#SHAREorg

SHARE
Technology · Connections · Results

Dynamic Features of Linux on System z

Richard Young
IBM STG Lab Services

Tuesday, August 7, 2012
09:30

Session 11827

SHARE in Anaheim 2012

Trademarks & Disclaimer

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DB2, e-business logo, ESCON, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/390, System Storage, System z9, VM/ESA, VSE/ESA, WebSphere, xSeries, z/OS, zSeries, z/VM.

The following are trademarks or registered trademarks of other companies

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries. LINUX is a registered trademark of Linux Torvalds in the United States and other countries. UNIX is a registered trademark of The Open Group in the United States and other countries. Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation. SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC. Intel is a registered trademark of Intel Corporation. * All other products may be trademarks or registered trademarks of their respective companies.

NOTES: Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply. All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions. This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography. References in this document to IBM products or services do not imply that IBM intends to make them available in every country. Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use. The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice. Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

Agenda

1. Value of Dynamic Resource Configuration
2. Dynamically Adding Disk Storage
3. Dynamically Adding Network Interfaces
4. Adding Hotplug Memory
5. Adding CPUs
6. Automatically Adjusting Memory and CPU Resources

Agenda

1. **Value of Dynamic Resource Configuration**
2. Dynamically Adding Disk Storage
3. Dynamically Adding Network Interfaces
4. Adding Hotplug Memory
5. Adding CPUs
6. Automatically Adjusting Memory and CPU Resources

Dynamic Resource Configuration

- Helps to avoid Linux guest restarts and potential outage/downtime resource allocation changes
- Accommodate unplanned increases in application workload demands or application “enhancements” that consume more resource
- It can allow for more efficient overall hypervisor operation (reduced operational overhead)
- Automated policy based reconfiguration more responsive than manual adjustments.

Agenda

1. Value of Dynamic Resource Configuration
2. **Dynamically Adding Disk Storage**
3. Dynamically Adding Network Interfaces
4. Adding Hotplug Memory
5. Adding CPUs
6. Automatically Adjusting Memory and CPU Resources

Dynamically Adding Disk Resources

- Disk Resource Types
  - ECKD
    - Full Volume
    - z/VM Minidisk
  - SCSI Luns
    - Via z/VM Emulated Device
    - Via Dedicated FCP Device
- All types can be dynamically added
- General Process
  - Add resource from hypervisor
  - Make new resource available
  - Bring virtual device online
  - Provision as usual

Dynamically Adding Disk Resources

```
dirm for rgylxws8 amdisk 201 3390 autog 3338 LINUX  MR
DVHXMT1191I Your AMDISK request has been sent for processing to DIRMAINT
DVHXMT1191I at POKLBS1.
Ready; T=0.01/0.02 19:16:54
 DVHREQ2288I Your AMDISK request for RGYLXWS8 at * has been accepted.
 DVHSCU3541I Work unit 15191655 has been built and queued f r processing.
 DVHSHN3541I Processing work unit 15191655 as RYOUNG1 from P KLBS1,
 DVHSHN3541I notifying RYOUNG1 at POKLBS1, request 614 for R LXWS8 SSI
 DVHSHN3541I node *; to: AMDISK 0201 3390 AUTOG 3338 LINUX M
 DVHBIU3450I The source for directory entry RGYLXWS8 has bee dated.
 DVHBIU3424I The next ONLINE will take place immediately.
 DVHDRC3451I The next ONLINE will take place via delta object ctory.
 DVHRLA3891I Your DSATCTL request has been relayed for process
 DVHBIU3428I Changes made to directory entry RGYLXWS8 have bee ed
 DVHBIU3428I online.
 DVHSHN3430I AMDISK operation for RGYLXWS8 address 0201 has fin
 DVHSHN3430I (WUCF 15191655).
 DVHREQ2289I Your AMDISK request for RGYLXWS8 at * has complete RC
 DVHREQ2289I = 0.
 DVHREQ2288I Your DSATCTL request for DIRMAINT at
 DVHREQ2288I * has been accepted.
 DVHREQ2289I Your DSATCTL request for DIRMAINT at
_
MA    B
```

- DIRM add minidisk disk shown
- Could be full volume or partial volume
- Disk could be added via a dedicate as well
- If not using dirmaint, edit user direct and DIRECTXA

Dynamically Adding Disk Resources

```
RGYLXWS8:/ # lscss
Device   Subchan.  DevType CU Type Use  PIM PAM POM  CHPIDs
-----------------------------------------------------------------------
0.0.1000 0.0.0000  1732/03 1731/03      80  80  ff   c4000000 00000000
0.0.1001 0.0.0001  1732/03 1731/03      80  80  ff   d1000000 00000000
0.0.1002 0.0.0002  1732/03 1731/03 yes  80  80  ff   c9000000 00000000
0.0.1003 0.0.0003  1732/03 1731/03 yes  80  80  ff   dd000000 00000000
0.0.0191 0.0.0004  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.0200 0.0.0006  3390/0c 3990/e9 yes  80  80  ff   ff000000 00000000
0.0.0192 0.0.0007  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.0009 0.0.0008  0000/00 3215/00 yes  80  80  ff   ff000000 00000000
0.0.0600 0.0.0009  1732/01 1731/01 yes  80  80  ff   00000000 00000000
0.0.0601 0.0.000a  1732/01 1731/01 yes  80  80  ff   00000000 00000000
0.0.0602 0.0.000b  1732/01 1731/01 yes  80  80  ff   00000000 00000000
0.0.000c 0.0.000c  0000/00 2540/00      80  80  ff   ff000000 00000000
0.0.000d 0.0.000d  0000/00 2540/00      80  80  ff   ff000000 00000000
0.0.000e 0.0.000e  0000/00 1403/00      80  80  ff   ff000000 00000000
0.0.0190 0.0.000f  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.019d 0.0.0010  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.019e 0.0.0011  3390/0c 3990/e9      80  80  ff   ff000000 00000000
RGYLXWS8:/ # vmcp q v dasd
DASD 0190 3390 P01RES R/O        214 CYL ON DASD  3F27 SUBCHANNEL = 000F
DASD 0191 3390 VM1US1 R/O        500 CYL ON DASD  3F10 SUBCHANNEL = 0004
DASD 0192 3390 LS3F18 R/W         50 CYL ON DASD  3F18 SUBCHANNEL = 0007
DASD 019D 3390 P01RES R/O        292 CYL ON DASD  3F27 SUBCHANNEL = 0010
DASD 019E 3390 P01RES R/O        500 CYL ON DASD  3F27 SUBCHANNEL = 0011
DASD 0200 3390 LS3F52 R/W      10015 CYL ON DASD  3F52 SUBCHANNEL = 0006
RGYLXWS8:/ #
```

- 201 minidisk still not available to Linux and not shown from a z/VM query virtual
- New storage must be `attached` or `linked` before it can be brought online

