Welcome back!

MEMORY: SWAPPING

Shivaram Venkataraman
CS 537, Spring 2023
Project 3 was due **Monday**.

Project 4: Scheduling new dates: Feb 22\(^{nd}\) to March 6\(^{th}\) → **March 2**

Midterm 1: In class midterm, Multiple choice.
No notes / calculators. (We will give a table of powers of 2)

→ Old exams on Canvas
→ Discussion: Practice problems
  → Handout

→ Textbook

→ Video Playlist
  → Some from past,
  → This year
OFFICE HOURS

1. One question per student at a time
2. Please be prepared before asking questions
3. The TAs might not be able to fix your problem
4. Limited time per student → 10 mins

Our

1. Increase number of TAs close to deadline
2. Study groups
   → Midterm
AGENDA / LEARNING OUTCOMES

Memory virtualization
  How we support virtual mem larger than physical mem?
  What are mechanisms and policies for this?
RECAP
Multilevel, Inverted Page Tables

PPN VPN Prot

Software defined TLB

Multi-level Page Table

PDBR

valid | PFN
---|---
1 | 201
0 | -
0 | -
1 | 204

Page Directory

valid | prot | PFN
---|---|---
1 | rx | 12
1 | rx | 13
0 | - | -
1 | rw | 100

[Page 1 of PT: Not Allocated]

[Page 2 of PT: Not Allocated]

valid | prot | PFN
---|---|---
0 | - | -
0 | - | -
1 | rw | 86
1 | rw | 15
HugePages saves TLB entries. But how does it affect page translation?

4KB pages: 4 levels $\rightarrow$ 4 memory accesses

<table>
<thead>
<tr>
<th>Level</th>
<th>Page Map Lvl 4</th>
<th>Page Pointer Dir.</th>
<th>Page Directory</th>
<th>Page Table</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>47-39</td>
<td>(9 bits)</td>
<td>(9 bits)</td>
<td>(9 bits)</td>
<td>(9 bits)</td>
<td>(12 bits)</td>
</tr>
</tbody>
</table>

2MB pages: 3 levels $\rightarrow$ 3 memory access on translation

<table>
<thead>
<tr>
<th>Level</th>
<th>Page Map Lvl 4</th>
<th>Page Pointer Dir.</th>
<th>Page Directory</th>
<th>Page Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>47-39</td>
<td>(bits)</td>
<td>(bits)</td>
<td>(bits)</td>
<td>(2^21 bits)</td>
</tr>
</tbody>
</table>

$\downarrow$ 2 MB
SUMMARY: BETTER PAGE TABLES

Problem: Simple linear page tables require too much contiguous memory.

Many options for efficiently organizing page tables:
- Inverted page tables (hashing) reduce page table size.

If Hardware handles TLB miss, page tables must follow specific format:
- Multi-level page tables used in x86 architecture
- Each inner page table fits within a page.

Large pages can reduce TLB use and number of accesses for translation.
SWAPPING
MOTIVATION

OS goal: Support processes when not enough physical memory
- Single process with very large address space
- Multiple processes with combined address spaces

User code should be independent of amount of physical memory
- Correctness, if not performance

Virtual memory: OS provides illusion of more physical memory

Why does this work?
- Relies on key properties of user processes (workload) and machine architecture (hardware)
Leverage locality of reference within processes

- **Spatial**: reference memory addresses near previously referenced addresses
- **Temporal**: reference memory addresses that have referenced in the past
- Processes spend majority of time in small portion of code
  - Estimate: 90% of time in 10% of code

Implication:

- Process only uses small amount of address space at any moment
- Only small amount of address space must be resident in physical memory
Leverage memory hierarchy of machine architecture.
Each layer acts as “backing store” for layer above.

- **registers**
  - few tens
  - ns or less
  - few MB

- **cache**
  - 1-3ns
  - few GB

- **main memory**
  - 100s of ns
  - few GB

- **disk storage**
  - micro/ ms
  - 100s of GB

Can use disk as a backing store.

