Address Translation for Virtual Machines

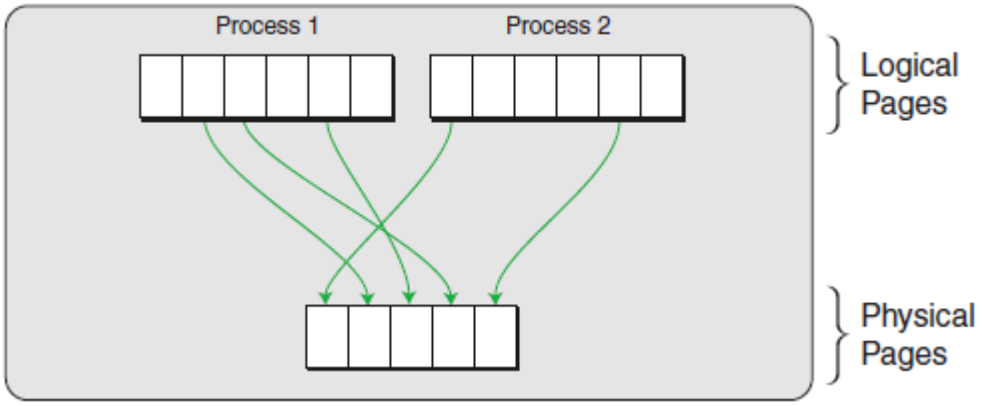
CSC456

Outline

- Address Translation for VM
  - Software Approach
    - Shadow Paging
    - Pro & Cons
  - Architectural Support
    - 2D Page Walk
    - Pro & Cons
    - Improvements

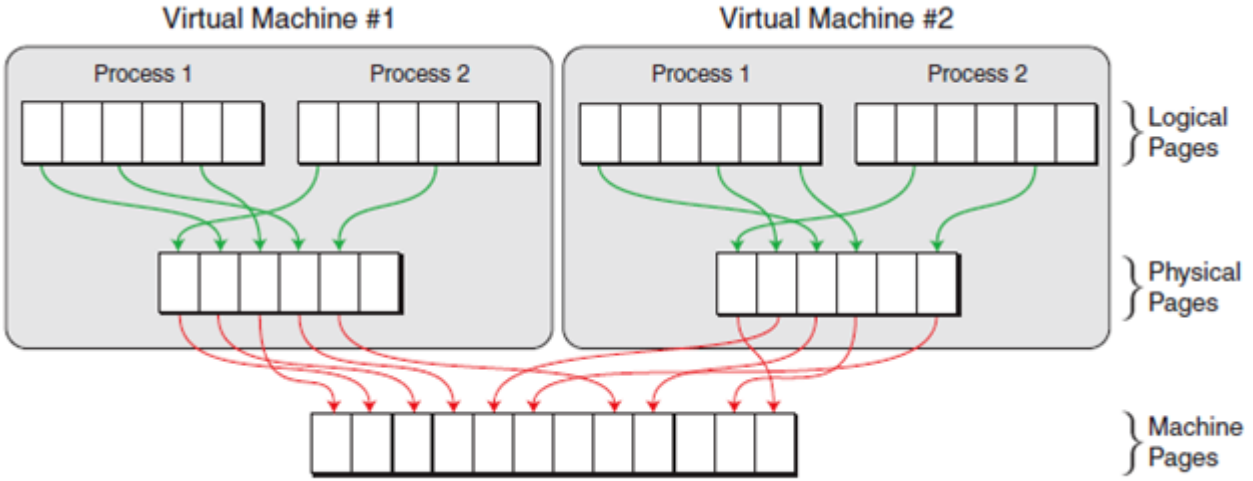
Address Translation

- Isolated process address space
- Provide processes with illusion of a large address space
- Architectural support for segmentation and paging