Dynamically Adding Disk Resources

```
RGYLXWS8:/ # vmcp link RGYLXWS8 201 201 MR
RGYLXWS8:/ # vmcp q v dasd
DASD 0190 3390 P01RES R/O        214 CYL ON DASD  3F27 SUBCHANNEL = 000F
DASD 0191 3390 VM1US1 R/O        500 CYL ON DASD  3F10 SUBCHANNEL = 0004
DASD 0192 3390 LS3F18 R/W         50 CYL ON DASD  3F18 SUBCHANNEL = 0007
DASD 019D 3390 P01RES R/O        292 CYL ON DASD  3F27 SUBCHANNEL = 0010
DASD 019E 3390 P01RES R/O        500 CYL ON DASD  3F27 SUBCHANNEL = 0011
DASD 0200 3390 LS3F52 R/W      10015 CYL ON DASD  3F52 SUBCHANNEL = 0006
DASD 0201 3390 LS3F18 R/W       3338 CYL ON DASD  3F18 SUBCHANNEL = 0005
RGYLXWS8:/ # lscss
Device   Subchan.  DevType CU Type Use  PIM PAM POM  CHPIDs
----------------------------------------------------------------------
0.0.1000 0.0.0000  1732/03 1731/03      80  80  ff   c4000000 00000000
0.0.1001 0.0.0001  1732/03 1731/03      80  80  ff   d1000000 00000000
0.0.1002 0.0.0002  1732/03 1731/03 yes  80  80  ff   c9000000 00000000
0.0.1003 0.0.0003  1732/03 1731/03 yes  80  80  ff   dd000000 00000000
0.0.0191 0.0.0004  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.0201 0.0.0005  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.0200 0.0.0006  3390/0c 3990/e9 yes  80  80  ff   ff000000 00000000
0.0.0192 0.0.0007  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.0009 0.0.0008  0000/00 3215/00 yes  80  80  ff   ff000000 00000000
0.0.0600 0.0.0009  1732/01 1731/01 yes  80  80  ff   00000000 00000000
0.0.0601 0.0.000a  1732/01 1731/01 yes  80  80  ff   00000000 00000000
0.0.0602 0.0.000b  1732/01 1731/01 yes  80  80  ff   00000000 00000000
0.0.000c 0.0.000c  0000/00 2540/00      80  80  ff   ff000000 00000000
0.0.000d 0.0.000d  0000/00 2540/00      80  80  ff   ff000000 00000000
0.0.000e 0.0.000e  0000/00 1403/00      80  80  ff   ff000000 00000000
0.0.0190 0.0.000f  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.019d 0.0.0010  3390/0c 3990/e9      80  80  ff   ff000000 00000000
0.0.019e 0.0.0011  3390/0c 3990/e9      80  80  ff   ff000000 00000000
RGYLXWS8:/ # chccwdev -e 201
Setting device 0.0.0201 online
Done
```

A z/VM "`link`" makes device available.

Can be performed from Linux via '`vmcp`"

Must still be brought online via "`chccwdev`"

Dynamically Adding Disk Resources

```
RGYLXWS8:/ # lsdasd
Bus-ID     Status      Name      Device  Type  BlkSz  Size      Blocks
==============================================================================
0.0.0200   active      dasda     94:0    ECKD  4096   7041MB    1802700
0.0.0201   active      dasdb     94:4    ECKD  4096   2347MB    600840

RGYLXWS8:/ # dasdfmt -b 4096  -f /dev/dasdb
Drive Geometry: 3338 Cylinders * 15 Heads =  50070 Tracks
I am going to format the device /dev/dasdb in the following way:
   Device number of device : 0x201
   Labelling device        : yes
   Disk label              : VOL1
   Disk identifier         : 0X0201
   Extent start (trk no)   : 0
   Extent end (trk no)     : 50069
   Compatible Disk Layout  : yes
   Blocksize               : 4096
--->> ATTENTION! <<---
All data of that device will be lost.
Type "yes" to continue, no will leave the disk untouched:
```

Dynamically Adding Disk Resources

```
RGYLXWS8:/ # fdasd -a /dev/dasdb
reading volume label ..: VOL1
reading vtoc ..........: ok
auto-creating one partition for the whole disk...
writing volume label...
writing VTOC...
rereading partition table...
RGYLXWS8:/ #
```

- Disk storage has been dynamically brought online, formatted, and partitioned
- Put file system on new device
  - `mkfs -t ext3 -c /dev/dasdb1`
- You could now add to a volume group and LVM to dynamically expand a filesystem without bring the Linux system down
  - `pvcreate /dev/dasdb1`
  - `vgextend VG00 /dev/dasdb1`
  - `lvextend -L+1G /dev/VG00/LV01 ; add one more GB to LV`
  - `ext2online /dev/VG00/LV01`
  - `resize2fs /dev/VG00/LV01`

Agenda

1. Value of Dynamic Resource Configuration
2. Dynamically Adding Disk Storage
3. **Dynamically Adding Network Interfaces**
4. Adding Hotplug Memory
5. Adding CPUs
6. Automatically Adjusting Memory and CPU Resources

Dynamically Adding Network Interfaces

- Much like dynamically adding disk resources a directory alone does not make the NIC available to Linux.
- Once the NIC is defined there are multiple ways to configure it and some methods vary by distro.
- Care and planning should be taking when adding additional NIC. When adding a new NIC mistakes can cause outages on existing functioning NICs in the same guest.

Dynamically Adding Network Interfaces

```
RGYLXWS8:~ # lsqeth
Device name                   : eth0
--------------------------------------------------
        card_type             : GuestLAN QDIO
        cdev0                 : 0.0.0600
        cdev1                 : 0.0.0601
        cdev2                 : 0.0.0602
        chpid                 : 00
        online                : 1
        portname              : NET172A
        portno                : 0
        state                 : UP (LAN ONLINE)
        priority_queueing     : always queue 2
        buffer_count          : 64
        layer2                : 1
        isolation             : none

RGYLXWS8:~ # znetconf -c
Device IDs                  Type    Card Type      CHPID Drv. Name        State
--------------------------------------------------------------------------------
0.0.0600,0.0.0601,0.0.0602 1731/01 GuestLAN QDIO     00 qeth eth0        online
RGYLXWS8:~ #
```

- This system that has only one NIC and a second NIC will be added

Dynamically Adding Network Interfaces

- New NIC added to the zVM user directory
  - Virtual device 700
  - Type QDIO
  - VSWITCH NET172B

```
dirm for rgylxws8 NICDEF 0700 TYPE QDIO DEV 3 LAN SYSTEM NET172B
DVHXMT1191I Your NICDEF request has been sent for processing to DIRMAINT
DVHXMT1191I at POKLBS1.
Ready; T=0.01/0.02 01:43:35
 DVHREQ2288I Your NICDEF request for RGYLXWS8 at * has been accepted.
 DVHBIU3450I The source for directory entry RGYLXWS8 has been updated.
 DVHBIU3424I The next ONLINE will take place immediately.
 DVHDRC3451I The next ONLINE will take place via delta object directory.
 DVHRLA3891I Your DSATCTL request has been relayed for processing.
 DVHBIU3428I Changes made to directory entry RGYLXWS8 have been placed
 DVHBIU3428I online.
 DVHREQ2289I Your NICDEF request for RGYLXWS8 at * has completed; with RC
 DVHREQ2289I = 0.
 DVHREQ2288I Your DSATCTL request for DIRMAINT at
 DVHREQ2288I * has been accepted.
 DVHREQ2289I Your DSATCTL request for DIRMAINT at
 DVHREQ2289I * has completed; with RC = 0.
```

Dynamically Adding Network Interfaces

- "DEFINE NIC" issued to make the new virtual NIC available to the guest
- Since it was already defined in the user directory it automatically coupled to its virtual switch
- znetconf now shows the new virtual NIC
- Since the NIC is yet unconfigured, it is still offline

```
RGYLXWS8:~ # vmcp define nic  0700 TYPE QDIO DEV 3
NIC 0700 is created; devices 0700-0702 defined
RGYLXWS8:~ # vmcp couple 700 to system net172b
HCPCPL2788E NIC 0700 not connected; already connected to VSWITCH SYSTEM NET172B
Error: non-zero CP response for command 'COUPLE 700 TO SYSTEM NET172B': #2788
RGYLXWS8:~ # znetconf -u
Scanning for network devices...
Device IDs                 Type    Card Type      CHPID Drv.
------------------------------------------------------------
0.0.0700,0.0.0701,0.0.0702 1731/01 OSA (QDIO)        01 qeth
RGYLXWS8:~ # znetconf -c
Device IDs                 Type    Card Type      CHPID Drv. Name        State
----------------------------------------------------------------------------
0.0.0600,0.0.0601,0.0.0602 1731/01 GuestLAN QDIO     00 qeth eth0        online
RGYLXWS8:~ #
```

Dynamically Adding Network Interfaces

- We could use tools such as Yast, netconfig, or redhat-config-network to configure the interface, but we will use znetconf from s390-tools
- znetconf allows you to configure many different possible attributes of the QDIO device
- Note: znetconf does not create a udev entry
- After executing znetconf the device (not the interface) will be online

```
RGYLXWS8:~ # znetconf -a 0700 -o layer2=1
Scanning for network devices...
Successfully configured device 0.0.0700 (eth1)
```

Dynamically Adding Network Interfaces

- To bring the network interface online you need to create an ifcfg-eth*x* script
- If you copy an existing file (such as ifcfg-eth0) you should have only two changes to make
  - IPADDR
  - _nm_name
- It is highly recommended to put a udev entry in place (/etc/udev/rules.d) so you have a persistent configuration across reboots

```
BOOTPROTO="static"
UNIQUE=""
STARTMODE="onboot"
IPADDR="172.110.100.38“
NETMASK="255.255.255.0"
NETWORK="172.110.100.0"
BROADCAST="172.110.100.255"
_nm_name='qeth-bus-ccw-0.0.0700‘
```

Dynamically Adding Network Interfaces

- You can activate your new configuration with `rcnetwork restart`
- If your new interface configuration breaks your existing network, logon to the 3270 console for the guest and move the ifcfg-eth*x* script to another directory and reissue your `rcnetwork restart` command.

