VM Performance 101

WAVV 2004

Bill Bitner
IBM Endicott
607-429-3286
bitnerb@us.ibm.com

Last Updated: April 24, 2004

Legal Stuff

Disclaimer

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "as is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environment do so at their own risk.

In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

Any performance data contained in this document was determined in a controlled environment and, therefore, the results which may be obtained in other operating environments may vary significantly.

Users of this document should verify the applicable data for their specific environments.

It is possible that this material may contain references to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country or not yet announced by IBM. Such references or information should not be construed to mean that IBM intends to announce such IBM products, programming, or services.

Should the speaker start getting too silly, IBM will deny any knowledge of his association with the corporation.

Trademarks

The following are trademarks of the IBM Corporation:

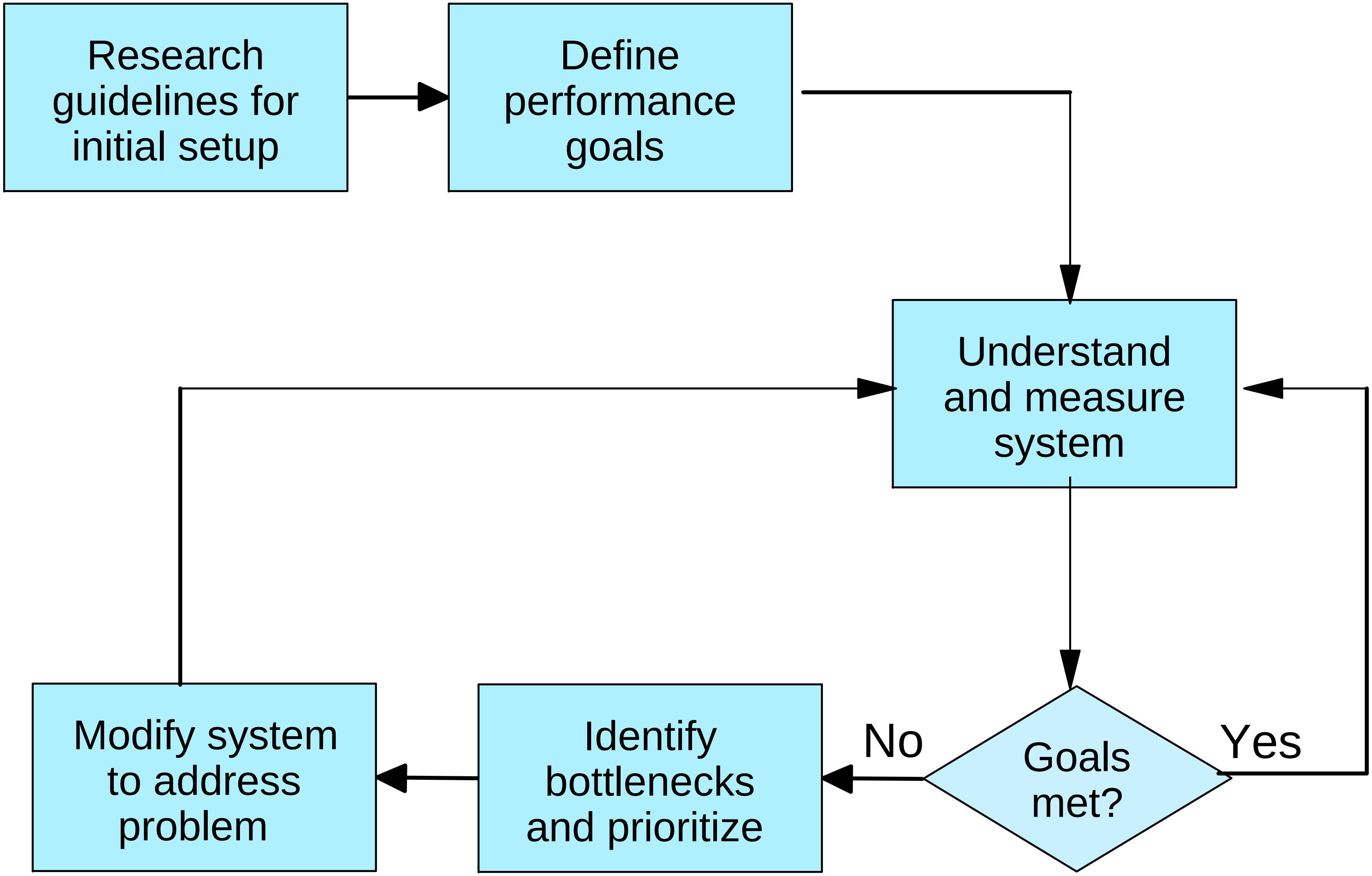
IBM, VM/ESA, z/VM

LINUX is a registered trademark of Linus Torvalds

Overview

- Performance process
- Performance definition
- Guidelines
- Native CP commands
- Other performance tools
- I/O performance concepts
- Case study
- Final thoughts