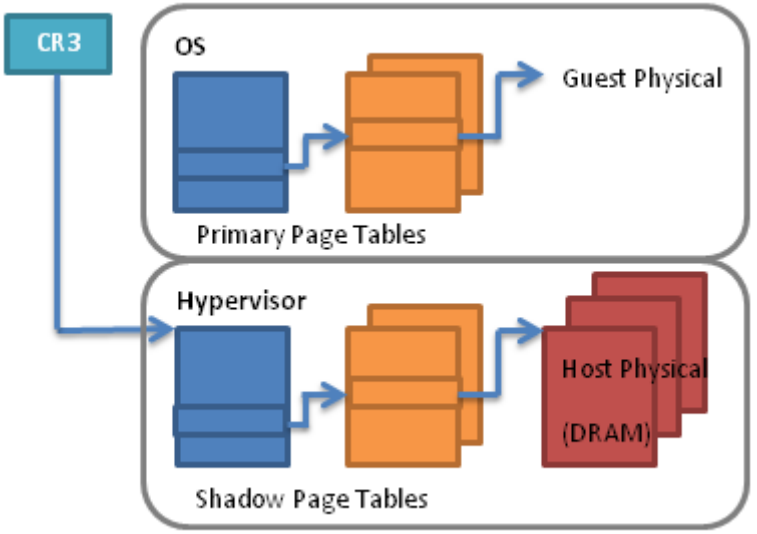
Address Translation for VMs

- Isolated Guest-Physical address space
  - A new layer of address translation
- Guest virtual address (gVA) to guest physical address (gPA) to system physical address (sPA)
  - But, there was no architectural support



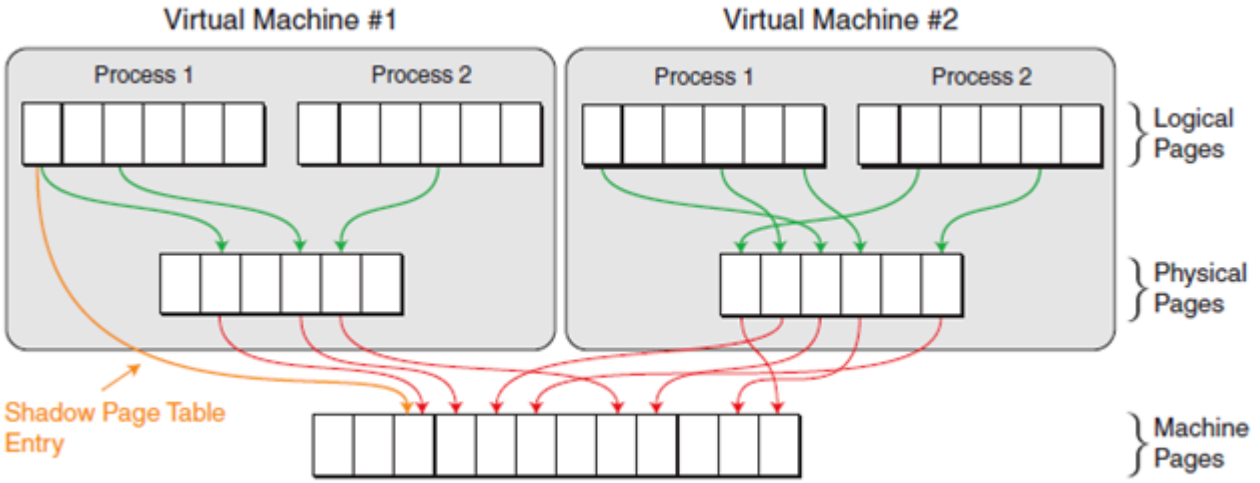
Shadow Paging

- Hypervisor must maintain a separate mapping from gVA to sPA
  - Intercept every attempt by guest to update or install a page table (performance)
  - Per process mapping (space)