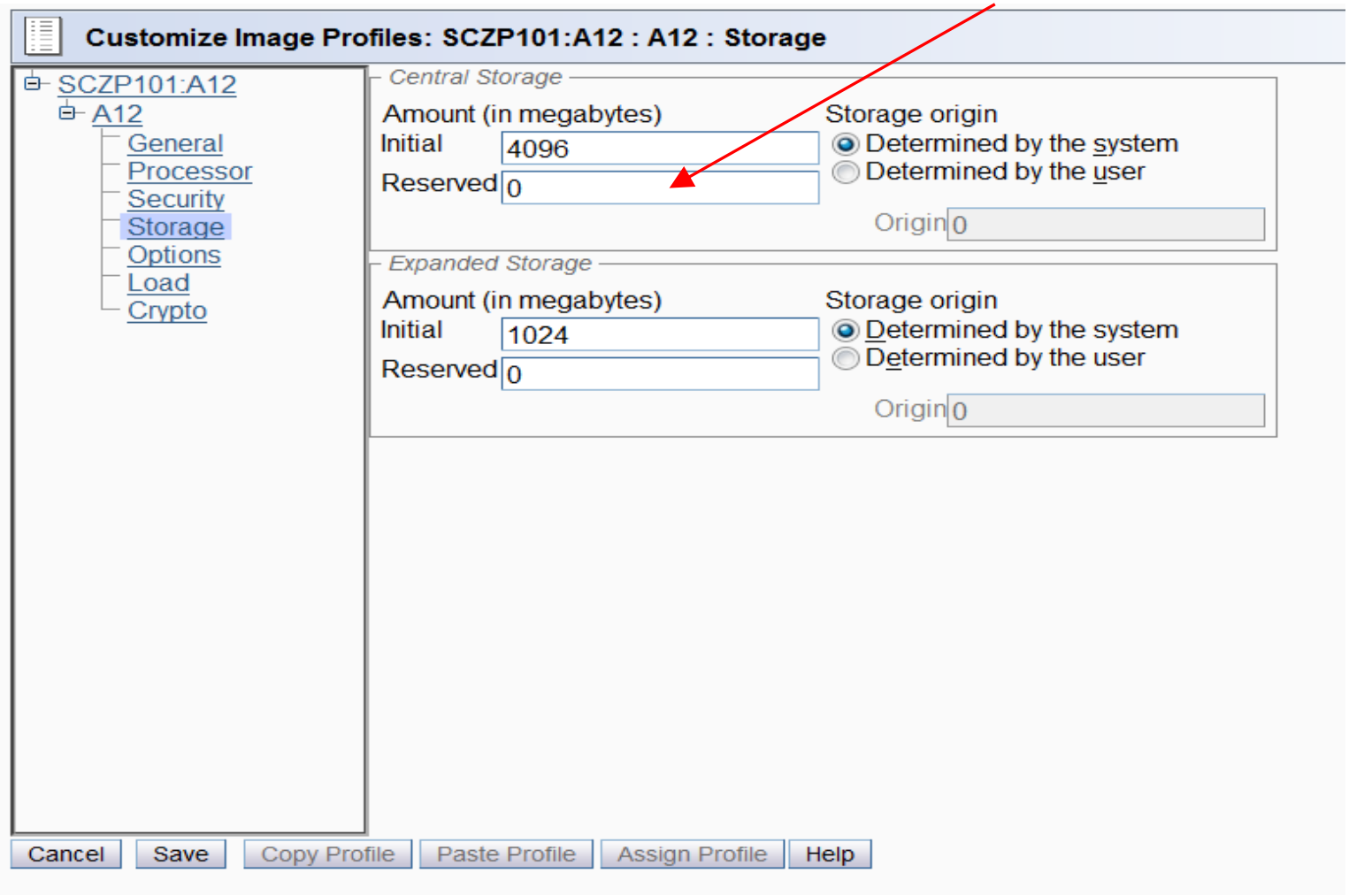
Agenda

1. Value of Dynamic Resource Configuration
2. Dynamically Adding Disk Storage
3. Dynamically Adding Network Interfaces
4. **Adding Hotplug Memory**
5. Adding CPUs
6. Automatically Adjusting Memory and CPU Resources

Hotplug Memory

- You can dynamically increase/decrease the memory for your running Linux guest system, making your penguins elastic.
- To make memory available as hotplug memory you must define it to your LPAR or z/VM **BEFORE** you IPL Linux.
- Hotplug memory is supported by z/VM 5.4 with `APAR VM64524` and by later z/VM versions.

Hotplug Memory – Reserved Storage



Dynamically Adding Memory - Planning

```
===== * * * Top of File * * *
===== USER RGYLX0E4 1GYLX0E4 1G 2G G
=====    INCLUDE LINDFLT
=====    CPU 00
=====    CPU 01
=====    CRYPTO    APVIRTUAL
=====    IUCV ANY
=====    LOADDEV PORTNAME 5005076306138411
=====    LOADDEV LUN 4011402E00000000
```

- This z/VM guest has a user directory entry with 1GB of initial memory and 2 GB of maximum memory
- In z/VM, changing the memory size or configuration of a guest causes a storage reset (all storage is cleared)
- If you are running Linux natively in an LPAR without z/VM, you would use reserved storage in the LPAR definition to set aside potential additional memory
- In z/VM, define the memory to be dynamically enabled as “standby” storage

Dynamically Adding Memory

```
21:15:04 Ready; T=0.01/0.02 21:15:04
21:15:14 define storage 1G standby 1G
21:15:14 00: STORAGE = 1G MAX = 2G INC = 2M STANDBY = 1G RESERVED = 0
21:15:14 00: Storage cleared - system reset.
```

- "**`DEFINE STORAGE 1G STANDBY 1G`**" issued for this guest
- Issuing a DEFINE STORAGE command causes storage to be cleared
- Anything running at the time of the reset will be immediately terminated without running any shutdown procedures
- This means if you issued this command from a CMS EXEC, CMS is no longer running because storage has been cleared.

