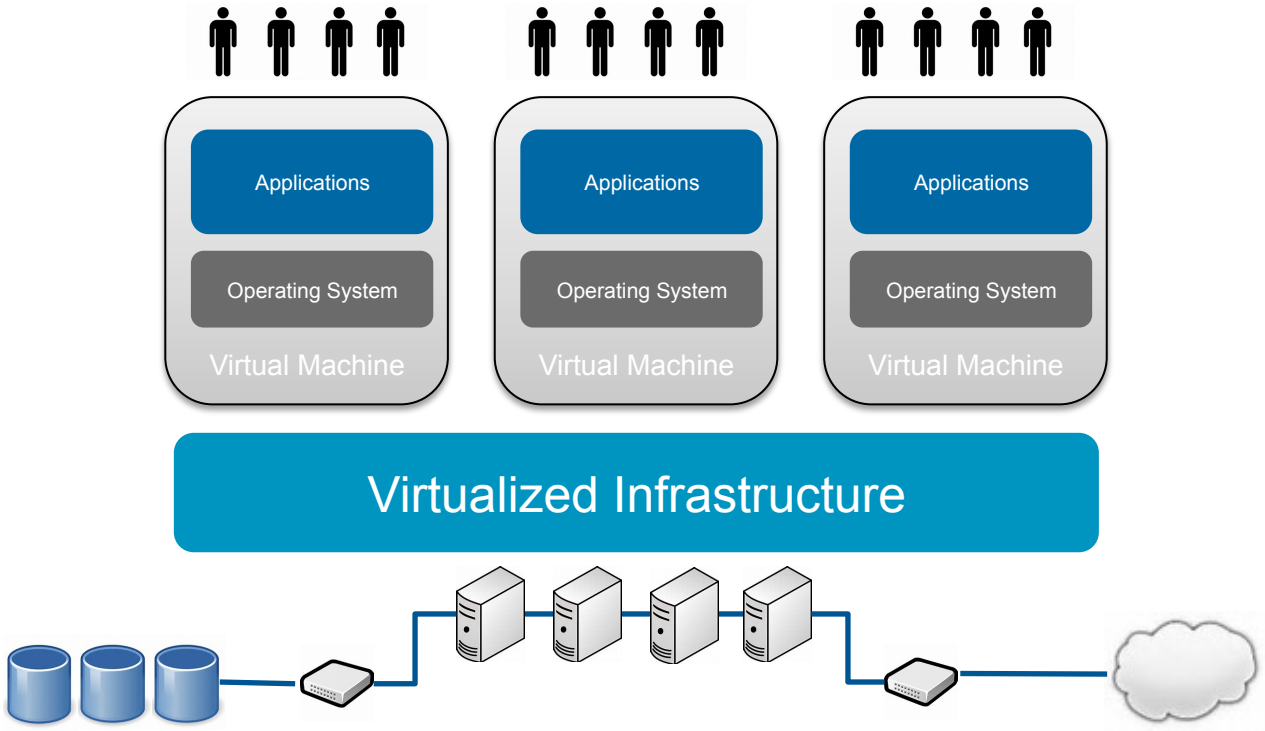
SGX memory oversubscription

Somnath Chakrabarti, Rebekah Leslie-Hurd, Mona Vij, Frank McKeen, Carlos Rozas, Dror Caspi, Ilya Alexandrovich, Ittai Anati

{somnath.chakrabarti, rebekah.leslie-hurd, mona.vij, frank.mckeen, carlos.v.rozas, dror.caspi, ilya.alexandrovich, ittai.anati}@intel.com

June 25, 2017

Virtualization Recap



What

- Virtualization allows multiple operating systems to run virtually on a single physical platform
- Virtualized infrastructure is responsible for management of physical resources and their allocation to various VMs

Why

- Virtual Infrastructure in a datacenter/cloud makes it possible to dynamically map resources to businesses
- Results in reduced cost and increased efficiency for businesses

How

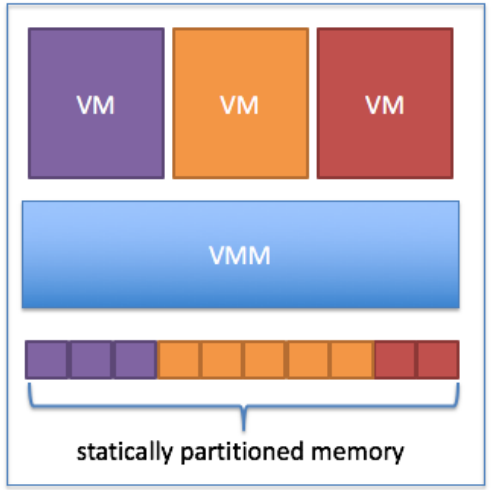
- VMMs partition the physical resources and let guests manage them on their own
- Modern VMMs use oversubscription mechanisms to allocate more resources than available and shares them between VMs