Shadow Paging

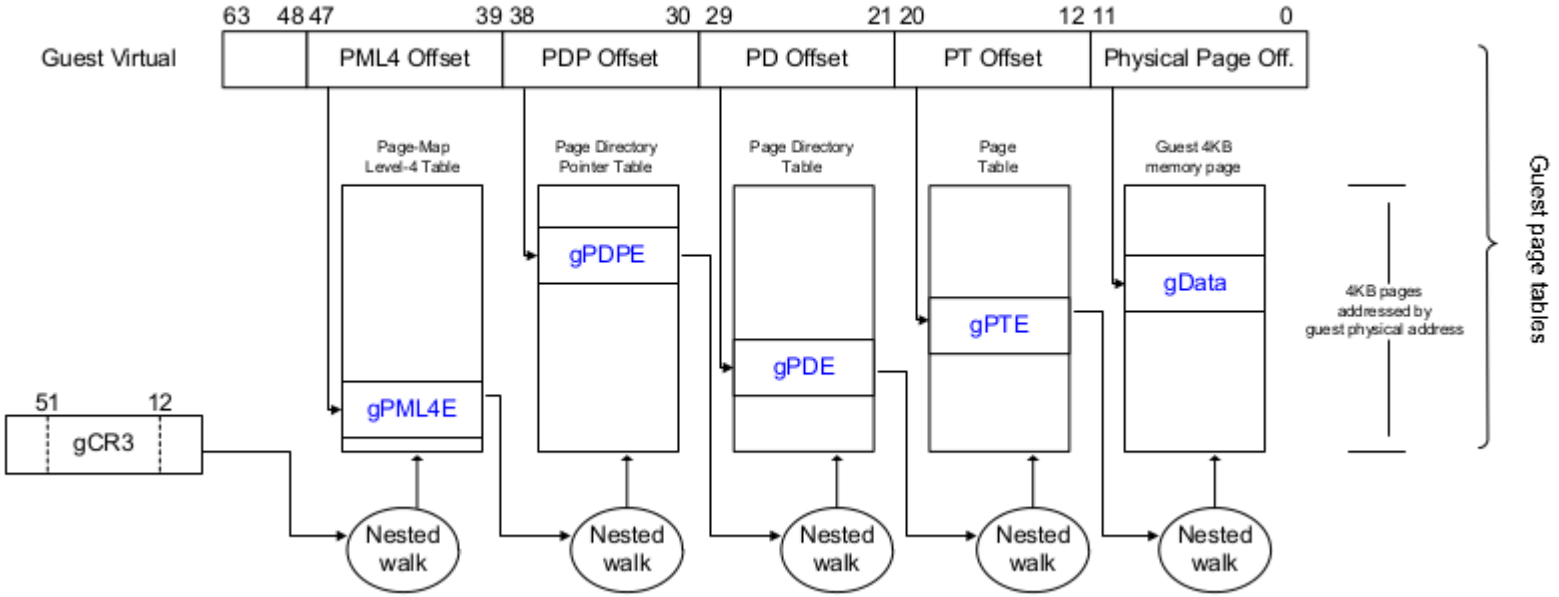
- Techniques:
  - Write-Protecting gPT
  - Virtual TLB
- Both incur lots of page faults