Dynamically Adding Memory

- Example COMMAND statement in User Directory

```
USER RGYLX0E1 RGYLX0E1 3G 8G G
   INCLUDE LINDFLT
   COMMAND DEFINE STORAGE 2G STANDBY 2G
   CPU 00
   CRYPTO    APVIRTUAL
   IUCV ANY
   OPTION MAXCONN 128
   LINK RGYLXMNT 0191 0191 RR
   MDISK 0200 3390 1 END LS20C8 MR READ WRITE MULTIPLE
```

Dynamically Adding Memory

```
 ICH70001I RGYLX0E1 LAST ACCESS AT 20:23:51 ON THURSDAY, SEPTEMBER 22,
  2011
00: NIC 0600 is created; devices 0600-0602 defined
00: z/VM Version 6 Release 1.0, Service Level 1002 (64-bit),
00: built on IBM Virtualization Technology
00: There is no logmsg data
00: FILES: 0001 RDR,   NO PRT,   NO PUN
00: LOGON AT 20:26:20 EDT THURSDAY 09/22/11
00: STORAGE = 2G MAX = 8G INC = 4M STANDBY = 2G RESERVED = 0
00: Storage cleared - system reset.
z/VM V6.1.0    2010-10-15 11:49
DMSACP723I A (191) R/O
20:26:20 DIAG swap disk defined at virtual address 101 (64989 4K pages
  of swap space)
20:26:20  Detected interactive logon
20:26:20  MUST BE LOGGING ON FROM TERMINAL
```

Dynamically Adding Memory

```
rgylx0e4:~ # cat /proc/meminfo
MemTotal:        2051920 kB
MemFree:         1877596 kB
Buffers:           10304 kB
Cached:            51160 kB
SwapCached:            0 kB
Active:            29788 kB
Inactive:          54872 kB
Active(anon):      23212 kB
Inactive(anon):      120 kB
Active(file):       6576 kB
Inactive(file):    54752 kB
Unevictable:           0 kB
Mlocked:               0 kB
SwapTotal:             0 kB
```

- After IPLing Linux in this guest, observe via /proc/meminfo that approximately 2GB of memory is available
- The “standby” memory is not reported by /proc/meminfo
- The /sys file system however has an awareness of this “standby” or “hot plug” memory
- With s390-tools, **lsmem** can be used to report this information and **chmem** to bring storage elements online or offline

Dynamically Adding Memory

```
rgylx0e4:~ # lsmem
Address Range                          Size (MB)  State    Removable  Device
==========================================================================
                                                                      Core Memory
0x0000000000000000-0x000000000fffffff       256  online   no         0-63
0x0000000010000000-0x000000006fffffff      1536  online   yes        64-447
0x0000000070000000-0x000000007fffffff       256  online   no         448-511
0x0000000080000000-0x00000000ffffffff      2048  offline  -          512-1023
                                                                      Hotplug Memory

Memory device size   : 4 MB
Memory block size    : 256 MB
Total online memory  : 2048 MB
```

Dynamically Adding Memory

```
rgylx0e4:~ # chmem -e 2g
rgylx0e4:~ # lsmem
Address Range                                Size (MB)  State    Removable  Device
==========================================================================================
0x0000000000000000-0x000000000fffffff         256  online   no         0-63
0x0000000010000000-0x000000006fffffff        1536  online   yes        64-447
0x0000000070000000-0x000000007fffffff         256  online   no         448-511
0x0000000080000000-0x00000000ffffffff        2048  online   yes        512-1023

Memory device size   : 4 MB
Memory block size    : 256 MB
Total online memory  : 4096 MB
Total offline memory: 0 MB
```

- Additional 2GB of memory now available for application use

Dynamically Removing Memory

```
rgylx0e4:~ # chmem -d 2g
rgylx0e4:~ # lsmem
Address Range                              Size (MB)  State    Removable  Device
==============================================================================
0x0000000000000000-0x000000000fffffff        256  online   no         0-63
0x0000000010000000-0x000000006fffffff       1536  online   yes        64-447
0x0000000070000000-0x000000007fffffff        256  online   no         448-511
0x0000000080000000-0x00000000ffffffff       2048  offline  -          512-1023

Memory device size  : 4 MB
Memory block size   : 256 MB
Total online memory : 2048 MB
Total offline memory: 2048 MB
```

- Storage no longer needed can also be removed to ensure efficient operation

Dynamic Memory - Considerations

- To add and remove memory takes some small advanced planning. Develop a standard policy around how you will handle memory needs.
- Memory can be added or removed whether you are running under z/VM or in a native LPAR
- zVM User Directory `COMMAND` statement provides an effective way to issue the `DEFINE STORAGE` command in an non-disruptive manner.
- Remember not all memory sections will be removable, and the removable state can change over time

Summary of Memory Hotplug

- Utilizing hotplug memory does require some requirements:
  - ✓ z/VM 5.4 with VM64524 or above
  - ✓ DEFINE STORAGE STANDBY issued before Linux is IPLed
  - ✓ For native LPAR, RESERVED STORAGE must be defined before the LPAR is activated
  - ✓ SLES 11 / RHEL 6 provide support in Linux
- Suspend/Resume restriction: The Linux instance must not have used any hotplug memory since it was last booted. (Has worked if freed in advance)
- You may not be able to disable hotplug memory that has been enabled
- Can be very helpful when exact future memory need is unknown, without over allocating online memory from the start.
- After a Linux reboot core memory is made available again and hotplug memory is freed

Agenda

1. Value of Dynamic Resource Configuration
2. Dynamically Adding Disk Storage
3. Dynamically Adding Network Interfaces
4. Adding Hotplug Memory
5. **Adding CPUs**
6. Automatically Adjusting Memory and CPU Resources

Dynamically Managing Virtual CPs

```
===== USER RGYLX0E4 1GYLX0E4 1G 2G G
=====    INCLUDE LINDFLT
=====    CPU 00
=====    CPU 01
=====    CRYPTO    APVIRTUAL
=====    IUCV ANY
=====    LOADDEV PORTNAME 5005076306138411
=====    LOADDEV LUN 4011402E00000000
=====    MACHINE ESA 4
=====    OPTION APPLMON MAXCONN 128
```

- The directory entry shows two initial virtual CPs
- The maximum potential virtual CPs shown is four
- z/VM does not make the additional potential virtual CPs available for Linux to enable on its own
- The additional potential virtual CPs must first be **defined** in the z/VM guest before dynamically enabling on Linux

Dynamically Managing Virtual CPs

```
rgylx0e4:~ # vmcp q v
STORAGE = 1G
XSTORE = none
CPU 00  ID  FF12EBBE20978000 (BASE) CP   CPUAFF ON
CPU 01  ID  FF12EBBE20978000 CP   CPUAFF ON
AP 51 CEX2A Queue 08 shared
CONS 0009 DISCONNECTED    TERM START
     0009 CL T NOCONT NOHOLD COPY 001    READY FORM STANDARD
     0009 TO RGYLX0E4 RDR DIST RGYLX0E4  FLASHC 000 DEST OFF
     0009 FLASH       CHAR       MDFY       0 FCB       LPP OFF
     0009 3215   NOEOF OPEN 0013 NOKEEP NOMSG NONAME
     0009 SUBCHANNEL = 000A
```

- The current z/VM guests virtual resources are displayed from within Linux
- The two initial and active virtual CPs are shown
- Notice there is no information displayed about the potential additional virtual CPs

Dynamically Managing Virtual CPs

```
rgylx0e4:~ # mpstat -A
Linux 2.6.32.29-0.3-default (rgylx0e4)  04/01/11        _s390x_

13:19:24     CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
13:19:24     all    1.43    0.00    0.65    0.30    0.00    0.02    0.06    0.00   97.53
13:19:24       0    1.62    0.00    0.67    0.29    0.00    0.02    0.03    0.00   97.37
13:19:24       1    1.25    0.00    0.64    0.30    0.00    0.02    0.08    0.00   97.70

13:19:24     CPU    intr/s
13:19:24     all      0.00
13:19:24       0      0.00
13:19:24       1      0.00
```

- Note the mpstat output from before defining the additional virtual CPs
- Observe the even distribution of idle time and usage

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # ls
cpu0  cpu1  dispatching  kernel_max  offline  online  perf_events  possible  present
rgylx0e4:/sys/devices/system/cpu # cat kernel_max
63
rgylx0e4:/sys/devices/system/cpu # cat online
0-1
rgylx0e4:/sys/devices/system/cpu # cat offline
2-63
rgylx0e4:/sys/devices/system/cpu # cat possible
0-63
rgylx0e4:/sys/devices/system/cpu # cat present
0-1
rgylx0e4:/sys/devices/system/cpu # cat sched_mc_power_savings
0
rgylx0e4:/sys/devices/system/cpu #
```

- The Linux sysfs file system can access information about the two active virtual CPs
- The kernel has a maximum potential of 64 processors
- No information about the two potential additional virtual CPs is shown yet

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # modprobe vmcp
rgylx0e4:/sys/devices/system/cpu # vmcp define CPU 03 type cp
CPU 03 defined
rgylx0e4:/sys/devices/system/cpu # vmcp define CPU 02 type cp
CPU 02 defined
rgylx0e4:/sys/devices/system/cpu # ls
cpu0  cpu1  dispatching  kernel_max  offline  online  perf_events  possible
rgylx0e4:/sys/devices/system/cpu #
```

- Using the **`vmcp`** command we pass the zVM **`CP DEFINE CPU`** commands on to our z/VM guest.
- Remember this is a class G guest enabling the additional resources previously defined in the user directory
- After defining the additional virtual CPs in z/VM we still do not see them in the Linux /sys filesystem.

