HP-UX Kernel Crashdump Improvements

Scott Millward

Architect

Hewlett Packard

(Event Management Lab)

HP WORLD 2003
Solutions and Technology Conference & Expo

What is a kernel crash?

- A kernel crash is an abnormal shutdown of the system. This can be the result of:
  - a kernel routine calling ‘Panic’ because it has encountered an unexpected state
  - a kernel routine hits an abnormal trap (illegal pointer trap)
  - someone initiates a Transfer Of Control (TOC).
- The ideal answer is to reduce the frequency of the abnormal shutdowns
  - Kernel Hardening
  - Quality processes
  - HP-UX 11.11 has seen 6X quality gains over previous releases
  - HP-UX defect density is in the top 1% of software Vendors

What is a kernel crashdump?

- The dump process exists so that you have a way of capturing what your system was doing at the time of a crash.
- This is *not* for recovery purposes
  - processes cannot resume where they left off, following a system crash. Rather this is for analysis purposes, to help determine why the system crashed in order to prevent it from happening again.

Faults can be reduced, but not eliminated

"If a problem has no solution, it may not be a problem, but a fact, not to be solved, but to be coped with over time"

— Shimon Peres

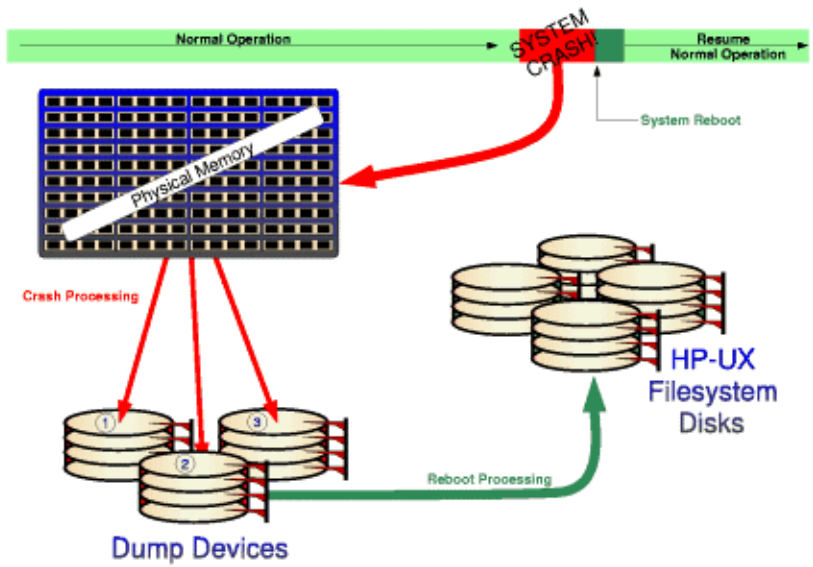
- Although the reduction in failures, as measured by Mean Time To Fail, will improve availability, some measure of failure is a fact to be coped with.
- If Availability can be quantified as

$$\frac{MTTF}{(MTTF + MTTR)}$$

- 1/3 Mean Time To Repair/Recover (MTTR) is just as good as 3X improvement in MTTF

Kernel Crash Process

- System calls panic (via trap or direct call to panic) or is stopped (via TOC)
- On the way down a memory dump is performed
- After reboot the dump image is moved from swap or the dump device to the file system during the system startup



Crashdump Configuration

- Configure Dump Devices
  - Dedicated
  - Shared with Swap
- Configure type of memory to be dumped
  - Full (ALL)
  - Selective (Subset of page classifications)
  - None
  - **Note: When the Dump Subsystem is first invoked at crash time, there is an opportunity to interact at the console to override and ask for no dump or a full dump.**

Crashdump Devices

- Three ways to configure
  - Specify in the kernel gen file
    - Run SAM's Kernel configuration area to configure dump devices
    - Manually edit the kernel configuration file
  - /etc/fstab
    - Dump devices will be dynamically configured at system initialization
    - Dump device configuration will be dynamically updated with crashconf –a, which reads the /etc/fstab file.
  - Interactively via crashconf
    - /sbin/**crashconf** [-arv] [-i|-e *class*] ... [*device*...]

