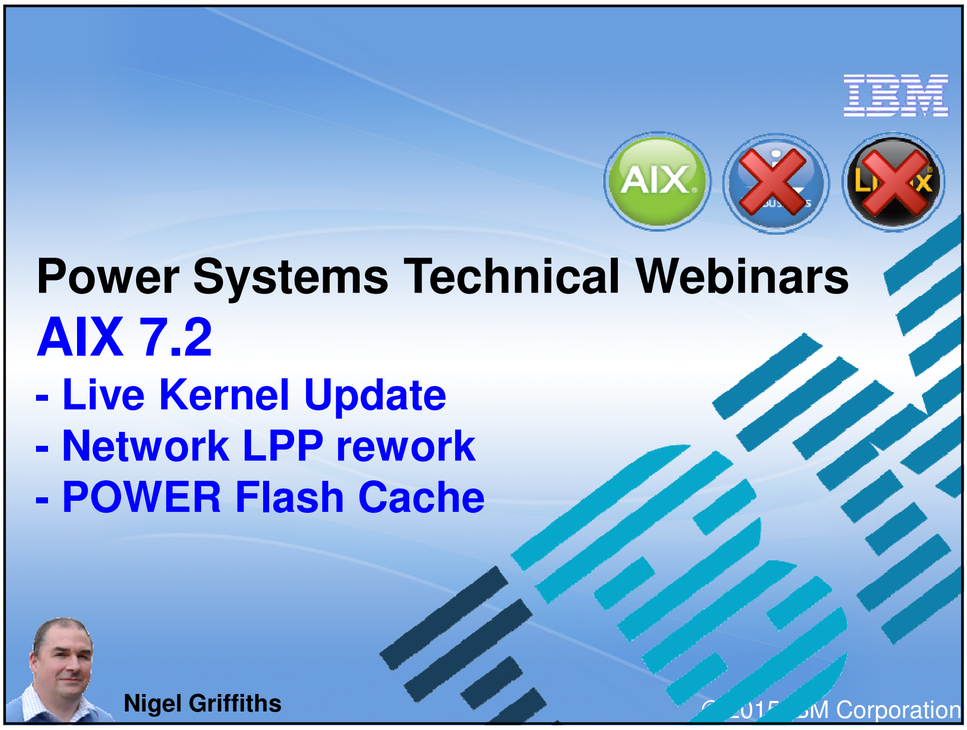
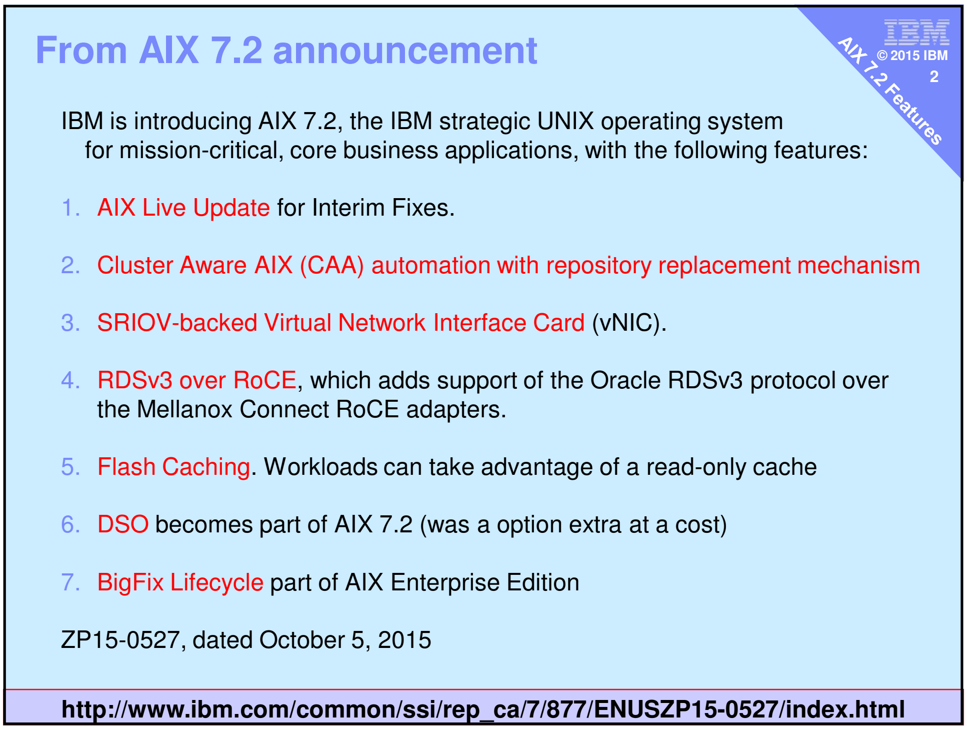


Power Systems Technical Webinars

AIX 7.2

- Live Kernel Update
- Network LPP rework
- POWER Flash Cache

Nigel Griffiths



From AIX 7.2 announcement

IBM is introducing AIX 7.2, the IBM strategic UNIX operating system for mission-critical, core business applications, with the following features:

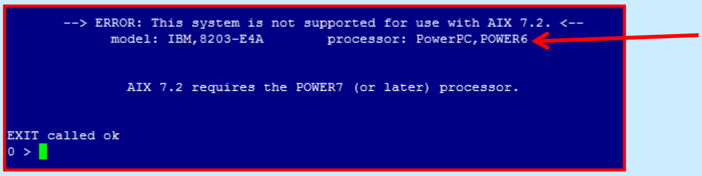
1. AIX Live Update for Interim Fixes.
2. Cluster Aware AIX (CAA) automation with repository replacement mechanism
3. SRIOV-backed Virtual Network Interface Card (vNIC).
4. RDSv3 over RoCE, which adds support of the Oracle RDSv3 protocol over the Mellanox Connect RoCE adapters.
5. Flash Caching. Workloads can take advantage of a read-only cache
6. DSO becomes part of AIX 7.2 (was a option extra at a cost)
7. BigFix Lifecycle part of AIX Enterprise Edition

ZP15-0527, dated October 5, 2015

http://www.ibm.com/common/ssi/rep_ca/7/877/ENUSZP15-0527/index.html

AIX 7.2 Pre-Reqs

- POWER7, Power7+ or POWER8 or higher
  - No support for POWER5 or POWER6 or older



```
       --> ERROR: This system is not supported for use with AIX 7.2. <--
             model: IBM,8203-E4A        processor: PowerPC,POWER6


               AIX 7.2 requires the POWER7 (or later) processor.


EXIT called ok
0 >
```

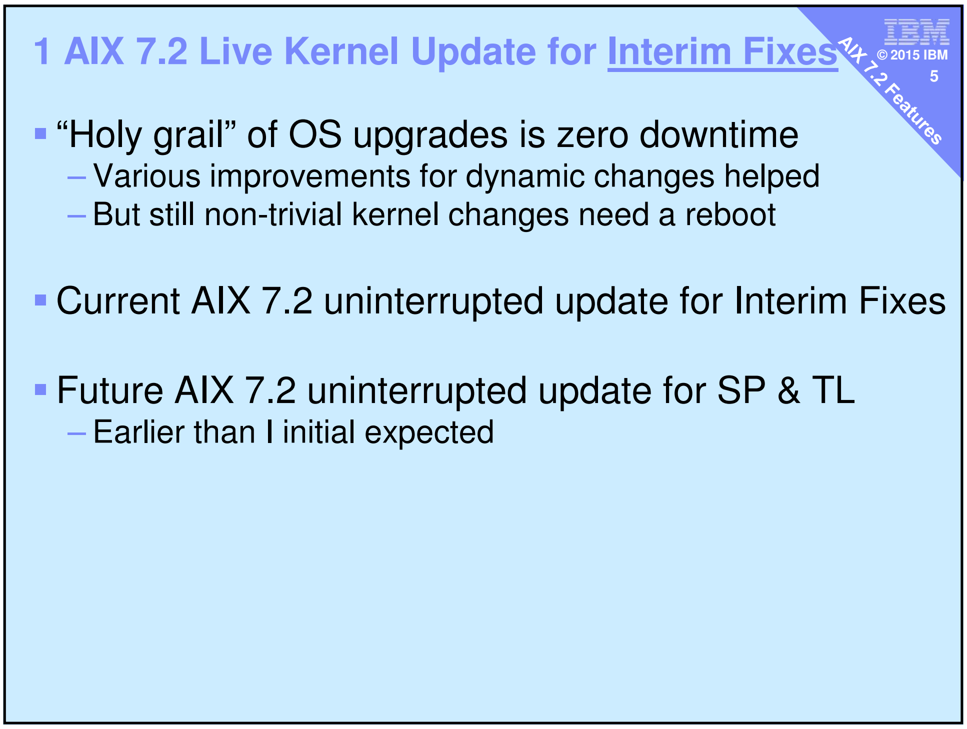
- AIX 7.2 arrived on 4th December 2015

1 AIX 7.2 Live Kernel Update for Interim Fixes

Chris Gibson

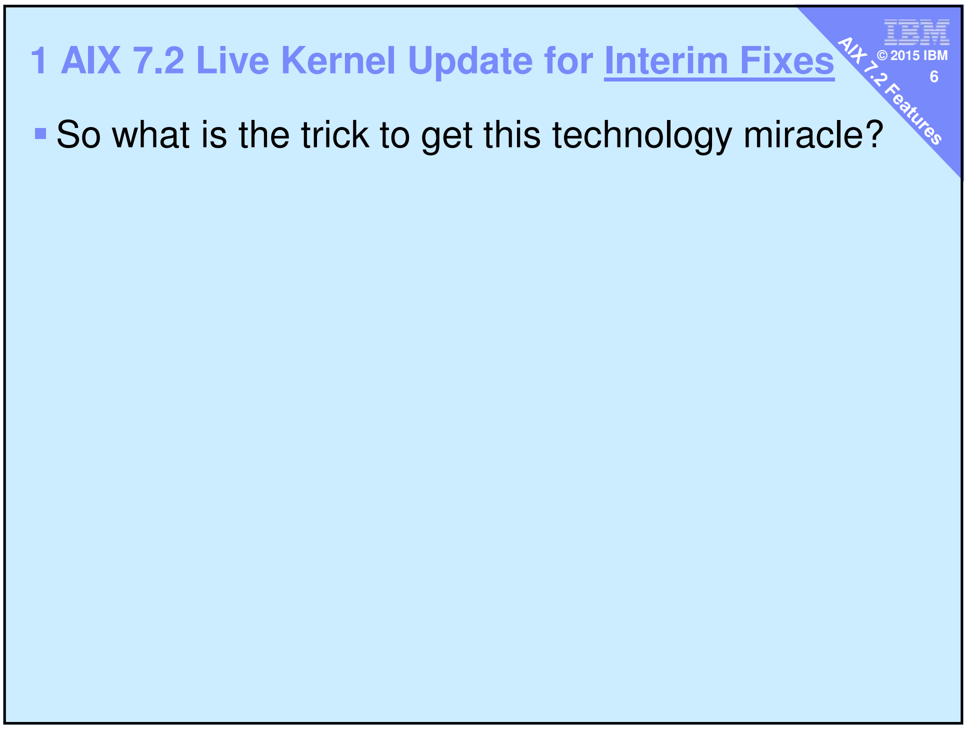
- Power Systems Client Technical Specialist
- Melbourne, Australia,

- Excellent Web Article / whitepaper
- https://www.ibm.com/developerworks/community/blogs/cgaix/resource/AIXLiveUpdateblog.pdf



1 AIX 7.2 Live Kernel Update for Interim Fixes

- “Holy grail” of OS upgrades is zero downtime
  - Various improvements for dynamic changes helped
  - But still non-trivial kernel changes need a reboot
- Current AIX 7.2 uninterrupted update for Interim Fixes
- Future AIX 7.2 uninterrupted update for SP & TL
  - Earlier than I initial expected



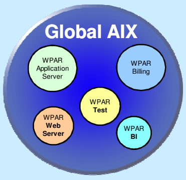
1 AIX 7.2 Live Kernel Update for Interim Fixes

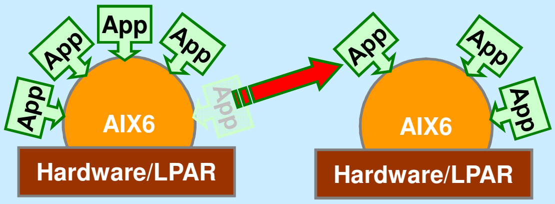
- So what is the trick to get this technology miracle?

1 AIX 7.2 Live Kernel Update for Interim Fixes

- So what is the trick to get this technology miracle?

Workload Partitions



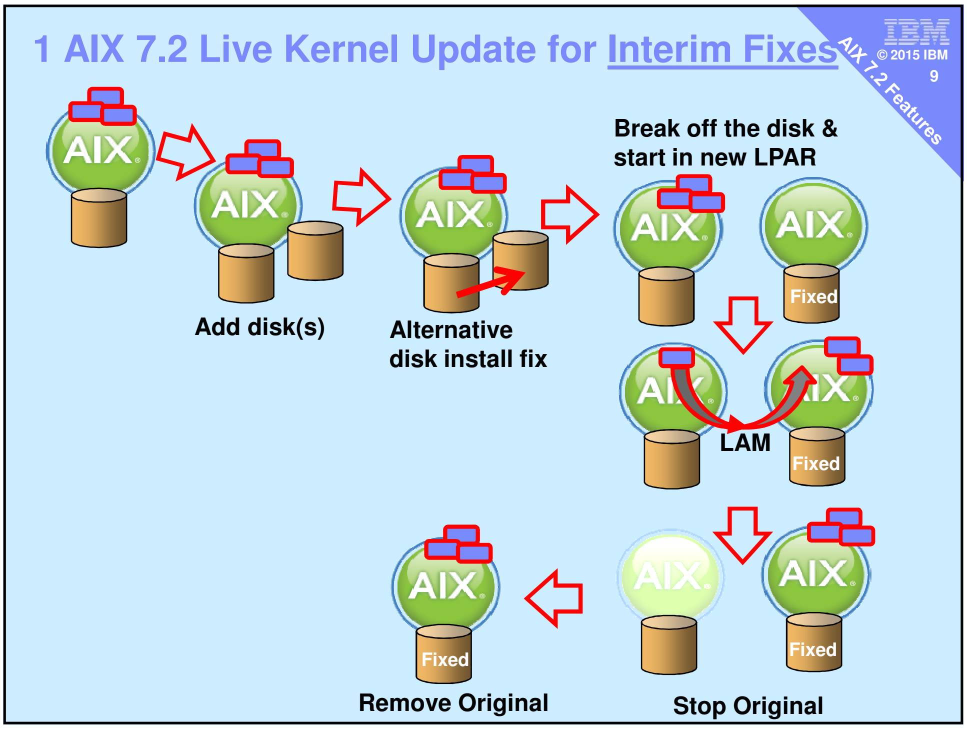


WPAR Reminder

- 2007 AIX 6.1 was release
- Workload Partitions was a major feature (WPAR)
  with RBAC for WPAR security

- All familiar with Live Partition Mobility (LPM)
  - Jumps a whole AIX + Apps to a different server
- WPAR has Live Application Mobility (LAM)
  - Jumps just the Apps to a different AIX image

- Now add
  - Clone the source AIX and add the iFix
  - The two AIX LPARs are on the same machine
  - LAM the running Apps between Source and Target AIX



1 AIX 7.2 Live Kernel Update for Interim Fixes

AIX 7.2 Features

© 2015 IBM

9

Add disk(s)

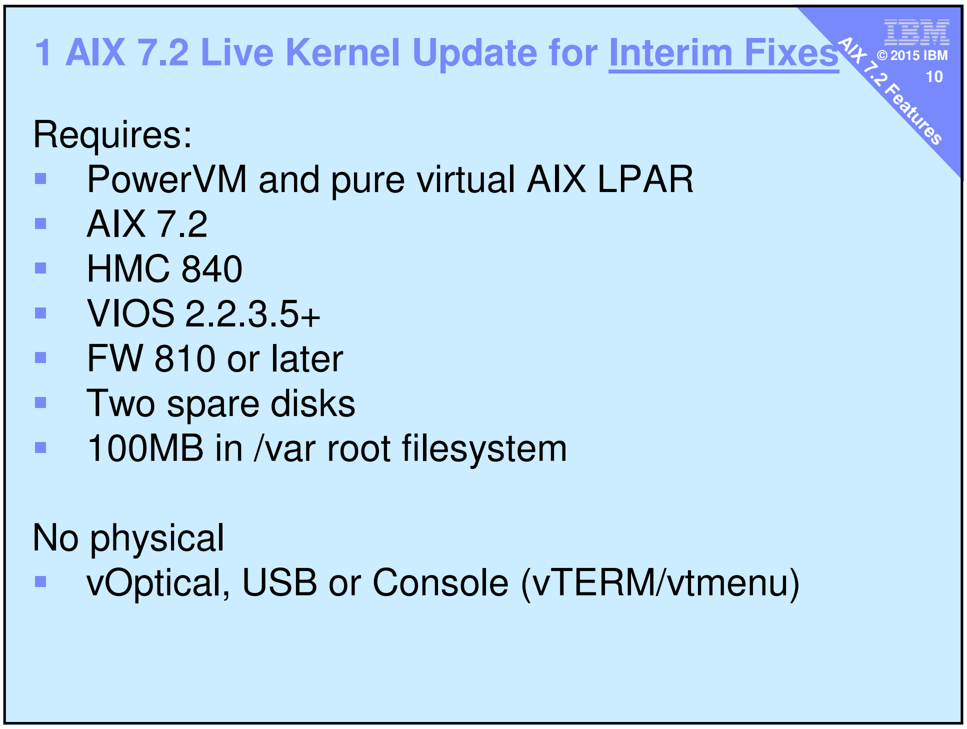
Alternative disk install fix

Break off the disk & start in new LPAR

LAM

Stop Original

Remove Original



1 AIX 7.2 Live Kernel Update for Interim Fixes

AIX 7.2 Features

© 2015 IBM

10

Requires:

- PowerVM and pure virtual AIX LPAR
- AIX 7.2
- HMC 840
- VIOS 2.2.3.5+
- FW 810 or later
- Two spare disks
- 100MB in /var root filesystem

No physical

- vOptical, USB or Console (vTERM/vtmenu)

1 AIX 7.2 Live Kernel Update for Interim Fixes

Control File: /var/adm/ras/liveupdate/lvupdate.data

- Two pages of comments at the top on the syntax options

```
general:
        mode = automated
        kext_check = no

disks:
        nhdisk  = hdisk1   # boot for surrogate
        mhdisk  = hdisk2   # mirror for surrogate
#       tohdisk =          # optional: non-rootvg paging space
#       tshdisk =          # optional: non-rootvg paging space

hmc:
        lpar_id  = 42
        management_console = hmc14
        user = hscroot
```