How long does it take to read?
SWAPPING INTUITION

Idea: OS keeps unreferenced pages on disk
   – Slower, cheaper backing store than memory

Process can run when not all pages are loaded into main memory
OS and hardware cooperate to make large disk seem like memory
   – Same behavior as if all of address space in main memory

Requirements:
   – OS must have mechanism to identify location of each page in address space → in memory or on disk
   – OS must have policy to determine which pages live in memory and which on disk
VIRTUAL ADDRESS SPACE MECHANISMS

Each page in virtual address space maps to one of three locations:
- Physical main memory: Small, fast, expensive
- Disk (backing store): Large, slow, cheap
- Nothing (error): Free

Extend page tables with an extra bit: present
- permissions \((r/w)\), valid, present
- Page in memory: present bit set in PTE
- Page on disk: present bit cleared
  - PTE points to block on disk
  - Causes trap into OS when page is referenced
  - Trap: page fault
What if access vpn 0xb?

- Read page from block 28
- Store page 12
- Update PTE
- Trap: Page fault
  - retry the translation
  - PFN valid 10 prot r-x present 1
VIRTUAL MEMORY MECHANISMS

First, hardware checks TLB for virtual address
   – if TLB hit, address translation is done; page in physical memory

Else
   ...  
   – Hardware or OS walk page tables
   – If PTE designates page is present, then page in physical memory (i.e., present bit is cleared)

 Else
   – Trap into OS (not handled by hardware)
   – OS selects victim page in memory to replace
     • Write victim page out to disk if modified (use dirty bit in PTE)
   – OS reads referenced page from disk into memory
   – Page table is updated, present bit is set
   – Process continues execution
SWAPPING POLICIES
SWAPPING POLICIES

Goal: Minimize number of page faults

- Page faults require milliseconds to handle (reading from disk)
- Implication: Plenty of time for OS to make good decision

OS has two decisions

- Page selection
  When should a page (or pages) on disk be **brought into** memory?

- Page replacement
  Which resident page (or pages) in memory should be **thrown out** to disk?

unlike TLBs

or evicted
PAGE SELECTION

Demand paging: Load page only when page fault occurs
- Intuition: Wait until page must absolutely be in memory
- When process starts: No pages are loaded in memory
- Problems: Pay cost of page fault for every newly accessed page

Prepaging (anticipatory, prefetching): Load page before referenced
- OS predicts future accesses (oracle) and brings pages into memory early
- Works well for some access patterns (e.g., sequential)

Hints: Combine above with user-supplied hints about page references
- User specifies: may need page in future, don’t need this page anymore, or sequential access pattern, ...
- Example: madvise() in Unix

When should we fetch a page from disk?

Virtual address

Array

I will access this page in the future
Which page in main memory should be selected as victim?

- Write out victim page to disk if modified (dirty bit set)
- If victim page is not modified (clean), just discard

**OPT:** Replace page not used for longest time in future

- Advantages: Guaranteed to minimize number of page faults
- Disadvantages: Requires that OS predict the future; Not practical, but good for comparison
PAGE REPLACEMENT

**FIFO:** Replace page that has been in memory the longest
- Intuition: First referenced long time ago, done with it now
- Advantages: Fair: All pages receive equal residency; Easy to implement
- Disadvantage: Some pages may always be needed

**LRU:** Least-recently-used: Replace page not used for longest time in past
- Intuition: Use past to predict the future
- Advantages: With locality, LRU approximates OPT
- Disadvantages:
  - Harder to implement, must track which pages have been accessed
  - Does not handle all workloads well