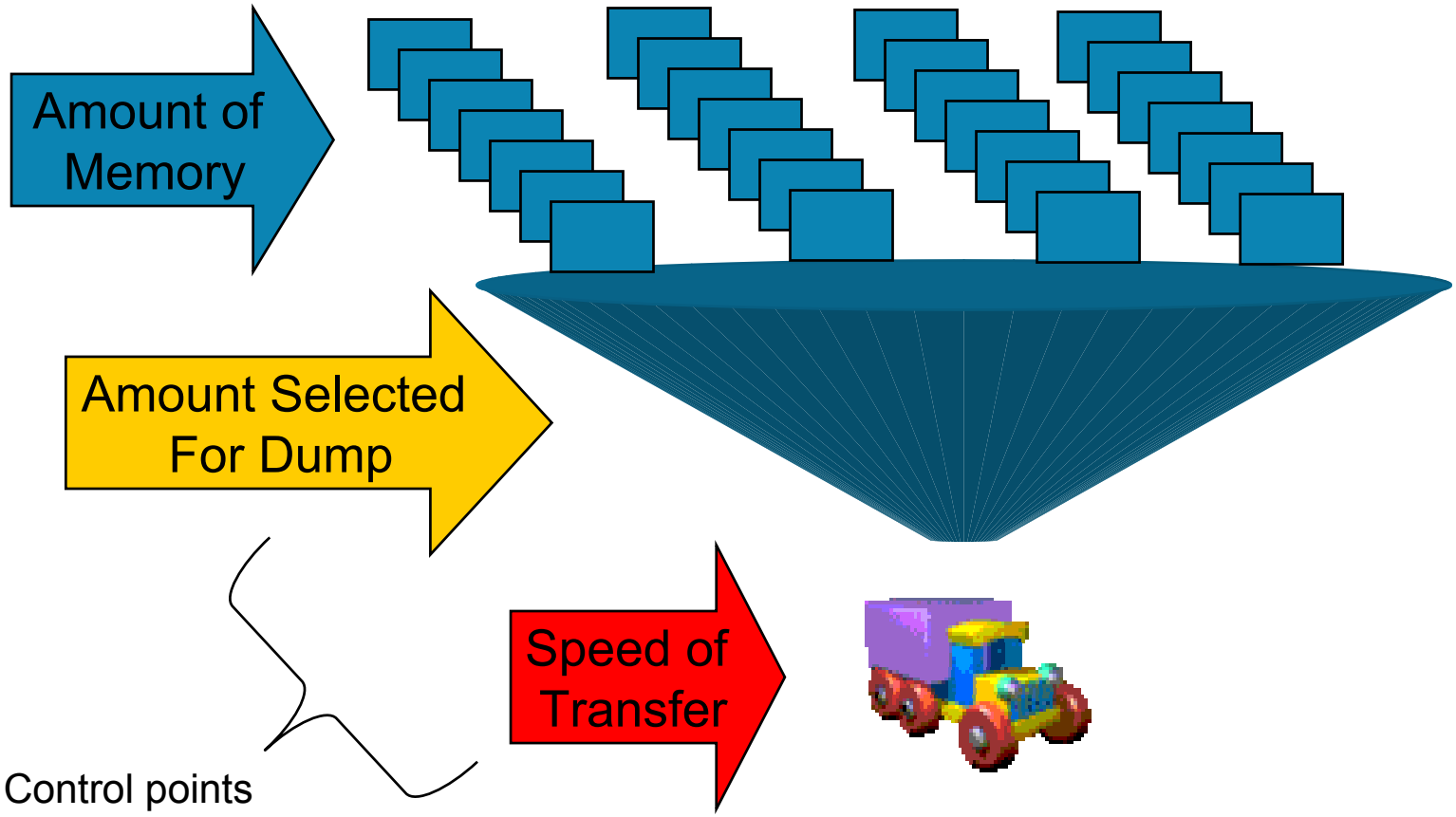
Dump device considerations

- Configuring dump devices that also page(Swap)
  - Have a device configured as both paging (Swapping) and as a dump device makes maximum usage of disks
  - Dump order: The last device configured is the first used, so it is best to configure the shared device first, so it is one of the last used.
- Configuring dump devices as dedicated dump devices
  - Savecrash will run in the foreground until all the dump devices that are also configured as paging devices, have completed their transfer to the file system space. Once the paging devices are completed, then the remaining dump devices will run in the background, allowing the system to continue the boot process.
  - If savecrash is not configured to run at bootup, any dump device that is also used as a paging device runs the risk of having the dump overwritten by the paging activity.
  - If system recovery time is critical to you, do not configure the primary paging device as a dump device

