# Kdump

## A Kexec Based Kernel Crash Dumping Mechanism

Vivek Goyal (vgoyal@in.ibm.com)
Eric W. Biederman (ebiederman@lnxi.com)
Hariprasad Nellitheertha (nharipra@gmail.com)

# Background

- Dump capture from crashing kernel's context
  - ✓ Resource lockup
  - ✓ Corrupt data structures
- Dedicated dump drivers
  - ✓ Limited number of target devices
  - ✓ Maintenance was a big issue
  - ✓ Dependency on crashing kernel reduced and not eliminated completely

# Background contd...

- Stand alone dumpers
  - ✓ Need to maintain low level hardware specific code
  - ✓ Filtering is not possible
- Kernel reboot based dumper
  - ✓ Memory constraint might prevent capturing full dump
  - ✓ Significant amount of code being run in crashing kernel context
  - ✓ Core kernel invasive code

# Design Goals of The New Solution

- Simple and minimally invasive into the kernel code
- Highly Reliable
- Available on most architectures
- Easy to Maintain
- Flexibility in terms of dump contents and targets
  - ✓ Full dump or kernel-pages only dump
  - ✓ Dump to disk or across the network
- Ease of Use

# Kdump – Overview

- A new kernel, often called capture kernel, is booted after the crash
- Previous kernel's memory is preserved
- Dump is captured from the context of capture kernel
- Kernel-to-kernel boot loader enables booting a new kernel after a crash
- Kexec is underlying kernel to kernel boot-loader

# Kernel-to-Kernel Boot Loader

- Running kernel acts as a loader for the new kernel
- System directly jumps from one kernel to another
  - ✓ Skips BIOS or Firmware stage
- Reboots are extremely fast (33% reduction in time)
- Memory can be preserved across reboots
  - ✓ Since BIOS is skipped, it is left to the OS to retain or erase memory
- Finds application in crash dumping tools

# Kexec

- Allows a Linux kernel to boot another kernel
- Currently available on i386, x86_64 and ppc64 platforms
- Two components
  - ✓ User space tool – kexec-tools
  - ✓ Kernel System Call (sys_kexec_load)
- Load a new kernel

  *kexec -l <kernel-image> --append=<options>*
- Exec new kernel

  *kexec -e*

# Kexec On Panic

- An extension of Kexec functionality
- Enables booting a new kernel after system crash
- Devices are not shutdown
- New kernel runs from a reserved memory location
    - ✓ Protection against on-going DMA at the time of crash

