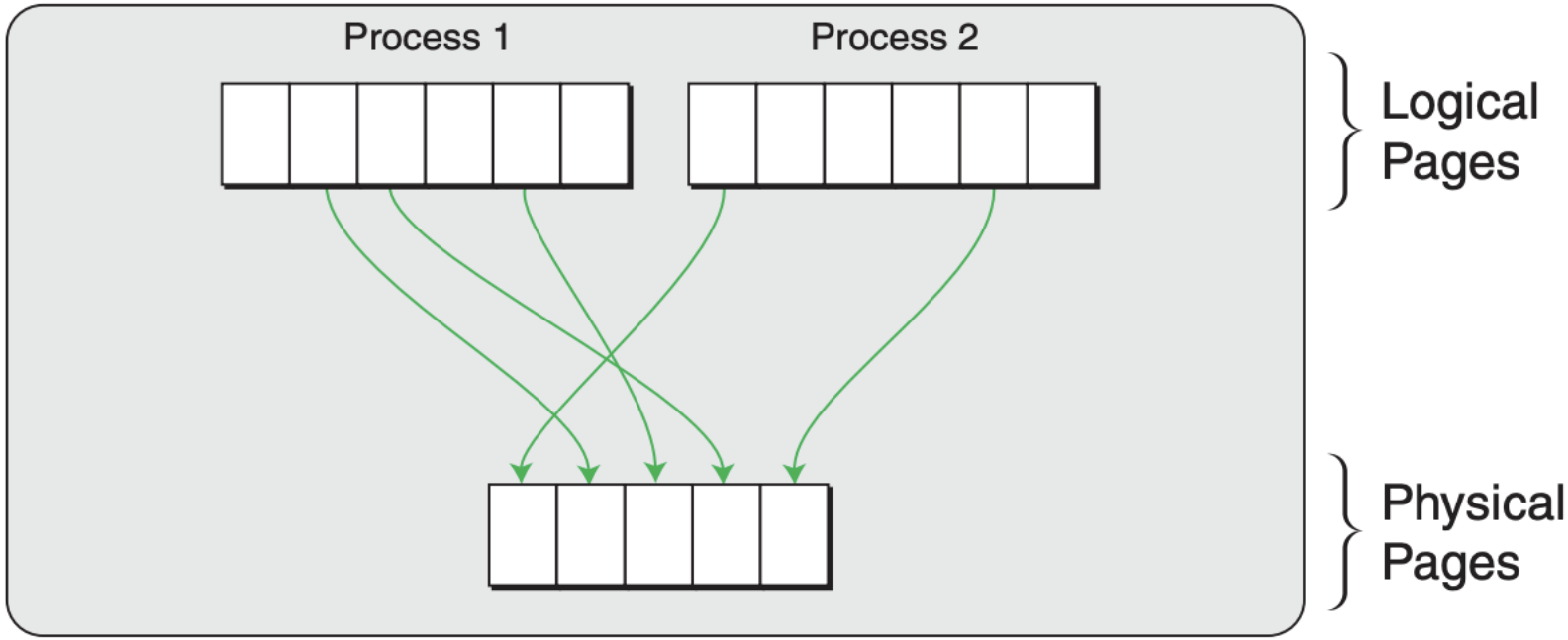
Virtualization II: MMU Virtualization

CS 423: Operating System Design

Peizhe Liu

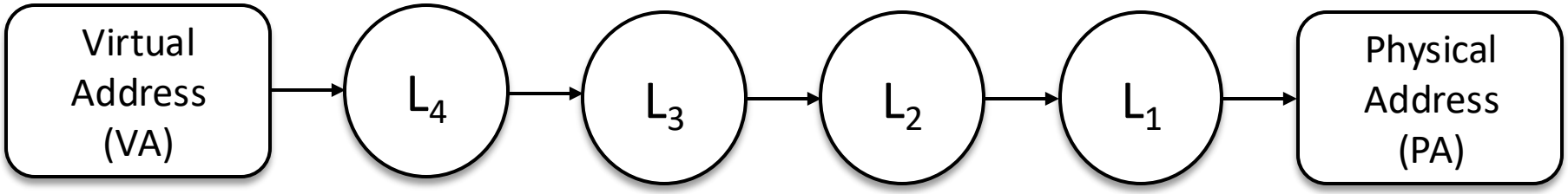
Recap: Native Memory Management



VMWare, Inc. Performance Evaluation of Intel EPT Hardware Assist. https://www.vmware.com/docs/perf_esx_intel-ept-eval.pdf. 2009.

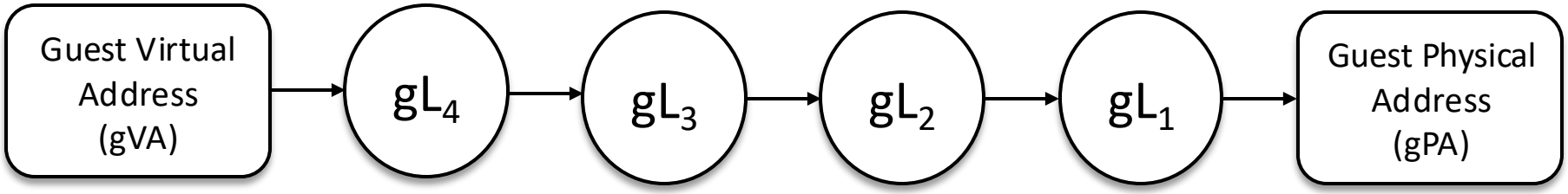
Recap: Address Translation

- Done by a hardware called MMU
- In case of a TLB miss…



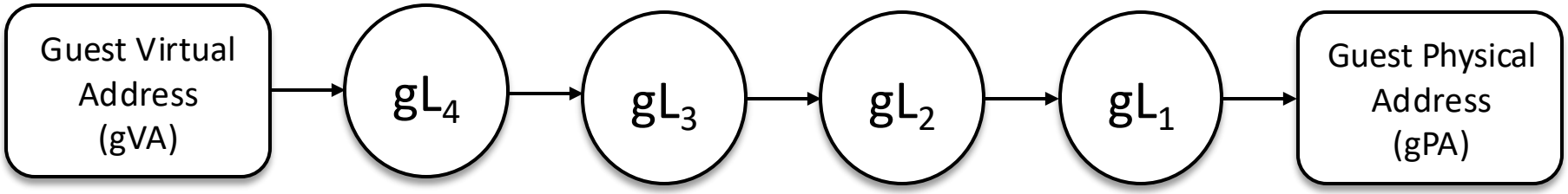
Address Translation in Guest OS

- Guest OS also have their own page tables.
- It translates from gVA to gPA only.

