第 4 讲：Optimization of Virtual Machine Monitor

第三节：Dune: Safe User-level Access to Privileged CPU Features

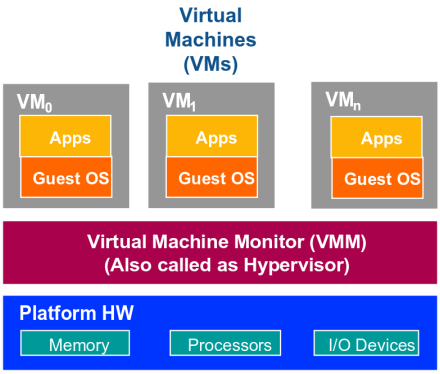
陈渝

清华大学计算机系

yuchen@tsinghua.edu.cn

2020 年 3 月 8 日

Requirement of DUNE

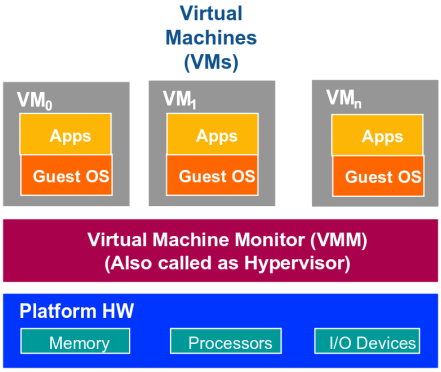


For MORE performance & features

Safe User -- level Access to Privileged CPU Features

Dune: Safe User--level Access to Privileged CPU Features, Adam Belay,etc., OSDI'12

Requirement of DUNE

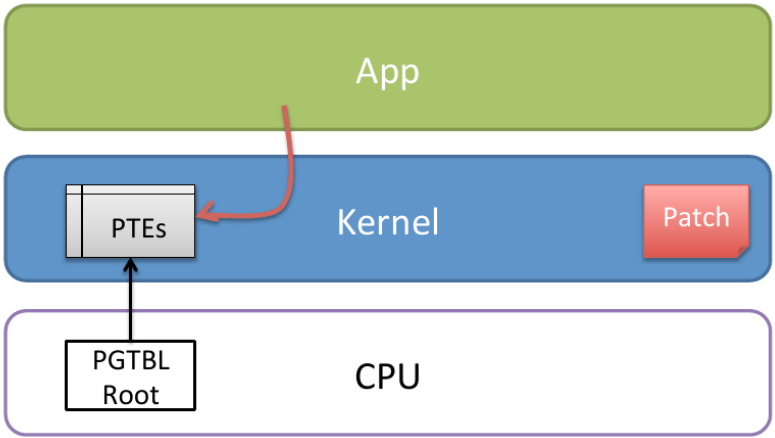


For MORE performance & features

- Speed up garbage collection (Azul C4)
  pagetable
- Privilege separation within a process (Palladium)
  MMU
- Safe native code in web browsers (Xax)
  Syscall handler

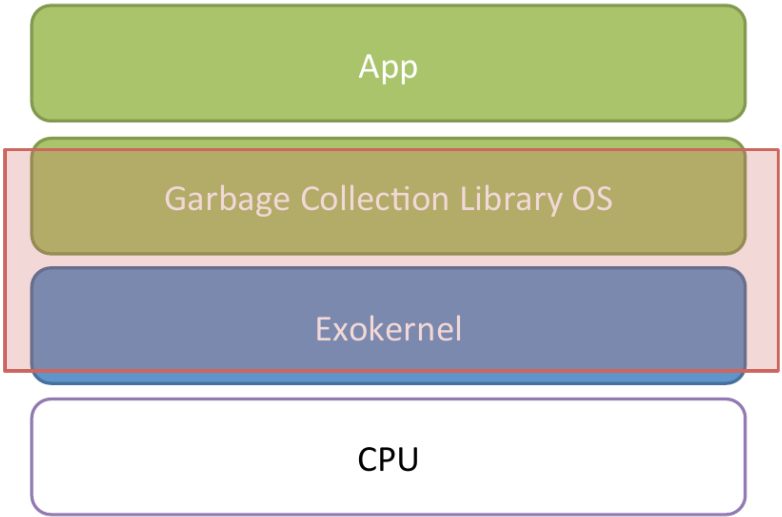
Some thoughts of DUNE – Change kernel

Problem: stability concerns, challenging to Optimization analysis distribute, composability concerns