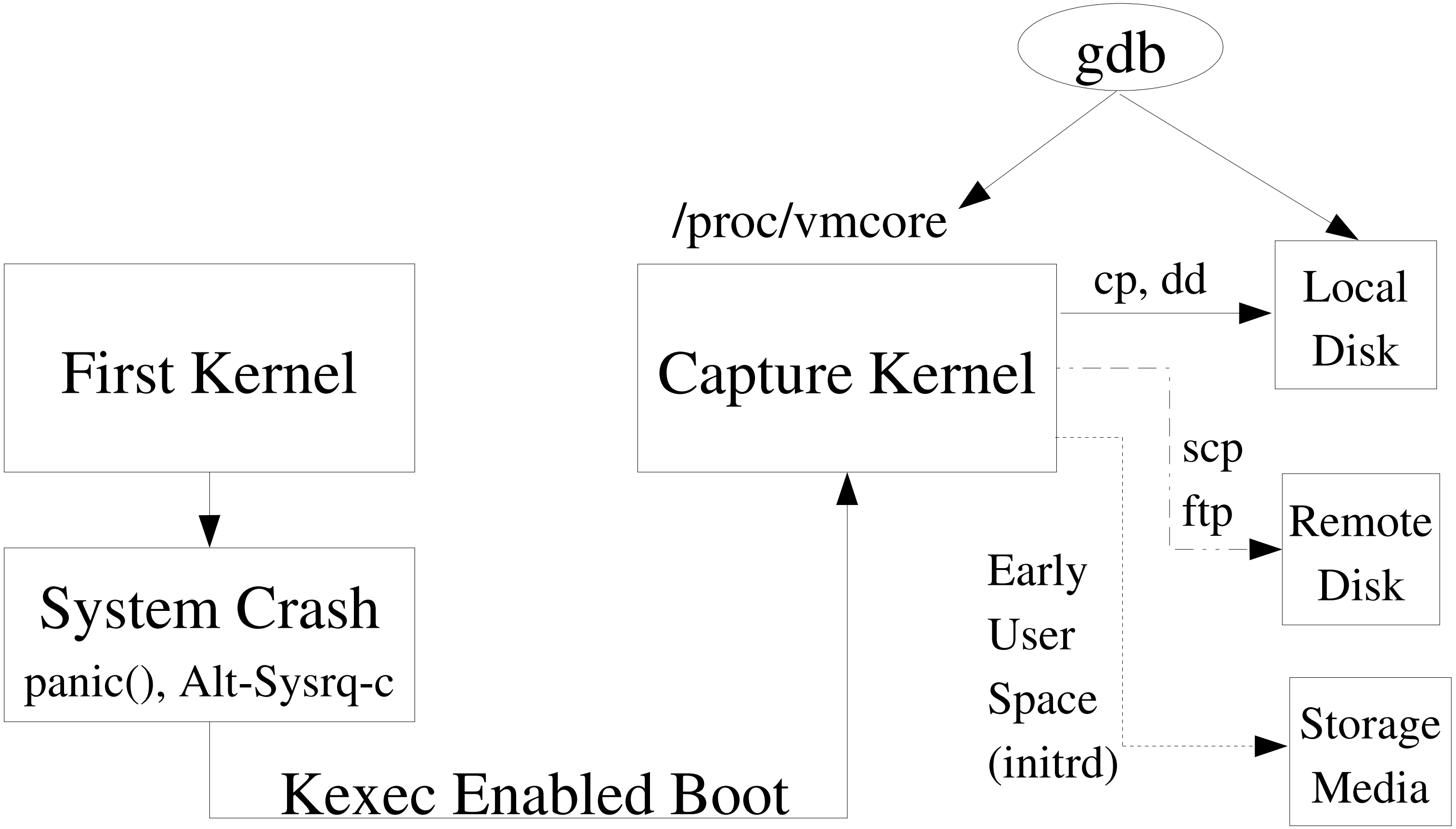
# Kexec On Panic Contd..

- Loading capture kernel

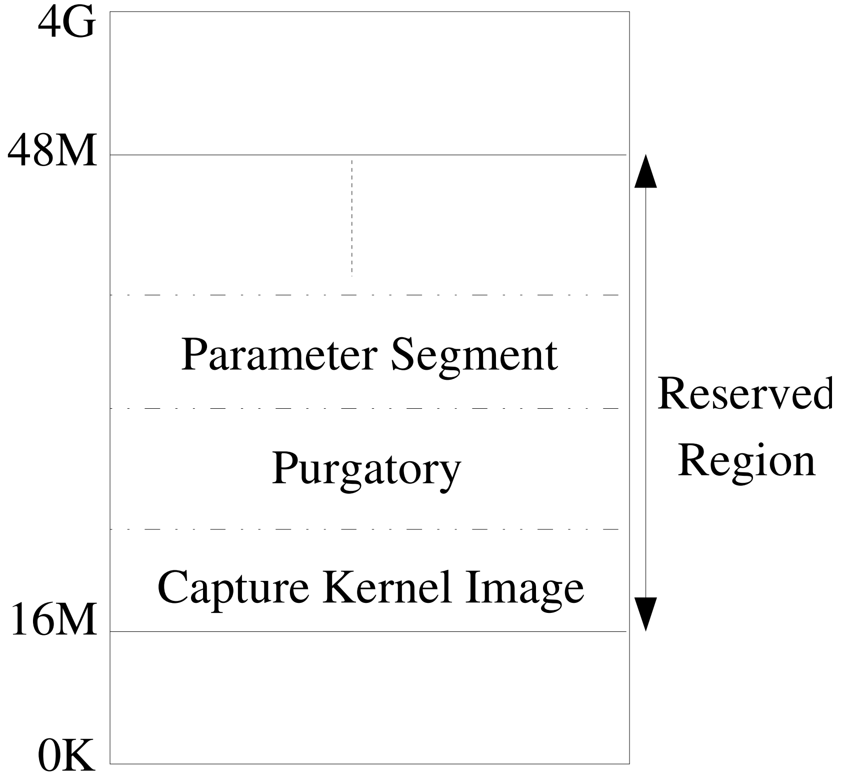
```
kexec -p <kernel-image> --append=<options>
```

