

ETH zürich spcl.inf.ethz.ch @spcl_eth

ADRIAN PERRIG & TORSTEN HOEFLER

Networks and Operating Systems (252-0062-00)

Chapter 6: Demand Paging

Samsung Galaxy Back-door <http://redmine.replicant.us/projects/replicant/wiki/SamsungGalaxyBackdoor> (2014)

This page contains a technical description of the back-door found in Samsung Galaxy devices, including the ones that are shipped with the devices. However, when Replicant is installed on the device, this back-door is not effective: Replicant does not cooperate with back-doors.

Abstract

Samsung Galaxy devices running proprietary Android versions come with a back-door that provides remote access to the data stored on the device. In particular, the proprietary software that is in charge of handling the communications with the modem, using the Samsung IPC protocol, implements a class of requests known as RFS commands, that allows the modem to perform remote I/O operations on the phone's storage. As the modem is running proprietary software, it is likely that it offers over-the-air remote control, that could then be used to issue the incriminated RFS messages and access the phone's file system.

Back-door sample

In order to investigate the back-door and check what it actually lets the modem do, some code was added to the modem kernel driver to make it craft and inject requests using the incriminated messages and check its results.

The following patch: 0001-modem-if-Inject-and-Intercept-RFS-I-O-messages-to-pe.patch (to apply to the SMDK4412 Replicant 4.2 kernel) implements a sample use of the back-door that will:

- open the /data/radio/test file
- read its content
- close the file

This demonstrates that the incriminated software will execute these operations upon modem request. Note that the software implementation appends /efs/root/ to the provided path, but it's fairly simple to escape that path and request any file on the file system (using ../..). Note that the files are opened with the incriminated software's user permissions, which may be root on some devices. On other cases, it runs as an unprivileged user that can still access the user's personal data (/sdcard). Finally, some devices may implement SELinux, which considerably restricts the scope of possible files that the modem can access, including the user's personal data (/sdcard/).

ETH zürich spcl.inf.ethz.ch @spcl_eth

#4 Inverted page table

- **One system-wide table now maps PFN -> VPN**
 - One entry for each real page of memory
 - Contains VPN, and which process owns the page
- **Bounds total size of all page information on machine**
 - Hashing used to locate an entry efficiently
- **Examples: PowerPC, ia64, UltraSPARC**

2

ETH zürich spcl.inf.ethz.ch @spcl_eth

Inverted page table architecture

The diagram illustrates the inverted page table architecture. On the left, a CPU provides a logical address consisting of process ID (pid), page number (p), and displacement (d). This address is used to search a page table. The page table contains entries with pid and p. Once the entry is found, it points to a physical address consisting of index (i) and displacement (d). This physical address is then used to access the physical memory.

3

ETH zürich spcl.inf.ethz.ch @spcl_eth

The need for more bookkeeping

- **Most OSes keep their own translation info**
 - Per-process hierarchical page table (Linux)
 - System wide inverted page table (Mach, MacOS)
- **Why?**
 - Portability
 - Tracking memory objects
 - Software virtual → physical translation
 - Physical → virtual translation

4

ETH zürich spcl.inf.ethz.ch @spcl_eth

Our Small Quiz

- **True or false (raise hand)**
 1. Base (relocation) and limit registers provide a full virtual address space
 2. Base and limit registers provide protection
 3. Segmentation provides a base and limit for each segment
 4. Segmentation suffers from external fragmentation
 5. Segmentation allows libraries to share their code
 6. Segmentation provides linear addressing
 7. Segment tables are set up for each process in the CPU
 8. Segmenting prevents internal fragmentation
 9. Paging prevents internal fragmentation
 10. Protection information is stored at the physical frame
 11. Pages can be shared between processes
 12. The same page may be writeable in proc. A and write protected in proc. B
 13. The same physical address can be referenced through different addresses from (a) two different processes – (b) the same process?
 14. Inverted page tables are faster to search than hierarchical (asymptotically)

5

ETH zürich spcl.inf.ethz.ch @spcl_eth

Today

- **TLB shutdown**
- **Uses for virtual memory**
- **Copy-on-write**
- **Demand paging**
 - Page fault handling
 - Page replacement algorithms
 - Frame allocation policies
 - Thrashing and working set
- **Book: OSPP Sections 9.5, 9.7 (all of 9 as refresh)**
 - As always, the book does not cover 100%!