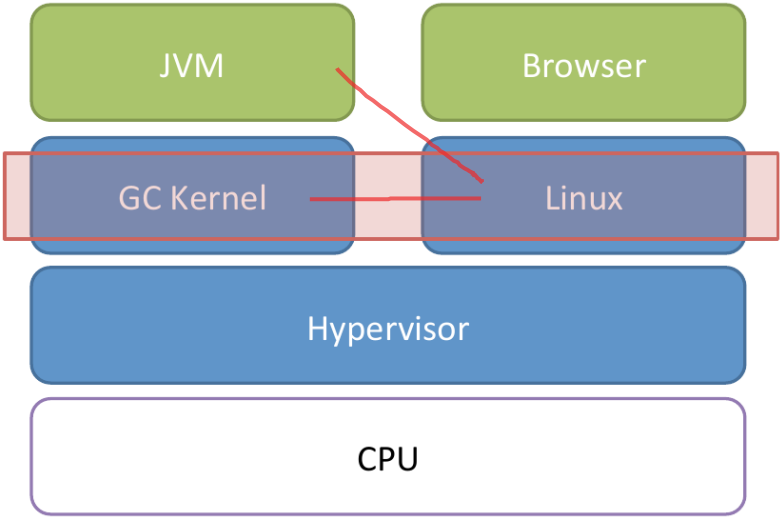
Some thoughts of DUNE – exokernel

Problem: must replace entire OS stack

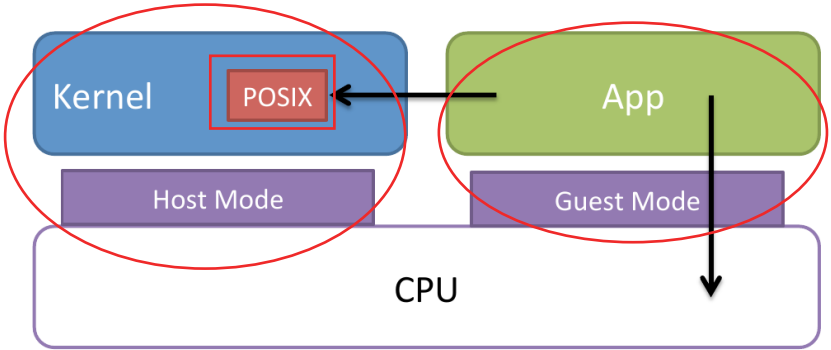


Some thoughts of DUNE – VMM

Problem: virtual machines have strict partitioning

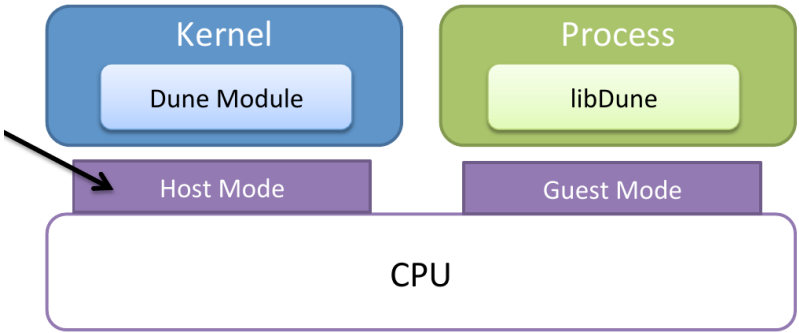


Some thoughts of DUNE – Dune in a Nutshell



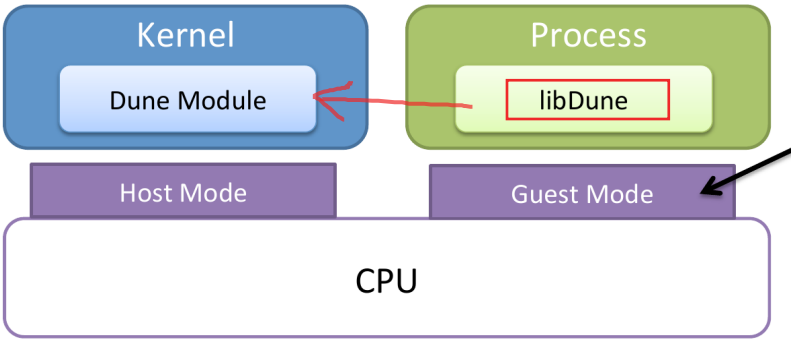
- Provide safe user--level access to privileged CPU features
- Still a normal process in all ways (POSIX API, etc)
- Key idea: leverage existing virtualization hardware (VT-x)

