Linux on z/VM Configuration Guidelines

Barton@VelocitySoftware.com
HTTP://VelocitySoftware.com
HTTP://LinuxVM.com

2014

"If you can't Measure it,
I am Just Not Interested ™"

VELOCITY SOFTWARE

Overview

Configuring z/VM for Linux on zSeries

- Must configure z/VM – many defaults incorrect
- Linux must be configured for shared resource environment
- Many actions not intuitive
- “Best Practices”

Infrastructure unknowns for “new” installations

- How to manage performance / capacity planning?
- Is chargeback important?
- Operational support for 1,000 servers?
- What are the limits of a configuration and how to measure
- How to share resources to INCREASE the ROI

Measurement and Tuning for z/VM IS Required

- Start with Proper Configurations

VELOCITY SOFTWARE

Configuration Options

General Storage Options

Linux Options

- Storage Sizes
- Swapping for Linux
- Linux virtual processors
- Network

z/VM Configuration

- Network, I/O, FTP Topics
- MDC
- Paging and Spooling for z/VM
- DASD/Cache/Channels
- z/VM System parameters
- Expanded Storage

Infrastructure

- Linux infrastructure – monitoring availability and performance

General Storage Requirements

Configuration requirements different for

- Small Infrastructure Servers – “Small” Systems
  - DNS, Apache, Samba
  - Low I/O rate
  - Real storage less than 2gb
  - Virtual servers sized 64mb to 256mb
- Medium (31bit) Application Servers - Small to Large Systems
  - Websphere, Domino, Oracle
  - z/VM Real storage greater than 2GB
  - High I/O rate potential
  - Typical 512MB to 2GB
- Large (64bit) Application Servers - Large Systems
  - Oracle, SAP
  - z/VM Real storage greater than 10GB
  - High I/O rate potential
  - Virtual Servers Typical 512MB to 16GB

VELOCITY SOFTWARE

Storage Considerations

z/VM is shared resource environment

- Over-committing storage improves costs per server
- Over-allocating storage reduces servers that can be supported
- **QDROP IS QUITE IMPORTANT**

Storage requirements of Linux very high

- Linux designed for dedicated storage, references all storage
- Linux is LRU, competing with VM's reference pattern
- High percent of referenced pages – what can z/VM page out?

Linux does not drop from queue –

- 100 timer pops per second was 1st problem, fixed.
- CP storage management bypassed, forces "emergency scan"
- **Current release of IBM JDK (WAS) polls 10 ms**

VELOCITY SOFTWARE

Qdrop issues

Storage in use by LINUX07 has 166K pages

- Server was active 8 hours prior, but idle for 8 hours
  - never dropped from queue, never gave up storage
- Active server LINUX02 must compete for reduced storage
- **Guideline: Force Linux Servers to drop from queue**

```
Report: ESAUSPG User Storage Analysis    STRESS TEST  ESAMAP 3.4.1 07/25/04  Page 149
------------------------------------------------------------------------------------
         <---Storage occupancy in pages---> <--Main Storage page Read/Write--> Pages
UserID   <---Main Storage---> <--Paging---> <-Page Writes to:--> <Page Reads:> Moved
/Class    Total   >2gb   <2GB  Xstor   DASD   Xsto   Disk   Migr  Xstor  Disk   <2GB
-------- ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------
*********************************************User Summary**********************
LINUX02  386858 277741 109118 128021      1 830202      1      0 122802      1  1147K
LINUX03  696040 691618   4422 506011      0  36439      0      0  36340      0   7398
ESAWRITE   1141   1124     17     17      0    882      0      0    846      0   1476
LINUX01   71478  67316   4162   1107      1  28815      0      0  28699      0    916
LINUX07  227393  60727 166666  28958      0 140013      0      0  23247      0    397
------------------------------------------------------------------------------------
Total     1881K  1577K 303851 733457      2  1154K      1      0 327905      1  1163K
```

Linux Storage (RAM)

z/VM Paging

- Over commitment of storage causes paging
- **Over commitment of storage reduces cost**
- Paging is common **(manageable)** performance problem

Linux Swapping

- Swapping result of over commitment of Linux storage
- Swapping to vdisk very fast, uses storage when it happens
- Swapping to dasd very slow, always noticeable

VELOCITY SOFTWARE

Linux Cache

- Linux avoids I/O by using cache
- Linux will cache gigabytes of data if allowed
- Oracle SGA MUST fit in linux page cache
- Swap historically was slow SCSI device

Reduce size of Linux Virtual Machine MAJOR Knob.

- Reducing virtual machine size reduces caching of old data
- Define virtual disk for swap
- Virtual Disk paged out when not in use - Unlike “Real” memory
- Experiment with Linux server swapped 40,000 per second.

Tailoring Linux Storage

Linux data shows
- Real storage
- Swap storage
- "cache"

Some Swapping is "good"