- Execution of capture kernel
  - ✓ *panic()*
  - ✓ *Alt-Sysrq-c*

# Kdump Overview

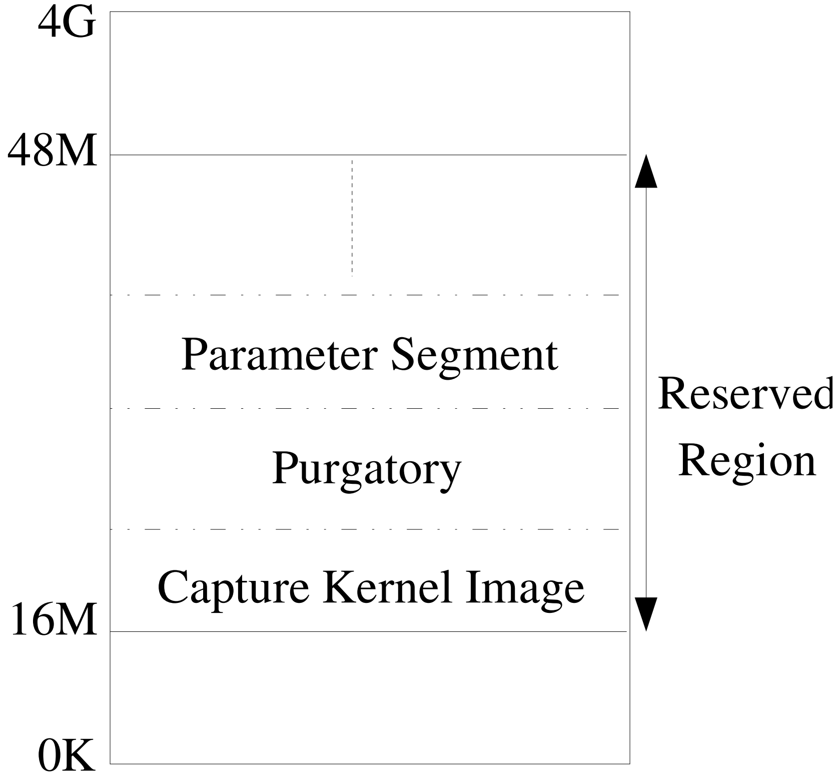


# Kexec on Panic - Pre-loading



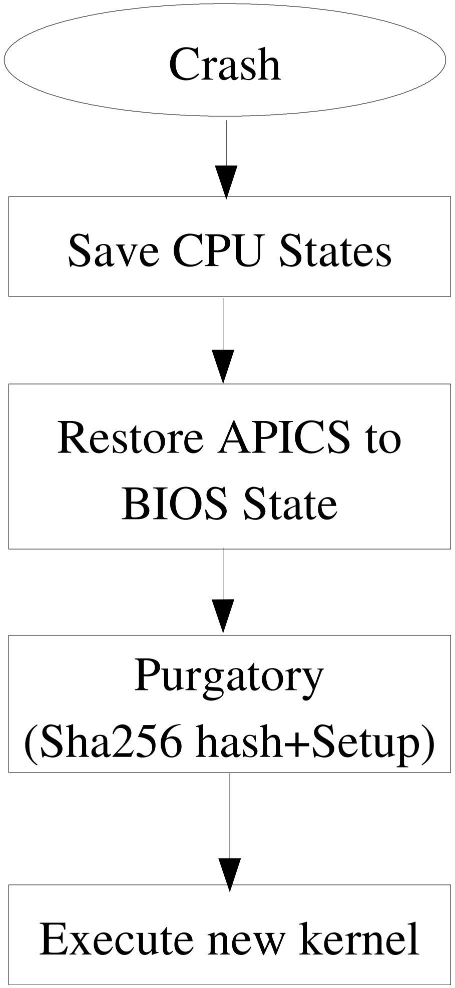
- Reserve memory for capture kernel (crashkernel=X@Y)
- Pre-load the capture kernel
- Capture kernel runs from reserved memory location