1 AIX 7.2 Live Kernel Update for Interim Fixes

```
# chfs -a size=512M /var
```

Commands:

Authenicate with HMC:

```
    hmcauth -u hscroot -a hmc_name
```

Preview check for   /tmp/dummy.150813.epkg.Z

```
    geninstall -k -p -d /tmp dummy.150813.epkg.Z
```

Note: the space character

Upgrade

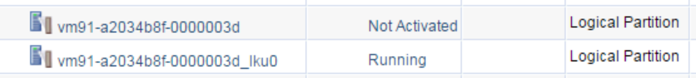
```
    geninstall -k    -d /tmp  dummy.150813.epkg.Z
```

AIX 7.2 LKU Experience in beta testing

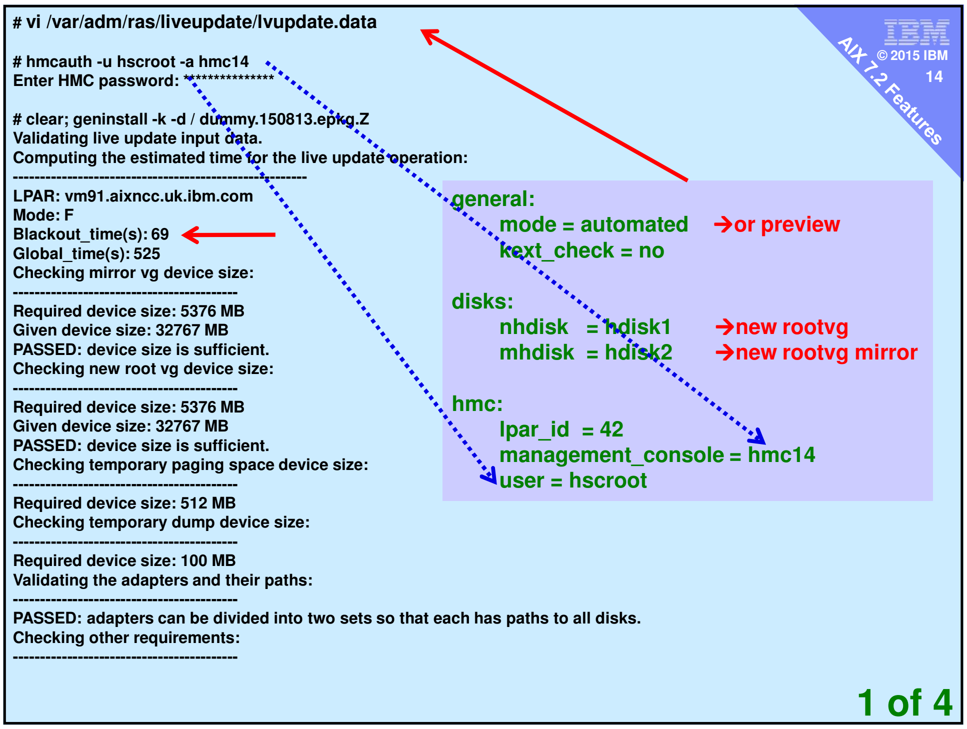
1. File & commands – easy to make mistakes/inconsistence
   - Like hmcauth one HMC but the other HMC in the file
2. Mandatory minimum of two other disks
3. All disks need to be multipath

```
# lspath
Enabled hdisk0 vscsi0
Enabled hdisk1 vscsi0
Enabled hdisk2 vscsi0
Enabled hdisk2 vscsi1
Enabled hdisk1 vscsi1
Enabled hdisk0 vscsi1
```

4. No Virtual optical attached
5. No VTERM console
6. Can't run the command from a VTERM console
7. Detailed logging is very good to work out the errors
8. Try a LPM Validate to help spot odd things!
9. Original LPAR renamed with added "_lku0" & finally removed



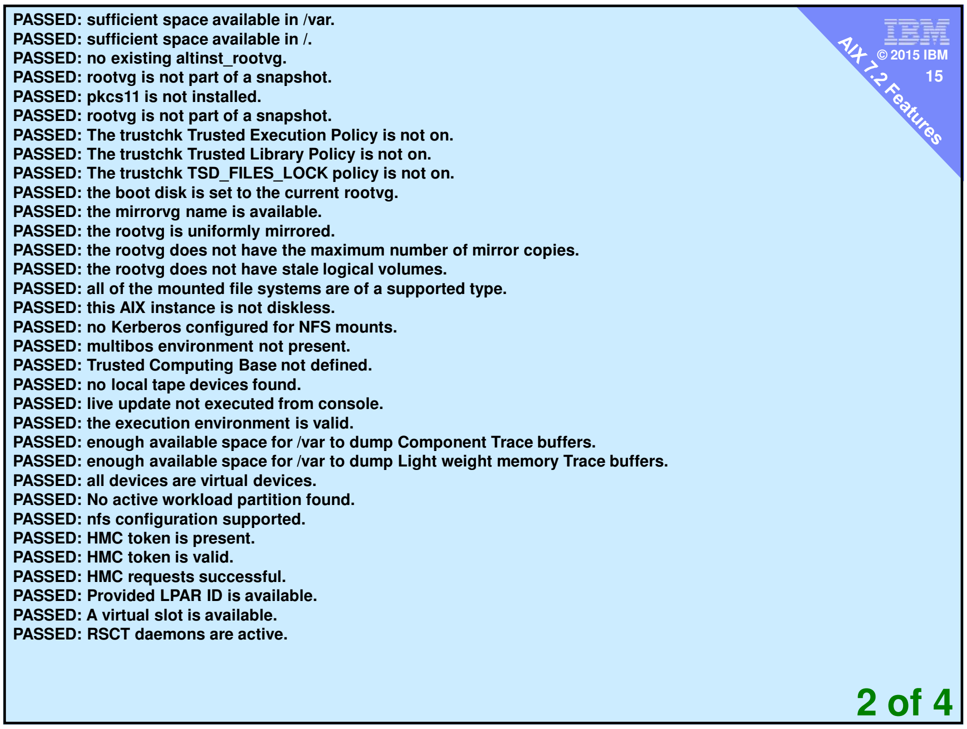
vm91-a2034b8f-0000003d	Not Activated		Logical Partition
vm91-a2034b8f-0000003d_lku0	Running		Logical Partition



```
# vi /var/adm/ras/liveupdate/lvupdate.data

# hmcauth -u hscroot -a hmc14
Enter HMC password: **************

# clear; geninstall -k -d / dummy.150813.epkg.Z
Validating live update input data.
Computing the estimated time for the live update operation:
-------------------------------------------------------
LPAR: vm91.aixncc.uk.ibm.com
Mode: F
Blackout_time(s): 69
Global_time(s): 525
Checking mirror vg device size:
------------------------------------------
Required device size: 5376 MB
Given device size: 32767 MB
PASSED: device size is sufficient.
Checking new root vg device size:
------------------------------------------
Required device size: 5376 MB
Given device size: 32767 MB
PASSED: device size is sufficient.
Checking temporary paging space device size:
------------------------------------------
Required device size: 512 MB
Checking temporary dump device size:
------------------------------------------
Required device size: 100 MB
Validating the adapters and their paths:
------------------------------------------
PASSED: adapters can be divided into two sets so that each has paths to all disks.
Checking other requirements:
------------------------------------------
```

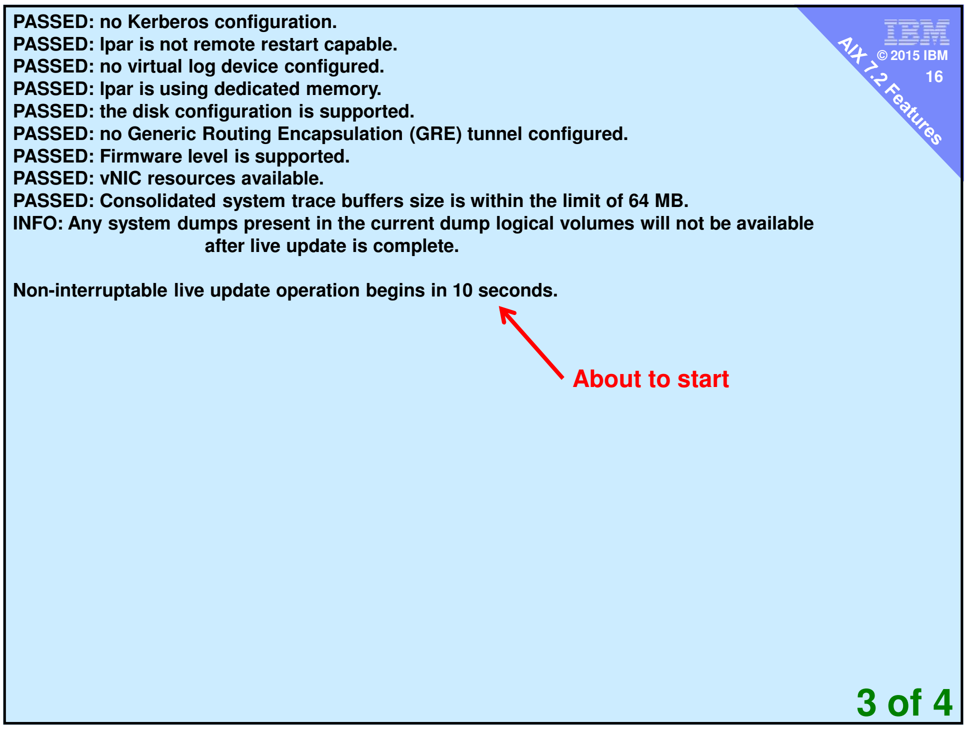
```
general:
    mode = automated       →or preview
    kext_check = no

disks:
    nhdisk  = hdisk1       →new rootvg
    mhdisk  = hdisk2       →new rootvg mirror

hmc:
    lpar_id  = 42
    management_console = hmc14
    user = hscroot
```

1 of 4



```
PASSED: sufficient space available in /var.
PASSED: sufficient space available in /.
PASSED: no existing altinst_rootvg.
PASSED: rootvg is not part of a snapshot.
PASSED: pkcs11 is not installed.
PASSED: rootvg is not part of a snapshot.
PASSED: The trustchk Trusted Execution Policy is not on.
PASSED: The trustchk Trusted Library Policy is not on.
PASSED: The trustchk TSD_FILES_LOCK policy is not on.
PASSED: the boot disk is set to the current rootvg.
PASSED: the mirrorvg name is available.
PASSED: the rootvg is uniformly mirrored.
PASSED: the rootvg does not have the maximum number of mirror copies.
PASSED: the rootvg does not have stale logical volumes.
PASSED: all of the mounted file systems are of a supported type.
PASSED: this AIX instance is not diskless.
PASSED: no Kerberos configured for NFS mounts.
PASSED: multibos environment not present.
PASSED: Trusted Computing Base not defined.
PASSED: no local tape devices found.
PASSED: live update not executed from console.
PASSED: the execution environment is valid.
PASSED: enough available space for /var to dump Component Trace buffers.
PASSED: enough available space for /var to dump Light weight memory Trace buffers.
PASSED: all devices are virtual devices.
PASSED: No active workload partition found.
PASSED: nfs configuration supported.
PASSED: HMC token is present.
PASSED: HMC token is valid.
PASSED: HMC requests successful.
PASSED: Provided LPAR ID is available.
PASSED: A virtual slot is available.
PASSED: RSCT daemons are active.
```

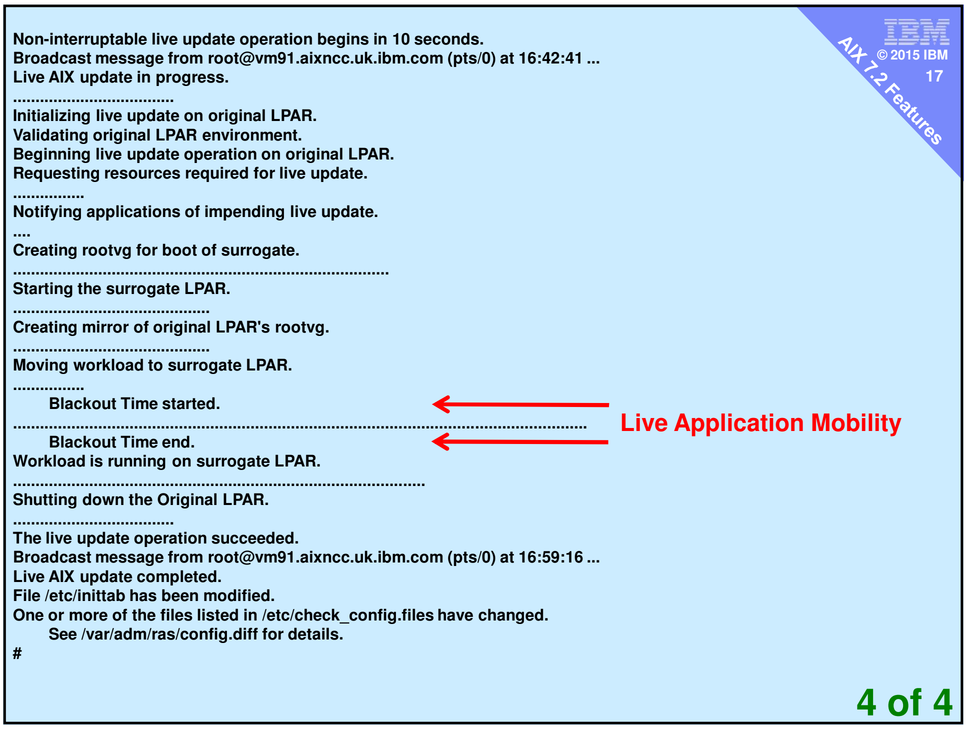
2 of 4



```
PASSED: no Kerberos configuration.
PASSED: lpar is not remote restart capable.
PASSED: no virtual log device configured.
PASSED: lpar is using dedicated memory.
PASSED: the disk configuration is supported.
PASSED: no Generic Routing Encapsulation (GRE) tunnel configured.
PASSED: Firmware level is supported.
PASSED: vNIC resources available.
PASSED: Consolidated system trace buffers size is within the limit of 64 MB.
INFO: Any system dumps present in the current dump logical volumes will not be available
                              after live update is complete.

Non-interruptable live update operation begins in 10 seconds.
```

About to start

3 of 4



AIX 7.2 Features

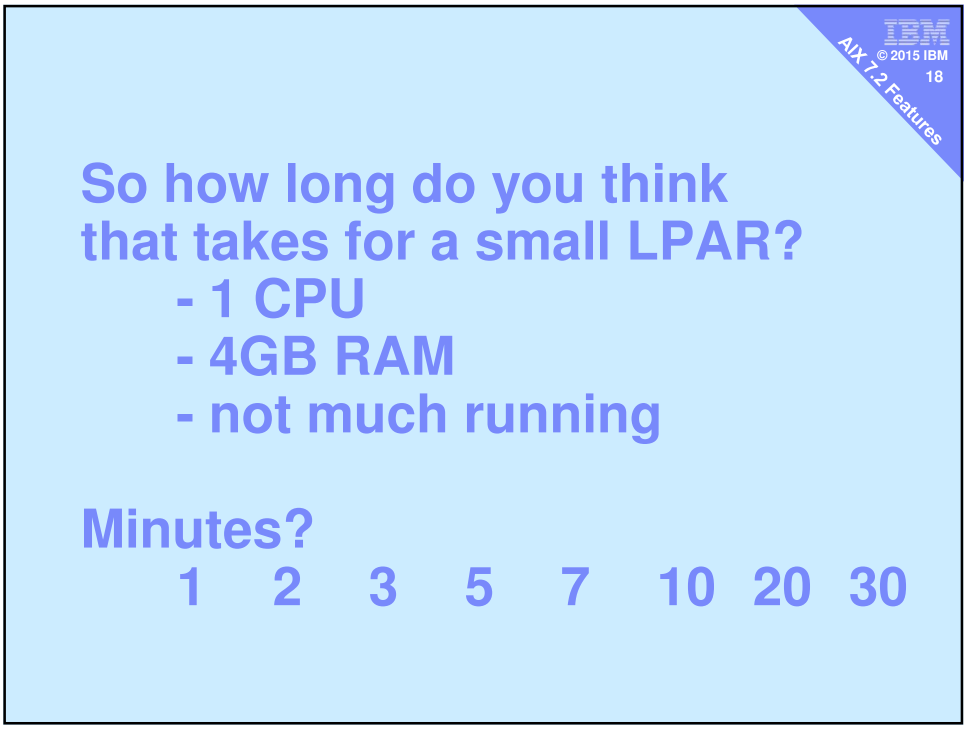
© 2015 IBM

17

```
Non-interruptable live update operation begins in 10 seconds.
Broadcast message from root@vm91.aixncc.uk.ibm.com (pts/0) at 16:42:41 ...
Live AIX update in progress.
....................................
Initializing live update on original LPAR.
Validating original LPAR environment.
Beginning live update operation on original LPAR.
Requesting resources required for live update.
................
Notifying applications of impending live update.
....
Creating rootvg for boot of surrogate.
....................................................................................
Starting the surrogate LPAR.
............................................
Creating mirror of original LPAR's rootvg.
............................................
Moving workload to surrogate LPAR.
................
        Blackout Time started.
..................................................................................................................................
        Blackout Time end.
Workload is running on surrogate LPAR.
............................................................................................
Shutting down the Original LPAR.
....................................
The live update operation succeeded.
Broadcast message from root@vm91.aixncc.uk.ibm.com (pts/0) at 16:59:16 ...
Live AIX update completed.
File /etc/inittab has been modified.
One or more of the files listed in /etc/check_config.files have changed.
        See /var/adm/ras/config.diff for details.
#
```

Live Application Mobility

4 of 4



AIX 7.2 Features

© 2015 IBM

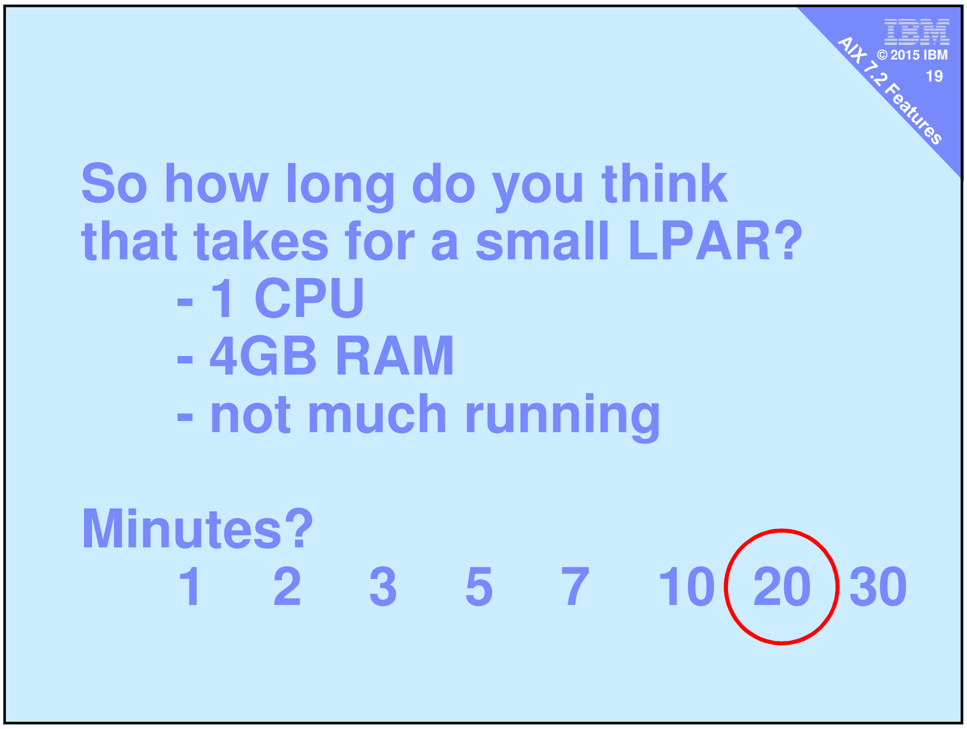
18

So how long do you think that takes for a small LPAR?