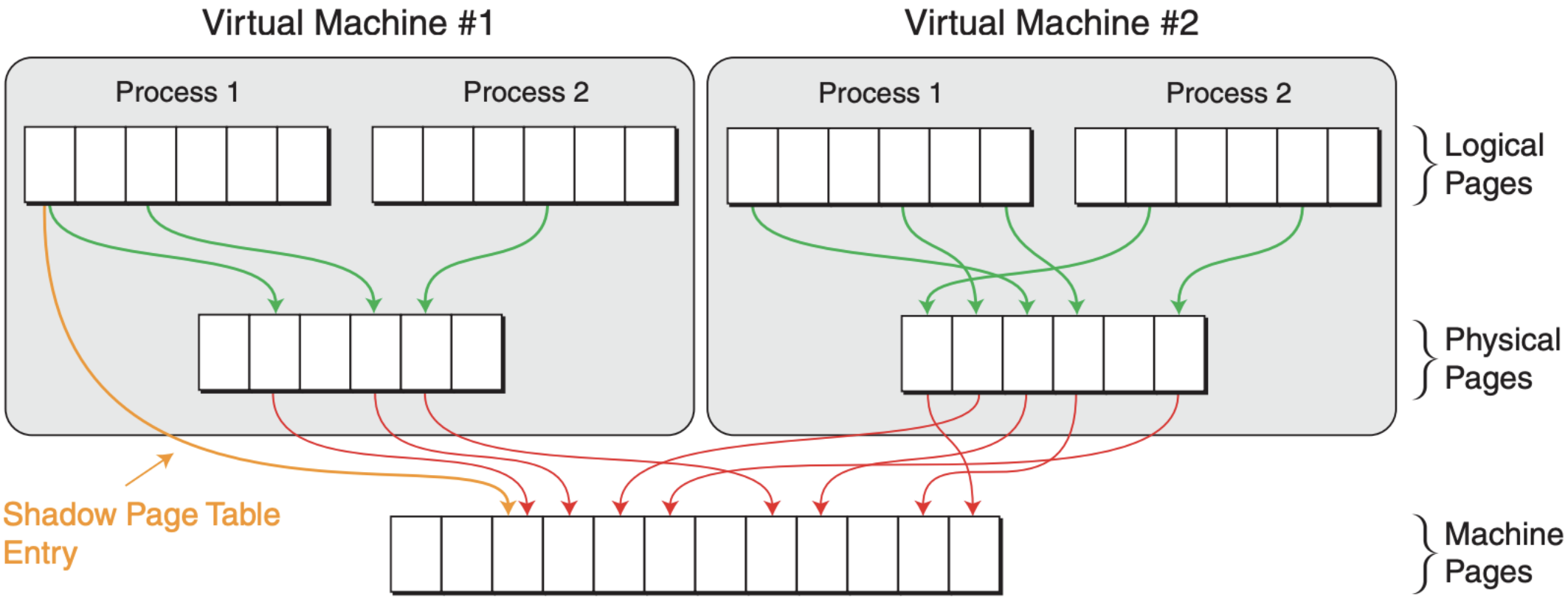


Address Translation in Guest OS

- Does this scheme work?
- **No! gPA (and every gL access) is not the actual PA.**



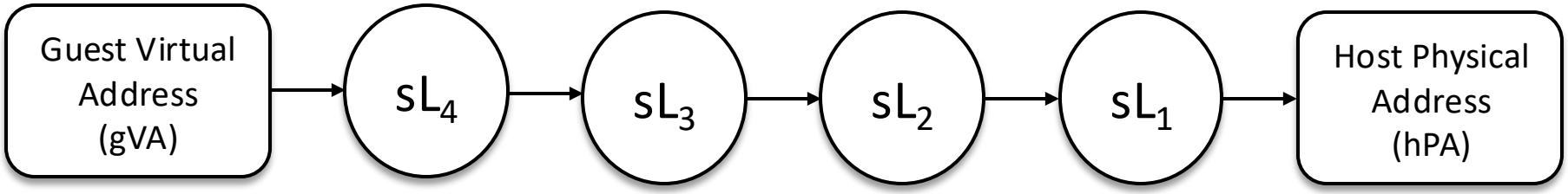
Shadow Page Table



VMWare, Inc. Performance Evaluation of Intel EPT Hardware Assist. https://www.vmware.com/docs/perf_esx_intel-ept-eval.pdf. 2009.

Shadow PT Address Translation

- Hypervisor maintains a shadow page table.
- Shadow page table can directly translate from gVA to hPA.

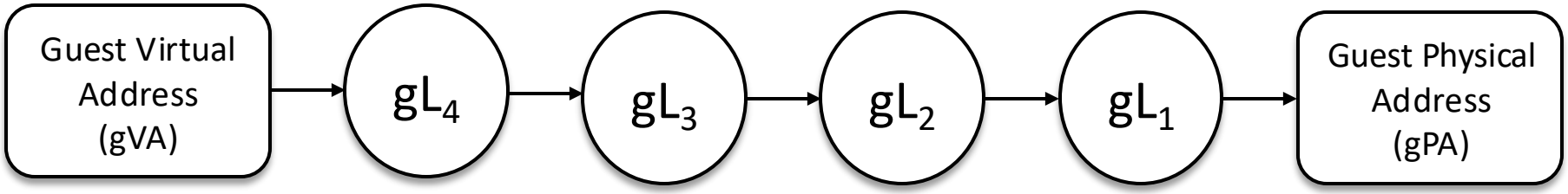


Maintaining the Shadow PT

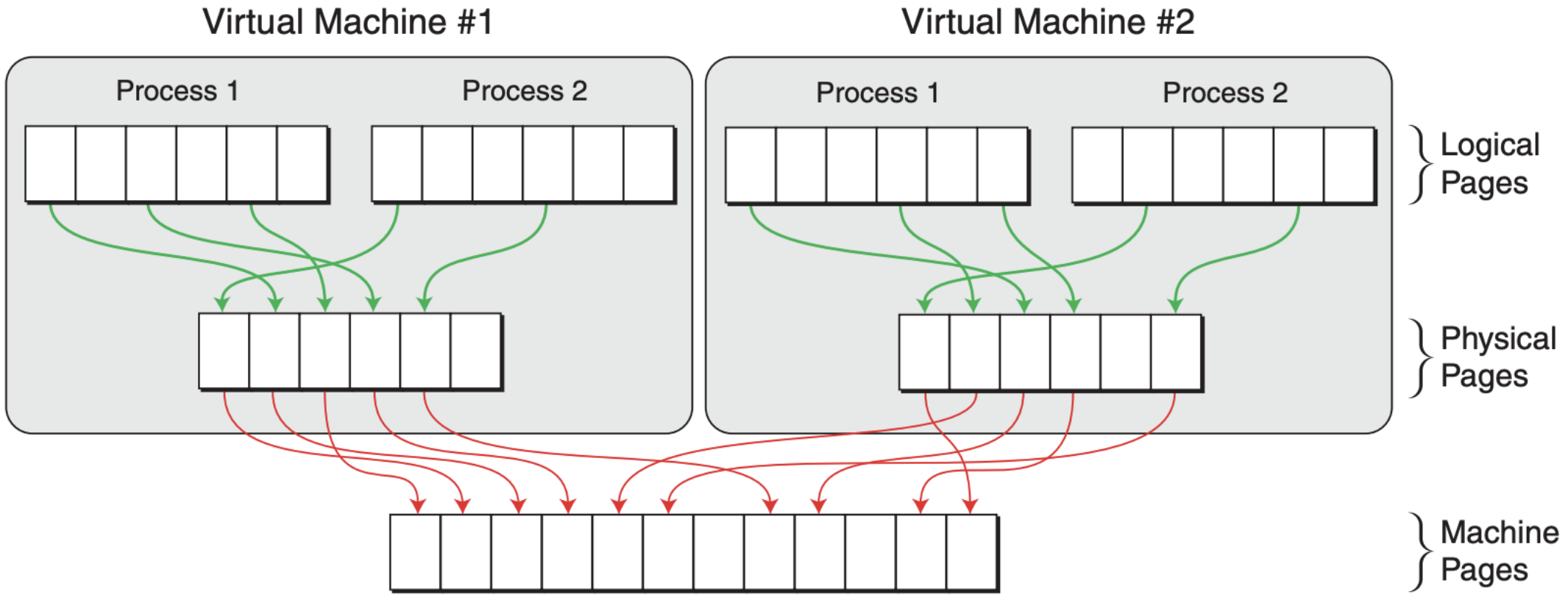
- When shadow PT #PF: VM exit and hypervisor create the entry.
- Let the guest OS handle #PF.
- When guest OS modify its PT: VM exit and hypervisor sync with the shadow PT.