Some thoughts of DUNE – Dune Simple Arch



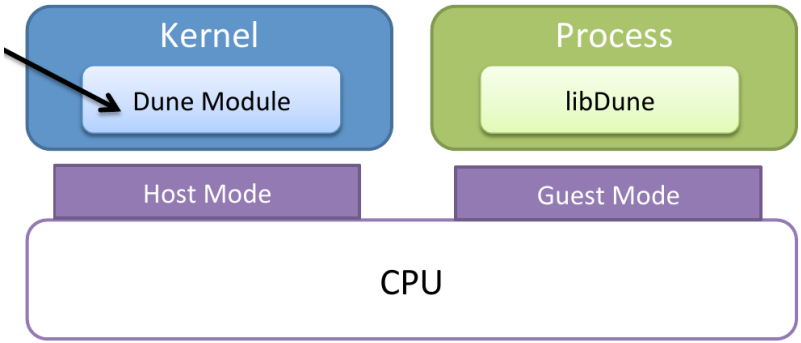
- Host mode --> VMX root mode on Intel
- Normally used for hypervisors
- In Dune, we run the kernel here，for access VT-x instructions.

Some thoughts of DUNE – Dune Simple Arch



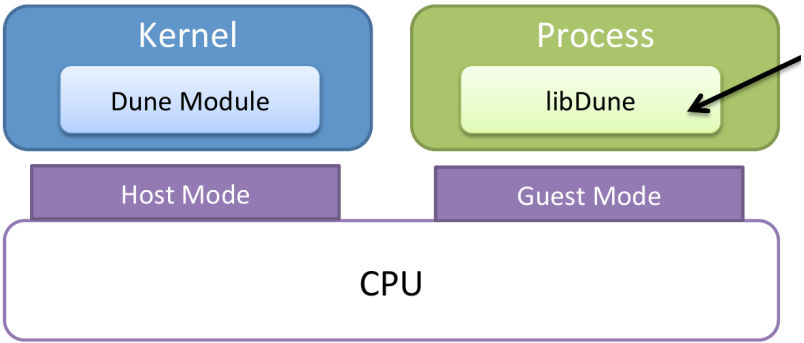
- Guest mode --> VMX non--root mode on Intel
- Normally used by the guest kernel
- In Dune, we run ordinary processes here, for access to privileged features

Some thoughts of DUNE – Dune Simple Arch



- Configures and manages virtualization hardware
- Provides integration with the rest of the kernel in order to support a process abstraction
- Uses Intel VT-x

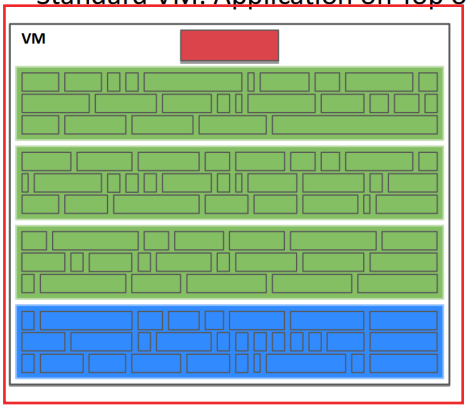
Some thoughts of DUNE – Dune Simple Arch



- A uAlity library to help applicaAons manage privileged hardware features
- Completely untrusted
- Exception handling, syscall handling, page allocator, page table management, ELF loader

