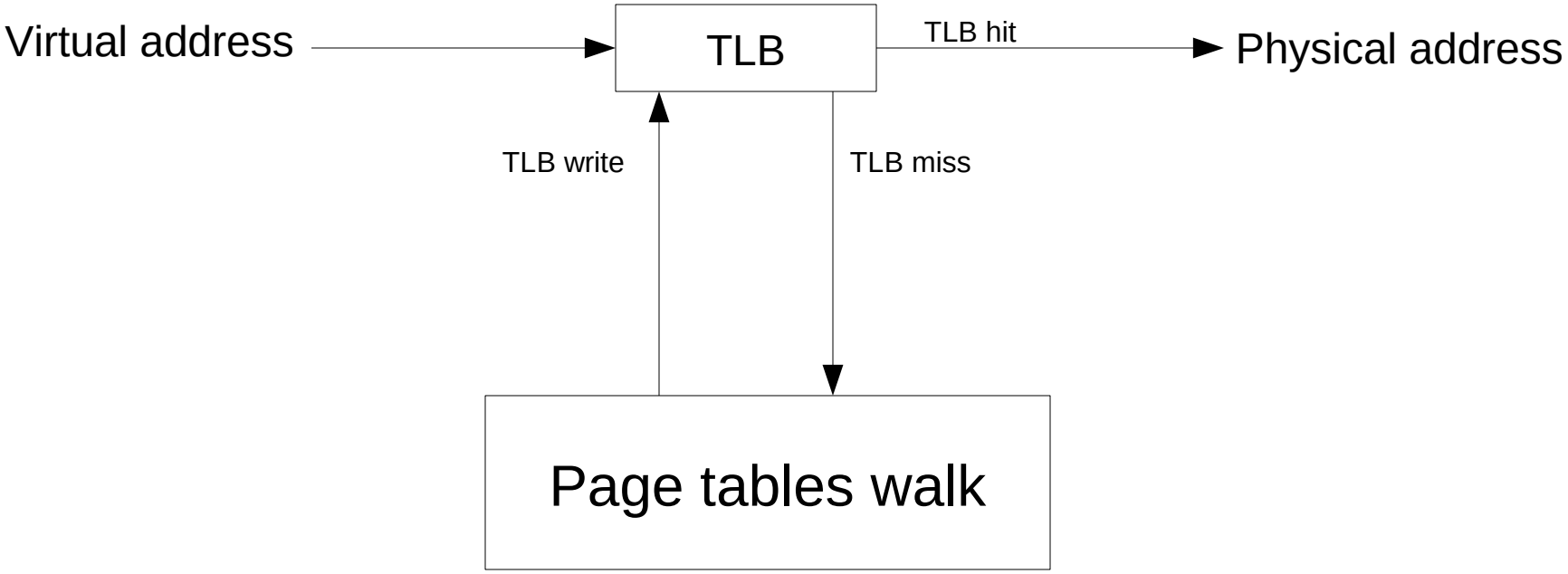
Transparent hugepages. Huge zero page.