Address Translation in Guest OS

- gPA (and every gL access) is not the actual PA.
- **However, we can let it translate to the actual PA (hPA).**



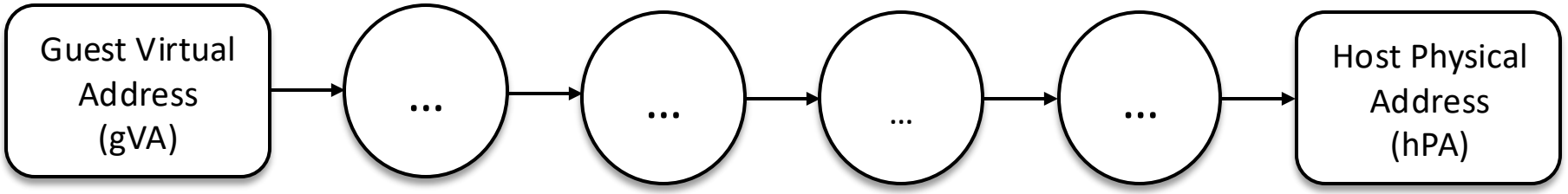
Extended Page Table (EPT)



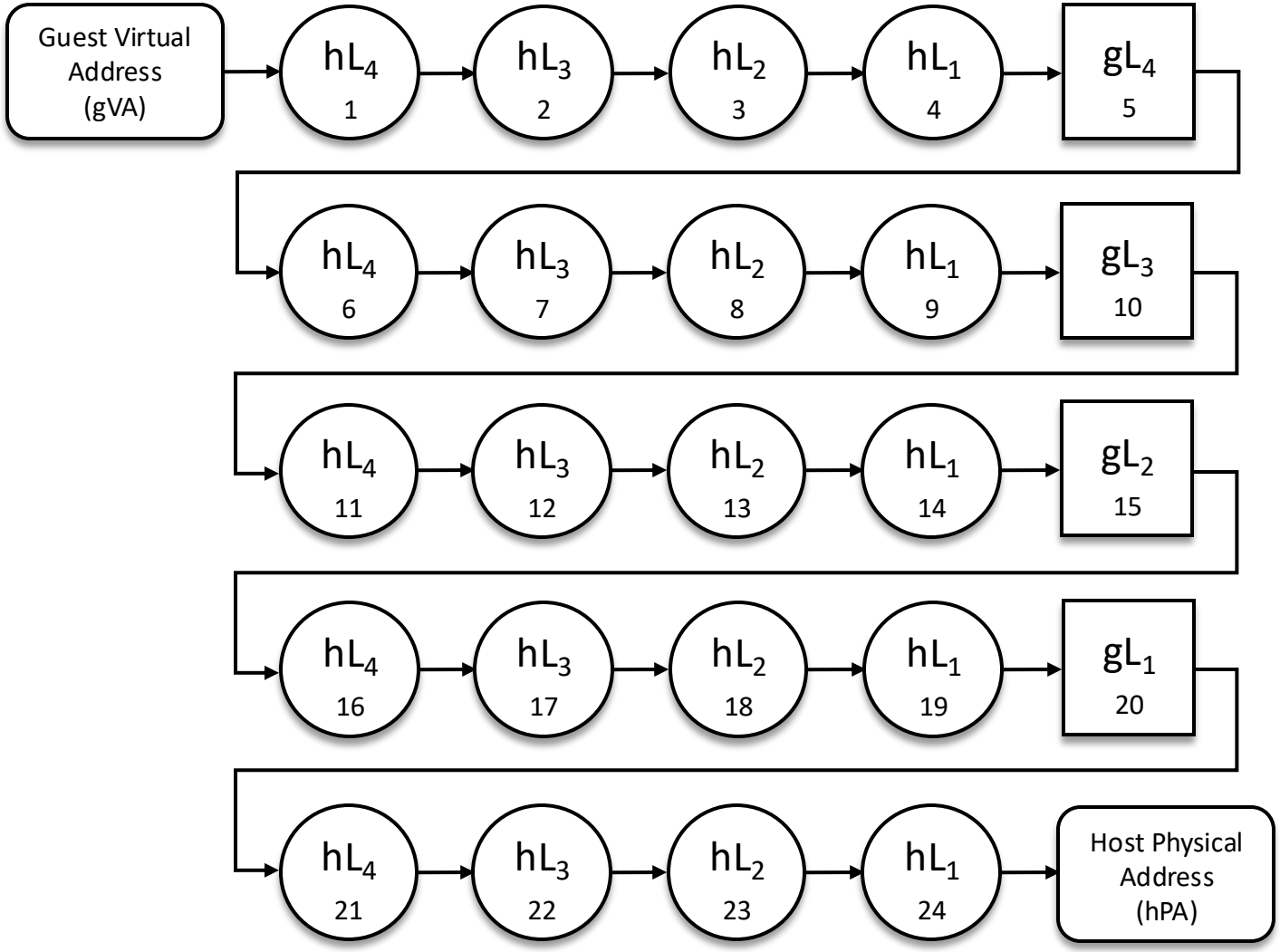
VMWare, Inc. Performance Evaluation of Intel EPT Hardware Assist. https://www.vmware.com/docs/perf_esx_intel-ept-eval.pdf. 2009.

EPT Address Translation

- For every gPA and gL access, extend the translation and find the actual hPA and hL.
- This is a 2-D page walk.