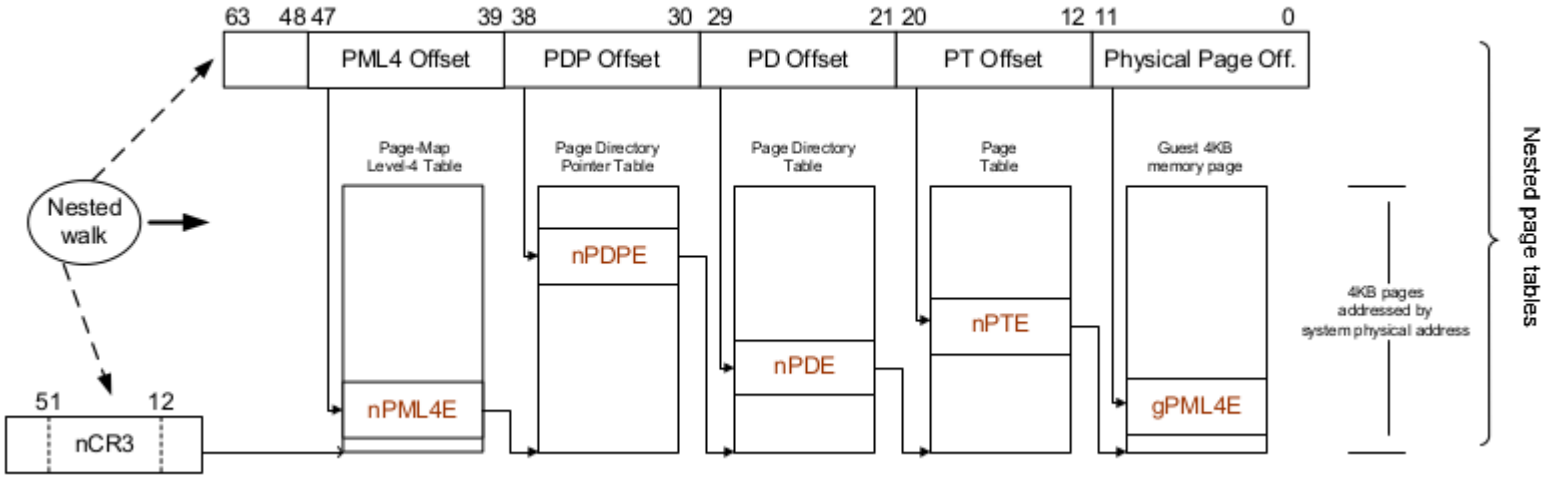
Hardware assisted virtualization

- Eliminate the need for shadow paging
- Provide architectural support for a new layer of address translation (also called Nested Level)
- A new hierarchy of paging which translates gPA to sPA
- Two dimensional Page Table
  - Nested Page Table (NPT) by AMD
  - Extended Page Table (EPT) by Intel
  - gCR3 and nCR3

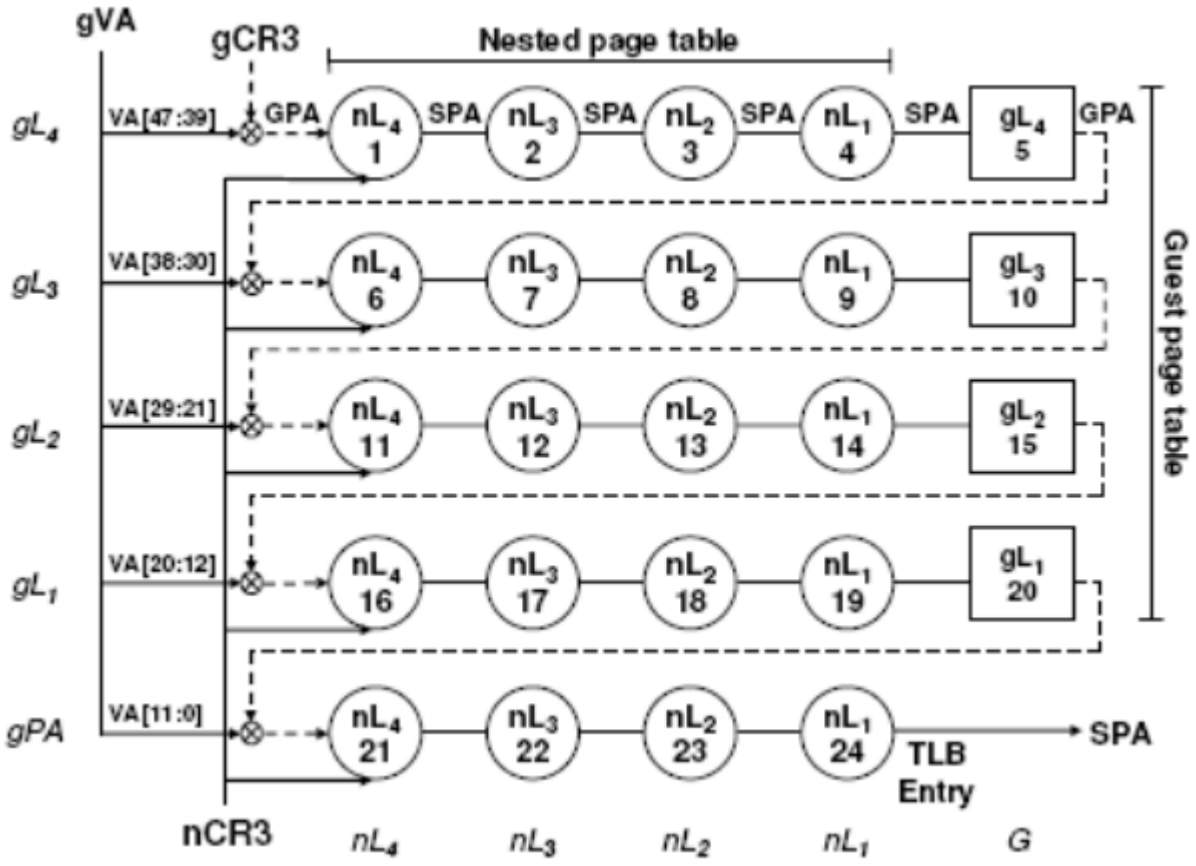
Two-Dimensional Page Walk
Guest Page Walk (1D)