- 1 CPU
- 4GB RAM
- not much running

Minutes?

1 2 3 5 7 10 20 30

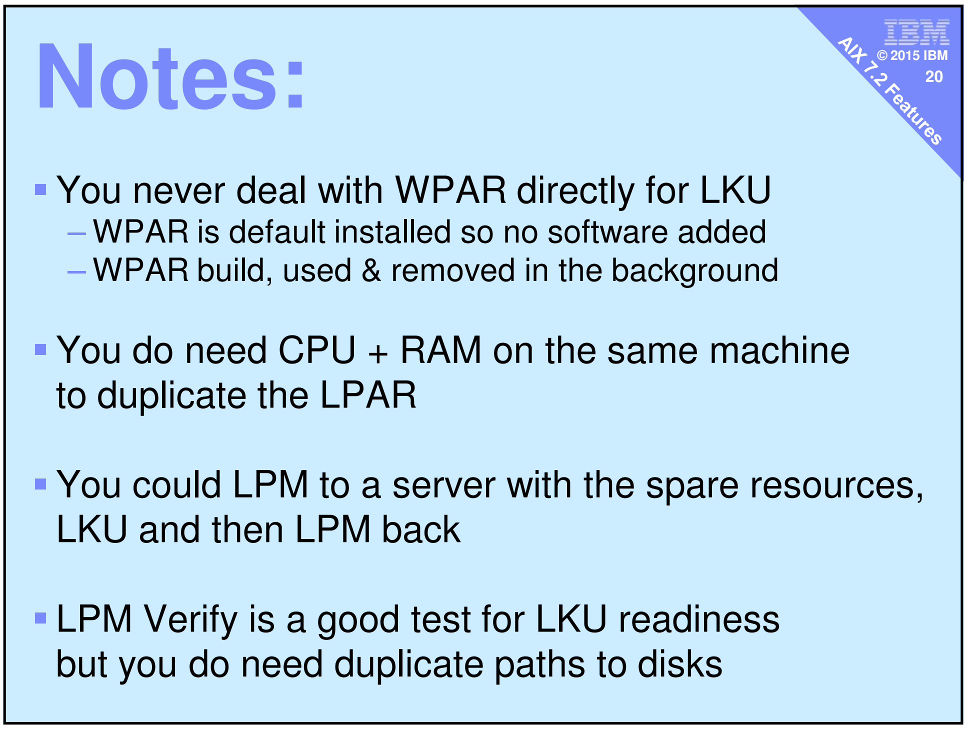


So how long do you think that takes for a small LPAR?

- 1 CPU
- 4GB RAM
- not much running

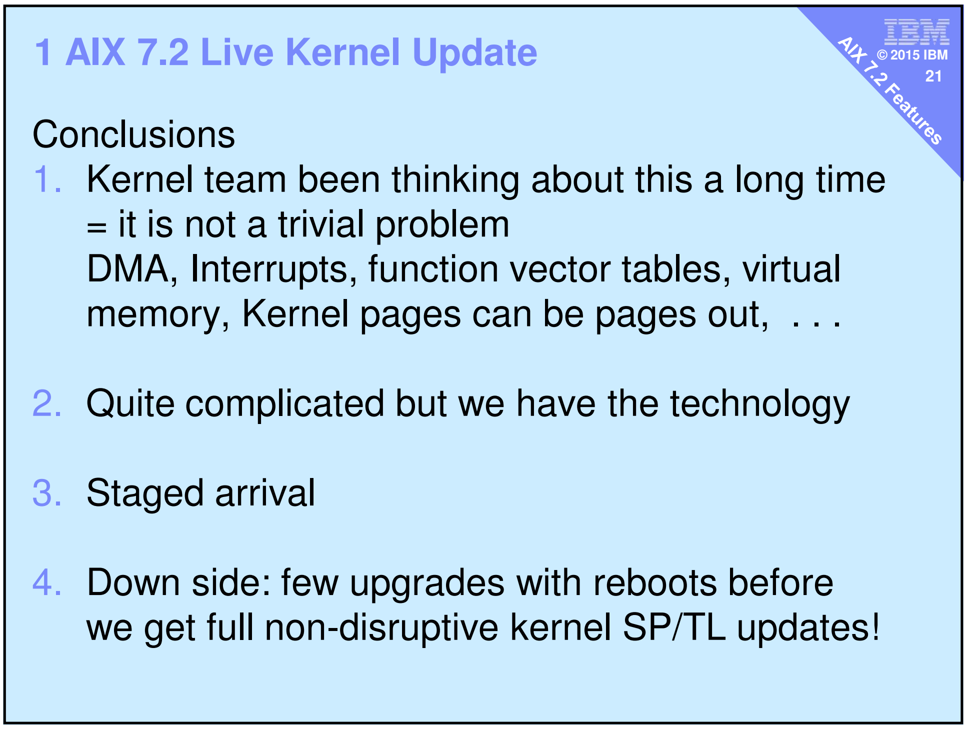
Minutes?

1 2 3 5 7 10 [✓ 20] 30



Notes:

- You never deal with WPAR directly for LKU
  - WPAR is default installed so no software added
  - WPAR build, used & removed in the background
- You do need CPU + RAM on the same machine to duplicate the LPAR
- You could LPM to a server with the spare resources, LKU and then LPM back
- LPM Verify is a good test for LKU readiness but you do need duplicate paths to disks



1 AIX 7.2 Live Kernel Update

Conclusions

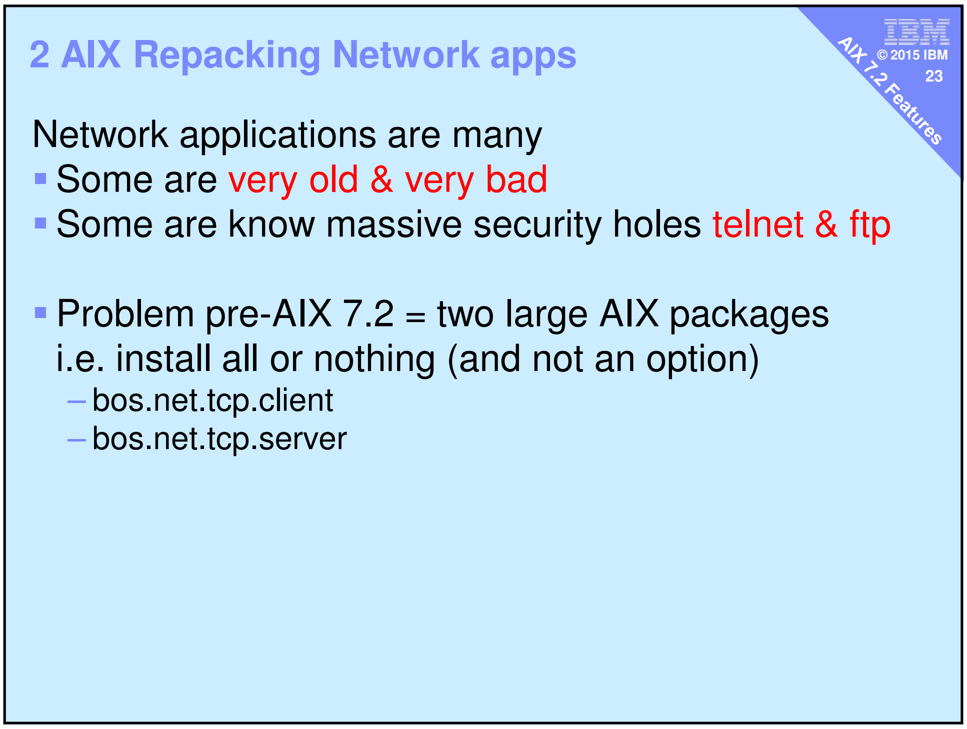
1. Kernel team been thinking about this a long time = it is not a trivial problem
   DMA, Interrupts, function vector tables, virtual memory, Kernel pages can be pages out, . . .

2. Quite complicated but we have the technology

3. Staged arrival

4. Down side: few upgrades with reboots before we get full non-disruptive kernel SP/TL updates!

2 AIX Repacking Network apps

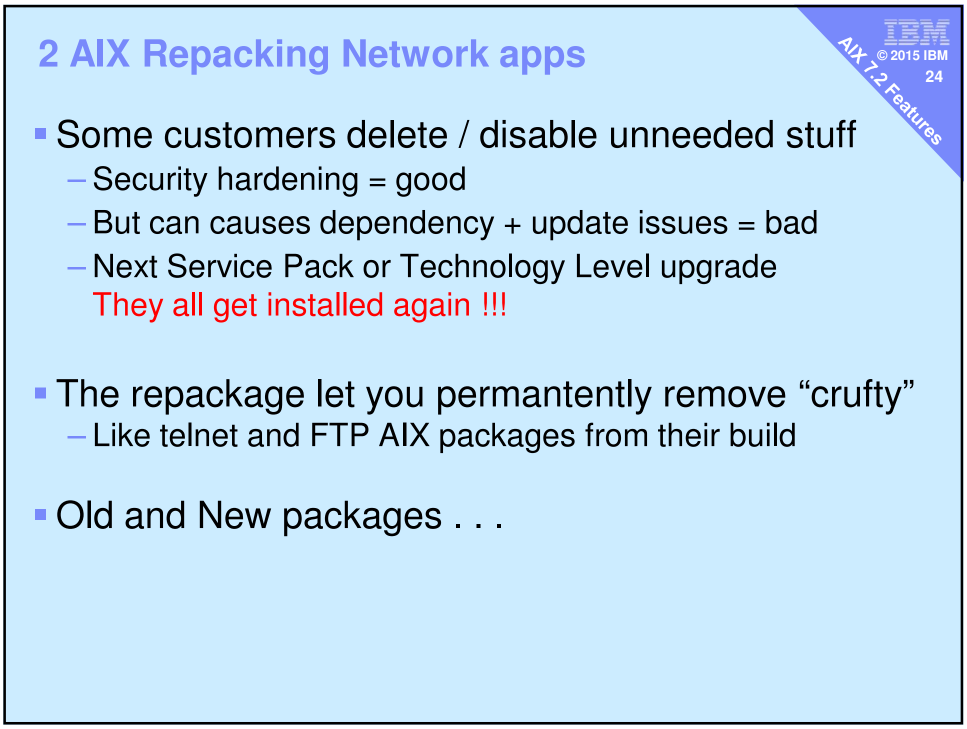


2 AIX Repacking Network apps

Network applications are many

- Some are very old & very bad
- Some are know massive security holes telnet & ftp

- Problem pre-AIX 7.2 = two large AIX packages i.e. install all or nothing (and not an option)
  - bos.net.tcp.client
  - bos.net.tcp.server

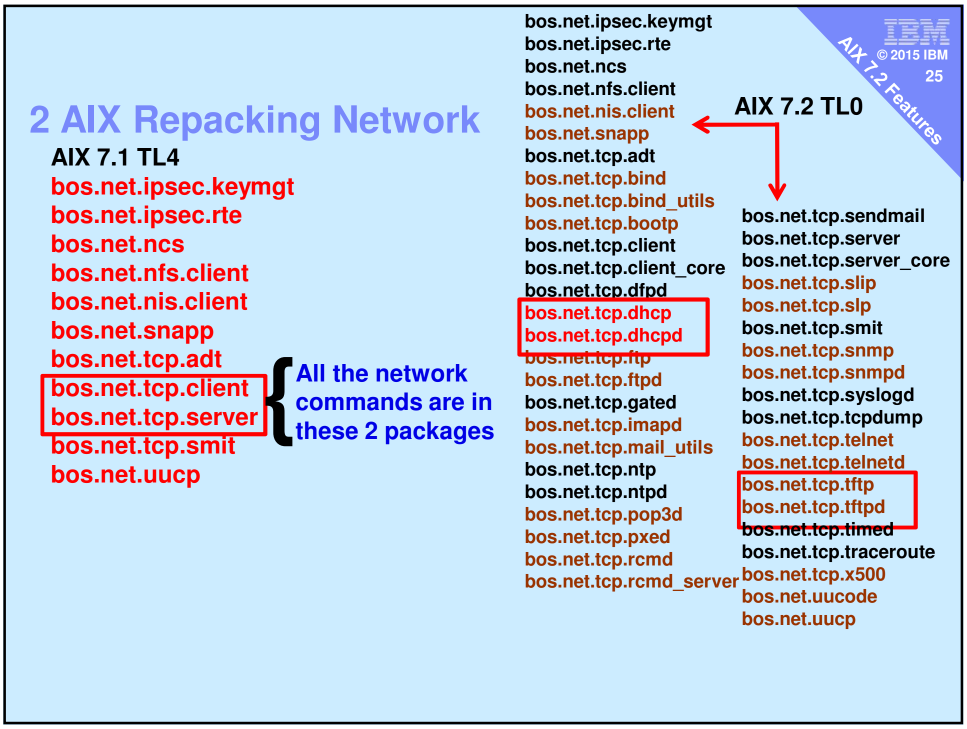


2 AIX Repacking Network apps

- Some customers delete / disable unneeded stuff
  - Security hardening = good
  - But can causes dependency + update issues = bad
  - Next Service Pack or Technology Level upgrade They all get installed again !!!

- The repackage let you permantently remove “crufty”
  - Like telnet and FTP AIX packages from their build

- Old and New packages . . .



IBM © 2015 IBM 25 AIX 7.2 Features

2 AIX Repacking Network

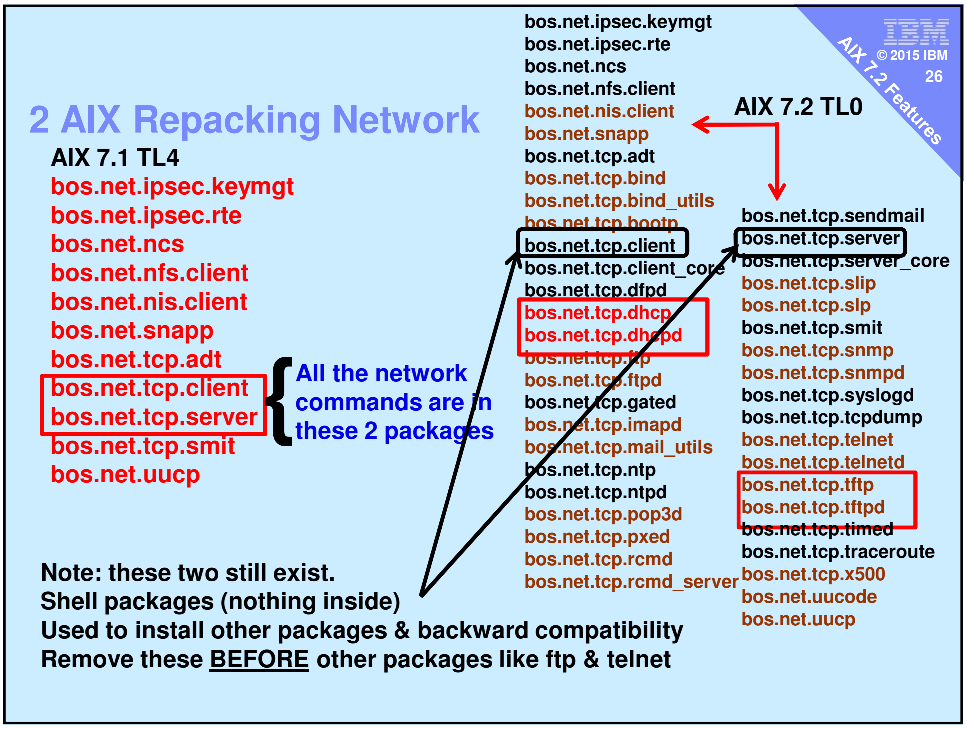
AIX 7.1 TL4

- bos.net.ipsec.keymgt
- bos.net.ipsec.rte
- bos.net.ncs
- bos.net.nfs.client
- bos.net.nis.client
- bos.net.snapp
- bos.net.tcp.adt
- bos.net.tcp.client
- bos.net.tcp.server
- bos.net.tcp.smit
- bos.net.uucp

All the network commands are in these 2 packages

AIX 7.2 TL0

- bos.net.ipsec.keymgt
- bos.net.ipsec.rte
- bos.net.ncs
- bos.net.nfs.client
- bos.net.nis.client
- bos.net.snapp
- bos.net.tcp.adt
- bos.net.tcp.bind
- bos.net.tcp.bind_utils
- bos.net.tcp.bootp
- bos.net.tcp.client
- bos.net.tcp.client_core
- bos.net.tcp.dfpd
- bos.net.tcp.dhcp
- bos.net.tcp.dhcpd
- bos.net.tcp.ftp
- bos.net.tcp.ftpd
- bos.net.tcp.gated
- bos.net.tcp.imapd
- bos.net.tcp.mail_utils
- bos.net.tcp.ntp
- bos.net.tcp.ntpd
- bos.net.tcp.pop3d
- bos.net.tcp.pxed
- bos.net.tcp.rcmd
- bos.net.tcp.rcmd_server
- bos.net.tcp.sendmail
- bos.net.tcp.server
- bos.net.tcp.server_core
- bos.net.tcp.slip
- bos.net.tcp.slp
- bos.net.tcp.smit
- bos.net.tcp.snmp
- bos.net.tcp.snmpd
- bos.net.tcp.syslogd
- bos.net.tcp.tcpdump
- bos.net.tcp.telnet
- bos.net.tcp.telnetd
- bos.net.tcp.tftp
- bos.net.tcp.tftpd
- bos.net.tcp.timed
- bos.net.tcp.traceroute
- bos.net.tcp.x500
- bos.net.uucode
- bos.net.uucp