TLB shutdown

TLB management

- Recall: the TLB is a **cache**.
- Machines have many MMUs on many cores
⇒ many TLBs
- Problem: TLBs should be coherent. Why?
 - Security problem if mappings change
 - E.g., when memory is reused

TLB management

	Process ID	VPN	PPN	acces
Core 1	0	0x0053	0x03	r/w
TLB:	1	0x20f8	0x12	r/w
Core 2	0	0x0053	0x03	r/w
TLB:	1	0x0001	0x05	read
Core 3	0	0x20f8	0x12	r/w
TLB:	1	0x0001	0x05	read

TLB management

	Process ID	VPN	PPN	acces
Core 1	0	0x0053	0x03	r/w
TLB:	1	0x20f8	0x12	r/w
Core 2	0	0x0053	0x03	r/w
TLB:	1	0x0001	0x05	read
Core 3	0	0x20f8	0x12	r/w
TLB:	1	0x0001	0x05	read

Change to read only

TLB management

	Process ID	VPN	PPN	acces
Core 1	0	0x0053	0x03	r/w
TLB:	1	0x20f8	0x12	r/w
Core 2	0	0x0053	0x03	r/w
TLB:	1	0x0001	0x05	read
Core 3	0	0x20f8	0x12	r/w
TLB:	1	0x0001	0x05	read

Change to read only

TLB management

	Process ID	VPN	PPN	acces
Core 1	0	0x0053	0x03	r/w
TLB:	1	0x20f8	0x12	r/w
Core 2	0	0x0053	0x03	r/w
TLB:	1	0x0001	0x05	read
Core 3	0	0x20f8	0x12	r/w
TLB:	1	0x0001	0x05	read

Change to read only

Process 0 on core 1 can only continue once shutdown is complete!

Keeping TLBs coherent

1. **Hardware TLB coherence**
 - Integrate TLB mgmt with cache coherence
 - Invalidate TLB entry when PTE memory changes
 - Rarely implemented
2. **Virtual caches**
 - Required cache flush / invalidate will take care of the TLB
 - High context switch cost!
⇒ Most processors use physical (last level) caches
3. **Software TLB shutdown**
 - Most common
 - OS on one core notifies all other cores - typically an IPI
 - Each core provides local invalidation
4. **Hardware shutdown instructions**
 - Broadcast special address access on the bus
 - Interpreted as TLB shutdown rather than cache coherence message
 - E.g., PowerPC architecture

Summary/recap: virtual memory

- **User logical memory ≠ physical memory.**
 - Only part of the program must be in RAM for execution
⇒ Logical address space can be larger than physical address space
 - Address spaces can be shared by several processes
 - More efficient process creation
- **Virtualize memory using software+hardware**

The many uses of address translation

- Process isolation
- IPC
- Shared code segments
- Program initialization
- Efficient dynamic memory allocation
- Cache management
- Program debugging
- Efficient I/O
- Memory mapped files
- Virtual memory
- Checkpoint and restart
- Persistent data structures
- Process migration
- Information flow control
- Distributed shared memory and many more ...

Copy-on-write (COW)



Photo by Josh Hammerling

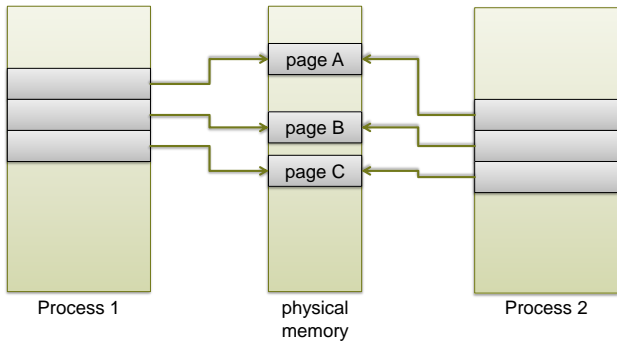
Recall fork ()

- Can be expensive to create a complete copy of the process' address space
 - Especially just to do `exec ()`!
- `vfork ()`: shares address space, doesn't copy
 - Fast
 - Dangerous – two writers to same heap
- **Better: only copy when you know something is going to get written**
 - Requires MMU/memory virtualization

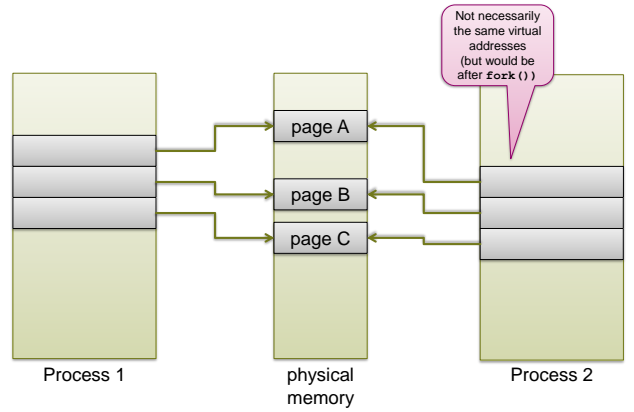
Copy-on-write

- **COW** allows both parent and child processes to initially *share* the same pages in memory
- If either process modifies a shared page, only then is the page copied
- **COW** allows more efficient process creation as only modified pages are copied
- Free pages are allocated from a **pool** of zeroed-out pages