# Kexec on Panic - Purgatory



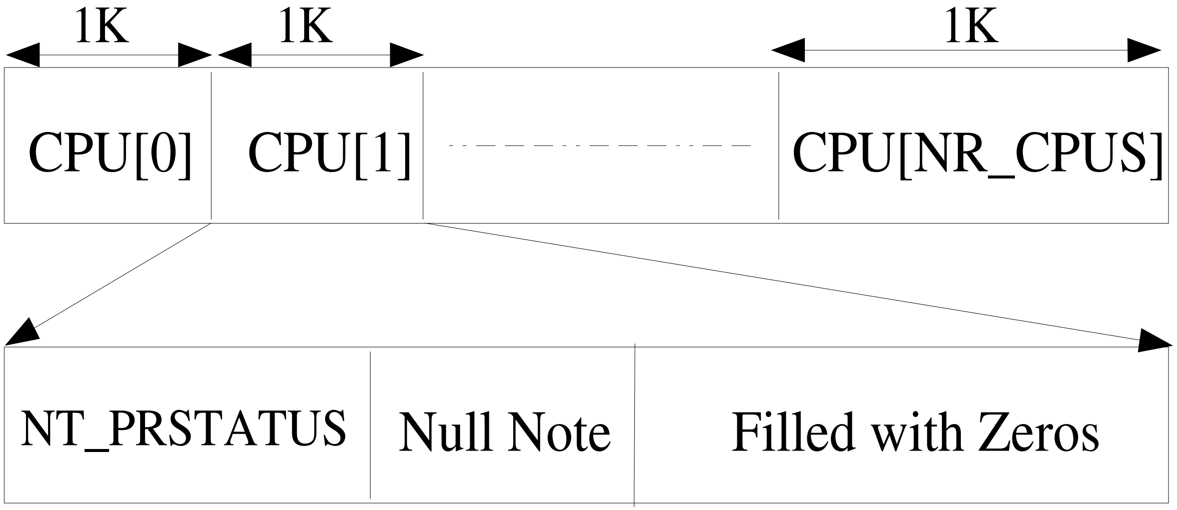
- Purgatory is an ELF relocatable object and it contains setup code and sha256 hash
- Sha256 hash ensures integrity of the new kernel's pre-loaded data