IBM © 2015 IBM 26 AIX 7.2 Features

2 AIX Repacking Network

AIX 7.1 TL4

- bos.net.ipsec.keymgt
- bos.net.ipsec.rte
- bos.net.ncs
- bos.net.nfs.client
- bos.net.nis.client
- bos.net.snapp
- bos.net.tcp.adt
- bos.net.tcp.client
- bos.net.tcp.server
- bos.net.tcp.smit
- bos.net.uucp

All the network commands are in these 2 packages

AIX 7.2 TL0

- bos.net.ipsec.keymgt
- bos.net.ipsec.rte
- bos.net.ncs
- bos.net.nfs.client
- bos.net.nis.client
- bos.net.snapp
- bos.net.tcp.adt
- bos.net.tcp.bind
- bos.net.tcp.bind_utils
- bos.net.tcp.bootp
- bos.net.tcp.client
- bos.net.tcp.client_core
- bos.net.tcp.dfpd
- bos.net.tcp.dhcp
- bos.net.tcp.dhcpd
- bos.net.tcp.ftp
- bos.net.tcp.ftpd
- bos.net.tcp.gated
- bos.net.tcp.imapd
- bos.net.tcp.mail_utils
- bos.net.tcp.ntp
- bos.net.tcp.ntpd
- bos.net.tcp.pop3d
- bos.net.tcp.pxed
- bos.net.tcp.rcmd
- bos.net.tcp.rcmd_server
- bos.net.tcp.sendmail
- bos.net.tcp.server
- bos.net.tcp.server_core
- bos.net.tcp.slip
- bos.net.tcp.slp
- bos.net.tcp.smit
- bos.net.tcp.snmp
- bos.net.tcp.snmpd
- bos.net.tcp.syslogd
- bos.net.tcp.tcpdump
- bos.net.tcp.telnet
- bos.net.tcp.telnetd
- bos.net.tcp.tftp
- bos.net.tcp.tftpd
- bos.net.tcp.timed
- bos.net.tcp.traceroute
- bos.net.tcp.x500
- bos.net.uucode
- bos.net.uucp

Note: these two still exist.
Shell packages (nothing inside)
Used to install other packages & backward compatibility
Remove these BEFORE other packages like ftp & telnet

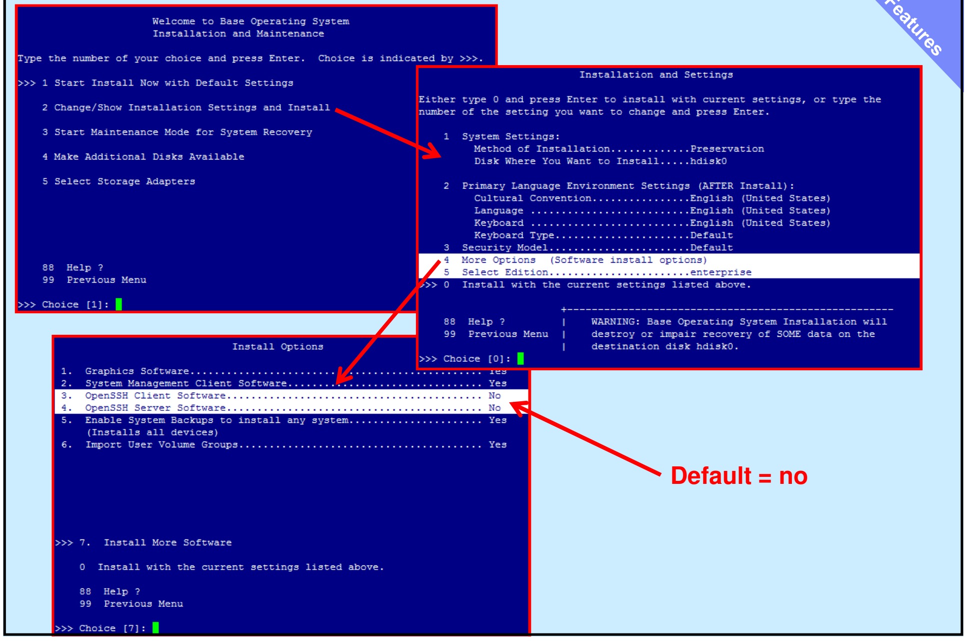
Other AIX 7.2 Packaging News

2 AIX Repacking Network apps

- AIX 7.2 by defaults installs 695 packages
  - Removing packages = faster install
  - Switched off graphics & “old” box support = device drivers and got down to 200 packages
  - AIX 7.2 DVD1 = 3.2GB [AIX 7.1 DVD1 = 4.3 GB]
- Sys. Admin can remove unwanted packages from:
  - NIM
  - mksysb installs
  - PowerVC clones
- Good news:

  ssh not a default install but is on the Installer menu

2 AIX Repacking Network apps



```
                     Welcome to Base Operating System
                      Installation and Maintenance

Type the number of your choice and press Enter.  Choice is indicated by >>>.

>>> 1 Start Install Now with Default Settings

    2 Change/Show Installation Settings and Install

    3 Start Maintenance Mode for System Recovery

    4 Make Additional Disks Available

    5 Select Storage Adapters




    88  Help ?
    99  Previous Menu

>>> Choice [1]:
```

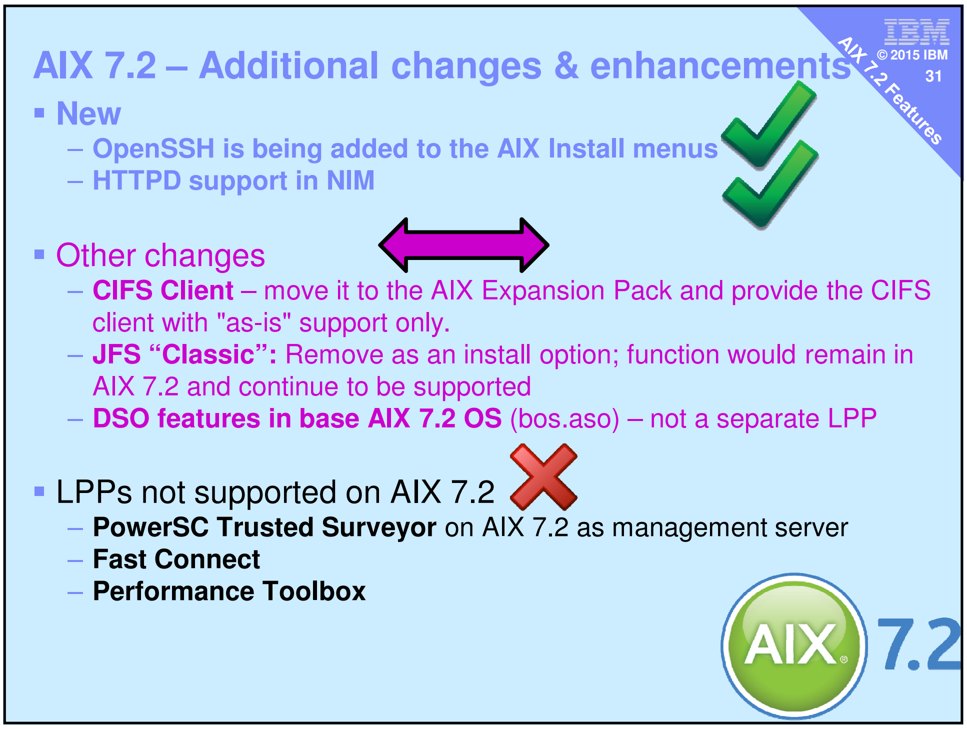
```
                         Installation and Settings

Either type 0 and press Enter to install with current settings, or type the
number of the setting you want to change and press Enter.

    1  System Settings:
         Method of Installation.............Preservation
         Disk Where You Want to Install.....hdisk0

    2  Primary Language Environment Settings (AFTER Install):
         Cultural Convention................English (United States)
         Language ..........................English (United States)
         Keyboard ..........................English (United States)
         Keyboard Type......................Default
    3  Security Model.......................Default
    4  More Options  (Software install options)
    5  Select Edition.......................enterprise
>>> 0  Install with the current settings listed above.

                       +-----------------------------------------------------
    88  Help ?         |    WARNING: Base Operating System Installation will
    99  Previous Menu  |    destroy or impair recovery of SOME data on the
                       |    destination disk hdisk0.
>>> Choice [0]:
```

```
                            Install Options

 1.  Graphics Software................................................ Yes
 2.  System Management Client Software................................ Yes
 3.  OpenSSH Client Software.......................................... No
 4.  OpenSSH Server Software.......................................... No
 5.  Enable System Backups to install any system...................... Yes
     (Installs all devices)
 6.  Import User Volume Groups........................................ Yes







>>> 7.  Install More Software

    0  Install with the current settings listed above.

    88  Help ?
    99  Previous Menu

>>> Choice [7]:
```

Default = no

AIX 7.2 Code Removal and LPP Changes

- No support for POWER6, POWER5, or POWER4
- Remove “Trusted Computing Base” → Trusted Execution
- Additional Code Removals from AIX 7.2
  - NIS+
  - NDAF
  - IBM Virtual Shared Disk (rsct.vsd)
  - IBM Systems Director Components; pConsole * Running Man!
  - Selected old adapters
  - Selected performance toolbox components & eclipse2.rte, including bos.perf.gtools and performance workbench GUI
  - IP over FC driver
  - Fcparray head driver
  - graPhigs
  - Java 5
  - Bos.INed * Worst editor on UNIX
  - Obsolete locales



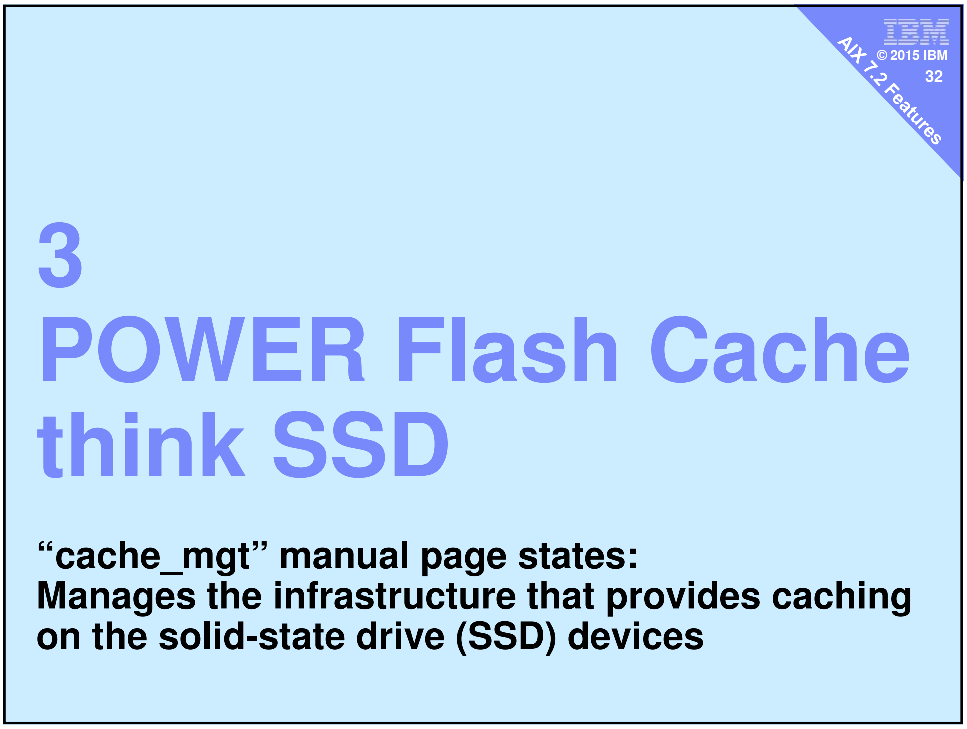
IBM © 2015 IBM 31

AIX 7.2 Features

AIX 7.2 – Additional changes & enhancements

- New
  - OpenSSH is being added to the AIX Install menus ✓
  - HTTPD support in NIM ✓
- Other changes ↔
  - **CIFS Client** – move it to the AIX Expansion Pack and provide the CIFS client with "as-is" support only.
  - **JFS "Classic":** Remove as an install option; function would remain in AIX 7.2 and continue to be supported
  - **DSO features in base AIX 7.2 OS** (bos.aso) – not a separate LPP
- LPPs not supported on AIX 7.2 ✗
  - **PowerSC Trusted Surveyor** on AIX 7.2 as management server
  - **Fast Connect**
  - **Performance Toolbox**

AIX® 7.2

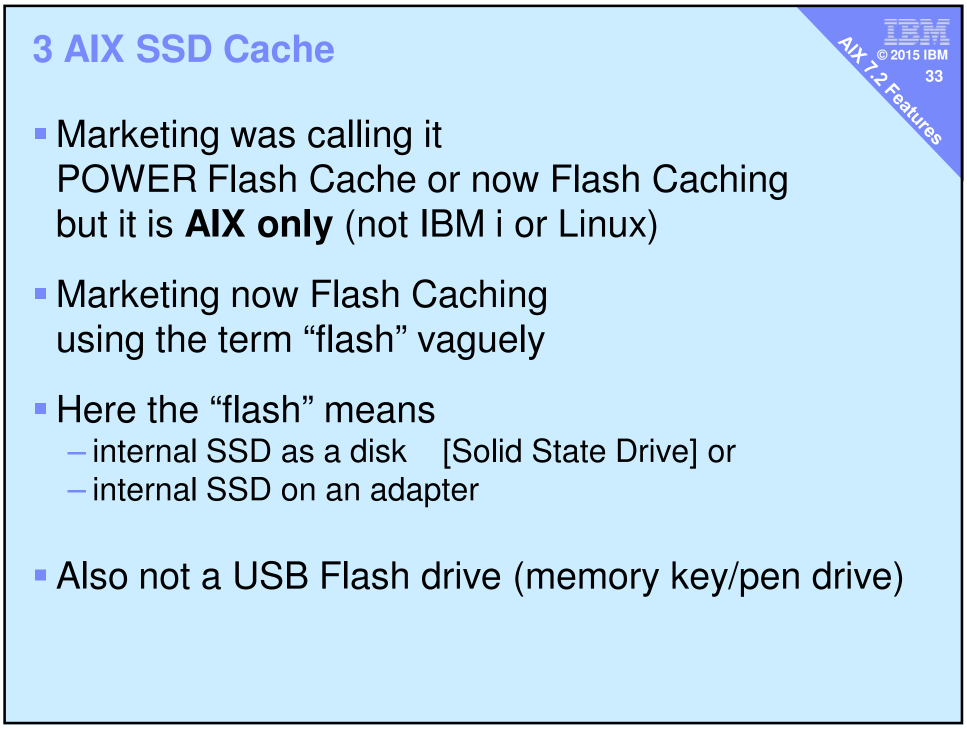


IBM © 2015 IBM 32

AIX 7.2 Features

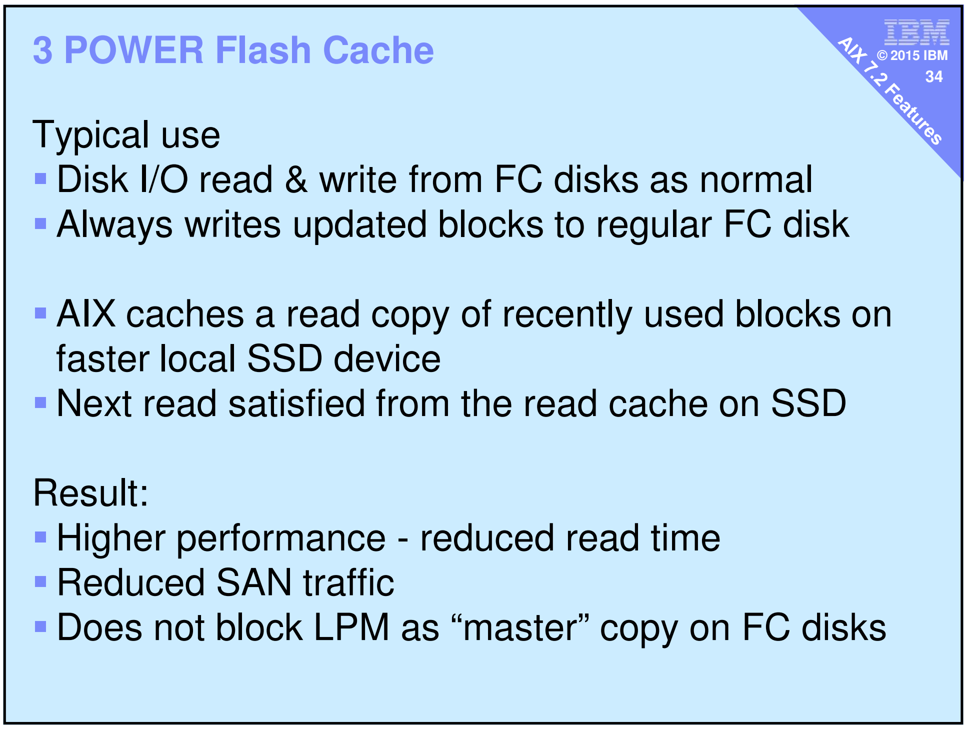
3
POWER Flash Cache think SSD

"cache_mgt" manual page states:
Manages the infrastructure that provides caching on the solid-state drive (SSD) devices



3 AIX SSD Cache

- Marketing was calling it
  POWER Flash Cache or now Flash Caching
  but it is **AIX only** (not IBM i or Linux)
- Marketing now Flash Caching
  using the term “flash” vaguely
- Here the “flash” means
  - internal SSD as a disk [Solid State Drive] or
  - internal SSD on an adapter
- Also not a USB Flash drive (memory key/pen drive)



3 POWER Flash Cache

Typical use

- Disk I/O read & write from FC disks as normal
- Always writes updated blocks to regular FC disk

- AIX caches a read copy of recently used blocks on faster local SSD device
- Next read satisfied from the read cache on SSD