First In First Out

If a page is popular, then it might not stay in memory

sort pages by access time
### Page Replacement

**Three pages of physical memory**

<table>
<thead>
<tr>
<th>Hit/Miss</th>
<th>OPT</th>
<th>FIFO</th>
<th>LRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>M D</td>
<td>D</td>
<td>M D</td>
<td></td>
</tr>
<tr>
<td>H D</td>
<td>D</td>
<td>H D</td>
<td></td>
</tr>
<tr>
<td>M B</td>
<td>D B</td>
<td>M D B</td>
<td></td>
</tr>
<tr>
<td>H B</td>
<td>D B</td>
<td>H D B</td>
<td></td>
</tr>
<tr>
<td>M A</td>
<td>D B A</td>
<td>M D B A</td>
<td></td>
</tr>
<tr>
<td>M C</td>
<td>D B C</td>
<td>M C B A</td>
<td></td>
</tr>
<tr>
<td>H B</td>
<td></td>
<td>H C B A</td>
<td></td>
</tr>
<tr>
<td>H D</td>
<td></td>
<td>M C D A</td>
<td></td>
</tr>
<tr>
<td>H B</td>
<td></td>
<td>M C D B</td>
<td></td>
</tr>
<tr>
<td>H D</td>
<td></td>
<td>H</td>
<td></td>
</tr>
</tbody>
</table>

**Page reference string:** DDBBACBBDDBD

**Metric:**

- **Miss count**
  - 4 misses

**Misses:**

- OPT: 6 misses
- FIFO: 6 misses
- LRU: 6 misses
Page reference string: ABCABDADADBCB

<table>
<thead>
<tr>
<th>Metric: Miss count</th>
<th>OPT</th>
<th>FIFO = 7</th>
<th>LRU = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three pages of physical memory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ABC</td>
<td>A B C</td>
<td>A B C</td>
</tr>
<tr>
<td>H A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>A B D</td>
<td>M D B C</td>
</tr>
<tr>
<td>H A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C B D</td>
<td>M C A B</td>
<td>M C B D</td>
</tr>
<tr>
<td>H B</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Add more physical memory, what happens to performance?

**LRU, OPT:**
- Guaranteed to have fewer (or same number of) page faults
- Smaller memory sizes are guaranteed to contain a subset of larger memory sizes
- Stack property: smaller cache always subset of bigger

**FIFO:**
- Usually have fewer page faults
- Belady’s anomaly: May actually have more page faults!
FIFO PERFORMANCE MAY DECREASE!

Consider access stream: ABCDABEABCD

Physical memory size: 3 pages vs. 4 pages

How many misses with FIFO?

Anomaly
IMPLEMENTING LRU

Software Perfect LRU
- OS maintains ordered list of physical pages by reference time
- When page is referenced: Move page to front of list
- When need victim: Pick page at back of list
- Trade-off: Slow on memory reference, fast on replacement

Hardware Perfect LRU
- Associate timestamp register with each page
- When page is referenced: Store system clock in register
- When need victim: Scan through registers to find oldest clock
- Trade-off: Fast on memory reference, slow on replacement (especially as size of memory grows)

In practice
LRU is an approximation anyway, so approximate more?
CLOCK ALGORITHM

Hardware
- Keep use (or reference) bit for each page frame
- When page is referenced: set use bit

Operating System
- Page replacement: Look for page with use bit cleared (has not been referenced for awhile)
- Implementation:
  - Keep pointer to last examined page frame
  - Traverse pages in circular buffer
  - Clear use bits as search
  - Stop when find page with already cleared use bit, replace this page
CLOCK: LOOK FOR A PAGE

Physical Mem:

- Use = 1, 1, 0, 1 to begin

Evict a page bring in 4
- select 2 for eviction

Page 0 is accessed
- Page 5 bring in
- select 1 for eviction
CLOCK EXTENSIONS

Replace multiple pages at once
  – Intuition: Expensive to run replacement algorithm and to write single block to disk
  – Find multiple victims each time and track free list

Use dirty bit to give preference to dirty pages
  – Intuition: More expensive to replace dirty pages
    Dirty pages must be written to disk, clean pages do not
  – Replace pages that have use bit and dirty bit cleared
SUMMARY: VIRTUAL MEMORY

Abstraction: Virtual address space with code, heap, stack

Address translation
- Contiguous memory: base, bounds, segmentation
- Using fixed sizes pages with page tables

Challenges with paging
- Extra memory references: avoid with TLB
- Page table size: avoid with multi-level paging, inverted page tables etc.

Larger address spaces: Swapping mechanisms, policies (LRU, Clock)
NEXT STEPS

Next class: New module on Concurrency!