# Kexec on Panic – Post Crash (x86)



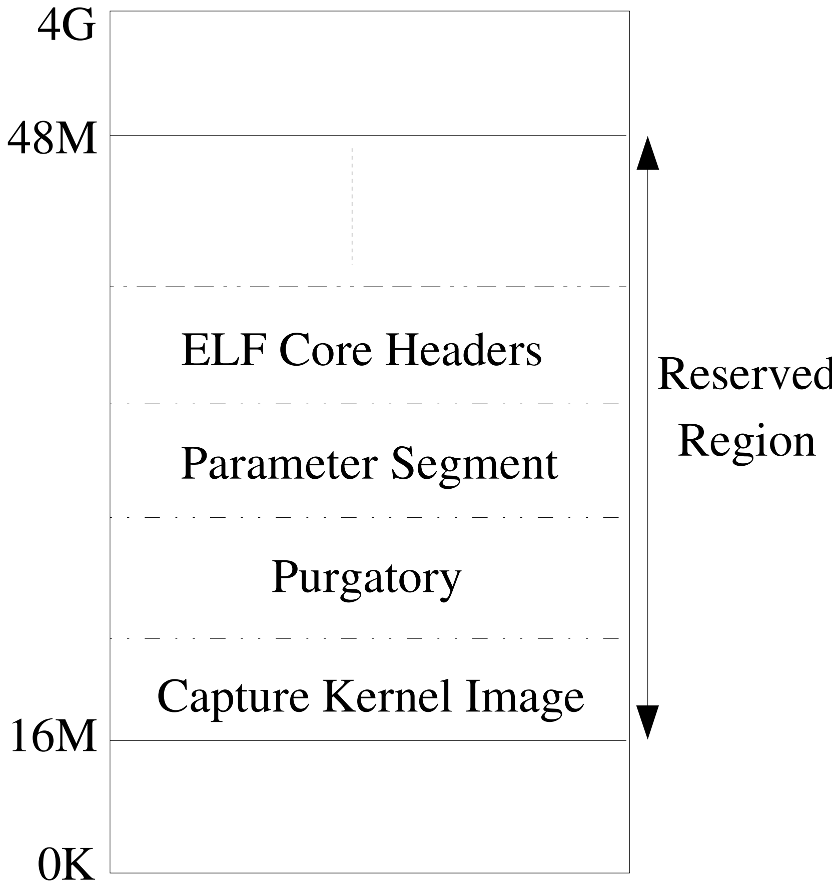
- CPU states are saved and other CPUs are halted using NMI
- LAPIC/IOAPIC are disabled and put back into PIC or virtual wire mode
- Purgatory is run and control is transferred to new kernel

# Kexec on Panic – Saving CPU States



- CPU register states are saved in ELF note format
- 1K of memory is reserved statically per CPU

# Kdump – ELF Header Generation

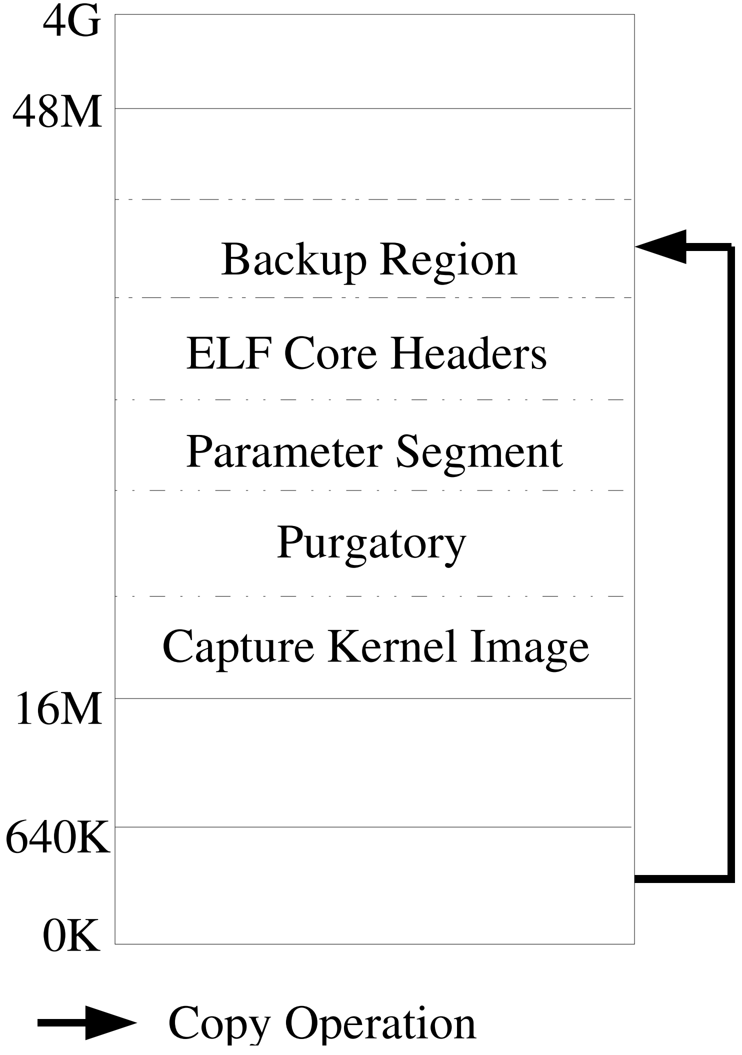


- Dump information across the kernel is exchanged in an ELF format Core file
- Kexec-tools prepare ELF headers and pre-load  them in the reserved region