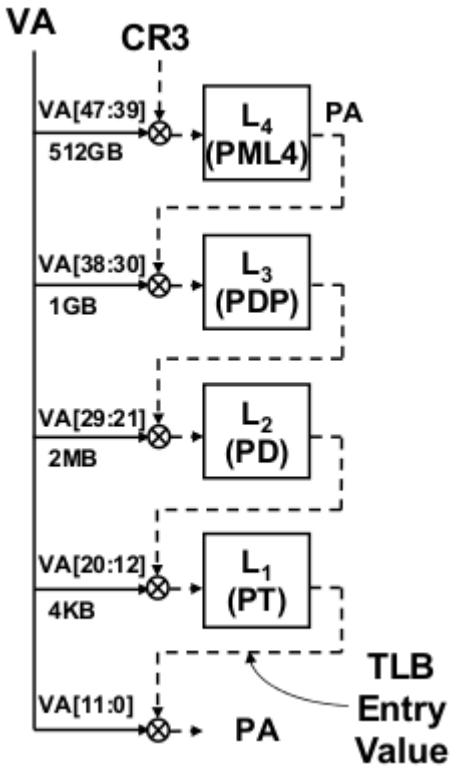
Two-Dimensional Page Walk
Nested Page Walk (2D)

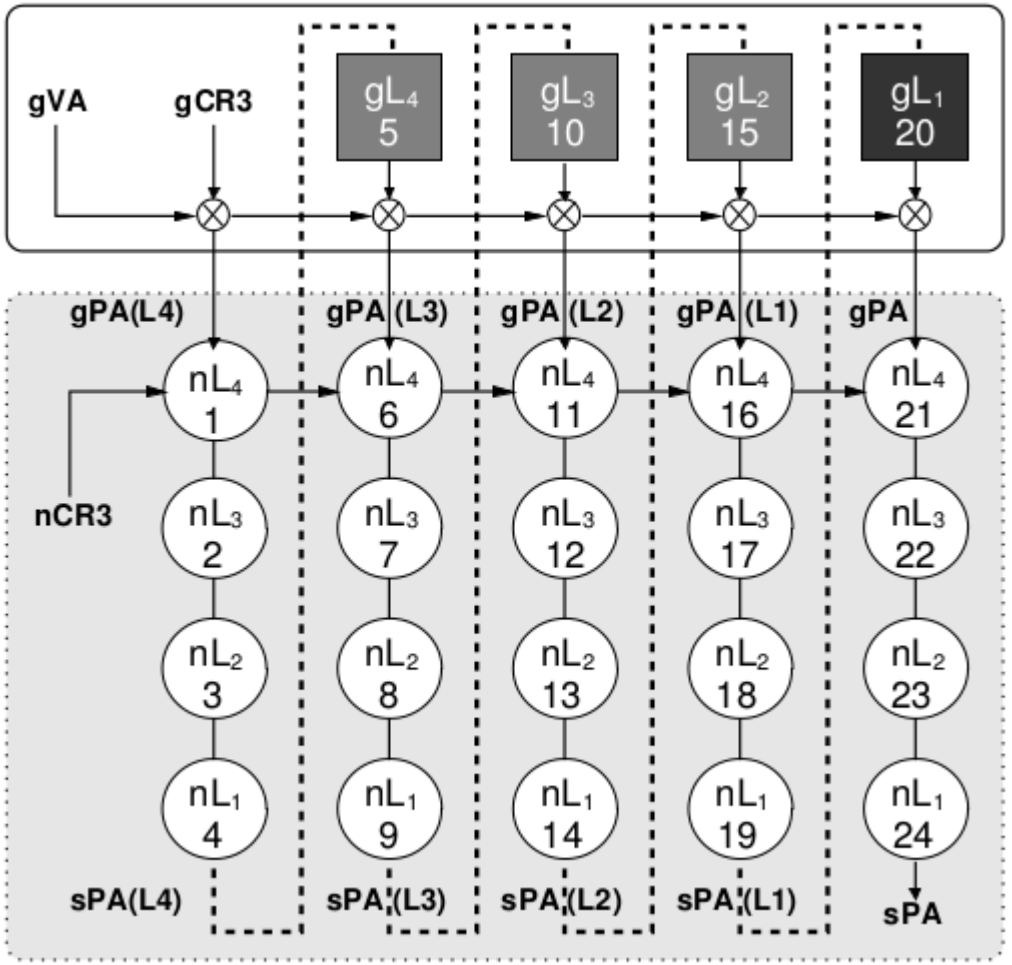


Two-Dimensional Page Walk Combined



Two Dimensional Page Walk Combined

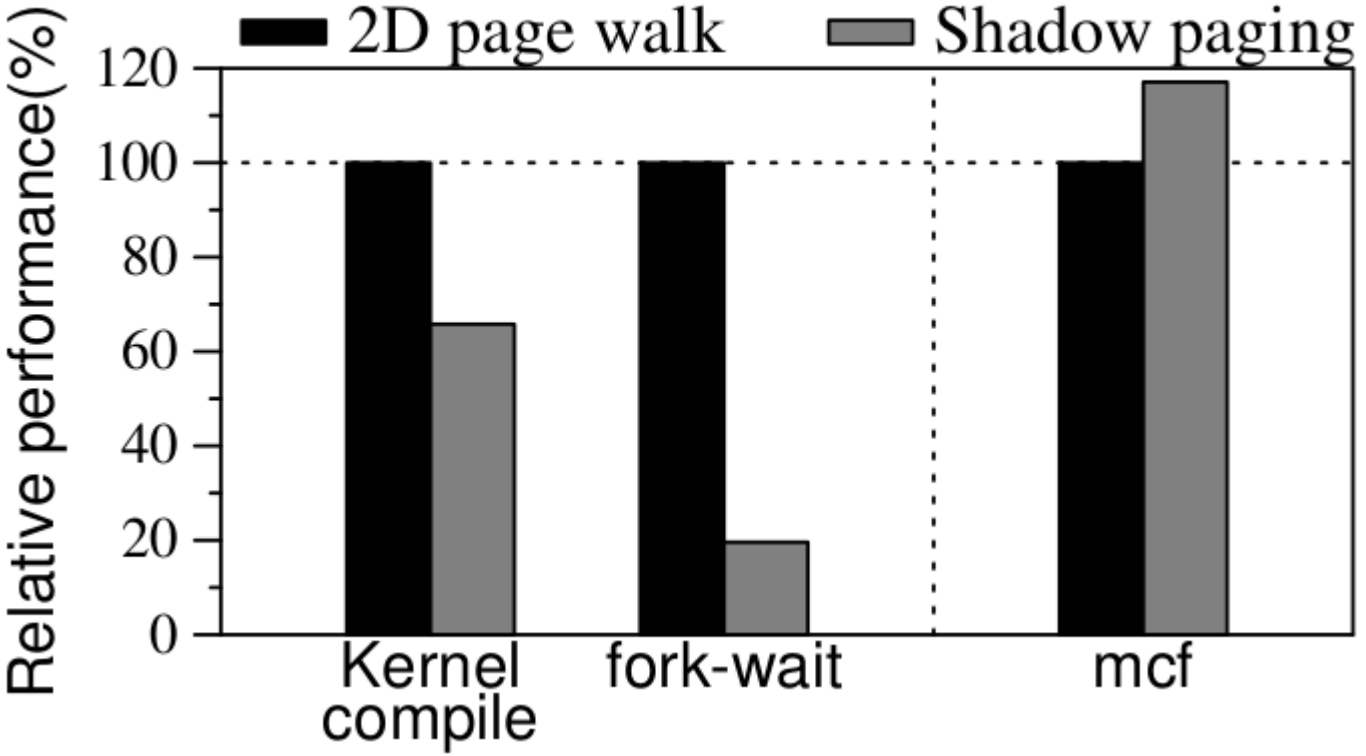




Two Dimensional Page Walk

- There is no need to keep a shadow per process, just a nPT per VM