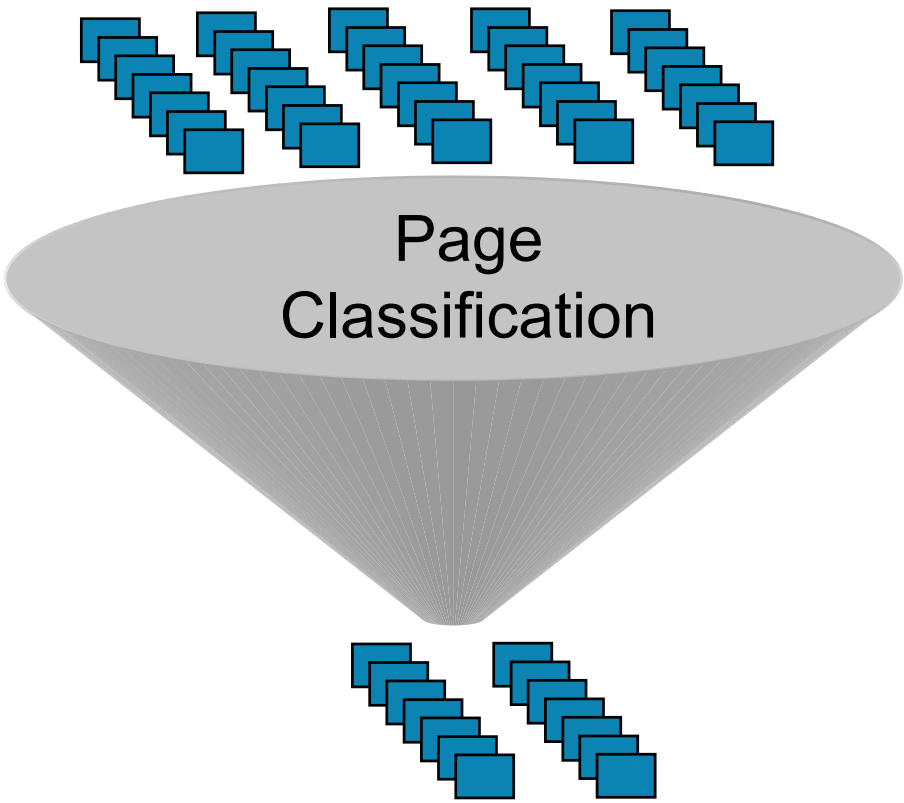
Problem

- Maximum memory configurations are increasing faster than the raw disk dump rates
  - Memory configurations are moving from 4 to 16 to 128 to 256 gigabytes and beyond.
- The customer cost of downtime for larger configurations tend to be higher than smaller configurations
- Average Legacy PA dump write speed is about 4 Megabytes/second

Factors in Dump Speed



Control Point #1
Amount Selected to Dump



- Select pages to be dumped based a static selection of page classifications

Page Class	Definition
UNUSED	Pages that have not been allocated yet
USERPG	Non-kernel pages
BCACHE	Buffer Cache Data
KCODE	Kernel Image (code)
USTACK ☆	User Stacks
FDDATA ☆	File System Metadata
KSDATA ☆	Kernel Static Data (Globals, etc)
FSDATA ☆	Kernel Dynamic Data

☆ Default Configuration

Selective Dump Configuration

- With 11.0, the system admin can specify the page classifications that are to be dumped via the crashconf command or the rc script

Page Class	Definition
UNUSED	Pages that have not been allocated yet
USERPG	Non-kernel pages
BCACHE	Buffer Cache Data
KCODE	Kernel Image (code)
USTACK ☆	User Stacks (Kernel related stacks)
FDDATA ☆	File System Metadata
KSDATA ☆	Kernel Static Data (Globals, etc)
FSDATA ☆	Kernel Dynamic Data

☆ Default Configuration

Displaying Page Classifications to be dumped

- Crashconf –v will display the current configuration

```
CLASS          PAGES  INCLUDED IN DUMP  DESCRIPTION
--------  ----------  ----------------  -----------------------------------
UNUSED           2036  no,  by default   unused pages
USERPG           6984  no,  by default   user process pages
BCACHE          15884  no,  by default   buffer cache pages
KCODE            1656  no,  by default   kernel code pages
USTACK            153  yes, by default   user process stacks
FSDATA            133  yes, by default   file system metadata
KDDATA           2860  yes, by default   kernel dynamic data
KSDATA           3062  yes, by default   kernel static data

Total pages on system:              32768
Total pages included in dump:        6208

DEVICE         OFFSET(kB)    SIZE (kB)  LOGICAL VOL.  NAME
------------  ----------  ----------  ------------  -------------------------
 31:0x00d000       52064      262144   64:0x000002  /dev/vg00/lvol2
                  ----------
```

