

Wirtualizacja wspomagana sprzętowo

Marek Adamczyk

Filip Grotkowski

Piotr Jaroszyński

Adam Witkowski

Wirtualizacja wspomagana sprzętowo

Plan prezentacji

- Wirtualizacja bez wspomagania sprzętowego
- Pierwsze podejścia do wspomagania sprzętowego
- Tablice stron
- Bezpośredni dostęp do urządzeń
- Uproszczenie dzięki wsparciu sprzętowemu
- Zagrożenia związane ze wsparciem sprzętowym
- Porównanie obu rodzajów wirtualizacji

Wirtualizacja wspomagana sprzętowo

Wirtualizacja

Zajmujemy się tym rodzajem wirtualizacji,
który symuluje działanie całego komputera.

Wirtualizacja wspomagana sprzętowo

Warto rozmawiać

„First dual-core in 2005, then quad-core in 2007: the multi-core snowball is rolling. The desktop market is still trying to find out how to wield all this power; meanwhile, the server market is eagerly awaiting the octal-cores in 2009. The difference is that the server market has a real killer application, hungry for all that CPU power: virtualization.”

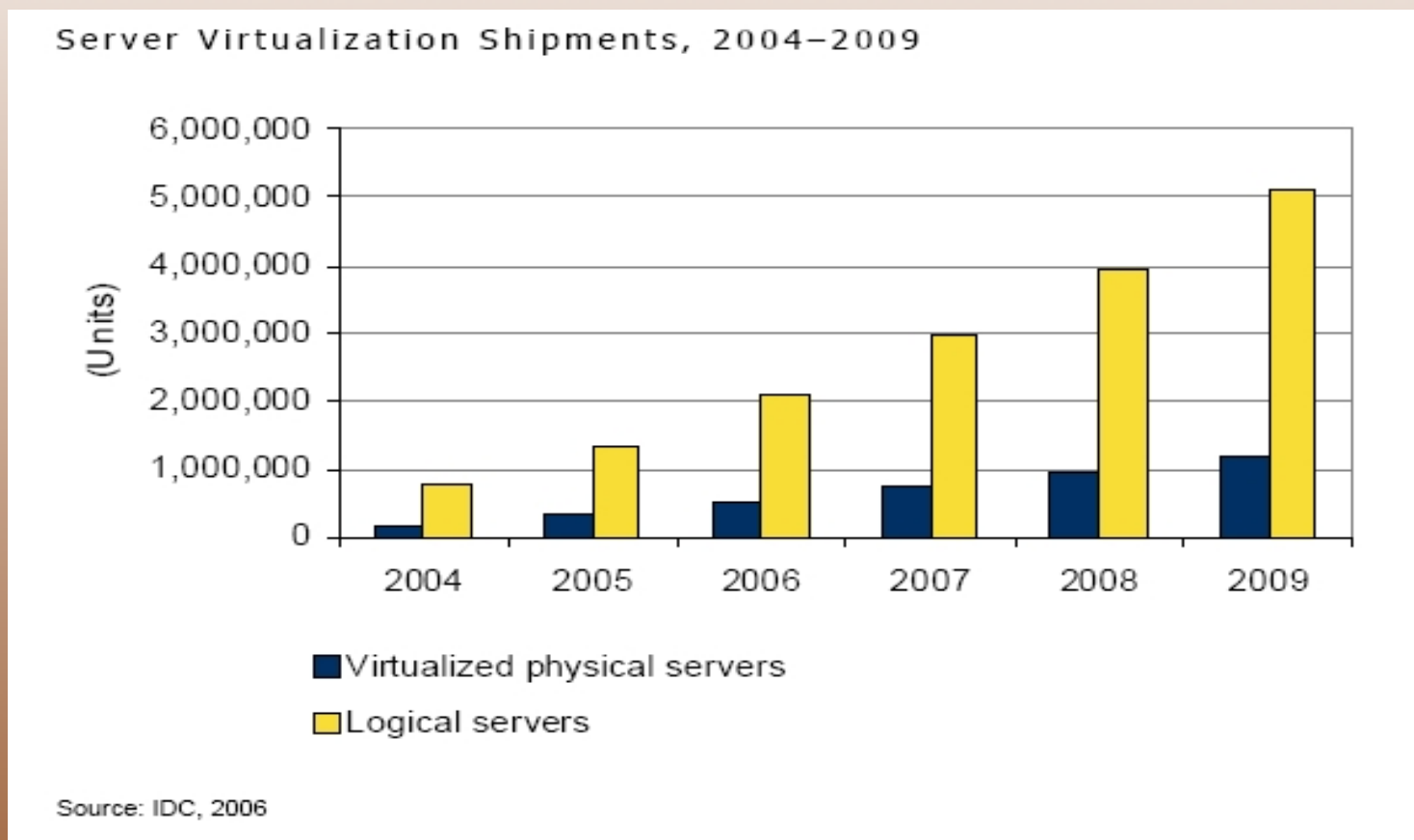
Wirtualizacja wspomagana sprzętowo

Warto rozmawiać

Wirtualizacja umożliwia lepsze wykorzystanie zasobów komputerowych.

Wirtualizacja wspomagana sprzętowo

Warto rozmawiać



Wirtualizacja wspomagana sprzętowo

Jest wiele rodzajów wirtualizacji:

- emulacja
- parawirtualizacja
- pełna wirtualizacja

Będziemy głównie mówić o pełnej wirtualizacji,
wspomnimy też o parawirtualizacji.

Wirtualizacja wspomagana sprzętowo

Czego należy wymagać od wirtualizacji?

Odpowiedniość

Kontrola zasobów

Wydajność

Wirtualizacja wspomagana sprzętowo

Odpowiedniość

Program działający na maszynie wirtualnej musi zachowywać się w dokładnie taki sam sposób, jak na rzeczywistym sprzęcie.

Wirtualizacja wspomagana sprzętowo

Kontrola zasobów

Wirtualna maszyna musi w pełni kontrolować wszystkie zasoby, które są wirtualizowane.

Tak naprawdę, nie maszyna wirtualna,
ale Virtual Machine Monitor.

VMM dla SO gościa, jest niczym zwykły SO
dla procesów i wątków.

Wirtualizacja wspomagana sprzętowo

Wydajność

Większa część instrukcji musi być wykonywana bez udziału maszyny wirtualnej.

Wirtualizacja wspomagana sprzętowo

Czy można spełnić wszystkie te warunki?

Okazuje się, że tak.

Wirtualizacja wspomagana sprzętowo

Jeśli tylko

Zbiór instrukcji wrażliwych
jest zawarty w zbiorze instrukcji uprzywilejowanych.

Wirtualizacja wspomagana sprzętowo

Instrukcje wrażliwe

Instrukcje wrażliwe to takie, których wykonanie ma związek z konfiguracją zasobów systemowych.

Instrukcje wrażliwe ze względu na kontrolę - mogą zmienić konfigurację zasobów systemowych.

Instrukcje wrażliwe ze względu na wykonanie - ich działanie jest zależne od konfiguracji systemu.

Wirtualizacja wspomagana sprzętowo

Instrukcje uprzywilejowane

Wykonanie instrukcji uprzywilejowanej w trybie użytkownika skutkuje przerwaniem lub wywołaniem systemowym.

Wirtualizacja wspomagana sprzętowo

Twierdzenie Popka-Goldberga, 1974

Dla każdego standardowego komputera trzeciej generacji wirtualna maszyna może zostać skonstruowana, jeśli zbiór instrukcji wrażliwych jest podzbiorem zbioru instrukcji uprzywilejowanych.

Wirtualizacja wspomagana sprzętowo

Szkic dowodu

System operacyjny gościa działa w trybie o mniejszym priorytecie niż system operacyjny gospodarza. Wykonuje się normalnie, dopóki nie wykona instrukcji wrażliwej. Z założenia instrukcja wrażliwa jest uprzywilejowana, więc procesor zaczyna krzyczeć. VMM przechwytuje ten wyjątek i wykonuje odpowiednie działania zastępcze.

Ta metoda to „klasyczna wirtualizacja”.

Wirtualizacja wspomagana sprzętowo

Twierdzenie Popka-Goldberga

Jedna z najpopularniejszych architektur nie spełnia założeń twierdzenia.

Wirtualizacja wspomagana sprzętowo

Problemy z procesorami x86

W procesorach x86 istnieją instrukcje wrażliwe, które nie są uprzywilejowane.

Wirtualizacja wspomagana sprzętowo

popf pushf

Jedne z najczęściej wykonywanych instrukcji.

Wirtualizacja wspomagana sprzętowo

Założenia nie są spełnione, ale może da się

Niestety nie. System operacyjny gościa może używać rejestru cs, z którego można odczytać aktualny priorytet. System operacyjny może się więc dowiedzieć, że nie działa w odpowiednio wysokim trybie. Nie jest więc spełniony warunek odpowiedniości. Jeśli obsługą takich sytuacji obarczylibyśmy VMM, to nie spełnilibyśmy warunku wydajności.

Wirtualizacja wspomagana sprzętowo

Jak obejść te trudności?

Można zmodyfikować system operacyjny.

Wirtualizacja wspomagana sprzętowo

Jak obejść te trudności?

Można zmodyfikować system operacyjny.

Można zmodyfikować sprzęt.

Wirtualizacja wspomagana sprzętowo

A jak bez ingerencji?

VMWare stosuje binarną translację.
Programy użytkownika działają normalnie.
Binarny kod jądra jest tłumaczony online.
Na podstawie kontekstu wykrywane są instrukcje
uprzywilejowane, a następnie zamieniane
na bezpieczniejsze odpowiedniki.