If not swapping,
- reduce vm size
- Use CMM to reduce

```
Report: ESAUCD2      LINUX UCD Memory Analysis Report                TEST MAP
-------------------------------------------------------------------------------
Node/    <-------------------------Storage Sizes (in MegaBytes)------------->
Time/    <--Real Storage--> <-----SWAP Storage----> Total <--Storage in Use->
Date     Total  Avail Used  Total Avail Used  MIN   Avail Shared Buffer Cache
-------- ------ ----- ----- ----- ----- ----- ----- ----- ------ ------ -----
20:58:35
LNXldap   122.4   4.6 117.8 511.4 501.2  10.2  15.6 505.8      0   17.1  49.6
LNXnfs    193.1   4.6 188.5 511.4 511.0   0.4  15.6 515.6      0   29.6  55.7
LNXzero   122.8   3.4 119.3 444.2 436.1   8.1  15.6 439.5      0   19.6  43.2
LNXdna2   499.6 182.9 316.8 317.3 317.3     0  15.6 500.1      0   25.7 164.5
LNXdna3   499.6  25.0 474.6 511.4 511.4     0  15.6 536.4      0   38.7 315.0
LNXtux    502.2   6.7 495.5 571.1 571.1     0  15.6 577.8      0  108.9 180.8
LNXPRbt0  499.6  22.9 476.7 511.4 511.4     0  15.6 534.3      0   94.6 241.5
LNXPRbt1  499.6  27.6 472.0 511.4 511.4     0  15.6 539.0      0   25.2 299.9
LNXPRbt2  287.4  18.5 268.9 511.4 511.4     0  15.6 529.9      0   30.7 106.3
LNXPRci1  499.6  10.1 489.5 511.4 358.6 152.9  15.6 368.7      0   20.6 269.4
LNXPRci2  499.6  21.3 478.4 511.4 449.8  61.7  15.6 471.0      0   17.7 164.5
LNXPRot1  499.6   8.5 491.1 511.4 394.6 116.8  15.6 403.1      0   39.0 164.5
LNXPRot3  704.0  12.1 691.8 511.4 511.4     0  15.6 523.6      0   28.9 239.9
LNXPRot5  499.6   4.0 495.6 511.4 451.3  60.1  15.6 455.3      0    4.4 426.5
LNXPRrg1  499.6  15.1 484.5 511.4 431.8  79.6  15.6 446.9      0   22.1 104.1
LNXPRrg2  499.6  24.6 475.0 511.4 465.3  46.1  15.6 489.9      0   23.1 127.1
LNXPRmk1  499.6  24.0 475.6 511.4 453.2  58.2  15.6 477.3      0    8.5 156.1
LNXPRmk2  499.6  27.2 472.4 511.4 465.2  46.3  15.6 492.4      0   13.6 136.3
LNXPRmx1  499.6  36.0 463.6 511.4 465.4  46.0  15.6 501.4      0   14.2 141.6
LNXPRic1  499.6  31.6 468.0 511.4 462.5  48.9  15.6 494.1      0   20.6 184.6
LNXPRic5  248.1   5.6 242.5 511.4 437.8  73.6  15.6 443.4      0    2.4 201.0
LNXPRic6  248.1   5.7 242.4 511.4 467.7  43.7  15.6 473.5      0    2.3 194.8
LNXPRic2  499.6  27.6 472.0 511.4 511.4     0  15.6 539.0      0   38.7 213.9
LNXPRiv1  499.6  16.0 483.6 511.4 316.7 194.7  15.6 332.7      0    2.8 281.7
LNXPRmx1  499.6  29.7 470.0 511.4 511.4     0  15.6 541.1      0   15.3 151.6
LNXPRmx2  499.6  27.8 471.8 511.4 459.2  52.3  15.6 487.0      0   14.6 143.1
LNXPRbq1  499.6  11.6 488.1 511.4 453.2  58.2  15.6 464.8      0   16.3  92.5
LNXPRsd1  499.6  23.7 475.9  1023  1023     0  15.6  1047      0    3.9 411.0
LNXPRkf1  499.6 161.8 337.8 511.4 511.4     0  15.6 673.2      0   22.7 178.9
LNXPRot2  751.1  13.5 737.6 511.4 511.4     0  15.6 524.9      0   47.3 235.8
LNXPRa8   502.3  21.5 480.8 507.7 507.7     0  15.6 529.2      0   18.9 292.3
```

Linux Swapping

Reducing virtual storage size may cause swap

- Linux does not swap until out of storage

Swapping to disk

- VERY VERY SLOW
- Other platforms increase storage size because disk is slow
- **Swap to disk if you want to penalize a server**
- Max swap rate maybe 200 on a very good day

Linux Swapping to Vdisk

- Not a performance degradation
- 40,000 / second is FAST

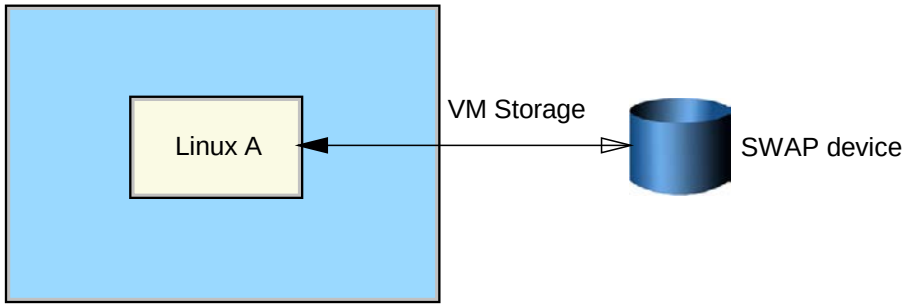
Swap Guideline:

- Define 2 virtual disks, prioritized swap
- Use DIAG driver instead of FBA - Reduces I/O by factor of 8