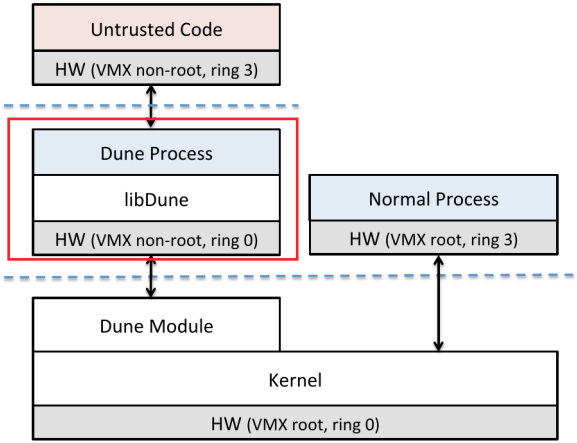
Diff Between VMM & DUNE

Standard VM: Application on Top of Distro

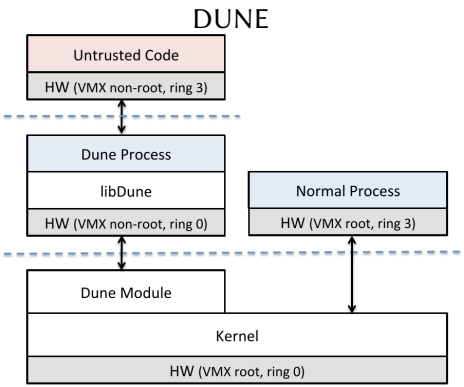


vmcall

DUNE: using virtualization hardware to provide a process

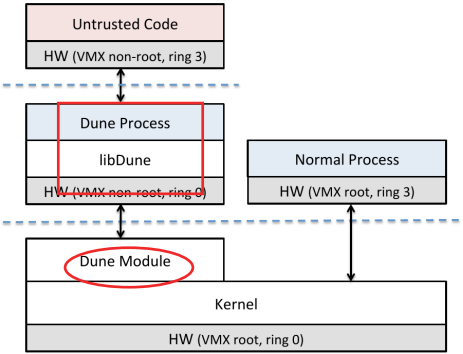


Contributions of DUNE



- a design that uses hardware-assisted virtualization to safely and efficiently expose privileged hardware features to user programs while preserving standard OS abstractions.
  - Memory management
  - System calls
  - POSIX Signals

Supported Hardware Features



Hardware features exposed by Dune and their corresponding privileged x86 instructions.

Mechanism	Privileged Instructions
Exceptions	LIDT, LTR, IRET, STI, CLI
Virtual Memory	MOV CRn, INVLPG, INVPCID
Privilege Modes	SYSRET, SYSEXIT, IRET
Segmentation	LGDT, LLDT

Supported Hardware Features – Exceptions

Hardware features exposed by Dune

Mechanism	Privileged Instructions
Exceptions	LIDT, LTR, IRET, STI, CLI
Virtual Memory	MOV CRn, INVLPG, INVPCID
Privilege Modes	SYSRET, SYSEXIT, IRET
Segmentation	LGDT, LLDT

- Normally, reporting an exception to a user program requires privilege mode transitions and an upcall mechanism (e.g., signals)
- Dune can reduce exception overhead because it uses VT-x to deliver exceptions directly in hardware.
- proves the speed of delivering page fault exceptions by more than 4 X

Supported Hardware Features – Virtual Memory

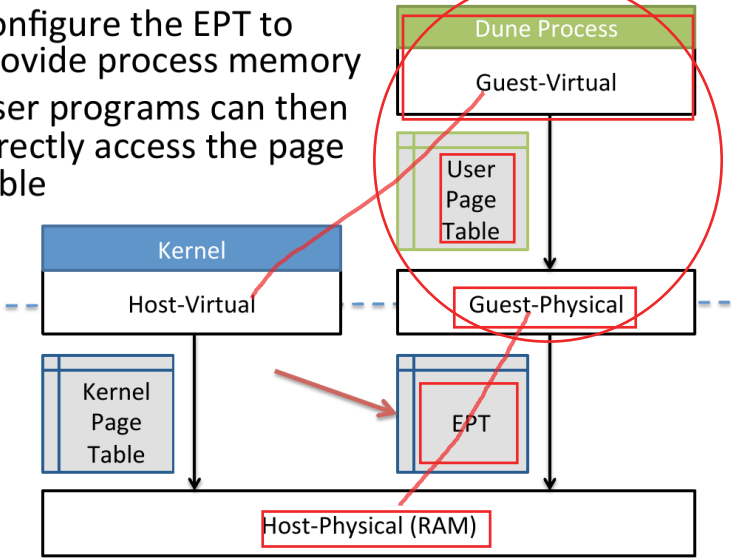
Hardware features exposed by Dune

Mechanism	Privileged Instructions
Exceptions	LIDT, LTR, IRET, STI, CLI
Virtual Memory	MOV CRn, INVLPG, INVPCID
Privilege Modes	SYSRET, SYSEXIT, IRET
Segmentation	LGDT, LLDT

- gives user programs the ability to manually control TLB invalidations.
- page table updates can be performed in batches when permitted by the application.
- Dune exposes TLB tagging by providing access to Intel's recently added process-context identifier (PCID) or virtual-processor identifiers (VPID) feature
- Dune results in a 7× speedup over Linux in the Appel and Li user-level virtual memory benchmarks