Kirill A. Shutemov

29 December 2012

Minsk, Belarus

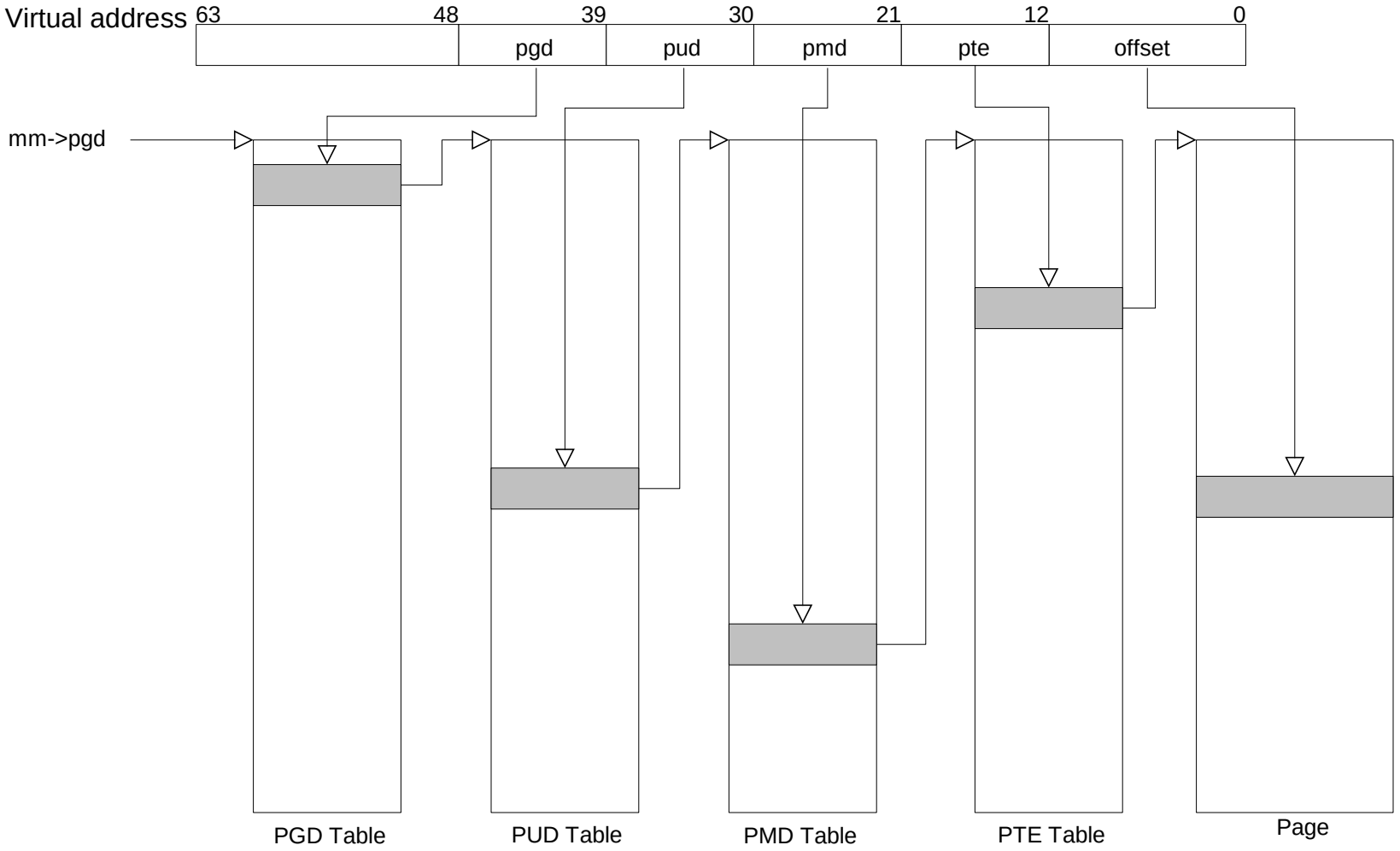
Virtual to physical address translation



TLB

- TLB is **Fast:**
  - Three virtual to physical address translations every cycle (2 load, 1 store);
  - TLB miss, STLB hit penalty is 7 cycles; can usually be hidden by OOO execution;
- Page table walk is **Slow:**
  - Hundreds cycles;
  - In the worst case a page table can be in swap;