- But, upon TLB miss, 4 vs. 24 memory reference
- mn+n+m memory reference per TLB miss
- Better for applications with frequent updates on guest page table (`fork`, `make`, …)
- But not for applications with lots of TLB misses

Comparison of 2D page walk and Shadow Paging

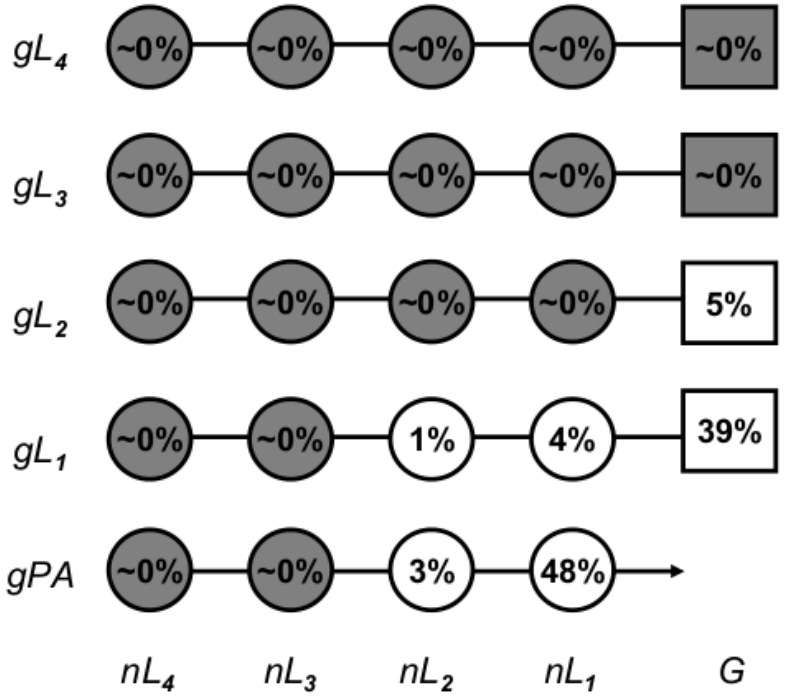


Large Pages

- Can reduce the paging level
  - If we can allot 2MB instead of 4KB in hypervisor, there will be 19 ref (%21 improvement)
  - The same holds for OS super-pages
- Can reduce TBL pressure
  - Not if guest PT use a large page which is mapped to smaller pages by nested PT
- Not always easy to find contiguous chunks