VELOCITY SOFTWARE

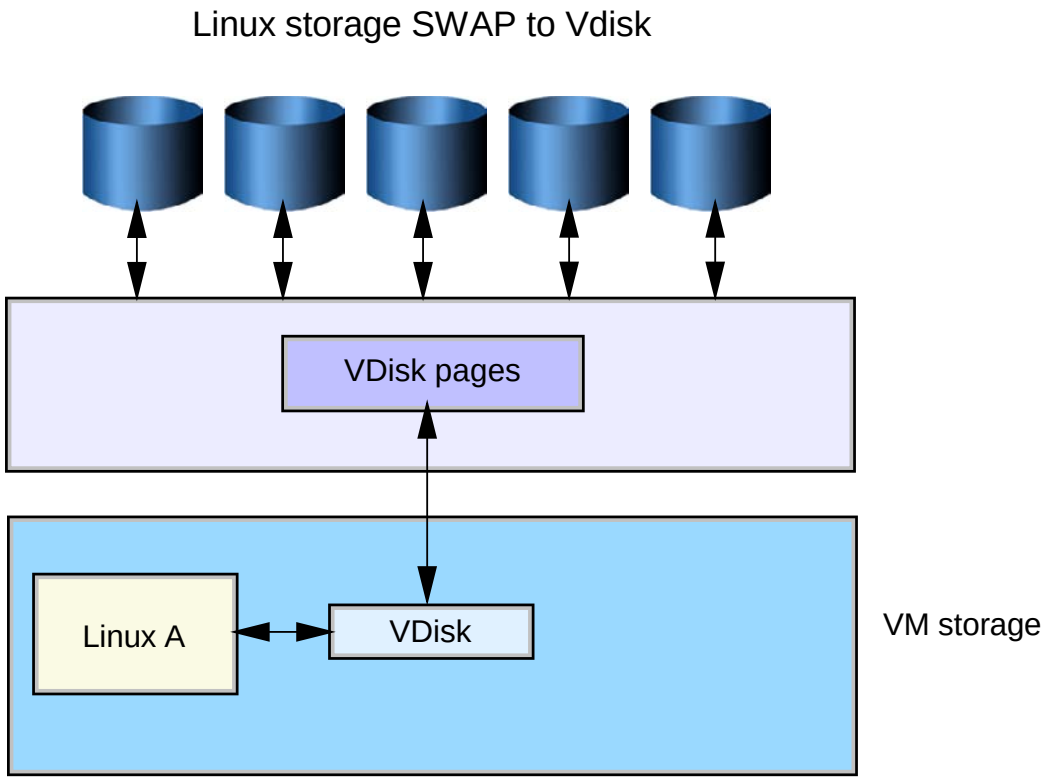
VM Storage Overview, Paging Hierarchy

Linux storage/SWAP



z/Linux Page / Swap Hierarchy

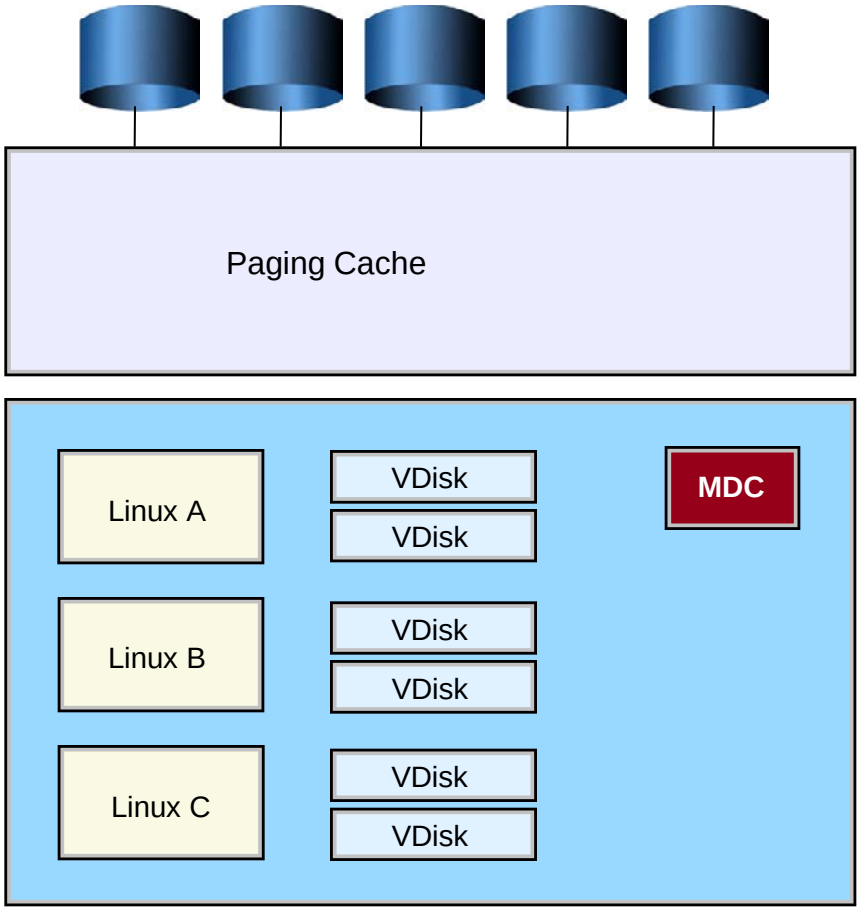
- Utilize features of z/VM – Virtual Disk
- Linux not limited in swap rate,
- z/VM manages storage, high band width



VELOCITY SOFTWARE

z/VM Paging Hierarchy

z/VM paging bandwidth very high, multi-level



Page volumes

Expanded storage

VM storage

Central processor storage

z/VM Paging bandwidth VERY HIGH

Linux Swap bandwidth VERY HIGH

VELOCITY SOFTWARE

Linux Storage Case Study

First case study:

- Process took hours, system paged significantly
- Reduced size of Linux Virtual Machine, 128mb to 24mb
- Defined 100MB Swap disk
- Linux reduces storage requirement
- Process took minutes

Virtual Disk paged out when not in use

- This works!!! Paging greatly reduced, Linux performance greatly improved!!!