EPT Address Translation



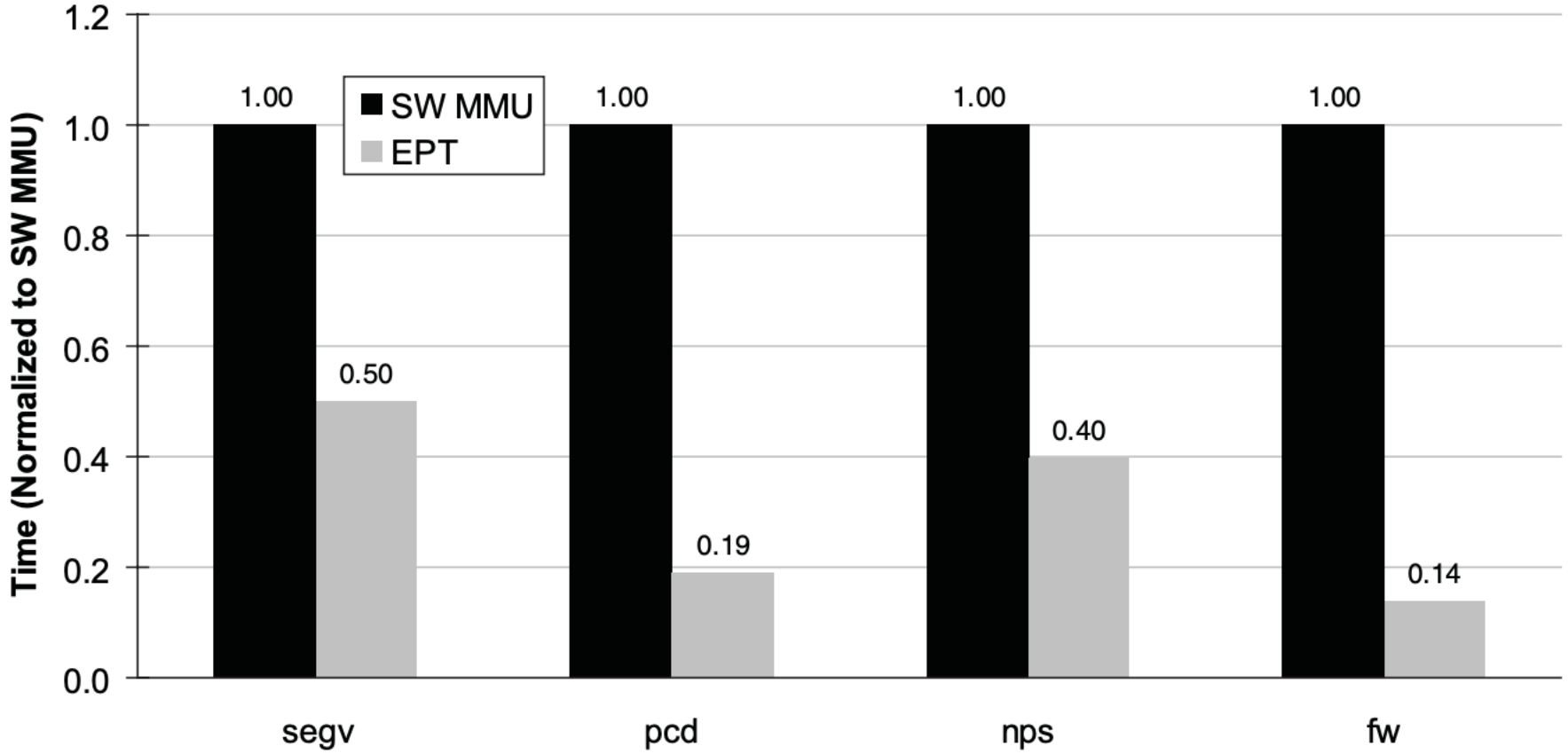
Shadow PT vs. EPT

- Shadow PT advantage:
  - EPT requires hardware (Intel VT-x/AMD-V) support.

- EPT advantages:
  - Eliminated the excessive VM exits, ctx switches, and TLB flushes.
  - Reduced memory footprint.
  - Simplified hypervisor design.

- **Today's virtualization widely used EPT.**

Shadow PT vs. EPT



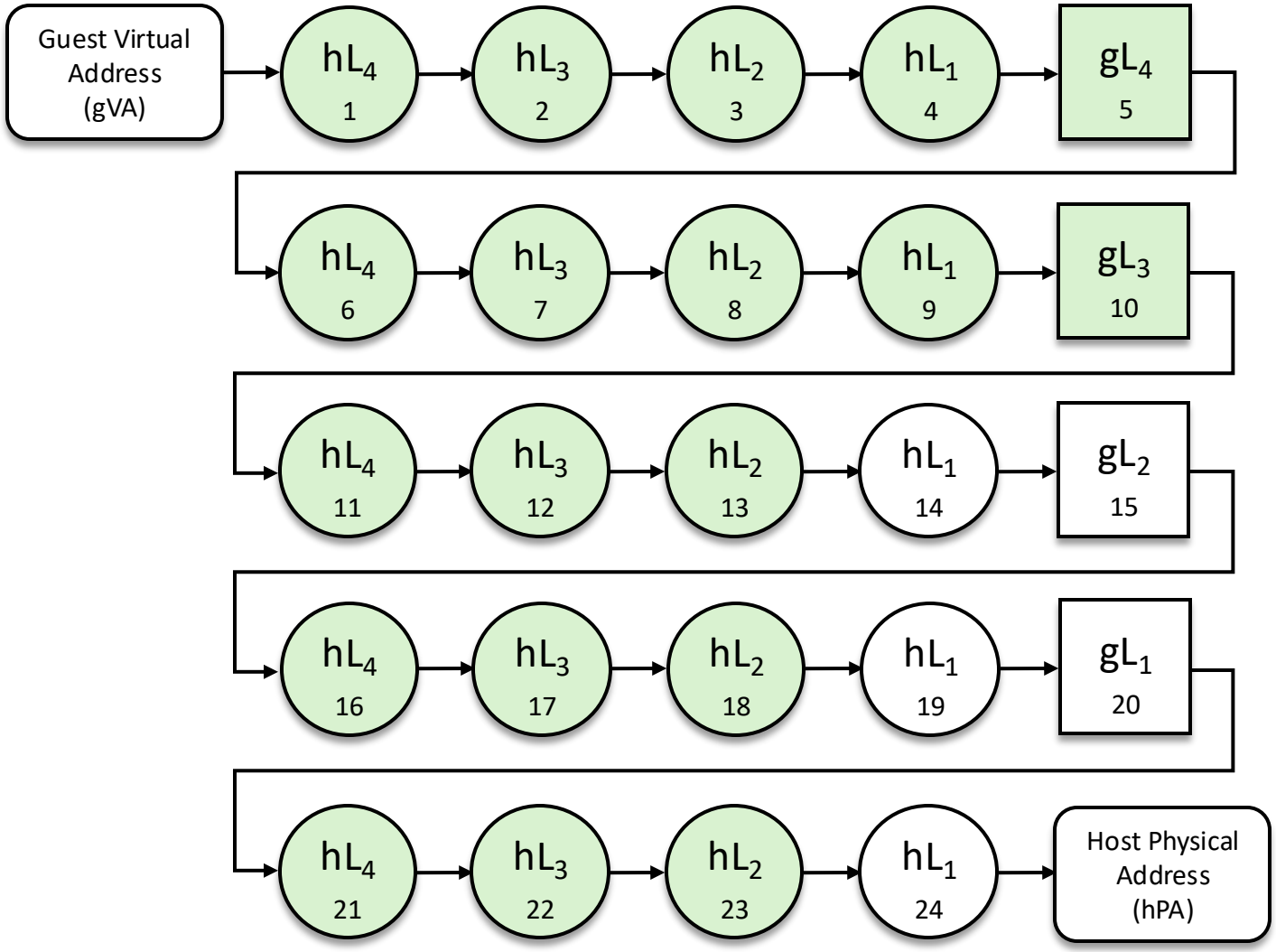
VMWare, Inc. Performance Evaluation of Intel EPT Hardware Assist. https://www.vmware.com/docs/perf_esx_intel-ept-eval.pdf. 2009.

EPT Overheads

- EPT introduced translation overheads (still better than VM exits!).

- Gets worse when scaling:
  - Nested virtualization.
  - Higher level page table.

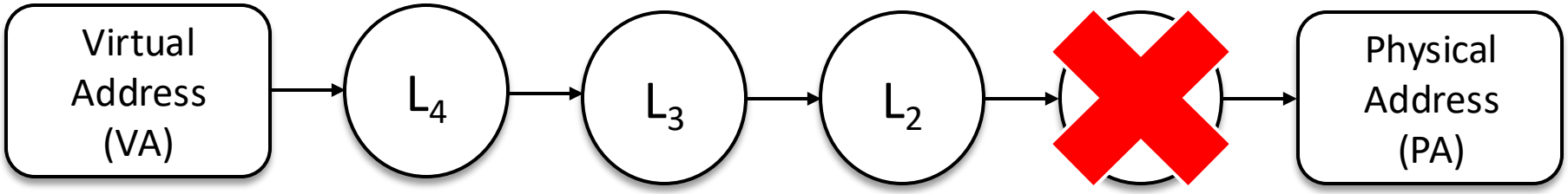
- Can mitigate with:
  - Page Walk Cache (PWC).
  - Various research projects.



