L1 Terminal Fault (L1TF)

Spectre & Meltdown Reloaded

Stefan.Lüders@cern.ch

Thx to the Cloud Team, BATCH, and Linux support
Special thanks to Domenico Giordano for performance studies

WLCG MB 20180918

~~**TIOC 20180905**~~
~~**ASDF 20180823**~~

L1 Terminal Fault (L1TF)

The Problem:

- Affecting Intel x86 processors
- **Break down the boundary that protects memory between processes and between a Hypervisor (HV) and the virtual machines (VMs) running on it:**
  - Variant 1 – SGX aka "Foreshadow" (CVE-2018-3615) – may allow unauthorized disclosure of information residing in L1 data cache from Intel's Software Guard Extensions (SGX) enclave;
  - Variant 2 – OS/SMM (CVE-2018-3620) – may allow unauthorized disclosure of information residing in L1 data cache from Operating System (OS) or System Management Mode (SMM);
  - Variant 3 – VMM (CVE-2018-3646) – may allow unauthorized disclosure of information residing in L1 data cache from a virtualized guest in Virtual Machine Monitor (VMM)
- **Any malicious actor with local access (e.g. LXPLUS/LXBATCH or rogue VMs) can bypass memory access security controls ordinarily imposed and managed by O/S or HV in order to read any physical memory location that is cached in the L1 data cache of the processor.**
- https://cern.ch/security/advisories/l1tf/l1tf.shtml
  https://access.redhat.com/security/vulnerabilities/L1TF
  https://www.kernel.org/doc/html/latest/admin-guide/l1tf.html
- However, exploits in the wild have not yet made public

General O/S Mitigations

CERN CentOS 7 and SLC 6:

- Fixes have been made available for testing through the CERN Linux software “test”/QA repositories. These fixes have been **released to production on 2018/8/16:**
  - CC7 - x86_64 [2018-08-20]
  - SLC6 - i686 [2018-08-20]
  - SLC6 - x86_64 [2018-08-20]
- In order to be active, they **require a restart** of the operating system

CERN centrally managed Windows devices:

- Updated through the **usual patch cycle**

Apple computers and individual, non-CERN managed devices:

- Should be updated using the **standard updating mechanisms** (Windows upgrade, yum update, ...) and subsequently **rebooted**

The three variants of "L1 Terminal Fault" are the 3rd iteration of CPU related hardware vulnerabilities. **Exploits or similar vulnerabilities may be published in the future and would require another iteration of applying patches and / or mitigation options.**

Mitigation for Hypervisors in CERN DC

Compute i.e. CERN LXBATCH (where we trust individual VMs):

- LXBATCH admins **fixed and rebooted of Batch VM** nodes with the latest kernel;
- HVs and SMT ("symmetric multi threading" or "HyperThreading") settings untouched;
- Scheduled to be **finished by end-of-August** with MD3/TS2 (2018/9/12) as fall-back date;
- Done: https://cern.service-now.com/service-portal/view-outage.do?n=OTG0045525

CERN IT Services (where we do not necessarily trust all VMs):

- **VM owners requested to patch their VMs** with latest kernel. Application not guaranteed;
- Cloud Team **disables** & monitors for "**SMT OFF**". Performance penalty deemed unlikely;
- Cloud Team **reboots HVs** after microcode update and deployment of new kernel;
- Scheduled to **start with MD3/TS2 (2018/9/12)** and taking about 8 working days for different availability zones (Remaining as of today: Meyrin critical area & Wigner DC);
- Progress: https://cern.service-now.com/service-portal/view-outage.do?n=OTG0045522

Impact on Performance

- Thorough tests at CERN using HS2006, SPEC2017, ATLAS KV. Results coherent with Intel®

Results — <=> no change within +/- 1%

HVs	VMs	Step		pnode Model name family-model-step CPU model OS project VMs	i69234828294880 Broadwell 06-3f-02 E5-2680 v4 @ 2.40GHz gva_project_035 2x vm14 SLC6 Ratio	 2x vm14 CC7 Ratio	p06253971a02939 Haswell 06-4f-01 E5-2630 v3 @ 2.40GHz wig_project_011 SLC6 2x vm8, Ratio	 SLC6 1x vm16 Ratio
i69234828294880.cern.ch 3.10.0-693.17.1.el7.x86_64 p06253971a02939.cern.ch 3.10.0-693.11.6.el7.x86_64 microcode_ctl-2.1-29.10.el7_5.x86_64	SLC6 2.6.32-696.18.7.el6.x86_64 microcode_ctl-1.17-33.3.el6_10.x86_64	Scores before microcode update for Speculative Store Bypass	HS06 SPEC2017 KV					
	SLC6 2.6.32-754.3.5.el6.x86_64 l1tf:Mitigation: PTE Inversion CC7 3.10.0-862.11.6.el7.x86_64 l1tf:Mitigation: PTE Inversion; VMX: SMT disabled, L1D conditional cache flushes	Scores after VM kernel patch for L1TF (reboot VMs)	HS06 SPEC2017 KV		1.00 1.00 0.99		1.00 1.01 0.95	1.00 1.00 1.00
3.10.0-862.11.6.el7.x86_64 l1tf:Mitigation: PTE Inversion; VMX: SMT vulnerable, L1D conditional cache flushes meltdown:Mitigation: PTI spec_store_bypass:Mitigation: Speculative Store Bypass disabled via prctl and seccomp spectre_v1:Mitigation: Load fences, __user pointer sanitization spectre_v2:Mitigation: Full retpoline		Scores after HV kernel patch for L1TF (reboot VMs and HVs)	HS06 SPEC2017 KV		0.99 0.99 0.98	0.99 1.01 1.02	0.99 0.99 1.00	0.99 1.00 0.98
		Overall Effect	HS06 SPEC2017 KV		1.00 1.00 0.98	0.99 0.99 1.00	1.01 1.01 0.94	1.02 1.01 1.01

Courtesy Domenico Giordano (CERN IT-CM)

- Despite this, SMT OFF **caused 20% performance drop** when testing with HS2006 (hence, as LXBATCH is trusted, SMT is here kept ON)
- OpenStack “service” VMs allocated by RAM, not CPU: **vCPUs usually overcommitted.** Only on HVs where all vCPUs run at their limit (i.e. HS06-like workloads), performance might drop. Deemed unlikely & tightly monitored during patching (Please report problems via SNOW).
- On one H/W type (high overcommit of 256GB RAM for 24CPUs) performance problems seen. Hence, CERN sticks here to SMT ON until issue fully understood to be mitigated.