Supported Hardware Features – Virtual Memory

- Configure the EPT to provide process memory
- User programs can then directly access the page table



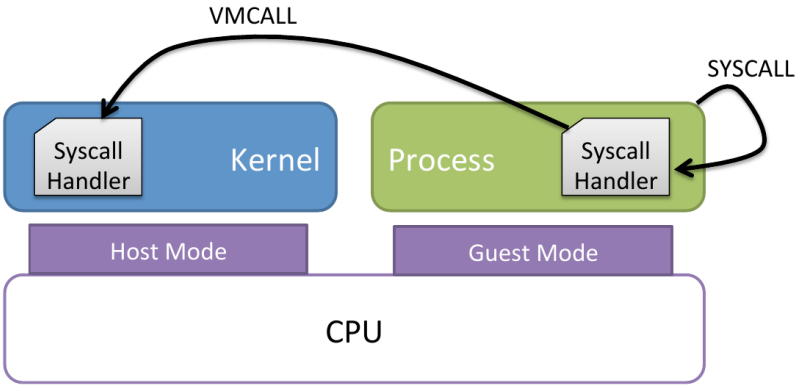
Supported Hardware Features – Privilege Modes

Hardware features exposed by Dune

Mechanism	Privileged Instructions
Exceptions	LIDT, LTR, IRET, STI, CLI
Virtual Memory	MOV CRn, INVLPG, INVPCID
Privilege Modes	SYSRET, SYSEXIT, IRET
Segmentation	LGDT, LLDT

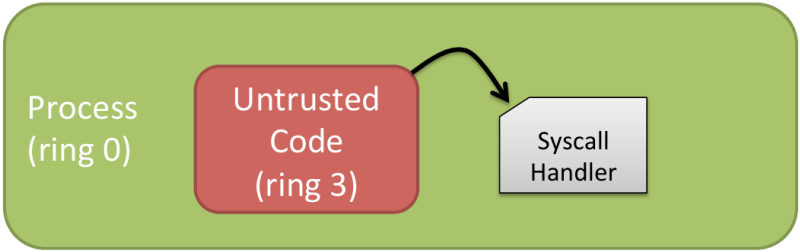
- Two motivating use cases for privilege modes are privilege separation and sandboxing of untrusted code.
- page table updates can be performed in batches when permitted by the application.
- system call instructions trap to the process itself, rather than to the kernel,
- can be used for system call interposition and to prevent untrusted code from directly accessing the kernel.
- Compared to ptrace in Linux, we show that Dune can intercept a system call with 25 X less overhead

Supported Hardware Features – Privilege Modes



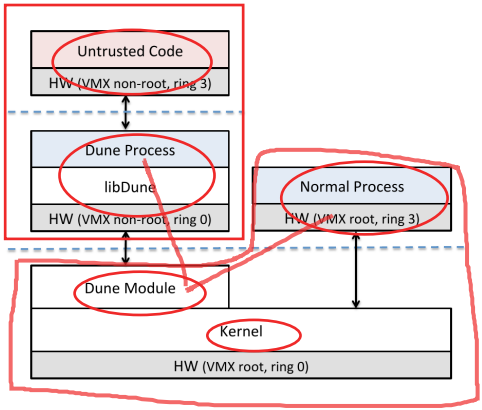
- SYSCALL will only trap back into the process
- Use VMCALL (i.e. a hypercall) to perform normal kernel system calls

Supported Hardware Features – Privilege Modes



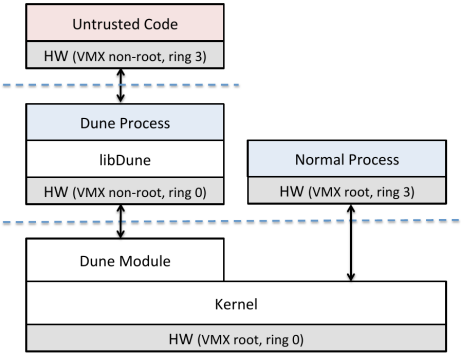
- Isolate untrusted code by running it in a less privileged mode (i.e. ring 3 on x86)
- Leverage the 'supervisor' bit in the page table to protect memory