Performance Process


Research guidelines for initial setup
Define performance goals
Understand and measure system
Goals met?
Yes
No
Identify bottlenecks and prioritize
Modify system to address problem

Definition of Performance

Performance definitions:

- ❑ Response time
- ❑ Batch elapsed time
- ❑ Throughput
- ❑ Utilization
- ❑ Users supported
- ❑ Phone ringing
- ❑ Consistency
- ❑ All of the above

Performance Guidelines

- Processor
- Storage
- Paging
- Minidisk cache
- Server machines

Processor Guidelines

- Dedicated processors - mostly political
  - A virtual machine should have all dedicated or all shared processors
- Share settings
  - Use absolute if you can judge percent of resources required
  - Use relative if difficult to judge and if lower share as system load increases is acceptable
  - Do not use LIMITHARD settings unnecessarily
- Small minor time slice keeps CP reactive.

Storage Guidelines

- Use SET RESERVE instead of LOCK to keep users pages in storage
- Define some processor storage as expanded storage to provide paging hierarchy (even when running a 64-bit CP)
- Exploit shared segments and SAVEFD where possible.
- SFS use of VM data spaces saves storage
- DB2 use of VM data spaces requires storage

Paging Guidelines

- DASD paging allocations less than or equal to 50%.
- Watch blocks read per paging request (keep >10)
- Multiple volumes and multiple paths
- Do not mix with other data types
- In a RAID environment, enable cache to mitigate write penalty.

Minidisk Cache Guidelines

- Configure some real storage for MDC.
- In general, enable MDC for everything.
- Disable MDC for
  - Minidisks mapped to VM data spaces
  - write-mostly or read-once disks (logs, accounting)
  - Backup applications
- In large storage environments, may need to bias against MDC.
- Better performer than vdisks for read I/Os

SVM Guidelines

- QUICKDSP ON to avoid eligible list
- Higher SHARE setting
- SET RESERVED to avoid paging
- NOMDCFS in directory option
- DIAG98 in directory where applicable
- Exploit DASD Fast Write for servers that do synchronous writes
- Potentially different CMS
  - Segment management
  - File buffers can be larger

Virtual Machine Guidelines

- Do not worry about 32 MB line. Pick a location above CMS/IBM segments and work up.
- Use SAVEFD where possible, or SFS dircontrol with data spaces
- Execs
  - Compile
  - Execload
  - In segment?
- Do not define more virtual CPUs than needed.