24 memory accesses in the worst case

Huge Page

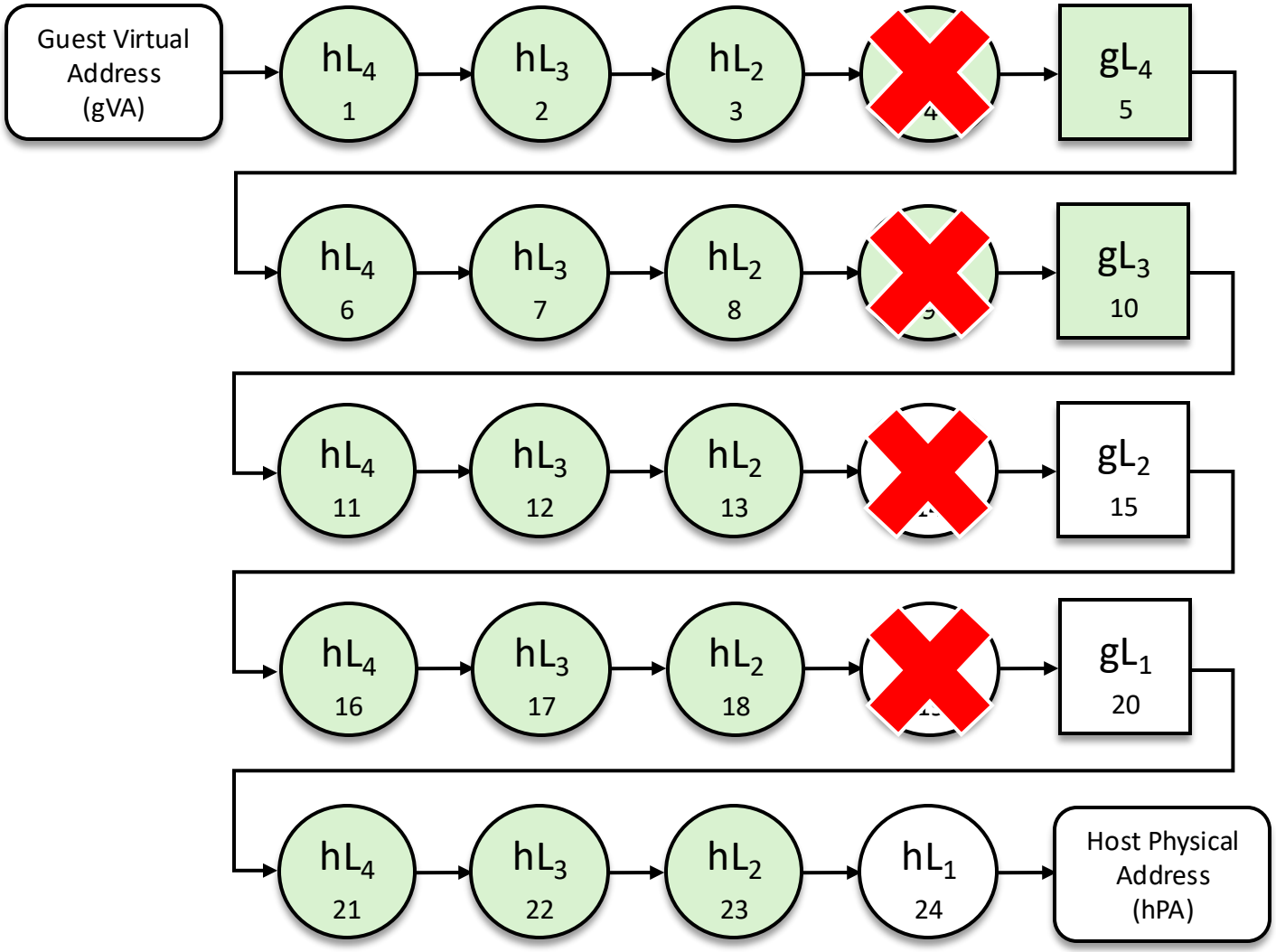
- Reduce the amount of TLB misses
- Reduce the amount of memory accesses during a page walk



3 memory accesses in the worst case

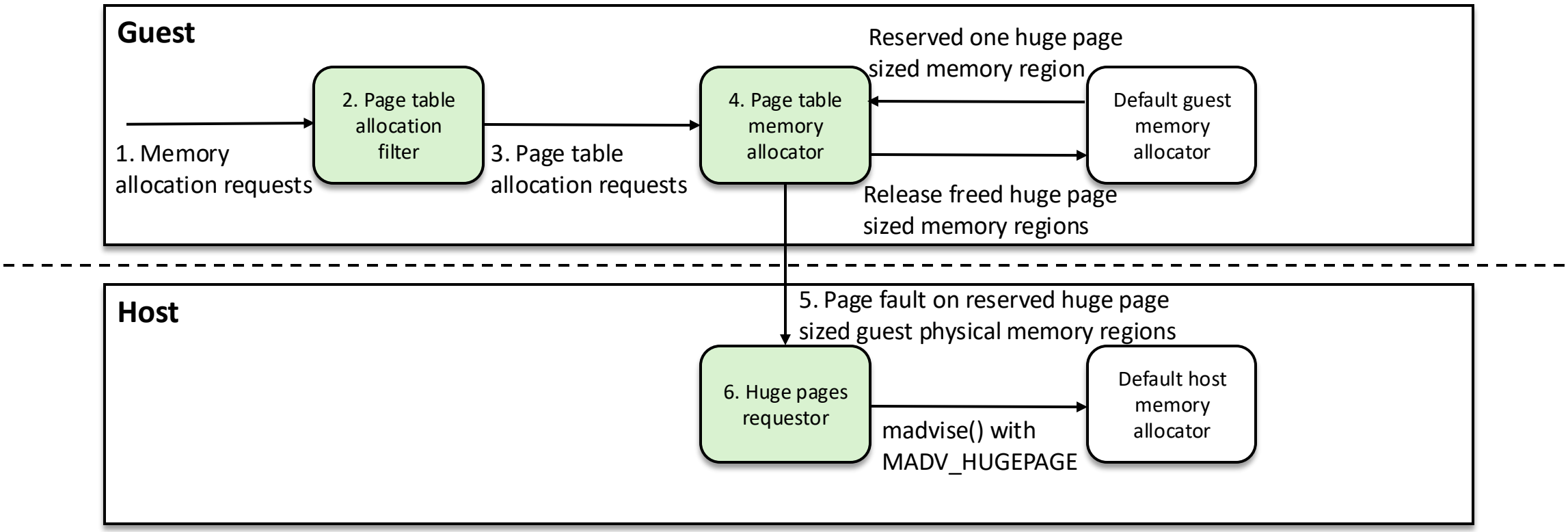
Research Project: HugeGPT

- Store the **G**uest **P**age **T**able on the host **Huge** page.
- Software approach to speed up EPT address translations.