This research critical to using Collaborative Memory Mgmt (CMM)

VELOCITY SOFTWARE

LINUX Swapping to VDISK

Change 128MB Server to 24MB with 100MB Swap

Reduction of Overall Storage Requirements of 100MB

- Unused VDISK is paged out

```
Screen: ESAVDSK  Velocity Software, Inc.          ESAMON V2.2  03/15 12:14-
                                                <--pages-->  DASD    X-
                                                Resi- Lock-  Page Store
Time     Owner    Space Name                    dent    ed  Slots  Blks
-------- -------- ------------------------  ----- ----- ----- -----
12:15:01 LINUX001 VDISK$LINUX001$0202$0009     36     0    50     0
12:16:01 LINUX001 VDISK$LINUX001$0202$0009     36     0    50     0
12:17:01 LINUX001 VDISK$LINUX001$0202$0009    173     0    50     0
12:18:01 LINUX001 VDISK$LINUX001$0202$0009    293     0    35     0
12:19:01 LINUX001 VDISK$LINUX001$0202$0009    293     0    35     0
…..
12:39:01 LINUX001 VDISK$LINUX001$0202$0009    259     0    35     0
12:40:01 LINUX001 VDISK$LINUX001$0202$0009    259     0    35     0
12:41:01 LINUX001 VDISK$LINUX001$0202$0009    207     0    86     0
12:42:01 LINUX001 VDISK$LINUX001$0202$0009    207     0    86     0
12:43:01 LINUX001 VDISK$LINUX001$0202$0009     13     0   280     0
12:44:01 LINUX001 VDISK$LINUX001$0202$0009     13     0   280     0
12:45:01 LINUX001 VDISK$LINUX001$0202$0009     13     0   280     0
```

VELOCITY SOFTWARE

Virtual Storage vs Virtual Disk tradeoffs

Virtual Disk I/O 838K / 900 seconds

- About 900 - 1,000 per second
- (NOTE MDISK HIT RATE!!!!)

```
Report: ESAUSR3      User Resource Utilization - Part 2 Domino Redbook ESAMAP 3.4.0
Monitor initialized: on 2066 serial 71CE3 First record analyzed: 08/21/03 12:00:00
--------------------------------------------------------------------------------
               DASD MDisk Virt Cache I/O    <---Virtual Device---->
UserID    DASD Block Cache Disk   Hit Prty   <----I/O Requests----->
/Class     I/O   I/O  Hits  I/O   Pct Queued  Cons   U/R  CTCA Other
-------- ----- ----- ----- ---- ------ ----- ----- ----- ----- -----
08/21/03
12:15:00  613K     0  248K 838K  74.8      0  1510     0   321     0
**Top User Analysis***
LINUXA    610K     0  246K 838K  74.8      0     1     0     0     0
--------------------------------------------------------------------------------
12:30:00  615K     0  250K 822K  74.6      0  1487     0   324     0
**Top User Analysis***
LINUXA    613K     0  248K 822K  74.6      0     0     0     0     0
--------------------------------------------------------------------------------
12:45:00  631K     0  260K 884K  75.5      0  1634     0   321     1
**Top User Analysis***
LINUXA    628K     0  258K 884K  75.5      0     0     0     0     0
```

VELOCITY SOFTWARE

Cost of Swap

Cost of Swap daemon about 10%

```
Report: ESAHSTA   LINUX HOST Application Report    Domino Redbook  ESAMAP 3.4.0 08/25/03
Monitor initialized:    on 2066 serial 71CE3 record analyzed: 08/21/03 12:00:00
-------------------------------------------------------------------------------
 Node/    Process/    <-Application Process Counts------> <-----Processor---->
 Date     Application                                     <---Utilization---->
 Time     name        Total active Running ResWait Loaded Percent seconds Avg
 -------- ----------- ----- ------ ------- ------- ------ ------- ------- ----
 08/21/03
 12:15:00
 LINUXA   java         15.0   15.0     2.0    13.0      0    10.3    92.6  0.7
          kswapd        1.0    1.0       0     1.0      0     9.1    82.2  9.1
          router       11.0   11.0       0    11.0      0    10.6    95.4  1.0
          server       67.0   67.0     1.0    63.0    3.0    63.2   568.5  0.9
          snmpd         1.0    1.0     1.0       0      0     3.3    29.3  3.3
          update        3.0    3.0     1.0     2.0      0    10.2    91.7  3.4
-------------------------------------------------------------------------------
 12:30:00
 LINUXA   java         17.0   17.0     2.0    15.0      0     9.5    85.9  0.6
          kswapd        1.0    1.0       0     1.0      0     8.8    79.5  8.8
          router       12.0   12.0     2.0     9.0    1.0    11.0    99.3  0.9
          server       61.0   61.0     4.0    55.0    2.0    62.7   563.9  1.0
          snmpd         1.0    1.0     1.0       0      0     3.2    28.8  3.2
          update        4.0    4.0       0     4.0      0    12.0   107.8  3.0
-------------------------------------------------------------------------------
 12:45:00
 LINUXA   java         16.0   16.0       0    16.0      0    10.3    92.4  0.6
          kswapd        1.0    1.0       0     1.0      0     9.5    85.6  9.5
          router       10.0   10.0       0    10.0      0    11.1    99.6  1.1
          server       67.0   67.0     9.0    53.0    5.0    64.3   578.6  1.0
          snmpd         1.0    1.0     1.0       0      0     2.4    21.9  2.4
          update        5.0    5.0       0     5.0      0    13.0   116.9  2.6
```

VDISK Case Study

VDisk for swap rules:

- Two small virtual disks for swap, prioritized

Breaking the rules increases storage:

Note vdisk large? WHY???

```
Report: ESASTR1                                  test                              ES
Monitor initialized: 032094 serial 9E14C     First record analyzed: 03/05/08
-------------------------------------------------------------------------------
         Users <-------Pages-----------------------------------------------------
         Loggd System   <Available> Systm  User  NSS/DCSS  <-AddSpace> VDISK
Time        On Storage  <2gb  >2gb  ExSpc Resdnt Resident  Systm User  Rsdnt
-------- ----- -------  ----- ----- ----- ------ --------  ----- -----  -----
03/05/08
02:15:00    28 1310719    802  4377  1124 967698     2950   230K 10866  229K
02:30:00    28 1310719    784  4635  1123 967458     2952   230K 10866  229K
02:45:00    28 1310719    806  3129  1124 967570     2950   230K 10867  229K
03:00:00    28 1310719    815  3669  1124 967954     2949   229K 10868  228K
03:15:00    28 1310719    874  3991  1124 967710     2953   230K 10868  229K
```