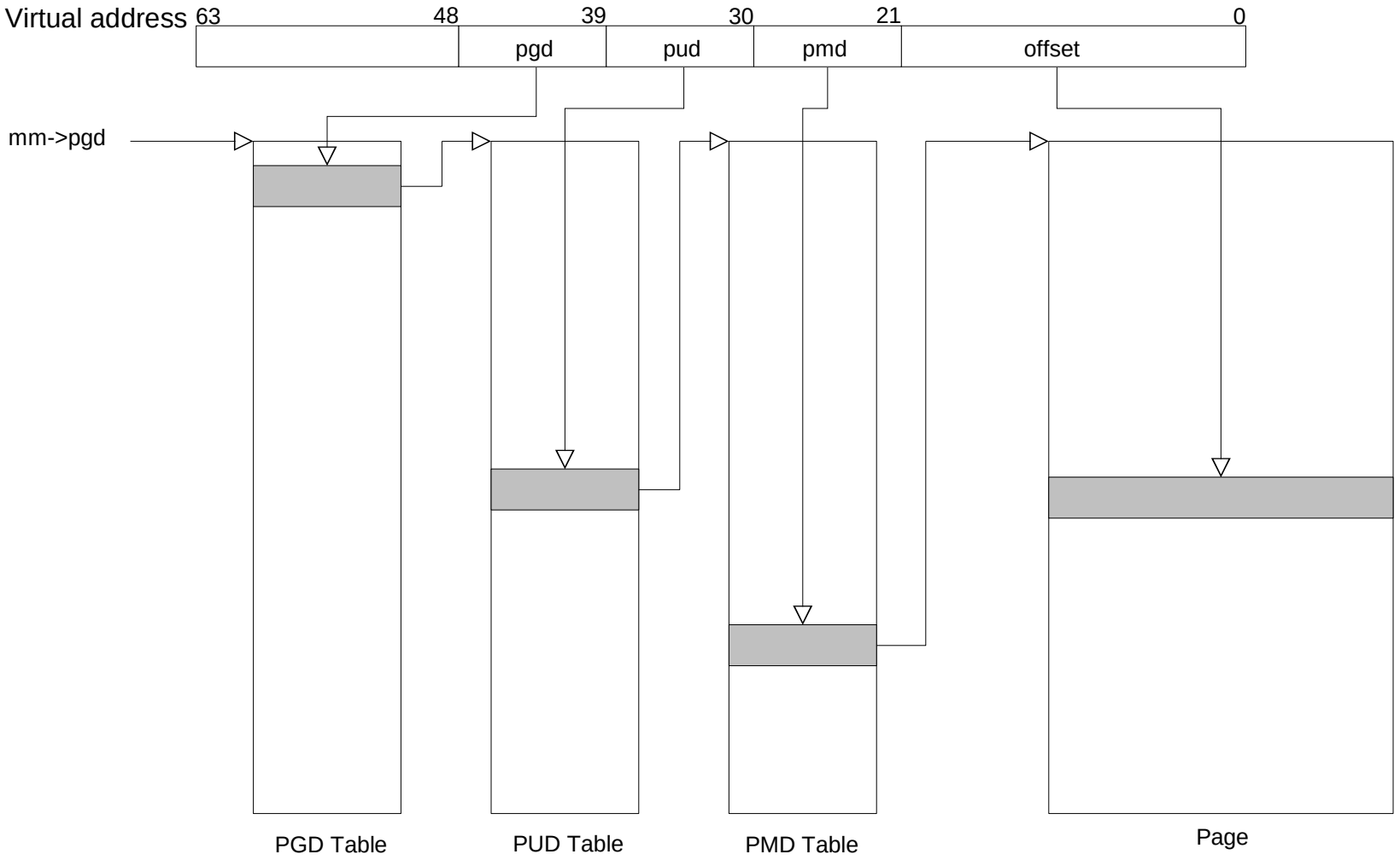
Page table walk, 4k pages



Huge pages

- Pros:
  - Reduce TLB pressure:
    - 1 TLB entry instead of 512 for every 2M;
    - 4k and 2M pages have separate TLB entires;
  - Faster page tables walk: 3 data access instead of 4;
  - Reduce cache pressure;
- Cons:
  - Memory footprint;
  - clear_page/copy_page is not cache friendly;

Page table walk, 2M pages



Hugetlbfs

- Cannot be swapped out;
- Need to be reserved on boot time;
- Requires to root access;
- Explicit application support is required;
- No fallback to 4k pages;

Transparent Hugepages

- No special enabling is required;
- Graceful fallback to 4k pages;
- Hugepage can be splitted to 4k pages at any time;
- Hugepages can be swapped out as 4k pages;
- MADV_HUGEPAGE and MADV_NOHUGEPAGE for tuning;

Zero page

```
#include <assert.h>
#include <stdlib.h>
#include <unistd.h>

#define MB (1024*1024)

int main(int argc, char **argv)
{
        char *p;
        int i;

        posix_memalign((void **)&p, 2 * MB, 200 * MB);
        for (i = 0; i < 200 * MB; i += 4096)
                assert(p[i] == 0);
        pause();
        return 0;
}
```

Huge zero page

- Two designs have been evaluated:
  - “physical” huge zero page;
  - “virtual” huge zero page;
- Lockless refcounting:
  - free huge zero page in shrinker callback;
  - ~1% slowdown on 40 parallel read page faulting processes;
- Merged in v3.8;
- LWN: http://lwn.net/Articles/517465/

Questions?


TURE INTEL LINUX WIRELESS GUPNP KVM POKY
OP OFONO LINUX KERNEL
CS YOCTO CONNMAN XEN
SYNCEVOLUTION SIMPLE FIRMWARE INTERFACE (SFI) ENTERPRISE SECURITY INFRASTRUCTURE
INTEL OPEN SOURCE
TECHNOLOGY CENTER