Agenda and Minutes: x86 Community Call August 2018

No new items were added to the agenda. Minutes are added in blue. Closed ACTIONS in green.

Attendees

Artem Mygaiev
Juergen Gross
Brian Woods
Christopher Clark
Jan Beulich
Wei, Sergey, Andrew
Julien Grall
Rich Persaud
Zhang, Yu C

4.10+ changes to Xen's memory scrubbing: Christopher

Discussion of the changes that made to it in recent versions of Xen (4.10+) - Christopher

Christopher: Asking on behalf of Star Lab who has a use case for ensuring scrubbing occurs on shutdown of VMs. Memory scrubbing is also of interest for OpenXT. Inspection of code shows it has changes done during 4.10: to defer scrubbing of pages to idle loop rather than synchronously during VM shutdown. Idle scrubbing may take indefinite amount of time to complete on a busy machine.

Checking: Is scrub only done if VM is explicitly killed?
What are the right expectations for us re former-VM memory content?

Andrew: What is in tree is only the 1st half of what was proposed
Boris didn't have time to do the second half of the series => XenServer has similar issues

Should never have a case where we don't scrub incoming RAM (leaving side channels aside). Everything handed to a guest is set to 0

Christopher: with the new code, when you shut down from a VM you have no longer been evacuated from the host + speculative side-channel vulnerabilities are a concern

Andrew: we would fix this if we permanently removed directmap => fixes the speculative angle => better in the long run

For very large VMs it can take up to 15 minutes to shut down very large VMs

Only case when scrubbing does not occur is if a VM releases a page. Citrix has removed even that exception in XenServer. Should do that upstream. Performance impact will be negligible.

Christopher: would code to re-introduce optional synchronous scrubbing be accepted in upstream?

Andrew doesn't think so: but also doesn't think that eagerly scrubbing will actually help as memory can be accessed through side channel ways.

Jan + Andrew: [discussion of vCPU use]

Christopher: Important issue is that once the decision is taken to evacuate by shutting down the VM, need to have a bounded amount of time for the memory to be erased and a verifiable indicator that it is complete. Previous version was indicated by the domain no longer being present.

Andrew: needs some kind of auditable point when memory is not available (and some bounded and verifiable point in time). Could put some kind of sync flag in kill domain to reinstate previous behaviour in domain destroy or use another approach (Jan had some ideas).

Christopher: agrees that this could be what he is looking for

ACTION: Christopher will follow up on IRC/xen-devel@

Unblocking Series - Jan

I think we need to talk about how we mean to unblock large chunks of work

- *VMX MSRs policy for Nested Virt: part 1 (I've looked over this, and I think it's okay, but I also think that in particular nested stuff wants both maintainers and Andrew to look over)*
- *vpci: add support for SR-IOV capability*
- *paravirtual IOMMU interface*
- *x86/domctl: Save info for one vcpu instance*
- *SSBD AMD via LS CFG Enablement and not to speak of "add vIOMMU support with irq remapping function of virtual VT-d". I'm however myself as well in an increasingly awkward position to do / post further work, due to there being patch series stalled in part from long before the 4.11 freeze (listing only series here, there are also individual stalled patches):*
- *x86: improve PDX <-> PFN and alike translations*
- *x86: assorted assembly related cleanup*
- *x86: indirect call overhead reduction*
- *x86/HVM: implement memory read caching*
- *x86: more power-efficient CPU parking*

Series is just a set of examples. Jan feels increasingly blocked by the number of series and volume of series. How do we get out of that huge backlog of patches that various people have to look at. Jan has not been able to reduce the backlog …

Andrew: This has been made worse by security work: Andrew is trying to pick up rather more review. It would be a useful thing to track the outstanding series - some series dont have outstanding actions (or are blocked on others).

We had issues where: patch is fine, but then has been deferred because XYZ has not yet done => but then XYZ had not materialized. Jan has some examples: 1st - where Jan fixed some boundary cases of error handling.

Another general problem: affects mostly individual patches when they have dependencies to do with rebase.

General workload: most of x86 series by Jan and Andrew need to be acked by each other. There has been improvements looking at x86 patches in some areas, but there are some areas such as AVX where where others can't help out.

Andrew also raises the issue of lack of comments from Intel, for serieus that cover vendor specific series.