Wirtualizacja wspomagana sprzętowo

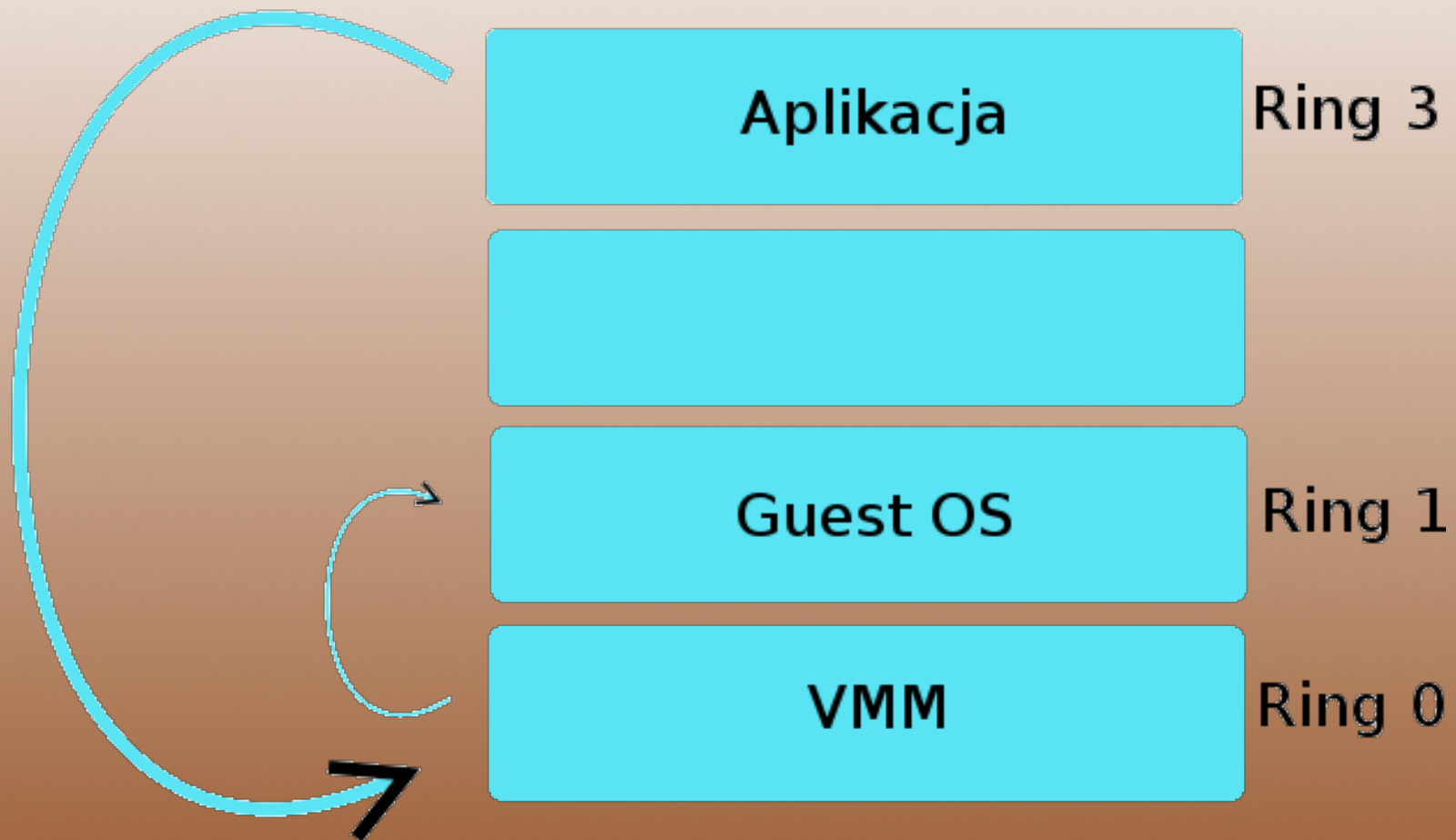
To nie rozwiązuje wszystkich problemów

wywołania systemowe
przerwania
tablice stron
DMA

Ich obsługa wymaga dodatkowej pracy i nie jest prosta,
co powoduje spadek wydajności.

Wirtualizacja wspomagana sprzętowo

Problemy z wywołaniami systemowymi



Wirtualizacja wspomagana sprzętowo

Problemy z DMA

Binarna translacja operuje na danych dla procesora.
Należy wykonać trochę gimnastyki,
skoro DMA chce procesor ominąć.

Wirtualizacja wspomagana sprzętowo

Problemy z tablicami stron

Nie dotyczy tylko binarnej translacji.

Gdzie zmieścić wirtualną tablicę stron?
Co zrobić z wynikami operacji SO gościa
na nieistniejącej tablicy stron?
O tym później.

Wirtualizacja wspomagana sprzętowo

A jak z modyfikacją SO?

Parawirtualizacja jest dość podobna do binarnej translacji. Niebezpieczny kod, który VMWare zamienia podczas binarnej translacji, można zamienić w kodzie źródłowym SO.

To jest oczywiście dość duże uproszczenie, bo modyfikacja źródeł pozwala na bardzo wiele.

Wirtualizacja wspomagana sprzętowo

A jak z modyfikacją sprzętu?

Intel VT-x i AMD SVM próbują spełnić założenia twierdzenia Popka-Goldberga.

Wirtualizacja wspomagana sprzętowo

Intel VT-x i AMD SVM

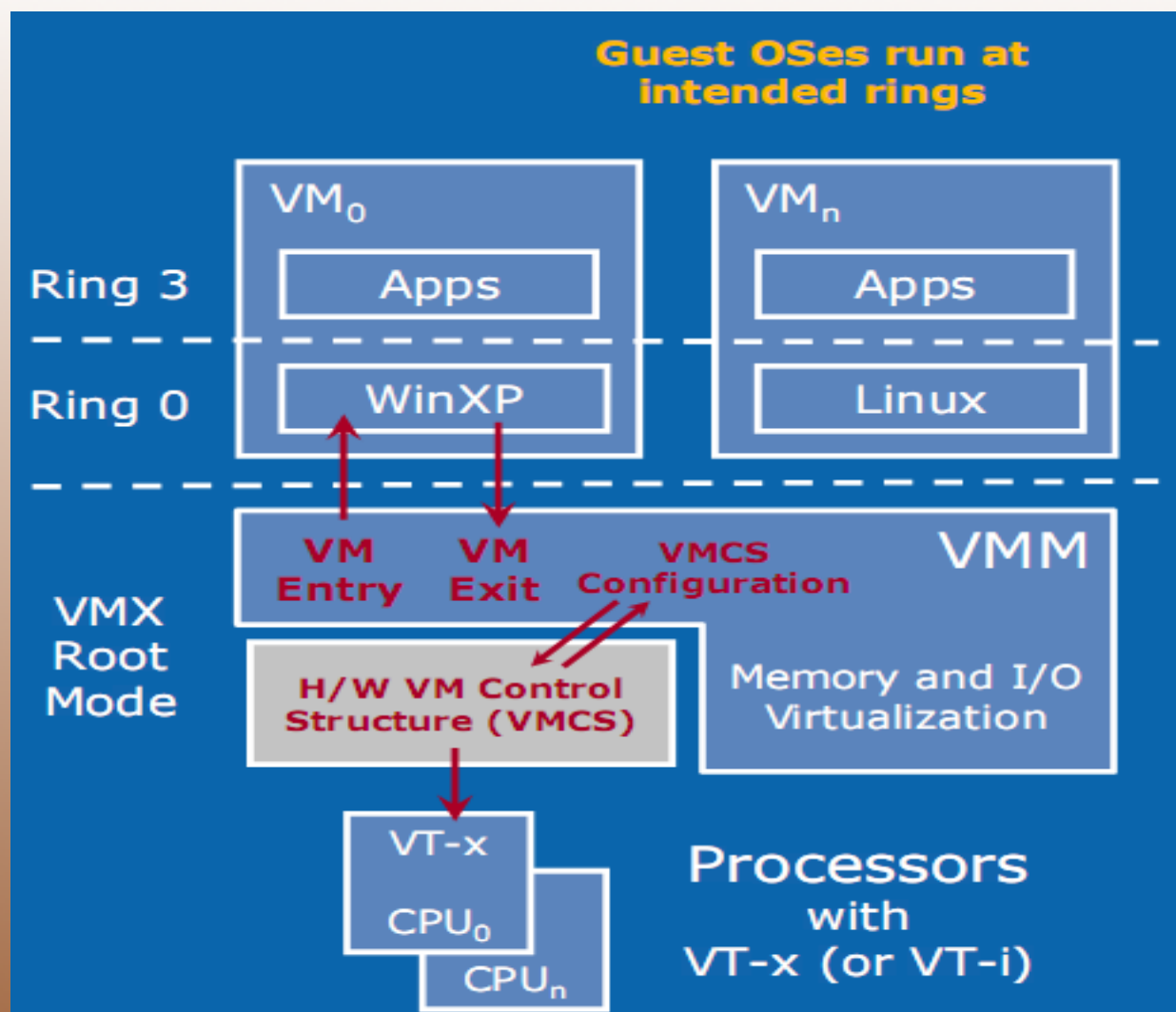
Tryb uprzywilejowania większy niż Ring 0,
tzw. Ring -1 lub Root Mode.

SO gościa działa w Ring 0, a VMM w Ring -1.

SO gościa bezpośrednio odpowiada swoim aplikacjom.

Kiedy SO gościa wykonuje niebezpieczne instrukcje,
VMExit oddaje sterowanie VMM.

Wirtualizacja wspomagana sprzętowo



Wirtualizacja wspomagana sprzętowo

Intel VT-x i AMD SVM

Przejścia VMEntry i VMExit są dość kosztowne.

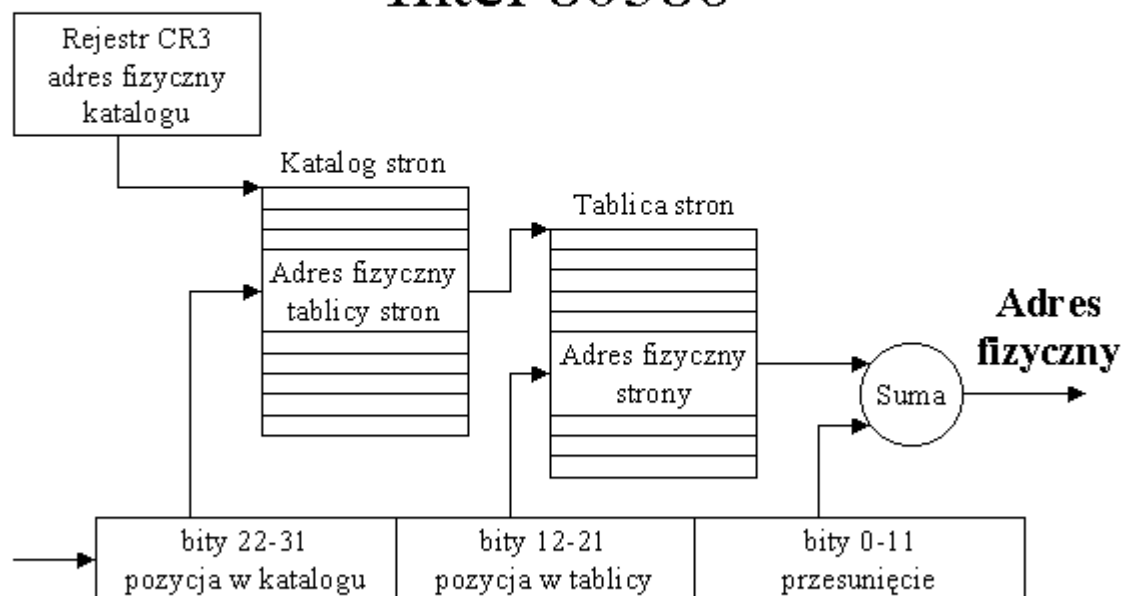
Nie jest to problem, gdy wykonujemy skomplikowane wywołania systemowe.

Jednak prostsze operacje(tworzenie procesów, przełączanie kontekstu, małe modyfikacje tablicy stron) wymagają proporcjonalnie mniej czasu.

Binarna translacja i parawirtualizacja radzą sobie z tym lepiej.

Wirtualizacja wspomagana sprzętowo

Stronicowanie na procesorze Intel 80386



Adres liniowy

Wirtualizacja wspomagana sprzętowo

Shadow Page Table (stronicowanie w tle)

Problem spójności sPT i gPT

Na początek dwa rozwiązania:

- Ochrona przed zapisem fizycznych stron gPT
- Zachłanna technika Virtual TLB

Obie metody owocują w wiele page faultów.

Page faulty indukowane przez gościa i VMM, ten ostatni musi rozróżniać.

Kiedy gość jest aktywny w sPT ustawiane są bity dirty i accessed.

Hypervisor musi te bity odwzorowywać w gPT.