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # ls
cpu0  cpu1  dispatching  kernel_max  offline  online  perf_events  possible  present  rescan
rgylx0e4:/sys/devices/system/cpu # vmcp q v
STORAGE = 1G
XSTORE = none
CPU 00  ID  FF12EBBE20978000 (BASE) CP   CPUAFF ON
CPU 01  ID  FF12EBBE20978000 CP   CPUAFF ON
CPU 03  ID  FF12EBBE20978000 STOPPED CP   CPUAFF ON
CPU 02  ID  FF12EBBE20978000 STOPPED CP   CPUAFF ON
AP 51 CEX2A Queue 08 shared
CONS 0009 DISCONNECTED    TERM START
     0009 CL T NOCONT NOHOLD COPY 001    READY FORM STANDARD
     0009 TO RGYLX0E4 RDR DIST RGYLX0E4  FLASHC 000 DEST OFF
     0009 FLASH       CHAR       MDFY       0 FCB       LPP OFF
     0009 3215   NOEOF OPEN 0013 NOKEEP NOMSG NONAME
     0009 SUBCHANNEL = 000A
RDR  000C CL * NOCONT NOHOLD   EOF       READY
     000C 2540         CLOSED    NOKEEP NORESCAN  SUBCHANNEL = 000E
```

- By using the z/VM QUERY VIRTUAL command we can see the additional virtual CPs have been defined to the guest
- The new virtual CPs are in a “stopped” state

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # mpstat -A
Linux 2.6.32.29-0.3-default (rgylx0e4)  04/01/11        _s390x_

13:23:58     CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
13:23:58     all    0.47    0.00    0.23    0.10    0.00    0.01    0.02    0.00   99.16
13:23:58       0    0.54    0.00    0.24    0.10    0.00    0.01    0.01    0.00   99.10
13:23:58       1    0.41    0.00    0.23    0.10    0.00    0.01    0.03    0.00   99.23
rgylx0e4:/sys/devices/system/cpu # ls
cpu0  cpu1  dispatching  kernel_max  offline  online  perf_events  possible  present  rescan  sched_mc_p
rgylx0e4:/sys/devices/system/cpu #  echo 1 > rescan
rgylx0e4:/sys/devices/system/cpu # ls
cpu0  cpu1  cpu2  cpu3  dispatching  kernel_max  offline  online  perf_events  possible  present  rescan
rgylx0e4:/sys/devices/system/cpu #
```

- **`mpstat`** is only reporting two CPUs
- The `rescan` operation is used to search for new available CPUs in the guest.
- After rescan, additional /sysfs entries exist

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # mpstat -A
Linux 2.6.32.29-0.3-default (rgylx0e4)  04/01/11        _s390x_

13:24:41     CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
13:24:41     all    0.43    0.00    0.21    0.09    0.00    0.01    0.02    0.00   99.23
13:24:41       0    0.49    0.00    0.22    0.09    0.00    0.01    0.01    0.00   99.18
13:24:41       1    0.37    0.00    0.21    0.09    0.00    0.01    0.02    0.00   99.29
13:24:41       2    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
13:24:41       3    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
```

- mpstat reports 0% use and 0% idle for the new CPUs. This is because they are stopped and offline
- The new CPUs must still be brought online to Linux

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu/cpu2 # echo 1 > online
rgylx0e4:/sys/devices/system/cpu/cpu2 # ls
address  capability  configure  crash_notes  idle_count  idle_time_us  online  polarization
rgylx0e4:/sys/devices/system/cpu/cpu2 # cat online
1
rgylx0e4:/sys/devices/system/cpu/cpu2 # echo 1 > ../cpu3/online
```

- Bring the new CPUs online to Linux by echoing 1 in to the "online" file for the given CPU

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # mpstat -A
Linux 2.6.32.29-0.3-default (rgylx0e4)  04/01/11        _s390x_

13:26:36     CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
13:26:36     all    0.33    0.00    0.17    0.07    0.00    0.01    0.02    0.00   99.41
13:26:36       0    0.39    0.00    0.18    0.07    0.00    0.01    0.01    0.00   99.33
13:26:36       1    0.30    0.00    0.17    0.07    0.00    0.01    0.02    0.00   99.43
13:26:36       2    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00
13:26:36       3    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00
```

- On a idle system, the new CPUs momentarily show 100% idle after being brought online
- Once a little bit of workload hits the system, this quickly changes

Dynamically Managing Virtual CPs

```
rgylx0e4:/sys/devices/system/cpu # ls
cpu0  cpu1  cpu2  cpu3  dispatching  kernel_max  offline  online  perf_events  possible
rgylx0e4:/sys/devices/system/cpu # echo 0 > cpu1/online
rgylx0e4:/sys/devices/system/cpu # echo 0 > cpu3/online
rgylx0e4:/sys/devices/system/cpu # mpstat -A
Linux 2.6.32.29-0.3-default (rgylx0e4)  04/01/11        _s390x_