# Kdump – ELF Header Generation

- One PT_LOAD type ELF program header is created for every contiguous memory chunk
- Kexec-tools use /proc/iomem to retrieve System RAM information on i386 platform
- Address of the start of ELF header is passed to the capture kernel using command line option “elfcorehdr=”

# Kdump – Backup Region



- Kernel uses some fixed memory locations to boot
- First 640K of memory is required for booting SMP capture kernel on i386
- Contents of first 640K of memory are backed up in Backup Region

# Kdump – Backup Region

- Kexec-tools reserve the memory for backup region while loading capture kernel.
- Purgatory contains the code for backing up first 640K of memory after crash.
- Other architectures can define their own backup region (If need be).

# Kdump – Booting into Capture Kernel

- Capture kernel uses limited amount of memory to boot
- Command line option “memmap=exactmap” is used to limit the memory regions capture kernel uses
- Kexec tools append memmap= command line options automatically

# Kdump – Capturing the Dump

- Accessing dump image in ELF Core format
  - ✓ /proc/vmcore
- Accessing dump image in linear raw format
  - ✓ /dev/oldmem

# Kdump – ELF Format Core File (/proc/vmcore)

| | | | | | |
|---|---|---|---|---|---|
| ELF Header | Program Header PT_NOTE | Program Header PT_LOAD | - - - - - - - - - | Per CPU Register States | Dump Image |

- ELF32/ELF64 format headers
- Physical addresses are filled for all the regions
- Virtual addresses are filled only for linearly mapped memory region

# Kdump – Analysis Tools

- gdb
  - ✓ Virtual view of memory
  - ✓ Can debug linearly mapped region of memory
  - ✓ User space utility to regenerate ELF headers to create the ELF headers for vmalloc regions
- crash
  - ✓ Physical view of memory

# Advantages

- Increased reliability
  - ✓ Dump is captured from a newly booted kernel
- Enhanced flexibility
  - ✓ Dump image can be saved to virtually any storage media supported by kernel
  - ✓ Filtering mechanism can be plugged in

# Advantages Contd..

- Ease of use
  - ✓ Standard utilities can be used to save the dump image either locally or remotely
  - ✓ Standard analysis tools like gdb can be directly used for limited debugging

# Limitations

- Devices are not shutdown/reset after a crash which might result in a driver initialization failure in capture kernel
- Non-disruptive dumping is not possible

# Current Status

- Initial i386 implementation is mainline now (2.6.13-rc1)
- Driver initialization issues are being addressed
  - ✓ Shared Interrupts
    - ➢ irqpoll commandline option, Disabling PCI interrupts etc.
  - ✓ Driver hardening
    - ➢ Reset the device if it is not reset already.

# ToDos

- Port kdump to other platforms like x86_64 and ppc64
- Modify “crash” tool to be able to analyze kdump generated dump images
- Implement kernel pages only filtering mechanism
- Relocatable Kernel for binary image unification
- Initialize APICs before timer initialization

# Downloads

- Kdump patches for kexec-tools and test reports are available at:

  http://lse.sourceforge.net/kdump/

Questions?

# Legal Statement

- Copyright ©2005 IBM.
- This work represents the view of the author and does not necessarily represent the view of IBM.
- IBM, and the IBM logo, are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Linux is a registered trademark of Linux Torvalds in the United States, other countries, or both.

# Legal Statement Contd...

- Other company, product, and service names may be trademarks or service marks of others.
- References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.
- This document is provided "AS IS" with no express or implied warranties. Use the information in this document at your own risk.

# Thank You