VDISK Case Study

VDisk for swap best practice: Two small disks, prioritized

Summary analysis:

Two disks per server, goodness

Should be 1 small swap disk, plus 2nd large disks, goodness

Prioritized backward though, badness….

```
********************************************************************************
                                <--Size--->  <--pages--> -----> DASD    X-
                                AddSpc VDSK  Resi- Lock-  Stg-> Page Store
Owner    Space Name             Pages  Blks   dent    ed T Migr Slots  Blks
-------- ---------------------- ----- -----  ----- ----- - ---- ----- -----
Average:
LINUX1   VDISK$LINUX1$$$0101$0041 65791  8738   3.0     0      0   568     0
LINUX1   VDISK$LINUX1$$$0112$0042  524K 69905   170     0    0.0 61212    11
LINUX2   VDISK$LINUX2$$$0101$0043 65791  8738   3.0     0      0   571     0
LINUX2   VDISK$LINUX2$$$0112$0044  524K 69905   85K     0    0.4  346K  2047
LINUX3   VDISK$LINUX3$$$0101$0045 65791  8738   3.0     0      0   571     0
LINUX3   VDISK$LINUX3$$$0112$0046  524K 69905   2.0     0      0  5767     0
LINUX4   VDISK$LINUX4$$$0101$0047 65791  8738   3.0     0      0   571     0
LINUX4   VDISK$LINUX4$$$0112$0048  524K 69905  147K     0    0.3  223K 35967
LINUX5   VDISK$LINUX5$$$0101$0049 65791  8738   3.0     0      0   568     0
LINUX5   VDISK$LINUX5$$$0112$004A  524K 69905   2.0     0      0  4321     0
LINUX6   VDISK$LINUX6$$$0101$004B 65791  8738   3.0     0      0   571     0
LINUX6   VDISK$LINUX6$$$0112$004C  524K 69905   771     0      0  5666   603
. . . . . .
                                ----- -----  ----- ----- - ---- ----- -----
System Totals:                  5901K 39321  233K     0    0.7  669K 38631
```

Additional Storage Performance

Named Saved System

- Fast IPL, shared kernel storage
- Saves 1mb per server, difficult to implement

DCSS with XIP File System

- Load all programs into shared DCSS,
- Saves 20-100mb/server, easy to implement

CMM: Collaborative memory management

- Dynamically manage storage size
- Saves GB/server, requires feedback

VELOCITY SOFTWARE

How many Virtual Processors?

- Linux is multiprocessor capable
- Global lock is large issue on older Linux
  - One processor acquires lock
  - Other processors attempt to spin
  - On 390 – spin converted to Diagnose 44 (now 9C)

- Problem easily detected
  - High Diagnose -> Instruction Simulation -> SIE
  - High TV ratio
  - Guideline: Minimize virtual processors

How many Virtual Processors?

```
Report: ESACPUA      CPU Utilization Analysis
-------------------------------------------------------------
            <CPU percents><--Internal (per second)--> SIGP
             Totl Ovrhead Diag Inst    SIE Fast  Page Rate
Time     CPU Util Usr Sys nose  Sim intrcp path fault /sec
-------- --- ---- --- --- ---- ---- ------ ---- ----- ----

16:01:00  0  66.6  12  25  80K  82K  83275 2108   0.1  350
          1  67.6  12  25  89K  91K  91879 1051     0  332
          2  62.3  12  24  83K  85K  85768 1219   0.1  383
          3  62.7  11  25  77K  78K  79354  776     0  293
          4  63.6  12  24  84K  85K  86175 1047   0.0  329
          5  63.1  11  26  82K  84K  85064 1188   0.0  297
          6  64.1  11  22  83K  84K  84874 1079   0.0  304
          7  57.3  10  22  73K  75K  75481 1044   0.0  323
          8  62.7  10  26  53K  57K  58761 1421   0.1  267
             ---- --- --- ---- ---- ------ ---- ----- ----
System:       570 101 218 704K 723K 730630  11K   0.2 2879
```

- CPU Performance typical of many Linux Apps:
  - High Diagnose -> Instruction Simulation -> SIE
  - z/VM 5.2 modifies logic

Mainframe I/O Expectation Issues

Mainframe I/O expectations OFTEN wrong

- I/O traditionally tuned to operate within limitations
- Separate I/O processors
- Competition not limited by ESCON channel speeds

Customer says "FTP on Linux under z/VM is slow"

- Benchmark was large FTP, problem NOT network
- Escon channels, 30ms "CONNECT" time
- 500K transfers

Questions:

- How fast are ESCON channels? FICON channels? Ficon Express?
- How fast are SCSI disks on other platforms?

PAV?