Wirtualizacja wspomagana sprzętowo

Shadow Page Table (stronicowanie w tle)

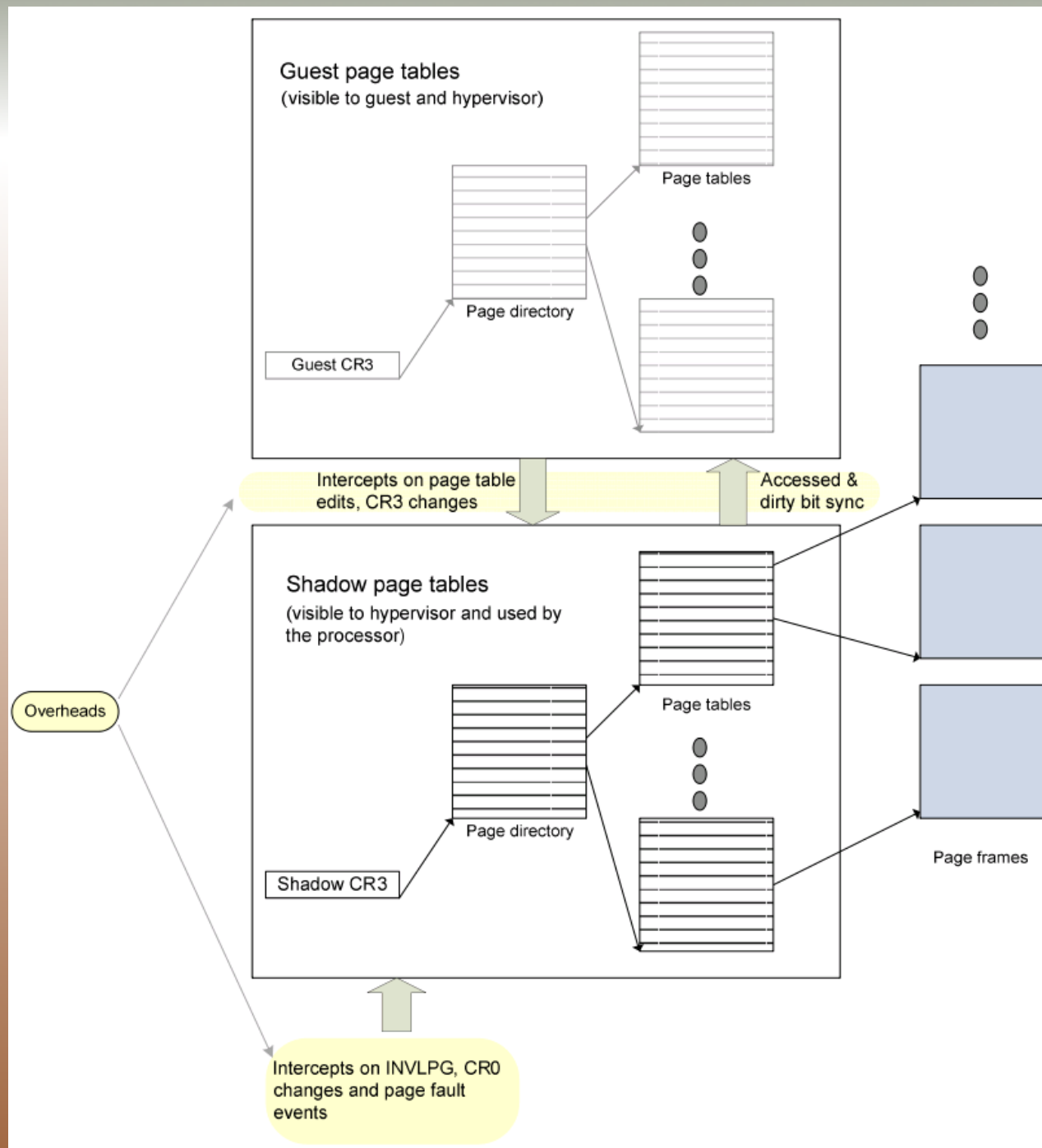
Aktualizacja podwójnych rejestrów, sPT i gPT przy contex switchu.

Przypadek SMP (symmetric multiprocessing):

- gPT może być użyte dla wielu procesorów
- Wtedy VMM ma za zadanie utrzymywać sPT w wielu kopiach dla każdego procesora wirtualnego, lub jednego dla wielu procesorów.

Narzut na synchronizację

Wniosek: aż do **75%** czasu działania VMM spędza na operacjach związanych z sPT.



Wirtualizacja wspomagana sprzętowo

AMD Nested Page Table

- Rozwiązanie dla stronicowania w tle
- Dodatkowe (zagnieżdżone) tablice stron (nPT), by tłumaczyć fizyczne adresy gościa na fizyczne adresy hosta
- Gość ma całkowitą kontrolę nad swoimi tablicami stron
- Nested Paging niweluje koszty związane ze stronicowaniem w tle
- Jedna instancja nPT dla jednego gościa

Wirtualizacja wspomagana sprzętowo

AMD Nested Page Table

Dwu krokowy algorytm tłumaczenia liniowego adresu gościa na fizyczny adres hosta:

- Tłumaczenie adresu liniowego gościa na adres fizyczny gościa (gPT)
- Tłumaczenie adresu fizycznego gościa na adres fizyczny hosta (nPT)

Cache'ujemy translacje w TLB.

W zależności od dwóch wymienionych wyżej kontekstów TLB zawiera tłumaczenia adresów z gPT lub nPT.

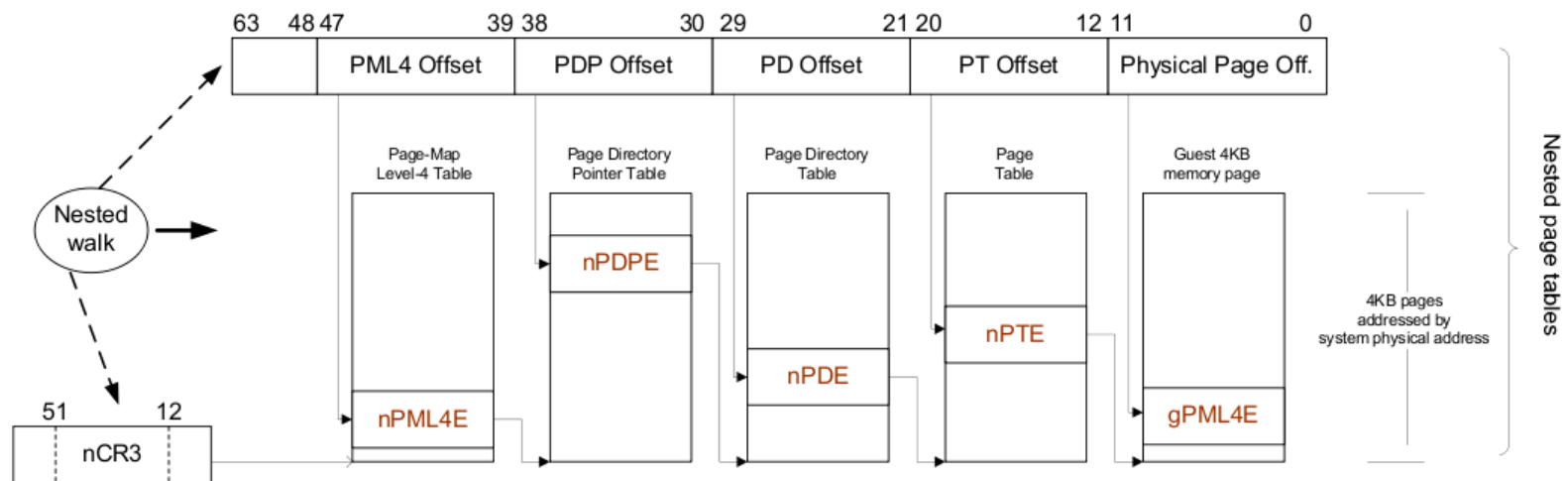
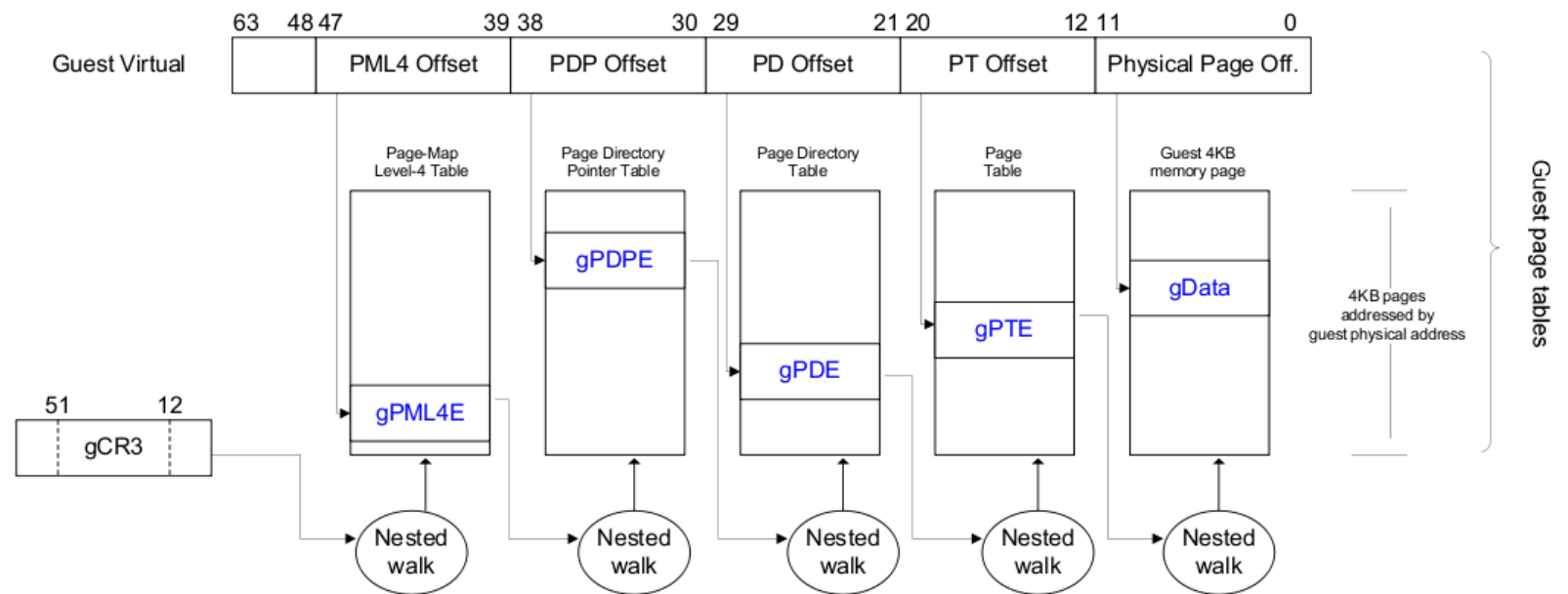
Wirtualizacja wspomagana sprzętowo

