Conversions content policy

For 2026-03-05 guest_memfd bi-weekly upstream call

Contact ackerleytng@google.com if you have questions/suggestions!

Background

- Discussion on guest_memfd conversions series RFC v2 [1]
- How will userspace tell KVM what to do with the content of the memory during conversions?

[1] https://lore.kernel.org/all/CAEvNRgFMNywpDRr+WeNsVj=MnsbhZp9H3j0QRDo_eOP+kGCNJw@mail.gmail.com/

Proposal

- For every conversion request aka SET_MEMORY_ATTRIBUTES guest_memfd IOCTL, provide a content policy enum
- “How memory will look for the next reader”
  - ZERO
    - After conversion, the memory contents will be zeroed for the next reader
  - READABLE
  - ENCRYPTED

ZERO

- TDX and SNP do nothing since firmware will zero on conversion
  - Sean: SNP doesn’t zero the contents
- pKVM and SW_PROTECTED_VM: software zeroing
  - Suzuki K.P.: CCA “scrubs” content (not necessarily zeros) because the encryption key changes, hence don’t support `ZERO`
    - Steven Price: spec says “wiped” doesn’t specify what wipe means
  - Documentation: ZEROing on shared to private is not supported

- This policy allows userspace to be sure that memory is zeroed after the conversion - no need to re-zero after conversion

~~READABLE~~ PRESERVED

- TDX and SNP return `EOPNOTSUPP`
- SW_PROTECTED_VM (our testing vehicle) will allow this, do nothing on conversion - memory remains as it is
- Sean: Shared to private, private to shared => host sees the same value it wrote
- Suzuki: populate at beginning: should we use PRESERVED?
  - Sean: no because the ordering among other ioctls is so vendor-specific, so it's better to keep it fully vendor specific
- pKVM will do the same as SW_PROTECTED_VM
  - No loss of confidentiality, since the conversion ioctl doesn't actually tell EL2 to make the memory accessible to the host
  - A host that makes the memory mappable without guest request
    - KVM_EXIT_MEMORY_FAULT on the next guest access
    - SIGBUS (from pKVM/EL2, not from guest_memfd) when accessing the memory?

ENCRYPTED (ARM - DONT_CARE (DEFAULT))

- “Memory contents will appear encrypted to the next reader”
- Does this make sense for SNP to use conversion for a special GHCB call
  - Goal: Guest wants to force TLB flush on the host in a verifiable way => guest must confirm that RMPUPDATE happened
  - Sean: Handle this completely within KVM, so no need for conversion ioctl call, don’t allow conversions, don’t update attributes in guest_memfd. guest_memfd thinks of it as private all the time
- TDX and SNP (other than the GHCB call) return `EOPNOTSUPP`
- SW_PROTECTED_VM (our testing vehicle) will write `0xff` for testing? `/dev/random`?
- pKVM will not support this, since there is no encryption