Page Walk Cache

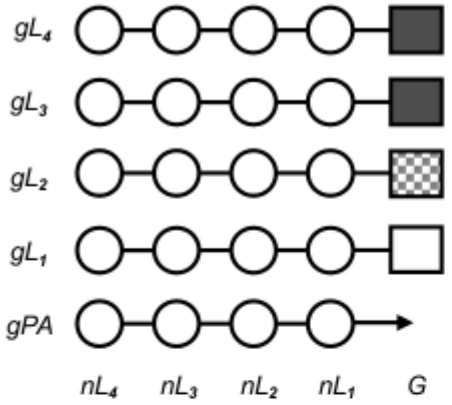
- Translation Caching: Skip, Don't Walk [Barr et. al. ISCA'10]
- Upper level intermediate translation exhibit high temporal localities
- Extra hardware table to cache intermediate translation



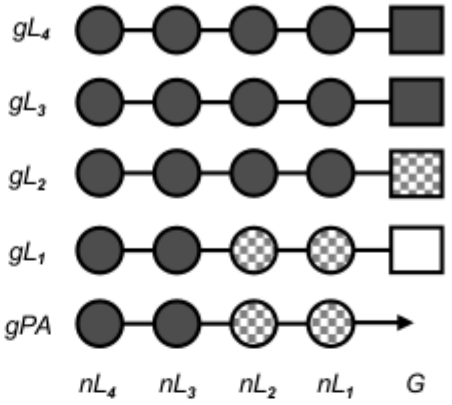
NTLB

- Guest Physical (gPA) to System Physical (sPA) cache
- Isolates the nested page table translation
- Design decisions (not to cache everything)

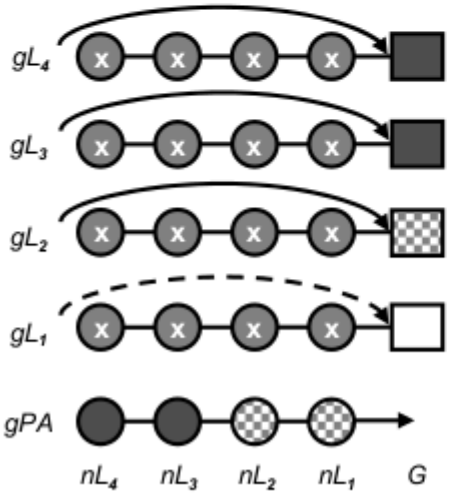
2D PW Cache Design



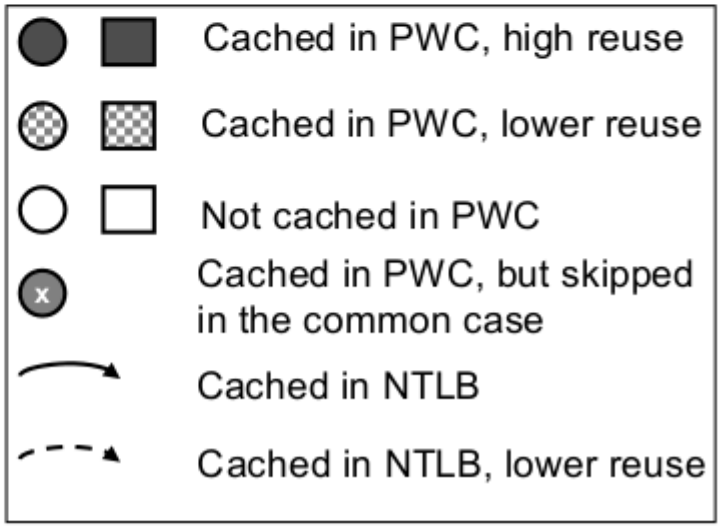
(a): 1D_PWC



(b): 2D_PWC



(c): 2D_PWC+NT



Summary

- Address Translation for VM
  - Software Techniques
    - Shadow Paging
    - Pros and Cons
  - Architectural Support
    - 2D Page Walk
    - Pros and Cons
      - Comparison with Shadow Pages
      - Improvements
        - Large Pages
        - Page Walk Cache
        - Nested TLB

References

- AMD-V Nested Paging, July 2008
- R. Bhargava et. al., “Accelerating Two-Dimentional Page Walks for Virtualized Systems” , ASPLOS'08
- J. Ahn et. al., “Revisiting Hardware-Assisted Page Walks for Virtualized Systems”, ISCA'12