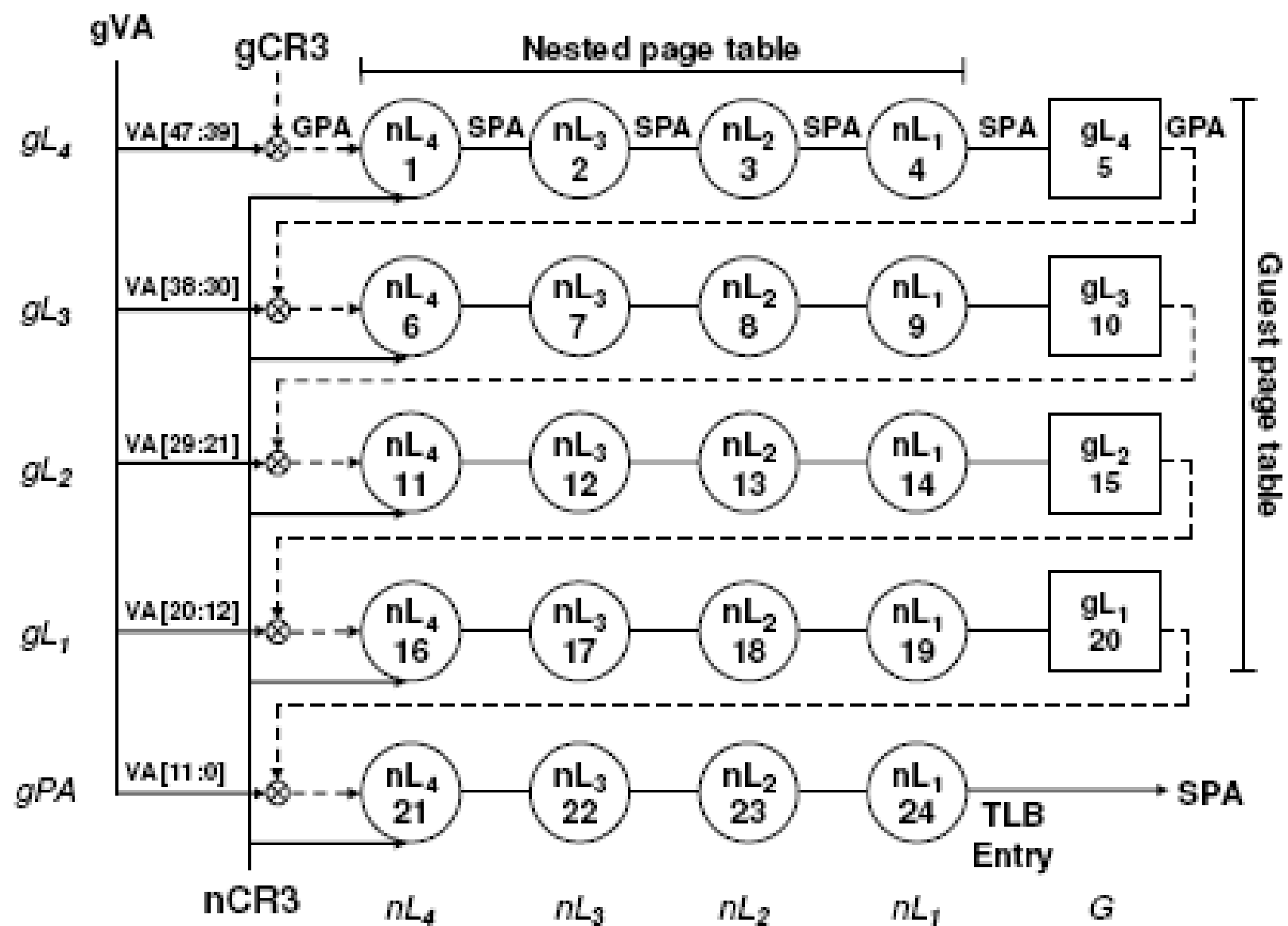
AMD Nested Page Table

Koszt translacji:

Nie trafienie szukanego wpisu w TLB jest bardziej kosztowne niż bez nested page'ingu.

Dodatkowe wsparcie sprzętowe w postaci Nested TLB do przechowywania tłumaczeń adresów fizycznych gościa na adresy fizyczne hosta.





Wirtualizacja wspomagana sprzętowo

AMD Nested Page Table

Podsumowanie:

- NPT jest funkcjonalnością tylko dla VMM, gość nie ma świadomości, że NPT działa.
- W szczególności nie trzeba modyfikować SO gościa
- NPT jest opcją w niektórych procesorach AMD-V
- CPUID stwierdza czy jest ona aktywna
- Jedna instancja nPT dla jednego gościa, a nie wiele sPT (dla każdej gPT)
- To procentuje także przy SMP (synchronizacja)
- Duże rozmiary stron przyspieszają przeszukiwanie

Wirtualizacja wspomagana sprzętowo

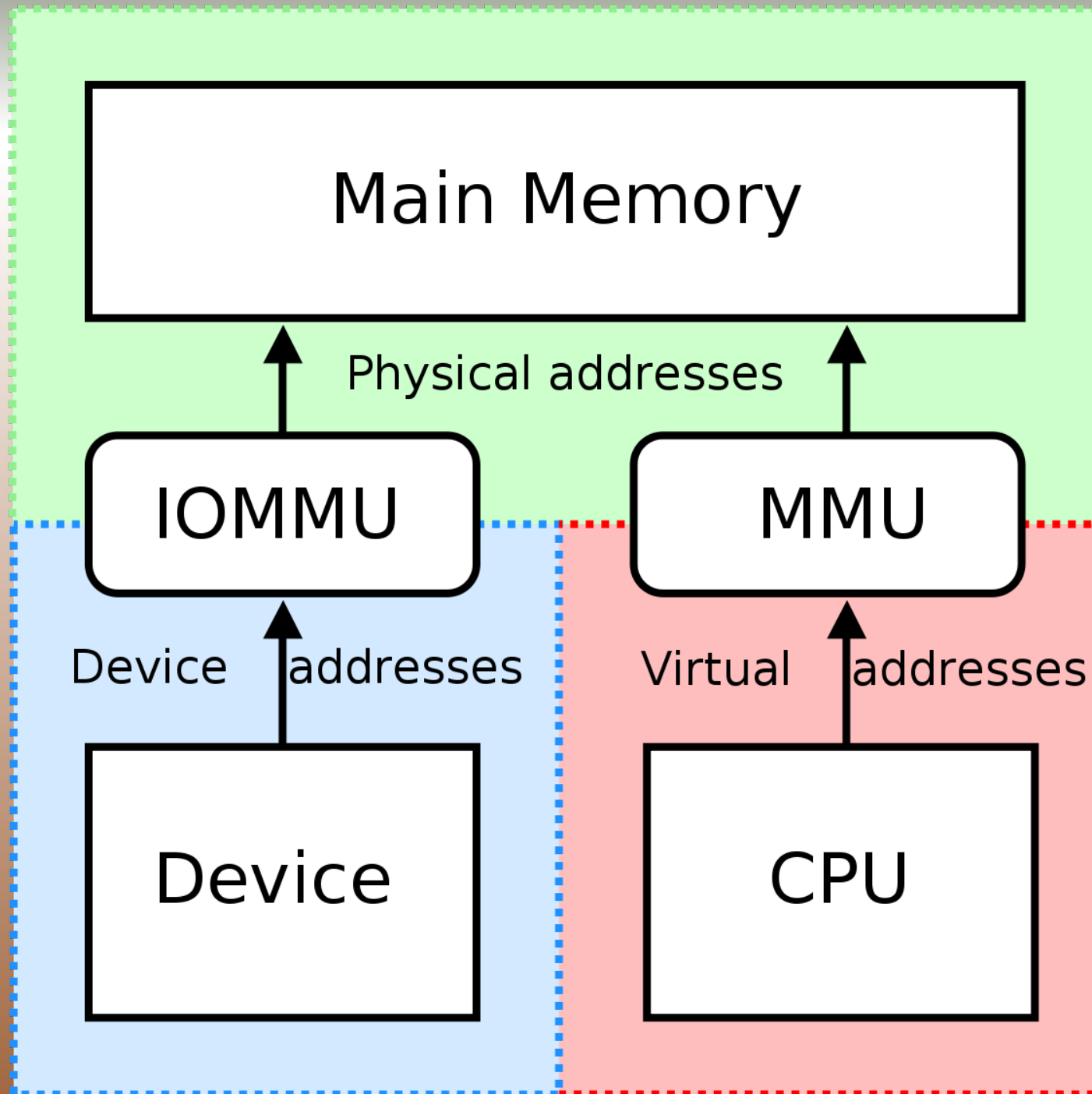
IOMMU - GART

Przykładem IOMMU jest Graphics Address Remapping Table (GART), używany przez AGP i PCI Express, które to ładują do pamięci kształty, siatki wielościanów, tekstury i inne dane za pomocą DMA.

Wirtualizacja wspomagana sprzętowo

IOMMU

- Input/output memory management unit
- Podobnie jak MMU, które tłumaczy adresy widziane przez procesor na adresy fizyczne, IOMMU tłumaczy adresy wirtualne widziane przez urządzenia (w tym kontekście urządzenia we/wy) na adresy fizyczne.
- IOMMU daje także pewne możliwości ochrony przed niebezpiecznymi zachowaniami urządzeń.



Wirtualizacja wspomagana sprzętowo

Intel VT-d

„Zwiększa stabilność, elastyczność oraz wydajność I/O w środowisku zwirtualizowanym, poprzez umożliwienie bezpośredniego podłączania urządzeń I/O do maszyn wirtualnych.”

Uproszczając Intel VT-d (Intel Virtualization Technology for Directed I/O) to:

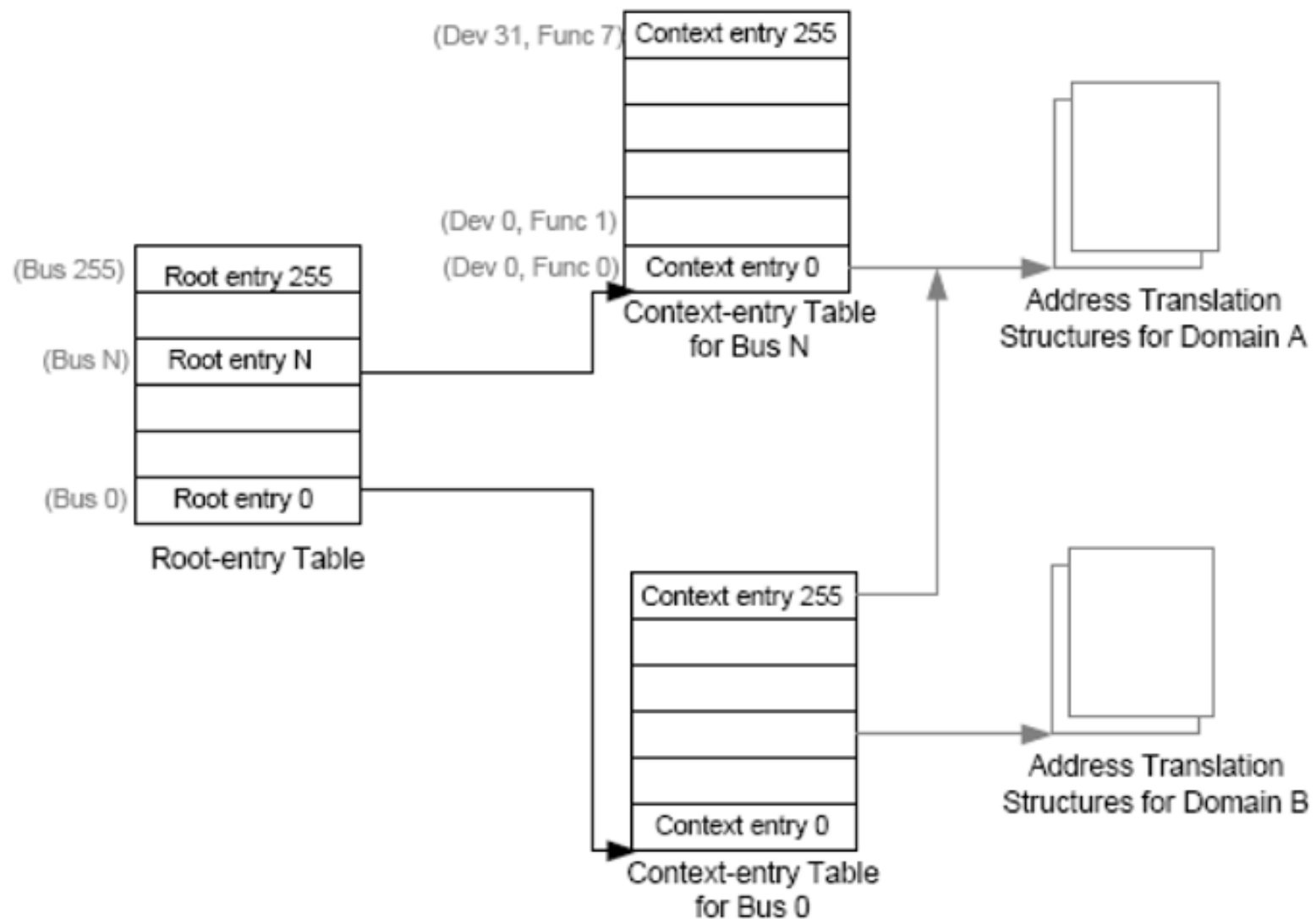
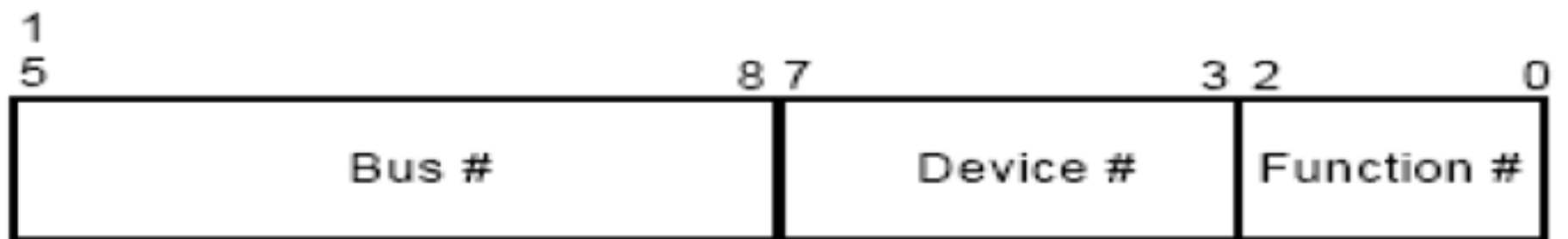
- Uogólnienie architektury IOMMU
- „Protection domains”

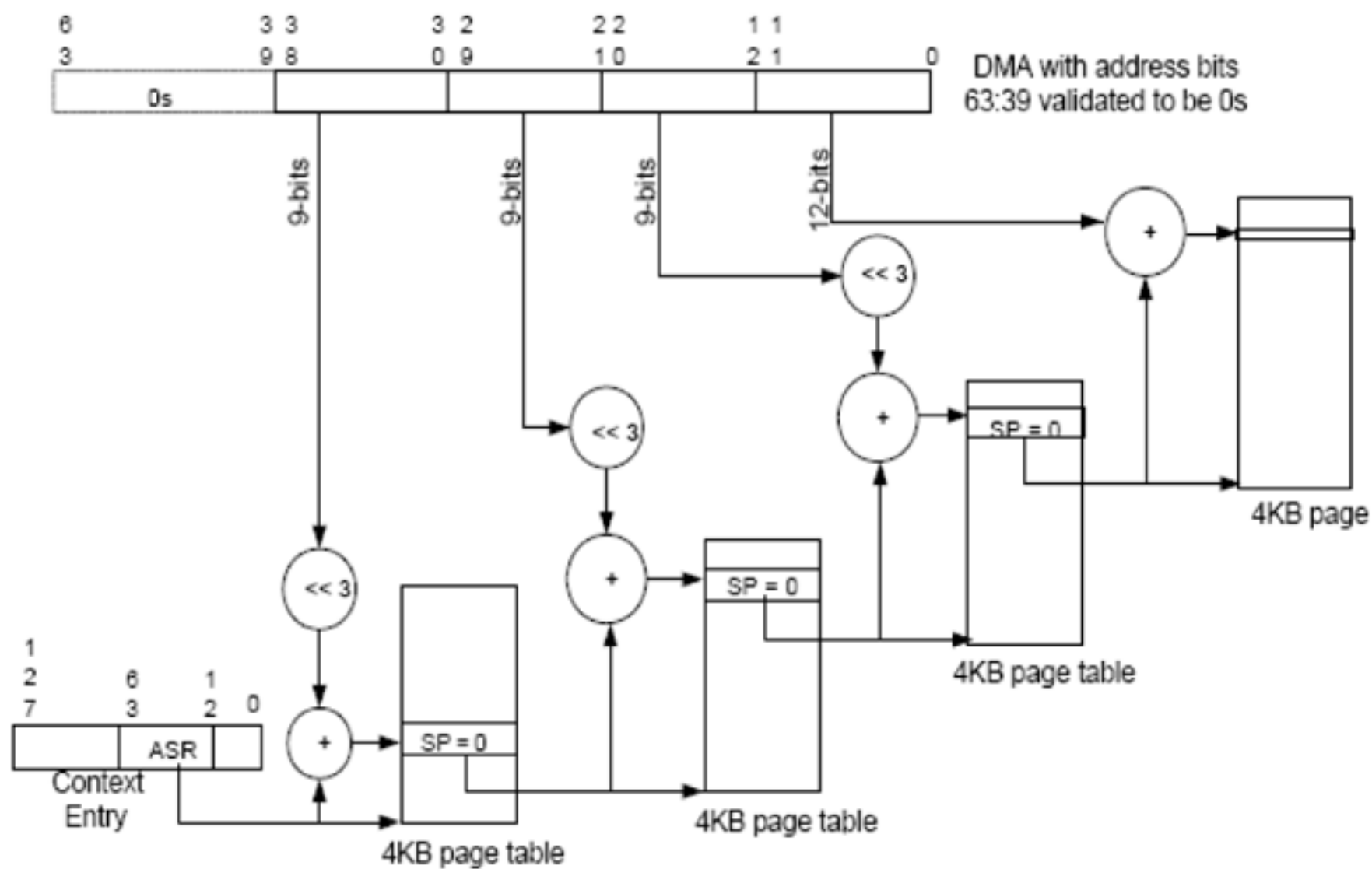
Wirtualizacja wspomagana sprzętowo

Intel VT-d

Rzutowanie urządzeń na chronione dziedziny:

- ID zlecenia dla DMA: szyna PCI / urządzenie / numer funkcji przypisany przez oprogramowanie konfiguracyjne PCI
- Architektura VT-d definiuje struktury danych, które są do tego potrzebne (Root-entry Table i Context-entry Table)
- Wielopoziomowe struktury tablic stron do tłumaczenia adresów DMA (podobnie jak w procesorze IA-32)





Wirtualizacja wspomagana sprzętowo

Intel VT-d

- Rzutowanie żądań DMA.
- Rzutowanie przerwań.

Pozwalają zredukować odpowiednio wystąpienia VMexit lub koszty obsługi przerwań, dla urządzeń, które są przypisane do jakiejś dziedziny.

Używa do tego specjalnych struktur danych i cache'uje ostatnio przetłumaczone żądania (przerwania).

Wirtualizacja wspomagana sprzętowo

Intel VT-d

Pamięci podręczne przyspieszające rzutowania żądań DMA i przerwań:

- Context Cache: wpisy tłumaczące urządzenia na chronione dziedziny
- PDE (Page Directory Entry) Cache: wpisy w katalogu stron
- IOTLB (I/O Translation Look-aside Buffer): wyniki translacji
- Interrupt Entry Cache: wpisy w strukturze interrupt-remapping table

Wirtualizacja wspomagana sprzętowo

Zastosowania IOMMU poza wirtualizacją

- Duże obszary pamięci mogą zostać zaalokowane, bez potrzeby ciągłości w pamięci fizycznej (mapowanie ciągłych obszarów pamięci wirtualnej na te nieciągłe fizycznej).
- Urządzenia, które nie wspierają odpowiednio długiego adresowania, są wyręczane przez IOMMU.
- W pewnych architekturach IOMMU zajmuje się także tłumaczeniem przerwań sprzętowych.
- IOMMU może również wspierać stronicowanie pamięci peryferyjnej.

Wirtualizacja wspomagana sprzętowo

Xen

Stworzony na uniwersytecie w Cambridge
pierwotnie część XenoSystems
open source
obecnie rozwijany zarówno przez Intela jak i AMD

Wirtualizacja wspomagana sprzętowo

Cechy Xen-a

1. Własne mikrojądro
2. Potrzebuje systemu bazowego do zarządzania I/O
3. Pierwotnie parawirtualizator
4. Stabilny
5. Może wykorzystywać wsparcie sprzętowe, ale nie musi

Wirtualizacja wspomagana sprzętowo

Xen, cd.