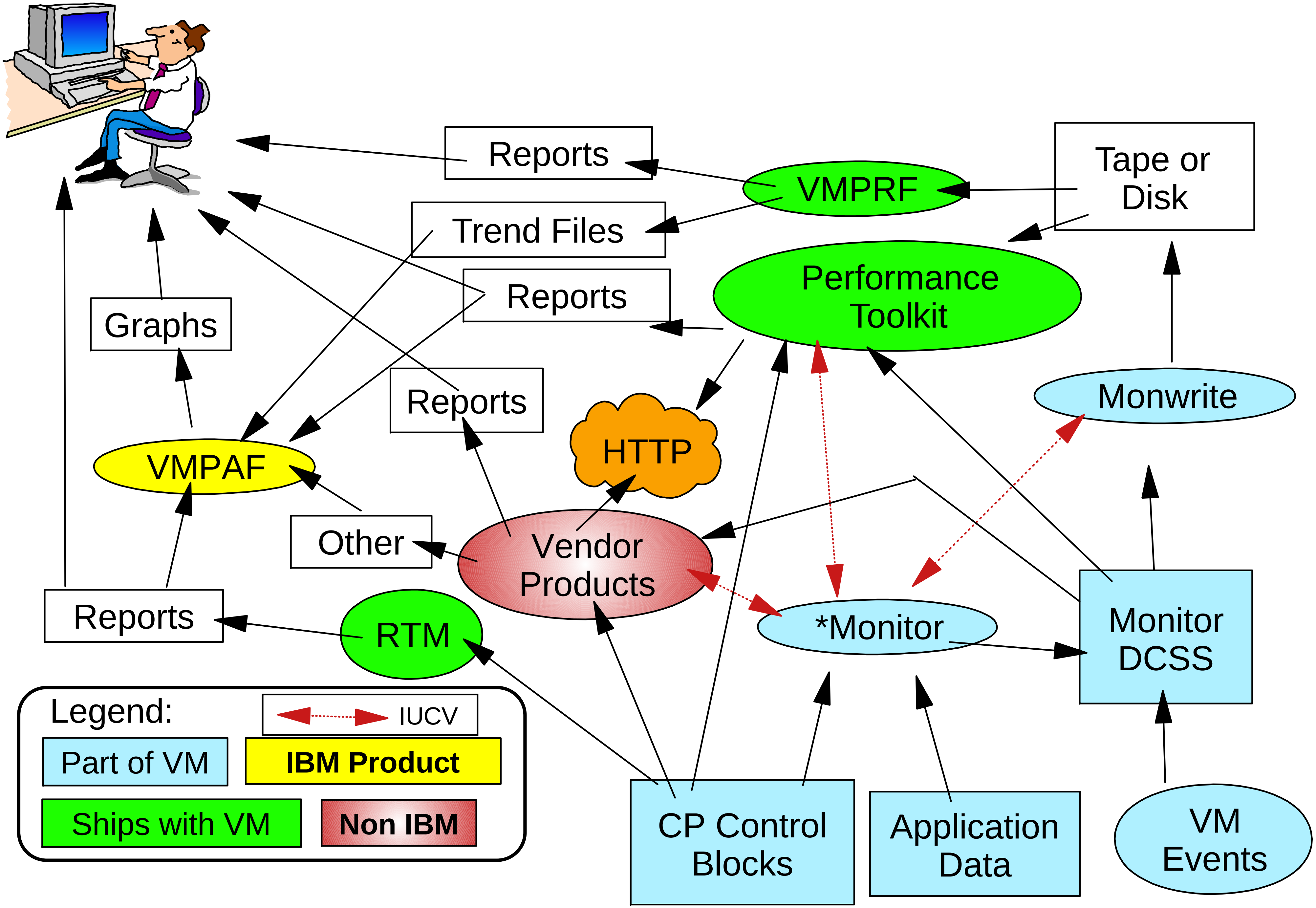
CP INDICATE Command

- LOAD: shows total system load. (STORAGE value not very meaningful)
- USER EXP: more useful than Indicate User
- QUEUES EXP: great for scheduler problems and quick state sampling
- PAGING: lists users in page wait.
- IO: lists users in I/O wait.
- ACTIVE: displays number of active users over given interval

Selected CP QUERY Commands

- Users: number and type of users on system
- SRM: scheduler/dispatcher settings
- SHARE: type and intensity of system share
- FRAMES: real storage allocation
- PATHS: physical paths to device and status
- ALLOC MAP: DASD allocation
- XSTORE: assignment of expanded storage
- MONITOR: current monitor settings
- MDC: MDC usage
- VDISK: virtual disk in storage usage

Performance Data Food Chain



State Sampling

- Find the state of given user or device
  - ►Consolidation of samples gives useful info
- Snap view:
  - ►INDICATE QUEUES
  - ►RTM Display User
- Low frequency:
  - ►RTM Display SRC
- High frequency:
  - ►Monitor: user, processor, and I/O domains
  - ►CP MONITOR SAMPLE RATE

I/O Response Time

Resp Time = Service Time + Queue Time

Service Time = Pending + Connect + Disconnect

- Queue Time: from hi-frequency sampling of queue in RDEV. Reported in monitor.
- Function Pending: time accumulated when a path to device cannot be obtained.
  - ► < 1 ms, unless contention at channels or control units.
- Connect: time device logically connected to channel path
  - ► proportional to amount of data per I/O

I/O Response Time *(continued)*

- Disconnect: time accumulated when device is logically disconnected from channel while subchannel system is active.
  - Cache miss
  - Seek on older devices
  - CU management
- Device Active: time accumulated between return of channel-end and device-end
  - Often reported as part of Disconnect Time

Other Sources

- SC24-5999-03 z/VM 4.4.0: Performance - Part of the z/VM Library
- http://www.vm.ibm.com/perf/
  - links to documents, tools, reference material
- http://www.vm.ibm.com/perf/tips/
  - common problems and solutions
  - guidelines
- http://www.vm.ibm.com/devpages/bitner/
  - presentations with speaker notes

A Case Study

The Grinch That Stole Performance

From VMPRF USER_STATES_BY_TIME PRF007 Report January 5:

```
<----Percent of True Non-Dormant Time Waiting on--------------->
                                         <---SVM and---->  I/O
    Load-              Inst  Test  Cons  Test Elig-  Dor-   Ac-
CPU   ing  Page   I/O   Sim  Idle  Func  Idle  ible  mant  tive

0.1   0.1   0.1  18.8   2.3  10.0   0.4   3.4     0  50.8   8.4
0.1     0   0.1  16.0   1.9   9.9   0.4   3.1     0  53.8   9.9
```

From VMPRF DASD_BY_ACTIVITY PRF012 Report January 5:

```
     SSCH  Pct   <-----------Time------------> <--Queue-->
Dev. Rate  Busy  Pend  Disc  Conn  Serv  Resp  Mean   Max

1742 26.7  65.4   1.3  18.4   4.7  24.5  69.0   1.2   8.5
```

Went to check VMPRF DASD_BY_ACTIVITY_EF PRF095 for control unit cache stats, but it didn't exist!