Implementation Challenges

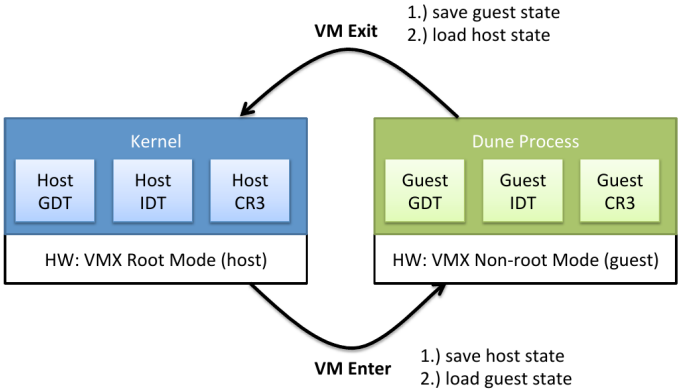


- Reducing VM exit and VM entry overhead
- Pthread and fork were tricky to integrate with the Linux kernel
- EPT does not support enough address space
- Signals should only be delivered to ring 0, but process is in ring 3

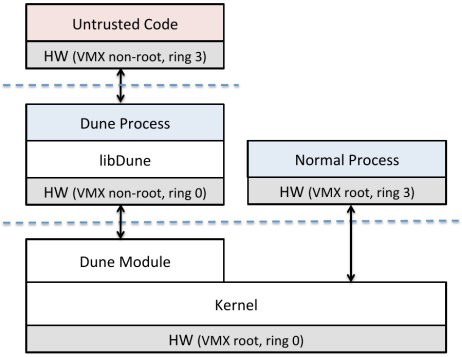
Implementation Challenges



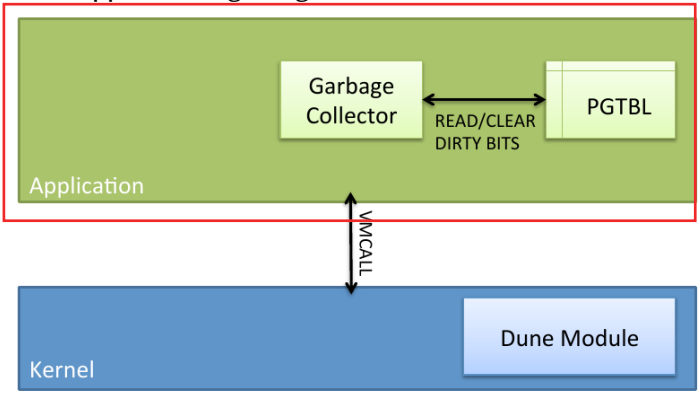
- Reducing VM exit and VM entry overhead