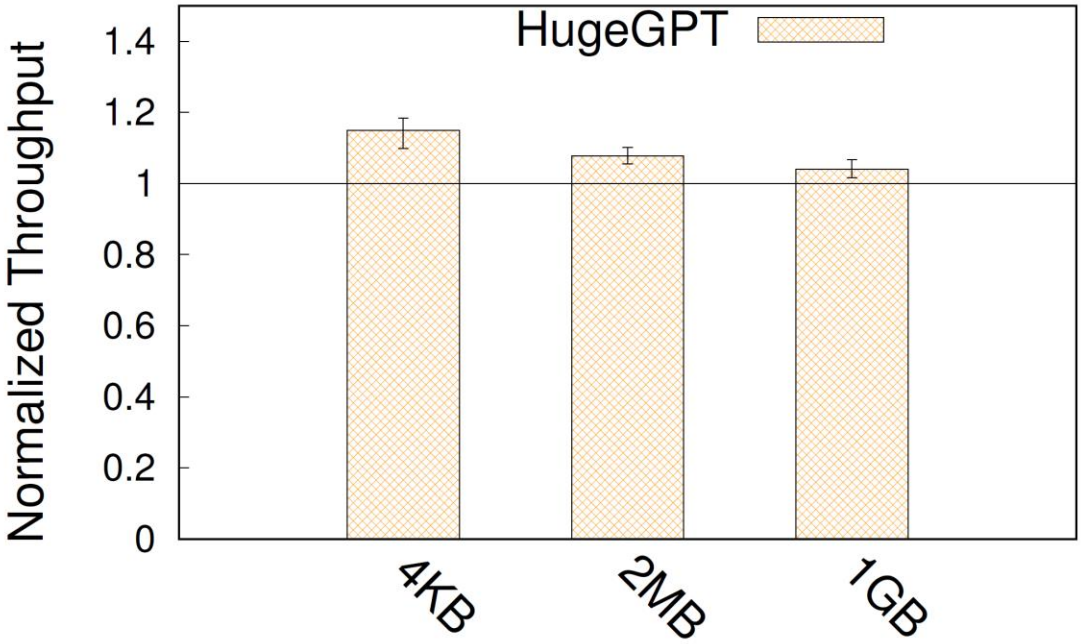


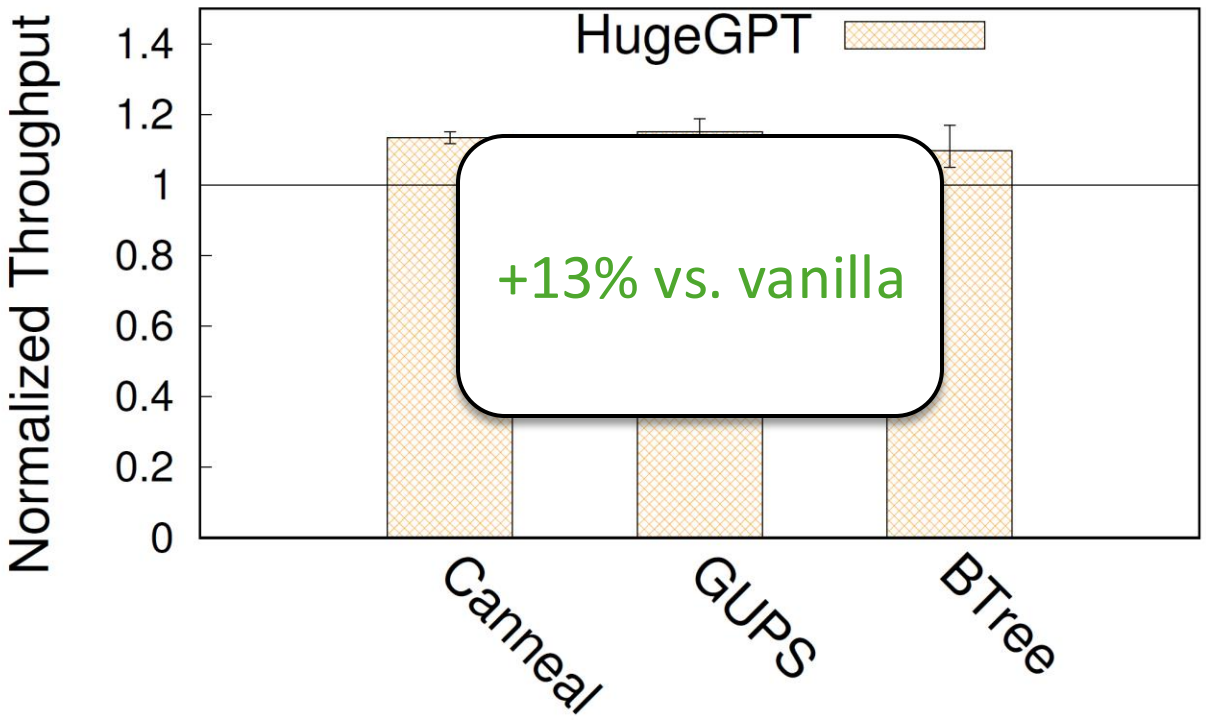
20 memory accesses in the worst case

HugeGPT Overview



HugeGPT Evaluation





Jia, W., Zhang, J., Shan, J., Du, Y., Ding, X., and Xu, T. HugeGPT: Storing Guest Page Tables on Host Huge Pages to Accelerate Address Translation. In *Proceedings of the 32nd International Conference on Parallel Architectures and Compilation Techniques (PACT'23)* (Oct. 2023).

HugeGPT Implementation

- Implemented on Linux v.6.1.81
- Guest: ~600 lines
- Host: ~200 lines

- Build and install https://github.com/xlab-uiuc/hugegpt-linux
- Requires `CONFIG_CMA=y`, `CONFIG_KVM=y`

Research Project: DMT

- **D**irect **M**emory **T**ranslation by mapping VMAs to last-level PTEs.
- Hardware-assisted approach to speed up native and EPT address translations.

VMA, TEA, and PTE

- VMA: Virtual Memory Region. Linux concept to manage the virtual memory region. Contains a continuous virtual memory region.

- DMT TEA: Translation Entry Area. DMT concept to manage last-level PTEs. Can be mapped by VMA and indexed in O(1).

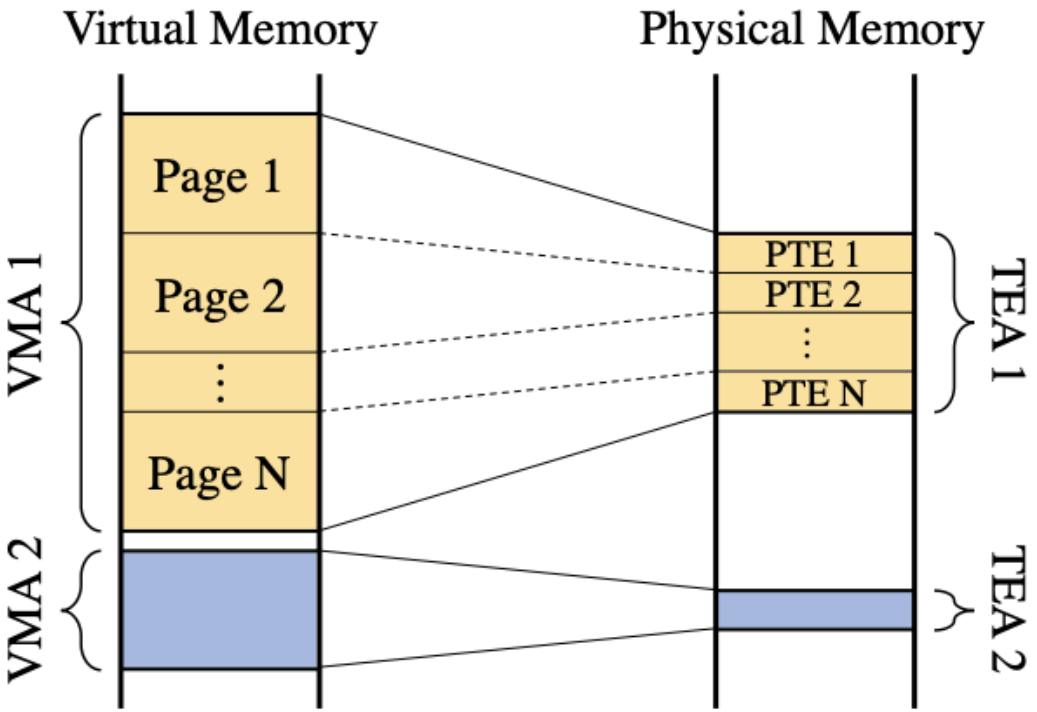
- Last-level PTE: Last-level Page Table Entry. General paging concept for address translation.