- What are options when high utilization on shared disks?
- PAV Available z/VM 5.2 - **Use for high activity shared devices ONLY**

VELOCITY SOFTWARE

FTP Benchmarks: Results NOT intuitive

- Benchmark 1: (G5 processor)
  - FTP through Linux router with OSA dedicated to Virtual Router
  - FTP to Linux on single (ESCON) device
  - **Throughput limited to 4mb / second: why?**
- Benchmark 2:
  - Eliminate router, dedicate OSA,
  - **Throughput increased to 8mb / second**, why?
  - **Guideline: Use dedicated OSA for high bandwidth**
- Benchmark 3:
  - Switch to LVM striped over 2 devices
  - Throughput reduced to 7mb / second, why?
  - **Guideline: Evaluate carefully use of striped LVM**
- **Answer to all questions: CPU was limiter, get a z196**

VELOCITY SOFTWARE

FTP Benchmarks: Results NOT intuitive

Compare Linux Asynchronous I/O vs synchronous I/O

- Asynchronous is default
- Synchronous writes data without buffering
- DASD response time
  - Asynchronous: 50ms (6 I/O / second, 512k / IO),
  - Synchronous: 1.5ms (300 I/O / second, 4k / IO)
- Which is better throughput?
- **Guideline: Use Asynchronous?**
- DASD Response time rot don't work

- **Guideline: Fight for FICON/(express)!!!!**

VELOCITY SOFTWARE

Benefit of MDC

```
Report: ESAMDC       Minidisk Cache Analysis       Domino Redbook      ESAMAP 3.4.0 08/25/03  Page 211
Monitor initialized:                 on 2066 serial 71CE3     First record analyzed: 08/24/03 18:00:00
------------------------------------------------------------------------------------------------------
         <----Load---->     <IO per><Insertions> <-----Main Storage MDC--> <-Expanded Storage MDC------>
         <-Users-> Tran Hit <second> Usr Per Not <-Sizes (MB)--> </Second> <-Sizes (MB)--> <Per Second >
Time     Actv In Q /sec Pct rds hits Max Min Ald Avg MIN MAX Obj Stls Delt Avg MIN MAX Obj Rds Wrts Stls
-------- ---- ---- ---- --- --- ---- --- --- --- --- --- --- --- ---- ---- --- --- --- --- --- ---- ----
08/24/03
19:00:00   20  7.9  4.0  48 376  181 17K 6.4   0 380   0  2K  1K    0    0  2K   0  2K  2K  2K 47.0  4.7
20:00:00   20  8.1  4.0  47 370  176 17K 6.4   0 373   0  2K  1K    0    0  2K   0  2K  2K  2K 46.0  4.9
21:00:00   20  7.7  3.9  48 377  179 16K 6.6   0 375   0  2K  1K    0    0  2K   0  2K  2K  2K 46.5  5.0
**********************************Summary*****************************************************
Average:   20  7.9  4.0  48 374  178 17K 6.5   0 376   0  2K  1K    0    0  2K   0  2K  2K  2K 46.5  4.9
```

- MINIDISK Cache very good for Linux under z/VM
  - **For SHARED disks**
- As servers get smaller, MDC takes over caching
  - Hit rate (48%) means 48% of MDC eligible I/O avoided
  - Use diagnose driver, record cache to best utilize MDC

VELOCITY SOFTWARE

MDC Pains

- CP algorithms VERY poor at sizing MDC Storage
- Control the size of MDC!

```
Report: ESAMDC       Minidisk Cache Analysis                          .     ESAMAP 3.6.1 02/08/07  Pg  2660
Monitor initialized: 02/07/07 at 00:00:05 on 2084 serial 447AA     First record analyzed: 02/07/07 00:00:05
------------------------------------------------------------------------------------------------------------------------
         <----Load---->     <IO per><Insertions> <-----Main Storage MDC--> <-Expanded Storage MDC------> <External>
         <-Users-> Tran Hit <second> Usr Per Not <-Sizes (MB)--> </Second> <-Sizes (MB)--> <Per Second > <I/O rate>
Time     Actv In Q /sec Pct rds hits Max Min Ald Avg MIN MAX Obj Stls Delt Avg MIN MAX Obj Rds Wrts Stls Pages DASD
-------- ---- ---- ---- --- --- ---- --- --- --- --- --- --- --- ---- ---- --- --- --- --- --- ---- ---- ----- ----

12:20:00   26 18.7  2.2  63  33 20.4  8K 7.5   0  2K   0  8K  2K  0.1  180  1K   0  3K  1K  55    0  0.1   253  261
12:35:00   26 19.1  2.1  63 8.5  5.4 10K 5.8   0  2K   0  8K  2K  0.0 69.9  1K   0  3K  1K  10    0  0.0    53  185
12:50:00   26 18.3  2.0  69 6.0  4.2 11K 4.7   0  1K   0  8K  2K  0.0 43.6  1K   0  3K  1K  12    0  0.0    33  167
13:05:00   27 19.5  2.2  38  29 11.0 12K 5.2 0.4  2K   0  8K  2K  1.2 1062  1K   0  3K  2K  63  0.0  1.3   571  406
13:26:00   31 17.4  1.7  28  28  8.0 14K  12 0.7  4K   0  8K  4K  2.8 1324 272   0  3K  2K 3.7  0.0  4.5  1090  356
13:41:00   25 19.9  2.9  69  60 41.5 14K 7.5   0  3K   0  8K  3K  0.5  483 727   0  3K  2K 2.0    0  0.2   742  422
```

Guidelines:

- SET MDC STORAGE 128M **128M**
- SET MDC XSTORE 0M 0M | OFF

VELOCITY SOFTWARE

z/VM Page Space

Overcommitting real storage is good, reduces cost

- Back up is Paging storage

If 40GB main storage

- Overcommit factor of 2 - How much paging storage needed?
- VM installations often very underconfigured
- **Guideline: Paging storage should still be 2 times requirement**

Number of paging devices? Number of channels?

- ROT not valid

Lack of page space planning is top reason for first installation z/VM outage

VELOCITY SOFTWARE

Expanded Storage

Expanded Storage required for paging performance

- True LRU, with page migration
- Page the correct pages
- Page rates to disk drop when converting real storage only system to real+expanded

How much expanded?