Example: processes sharing an area of memory



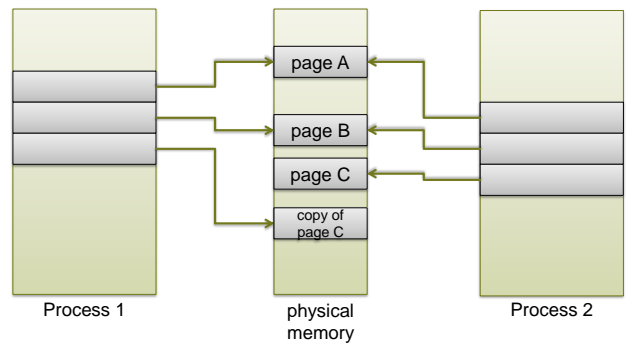
Example: processes sharing an area of memory



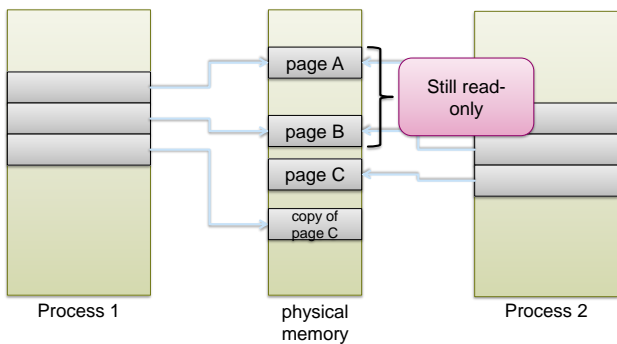
How does it work?

- Initially mark all pages as read-only
- Either process writes ⇒ page fault
 - Fault handler allocates new frame
 - Makes copy of page in new frame
 - Maps each copy into resp. processes writeable
- Only modified pages are copied
 - Less memory usage, more sharing
 - Cost is page fault for each mutated page

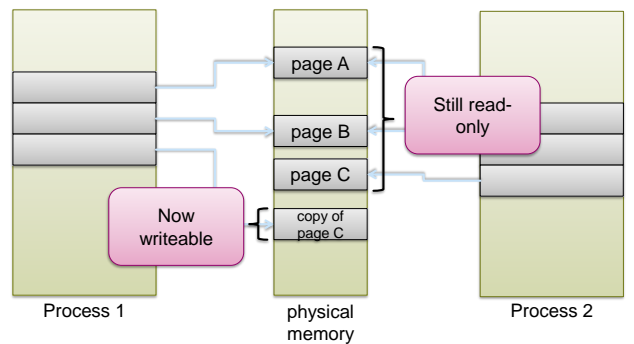
After process 1 writes to page C



After process 1 writes to page C



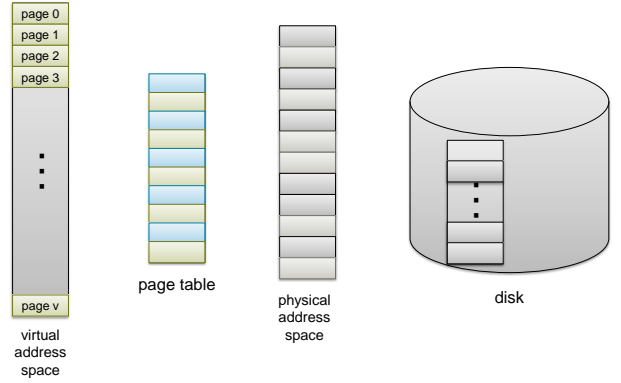
After process 1 writes to page C



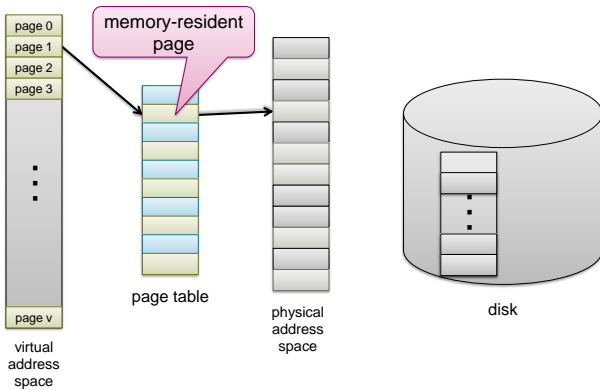
General principle

- Mark a VPN as invalid or read-only
⇒ trap indicates attempt to read or write
- On a page fault, change mappings somehow
- Restart instruction, as if nothing had happened
- General: allows **emulation** of memory as well as **multiplexing**.
 - E.g. on-demand zero-filling of pages
 - And...