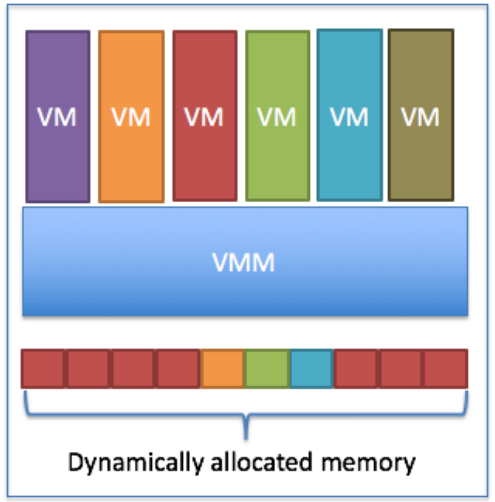
Motivation

- SGX 1.0 primarily useful in client and limited server scenarios
  - More advanced usages possible, but often with significant software complexity and performance limitations
- The vast majority of the 60+ papers addressing SGX so far focus on potential uses in the datacenter
- Cloud and datacenter platforms have unique challenges with regard to virtualization and shared platform resources
- We are introducing SGX extensions to make it more useful in the datacenter, today we will discuss platform memory oversubscription

VMM Memory Oversubscription

VMM Memory Oversubscription allocates more memory to virtual machines than what is actually available on the platform





Memory Allocation Schemes

- Partitioning (No Oversubscription)
  - VMM statically partitions the memory
  - No guest involvement

✓ SGX Supported

- Ballooning
  - VMM can dynamically move the memory between guests
  - Guest explicitly requests and releases memory

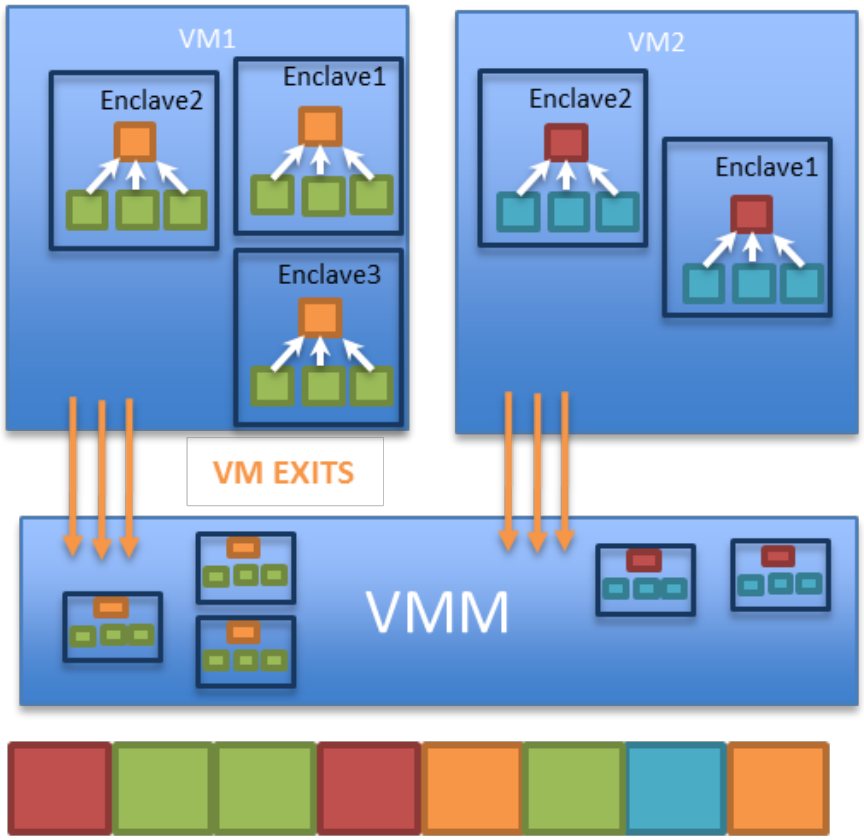
✓ SGX Supported

- Paging
  - VMM can dynamically move the memory between guests
  - No guest involvement