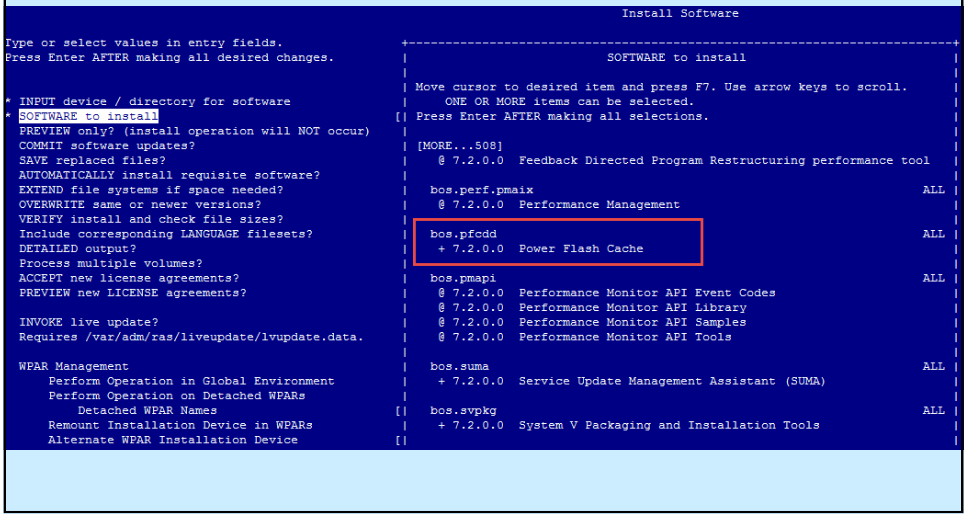
Result:

- Higher performance - reduced read time
- Reduced SAN traffic
- Does not block LPM as “master” copy on FC disks

3 POWER Flash Cache

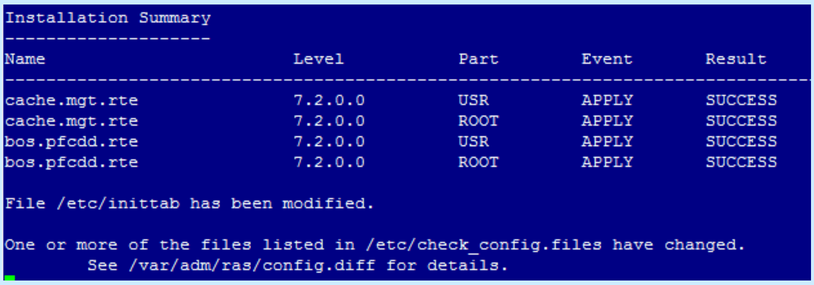
- Details
  - Workloads can be using physical storage or storage provisioned through FC, VIOS+vSCSI or VIOS+NPIV
  - Cache devices can be attached directly or provisioned through VIOS (vSCSI)
  - User may target individual or group of disks to be cached on AIX 7.2
  - Partition using a cache may use LPM with or without a locally attached flash
- Benefits
  - Most applications - higher throughput & lower latency
  - Completely invisible to applications

Non-default package
If cache_mgt “not found”
then install it from AIX DVD media



```
                                                                                   Install Software

Type or select values in entry fields.               +---------------------------------------------------------------------------+
Press Enter AFTER making all desired changes.        |                           SOFTWARE to install                             |
                                                     |                                                                           |
                                                     | Move cursor to desired item and press F7. Use arrow keys to scroll.       |
* INPUT device / directory for software              |     ONE OR MORE items can be selected.                                    |
* SOFTWARE to install                               [| Press Enter AFTER making all selections.                                  |
  PREVIEW only? (install operation will NOT occur)   |                                                                           |
  COMMIT software updates?                           | [MORE...508]                                                              |
  SAVE replaced files?                               |   @ 7.2.0.0  Feedback Directed Program Restructuring performance tool     |
  AUTOMATICALLY install requisite software?          |                                                                           |
  EXTEND file systems if space needed?               |  bos.perf.pmaix                                                       ALL |
  OVERWRITE same or newer versions?                  |   @ 7.2.0.0  Performance Management                                       |
  VERIFY install and check file sizes?               |                                                                           |
  Include corresponding LANGUAGE filesets?           |  bos.pfcdd                                                            ALL |
  DETAILED output?                                   |   + 7.2.0.0  Power Flash Cache                                            |
  Process multiple volumes?                          |                                                                           |
  ACCEPT new license agreements?                     |  bos.pmapi                                                            ALL |
  PREVIEW new LICENSE agreements?                    |   @ 7.2.0.0  Performance Monitor API Event Codes                          |
                                                     |   @ 7.2.0.0  Performance Monitor API Library                              |
  INVOKE live update?                                |   @ 7.2.0.0  Performance Monitor API Samples                              |
  Requires /var/adm/ras/liveupdate/lvupdate.data.    |   @ 7.2.0.0  Performance Monitor API Tools                                |
                                                     |                                                                           |
  WPAR Management                                    |  bos.suma                                                             ALL |
      Perform Operation in Global Environment        |   + 7.2.0.0  Service Update Management Assistant (SUMA)                   |
      Perform Operation on Detached WPARs            |                                                                           |
          Detached WPAR Names                       [|  bos.svpkg                                                            ALL |
      Remount Installation Device in WPARs           |   + 7.2.0.0  System V Packaging and Installation Tools                    |
      Alternate WPAR Installation Device            [|                                                                           |
```

Also brings in cache.mgt.rte package
No reboot needed

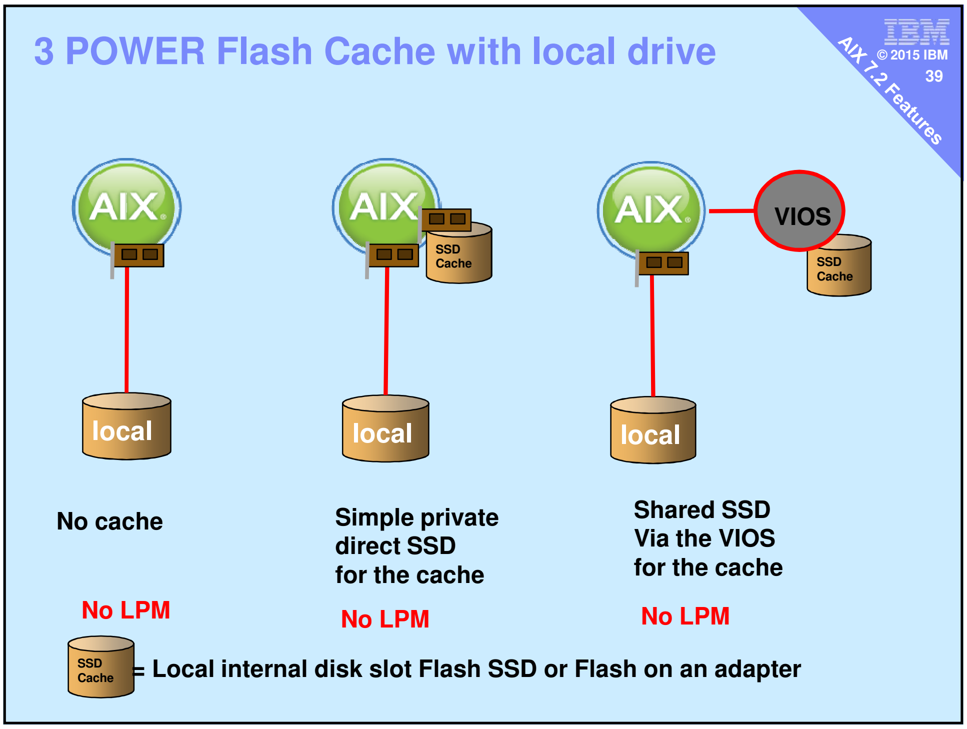


```
Installation Summary
--------------------
Name                        Level           Part        Event       Result
-------------------------------------------------------------------------------
cache.mgt.rte               7.2.0.0         USR         APPLY       SUCCESS
cache.mgt.rte               7.2.0.0         ROOT        APPLY       SUCCESS
bos.pfcdd.rte               7.2.0.0         USR         APPLY       SUCCESS
bos.pfcdd.rte               7.2.0.0         ROOT        APPLY       SUCCESS

File /etc/inittab has been modified.

One or more of the files listed in /etc/check_config.files have changed.
        See /var/adm/ras/config.diff for details.
```

Manual page for cache_mgt

- https://www-01.ibm.com/support/knowledgecenter/ssw_aix_72/com.ibm.aix.cmds1/cache_mgt.htm

- Google: **cache_mgt**



3 POWER Flash Cache with local drive

No cache

No LPM

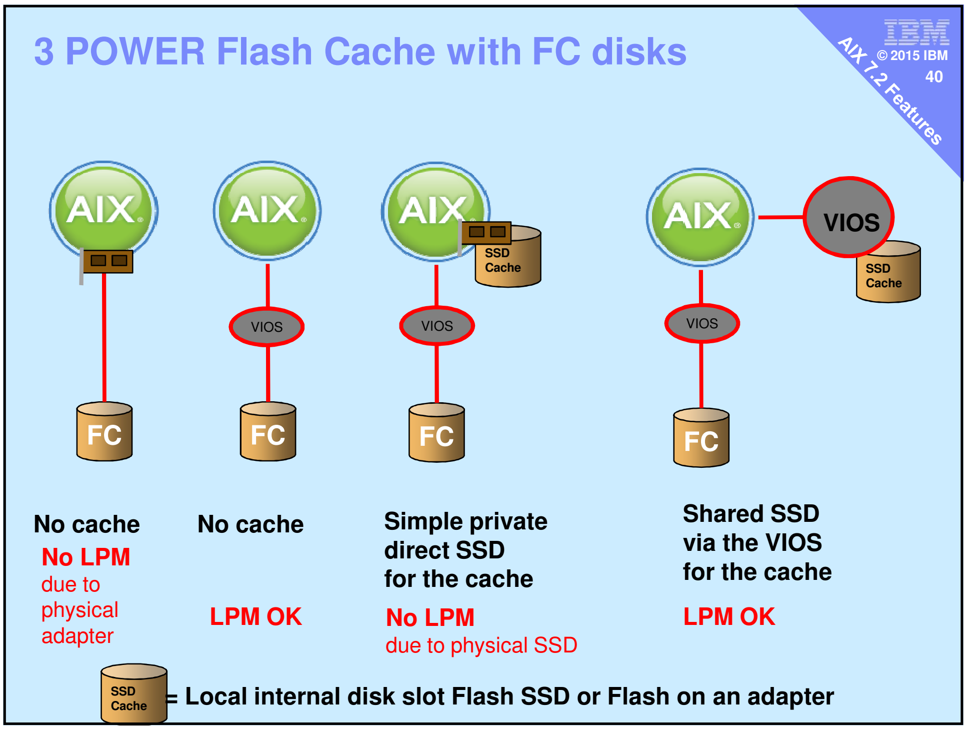
Simple private
direct SSD
for the cache

No LPM

Shared SSD
Via the VIOS
for the cache

No LPM

SSD Cache = Local internal disk slot Flash SSD or Flash on an adapter



3 POWER Flash Cache with FC disks

No cache

No LPM
due to
physical
adapter

No cache

LPM OK

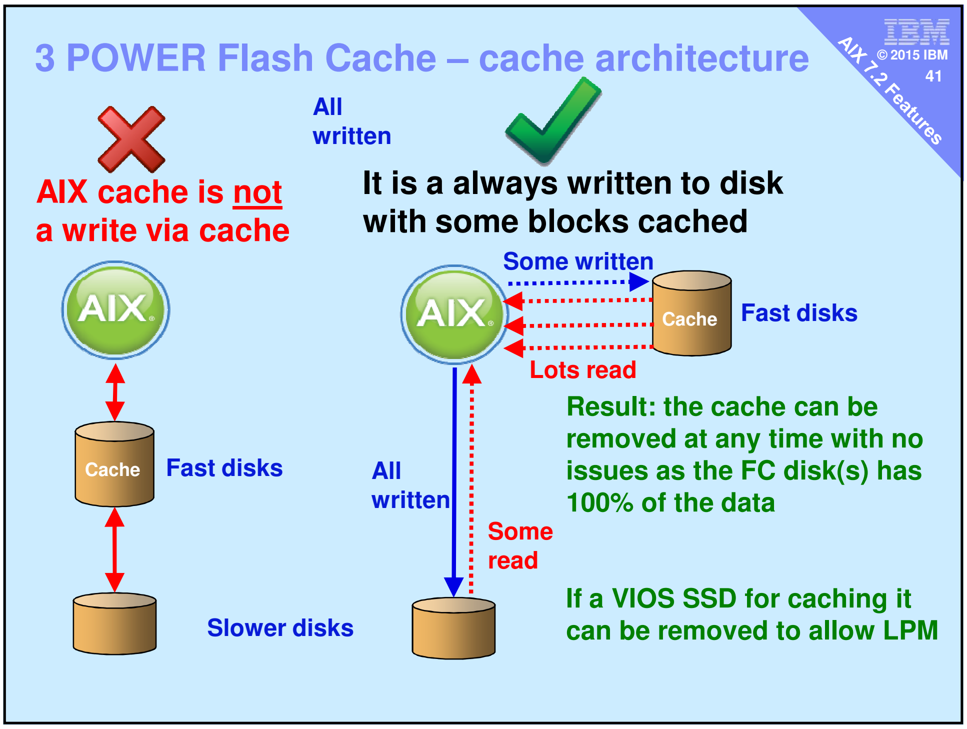
Simple private
direct SSD
for the cache

No LPM
due to physical SSD

Shared SSD
via the VIOS
for the cache

LPM OK

SSD Cache = Local internal disk slot Flash SSD or Flash on an adapter



3 POWER Flash Cache – cache architecture

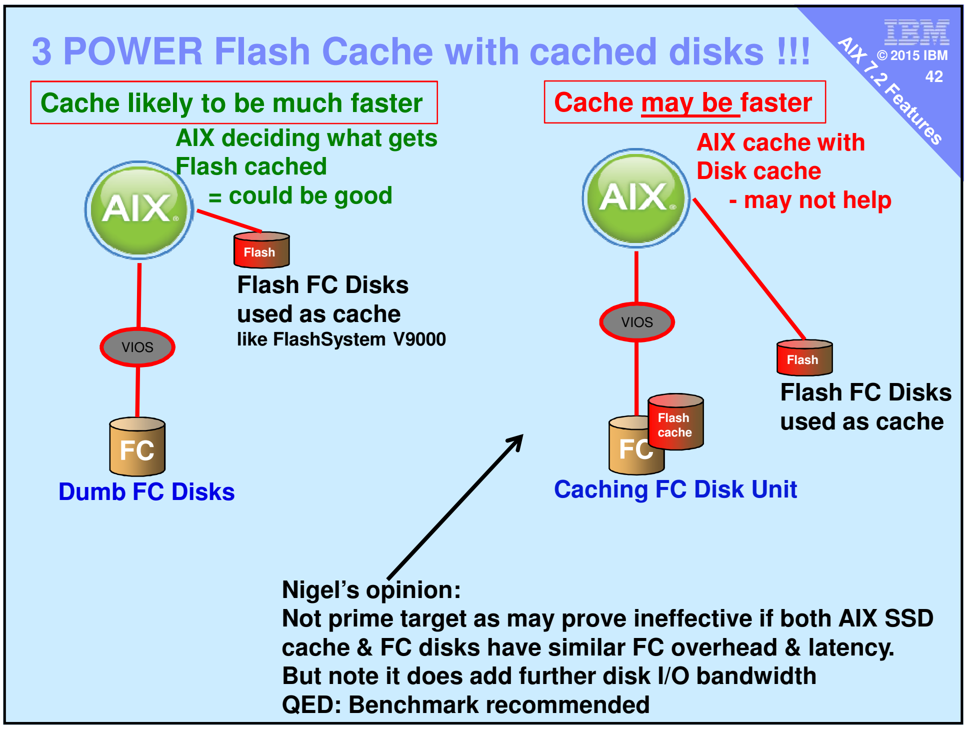
IBM © 2015 IBM 41 AIX 7.2 Features

AIX cache is not a write via cache

It is a always written to disk with some blocks cached

Result: the cache can be removed at any time with no issues as the FC disk(s) has 100% of the data

If a VIOS SSD for caching it can be removed to allow LPM



3 POWER Flash Cache with cached disks !!!

IBM © 2015 IBM 42 AIX 7.2 Features

Cache likely to be much faster

AIX deciding what gets Flash cached = could be good

Flash FC Disks used as cache like FlashSystem V9000

Dumb FC Disks

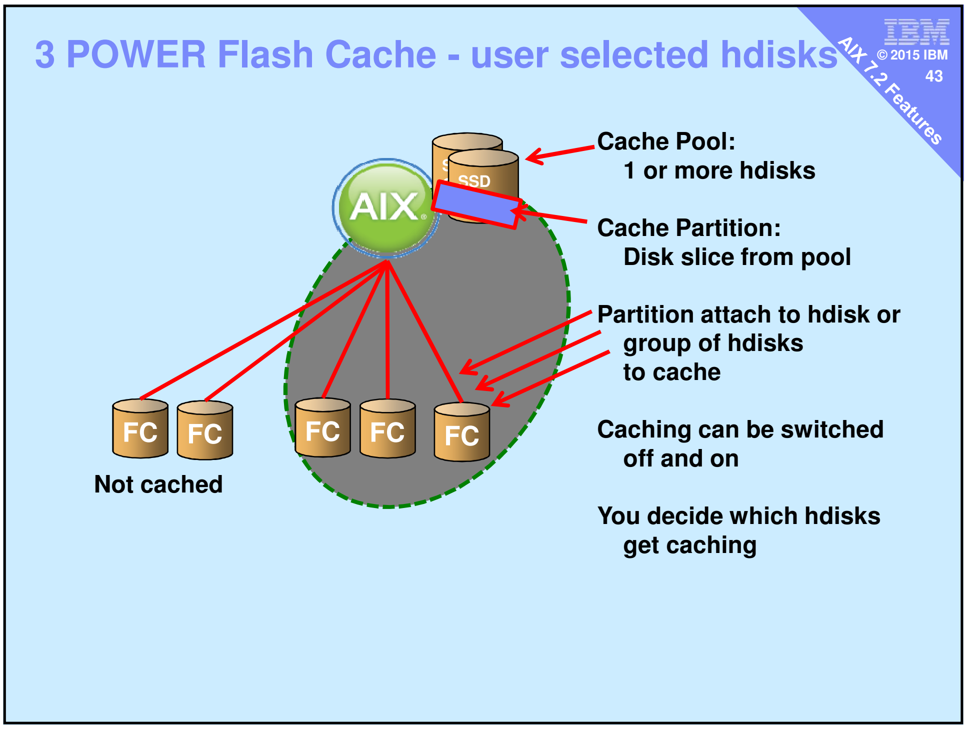
Cache may be faster

AIX cache with Disk cache - may not help

Flash FC Disks used as cache

Caching FC Disk Unit

Nigel's opinion:
Not prime target as may prove ineffective if both AIX SSD cache & FC disks have similar FC overhead & latency.
But note it does add further disk I/O bandwidth
QED: Benchmark recommended