Changing Page Classifications to be dumped Interactively

- Interactive changes are only good until next reboot or until crashconf is run again to alter it

  Crashconf –i <class> : This includes the classification

  Crashconf –e <class> : This excludes the classification

- Examples
  - Include buffer cache pages to the pages that are to be dumped:

  $ Crashconf –i bcache

  The crashconf –v will show the BCACHE entry as 'Yes, forced' in the INCLUDED IN DUMP column.

  - Exclude buffer cache pages from the dump

  $ Crashconf –e bcache

  The crashconf –v will show the BCACHE entry as 'No, forced' in the INCLUDED IN DUMP column

Changing Page Classifications to be dumped persistently

- The /etc/rc.config.d/crashconf file is where the page classifications can be forced. The crashconf command is run during system boot and uses this file to override the default.

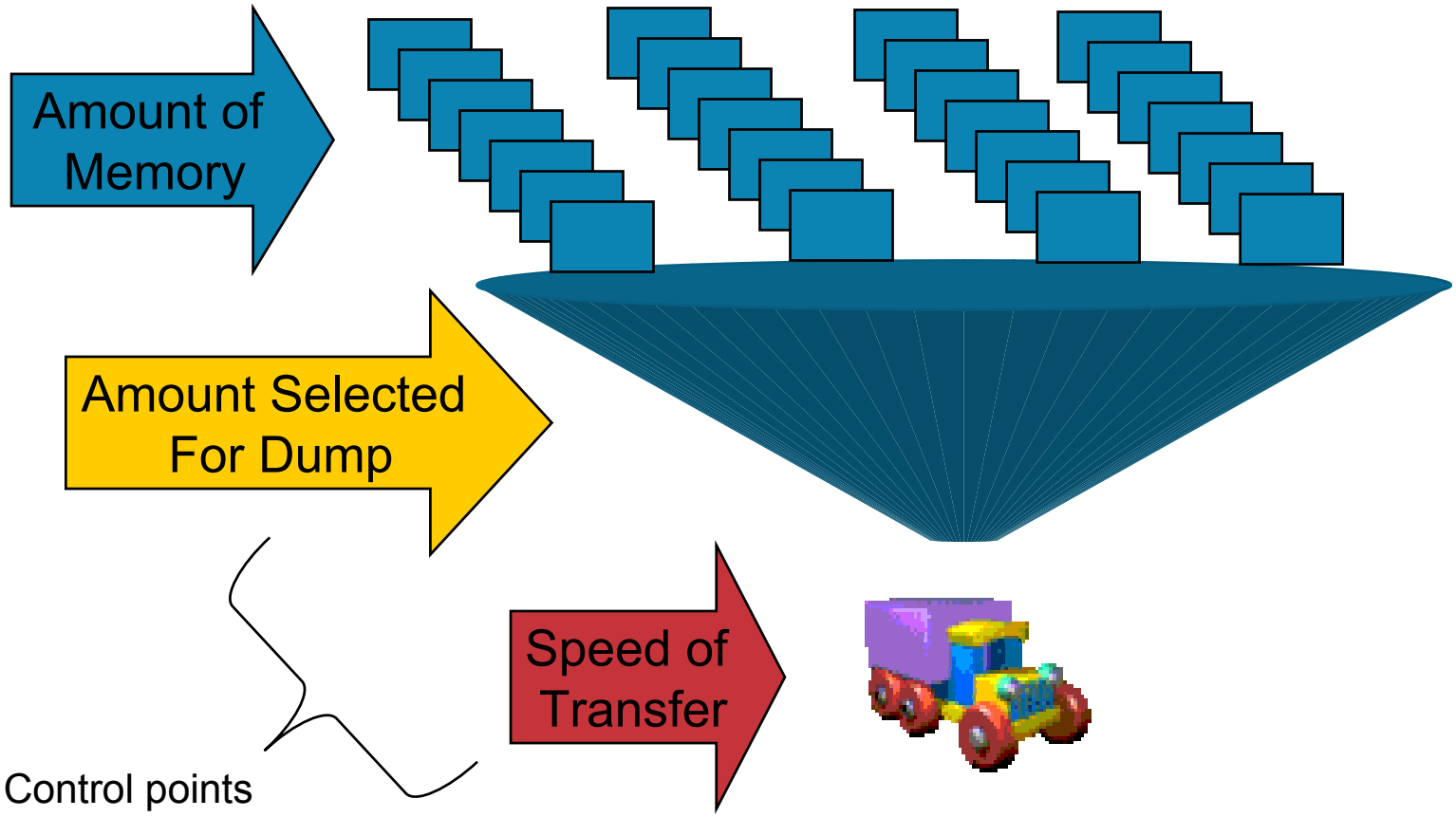
```
# CRASH_INCLUDED_PAGES: A blank-separated list of page classes that must be
#           included in any dump.  crashconf -v will give you a list of
#           valid page classes.  Specify "all" to force full dumps.
CRASH_INCLUDED_PAGES=""


# CRASH_EXCLUDED_PAGES: A blank-separated list of page classes to omit from
#           selective dumps.  crashconf -v will give you a list of valid
#           page classes.  Specify "all" to disable crash dumps.
CRASH_EXCLUDED_PAGES=""
```

