Kexec Issues & Ideas

Contacts: xuehaohu@google.com, rminnich@google.com
Last Updated: Sept 1, 2021
Status: Curating
Update: Weekly update Updates / Newsletter.

The goal of this doc is to collect all pains / issues / problems / ideas people have with kexec nowadays, and then come up with an enhancement plan to address them. Feel free to add yours following this template!

Problems / Ideas

Problems that kexec did not work for you, or improvement ideas.

- Kexec transitioning into crash kernel can be unsafe in case of a hailstorm of NMIs
    - More Details: This is a common problem with GPUs in servers. Under high load, they can fail off the PCI bus. In case of a subsequent kdump, the transition into the crash kernel tends to happen under a hailstorm of NMIs. That transition today involves creating a binary object (called purgatory, https://lwn.net/Articles/582711/), which runs between two kernels. Purgatory is not NMI safe.
    - Repro: <TODO>
    - How Should it Behave:<TODO>
    - Benefits / Impact: <TODO>
    - Your Contact Info: Andrew Cooper, <TODO: email>
- There is not a way to hand off the kernel keyring to a new kernel.
- More details: Current hack is to align the existing initrd to 512 bytes and append a new cpio with the secrets, although this leaves them in user space for a little while
        - Kexec does not support EFI target
        - Speculative case study:
            - Run the bootloader in a simple VM that implements UEFI boot stubs as services, see https://github.com/linuxboot/voodoo, which can run GNUEFI “hello” and partly run grub; all in Go including the VMM.
        - Contacts: rminnich@google.com, hudson@lowerlayer.nl
- Kexec can leave devices (semi-)initialized during transition
    - More Details: This can cause issues when the following (Linux, Windows, or other) kernel tries to initialize that device again.
      The issue isn’t actually in kexec code. It’s the device drivers that need to be fixed.
    - How it should behave: All drivers should fully deinitialize the devices they manage.
    - Drivers should be tagged as KEXEC_SAFE, similar to the way they are today as GPL compliant.
    - References: Ron Minnich, kexec based bootloaders on RISC V Use cases and Advantages
- Kexec from one kernel version to another (even if LTS) does not always work
    - We want any LTS kernel, starting with 5.4, to be able to boot any LTS kernel.
    - Kexec hangs sometimes (Linuxboot kernel: 5.10.x; target kernel: 5.14-rc1)
        - As another data point, our experience shows that when Linuxboot kernel is v5.10.50, the target Linux kernel is v5.2.9, there are issues on EagleSteram servers, but not on CooperLake servers. I wonder if you could replicate such issue on QEMU, it might have something to do with the server (processor). If you are not able to replicate on QEMU, you should try to replicate on Intel Archercity CRB, for which there is fsp/coreboot/Linuxboot based firmware available for Google, as per the MP-NDA.
        - https://osfw.slack.com/files/UEZE0ANSC/F02CTJ484M8/fbnetboot_kexec_stuck.zip
        - I have not tried this on other platforms such as QEMU. It happens randomly on ArcherCity CRB (you have everything needed to repro it), and we do not experience a similar kexec problem on DeltaLake which is based on CooperLake processor.
        - Hang issue screenshot: https://osfw.slack.com/files/UEZE0ANSC/F02CWQHLYAJ/screen_shot_2021-09-01_at_10.09.49_am.png
        - contacts: jonzhang@fb.com
    -
- Raw kexec_load
    - More details: A command line option that allows arbitrary binary blobs and address ranges to be passed to kexec_load segments without any parsing.
    - Benefits/Impact
        - It would make it easier to prototype new file formats
        - Can move file format parsing out of kexec binary
        - TODO: how does this interact with signature checks?
    - Contact: Trammell Hudson hudson@lowerlayer.nl
- Multiple initrd options

    - More details: the kernel can handle multiple initrd's aligned to 512 byte sectors and concatenated, but kexec only allows a single --initrd argument.
    - Benefits/Impact
        - Tools that generate multiple initrd have to make an in-memory copy of all of the initrd in /tmp, which is then copied into memory for the kexec. This would avoid the extra copy.
        - Tools have to be aware of the alignment issue and use tools like dd to ensure 512-byte sizes, which requires an extra tool in the runtime.
    - Contact: Trammell Hudson hudson@lowerlayer.nl
- Inconsistent configuration between source and target kernel lead to unexpected driver states
    - One thing, although not directly related to kexec, is that the boot Linux and the runtime (target) Linux may have inconsistent configuration of certain drivers. For example, if Intel Pstate driver in the boot Linux is enabled but is disabled in runtime Linux, the Intel_Pstate driver is still somewhat enabled in the target Linux, it won't get disabled/re-init after kexec because there are certain MSR bits that cannot be unset. Just a little caveat.
        - **Johnny Lin (Wiwynn)** - OSF - osfw.slack.com
- <Add a new problem/idea>
    - [Briefly Describe the Problem]
        - **More Details**: <more details about the problem>
            - Link as much context / links / pointers as possible.
        - **Repro:** < Is there a way / setup you can share for others to repro your problem? >
            - Pls provide as many details as possible.
        - **How Should it Behave**: < How should it behave ? describe the correct behavior >
            - Alternative best? - In case an ideal behavior is exponentially hard to tackle, what should the best alternative behavior look like ?
        - **Benefits / Impact:** < Benefits / Impact: What benefits / impact would it have if respective improvements are made? >
        - **Contact:** <Name and email. It is likely people will need to reach out to ask questions to understand more about your problems >

Open Questions

- Andrew Cooper: Is this something which the Linux and kexec-tools folk are on board with? Getting them on board is surely a prerequisite
    - https://github.com/horms ?
- Collaborate with petitboot? Is a kexec bootloader and seems to be the main method of booting (Open)POWER machines.

Updates / Newsletter

Sept 1, 2021

- Added issues reported by Jonathan @fb

July 26, 2021

- xuehaohu@google: bringing up internal kexec integration tests to run continuously.
    - Continuous kexec integration tests in prod machines across archs, and LTS kernels.
        - LTS1 -> LTS2 across different arches.

Aug 10, 2021

- xuehaohu@google:
    - Continuous kexec tests up and running on 2 google prod platforms
        - 2500+ iterations
    - To add in comprehensive checks to validate the system are in the same state after each kexecs. E.g.
        - Does memory available decrease over time ?
        - Does the number of cores not change over reboots ?