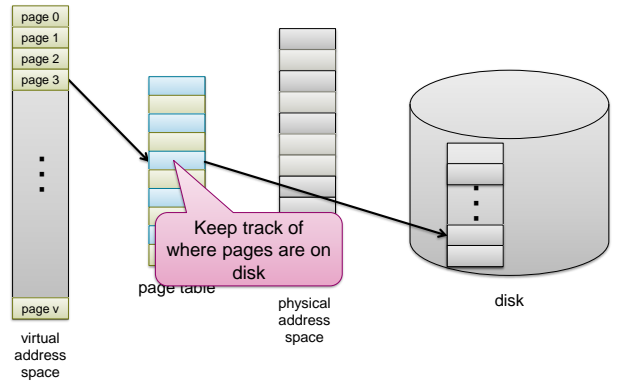
Paging concepts



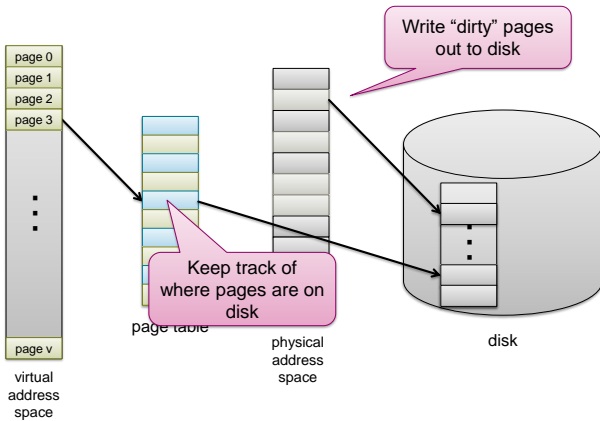
Paging concepts



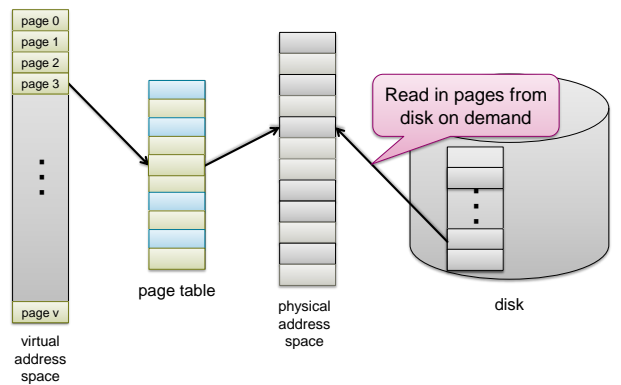
Paging concepts



Paging concepts



Paging concepts



Demand paging

- **Bring a page into memory only when it is needed**
 - Less I/O needed
 - Less memory needed
 - Faster response
 - More users
- **Turns RAM into a *cache* for processes on *disk*!**

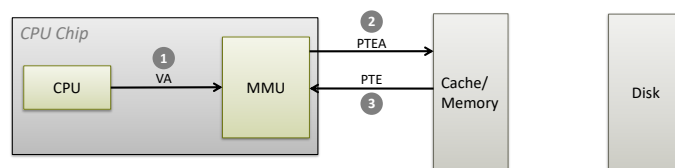
Demand paging

- **Page needed \Rightarrow reference (load or store) to it**
 - invalid reference \Rightarrow abort
 - not-in-memory \Rightarrow bring to memory
- **Lazy swapper – never swaps a page into memory unless page will be needed**
 - Swapper that deals with pages is a pager
 - Can do this with segments, but more complex ... or whole processes
- **Strict demand paging: only page in when referenced**

Page fault

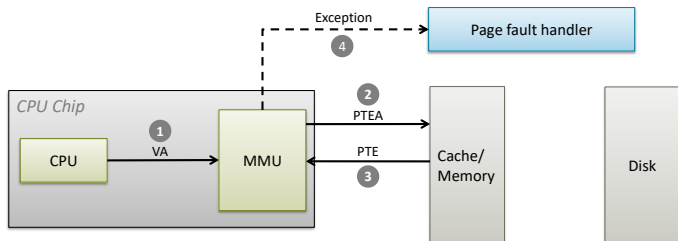
- **If there is a reference to a page, first reference to that page will trap to operating system:**
page fault
- 1. **Operating system looks at another table to decide:**
 - Invalid reference \Rightarrow abort
 - Just not in memory
- 2. **Get empty frame**
- 3. **Swap page into frame**
- 4. **Reset tables**
- 5. **Set valid bit \checkmark**
- 6. **Restart the instruction that caused the page fault**

Recall: handling a page fault



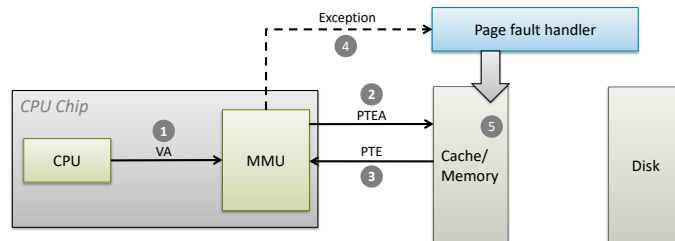
- 1) Processor sends virtual address to MMU
- 2-3) MMU fetches PTE from page table in memory