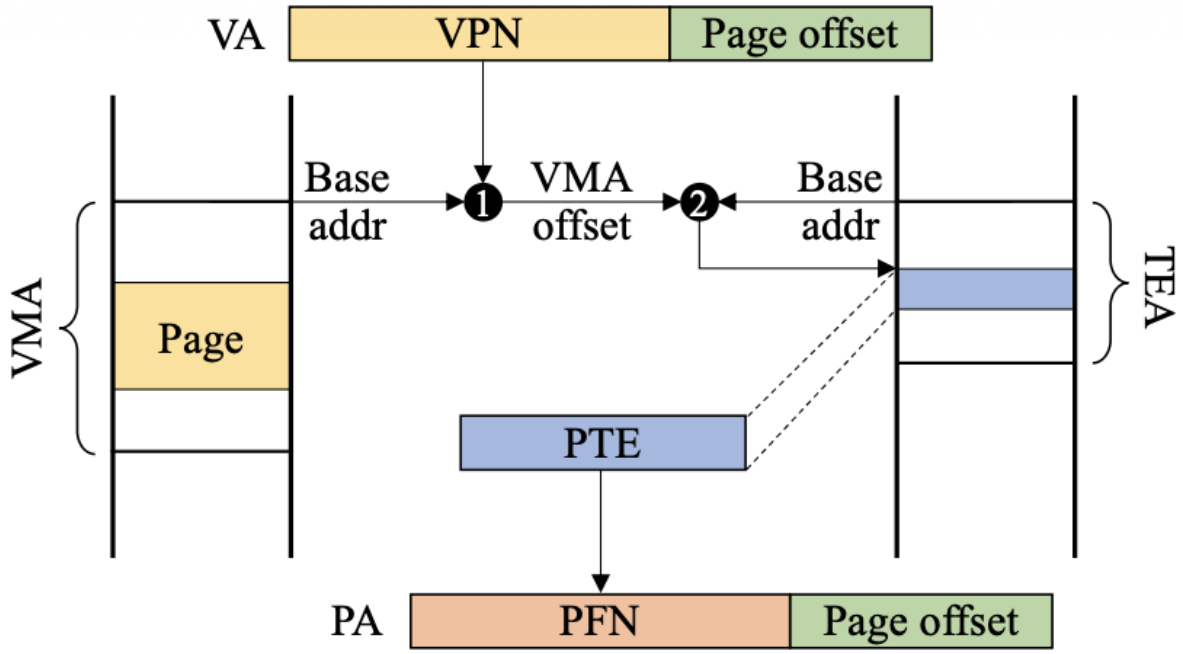
pvDMT

- DMT can accelerate native systems.
- DMT extends to pvDMT to accelerate virtualized and even nested virtualized systems.
- pvDMT works with KVM to maintain gTEAs for Guest OS.

- gTEA contains L1PTE, which maps to gPA, and we will need another map to hPA.
- Nested: L2PTE, L1PTE, and PTE.