13:27:53     CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
13:27:53     all    0.27    0.00    0.14    0.06    0.00    0.01    0.01    0.00   99.52
13:27:53       0    0.35    0.00    0.16    0.06    0.00    0.01    0.01    0.00   99.40
13:27:53       1    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
13:27:53       2    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00
13:27:53       3    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
```

- You can take dynamically added CPUs offline again
- You can take offline CPUs that were initially online as well

Dynamically Managing Virtual CPs

- Considerations
  - Multithreaded application or multiple applications in a single virtual server could potentially benefit from additional virtual CPs
  - Adding or removing virtual CPs could impact monitoring applications or middleware that might query the number of processors on startup (ie the Java Virtual Machine)
  - zVM “DEFINE CPU” is a Class G command
  - This does NOT add additional capacity to the LPAR, it simply makes resources available to the guest
  - Watch the runnable queue (vmstat column 1)
  - (R.O.T.) Don’t add unnecessary virtual CPs or more virtual CPs than logical processors available.

Agenda

1. Value of Dynamic Resource Configuration
2. Dynamically Adding Disk Storage
3. Dynamically Adding Network Interfaces
4. Adding Hotplug Memory
5. Adding CPUs
6. **Automatically Adjusting Memory and CPU Resources**

cpuplugd – What is cpuplugd and why should I use it?

- Manually adjusting the quantity of CPU and memory configured to virtual guests is not the most effective approach, especially when managing thousands of virtual servers.
- The daemon (cpuplugd) can dynamically offline and re-online processors in Linux based on a rules based policy
- The daemon can also add and remove memory via CMM1
- The cpuplug daemon checks the system at user configurable intervals
- You must configure the plug and unplug rules for it to operate
- You must activate the cpuplug daemon to use it, by default it is inactive
- New capabilities have recently been added to cpuplugd with s390-tools 1.15

11.2

6.2

cpuplugd – Planning and evaluating

- The default rules are NOT recommendations, they are syntax examples.
- You should customize the configuration to fit your environment. Each virtual server may have different needs based on workload, middleware, and other factors.
- cpuplugd -V -f -c /etc/sysconfig/cpuplugd - This invokes cpuplugd in the foreground with verbose messaging to help you understand its operation. It is highly recommend you use this to understand how cpuplugd is functioning
  - Send to logfile: cpuplugd -c <config file> -f -**V**>&<logname> &
- When building rules for cpuplugd, it is important to understand what state you will be in after you execute a "plug" or "unplug" operation when writing the rules.
- Suggested Reading: May 2012 *Paper ZSW03228* "**Using the Linux cpuplugd Daemon to manage CPU and memory resources from z/VM Linux guests**"

cpuplugd – CP plug/unplug considerations

- Ensure you can grow CPU capacity to what the application requires to perform well (don't artificially limit). Use other mechanisms to throttle MIP usage based on priorities.
- Rules based on the last couple of sample intervals are more responsive than ones based on averages over minutes. Slower responses to change can mean lower throughput for your applications
- Keep in mind you can only add/remove a full virtual CP of capacity.
- Avoid rules that plug and immediately unplug CPUs continuously
  - Plug = idle < 50
  - Unplug = idle > 50
- This means at times you might have > 1 virtual CPs of idle capacity as an acceptable state.

cpuplugd - what if I run with the default rules?

- CPU_MIN= 1 and CPU_MAX= 0 (maximum available)
- UPDATE= 5
- HOTPLUG="(loadavg > onumcpus + 0.75) & (idle < 10.0)“
- HOTUNPLUG="(loadavg < onumcpus - 0.25) | (idle > 50)“
- Basic variables can be defined as:
  - loadavg: The load average over the past minute
  - onumcpus: The number of cpus which are online now
  - runable_proc: The current quantity of runable processes
  - idle: The current idle percentage
- Unplug at 51% idle? After unplug, what is my cpu busy?
- Plug only at 90% busy? What if my runable processes are growing high?

cpuplugd – understand what the variable represents

Where:

- **idle:** Current idle – Where 1 idle processor = 100 and 4 idle processors = 400 (/proc/stat 4th value). Idle does NOT stop at 100!

- **loadavg:** The current load average – The first /proc/loadavg value. The average number of runnable process. Not average CPU utilization! One looping process on a system would cause this to approach 1.0 Five looping processes on a single CPU system would cause this to approach 5.0

- **onumcpus:** The actual number of cpus which are online
  (Via: /sys/devices/system/cpu/cpu%d/online)

- **runable_proc**: The current quantity of runnable processes (The 4th /proc/loadavg value)

cpuplugd – New variables and rule capabilities for CPU

- New predefined keywords
  - `user` - the current CPU user percentage
  - `nice` - the current CPU nice percentage
  - `system` - the current CPU system percentage
  - `idle` - the current CPU idle percentage
  - `iowait` - the current CPU iowait percentage
  - `irq` - the current CPU irq percentage
  - `softirq` - the current CPU softirq percentage
  - `steal` - the current CPU steal percentage
  - `guest` - the current CPU guest percentage
  - `guest_nice` - the current CPU guest_nice percentage
  - `cpustat.<name>` - data from /proc/stat and /proc/loadavg
  - `time` - floating point timestamp in "seconds.microseconds" since Unix Epoch

11.2

6.2

- Historical function available and extremely useful
  - 0 is current interval
  - `cpustat.idle[0]` …. `cpustat.idle[99]`

- User Defined Variables Now Supported (See examples next slide)

User Define Variables Example for CPU

- user_0="(cpustat.user[0] - cpustat.user[1])"
- nice_0="(cpustat.nice[0] - cpustat.nice[1])"
- system_0="(cpustat.system[0] - cpustat.system[1])"
- user_2="(cpustat.user[2] - cpustat.user[3])"
- nice_2="(cpustat.nice[2] - cpustat.nice[3])"
- system_2="(cpustat.system[2] - cpustat.system[3])"
- CP_Active0="(user_0 + nice_0 + system_0)/ (cpustat.total_ticks[0] - cpustat.total_ticks[1])"
- CP_Active2="(user_2 + nice_2 + system_2)/ (cpustat.total_ticks[2] - cpustat.total_ticks[3])"
- **CP_ActiveAVG="(CP_Active0+CP_Active2) / 2“**

- idle_0="(cpustat.idle[0] - cpustat.idle[1])"
- iowait_0="(cpustat.iowait[0] - cpustat.iowait[1])"
- idle_2="(cpustat.idle[2] - cpustat.idle[3])"
- iowait_2="(cpustat.iowait[2] - cpustat.iowait[3])"
- CP_idle0="(idle_0 + iowait_0)/ (cpustat.total_ticks[0] - cpustat.total_ticks[1])"
- CP_idle2="(idle_2 + iowait_2)/ (cpustat.total_ticks[2] - cpustat.total_ticks[3])"
- **CP_idleAVG="(CP_idle0 + CP_idle2) / 2"**

cpuplugd – New variables and rule capabilities for CPU

- Valid operators for HOTPLUG/HOTUNPLUG rules

  + * ( ) / - < > & | !
- If HOTPLUG and HOTUNPLUG are true, only HOTPLUG is executed

- Additional features available for memory (discussed in the next section on memory)

Potential Starting Point for CPU Management

- Refer to paper ZSW03228
- Uses the CPU load values (from /proc/stat). The values of user, system, and nice are counted as active CPU use. *idle*, and *iowait* are considered as unused CPU capacity.
- The averages over the last three intervals are taken and divided by the corresponding time interval. The resulting values are stored in the variables `CP_ActiveAVG` and `CP_idleAVG`. The corresponding rules are as follows:
- `HOTPLUG="((1 - CP_ActiveAVG) * onumcpus) < 0.08"`
- `HOTUNPLUG="(CP_idleAVG * onumcpus) > 1.15"`

Potential Starting Point for CPU Management

- The values of *CP_ActiveAVG* and *CP_idleAVG* are between 0 and 1.
- Therefore, 1 - *CP_ActiveAVG* is the unused CPU capacity. When multiplied by the number of active CPUs, it is specified in CPUs.
- When the total unused CPU capacity falls below 8% of a single CPU, a new CPU is added. If the total amount of idle capacity is larger than 115% (this is 15% more than one CPU free), a CPU is withdrawn.
- The resulting automated sizing values are the same as the manual sizing settings in the test results documented in the paper. The system reacts quickly to load variations. The throughput closely approximates that of the manual sizing.
- Remember this is only a starting point. You must monitor results and adjust to what works well for the specific server, application, and workload.

cpuplugd memory management features

Automated Adjustments of Memory

- cpuplugd memory management utilizes CMM (CMM1)
- The cpuplug daemon determines how much memory to add or remove based upon the rules you put in place
- It is based on the same configurable interval you set for CPU rules
- The memory increment added or removed is configurable (and you should)
- Separate plug and unplug rules are used for memory management
- There are NO default memory plug and unplug rules
- If you start cpuplugd without any configuration changes it will manage CPUs but NOT memory.
- Be sure to have the following z/VM PTFs on:
  - APAR VM65060 REQUIRED!
    - **540 UM33537**
    - **610 UM33538**
    - **620 UM33539**

Linux Memory Management at a High Level

- Application requests for memory are managed as follows:
  - With sufficient free pages, the request is fulfilled immediately
  - If that causes the amount of free memory to fall below a high water mark, an **asynchronous** page scan by kswapd is triggered in the background.
  - If serving the request would cause the amount of free memory to fall below a low water mark, a so called direct scan is triggered, and the application **waits** until this scan provides the required pages.
  - The system may decide to mark anonymous pages (pages that are not related with files on disks) for swapping and initiate that these pages be written to swap asynchronously.

- The kswapd process is in an early indicator of a memory shortage
- Direct scans are more costly in terms of application performance
- Writing rules based on the scans can be more responsive than waiting until some paging activity occurs.

Automated Adjustments of Memory

- Basic variables for writing memory plug and unplug rules
  - **`apcr:`** the amount of page cache reads listed in vmstat bi/bo
  - **`Freemem:`** the amount of free memory (in megabyte)
  - **`swaprate`** the number of swapin & swapout operations
- CMM pool size and increment
  - **`CMM_MIN`** min size of static page pool (default 0)
  - **`CMM_MAX`** max size of static page pool
    - default was 32MB, now 512MB
  - **`CMM_INC`** amount for memunplug only (previously for plug and unplug)
    - default was 1MB, now 10% of free memory + cache, in pages
  - **`CMM_DEC`** amount for memplug operation ** New **
    - default 10% of total memory in pages
- **`apcr`** can be used to gauge the IO load on Linux system. With heavier IO rates you may want to allow the system to utilize more memory to help improve performance. This memory would get utilized by pagecache.
- Looking at "cache" for free memory might be skewed if you have a lot of shared memory (databases or java for example)

cpuplugd – New Variable & Rule Capabilities for Memory

- **New predefined keywords**
  - meminfo.<name> - any value from /proc/meminfo
  - vmstat.<name> - any value from /proc/vmstat
  - time - floating point timestamp in "seconds.microseconds"
- **Pre-defined dynamic variables - can be set to static value or algebraic expression:**
  - CMM_INC - pages the CMM page pool is increased for MEMUNPLUG
  - CMM_DEC - pages the CMM page pool is decreased for MEMPLUG
  - Operators for dynamic variable expressions: + * ( ) / - < >
- **History function available**
  - cpustat.<name> - from /proc/stat and /proc/loadavg
  - meminfo.<name> - any value from /proc/meminfo
  - vmstat.<name> - any value from /proc/vmstat
  - time - floating point timestamp in "seconds.microseconds"
- **User-defined variables (examples next slide)**

User Defined Variable Example for Memory

- The page scan rate can be calculated as the sum of:
  - vmstat.pgscan_kswapd_dma
  - vmstat.pgscan_kswapd_normal
  - vmstat.pgscan_kswapd_movable
    - pgscan_k="vmstat.pgscan_kswapd_dma[0] + vmstat.pgscan_kswapd_normal[0] + vmstat.pgscan_kswapd_movable[0]“
- The direct page scan rate can be calculated as the sum of:
  - vmstat.pgscan_direct_dma
  - vmstat.pgscan_direct_normal
  - vmstat.pgscan_direct_movable
    - pgscan_d="vmstat.pgscan_direct_dma[0] + vmstat.pgscan_direct_normal[0] + vmstat.pgscan_direct_movable[0]“
- The available part of the cache can be calculated as the:
  - meminfo.Cached -meminfo.Shmem
    - avail_cache="meminfo.Cached -meminfo.Shmem"

Automated Adjustments of Memory

- cpuplugd and CMM1 currently will NOT release pagecache memory. Consider writing a simple script of your own to perform this function if desired
- With the previous defaults, interval of 10 seconds and increment of 1MB, in a memory constrained situation you will only add 6MB/min or 360MB/hr
- With instantaneous allocations in GB by some application environments this has the potential to impact application performance, unless increased

CPU Hotplug Summary

- CPU Hotplug memory management will NOT release page cache memory on its own
- The CMM module has to be loaded before starting cpuplugd
- Understand how much memory you want to allow CMM to claim and the rate at which you will return memory to the system for use. The last thing you want is a failing memory allocation, or adverse performance impact.
- Under heavier IO load you might want to make more free memory available to Linux
- The goal is to allow the Linux to dynamically return pages of memory to z/VM when they are not in use, and to allow the entire system to operate more efficiently
- The amount of memory required an application to run is a function of the application program code, the workload volume, and any other software added to monitor or manage the environment.
- Improvement continue to be made to CMM and CPU Hotplug.

References

- Linux on System z Device Drivers, Features, and Commands
  - SC33-8411-09
- z/VM CP Commands and Utilities Reference
  - SC24-6175-01
- z/VM Directory Maintenance Facility Commands Reference
  - SC24-6188-01
- Using the Linux cpuplugd Daemon to manage CPU and memory resources from z/VM Linux guests
  - ZSW03228-USEN-00
  - http://www.ibm.com/developerworks/linux/linux390/perf/tuning_cpuhotplug.html#cpuplugd

Session Evaluations

- Dynamic Features of Linux on System z
- Session 11827
- www.SHARE.org/AnaheimEval

Richard G. Young

Senior Certified I.T. Specialist

IBM STG Lab Services

Virtualization & Linux on zEnterprise Team Lead

777 East Wisconsin Ave

Milwaukee, WI 53202

Tel 262 893 8662

Email: ryyoung1@us.ibm.com

Additional Material

Agenda

1. Value of Dynamic Resource Configuration
2. Dynamically Adding Disk Storage
3. Dynamically Adding Network Interfaces
4. Adding Hotplug Memory
5. Adding CPUs
6. Automatically Adjusting Memory and CPU Resources
7. **Suspend and Resume Functions**

Suspend and Resume Uses

- Possible Uses:
  - **Linux instance with middleware that has long startup or initialization time.**
  - **Instances with long idle periods where the server is not used. (development servers?) Use to free memory and processor resources while suspended**
  - **Resume a guest to central storage, moments before it is needed. (Assumes you know when it will be needed again)**
  - **Sync() provides OS/Filesystem consistency during backup.**
    - Suspend, FlashCopy, and Resume ?
    - Backup swap file with suspended image
    - Consistency of middleware such as databases must be handled through other means

Suspend and Resume Planning

- Planning for Suspend and Resume
  - Kernel 2.6.31 or higher
  - RHEL 6 / SLES 11 or higher
  - Suspended Linux is written to the designed swap disk
  - Must be large enough to hold the memory foot print of the Linux server
- Restrictions
  - No hotplug memory since the last boot
  - No CLAW Device Driver
  - All tape devices closed and unloaded
  - No DCSS with exclusive writable access
- While suspended:
  - Don't alter the data on the swap device with the suspend Linux
  - DCSSs and NSSs used must remain unchanged
  - Avoid real and virtual hardware configuration changes
- For all the restrictions and configuration information see:
  - Linux on System z Device Drivers, Features, and Commands SC33-8411-x

Suspend and Resume Planning

- Kernel Parameters
  - resume=<device node for swap partition>
  - no_console_suspend - Allows you to see console messages longer in to the suspend process
  - noresume -Skip resume of previously suspended system
- Consider swap file priorities
  - You might want to make swap partition for suspend the lowest priority
- Utilize echo disk > /sys/power/state
- Utilize SIGNAL SHUTDOWN and /etc/inittab CTRL-ALT-DELETE to suspend your system

Suspend and Resume Planning