Recall: handling a page fault



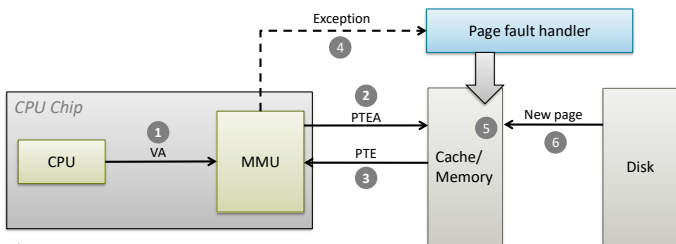
- 1) Processor sends virtual address to MMU
- 2-3) MMU fetches PTE from page table in memory
- 4) Valid bit is zero, so MMU triggers page fault exception

Recall: handling a page fault



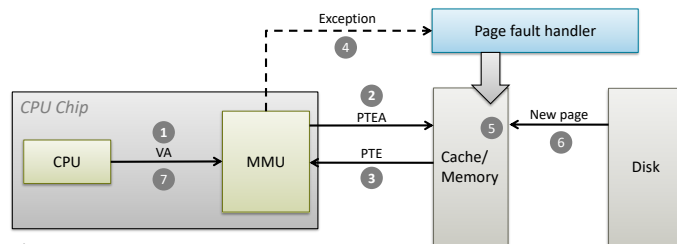
- 1) Processor sends virtual address to MMU
- 2-3) MMU fetches PTE from page table in memory
- 4) Valid bit is zero, so MMU triggers page fault exception
- 5) Handler finds a frame to use for missing page

Recall: handling a page fault



- 1) Processor sends virtual address to MMU
- 2-3) MMU fetches PTE from page table in memory
- 4) Valid bit is zero, so MMU triggers page fault exception
- 5) Handler finds a frame to use for missing page
- 6) Handler pages in new page and updates PTE in memory

Recall: handling a page fault



- 1) Processor sends virtual address to MMU
- 2-3) MMU fetches PTE from page table in memory
- 4) Valid bit is zero, so MMU triggers page fault exception
- 5) Handler finds a frame to use for missing page
- 6) Handler pages in new page and updates PTE in memory
- 7) Handler returns to original process, restarting faulting instruction

Performance of demand paging

- **Page fault rate** $0 \leq p \leq 1.0$
 - if $p = 0$: no page faults
 - if $p = 1$: every reference is a fault
- **Effective Access Time (EAT)**

$$\text{EAT} = (1 - p) \times \text{memory access} + p (\text{page fault overhead} + \text{swap page out} + \text{swap page in} + \text{restart overhead})$$

Demand paging example

- **Memory access time = 50 nanoseconds**
- **Average page-fault service time = 4 milliseconds**
- **EAT = $(1 - p) \times 50 + p (4 \text{ milliseconds})$**

$$= (1 - p) \times 50 + p \times 4,000,000 = 50 + p \times 3,999,950$$
- **If one access out of 1,000 causes a page fault, then EAT = 4 microseconds. This is a slowdown by a factor of 80!!**

Page Replacement



Photo: Urs Flueeler, source: <https://aeon.co/essays/swiss-flying-cows-is-this-the-future>

What happens if there is no free frame?

- **Page replacement** – find “little used” resident page to discard or write to disk
 - “victim page”
 - needs selection algorithm
 - performance – want an algorithm which will result in minimum number of page faults
- **Same page may be brought into memory several times**

Page replacement

- Try to pick a victim page which won't be referenced in the future
 - Various heuristics – but ultimately it's a guess
- Use “modify” bit on PTE
 - Don't write “clean” (unmodified) page to disk
 - Try to pick “clean” pages over “dirty” ones (save a disk write)