- Enough for 30 second window
- **Guideline: Enough so STEAL does not page to disk**
- 20% a good target, usually enough – NO arbitrary LIMITS!!!
- Measure on ESABLKP
- **MEASURE BLOCK SIZE!!!!**

VELOCITY SOFTWARE

Original problem documented in 1992

- If VTAM has (recommended) REL SHARE 10000, looping user consumed CPU
- If VTAM had ABS 5%, looping user constrained
- Velocity recommended ABS shares for critical servers

Creating EXCESS SHARE

- Setting SHARE to 10000 (compare 100 servers at REL 100)
- Linux servers that are idle, but inqueue servers count
- NOTE: VMRM often used SET REL 10000

Excess Share Analysis

Starting with 3 looping users REL 100.

- They all get equal share of the resources
- this is as we expected.

```
 Screen: ESAUSP2  Velocity Software-Test VSIVM4  ESAMON 3.778
 1 of 3  User Percent Utilization                CLASS * USER
                              <-------Main Storage-------->
         UserID   <Processor> <Resident->  Lock <-WSSize-->
Time     /Class   Total  Virt Total  Actv   -ed Total  Actv
-------- -------- ----- ----- ----- ----- ----- ----- -----
00:11:00 ROBLNX1  32.39 32.38 15862 15862    11 15536 15536
         ROBLX2   32.12 32.11 66136 66136   259 78478 78478
         ROBLX1   32.02 32.01 38219 38219   176 37790 37790
         ROB2LV    0.01  0.00  2246  2246     0  2246  2246
```

Excess Share Analysis

We now give ROBLX2 a REL 200

- because that is a more important service
- (nothing with virtual 2-way).
- Not as expected, it gets the excess share

```
 Screen: ESAUSP2  Velocity Software-Test VSIVM4  ESAMON 3.778
1 of 3  User Percent Utilization                  CLASS * USER
                                <-------Main Storage-------->
         UserID   <Processor> <Resident->  Lock <-WSSize-->
Time     /Class   Total  Virt Total  Actv   -ed Total  Actv
-------- -------- ----- ----- ----- ----- ----- ----- -----
00:14:00 ROBLX2   68.71 68.68 66211 66211   258 78478 78478
         ROBLX1   14.00 14.00 38245 38245   256 37790 37790
         ROBLNX1  13.99 13.99 15879 15879    11 15536 15536
         ROB2LV    0.01  0.00  2246  2246     0  2246  2246
```

VELOCITY SOFTWARE

Excess Share Analysis

Now for the experiment

- we reduce the relative share for all idle users down to 1
- (using the allocated share computation below and showing how much allocated / consumed share is).
- This ELIMINATES “EXCESS” bucket

```
 Screen: ESAUSP2  Velocity Software-Test VSIVM4  ESAMON 3.778
1 of 3  User Percent Utilization                    CLASS * USER
                                 <-------Main Storage-------->
         UserID   <Processor> <Resident->  Lock <-WSSize-->
Time     /Class   Total  Virt Total  Actv   -ed Total  Actv
-------- -------- ----- ----- ----- ----- ----- ----- -----
00:20:00 ROBLX2   48.39 48.37 67141 67141   292 80047 80047
         ROBLNX1  24.19 24.19 16168 16168    11 15536 15536
         ROBLX1   24.19 24.18 39006 39006   241 37790 37790
         ROB2LV    0.01  0.00  2246  2246     0  2246  2246
```

Excess Share Analysis

And when we set ROBLNX1 to REL 300

- it works again: 48% 32% and 16%
- exactly like the REL 300, 200 and 100 we set.

```
Screen: ESAUSP2  Velocity Software-Test VSIVM4  ESAMON 3.778
1 of 3  User Percent Utilization                CLASS * USER
                                <-------Main Storage-------->
         UserID   <Processor> <Resident->  Lock <-WSSize-->
Time     /Class   Total  Virt Total  Actv   -ed Total  Actv
-------- -------- ----- ----- ----- ----- ----- ----- -----
00:23:00 ROBLNX1  48.15 48.14 16170 16170    11 15536 15536
         ROBLX2   32.86 32.86 67190 67190   211 80047 80047
         ROBLX1   16.44 16.43 39016 39016   193 37790 37790
         ROB2LV    0.01  0.00  1680  1680     0  1680  1680
```

VELOCITY SOFTWARE

z/VM System Parameters

SET SHARE

- Use RELATIVE 100 for single virtual CPU
- Use RELATIVE 200 for two virtual CPU
- Use ABSOLUTE for shared or critical resource servers

SET SRM STORBUF – allow overcommit

- SET SRM STORBUF 350 300 300

SET SRM LDUBUF – DO NOT allow overcommit

- You can NOT run paging devices at more than 100% busy!!!
- SET SRM LDUBUF 100 80 60 (or lower)

SET QUICKDSP

- **Use for only absolutely critical servers**

VELOCITY SOFTWARE

Infrastructure Requirements

Requirements:

- **Performance management**
- **Capacity planning**
- **Chargeback**
- **Operations**

Shared resource environment:

- Avoid unnecessary work
- Avoid “waking up Linux”

Availability Monitoring – necessary?