✗ **SGX support Challenging**

VMM SGX Oversubscription Challenges

- SGX memory has a hierarchical structure
  - VMM paging needs a way to efficiently track that hierarchy
- Tracking SGX memory hierarchy is complex
  - VMM intercepts guest SGX operations
  - Emulates ENCLS instructions
  - Constructs SGX memory map
- Prevent paging operations in VMM and guest from occurring at the same time
  - Simultaneous paging operations can cause unexpected fault conditions in the guest
  - VMM intercepts guest SGX operations to prevent this condition
- Overheads
  - VM Exits and emulation add to execution time - ~60% overhead for paging, ~100% overhead for enclave build/teardown
  - SGX memory map consumes significant host memory



Cloud customers have requested to simplify the oversubscription of SGX memory

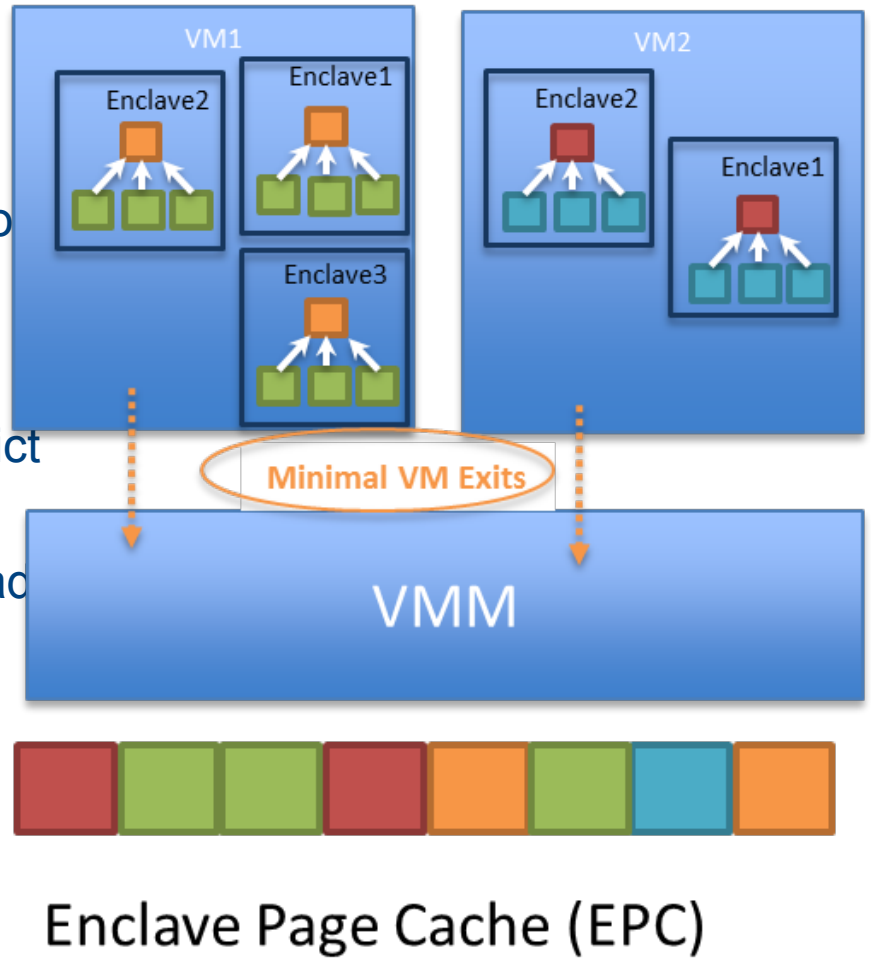
SGX Oversubscription Architecture Overview

- Built an architecture to avoid VM exits, emulation and guest pausing
- SGX Extensions
  - Provide new instructions to discover and virtualize parent/child relationships
  - Extended SGX architecture to provide conflict free paging
- VT Extensions
  - Added a new opcode for VMM only execution (ENCLV)
  - New exit support for virtualization and conflict handling

VMM SGX Oversubscription Architecture benefits

Architectural extensions (New SGX instructions and VT controls) to avoid VM exits, emulation and guest pausing

- Memory Savings
  - Architecture maintains the hierarchy and provides new instructions to discover and virtualize parent/child relationships
- Performance Benefit
  - Guest and VMM can do paging operations simultaneously with conflict free SGX architecture extensions
  - Minimal VM Exits only in case of lock conflicts - ~0% paging overhead
  - No overhead for enclave build/teardown



SGX Oversubscription ISA

ENCLS[ERDINFO] instruction

- Provides the VMM with information about a given EPC page (type, EPCM attributes, SECS context)
- For SECS pages, indicates whether or not the enclave has resident children

ENCLV[ESETCONTEXT] instruction

- Provides a mechanism for the VMM to store context specific value in SECS.ENCLAVECONTEXT field.
- Enables VMM to keep track of enclave Parent/Child relationship

