100.000	Avg. Disk Queue Length	4 V:
🟨	1000.000	Avg. Disk sec/Read	4 V:
🟦	1000.000	Avg. Disk sec/Write	4 V:
⬛	0.0001000	Disk Bytes/sec	4 V:
⬛	1.000	Disk Reads/sec	4 V:



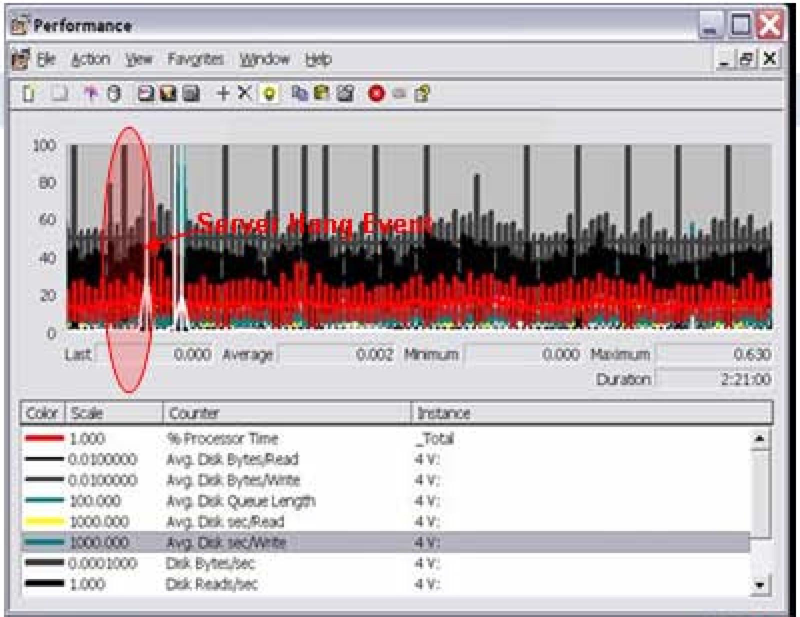
Last 0.000 Average 2526.014 Minimum 0.000 Maximum 4375.273
Duration 2:21:00

Color	Scale	Counter	Instance
🟥	1.000	% Processor Time	_Total
⬛	0.0100000	Avg. Disk Bytes/Read	4 V:
⬛	0.0100000	Avg. Disk Bytes/Write	4 V:
🟦	100.000	Avg. Disk Queue Length	4 V:
🟨	1000.000	Avg. Disk sec/Read	4 V:
🟦	1000.000	Avg. Disk sec/Write	4 V:
⬛	0.0001000	Disk Bytes/sec	4 V:
⬛	1.000	Disk Reads/sec	4 V:

Disk Performance Analysis

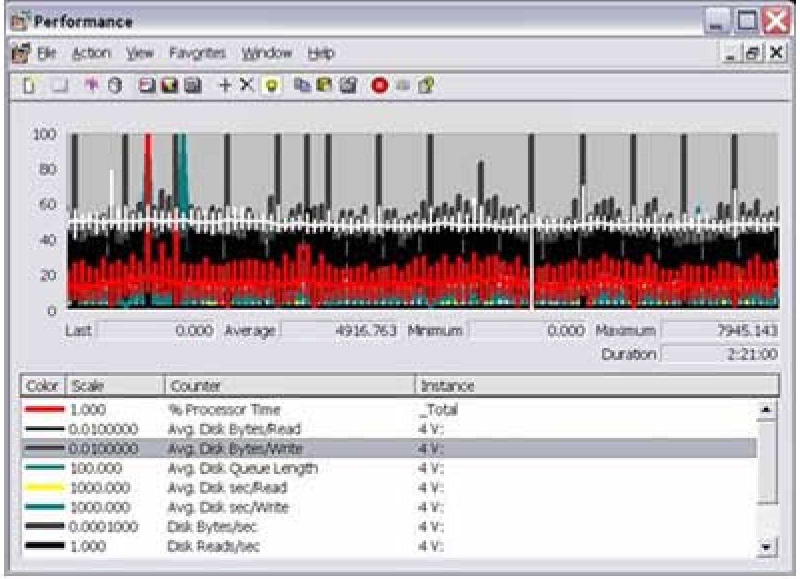
- Avg. Disk Sec/Write is good except for the bottleneck correlated with spike in CPU utilization (red)
  - Avg. Disk Sec/Read is only 3mSec (good)
  - White spike on top chart indicating 630mSec latency
  - Write bottleneck during spike but may not be disk subsystem

- Avg. Bytes/Write (white line) indicate about 5KB, with 7KB maximum so 8KB stripe size should be used for this device
  - Recommendations
    - Disk V: should have at least an 8KB stripe size



Last 0.000 Average 0.002 Minimum 0.000 Maximum 0.630
Duration 2:21:00

Color	Scale	Counter	Instance
🟥	1.000	% Processor Time	_Total
⬛	0.0100000	Avg. Disk Bytes/Read	4 V:
⬛	0.0100000	Avg. Disk Bytes/Write	4 V:
🟩	100.000	Avg. Disk Queue Length	4 V:
🟨	1000.000	Avg. Disk sec/Read	4 V:
🟩	1000.000	Avg. Disk sec/Write	4 V:
⬛	0.0001000	Disk Bytes/sec	4 V:
⬛	1.000	Disk Reads/sec	4 V:

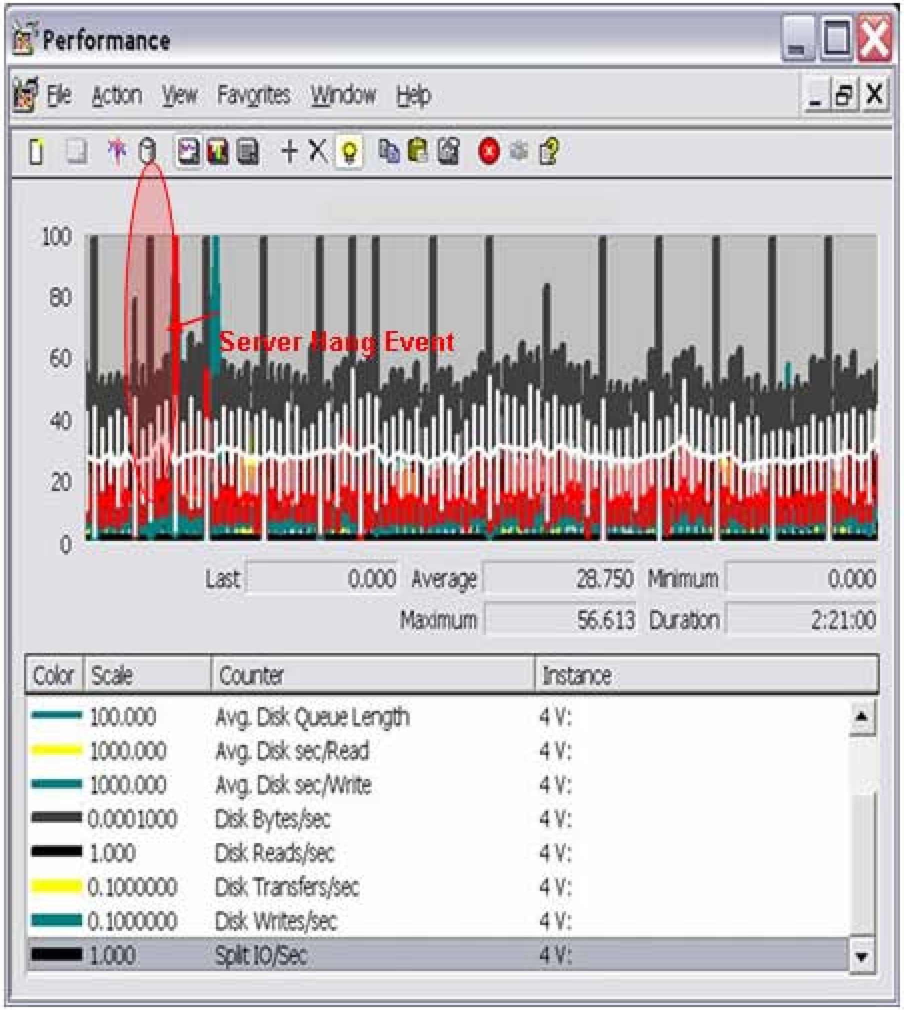


Last 0.000 Average 4916.763 Minimum 0.000 Maximum 7945.143
Duration 2:21:00

Color	Scale	Counter	Instance
🟥	1.000	% Processor Time	_Total
⬛	0.0100000	Avg. Disk Bytes/Read	4 V:
⬛	0.0100000	Avg. Disk Bytes/Write	4 V:
🟩	100.000	Avg. Disk Queue Length	4 V:
🟨	1000.000	Avg. Disk sec/Read	4 V:
🟩	1000.000	Avg. Disk sec/Write	4 V:
⬛	0.0001000	Disk Bytes/sec	4 V:
⬛	1.000	Disk Reads/sec	4 V:

Disk Performance Analysis

- Excessive split I/O rate to V disk
  - Avg. about 30/Sec
- Recommend to investigate cause for split I/O as this should not be occurring and reduces SAN performance
  - Split IO/Sec reports the rate at which I/Os to the disk were split into multiple I/Os. A split I/O may result from requesting data of a size that is too large to fit into a single I/O or that the disk is fragmented



Last 0.000 Average 28.750 Minimum 0.000
Maximum 56.613 Duration 2:21:00

Color	Scale	Counter	Instance
🟩	100.000	Avg. Disk Queue Length	4 V:
🟨	1000.000	Avg. Disk sec/Read	4 V:
🟩	1000.000	Avg. Disk sec/Write	4 V:
⬛	0.0001000	Disk Bytes/sec	4 V:
⬛	1.000	Disk Reads/sec	4 V:
🟨	0.1000000	Disk Transfers/sec	4 V:
🟩	0.1000000	Disk Writes/sec	4 V:
⬛	1.000	Split IO/Sec	4 V:

Some Final Lessons Learned…

- Performance analysis techniques are a synergy between Windows and mainframe techniques
- It's fairly easy to reassign ESX hosting resources but difficult and costly to modify poorly designed storage and network subsystem components
- Develop a process to identify performance problems and share it with others
- DO NOT underestimate the importance and challenges associated with the underlying storage and network infrastructure design and performance
- DO NOT try to put too much additional load into an existing, heavily utilized ESX complex
- Perseverance is the key to a successful conversion to virtual environment

Performance Monitoring & Capacity Planning

John Paul – johnathan.paul@siemens.com

Session: ADC9398

Questions?

Presentation Download

Please remember to complete your
session evaluation form
and return it to the room monitors
as you exit the session

The presentation for this session can be downloaded at
http://www.vmware.com/vmtn/vmworld/sessions/

Enter the following to download (case-sensitive):

Username: cbv_rep
Password: cbvfor9v9r

VMWORLD 2006