Page replacement

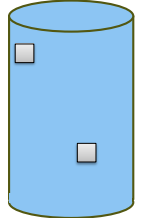
frame valid

0	i
f	v

Page table



Physical memory



Page replacement

frame valid

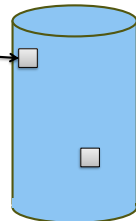
0	i
f	v

Page table



Physical memory

1. Swap victim page to disk

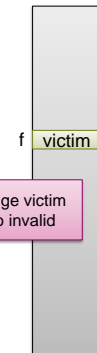


Page replacement

frame valid

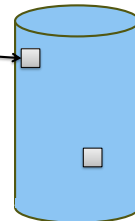
0	i
0	i

Page table



Physical memory

1. Swap victim page to disk



2. Change victim PTE to invalid

Page replacement

frame valid

0	i
0	i

Page table



Physical memory

3. Load desired page in from disk



Page replacement

frame valid

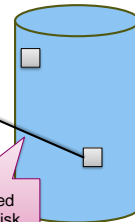
f	v
0	i

Page table



Physical memory

3. Load desired page in from disk

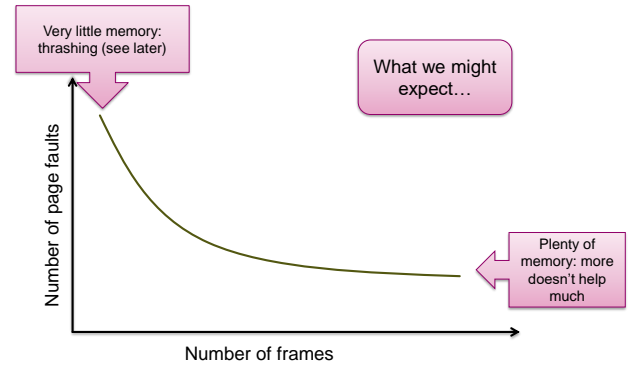


4. Change fault PTE to valid

Page replacement algorithms

- Want lowest page-fault rate
- Evaluate algorithm by running it on a particular string of memory references (**reference string**) and computing the number of page faults on that string
- E.g.
7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1, 7, 0, 1

Page faults vs. number of frames



FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page
frames:

FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page
frames: 7

FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page
frames: 7 7
0

FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page
frames: 7 7 7
0 0
1



FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page	7	7	7	2
frames:	0	0	0	
			1	1



FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page	7	7	7	2	2
frames:	0	0	0		3
			1	1	1



FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page	7	7	7	2	2	2	4	4	4	0
frames:	0	0	0		3	3	3	2	2	2
			1	1	1	0	0	0	3	3



FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page	7	7	7	2	2	2	4	4	4	0	0	0
frames:	0	0	0		3	3	3	2	2	2	1	1
			1	1	1	0	0	0	3	3	3	2



FIFO (First-In-First-Out) page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

page	7	7	7	2	2	2	4	4	4	0	0	0	7	7	7
frames:	0	0	0		3	3	3	2	2	2	1	1	1	0	0
			1	1	1	0	0	0	3	3	3	2	2	2	1

Here, 15 page faults.



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5

- 3 frames (3 pages can be in memory):

1	1	1
	2	2
		3



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5

- 3 frames (3 pages can be in memory):

1	1	1	4
	2	2	2
		3	3



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5

- 3 frames (3 pages can be in memory):

1	1	1	4	4	4	5
	2	2	2	1	1	1
		3	3	3	2	2



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5

- 3 frames (3 pages can be in memory):

1	1	1	4	4	4	5	5
	2	2	2	1	1	1	3
		3	3	3	2	2	2



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5

- 3 frames (3 pages can be in memory):

1	1	1	4	4	4	5	5	5
	2	2	2	1	1	1	3	3
		3	3	3	2	2	2	4

9 page faults



More memory is better?

Reference string: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5

- 3 frames (3 pages can be in memory):

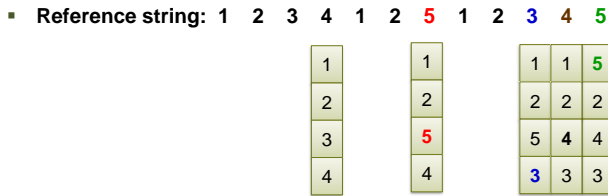
1	1	1	4	4	4	5	5	5
	2	2	2	1	1	1	3	3
		3	3	3	2	2	2	4

9 page faults

- 4 frames:

1	1	1	1
	2	2	2
		3	3
			4

Least Recently Used (LRU) algorithm



- Counter implementation
 - Every page entry has a counter; every time page is referenced through this entry, copy the clock into the counter
 - When a page needs to be changed, look at the counters to determine which are to change

LRU page replacement

reference string: 7 0 1 2 0 3 0 4 2 3 0 3 2 1 2 0 1 7 0 1

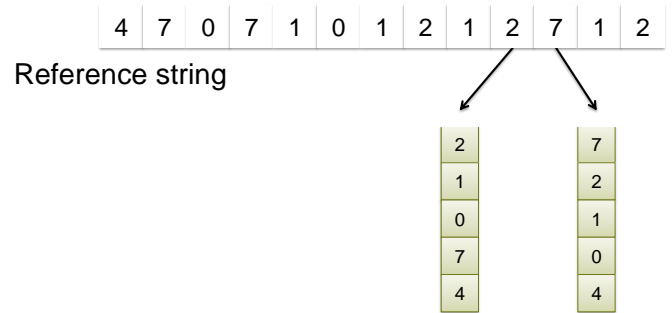


Here, 12 page faults.