3 POWER Flash Cache - user selected hdisks

Not cached

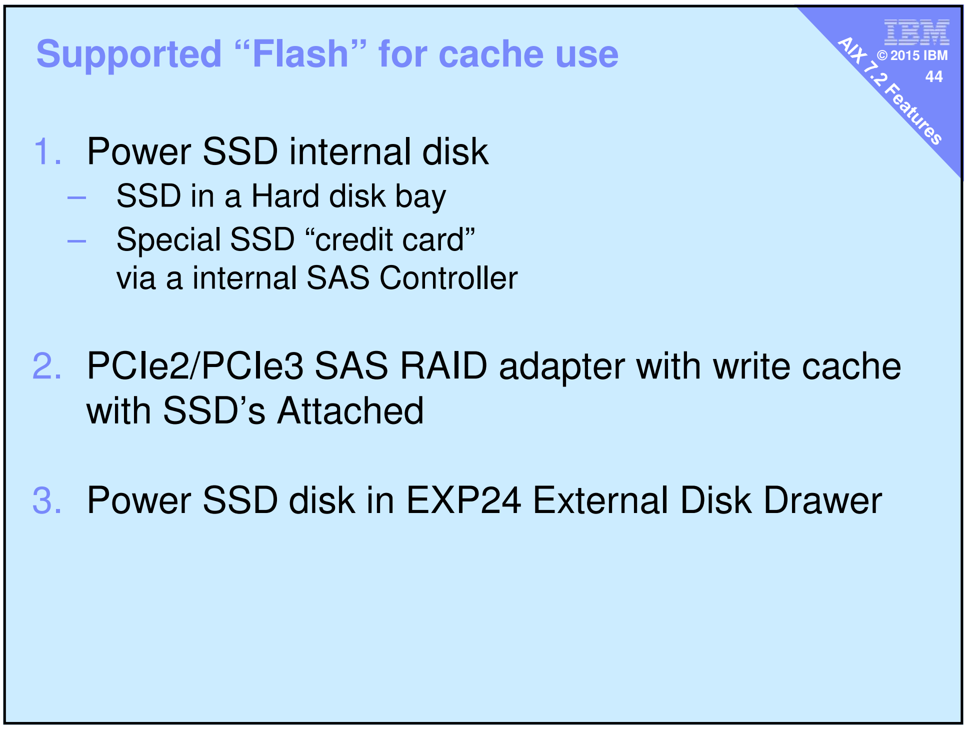
Cache Pool:
1 or more hdisks

Cache Partition:
Disk slice from pool

Partition attach to hdisk or
group of hdisks
to cache

Caching can be switched
off and on

You decide which hdisks
get caching

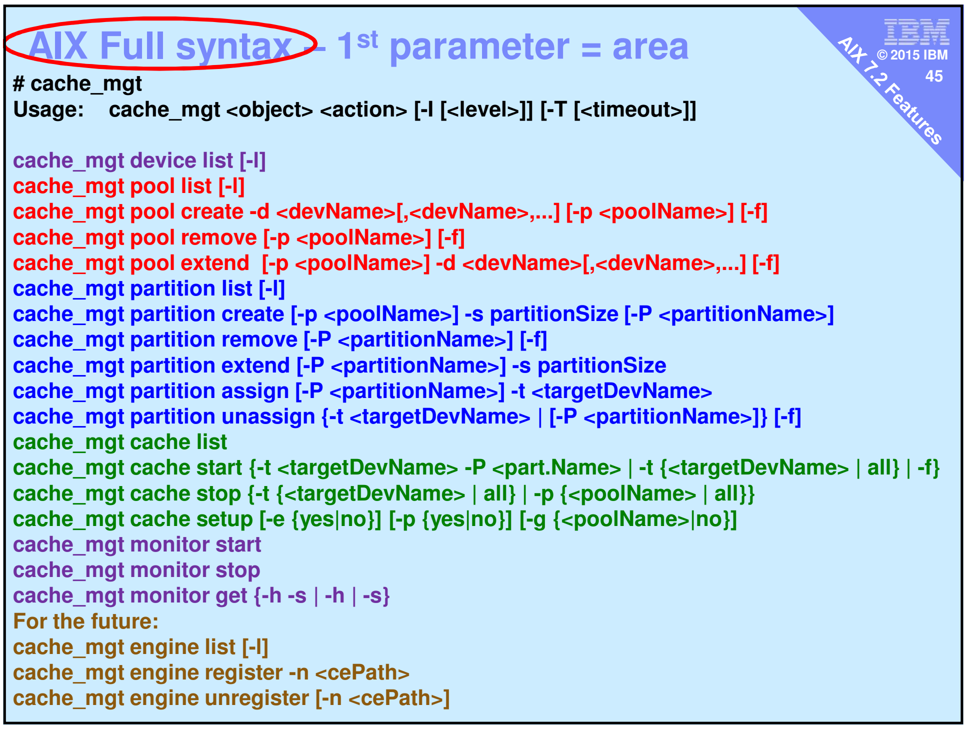


Supported “Flash” for cache use

1. Power SSD internal disk
   - SSD in a Hard disk bay
   - Special SSD “credit card”
     via a internal SAS Controller

2. PCIe2/PCIe3 SAS RAID adapter with write cache with SSD’s Attached

3. Power SSD disk in EXP24 External Disk Drawer



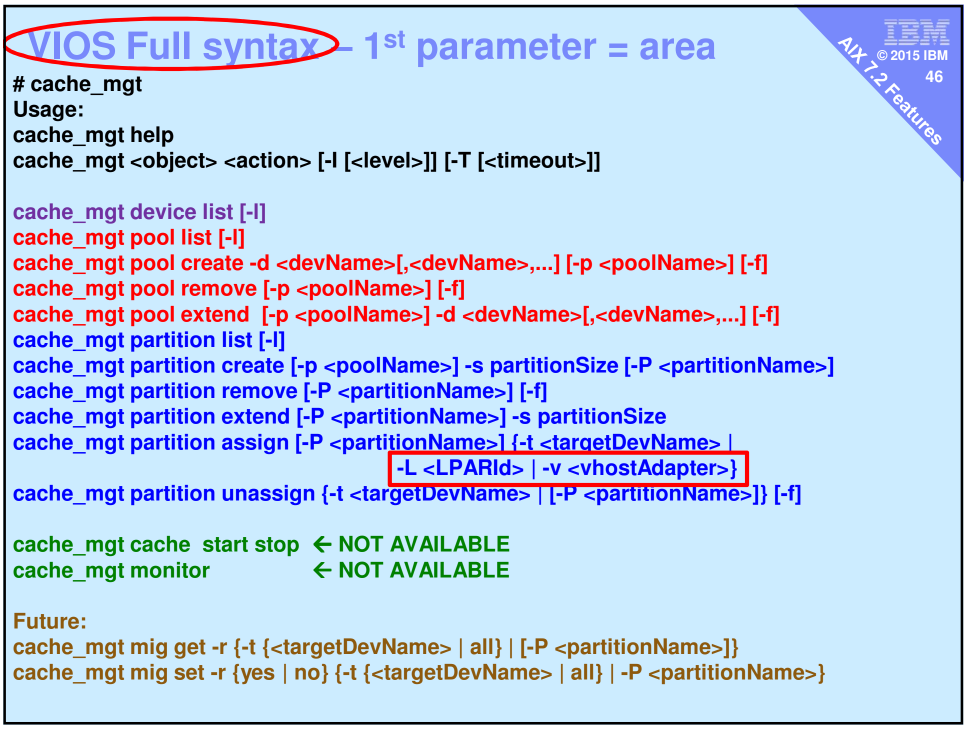
AIX Full syntax – 1st parameter = area

AIX 7.2 Features

© 2015 IBM

45

```
# cache_mgt
Usage:    cache_mgt <object> <action> [-l [<level>]] [-T [<timeout>]]

cache_mgt device list [-l]
cache_mgt pool list [-l]
cache_mgt pool create -d <devName>[,<devName>,...] [-p <poolName>] [-f]
cache_mgt pool remove [-p <poolName>] [-f]
cache_mgt pool extend  [-p <poolName>] -d <devName>[,<devName>,...] [-f]
cache_mgt partition list [-l]
cache_mgt partition create [-p <poolName>] -s partitionSize [-P <partitionName>]
cache_mgt partition remove [-P <partitionName>] [-f]
cache_mgt partition extend [-P <partitionName>] -s partitionSize
cache_mgt partition assign [-P <partitionName>] -t <targetDevName>
cache_mgt partition unassign {-t <targetDevName> | [-P <partitionName>]} [-f]
cache_mgt cache list
cache_mgt cache start {-t <targetDevName> -P <part.Name> | -t {<targetDevName> | all} | -f}
cache_mgt cache stop {-t {<targetDevName> | all} | -p {<poolName> | all}}
cache_mgt cache setup [-e {yes|no}] [-p {yes|no}] [-g {<poolName>|no}]
cache_mgt monitor start
cache_mgt monitor stop
cache_mgt monitor get {-h -s | -h | -s}
For the future:
cache_mgt engine list [-l]
cache_mgt engine register -n <cePath>
cache_mgt engine unregister [-n <cePath>]
```



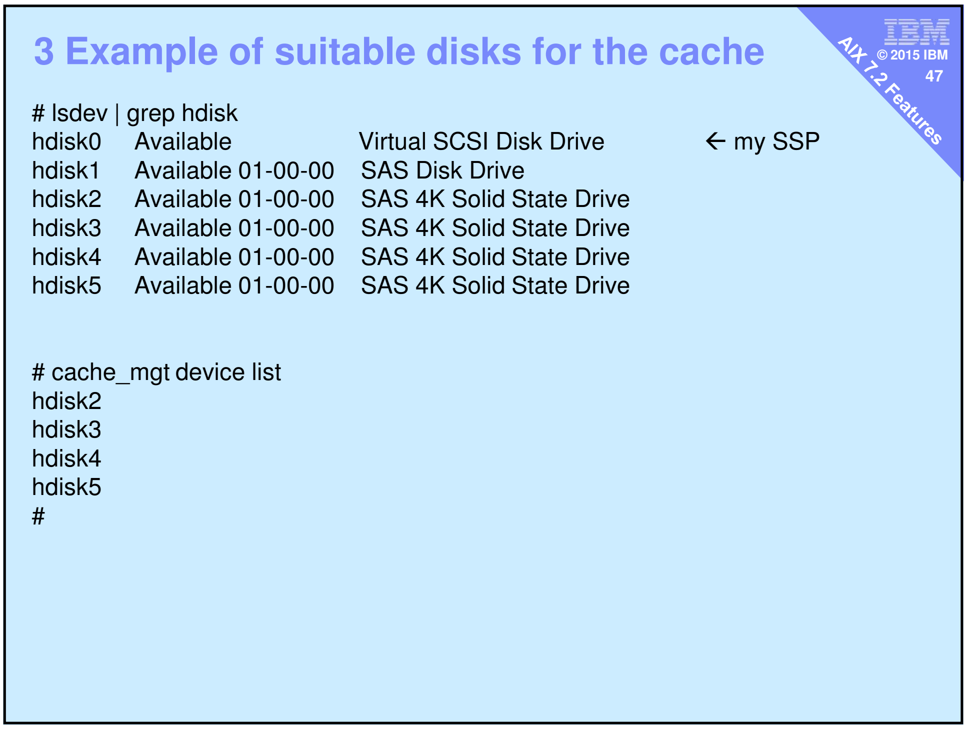
VIOS Full syntax – 1st parameter = area

AIX 7.2 Features

© 2015 IBM

46

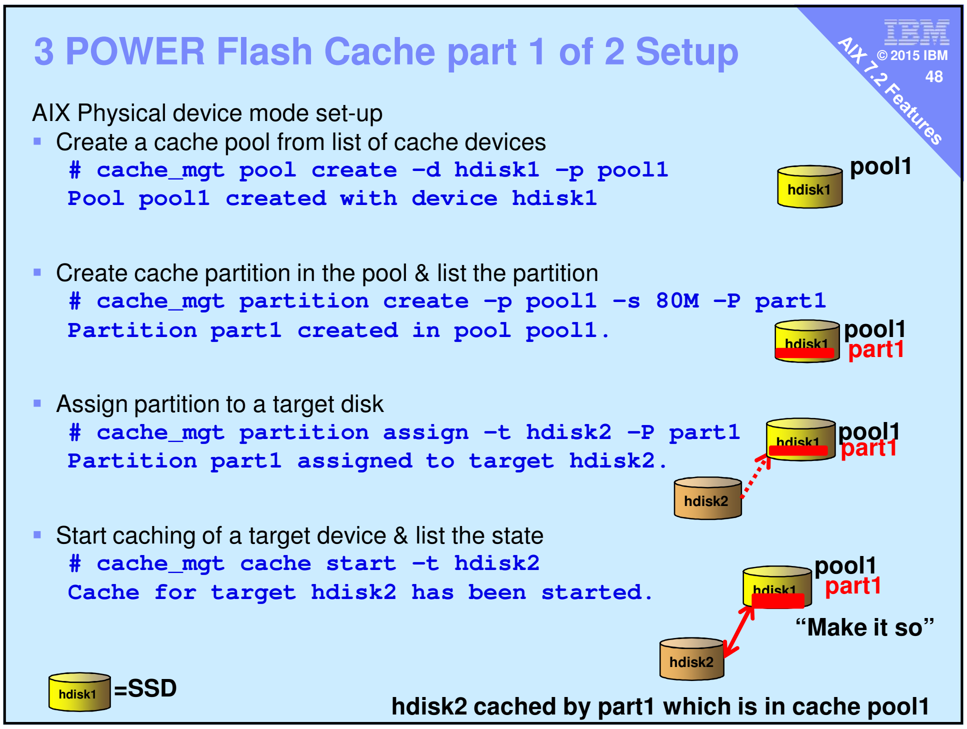
```
# cache_mgt
Usage:
cache_mgt help
cache_mgt <object> <action> [-l [<level>]] [-T [<timeout>]]

cache_mgt device list [-l]
cache_mgt pool list [-l]
cache_mgt pool create -d <devName>[,<devName>,...] [-p <poolName>] [-f]
cache_mgt pool remove [-p <poolName>] [-f]
cache_mgt pool extend  [-p <poolName>] -d <devName>[,<devName>,...] [-f]
cache_mgt partition list [-l]
cache_mgt partition create [-p <poolName>] -s partitionSize [-P <partitionName>]
cache_mgt partition remove [-P <partitionName>] [-f]
cache_mgt partition extend [-P <partitionName>] -s partitionSize
cache_mgt partition assign [-P <partitionName>] {-t <targetDevName> |
                                       -L <LPARId> | -v <vhostAdapter>}
cache_mgt partition unassign {-t <targetDevName> | [-P <partitionName>]} [-f]

cache_mgt cache  start stop  ← NOT AVAILABLE
cache_mgt monitor            ← NOT AVAILABLE

Future:
cache_mgt mig get -r {-t {<targetDevName> | all} | [-P <partitionName>]}
cache_mgt mig set -r {yes | no} {-t {<targetDevName> | all} | -P <partitionName>}
```



3 Example of suitable disks for the cache

AIX 7.2 Features
© 2015 IBM
47

```
# lsdev | grep hdisk
hdisk0    Available              Virtual SCSI Disk Drive        ← my SSP
hdisk1    Available 01-00-00     SAS Disk Drive
hdisk2    Available 01-00-00     SAS 4K Solid State Drive
hdisk3    Available 01-00-00     SAS 4K Solid State Drive
hdisk4    Available 01-00-00     SAS 4K Solid State Drive
hdisk5    Available 01-00-00     SAS 4K Solid State Drive


# cache_mgt device list
hdisk2
hdisk3
hdisk4
hdisk5
#
```



3 POWER Flash Cache part 1 of 2 Setup

AIX 7.2 Features
© 2015 IBM
48

AIX Physical device mode set-up

- Create a cache pool from list of cache devices

```
# cache_mgt pool create -d hdisk1 -p pool1
Pool pool1 created with device hdisk1
```

- Create cache partition in the pool & list the partition

```
# cache_mgt partition create -p pool1 -s 80M -P part1
Partition part1 created in pool pool1.
```

- Assign partition to a target disk

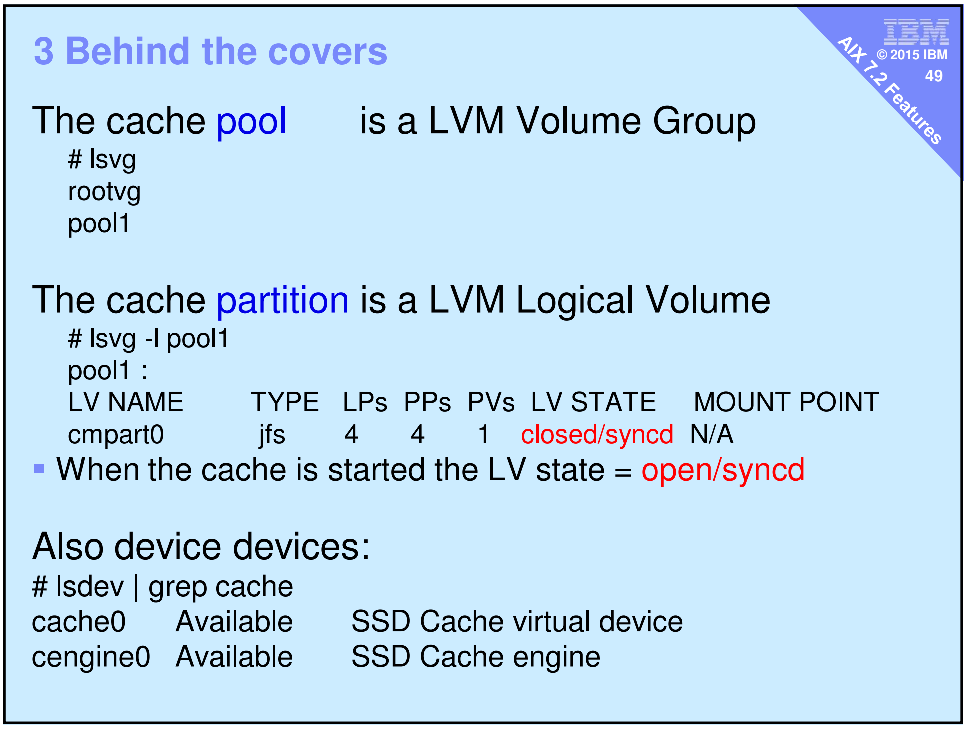
```
# cache_mgt partition assign -t hdisk2 -P part1
Partition part1 assigned to target hdisk2.
```