It is a good thing I keep historical data -- let's go back and see what's going on...

When Did We Last See Cache?

From VMPRF DASD_BY_ACTIVITY PRF012 Report from December 8:

```
      SSCH  Pct   <-----------Time------------> <--Queue-->
 Dev. Rate  Busy  Pend  Disc  Conn  Serv  Resp  Mean   Max
1742  41.0  10.5   0.3   0.2   2.0   2.6   2.9   0.0   0.3
Jan5: 26.7  65.4   1.3  18.4   4.7  24.5  69.0   1.2   8.5
```

VMPRF DASD_BY_ACTIVITY_EF PRF095 Report for 1742 on Dec 8:

```
<---------Rate--------> <-----Percent----------->
Total  Read  Read Write      <---------Hits----->
  I/O NonSq   Seq      FW Read  Tot Read  Wrt  DFW

 53.0  52.3     0   0.6    99   99   99   96   96
```

Down for the 3-Count

```
q dasd details 1742
1742 CUTYPE = 3990-EC, DEVTYPE = 3390-06, VOLSER= USE001
      CACHE DETAILS:  CACHE NVS CFW DFW PINNED CONCOPY
           -SUBSYSTEM   F    Y   Y   -    Y
           -DEVICE      Y    -   -   Y    N       N
      DEVICE DETAILS: CCA = 02, DDC = 02
      DUPLEX DETAILS: SIMPLEX
```

Pinned data! Yikes! I had never seen that before!

Performance Toolkit Device Report

```
FCX110      CPU 2003   GDLVM7    Interval INITIAL. - 13:08:47     Remote Data

Detailed Analysis for Device 1742 ( SYSTEM )
Device type :  3390-2      Function pend.:     .8ms      Device busy   :   27%
VOLSER      :  USE001      Disconnected  :   20.3ms      I/O contention:    0%
Nr. of LINKs:     404      Connected     :    5.4ms      Reserved      :    0%
Last SEEK   :    1726      Service time  :   26.5ms      SENSE SSCH    :  ...
SSCH rate/s :    10.5      Response time :   26.5ms      Recovery SSCH :  ...
Avoided/s   :    ....      CU queue time :     .0ms      Throttle del/s:  ...
Status: SHARABLE

Path(s) to device 1742:    0A    2A    4A
Channel path status   :    ON    ON    ON

Device            Overall CU-Cache Performance           Split
DIR ADDR VOLSER   IO/S %READ %RDHIT %WRHIT ICL/S BYP/S   IO/S %READ %RDHIT
08  1742 USE001     .0     0      0      0    .0    .0   'NORMAL' I/O only
```

Performance Toolkit Device Report

```
  MDISK Extent    Userid   Addr IO/s VSEEK Status    LINK MDIO/s
+----------------------------------------------------------------+
|   101 -   200    EDLSFS   0310   .0     0 WR            1     .0 |
|   201 -   500    EDLSFS   0300   .0     0 WR            1     .0 |
|   501 -   600    EDLSFS   0420   .0     0 WR            1     .0 |
|   601 - 1200    EDLSFS   0486   .0     0 WR            1     .0 |
|  1206 - 1210    RAID     0199   .0       owner                 |
|                 BRIANKT  0199   .0     0 RR            5     .0 |
|  1226 - 1525    DATABASE 0465   .0       owner                 |
|                 K007641  03A0   .0     0 RR            3     .0 |
|  1526 - 1625    DATABASE 0269   .0       owner                 |
|                 BASILEMM 0124   .0     0 RR           25     .0 |
|  1626 - 1725    DATABASE 0475   .0       owner                 |
|                 SUSANF7  0475   .0     0 RR            1     .0 |
|  1726 - 2225    DATABASE 0233   .0     0 owner        366   10.5 |
+----------------------------------------------------------------+
```

Solution

- Use **Q PINNED** CP command to check for what data is pinned.
- Discussion with Storage Management team.
- Moved data off string until corrected.

Pinned data is very rare, but when it happens it is serious.

Some Final Thoughts

- Collect data for a base line of good performance.
- Implement change management process.
- Make as few changes as possible at a time.
- Performance is often only as good as the weakest component.
- Relieving one bottleneck will reveal another. As attributes of one resource change, expect at least one other to change as well.
- Latent demand is real.