LRU stack algorithm

- Stack implementation – keep a stack of page numbers in a double link form:
 - Page referenced:
 - move it to the top
 - requires 6 pointers to be changed
 - No search for replacement
- General term: *stack algorithms*
 - Have property that adding frames always reduces page faults (no Belady's Anomaly)

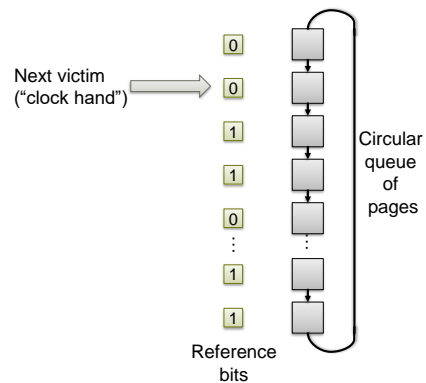
Use a stack to record most recent page references



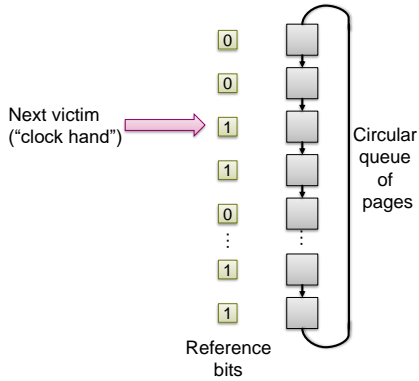
LRU approximation algorithms

- Reference bit
 - With each page associate a bit, initially = 0
 - When page is referenced bit set to 1
 - Replace a page which is 0 (if one exists)
 - We do not know the order, however
- Second chance
 - Need reference bit
 - Clock replacement
 - If page to be replaced (in clock order) has reference bit = 1 then:
 - set reference bit 0
 - leave page in memory
 - replace next page (in clock order), subject to same rules

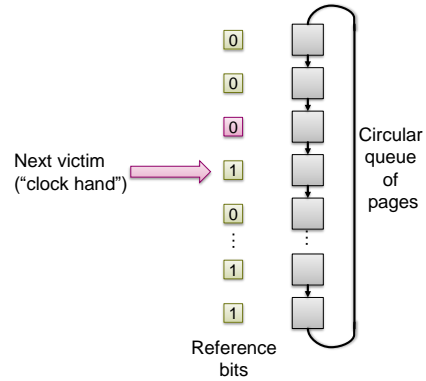
Second-chance (clock) page replacement algorithm



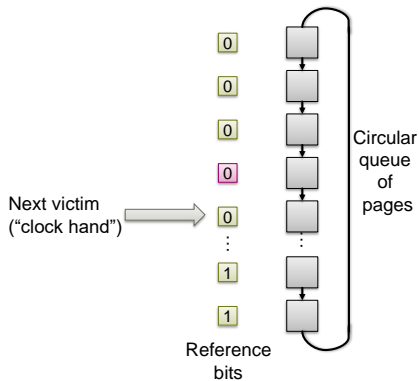
Second-chance (clock) page replacement algorithm



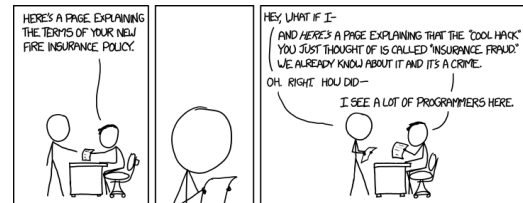
Second-chance (clock) page replacement algorithm



Second-chance (clock) page replacement algorithm



Frame allocation policies (multi-process)



Allocation of frames

- Each process needs minimum number of pages
- Example: IBM 370 – 6 pages to handle SS MOVE instruction:
 - instruction is 6 bytes, might span 2 pages
 - 2 pages to handle from
 - 2 pages to handle to
- Two major allocation schemes
 - fixed allocation
 - priority allocation

Fixed allocation

- Equal allocation
 - all processes get equal share
- Proportional allocation
 - allocate according to the size of process

 $s_i = \text{size of process } p_i$
 $m = 64$
 $S = \sum s_i$
 $s_1 = 10$
 $m = \text{total number of frames}$
 $s_2 = 127$
 $a_i = \text{allocation for } p_i = \frac{s_i}{S} \times m$
 $a_1 = \frac{10}{137} \times 64 \approx 5$
 $a_2 = \frac{127}{137} \times 64 \approx 59$

ETH zürich spcl.inf.ethz.ch
@spcl_eth

Priority allocation

- Proportional allocation scheme
- Using priorities rather than size
- If process P_i generates a page fault, select:
 1. one of its frames, or
 2. frame from a process with lower priority