Selective Dump Results

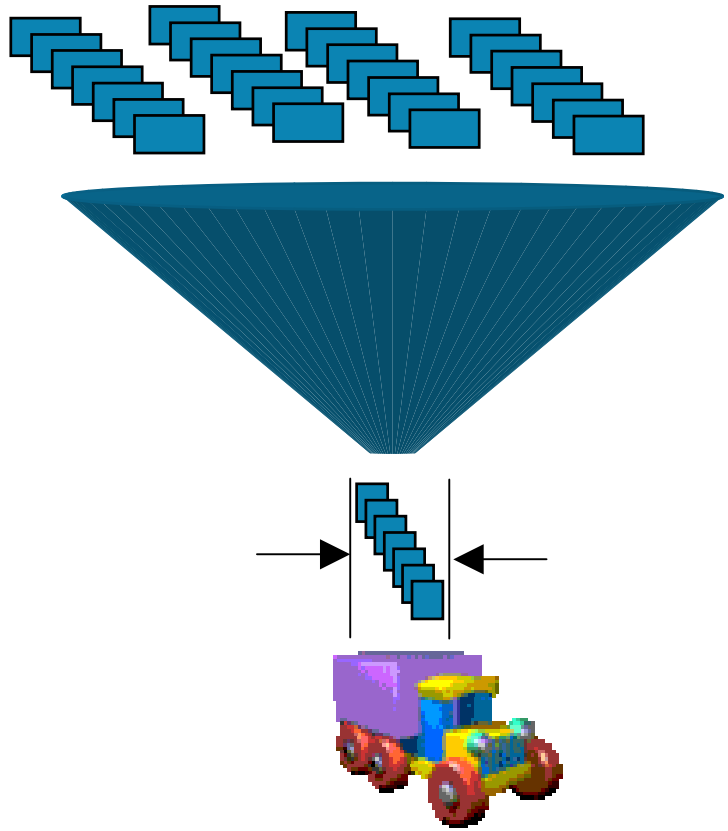
- Introduced in 11.0
- A study of customer dumps (over 100) has shown that this approach has resulted In:
  - dump sizes are now only 15% (+/- 5% confidence) of physical memory.
  - Except in extreme corner cases, the dump contains as much data required to resolve the defect as a FULL dump.

Factors in Dump Speed



Control point #2
Speed of the transfer

- New for 11.23 and 11.11
  - 11.11 requires June ’03 Software Pack
- Dump Compression
  - Compressing the data before it is written to disk increases the effective transfer rate
  - Primary target is large memory systems with multiple processors



Compressed Dump Goals

- Significantly increase the effective transfer rate of dump pages to disk, reducing the downtime associated with a system crash.
  - Minimum goal of increasing effective transfer rate by a factor of 3.
- Compression is done in parallel with write to the disk.
- Primary Target is Large Memory systems with Multiple processors
  - Compression can be used when
    - Physical memory is Two gigabytes or larger
    - There are more than Four Processors