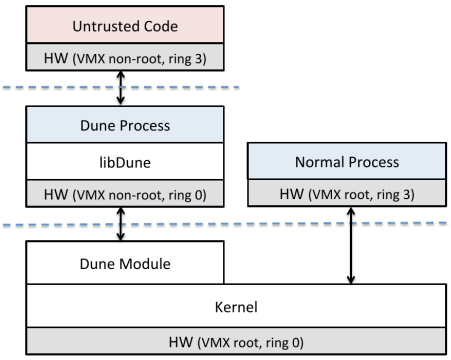
Implementation Challenges



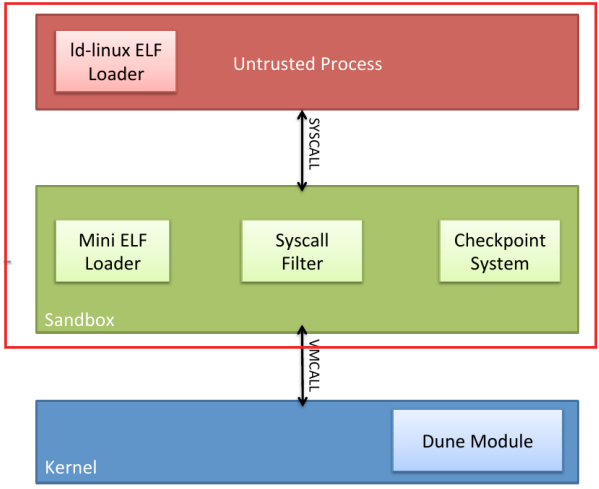
- Application: garbage collection



Implementation Challenges



- Application: sandbox



Performance

Overhead analysis : VMX trans, EPT trans

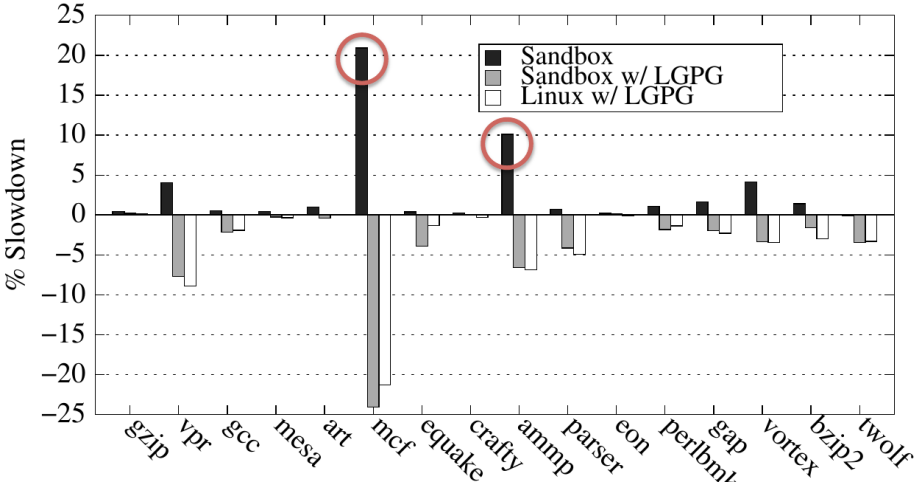
(cycles)	Getpid	Page fault	Page walk
Linux	138	2,687	36
Dune	895	5,093	86

Optimization analysis : Faster system call，Virt Mem manipulation

(cycles)	ptrace (getpid)	trap	Appel 1 (TRAP, PROT1, UNPROT)	Appel 2 (PROTN, TRAP, UNPROT)
Linux	27,317	2,821	701,413	684,909
Dune	1,091	587	94,496	94,854

Performance

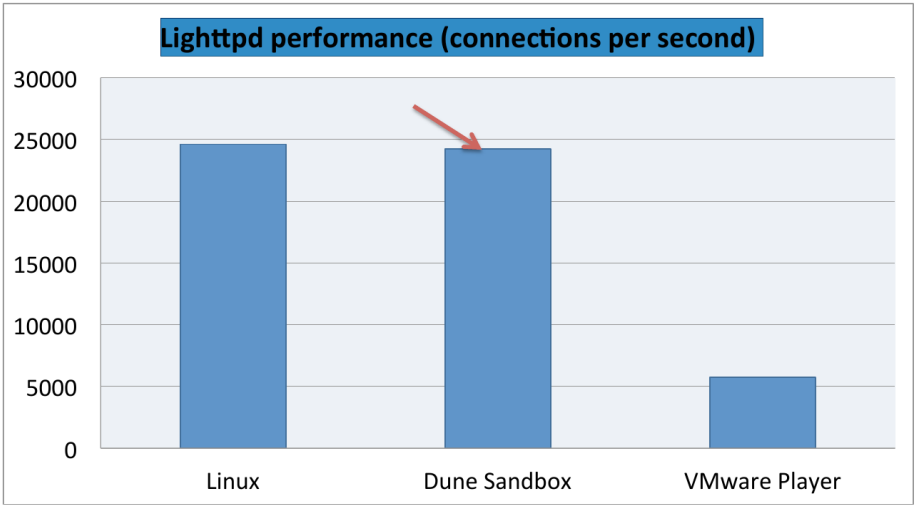
Sandbox: SPEC2000 performance



EPT overhead: use of large pages

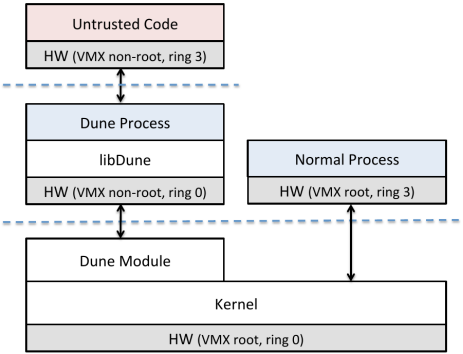
Performance

Sandbox: Lighttpd performance



Slight reduction in throughput (less than 2%) due to VMCALL overhead

Conclusions



- Applications can benefit from access to privileged CPU features
- Virtualization hardware allows us to provide such access safely
- Dune creates new opportunities to build and improve applications without kernel changes
- Dune has modest performance overhead