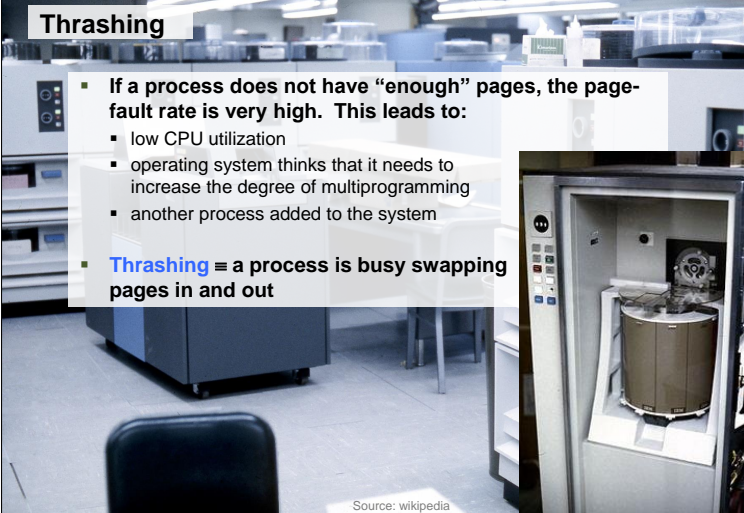
ETH zürich spcl.inf.ethz.ch
@spcl_eth

Global vs. local allocation

- **Global replacement** – process selects a replacement frame from the set of all frames; one process can take a frame from another
- **Local replacement** – each process selects from only its own set of allocated frames

ETH zürich spcl.inf.ethz.ch
@spcl_eth

Thrashing

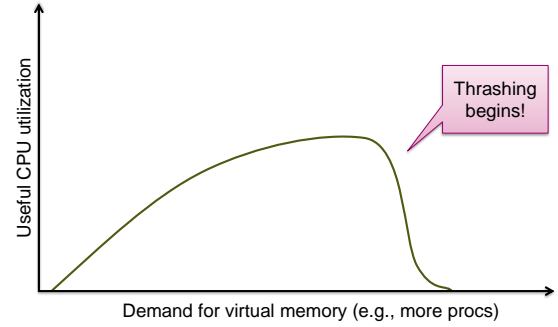


- If a process does not have “enough” pages, the page-fault rate is very high. This leads to:
 - low CPU utilization
 - operating system thinks that it needs to increase the degree of multiprogramming
 - another process added to the system
- **Thrashing** = a process is busy swapping pages in and out

Source: wikipedia

ETH zürich spcl.inf.ethz.ch
@spcl_eth

Thrashing



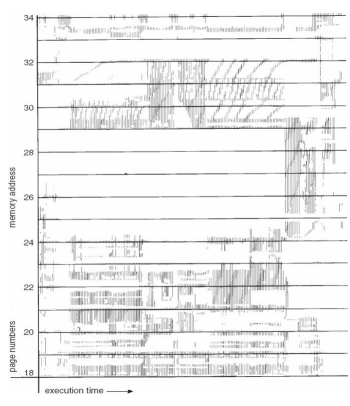
ETH zürich spcl.inf.ethz.ch
@spcl_eth

Demand paging and thrashing

- **Why does demand paging work?**
 - **Locality model**
 - Process migrates from one locality to another
 - Localities may overlap
- **Why does thrashing occur?**
 - Σ size of localities > total memory size

ETH zürich spcl.inf.ethz.ch
@spcl_eth

Locality in a memory reference pattern



Working-set model

- Δ \equiv working-set window
 \equiv a fixed maximum number of page references
 - Example: 10,000 instruction
- WSS_i (working set of process P_i) = total number of pages referenced in the most recent Δ (varies in time)
 - Δ too small \Rightarrow will not encompass entire locality
 - Δ too large \Rightarrow will encompass several localities
 - $\Delta = \infty \Rightarrow$ will encompass entire program

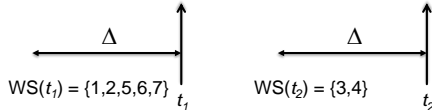
Allocate demand frames

- $D = \sum WSS_i \equiv$ total demand frames
 - Intuition: how much space is really needed
- $D > m \Rightarrow$ Thrashing
- Policy: if $D > m$, suspend some processes

Working-set model

Page reference string:

... 2 6 1 5 7 7 7 5 1 6 2 3 4 1 2 3 4 4 4 3 4 3 4 4 4 1 3 2 3 4 4 4 3 4 4 4 ...



Keeping track of the working set

- Approximate with interval timer + a reference bit
- Example: $\Delta = 10,000$
 - Timer interrupts after every 5,000 time units
 - Keep in memory 2 bits for each page
 - Whenever a timer interrupts shift+copy and sets the values of all reference bits to 0
 - If one of the bits in memory = 1 \Rightarrow page in working set
- Why is this not completely accurate?
 - Hint: Nyquist-Shannon!

Keeping track of the working set

- Approximate with interval timer + a reference bit
- Example: $\Delta = 10,000$
 - Timer interrupts after every 5000 time units
 - Keep in memory 2 bits for each page
 - Whenever a timer interrupts shift+copy and sets the values of all reference bits to 0
 - If one of the bits in memory = 1 \Rightarrow page in working set
- Why is this not completely accurate?
- Improvement = 10 bits and interrupt every 1,000 time units

Page-fault frequency scheme

- Establish "acceptable" page-fault rate
 - If actual rate too low, process loses frame
 - If actual rate too high, process gains frame