Systemy rezydują w tzw. domenach
w domenie 0 jest system zarządzający

System zarządzający: FreeBSD, NetBSD, Linux

Systemy gości: Windows XP, Plan 9, Linux, Solaris

Wirtualizacja wspomagana sprzętowo

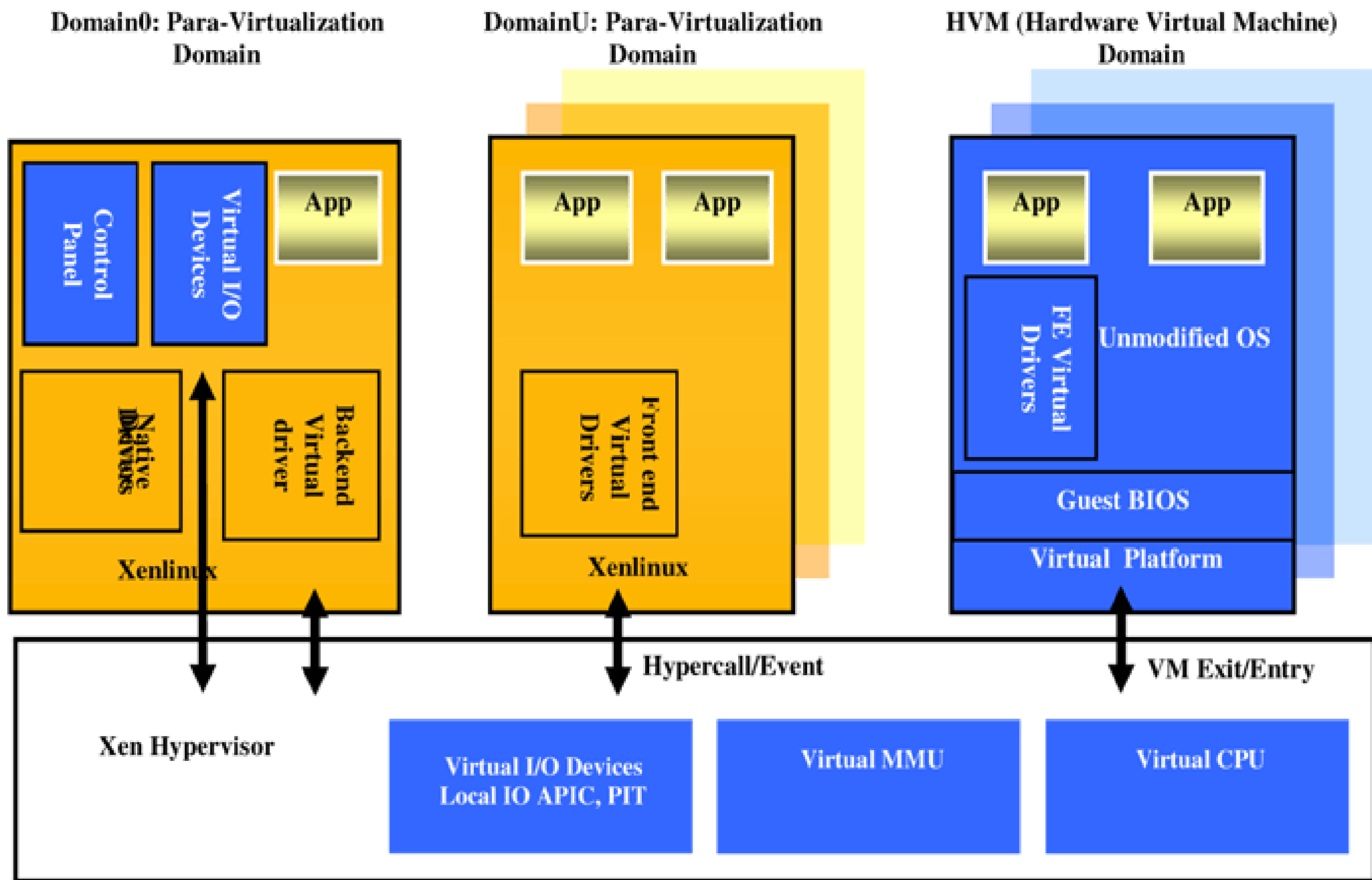
Jak to działa?

Mikrojądro Xen-a zajmuje się prawie wszystkim

System zarządzający jest potrzebny do:

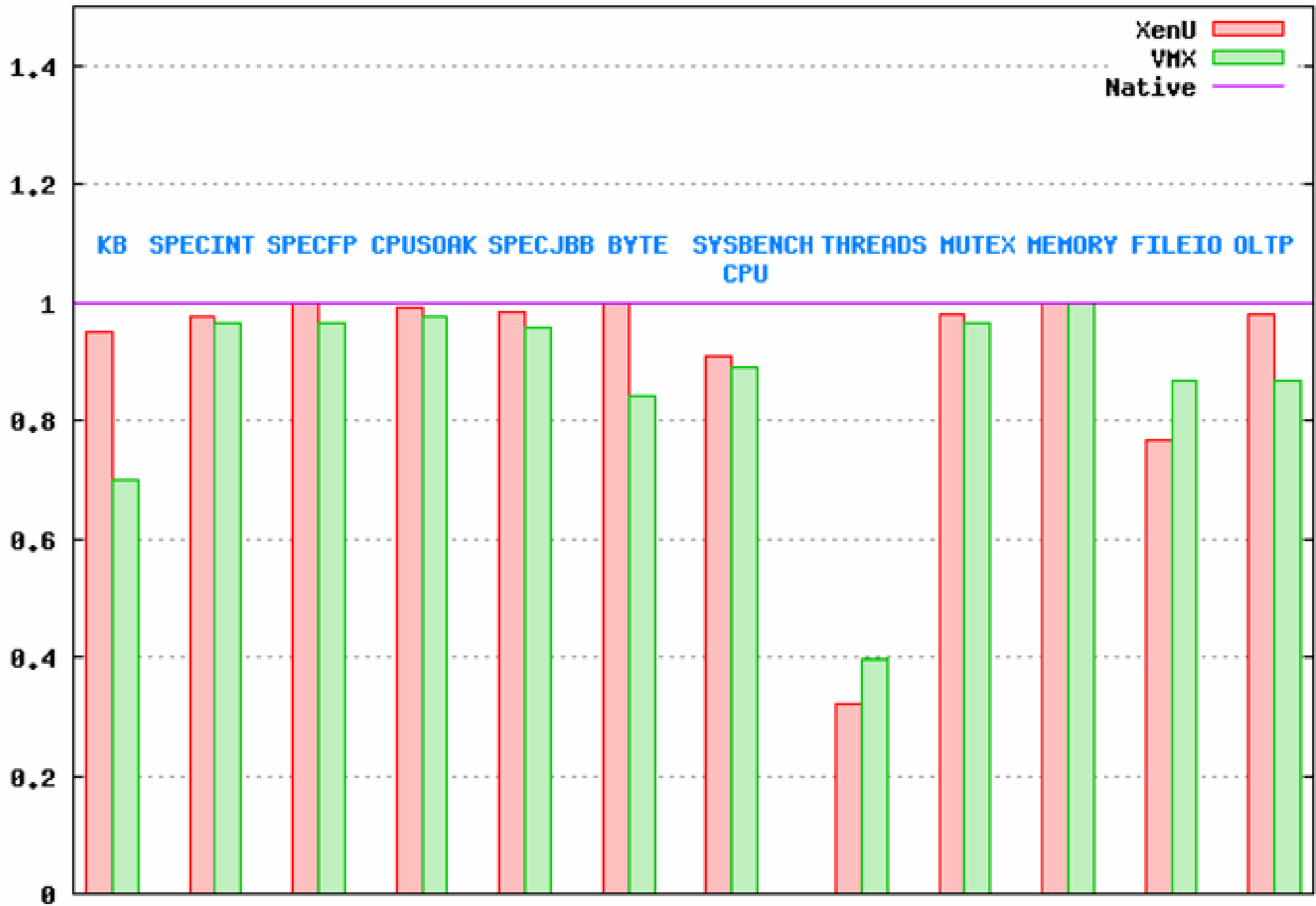
1. Obsługi urządzeń I/O
2. Zarządzania domenami

QEMU emuluje w systemach gości urządzenia I/O



Platform with Hardware-Based Virtualization (e.g. Intel® Virtualization Technology on IA-32, EM64T, IPF, aka IA-64)

Performance Testing Summary --IA32 Guest/IA32 HV
Test Machine: Intel S3E2340, Dual-Core Xeon Processor
Change Set:10031



Wirtualizacja wspomagana sprzętowo

KVM

system pełnej wirtualizacji dla x86
moduł kernela linuxa

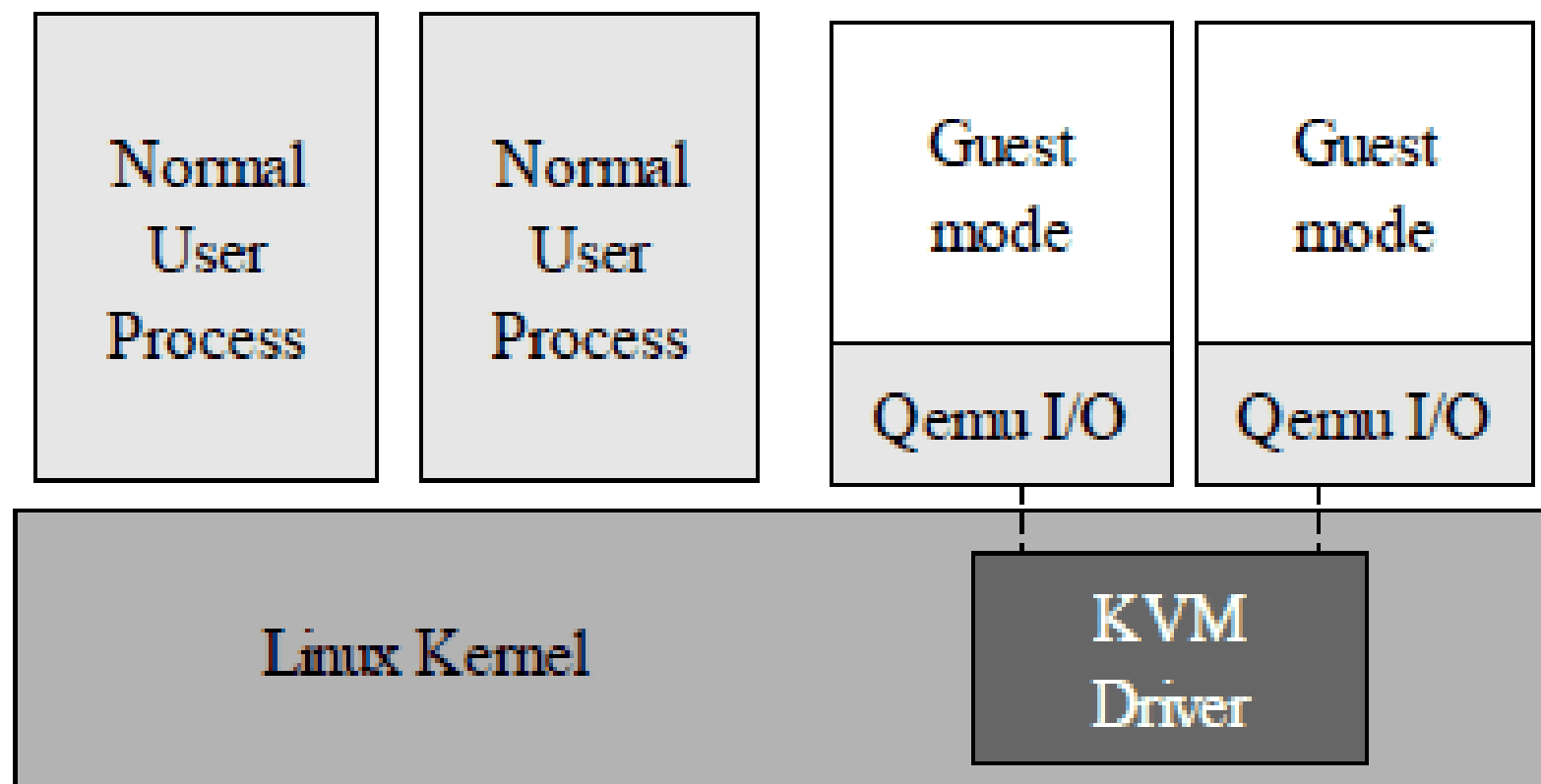
Wirtualizacja wspomagana sprzętowo

KVM, właściwości

Potrzebuje wsparcia sprzętowego
W jądrze linuxa od 2.6.20
samo KVM nie zapewnia wirtualizacji
potrzeba user-space emulatora

Wirtualizacja wspomagana sprzętowo

Jak to wygląda?



Wirtualizacja wspomagana sprzętowo

KVM, właściwości

Potrzebuje wsparcia sprzętowego
W jądrze linuxa od 2.6.20
samo KVM nie zapewnia wirtualizacji
potrzeba user-space emulatora

Wirtualizacja wspomagana sprzętowo

KVM, cd.

QEMU jako user-space emulator
zarządzanie zasobami(pamięcią, CPU)
jest wykonywane przez jądro Linuxa – gospodarza

Wirtualizacja wspomagana sprzętowo

Jacy goście są wpuszczani?

Windows od wersji 2000

Fedora 9, Ubuntu 8.04, Debian Etch 4
RedHat, Slackware, Gentoo, Suse...