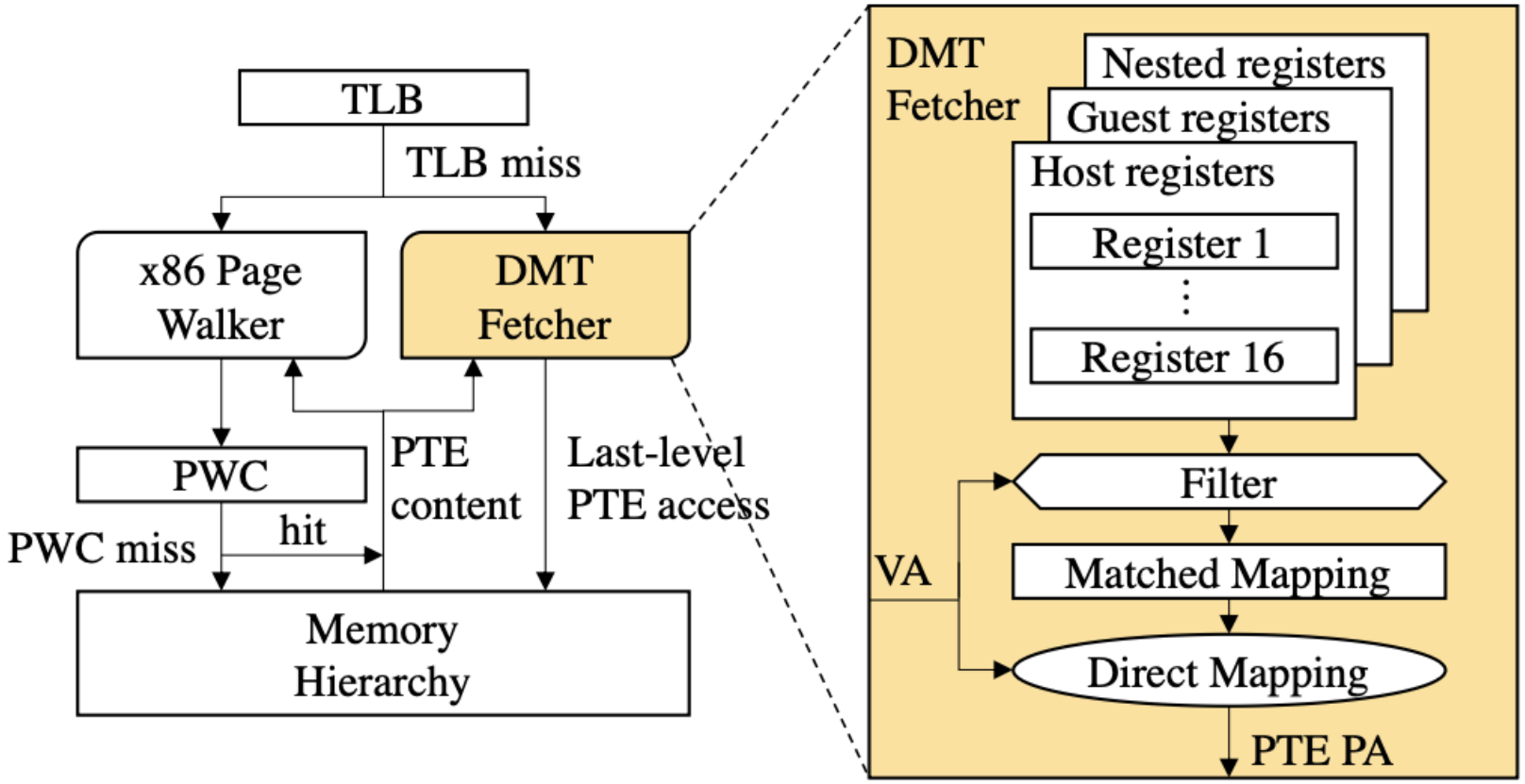
DMT Overview





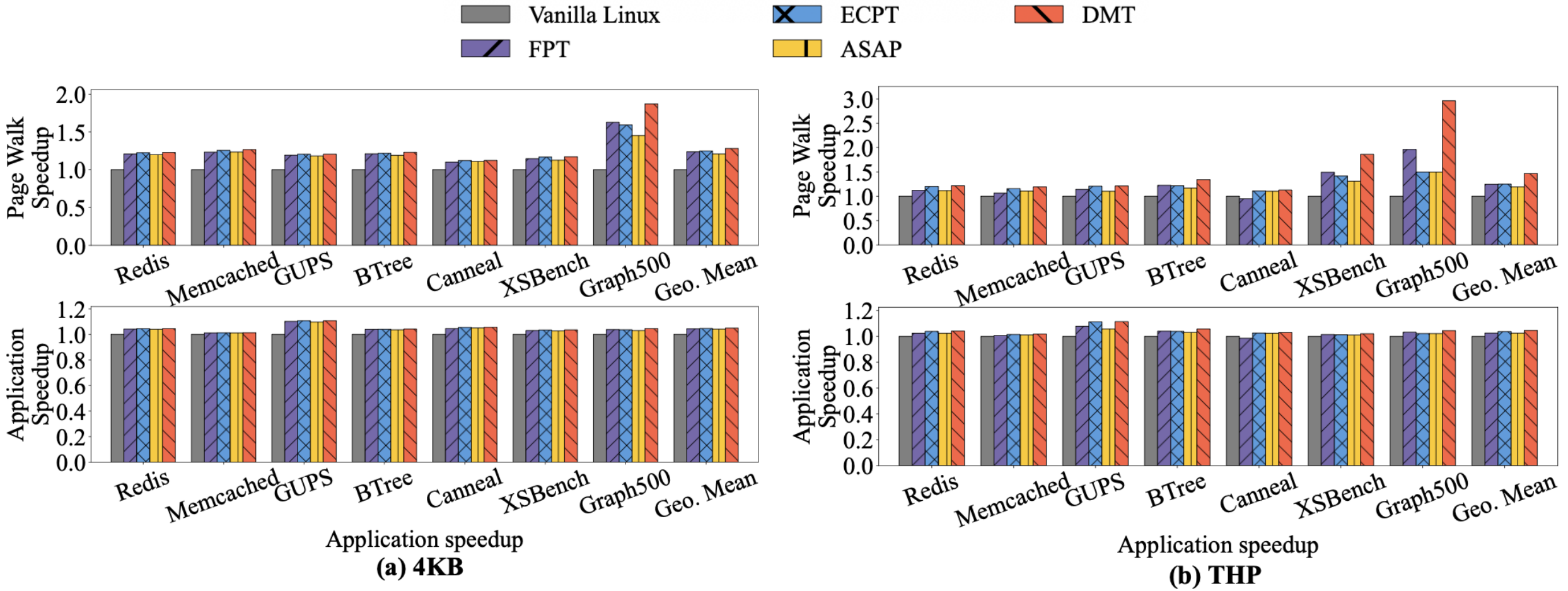
Zhang, J., Jia, W., Chai, S., Liu, P., Kim, J., and Xu, T. Direct Memory Translation for Virtualized Clouds. *In Proceedings of the 29th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'24)* (Apr. 2024).

DMT Overview



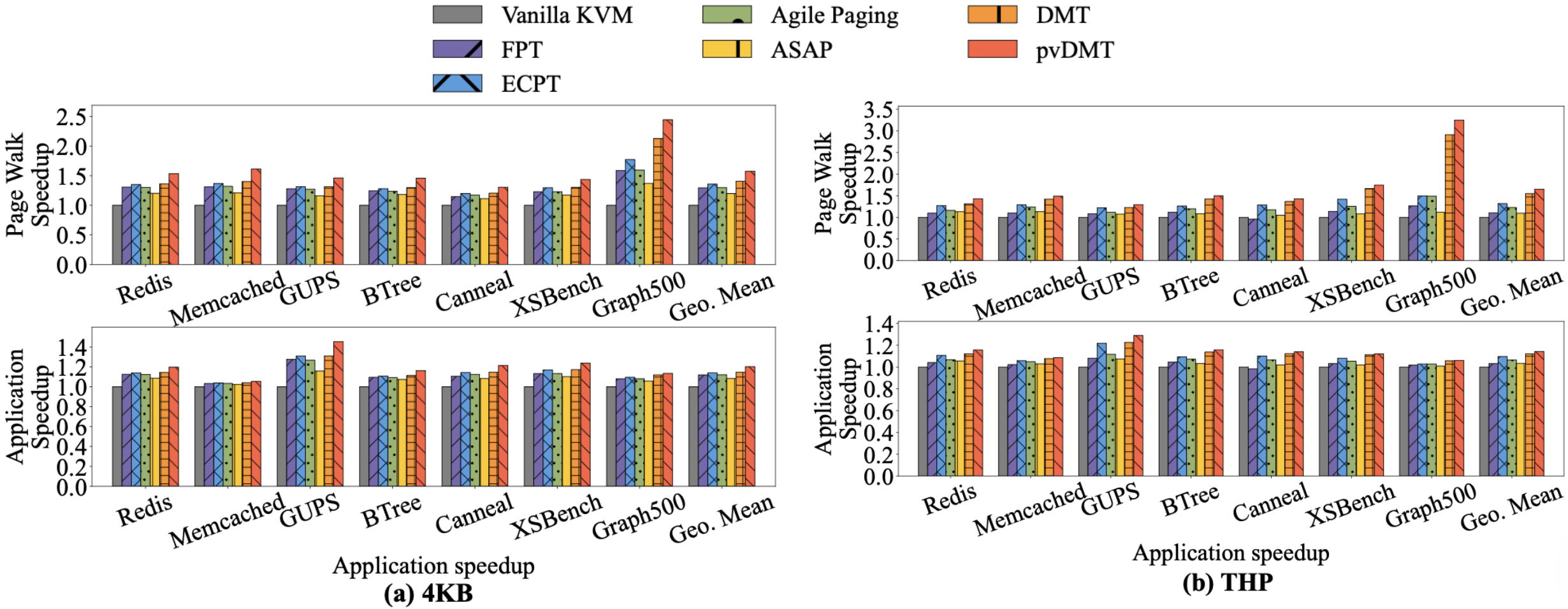
Zhang, J., Jia, W., Chai, S., Liu, P., Kim, J., and Xu, T. Direct Memory Translation for Virtualized Clouds. *In Proceedings of the 29th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'24)* (Apr. 2024).

DMT Evaluation (Native)



Zhang, J., Jia, W., Chai, S., Liu, P., Kim, J., and Xu, T. Direct Memory Translation for Virtualized Clouds. *In Proceedings of the 29th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'24)* (Apr. 2024).

DMT and pvDMT Evaluation (Virtualized)



Zhang, J., Jia, W., Chai, S., Liu, P., Kim, J., and Xu, T. Direct Memory Translation for Virtualized Clouds. *In Proceedings of the 29th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'24)* (Apr. 2024).

DMT Implementation

- Implemented on Linux v.5.15.127

- Artifact https://github.com/xlab-uiuc/dmt

Takeaways

- **Shadow PT:** original software approach to build a shadow PT.
- **EPT:** hardware approach to extend the PW and removed the excessive VM exits.

- **HugeGPT:** software approach to put guest PTs on hugepages, improved the address translation overhead of EPT.
- **DMT:** hardware approach to build VMA to last-level PTE mappings, support both native, virtualized, and nested virtualized systems; improved the address translation overhead of EPT.