```
rgylxd85:/etc # cat /etc/zipl.conf
# Modified by YaST2. Last modification on Sat Apr 23 15:48:27 EDT 2011
[defaultboot]
defaultmenu = menu

###Don't change this comment - YaST2 identifier: Original name: linux###
[SLES11_SP1V1]
    image = /boot/image-2.6.32.29-0.3-default
    target = /boot/zipl
    ramdisk = /boot/initrd-2.6.32.29-0.3-default,0x2000000
    parameters = "root=/dev/disk/by-path/ccw-0.0.0200-part1 resume=/dev/sda2 no_console_suspend
```

Suspend and Resume – Suspending

```
rgylxd85:~ # cat /proc/swaps
Filename                                Type            Size    Used    Priority
/dev/sda1                               partition       5237148 0       -1
/dev/sda2                               partition       5245212 0       1
rgylxd85:~ # vmstat 1
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  0      0 2957980   6424  43892    0    0   390    23    0  164  2  1 94  2  0
 0  0      0 2957980   6424  43892    0    0     0     0    0    8  0  0 100  0  0
 0  0      0 2957964   6424  43932    0    0     0     0    0   10  0  0 100  0  0
^C
rgylxd85:~ # echo disk > /sys/power/state
```

Suspend and Resume – Suspending

```
16:21:15 PM: Syncing filesystems ... 16:21:15 done.
16:21:15 Freezing user space processes ... (elapsed 0.00 seconds) done.
16:21:15 Freezing remaining freezable tasks ... (elapsed 0.00 seconds) done.
16:21:15 PM: Preallocating image memory... 16:21:15 done (allocated 45601 pages)