FreeBSD, Solaris, Haiku,
AROS, ReactOS, FreeDos

Wirtualizacja wspomagana sprzętowo

Ale są też problemy z:

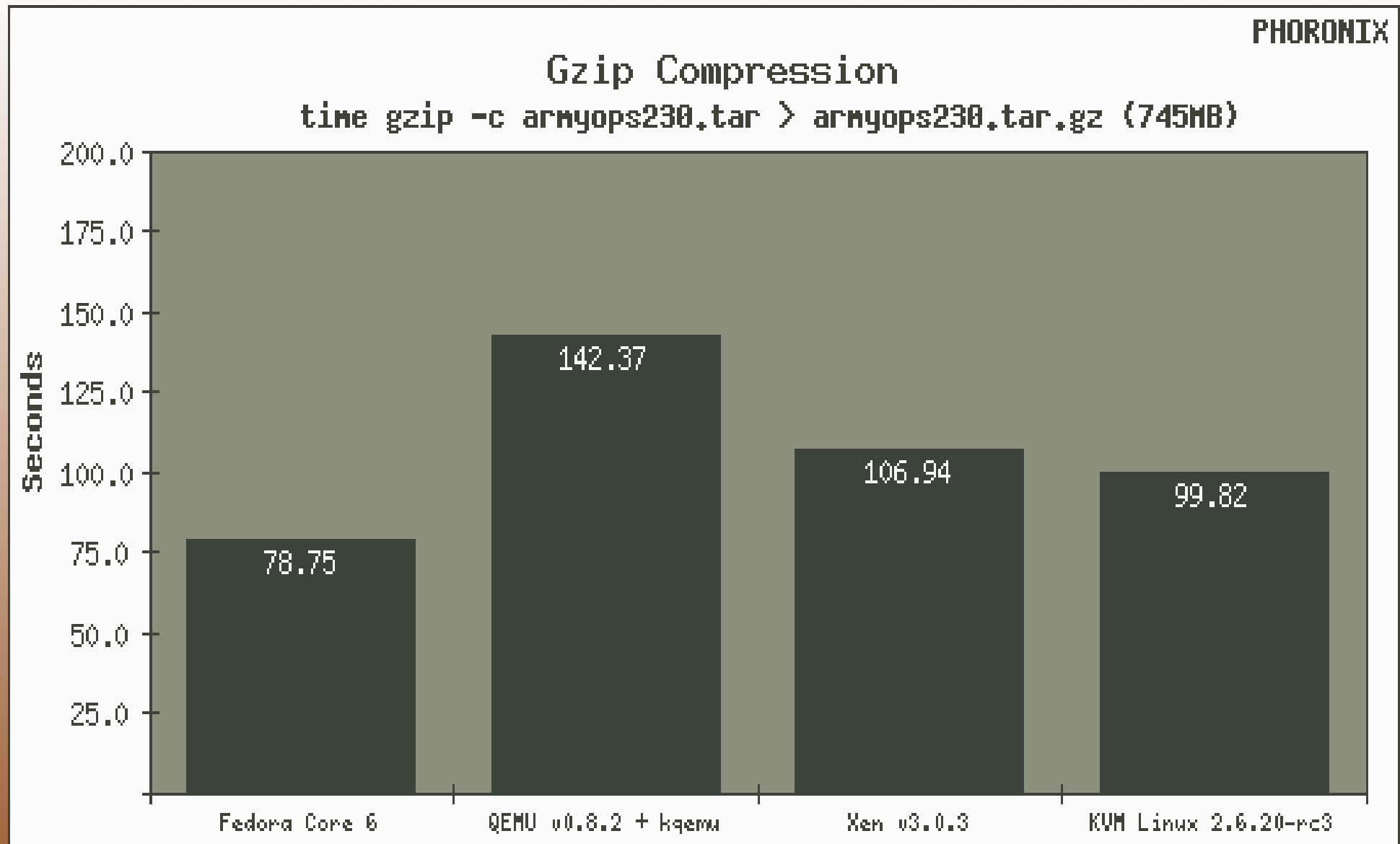
- Mandriva
- Ubuntu
- OpenSuse
- RedHat

Wirtualizacja wspomagana sprzętowo

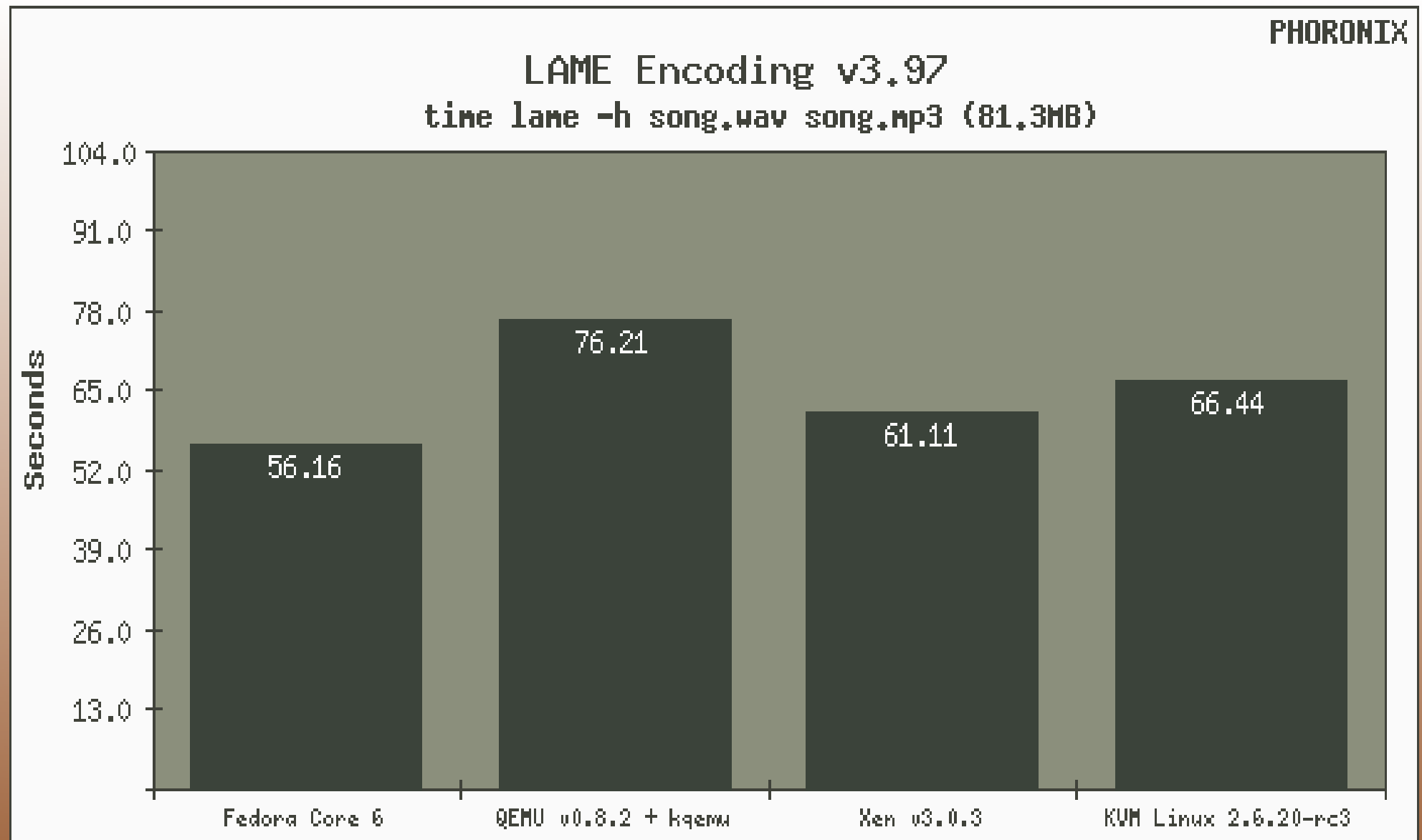
Porównanie:

KVM vs. Xen

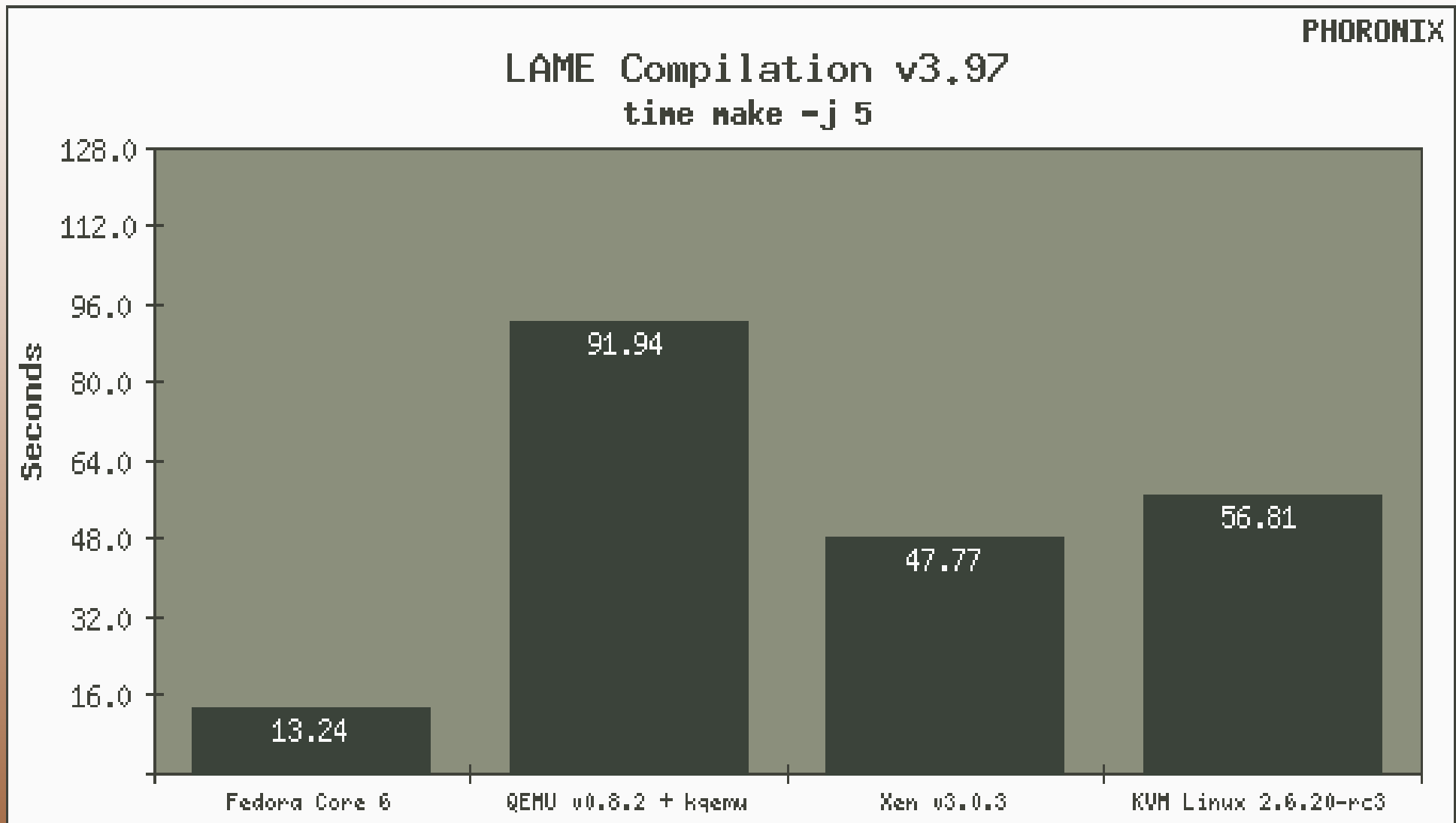
Wirtualizacja wspomagana sprzętowo



Wirtualizacja wspomagana sprzętowo



Wirtualizacja wspomagana sprzętowo



Wirtualizacja wspomagana sprzętowo

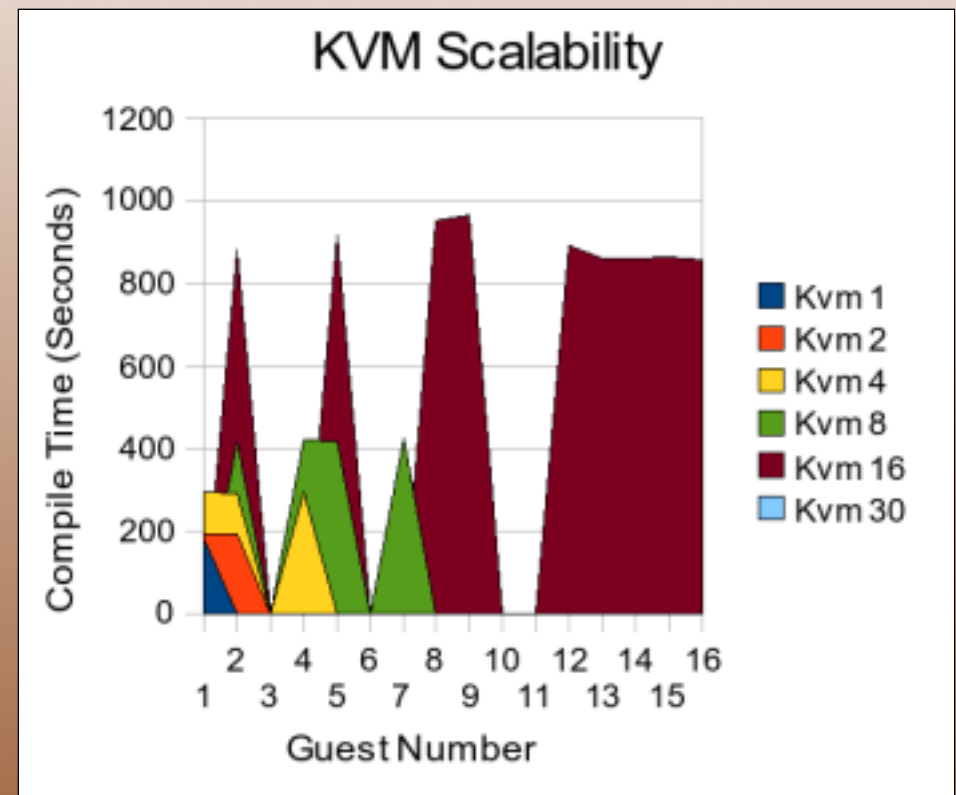
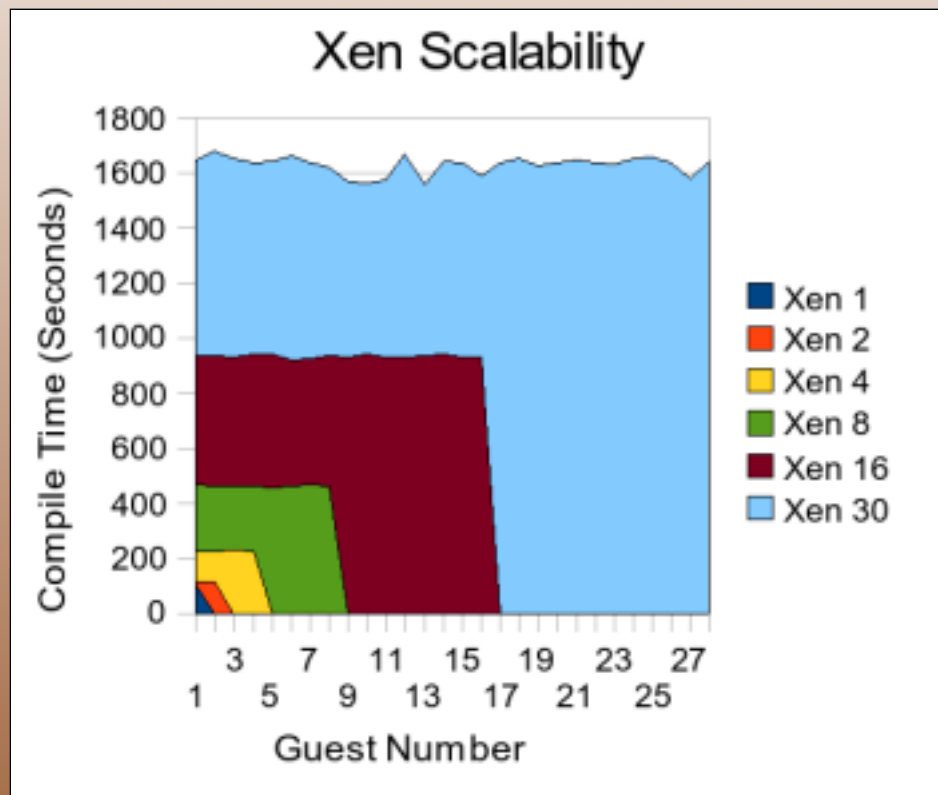
A teraz wyniki z Xen Summit z 2008 roku
2,4 GHz Intel Core 2, 4 GB RAM, Ubuntu 8.04

	Linux	Xen	KVM
CPU	1.000	0.999	0.993
Kernel Compile	1.000	0.487	0.384
IOzone Write	1.000	0.855	0.934
IOzone Read	1.000	0.852	0.994

Wirtualizacja wspomagana sprzętowo

Skalowalność

kompilacja Apache'a na 1, 2, 4, 8, 16 i 30 gościach



Wirtualizacja wspomagana sprzętowo

Zagrożenia

Do tej pory wszystko wyglądało różowo..

Jednak pełna wirtualizacja niesie pewne zagrożenia

Wirtualizacja wspomagana sprzętowo

BLUE PILL

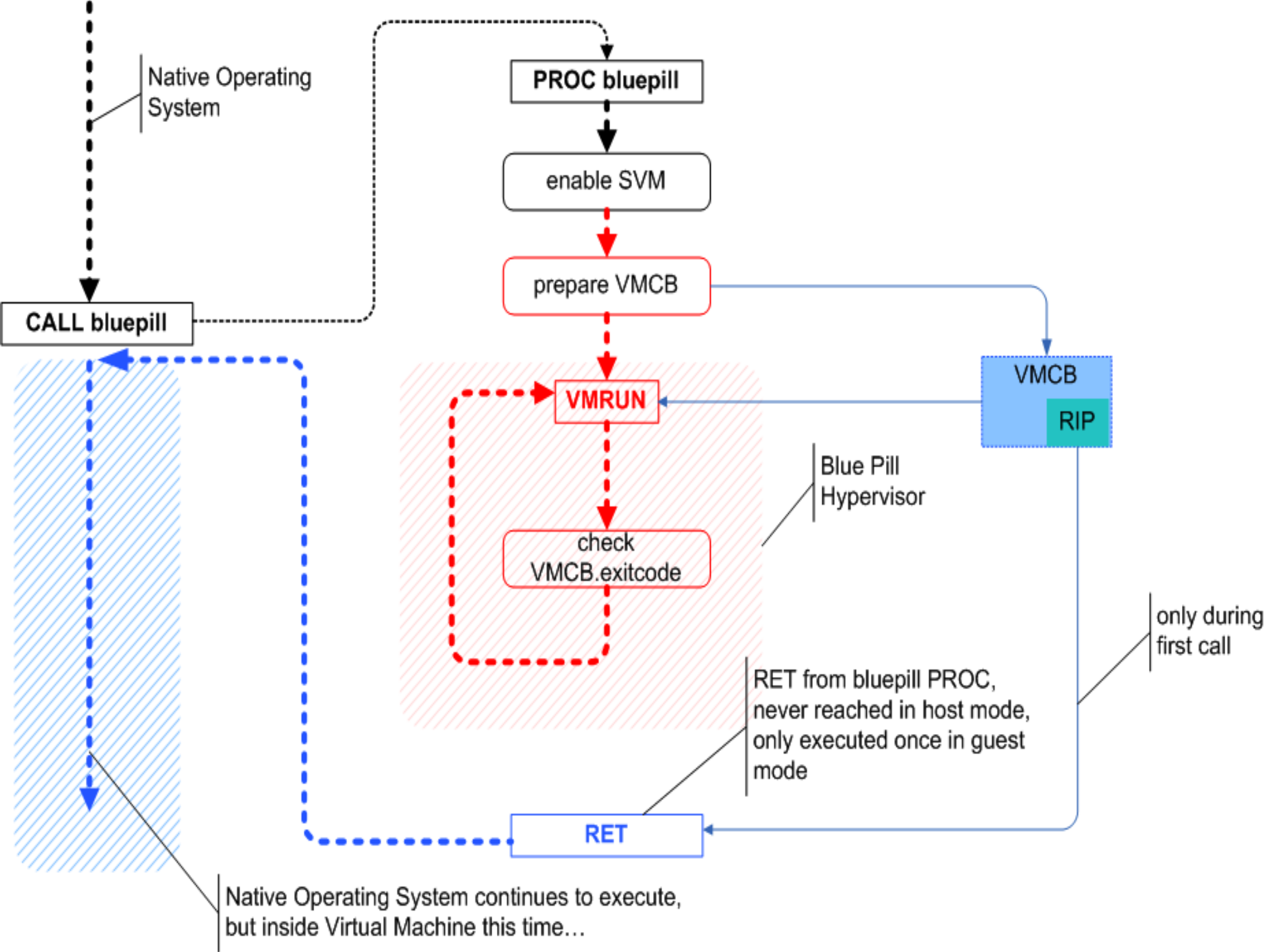
stworzony przez Joanne Rutkowską
działa na AMD Pacifica

Wirtualizacja wspomagana sprzętowo

Jak działa Blue Pill?

Instalacja „on the fly” pod Vistą,
umieszcza system operacyjny w VMM

Blue Pill pierwszej generacji jest niedoskonały



Wirtualizacja wspomagana sprzętowo

Jeśli wirtualizacja rzeczywiście jest pełna, to taki rootkit jest
niewykrywalny!

Wirtualizacja wspomagana sprzętowo

Czy można temu zapobiec?

Postulowane rozwiązania:

1. Wyłączenie wirtualizacji sprzętowej
2. „Timing attacks”
3. Uniemożliwienie instalacji „on the fly”

Wirtualizacja wspomagana sprzętowo

Inne tego typu rootkity:

Virtioli (dla Vanderpoola)

Sysvirt

Wirtualizacja wspomagana sprzętowo

Zalety:

- Uruchamianie kilku systemów na jednej maszynie
- System operacyjny nie będzie uzależniony od hardware'u

Wirtualizacja wspomagana sprzętowo

Wady:

- System wirtualizowany zawsze jest wolniejszy
- Możliwość ataku type Blue Pill