- Start caching of a target device & list the state

```
# cache_mgt cache start -t hdisk2
Cache for target hdisk2 has been started.
```

“Make it so”

hdisk2 cached by part1 which is in cache pool1



3 Behind the covers

AIX 7.2 Features
© 2015 IBM
49

The cache pool is a LVM Volume Group

```
# lsvg
rootvg
pool1
```

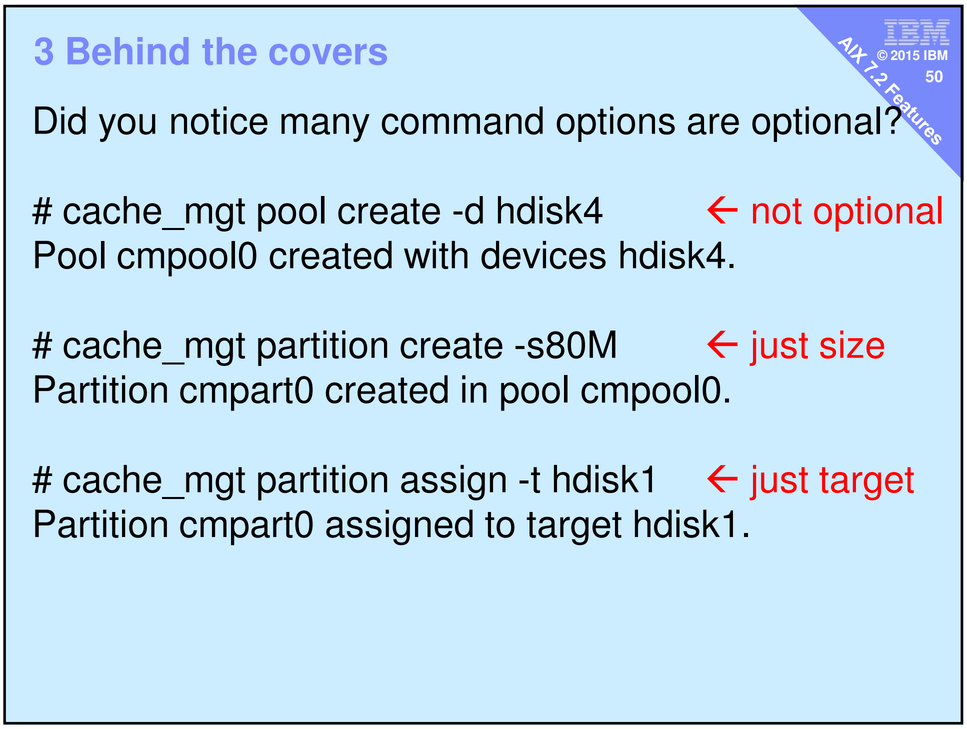
The cache partition is a LVM Logical Volume

```
# lsvg -l pool1
pool1 :
LV NAME        TYPE   LPs  PPs  PVs  LV STATE      MOUNT POINT
cmpart0         jfs     4     4     1   closed/syncd  N/A
```

- When the cache is started the LV state = open/syncd

Also device devices:

```
# lsdev | grep cache
cache0      Available      SSD Cache virtual device
cengine0   Available      SSD Cache engine
```



3 Behind the covers

AIX 7.2 Features
© 2015 IBM
50

Did you notice many command options are optional?

```
# cache_mgt pool create -d hdisk4
Pool cmpool0 created with devices hdisk4.
```

← not optional

```
# cache_mgt partition create -s80M
Partition cmpart0 created in pool cmpool0.
```

← just size

```
# cache_mgt partition assign -t hdisk1
Partition cmpart0 assigned to target hdisk1.
```

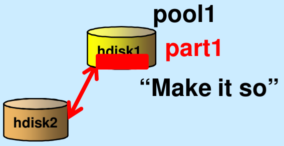
← just target

3 POWER Flash Cache part 2 of 2 Admin

AIX Physical device mode admin

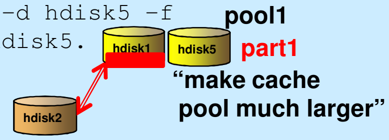
- List the state

```
# cache_mgt pool list
pool1, hdisk1
# cache_mgt partition list -l
part1,pool1
# cache_mgt cache list
hdisk2,part1 active
```



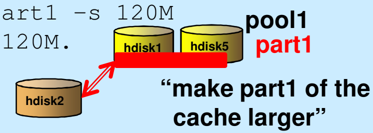
- Grow the cache pool

```
# cache_mgt pool extend -p pool1 -d hdisk5 -f
Pool pool1 extended with device hdisk5.
```



- Extend an existing cache partition size

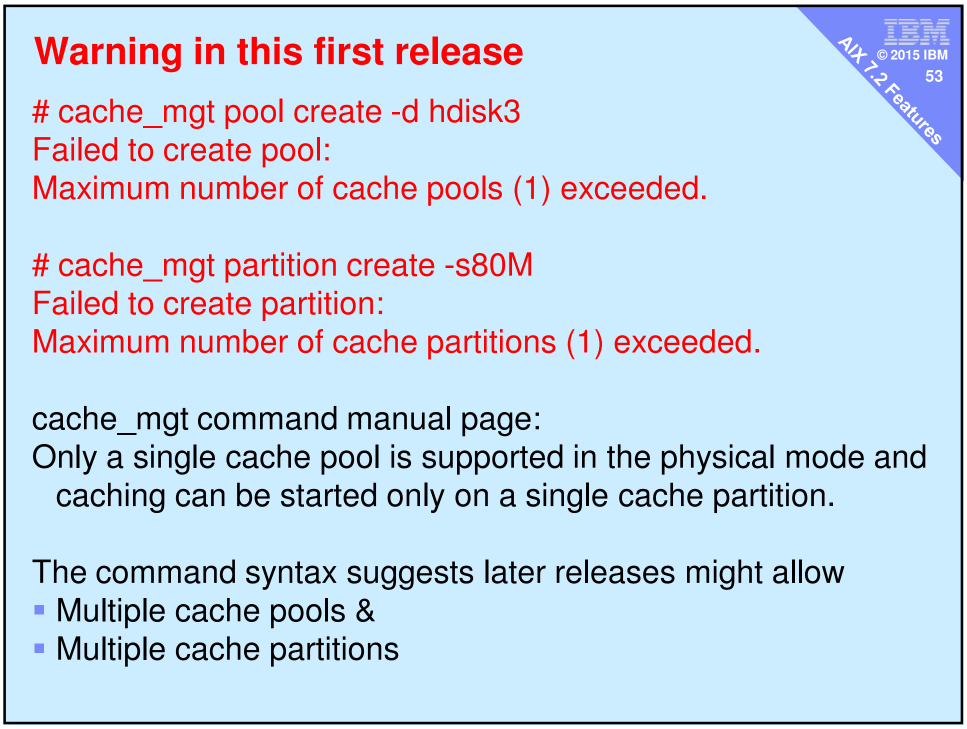
```
# cache_mgt partition extend -P part1 -s 120M
Partition part1 extended to size 120M.
```



3 The un-make commands

There is also the undo commands

```
# cache_mgt cache stop -t  hdisk2
# cache_mgt cache stop -t  all

# cache_mgt partition unassign -t hdisk2
# cache_mgt partition remove ...

# cache_mgt pool remove ...
```



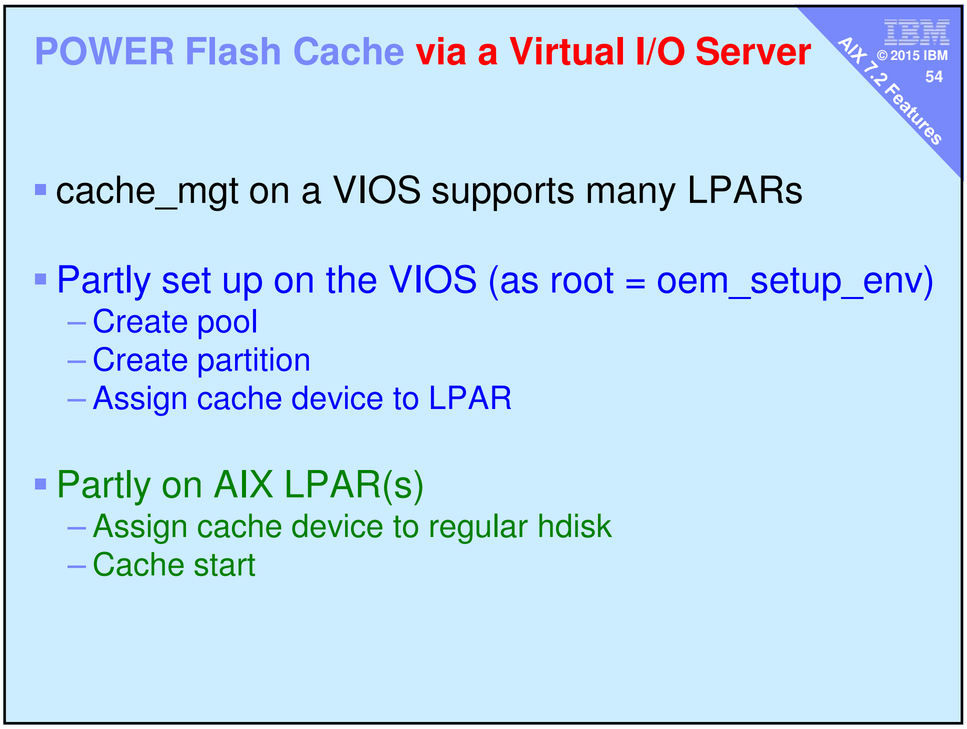
Warning in this first release

```
# cache_mgt pool create -d hdisk3
Failed to create pool:
Maximum number of cache pools (1) exceeded.

# cache_mgt partition create -s80M
Failed to create partition:
Maximum number of cache partitions (1) exceeded.
```

cache_mgt command manual page:
Only a single cache pool is supported in the physical mode and caching can be started only on a single cache partition.

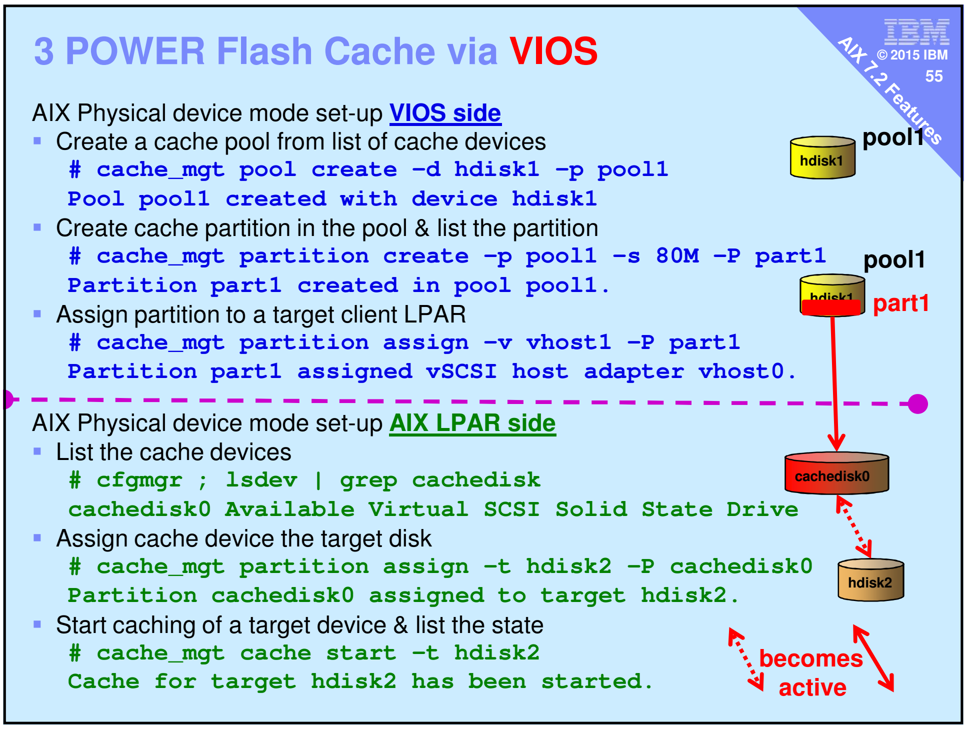
The command syntax suggests later releases might allow

- Multiple cache pools &
- Multiple cache partitions



POWER Flash Cache via a Virtual I/O Server

- cache_mgt on a VIOS supports many LPARs
- Partly set up on the VIOS (as root = oem_setup_env)
  - Create pool
  - Create partition
  - Assign cache device to LPAR
- Partly on AIX LPAR(s)
  - Assign cache device to regular hdisk
  - Cache start



3 POWER Flash Cache via VIOS

AIX Physical device mode set-up **VIOS side**

- Create a cache pool from list of cache devices

```
# cache_mgt pool create -d hdisk1 -p pool1
Pool pool1 created with device hdisk1
```

- Create cache partition in the pool & list the partition

```
# cache_mgt partition create -p pool1 -s 80M -P part1
Partition part1 created in pool pool1.
```

- Assign partition to a target client LPAR

```
# cache_mgt partition assign -v vhost1 -P part1
Partition part1 assigned vSCSI host adapter vhost0.
```

AIX Physical device mode set-up **AIX LPAR side**

- List the cache devices

```
# cfgmgr ; lsdev | grep cachedisk
cachedisk0 Available Virtual SCSI Solid State Drive
```

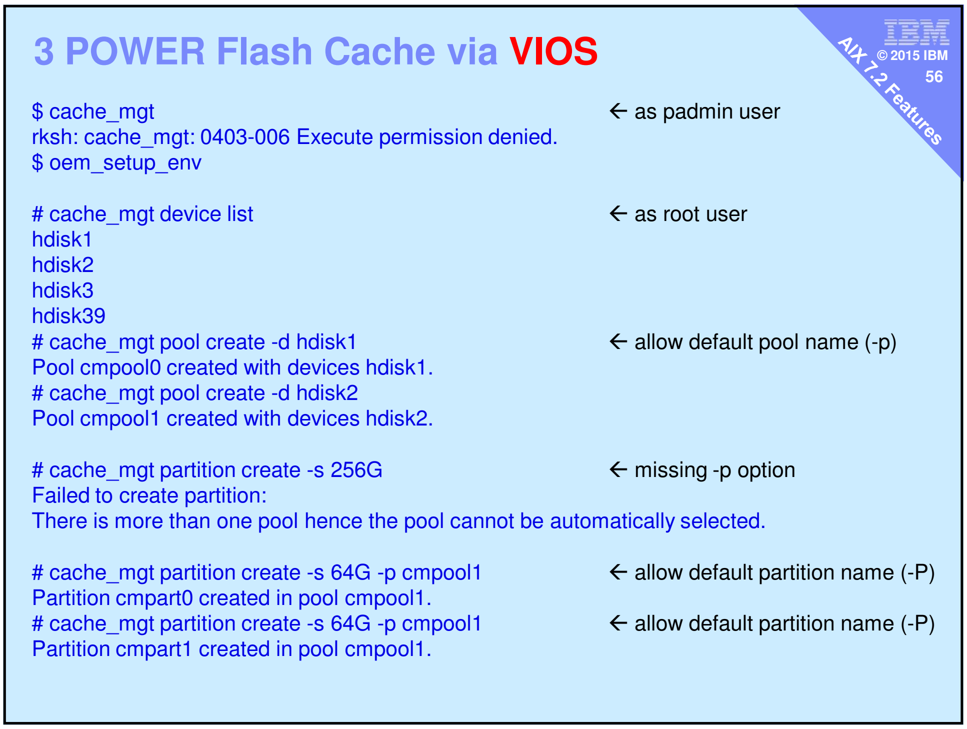
- Assign cache device the target disk

```
# cache_mgt partition assign -t hdisk2 -P cachedisk0
Partition cachedisk0 assigned to target hdisk2.
```