ACTION: Lars to bring up at AB call

L1 Terminal Fault - Andrew

Sent out an update mail to xen-devel@: see
https://lists.xenproject.org/archives/html/xen-devel/2018-08/threads.html#01160

We still have a lot of work to do in this area: we could not do these during the embargo

Cannot remove all the cache load gadgets in the Hypervisor
Properly undertake work such as removing directmap, …

Need to start tracking issues related to this in JIRA: may need to sort this into different buckets

ACTION: Andrew will have a poke at these => classify correctly or close
Juergen is OK with using Jira

David Woodhouse from AWS has some proposal, which he will put out publicly

There same issues with page table writable in Linux: stop requesting to use it from Linux. It is nobbling one of the L feature flags

Andrew, Jan, Juergen: Question of 32 bit Xen support in the kernel: 32 bit PV guests will get increasingly crippled. If we drop 32 bit support from Linux, we should also do this for the hypervisor

ACTION: Juergen will set up an official email to see whether anyone is still using 32 bit PV guests. Need to check: netbsd - still have that linear page table issues (only reproducible with 32 bit guests) => Other options:

1) Declare it security unsupported
2) Limit to run inside a shim (need to keep it in the hypervisor), have a config pv 32, …
3) Deprecate before removal, then remove

Pose the question for Linux and others: important aspect is to reach out to users, developers are the wrong audience. **Lars to help out**

Project Management stuff to keep the Momentum going

Did not get to this section
Zhang: not sure about his colleagues - looking at some other issues and seek help.

This is something where Andrew and and Zhang should talk:

ACTION: Zhang to send a mail explaining the issue, possible set up

We have made significant progress on design related questions at the developer summit. Although not all the notes for these have been published (SGX and NVDIMM are missing, the former are on my plate). The series, which have been discussed at the summit and where I believe that good progress has been made were.

In other words, we should expect new versions of these series

Series	**Stakeholders**
Add vNVDIMM support to HVM domains *As far as I understand a simple and clean way to implement this has been found, but the design session notes are still missing* *We spent almost two days on NVDIMM related discussions: we have something that should be fairly simple and easy to implement. Dan Williams is happy to take changes into upstream as long as they are sensible.* *George: the key behind the discussion was to be able to deliver a functional solution soon. We can make it nicer incrementally.*	Zhang Yi, Intel Zhang Yu, Intel George Dunlap, Citrix

ACTION: George will update and re-submit the NVDIMM doc (he didn't take any notes during the discussion - we are going to have to reconstruct some of the discussion) *Andrew: Yi & Yu were taking notes in the meeting* *ACTION: Lars to reach out to Yi & Yu and see what they have* *See https://lists.xenproject.org/archives/html/xen-devel/2018-07/threads.html#01592*	
Intel Processor Trace virtualization enabling *See https://www.slideshare.net/xen_com_mgr/xpdds18-intel-processor-trace-for-xen-hypervisor-luwei-kang-intel* *Partly blocked on CPUID & MSR* *Discussed the corner cases - these are in a PPT from Intel which Lars is waiting for. There was an open question re nested virt and a recognition that both cannot co-exist.* *See https://lists.xenproject.org/archives/html/xen-devel/2018-07/threads.html#01592*	Luwei Kang, Intel
Extend resources to support more vcpus in single VM *Also depends on the topology work* *IOREQ work needs another iteration* *Virtual IOMMU needs to be done* *See https://lists.xenproject.org/archives/html/xen-devel/2018-07/threads.html#01592*	Chao Gao, Intel
EPT-Based Sub-page Write Protection Support See https://www.slideshare.net/xen_com_mgr/xpdds18-eptbased-subpage-write-protection-on-xenc-yi-zhang-intel *Intel posted series and doesn't know what to do next due to lack of feedback. We were also lacking a plausible use-case: Intel and BitDefender are talking together to clarify the*	Zhang Yi, Intel

use-case. Still largely blocked on reviews. *Also see https://lists.xenproject.org/archives/html/xen-devel/2018-07/threads.html#01592*	
SGX Virtualization design and draft patches *Kai sent Lars some notes, which are published here.* *Partly blocked on CPUID & MSR*	Kai Huang, Intel
5 Level Paging *XPTI would become very problematic with 5 level paging.* *Currently Intel's lowest priority.*	

Then there were series which were blocked on CPUID and related work

Series	**Stakeholders**
Add guest CPU topology support *[PATCH 00/13] x86: CPUID and MSR policy marshalling support* has been posted on which this series depends on, but it is only covering ⅓ of the needed patches and requires some fixes. Sergey is working on the libxc side and Andrew on the hypervisor auditing/checking. Roger is working on topology support, which depends on the other three pieces.	Zhang Yi, Intel Andrew Cooper, Citrix Sergey Dyasli, Citrix Roger Pau Monne, Citrix

And other series, which are moving forward

Series	**Stakeholders**
paravirtual IOMMU interface *v2 posted recently*	Paul Durrant, Citrix
x86/cpuid: enable new cpu features *Waiting for v5*	Yang Zhong, Intel

add vIOMMU support with irq remapping function of virtual VT-d *Waiting for v5*	Chao Gao, Intel
AMD Avic Series *Waiting for v3*	Janakarajan Natarajan, AMD
MSR Spec Support for AMD speculative store bypass mitigations *Work has just started*	Brian Woods, AMD
Dom B *Waiting for Christopher's reply*	Christopher Clark, OpenXT
XSM Daniel De Graf on sabbatical - not sure for how long ACTION: Rich to follow up with committers@xenproject.org	