Compressed Dump Configuration Interactively

- To enable the compressed dump ‘hint’, crashconf has added a new –c switch

```
$ crashconf –c on
$ crashconf –v
Crash dump configuration has been changed since boot.

CLASS          PAGES  INCLUDED IN DUMP  DESCRIPTION
--------  ----------  ----------------  -----------------------------------
UNUSED          46397  no,  by default   unused pages
USERPG            690  no,  by default   user process pages
BCACHE          80794  no,  by default   buffer cache pages
KCODE            5595  no,  by default   kernel code pages
USTACK              1  yes, by default   user process stacks
FSDATA              0  yes, by default   file system metadata
KDDATA         101186  yes, by default   kernel dynamic data
KSDATA           2311  yes, by default   kernel static data
SUPERPG         23476  no,  by default   unused kernel super pages

Total pages on system:            260450
Total pages included in dump:     103498

Dump compressed:    ON

DEVICE        OFFSET(kB)   SIZE (kB)  LOGICAL VOL.  NAME
------------  ----------  ----------  ------------  -------------------------
 31:0x001000      826228     2097148   64:0x000002  /dev/vg00/lvol2
                           ----------
                              2097148
```

Note: The compression switch is a hint to the OS. Compression will occur when physical memory is at least 2 gigabytes and the number of processors is greater than 4

In this case, the hint is on, but compression would not happen (1 Gig system)

Compressed Dump Configuration Persistently

- In 11.11, The /etc/rc.config.d/crashconf file is where the compressed dump hint can be set persistently. The crashconf command is run during system boot and uses this file to override the default.

```
# CRASHCONF_COMPRESS: CRASHCONF_COMPRESS set to 1 enables the hint
#                     CRASHCONF_COMPRESS set to 0 will disable the compression hint.
#                     leaving this blank will result in the default, which is enabled.
CRASHCONF_COMPRESS=1
```

Compressed Dump Configuration Persistently

- In 11.23 and forward, the new dynamic kernel tunables interface is used to create persistence across boots.
- The dump subsystem parameters are now kernel tunables and can be set via crashconf with the –t switch.
- The kernel tunable is called ‘dump_compress_on’ and can also be managed via the kctune interface.

```
$ crashconf –t –c on
$ kctune -v dump_compress_on
Tunable             dump_compress_on
Description         Enable compressed dump functionality
Module              dump
Current Value       1
Value at Next Boot  1
Value at Last Boot  1
Default Value       1
Constraints         dump_compress_on >= 0
                    dump_compress_on <= 1
Can Change          Immediately or at Next Boot
```

Compressed Dump Results

- Lab testing has shown
  - Average effective dump write speed improvement met or exceeded the 3X goal, except when UNUSED Pages were dumped.
  - dump speed write speed improvements of more than 4.5 times have been observed.
    - Write times will vary on the nature of the data being dumped
    - UNUSED pages do not compress well, so dumping these pages reduces the efficiency of the compression algorithm.

Compressed Dump Considerations for Analysis

- The compressed dump results in an image file that is in a different format than legacy dump images. Crashutil can now convert formats

```
$ crashutil -v COREDIR /var/adm/crash/crash.2 /tmp/oldcoredir

This translates a new coredump to an older version The versions include:
```

COREFILE – Version 0	This format, used up through HP-UX 10.01, consists of a single file containing the physical memory image, with a 1-to-1 correspondence between file offset and memory address. Normally there is an associated file containing the kernel image. sources or destinations of this type must be specified as two pathnames to plain files, separated by whitespace; the first is the core image file and the second is the kernel image file.
COREDIR – Version 1	This format, used in HP-UX 10.10, 10.20, and 10.30, consists of a core.n directory containing an INDEX file, the kernel (vmunix) file, and numerous core.n.m files, which contain portions of the physical memory image. sources or destinations of this type should be specified as the pathname to a core directory.
CRASHDIR – Version 2	(Version 2) This format, used in HP-UX 11.00 and later, consists of a crash.n directory containing an INDEX file, the kernel and all dynamically loaded kernel module files, and numerous image.m.p files, each of which contain portions of the physical memory image and metadata describing which memory pages were dumped and which were not. sources or destinations of this typeshould be specified as the pathname to a crash directory.
PARDIR – Version 3	This format is used in HP-UX Release 11i Standard input to the CRASHDIR format in that it consists of a crash.n directory containing an INDEX file, the kernel and all dynamically loaded kernel module files, and numerous

How do I get Compressed Dump?

- Included in 11.23
- For 11.11
  - June 2003 Software Pack CD
    - Part #5012-7841
    - SPK0306-11.11
  - Web
    - http://www.software.hp.com/
    - Search for CDUMP11i

Interex, Encompass and HP bring you a powerful new HP World.