- Start caching of a target device & list the state

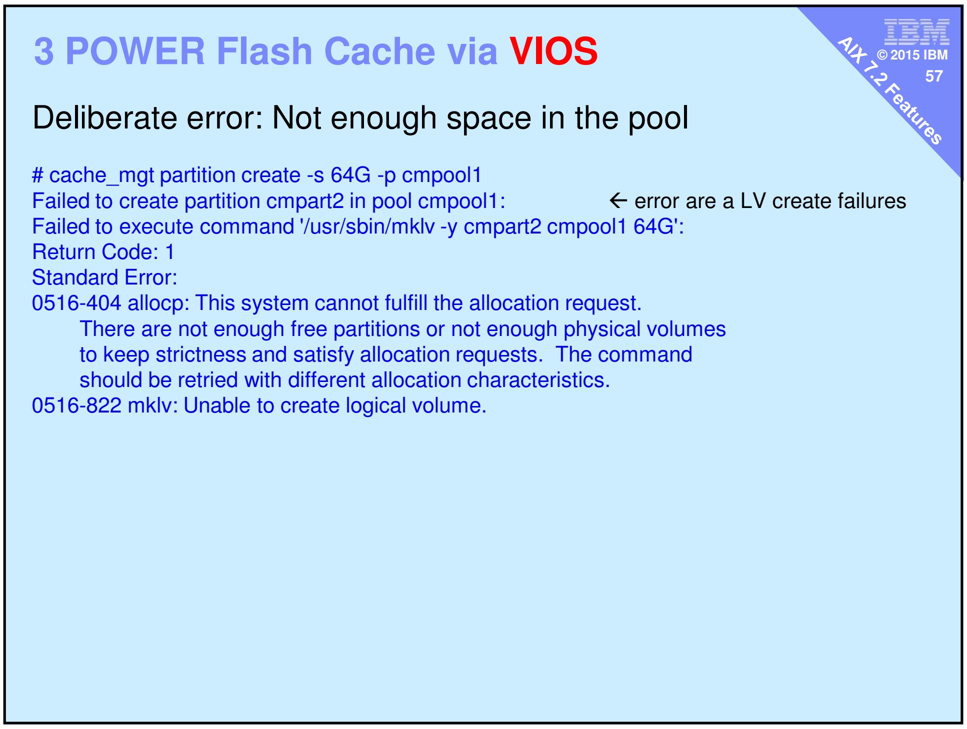
```
# cache_mgt cache start -t hdisk2
Cache for target hdisk2 has been started.
```

hdisk1
pool1
pool1
hdisk1
part1
cachedisk0
hdisk2
becomes
active



3 POWER Flash Cache via VIOS

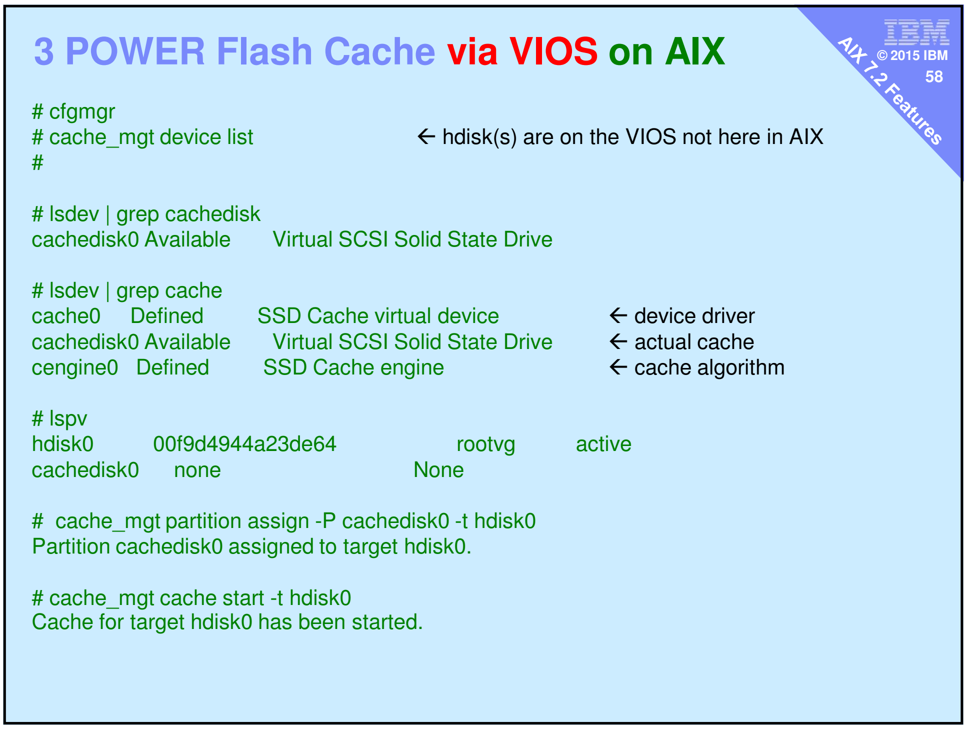
```
$ cache_mgt                                                ← as padmin user
rksh: cache_mgt: 0403-006 Execute permission denied.
$ oem_setup_env

# cache_mgt device list                                    ← as root user
hdisk1
hdisk2
hdisk3
hdisk39
# cache_mgt pool create -d hdisk1                          ← allow default pool name (-p)
Pool cmpool0 created with devices hdisk1.
# cache_mgt pool create -d hdisk2
Pool cmpool1 created with devices hdisk2.

# cache_mgt partition create -s 256G                       ← missing -p option
Failed to create partition:
There is more than one pool hence the pool cannot be automatically selected.

# cache_mgt partition create -s 64G -p cmpool1             ← allow default partition name (-P)
Partition cmpart0 created in pool cmpool1.
# cache_mgt partition create -s 64G -p cmpool1             ← allow default partition name (-P)
Partition cmpart1 created in pool cmpool1.
```



3 POWER Flash Cache via VIOS

Deliberate error: Not enough space in the pool

```
# cache_mgt partition create -s 64G -p cmpool1
Failed to create partition cmpart2 in pool cmpool1:              ← error are a LV create failures
Failed to execute command '/usr/sbin/mklv -y cmpart2 cmpool1 64G':
Return Code: 1
Standard Error:
0516-404 allocp: This system cannot fulfill the allocation request.
        There are not enough free partitions or not enough physical volumes
        to keep strictness and satisfy allocation requests.  The command
        should be retried with different allocation characteristics.
0516-822 mklv: Unable to create logical volume.
```

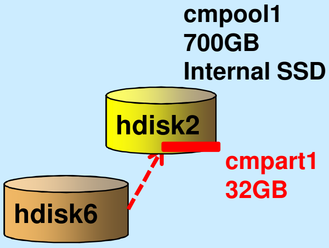


3 POWER Flash Cache via VIOS on AIX

```
# cfgmgr
# cache_mgt device list                    ← hdisk(s) are on the VIOS not here in AIX
#

# lsdev | grep cachedisk
cachedisk0 Available       Virtual SCSI Solid State Drive

# lsdev | grep cache
cache0     Defined        SSD Cache virtual device            ← device driver
cachedisk0 Available       Virtual SCSI Solid State Drive     ← actual cache
cengine0   Defined         SSD Cache engine                   ← cache algorithm

# lspv
hdisk0          00f9d4944a23de64               rootvg         active
cachedisk0      none                           None

#  cache_mgt partition assign -P cachedisk0 -t hdisk0
Partition cachedisk0 assigned to target hdisk0.

# cache_mgt cache start -t hdisk0
Cache for target hdisk0 has been started.
```

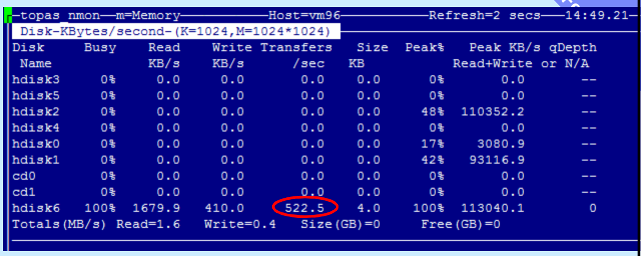
cache_mgt Cheat Sheet

- **Test Config: HDD = hdisk6 and SSD = hdisk2**
- cache_mgt device list → Output your online SSD's suitable for caching
- cache_mgt pool create -d hdisk2 -p cmpool1
- cache_mgt pool list -l
- cache_mgt partition create -p cmpool1 -s 32G -P cmpart1
- cache_mgt partition list -l
- cache_mgt partition assign -t hdisk6 -P cmpart1
- cache_mgt cache start -t hdisk6
- Cache_mgt cache stop -t hdisk6
- cache_mgt cache list
- cache_mgt monitor get



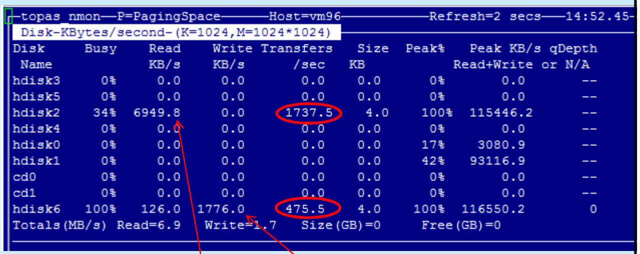
Simplistic "Does it work? test

- **Local HHD**
- 100 GB on JFS2
- Direct I/O
- 4KB blocks
- 80% read + 20% write
- Random
- 8 processes doing I/O
- 8 x 1 GB file
- at 522 IOPS



```
topas nmon——m=Memory——————————Host=vm96——————————Refresh=2 secs——14:49.21
 Disk-KBytes/second-(K=1024,M=1024*1024)
Disk      Busy    Read    Write Transfers  Size  Peak%   Peak KB/s qDepth
 Name             KB/s    KB/s      /sec   KB           Read+Write or N/A
hdisk3      0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk5      0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk2      0%     0.0     0.0       0.0   0.0    48%   110352.2      --
hdisk4      0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk0      0%     0.0     0.0       0.0   0.0    17%     3080.9      --
hdisk1      0%     0.0     0.0       0.0   0.0    42%    93116.9      --
cd0         0%     0.0     0.0       0.0   0.0     0%        0.0      --
cd1         0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk6    100%  1679.9   410.0     522.5   4.0   100%   113040.1       0
Totals(MB/s) Read=1.6  Write=0.4   Size(GB)=0    Free(GB)=0
```

- Switch on cache
- 32 GB
- At 1737 + 475 IOPS
- = 2212 IOPS



```
topas nmon——P=PagingSpace————Host=vm96——————————Refresh=2 secs——14:52.45
 Disk-KBytes/second-(K=1024,M=1024*1024)
Disk      Busy    Read    Write Transfers  Size  Peak%   Peak KB/s qDepth
 Name             KB/s    KB/s      /sec   KB           Read+Write or N/A
hdisk3      0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk5      0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk2     34%  6949.8     0.0    1737.5    4.0  100%   115446.2      --
hdisk4      0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk0      0%     0.0     0.0       0.0   0.0    17%     3080.9      --
hdisk1      0%     0.0     0.0       0.0   0.0    42%    93116.9      --
cd0         0%     0.0     0.0       0.0   0.0     0%        0.0      --
cd1         0%     0.0     0.0       0.0   0.0     0%        0.0      --
hdisk6    100%   126.0  1776.0     475.5   4.0   100%   116550.2       0
Totals(MB/s) Read=6.9  Write=1.7   Size(GB)=0    Free(GB)=0
```

Writes to real disk

Reads from cache disk

- **4.23 times faster**
- Have seen that the cache warms up over time – wait at least 5 minutes or longer

cache_mgt monitor -s

```
# cache_mgt monitor get -s

ETS Device I/O Statistics -- hdisk6
Start time of Statistics -- Sun Oct 25 14:49:53 2015
------------------------------------------------------
 Read Count:                                    788298
 Write Count:                                   197124
 Read Hit Count:                                751278
 Partial Read Hit Count:                             0
 Read Bytes Xfer:                           3228868608
 Write Bytes Xfer:                           807419904
 Read Hit Bytes Xfer:                       3077234688
 Partial Read Hit Bytes Xfer:                        0
 Promote Read Count:                       14017363968
 Promote Read Bytes Xfer:                        13368
```

Most read I/O from cache

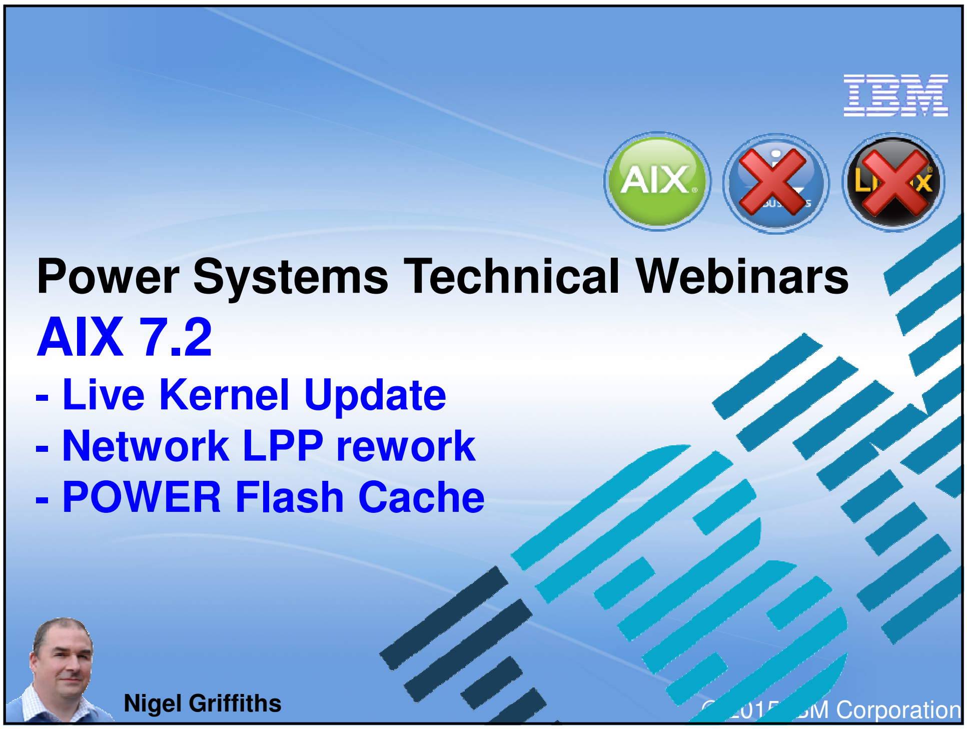
**I need to study these more during a real workload or benchmark!
- Get in touch.**

POWER Flash Cache Conclusions:

- Flexible design with cache pool & cache partitions
- SSD’s directly physical at AIX level or via VIOS
- Cache is Transparent (no application changes)
  - Always writes to the real disks as the master copy
- Cache target can be single disk or a disk group
- LPM possible
  - SSD Cache on the VIOS – just works
  - SSD Cache on AIX – remove SSD(s) before LPM
- The slower the normal disks, the bigger the effect

Notes:

- AIX LPAR min 4GB
- No “shared disks” for Workload data or for the cache
  - “Shared disk” meaning online to more then one VIOS or AIX



Power Systems Technical Webinars

AIX 7.2

- Live Kernel Update

- Network LPP rework

- POWER Flash Cache

Nigel Griffiths

Are you keeping up to date?

mr_nmon on twitter

– Only used to POWER / AIX technical content, hints, tips and links

™ 131 techie hands-on videos on **YouTube** at http://www.youtube.com/nigelargriffiths

AIXpert Blog

– Lots of mini articles & thoughts

– http://tinyurl.com/AIXpert

Also:

– http://tinyurl.com/ibmAIXVUG

– http://tinyurl.com/PowerSystemsTechnicalWebinars

An aside on internal SSD disks on my E850

- Might save you some time!

smitty→devices→DiskArray→ SASDiskArray → IBMSASDiskArrayManager →List → sissas0

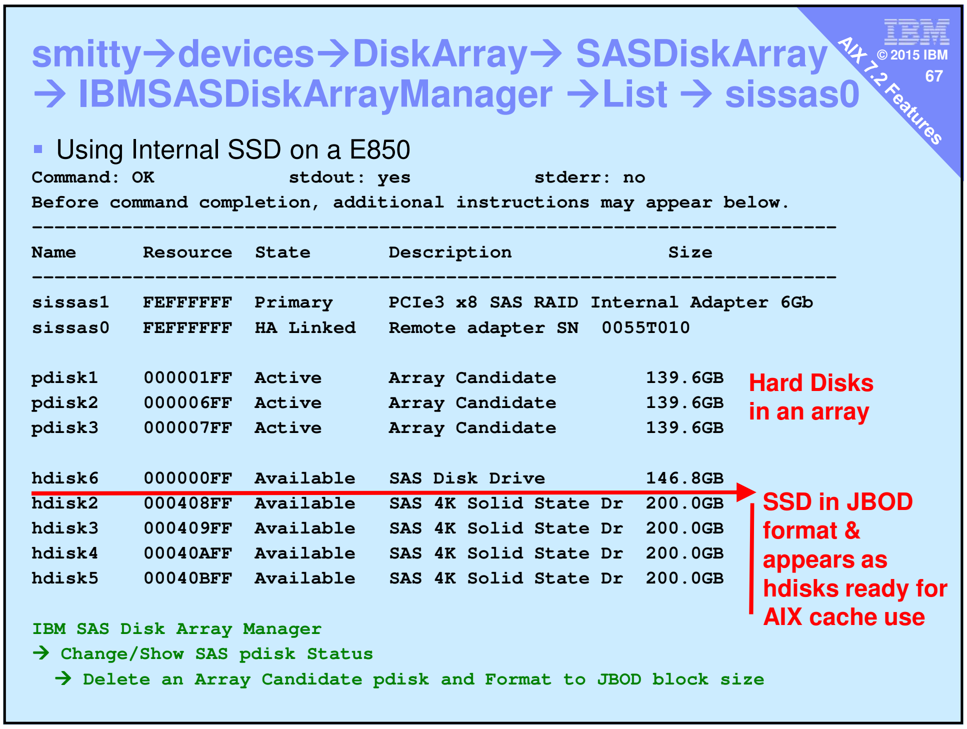
- Using Internal SSD on a E850

```
Command: OK            stdout: yes           stderr: no
Before command completion, additional instructions may appear below.
------------------------------------------------------------------------
Name      Resource  State       Description              Size
------------------------------------------------------------------------
sissas0   FEFFFFFF  Secondary   PCIe3 x8 SAS RAID Internal Adapter 6Gb
tmscsiX   FEFFFFFF  HA Linked   Remote adapter SN  0068412E
hdisk11   FC0000FF  Optimal     RAID 0 Array            139.6GB
 pdisk2   000006FF  Active      Array Member                N/A
hdisk12   FC0100FF  Optimal     RAID 0 Array            139.6GB
 pdisk3   000007FF  Active      Array Member                N/A
hdisk13   FC0200FF  Optimal     RAID 0 Array            139.6GB
 pdisk1   000001FF  Active      Array Member                N/A
hdisk14   FC0300FF  Optimal     RAID 0 Array            139.6GB
 pdisk0   000000FF  Active      Array Member                N/A

pdisk4    000408FF  Active      4K RI Array Candidate       N/A
pdisk5    000409FF  Active      4K RI Array Candidate       N/A
pdisk6    00040AFF  Active      4K RI Array Candidate       N/A
pdisk7    00040BFF  Active      4K RI Array Candidate       N/A
```

Hard Disks in an array as members (in use)

SSD's in RAID format + Array Candidate **but can't be added to an array**

=RAID Array Candidate



smitty→devices→DiskArray→ SASDiskArray → IBMSASDiskArrayManager →List → sissas0

- Using Internal SSD on a E850

```
Command: OK            stdout: yes           stderr: no
Before command completion, additional instructions may appear below.
------------------------------------------------------------------------
Name      Resource  State      Description              Size
------------------------------------------------------------------------
sissas1   FEFFFFFF  Primary    PCIe3 x8 SAS RAID Internal Adapter 6Gb
sissas0   FEFFFFFF  HA Linked  Remote adapter SN  0055T010

pdisk1    000001FF  Active     Array Candidate        139.6GB
pdisk2    000006FF  Active     Array Candidate        139.6GB
pdisk3    000007FF  Active     Array Candidate        139.6GB

hdisk6    000000FF  Available  SAS Disk Drive         146.8GB
hdisk2    000408FF  Available  SAS 4K Solid State Dr  200.0GB
hdisk3    000409FF  Available  SAS 4K Solid State Dr  200.0GB
hdisk4    00040AFF  Available  SAS 4K Solid State Dr  200.0GB
hdisk5    00040BFF  Available  SAS 4K Solid State Dr  200.0GB

IBM SAS Disk Array Manager
→ Change/Show SAS pdisk Status
  → Delete an Array Candidate pdisk and Format to JBOD block size
```

Hard Disks in an array

SSD in JBOD format & appears as hdisks ready for AIX cache use