16:21:15 PM: Allocated 182404 kbytes in 0.12 seconds (1520.03 MB/s)
16:21:15 sd 1:0:3:1077035025: [sdb] Synchronizing SCSI cache
16:21:15 sd 0:0:5:1077035025: [sda] Synchronizing SCSI cache
16:21:16 01: HCPGSP2629I The virtual machine is placed in CP mode due to a SIGP
stop from CPU 01.
16:21:16 01: HCPGSP2627I The virtual machine is placed in CP mode due to a SIGP
initial CPU reset from CPU 00.
16:21:16 Disabling non-boot CPUs ...
16:21:16 cpu.f76a91: Processor 1 stopped
16:21:16 PM: Creating hibernation image:
16:21:16 PM: Need to copy 45066 pages
16:21:16 PM: Hibernation image created (45066 pages copied)
16:21:16 Enabling non-boot CPUs ...
16:21:16 cpu.17772b: Processor 1 started, address 0, identification 12EBBE
16:21:16 CPU1 is up
16:21:16 qdio: 0.0.2000 ZFCP on SC 1 using AI:1 QEBSM:1 PCI:1 TDD:1 SIGA: W AO
16:21:16 qdio: 0.0.1000 ZFCP on SC 0 using AI:1 QEBSM:1 PCI:1 TDD:1 SIGA: W AO
```

Suspend and Resume - Suspending

```
16:21:16 qdio: 0.0.0602 OSA on SC e using AI:1 QEBSM:0 PCI:1 TDD:1 SIGA:RW AO
16:21:16 qeth.736dae: 0.0.0600: Device is a Guest LAN QDIO card (level: V611)
16:21:16 with link type GuestLAN QDIO (portname: )
16:21:16 qeth.47953b: 0.0.0600: Hardware IP fragmentation not supported on eth0
16:21:16 qeth.066069: 0.0.0600: Inbound source MAC-address not supported on eth0

16:21:16 qeth.d7fdb4: 0.0.0600: VLAN enabled
16:21:16 qeth.e90c78: 0.0.0600: Multicast enabled
16:21:16 qeth.5a9d02: 0.0.0600: IPV6 enabled
16:21:16 qeth.184d8a: 0.0.0600: Broadcast enabled
16:21:16 qeth.dac2aa: 0.0.0600: Using SW checksumming on eth0.
16:21:16 qeth.9c4c89: 0.0.0600: Outbound TSO not supported on eth0
16:21:16 PM: Saving image data pages (45155 pages) ...          0%      1%
 2%      3%      4%      5%      6%      7%16:21:21       8%      9%      10%
 11%     12%     13%     14%     15%     16%     17%     18%     19%     20%
 21%     22%     23%     24%     25%     26%     27%     28%     29%     30%
 31%     32%     33%     34%     35%     36%     37%     38%     39%     40%
 41%     42%     43%     44%     45%     46%     47%     48%     49%     50%
 51%     52%     53%     54%     55%     56%     57%     58%     59%     60%
 61%     62%     63%     64%     65%     66%     67%     68%     69%     70%
 71%     72%     73%     74%     75%     76%     77%     78%     79%     80%
 81%     82%     83%     84%     85%     86
```

Suspend and Resume

```
%     87%     88%     89%     90%     91%     92%     93%     94%     95%     96
%     97%     98%     99%    100%    done
16:21:21 PM: Wrote 180620 kbytes in 1.18 seconds (153.06 MB/s)
16:21:21 PM: S
16:21:21 md: stopping all md devices.
16:21:25 sd 1:0:3:1077035025: [sdb] Synchronizing SCSI cache
16:21:25 sd 0:0:5:1077035025: [sda] Synchronizing SCSI cache
16:21:25 01: HCPGSP2629I The virtual machine is placed in CP mode due to a SIGP
stop from CPU 01.
16:21:25 00: HCPGSP2629I The virtual machine is placed in CP mode due to a SIGP
stop from CPU 00.
16:21:33 00: IPL 200 CLEAR
16:21:33 00: zIPL v1.8.0-44.45.2 interactive boot menu
16:21:33 00:
16:21:33 00:  0. default (SLES11_SP1V1)
16:21:33 00:
16:21:33 00:  1. SLES11_SP1V1
16:21:33 00:  2. FailsafeV2
16:21:33 00:  3. ipl
16:21:33 00:
16:21:33 00: Note: VM users please use '#cp vi vmsg <number> <kernel-parameters>
'
```

Suspend and Resume – Resuming

```
oesn't support DPO or FUA
16:21:54  sda: sda1 sda2
16:21:54 sd 0:0:5:1077035025: [sda] Attached SCSI disk
16:21:54 mount: devpts already mounted or /dev/pts busy
16:21:54 mount: according to mtab, devpts is already mounted on /dev/pts
16:21:54  Boot logging started on /dev/ttyS0(/dev/console) at Sat Apr 23 16:21:4
5 2011
16:21:54 PM: Starting manual resume from disk
16:21:54 Freezing user space processes ... (elapsed 0.00 seconds) done.
16:21:54 Freezing remaining freezable tasks ... (elapsed 0.00 seconds) done.
16:21:54 PM: Loading image data pages (45155 pages) ...          0%      1%
  2%      3%      4%      5%      6%      7%      8%      9%     10%     11%
 12%     13%     14%     15%     16%     17%     18%     19%     20%     21%
 22%     23%     24%     25%     26%     27%     28%     29%     30%     31%
 32%     33%     34%     35%     36%     37%     38%     39%     40%     41%
 42%     43%     44%     45%     46%     47%     48%     49%     50%     51%
 52%     53%     54%     55%     56%     57%     58%     59%     60%     61%
 62%     63%     64%     65%     66%     67%     68%     69%     70%     71%
 72%     73%     74%     75%     76%     77%     78%     79%     80%     81%
 82%     83%     84%     85%     86%     87%     88%     89%     90%     91%
 92%     93%     94%     95%     96%     97%     98%     99%    100%    done
16:21:54 PM: Read 180620 kbytes in 1.31 seconds (137.87 MB/s)
```

Suspend and Resume

```
rgylxd85:~ # cat /proc/swaps
Filename                                Type            Size    Used    Priority
/dev/sda1                               partition       5237148 0       -1
/dev/sda2                               partition       5245212 0       1
rgylxd85:~ # vmstat 1
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  0      0 2957980   6424  43892    0    0   390    23    0  164  2  1 94  2  0
 0  0      0 2957980   6424  43892    0    0     0     0    0    8  0  0 100  0  0
 0  0      0 2957964   6424  43932    0    0     0     0    0   10  0  0 100  0  0
^C
rgylxd85:~ # echo disk > /sys/power/state
rgylxd85:~ # uptime
  4:22pm  up   0:02,  1 user,  load average: 0.05, 0.02, 0.00
rgylxd85:~ #
```

Suspended

Resumed

If the suspend and resume are completed fast enough your TCP connections may not even drop. The above ssh session is an example of that.

Using “Signal Shutdown” to trigger a suspend

Suspend and Resume - /etc/inittab

```
#3:2345:respawn:/sbin/mingetty --noclear /dev/3270/ttycons dumb
# KVM hypervisor console:
#1:2345:respawn:/sbin/mingetty --noclear /dev/hvc0 linux

# what to do when CTRL-ALT-DEL is pressed
#<F12>ca::ctrlaltdel:/sbin/shutdown -r -t 4 now
ca::ctrlaltdel:/bin/sh -c "/bin/echo disk > /sys/power/state || /sbin/shutdown -t3 -h now"

# not used for now:
pf::powerwait:/etc/init.d/powerfail start
pn::powerfailnow:/etc/init.d/powerfail now
#pn::powerfail:/etc/init.d/powerfail now
po::powerokwait:/etc/init.d/powerfail stop
sh:12345:powerfail:/sbin/shutdown -h now THE POWER IS FAILING
```

- By adding the modified **`ctrlaltdel`** entry to /etc/inittab you can suspend your Linux guest to a swap file when it receive a "Signal shutdown".
- In the event the suspend fails, a "regular" shutdown would occur.

Suspend and Resume - Signal

```
signal shutdown user rgylxd85 within 60
Ready; T=0.01/0.01 17:02:06
```

- Trigging a suspend from z/VM is easy once the Linux inittab update is in place.
- The standard signal shutdown command should very quickly suspend the guest

Suspend and Resume - Summary

- Great option for middleware with long startup times
- Using Linux hotplug memory should be avoided with suspend / resume
- Ensure your swap file has adequate space to store the Linux instance
- If the resume fails, a normal IPL will occur