ENCLV[EINCVIRTCHILD/EDECVIRTCHILD]

- Enables VMM to pin a SECS page in EPC memory, even when all child pages are evicted out
- Increments/Decrements VIRTCHILDCOUNT inside SECS, checked when guest executes EREMOVE and EWB

VIRTCHILDCOUNT tracking opt-in

- Allows VMM to enable VIRTCHILDCOUNT check by EWB and EREMOVE when executed inside guests
- Changes in EWB and EREMOVE to check VIRTCHILDCOUNT

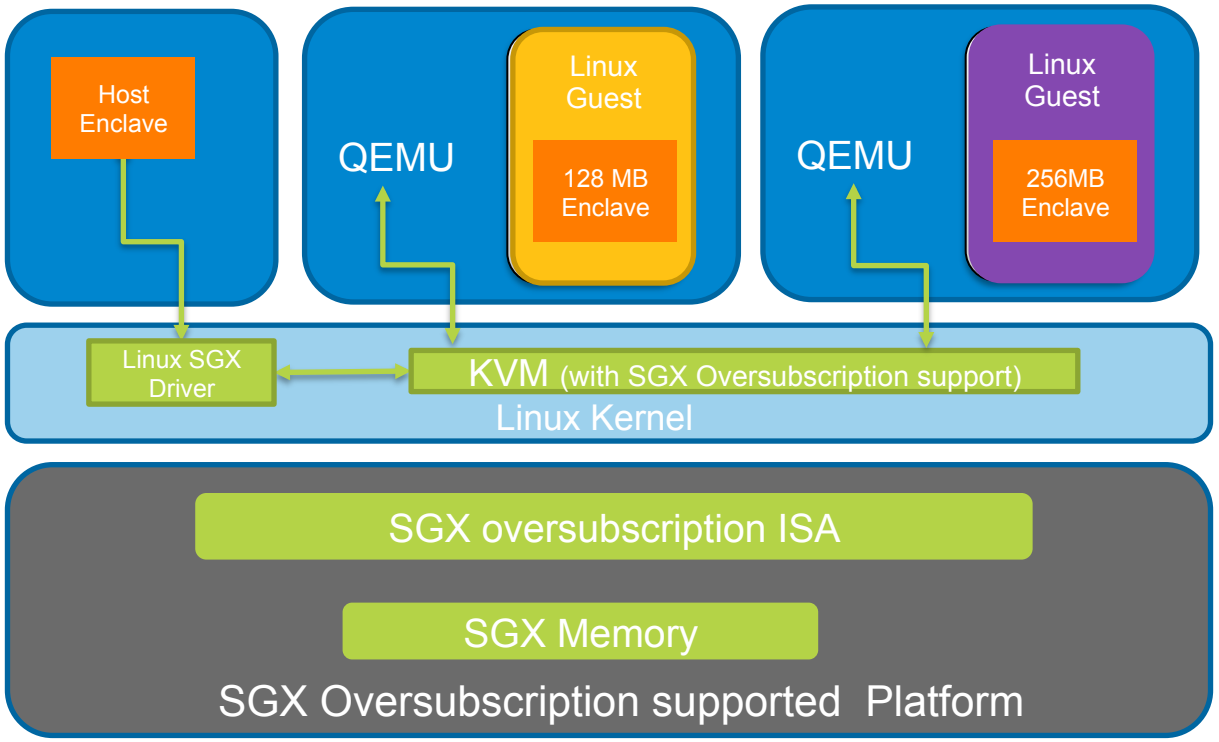
ENCLS[ETRACKC & ELDC)]

- New concurrent ETRACK and ELD variant that supports lock conflict handling by the VMM
- Lock conflicts encountered by VMM is reported as an error code rather than a #GP

SGX CONFLICT VM exiting

- Allows VMM to receive VM exit when guest encounters an unexpected failure in executing any SGX instructions
- Failure in guest may have been caused by VMM interference

SGX Oversubscription Linux setup



Built an internal prototype for architectural and software feasibility of new architecture

- Exercising all the new SGX and VT extensions
- Demonstrating interoperability with current SGX instructions

- With new Architecture
  - **Architecture** maintains parent-child relationship
  - Guest and VMM can do **paging simultaneously**
  - SGX Instructions in guest **do not** cause VM exits
  - VMM gets **VM exit only on conflicting scenarios without impacting guest flow**

SGX Oversubscription feature availability

- Look in the paper for more details
- Planned to intercept future generation Intel CPUs
- Reference KVM and XEN implementation will be made available
- More details about the architecture will be published in future version of Intel® SDM

Thank
You