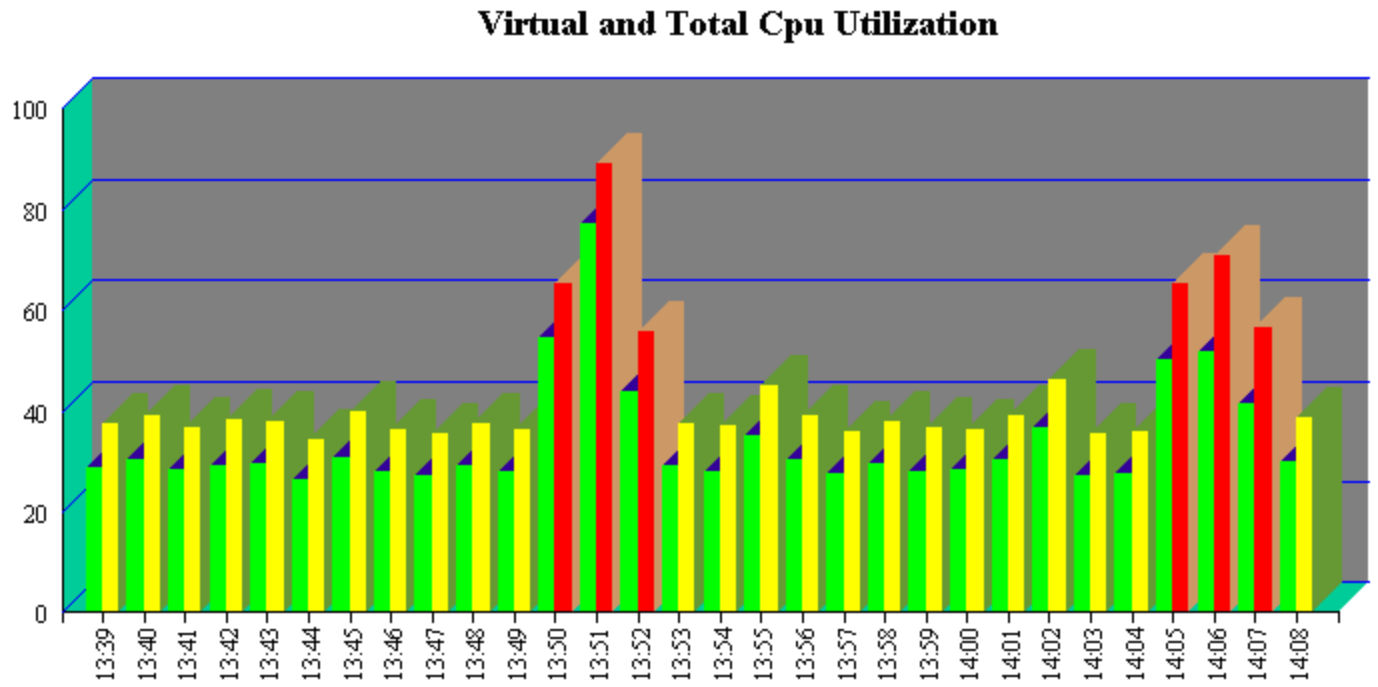
High Availability – cost? (DB2, RAC)

Using Encryption - necessary if on virtual lan?

Measure your infrastructure and determine scalability!

VELOCITY SOFTWARE

Infrastructure: SOP Valid?



Question:

- Why always hit every 15 minutes?

SOP: Standard Operating Procedure???

VELOCITY SOFTWARE

Detect and alert looping processes

```
Report: ESAHST1      LINUX HOST Software Analysis Report
Monitor initialized:                        on 2066 serial 71CE3
-------------------------------------------------------------------------
Node/    <-----Software Program-------> <CPU Seconds> CPU  StgSize
Time     Name      ID    Type   Status  Total Intrval Pct  (Bytes)
-------- --------  ----  ------ ------- ----- ------- ---- -------
08:32:00
LINUXA
         init         1  Applic ResWait   0.9     0.0  0.0   61440
         kjournal    95  Applic ResWait   2.5     0.0  0.0       0
         db2fmd     596  Applic ResWait   0.3     0.0  0.0  573440
         sshd      1081  Applic ResWait   0.4     0.0  0.0  204800
         event    10787  Applic ResWait  19.5     0.0  0.0  11188K
         snmpd    10861  Applic Running 193.4     4.2  7.1   1492K
         adminp   11452  Applic ResWait  58.5     0.0  0.1  13848K
         server   11525  Applic ResWait   1.0     0.1  0.1  35720K
         server   11533  Applic ResWait   4.3     0.0  0.0  35720K
         server   11537  Applic Running 44697    58.3 99.2  35720K
         java     13024  Applic ResWait   0.0     0.0  0.0   6632K
         java     24016  Applic ResWait   1.9     0.0  0.0   6632K
         java     24024  Applic ResWait   4.9     0.0  0.0   6632K
         server   24192  Applic ResWait  19.0     0.1  0.1  35720K
         java     26352  Applic ResWait   0.4     0.0  0.0   7320K
         sshd     26477  Applic ResWait   0.2     0.0  0.1   2028K
```

Show process by ID

- Status
- Total CPU
- Percent CPU
- Storage

(Non-velocity mib)

Performance Instrumentation

Performance Instrumentation

- Cost of instrumentation often excessive
- “Native Linux” tools will not detect many problems
- Agents may take 5-10% of a processor (Per server)

Cost of instrumentation should be < .1% (of ONE CPU) per server

- **Performance instrumentation should not change performance**

Active agents vs Passive agents

- Active agent wakes up at constant interval and records data
- Passive agent only responds to external request

Dynamically turn off monitoring of idle servers!!!!

- If z/VM data shows server is idle, should agent wake up to find out what is running?

VELOCITY SOFTWARE

Linux Configuration Summary

Virtual machine size

- Minimize until some swap

Swapping

- Swap to virtual disk
- Define 2 virtual disks,
  - One to meet the average requirement
  - Second one for overflow
- Use DIAG driver instead of FBA
  - Reduces I/O by factor of 8

Virtual processors

- Minimize to meet the workload/application requirement

Infrastructure costs

- Minimize – shared resource architecture

VELOCITY SOFTWARE

z/VM Subsystem Configuration

DASD Channels

- ESCON channels are 17MByte / second
- Ficon channels 100MB, Ficon Express 200MB
- Ficon compares to SCSI disks on other platforms

Paging

- How much paging is required to support 2 times over commitment of 40GB z/VM system?

MDC

- Caches data – read-ahead, often used data
- Default too high
- SET MDC STORAGE 128M 128M
- SET MDC XSTORE 0M 0M | OFF

VELOCITY SOFTWARE